

Introduction to

STATISTICS & DATA ANALYSIS

Roxy Peck • Tom Short • Chris Olsen



Sixth Edition

Index of Applications in Examples and Activities

Act: Activity; Ex: Example

Agriculture

Grape production: Ex 3.6
Tomato yield and planting density: Ex 15.12,
Ex 15.13 (online)

Price of fish: Ex 14.13 (online)
Prices of industrial properties: Ex 14.17, Ex 14.19 (online)
Resumé typos: Ex 3.3
Starting salaries of business school graduates: Ex 16.11,
Ex 16.12 (online)

Biology

Age of a lobster: Ex 5.22
Barking treefrog behavior: Ex 13.8
Bee mating behavior: Ex 3.12, Ex 3.13
Black bear habitat selection: Ex 5.9
Calling behavior of Amazonian frogs: Ex 5.25 (online)
Cannibalism in wolf spiders: Ex 5.23, Ex 5.24 (online)
Charitable behavior of chimpanzees: Ex 9.10, Ex 9.16,
Ex 11.7, Ex 11.14
Chirp rate for crickets: Ex 10.15
Compression strength of wood: Ex 14.10
Distance deer mice will travel for food: Ex 5.7, Ex 5.10,
Ex 5.11
Distinguishing pâté from dog food: Ex 12.3
Dominant and nondominant hands: Act 3.2
Effect of human activity on bears: Ex 5.15, Ex 5.16
Egg weights: Ex 7.31
Head circumference at birth: Ex 4.19
Hitchhiker's thumb: Ex 6.17
Loon chick survival factors: Ex 5.19, Ex 5.20
Predator inspection in guppies: Ex 6.18
Recognizing your roommate's scent: Ex 7.18
Reflexes with dominant and nondominant hands: Act 11.2
Salamander behavior: Ex 5.21
Scorpionfly courtship: Ex 8.4
Shark length and jaw width: Ex 13.10, Ex 13.11 (online)

College Life

Academic success of college sophomores: Ex 14.1
Advantages of multiple SAT scores in college admissions:
Act 8.3
Asking questions in seminar class: Ex 6.5
Back-to-college spending: Ex 3.7
College attendance: Ex 3.15, Ex 10.12
College choice do-over: Ex 1.4
Comparing job offers: Ex 4.18
Cost of textbooks: Ex 9.11
Detecting plagiarism: Ex 10.9
Graduation rates: Ex 1.9, Ex 5.13, Ex 13.5
Graduation rates at small colleges: Ex 14.6, Ex 14.7,
Ex 14.8, Ex 14.9
How safe are college campuses?: Ex 1.5
Impact of internet and television use on college student
reading habits: Ex 9.2
Importance of college education: Ex 9.4, Ex 9.14
Internet use by college students: Ex 9.1
Math SAT score distribution: Ex 3.14
Money spent on textbooks: Ex 8.1
Predicting graduation rates: Ex 5.13
STEM college students: Ex 8.7
Student college housing choices: Ex 6.11
Student debt on the rise: Ex 3.17
Students with jumper cables: Ex 7.22
Study habits of college seniors: Ex 3.5
Time required to complete registration: Ex 7.29
Tuition at public universities: Ex 3.9
Visits to class web site: Ex 4.3, Ex 4.4

Business and Economics

Application processing times: Ex 7.8
Book sales: Ex 7.1
Business of baseball: Ex 13.9
Cable services: Ex 6.22
Christmas Price Index: Ex 3.23
Comparing gasoline additives: Ex 2.10
Cost of Big Macs: Ex 4.7, Ex 4.8, Ex 4.12
Cost of energy bars: Ex 14.11
Cost of residential air-conditioning: Ex 15.8,
Ex 15.9 (online)
Daily wasted time at work: Ex 10.14, Ex 10.26
Education level and income: Ex 3.24
Express mail volume: Ex 7.34
Licensing exam attempts: Ex 7.9
Mortgage choices: Ex 6.16
Predicting house prices: Ex 14.5

Demography and Population Characteristics

County population sizes: Ex 4.2
Head circumferences: Act 1.1
Heights of college athletes: Ex 1.1
Heights of mothers: Ex 4.17
Household phone service: Ex 3.8
Median ages in 2030: Ex 3.10
Newborn birth weights: Ex 7.27
Two-child families: Ex 6.14
Voting registration: Ex 10.20
Women's heights and number of siblings: Act 13.1

Education and Child Development

Childcare for preschoolers: Ex 4.15
College plans of high school seniors: Ex 7.4
Combining exam scores: Ex 7.16
Hand gestures in learning: Ex 2.7, Ex 2.9
IQ scores: Ex 4.16, Ex 7.28
Is education worth the cost?: Ex 3.1
Note-taking methods: Ex 2.4
Predictors of writing competence: Ex 14.4
School enrollment in Northern and Central Africa: Ex 3.11
Standardized test scores: Ex 4.14, Ex 10.16
Students' knowledge of geography: Act 3.1

Durable press rating of cotton fabric: Ex 14.18 (online)
Engineering stress test: Ex 7.3
Ergonomic characteristics of stool designs: Ex 15.10,
Ex 15.11 (online)
Garbage truck processing times: Ex 7.30
GFI switches: Ex 6.12
Lifetime of compact fluorescent lightbulbs: Ex 10.2
On-time package delivery: Ex 10.18
Paint flaws: Ex 7.6
Smart phone warranties: Ex 6.24
Strength of bark board: Ex 16.4 (online)
Testing for flaws: Ex 7.11, Ex 7.12

Environmental Science

Bison of Yellowstone Park: Ex 13.4
Cosmic radiation: Ex 9.7
Lead in tap water: Ex 10.7
Rainfall frequency distributions for Albuquerque: Ex 3.19
River water velocity and distance from shore: Ex 5.17
Soil and sediment characteristics: Ex 14.12, Ex 14.14,
Ex 14.16 (online)
Water conservation: Ex 10.10
Water quality: Ex 1.2

Marketing and Consumer Behavior

Budgets and tracking spending: Ex 7.21
Car choices: Ex 6.10
Energy efficient refrigerators: Ex 7.5
High-pressure sales tactics: Ex 16.13 (online)
Impact of food labels: Ex 10.8
Satisfaction with cell phone service: Ex 4.6

Food Science

Calorie consumption at fast food restaurants: Ex 2.2
Effect of exercise on food intake: Ex 4.1
Fat content of hot dogs: Ex 8.6

Medical Science

Affect of long work hours on sleep: Ex 11.10
Anti-clotting medications after hip or knee surgery: Ex 11.15
Apgar scores: Ex 7.10, Ex 7.13
Blood platelet volume: Ex 8.2
Blood pressure and kidney disease: Ex 16.5 (online)
Blue light exposure and blood glucose level: Ex 11.12
Body mass index (BMI): Ex 7.14
Chronic airflow obstruction: Ex 16.9 (online)
Contracting hepatitis from blood transfusion: Ex 8.8, Ex 8.9
Cooling treatment after oxygen deprivation in newborns:
Ex 2.6
Diagnosing tuberculosis: Ex 6.15
Drive-through medicine: Ex 9.8
Early detection of lung cancer: Ex 10.6
Effects of ethanol on sleep time: Ex 15.6
Effect of school start time on sleep patterns: Ex 3.16
Evaluating disease treatments: Ex 10.3
Exercise and sleep quality: Ex 12.8
Facial expression and self-reported pain level: Ex 12.7
Growth hormone levels and diabetes: Ex 16.10 (online)
Heart attacks in high-rise buildings: Ex 12.4, Ex 12.5
Hip-to-waist ratio and risk of heart attack: Ex 14.2
Hormones and body fat: Ex 15.4, Ex 15.5
Lead exposure and brain volume: Ex 5.12
Liver injuries in newborns: Ex 9.15
Lyme disease: Ex 6.27
Markers for kidney disease: Ex 7.33
Maternal age and baby's birth weight: Ex 13.2
Medical errors: Ex 6.9
Oxytocin nasal spray and social interaction: Ex 11.16
Parental smoking and infant health: Ex 16.2,
Ex 16.3 (online)

Leisure and Popular Culture

Babies on social media: Ex 9.5
Car preferences: Ex 6.1
Do U Txt?: Ex 1.6
Facebook and academic performance: Ex 11.1
iPod shuffles: Ex 7.7
Jeopardy! nerds: Ex 12.1, Ex 12.2
Life insurance for cartoon characters: Ex 3.2
Number of trials required to complete game: Ex 7.2
Probability a Hershey's Kiss will land on its base: Act 6.1
Selecting cards: Ex 6.20
Selection of contest winners: Ex 6.7
Tossing a coin: Ex 6.8
Video game performance: Ex 4.9, Ex 4.10, Ex 6.2, Ex 6.4
Word cloud representation of a document: Ex 11.5,
Ex 11.6, Ex 11.8

Manufacturing and Industry

Bottled soda volumes: Ex 8.5
Computer configurations: Ex 6.19
Computer sales: Ex 7.19
Corrosion of underground pipe coatings: Ex 15.14 (online)
Customer hold time: Ex 10.17

Pediatric tracheal tube: Ex 13.7
Platelet volume and heart attack risk: Ex 15.1, Ex 15.2, Ex 15.3
Pomegranate juice and tumor growth: Ex 5.4, Ex 5.5
Premature births: Ex 7.35
Sleep duration and blood leptin level: Ex 13.12 (online)
Slowing the growth rate of tumors: Ex 10.5
Sniffing out cancer: Ex 9.6
Surviving a heart attack: Ex 6.13
Time perception and nicotine withdrawal: Ex 10.13
Treating dyskinesia: Ex 16.8 (online)
Treatment for acute mountain sickness: Act 2.5
Video games and pain management: Act 2.4
Vitamin B12 levels in human blood: Ex 16.7 (online)
Waiting time for hip surgery: Ex 9.9

Physical Sciences

Electromagnetic radiation: Ex 5.18
Rainfall data: Ex 7.32
Wind chill factor: Ex 14.3

Politics and Public Policy

Fair hiring practices: Ex 6.29
Predicting election outcomes from facial appearance: Ex 13.3, Ex 13.6
Recall petition signatures: Act 9.3
Requests for building permits: Ex 6.31
Scientists and nonscientists: Ex 3.4

Psychology, Sociology, and Social Issues

Color and perceived taste: Act 12.2, Ex 15.7
Estimating sizes: Act 1.2
Extrasensory perception: Ex 6.33
Face-to-height ratio: Ex 5.6
Facial cues and trustworthiness: Ex 5.1, Ex 5.6
The “Freshman 15”: Ex 11.4, Ex 11.13
Golden rectangles: Ex 4.11
Hand-holding couples: Ex 6.30
Internet addiction: Ex 6.28
Morality in the morning: Ex 2.5
One-boy family planning: Ex 6.32
Stroop effect: Act 2.2
Weight regained proportions for three follow-up methods: Ex 12.6

Public Health and Safety

Age and flexibility: Act 5.1
Chew more, eat less?: Ex 1.3

Crime scene investigators: Ex 5.14
Effect of cell phone distraction: Ex 2.8
Effects of McDonald’s hamburger sales: Act 2.3
Emotional health and work environment: Ex 3.21
Exercise on the rise: Ex 3.22
Fitness trackers and weight loss: Ex 11.3
Safety of bicycle helmets: Ex 5.2
Salmonella in restaurant eggs: Act 7.2
Teenage driver citations and traffic school: Ex 6.23

Sports

Age and marathon times: Ex 5.3
Calling a toss at a football game: Ex 6.6
Fairness of Euro coin-flipping in European sports: Act 6.2
Helium-filled footballs: Act 11.1
“Hot hand” in basketball: Act 6.3
NBA player salaries: Ex 4.5, Ex 4.13
Olympic figure skating: Ex 3.20
Racing starts in competitive swimming: Ex 16.6 (online)
Soccer goal keepers, action bias among: Ex 6.26
Tennis ball diameters: Ex 10.1
Time to first goal in hockey: Ex 8.3
Wrestlers’ weight loss by headstand: Ex 13.1

Surveys and Opinion Polls

Are cell phone users different?: Ex 2.1
Cell phone usage: Ex 11.11
Designing a sampling plan: Facebook friending: Act 2.1
Jump distance of frogs: Ex 11.2
Love for cell phones: Ex 10.11, Ex 10.21, Ex 10.24
Vaccination coverage: Ex 10.22, Ex 10.25
Wedding vows: Ex 7.20

Transportation

Accidents by bus drivers: Ex 3.18
Airborne times for San Francisco to Washington, D.C. flight: Ex 9.3
Airline luggage weights: Ex 7.17
Airline passenger weights: Act 4.2
Electric cars: Ex 11.9
Freeway traffic: Ex 7.15
Fuel efficiency of automobiles: Ex 16.1 (online)
Lost airline luggage: Ex 6.25
Motorcycle helmets: Ex 1.7, Ex 1.8
On-time airline flights: Ex 10.4
Predicting transit times: Ex 14.15 (online)
Turning directions on freeway off-ramp: Ex 6.3

EDITION

6

Introduction to Statistics and Data Analysis



Roxy Peck

California Polytechnic State University, San Luis Obispo

Tom Short

West Chester University of Pennsylvania

Chris Olsen

Grinnell College



Australia • Brazil • Mexico • Singapore • United Kingdom • United States

**Introduction to Statistics and Data Analysis,
Sixth Edition**
Roxy Peck, Tom Short, Chris Olsen

Product Director: Mark Santee
Product Manager: Catherine Van Der Laan
Product Assistant: Amanda Rose
Marketing Manager: Mike Saver
Learning Designer: Elinor Gregory
Subject Matter Expert: Morgan Johnson
Content Managers: Brendan Killion and Abby DeVeuve
Manufacturing Planner: Doug Bertke
IP Analyst: Reba Frederics
IP Project Manager: Carly Belcher
Art Director: Vernon T. Boes
Design and Production Services/Compositor:
MPS Limited
Cover Image: Tawatchai Prakobit/EyeEm/
Getty Images

© 2020, 2016, 2012 Cengage Learning, Inc.

Unless otherwise noted, all content is © Cengage.

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced or distributed in any form or by any means, except as permitted by U.S. copyright law, without the prior written permission of the copyright owner.

For product information and technology assistance, contact us at
Cengage Customer & Sales Support, 1-800-354-9706
or support.cengage.com.

For permission to use material from this text or product, submit all requests online at www.cengage.com/permissions.

Library of Congress Control Number: 2018949851

Student Edition:

ISBN: 978-1-337-79361-2

Loose-leaf Edition:

ISBN: 978-1-337-79432-9

Annotated Instructor's Edition:

ISBN: 978-1-337-79415-2

Cengage

20 Channel Street

Boston, MA 02210

USA

Cengage is a leading provider of customized learning solutions with employees residing in nearly 40 different countries and sales in more than 125 countries around the world. Find your local representative at: www.cengage.com.

Cengage products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage platforms and services, register or access your online learning solution, or purchase materials for your course, visit www.cengage.com.

To Lygia and Kyle

Roxy Peck

To my children: Bob, Kathy, Peter, and Kellen

Tom Short

To my wife, Sally, and my daughter, Anna

Chris Olsen

Author Bios



ROXY PECK is Emerita Associate Dean of the College of Science and Mathematics and Professor of Statistics Emerita at California Polytechnic State University, San Luis Obispo. A faculty member at Cal Poly from 1979 until 2009, Roxy

served for 6 years as Chair of the Statistics Department before becoming Associate Dean, a position she held for 13 years. She received an M.S. in Mathematics and a Ph.D. in Applied Statistics from the University of California, Riverside. Roxy is nationally known in the area of statistics education, and she was presented with the Lifetime Achievement Award in Statistics Education at the U.S. Conference on Teaching Statistics in 2009. In 2003 she received the American Statistical Association's Founder's Award, recognizing her contributions to K–12 and undergraduate statistics education. She is a Fellow of the American Statistical Association and an elected member of the International Statistics Institute. Roxy served for 5 years as the Chief Reader for the Advanced Placement (AP®) Statistics Exam and has chaired the American Statistical Association's Joint Committee with the National Council of Teachers of Mathematics on Curriculum in Statistics and Probability for Grades K–12 and the Section on Statistics Education. In addition to her texts in introductory statistics, Roxy is also co-editor of *Statistical Case Studies: A Collaboration Between Academe and Industry* and a member of the editorial board for *Statistics: A Guide to the Unknown*, 4th edition. Outside the classroom, Roxy likes to travel and spends her spare time reading mystery novels. She also collects Navajo rugs and heads to Arizona and New Mexico whenever she can find the time.



TOM SHORT is an Associate Professor in the Statistics Program within the Department of Mathematics at West Chester University of Pennsylvania. He previously held faculty positions at Villanova University, Indiana University of Pennsylvania, and John Carroll University. He is a Fellow of the American Statistical Association and received the 2005 Mu Sigma Rho Statistics Education Award. Tom is part of the leadership team for readings of the Advanced Placement® Statistics Exam, and was a

member of the AP® Statistics Development Committee. He has also served on the Board of Directors of the American Statistical Association. Tom treasures the time he shares with his children and the many adventures experienced with his wife, Darlene.



CHRIS OLSEN taught statistics in Cedar Rapids, Iowa, for over 25 years, and at Cornell College and Grinnell College. Chris is a past member (twice!) of the Advanced Placement® Statistics Test Development Committee and has been a table leader and question leader at the AP® Statistics reading for 11 years. He is a long-time consultant to the College Board, and Chris has led workshops and institutes for AP® Statistics teachers in the United States and internationally. Chris was the Iowa recipient of the Presidential Award for Excellence in Science and Mathematics Teaching in 1986, a regional awardee of the IBM Computer Teacher of the Year in 1988, and received the Siemens Award for Advanced Placement in mathematics in 1999. Chris is a frequent contributor to and is moderator of the AP® Statistics Teacher Community online. He is currently a member of the editorial board of *Teaching Statistics*. Chris graduated from Iowa State University with a major in mathematics. While acquiring graduate degrees at the University of Iowa, he concentrated on statistics, computer programming and psychometrics. In his spare time he enjoys reading and hiking. He and his wife have a daughter, Anna, a Caltech graduate in Civil Engineering. She is a risk modeler at RMS, the world's leading catastrophe risk modeling company.

Brief Contents

- 1** The Role of Statistics and the Data Analysis Process 1
 - 2** Collecting Data Sensibly 28
 - 3** Graphical Methods for Describing Data 77
 - 4** Numerical Methods for Describing Data 148
 - 5** Summarizing Bivariate Data 197
 - 6** Probability 277
 - 7** Random Variables and Probability Distributions 343
 - 8** Sampling Variability and Sampling Distributions 427
 - 9** Estimation Using a Single Sample 453
 - 10** Hypothesis Testing Using a Single Sample 507
 - 11** Comparing Two Populations or Treatments 576
 - 12** The Analysis of Categorical Data and Goodness-of-Fit Tests 654
 - 13** Simple Linear Regression and Correlation: Inferential Methods 689
 - 14** Multiple Regression Analysis 730
 - 15** Analysis of Variance 759
 - 16** Nonparametric (Distribution-Free) Statistical Methods 16-1
- Appendix: Statistical Tables 785
- Answers to Selected Odd-Numbered Exercises 805
- Index 845

Sections and/or chapter numbers shaded in color can be found at
www.cengage.com

Contents

CHAPTER 1 The Role of Statistics and the Data Analysis Process 1

- 1.1** Why Study Statistics? 2
- 1.2** The Nature and Role of Variability 3
- 1.3** Statistics and the Data Analysis Process 5
- 1.4** Types of Data and Some Simple Graphical Displays 9
- Activity 1.1** Head Sizes: Understanding Variability 22
- Activity 1.2** Estimating Shape Sizes 23
- Activity 1.3** A Meaningful Paragraph 24
- Summary: Key Concepts and Formulas 24
- Chapter Review 24
- Technology Notes 26

CHAPTER 2 Collecting Data Sensibly 28

- 2.1** Statistical Studies: Observation and Experimentation 29
- 2.2** Sampling 34
- 2.3** Simple Comparative Experiments 45
- 2.4** More on Experimental Design 60
- 2.5** Interpreting and Communicating the Results of Statistical Analyses 65
- Activity 2.1** Facebook Friending 68
- Activity 2.2** An Experiment to Test for the Stroop Effect 68
- Activity 2.3** McDonald's and the Next 100 Billion Burgers 69
- Activity 2.4** Video Games and Pain Management 69
- Activity 2.5** Be Careful with Random Assignment! 70
- Summary: Key Concepts and Formulas 70
- Chapter Review 71
- Technology Notes 73

See Chapter 2 online materials for More on Observational Studies: Designing Surveys.

CHAPTER 3 Graphical Methods for Describing Data 77

- 3.1** Displaying Categorical Data: Comparative Bar Charts and Pie Charts 78
- 3.2** Displaying Numerical Data: Stem-and-Leaf Displays 88
- 3.3** Displaying Numerical Data: Frequency Distributions and Histograms 97
- 3.4** Displaying Bivariate Numerical Data 116
- 3.5** Interpreting and Communicating the Results of Statistical Analyses 125
- Activity 3.1** Locating States 134
- Activity 3.2** Bean Counters! 134
- Summary: Key Concepts and Formulas 135
- Chapter Review 135

- Technology Notes 139
Cumulative Review Exercises 144

CHAPTER 4 Numerical Methods for Describing Data 148

- 4.1** Describing the Center of a Data Set 149
4.2 Describing Variability in a Data Set 159
4.3 Summarizing a Data Set: Boxplots 168
4.4 Interpreting Center and Variability: Chebyshev's Rule, the Empirical Rule, and z Scores 175
4.5 Interpreting and Communicating the Results of Statistical Analyses 183
Activity 4.1 Collecting and Summarizing Numerical Data 188
Activity 4.2 Airline Passenger Weights 188
Activity 4.3 Boxplot Shapes 188
Summary: Key Concepts and Formulas 189
Chapter Review 189
Technology Notes 191

CHAPTER 5 Summarizing Bivariate Data 197

- 5.1** Correlation 198
5.2 Linear Regression: Fitting a Line to Bivariate Data 209
5.3 Assessing the Fit of a Line 221
5.4 Nonlinear Relationships and Transformations 241
5.5 Interpreting and Communicating the Results of Statistical Analyses 259
Activity 5.1 Age and Flexibility 265
Summary: Key Concepts and Formulas 265
Chapter Review 266
Technology Notes 269
Cumulative Review Exercises 273

See Chapter 5 online materials for coverage of Logistic Regression.

CHAPTER 6 Probability 277

- 6.1** Chance Experiments and Events 278
6.2 Definition of Probability 285
6.3 Basic Properties of Probability 290
6.4 Conditional Probability 297
6.5 Independence 307
6.6 Some General Probability Rules 315
6.7 Estimating Probabilities Empirically and Using Simulation 327
Activity 6.1 Kisses 337
Activity 6.2 A Crisis for European Sports Fans? 338
Activity 6.3 The “Hot Hand” in Basketball 338
Summary: Key Concepts and Formulas 339
Chapter Review 339

CHAPTER 7 Random Variables and Probability Distributions 343

- 7.1** Random Variables 344
7.2 Probability Distributions for Discrete Random Variables 347
7.3 Probability Distributions for Continuous Random Variables 353
7.4 Mean and Standard Deviation of a Random Variable 358

- 7.5** Binomial and Geometric Distributions 371
- 7.6** Normal Distributions 383
- 7.7** Checking for Normality and Normalizing Transformations 400
- 7.8** Using the Normal Distribution to Approximate a Discrete Distribution (Optional) 410
 - Activity 7.1** Is It Real? 415
 - Activity 7.2** Rotten Eggs? 416
- Summary: Key Concepts and Formulas 416
- Chapter Review 417
- Technology Notes 420
- Cumulative Review Exercises 423

CHAPTER 8 Sampling Variability and Sampling Distributions 427

- 8.1** Statistics and Sampling Variability 428
- 8.2** The Sampling Distribution of a Sample Mean 432
- 8.3** The Sampling Distribution of a Sample Proportion 441
 - Activity 8.1** Sampling Distribution of the Sample Mean 447
 - Activity 8.2** Sampling Distribution of the Sample Proportion 449
 - Activity 8.3** Do Students Who Take the SATs Multiple Times Have an Advantage in College Admissions? 450
- Summary: Key Concepts and Formulas 452
- Chapter Review 452

CHAPTER 9 Estimation Using a Single Sample 453

- 9.1** Point Estimation 454
- 9.2** Large-Sample Confidence Interval for a Population Proportion 459
- 9.3** Confidence Interval for a Population Mean 472
- 9.4** Interpreting and Communicating the Results of Statistical Analyses 484
- 9.5** Bootstrap Confidence Intervals for a Population Proportion (Optional) 489
- 9.6** Bootstrap Confidence Intervals for a Population Mean (Optional) 496
 - Activity 9.1** Getting a Feel for Confidence Level 500
 - Activity 9.2** An Alternative Confidence Interval for a Population Proportion 501
 - Activity 9.3** Verifying Signatures on a Recall Petition 502
 - Activity 9.4** A Meaningful Paragraph 502
- Summary: Key Concepts and Formulas 502
- Chapter Review 503
- Technology Notes 504

CHAPTER 10 Hypothesis Testing Using a Single Sample 507

- 10.1** Hypotheses and Test Procedures 508
- 10.2** Errors in Hypothesis Testing 512
- 10.3** Large-Sample Hypothesis Tests for a Population Proportion 517
- 10.4** Hypothesis Tests for a Population Mean 530
- 10.5** Power and Probability of Type II Error 541
- 10.6** Interpreting and Communicating the Results of Statistical Analyses 549
- 10.7** Randomization Test and Exact Binomial Test for a Population Proportion (Optional) 552

- 10.8** Randomization Test for a Population Mean (Optional) 562
Activity 10.1 Comparing the t and z Distributions 567
Activity 10.2 A Meaningful Paragraph 568
Summary: Key Concepts and Formulas 568
Chapter Review 569
Technology Notes 571
Cumulative Review Exercises 573

CHAPTER 11

Comparing Two Populations or Treatments 576

- 11.1** Inferences Concerning the Difference Between Two Population or Treatment Means Using Independent Samples 577
11.2 Inferences Concerning the Difference Between Two Population or Treatment Means Using Paired Samples 595
11.3 Large-Sample Inferences Concerning the Difference Between Two Population or Treatment Proportions 608
11.4 Interpreting and Communicating the Results of Statistical Analyses 619
11.5 Simulation-Based Inference for Two Means (Optional) 623
11.6 Simulation-Based Inference for Two Proportions (Optional) 633
Activity 11.1 Helium-Filled Footballs? 641
Activity 11.2 Thinking About Data Collection 642
Activity 11.3 A Meaningful Paragraph 642
Summary: Key Concepts and Formulas 642
Chapter Review 643
Technology Notes 646

CHAPTER 12

The Analysis of Categorical Data and Goodness-of-Fit Tests 654

- 12.1** Chi-Square Tests for Univariate Data 655
12.2 Tests for Homogeneity and Independence in a Two-way Table 665
12.3 Interpreting and Communicating the Results of Statistical Analyses 679
Activity 12.1 Pick a Number, Any Number ... 683
Activity 12.2 Color and Perceived Taste 683
Summary: Key Concepts and Formulas 684
Chapter Review 684
Technology Notes 685

CHAPTER 13

Simple Linear Regression and Correlation: Inferential Methods 689

- 13.1** Simple Linear Regression Model 690
13.2 Inferences About the Slope of the Population Regression Line 702
13.3 Checking Model Adequacy 713
Activity 13.1 Are Tall Women from “Big” Families? 724
Summary: Key Concepts and Formulas 725
Technology Notes 725
Cumulative Review Exercises 726
13.4 Inferences Based on the Estimated Regression Line 13-1
13.5 Inferences About the Population Correlation Coefficient 13-8
13.6 Interpreting and Communicating the Results of Statistical Analyses 13-11

CHAPTER 14 Multiple Regression Analysis 730

- 14.1** Multiple Regression Models 731
- 14.2** Fitting a Model and Assessing Its Utility 742
 - Activity 14.1** Exploring the Relationship Between Number of Predictors and Sample Size 758
 - Summary: Key Concepts and Formulas 758
- 14.3** Inferences Based on an Estimated Model 14-1
- 14.4** Other Issues in Multiple Regression 14-12
- 14.5** Interpreting and Communicating the Results of Statistical Analyses 14-22
 - Chapter Review 14-23

CHAPTER 15 Analysis of Variance 759

- 15.1** Single-Factor ANOVA and the *F* Test 760
- 15.2** Multiple Comparisons 772
 - Activity 15.1** Exploring Single-Factor ANOVA 780
 - Summary: Key Concepts and Formulas 782
 - Technology Notes 782
- 15.3** The *F* Test for a Randomized Block Experiment 15-1
- 15.4** Two-Factor ANOVA 15-7
- 15.5** Interpreting and Communicating the Results of Statistical Analyses 15-17
 - Chapter Review 15-21

CHAPTER 16 Nonparametric (Distribution-Free) Statistical Methods 16-1

- 16.1** Distribution-Free Procedures for Inferences About a Difference Between Two Population or Treatment Means Using Independent Samples 16-2
- 16.2** Distribution-Free Procedures for Inferences About a Difference Between Two Population or Treatment Means Using Paired Samples 16-9
- 16.3** Distribution-Free ANOVA 16-19
 - Summary: Key Concepts and Formulas 16-26
 - Appendix: Tables 16-27

Appendix: Statistical Tables 785

Answers to Selected Odd-Numbered Exercises 805

Index 845

Sections and/or chapter numbers shaded in color can be found at
www.cengage.com

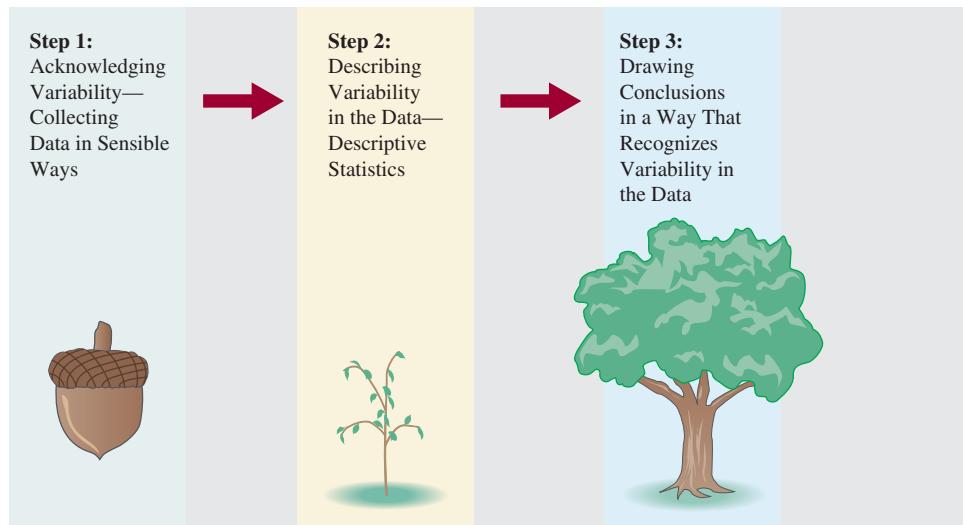
Preface

In a nutshell, statistics is about understanding the role that variability plays in drawing conclusions based on data. *Introduction to Statistics and Data Analysis*, Sixth Edition, develops this crucial understanding of variability through its focus on the data analysis process.

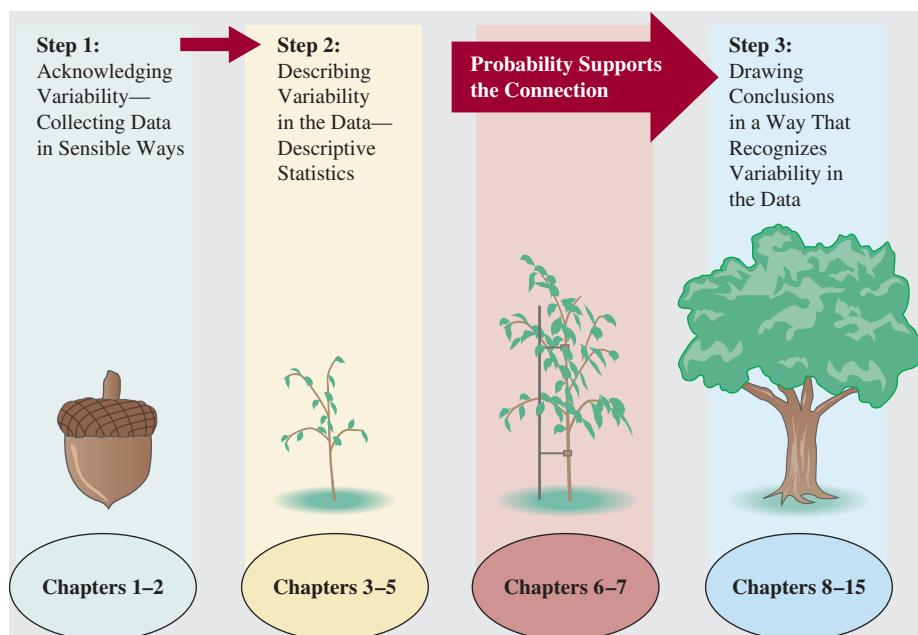
An Organization That Reflects the Data Analysis Process

Students are introduced early to the idea that data analysis is a process that begins with careful planning, followed by data collection, data description using graphical and numerical summaries, data analysis, and finally interpretation of results. This process is described in detail in Chapter 1, and the ordering of topics in the first ten chapters of the book mirrors the process: data collection, then data description, then statistical inference.

The logical order in the data analysis process can be pictured as shown in the following figure.



Unlike many introductory texts, *Introduction to Statistics and Data Analysis*, Sixth Edition, is organized in a way that is consistent with the natural order of the data analysis process:



The Importance of Context and Real Data

Statistics is not about numbers; it is about data—numbers in context. It is the context that makes a problem meaningful and something worth considering. For example, exercises that ask students to calculate the mean of 10 numbers or to construct a dotplot or boxplot of 20 numbers without context are arithmetic and graphing exercises. They become statistics problems only when a context gives them meaning and allows for interpretation. While this makes for a text that may appear “wordy” when compared to traditional mathematics texts, it is a critical and necessary component of a modern statistics text.

Examples and exercises with overly simple settings do not allow students to practice interpreting results in authentic situations or give students the experience necessary to be able to use statistical methods in real settings. We believe that the exercises and examples are a particular strength of this text, and we invite you to compare the examples and exercises with those in other introductory statistics texts.

Many students are skeptical of the relevance and importance of statistics. Contrived problem situations and artificial data often reinforce this skepticism. A strategy that we have employed successfully to motivate students is to present examples and exercises that involve data extracted from journal articles, newspapers, and other published sources. Most examples and exercises in the book are of this nature; they cover a very wide range of disciplines and subject areas. These include, but are not limited to, health and fitness, consumer research, psychology and aging, sports, environmental research, law and criminal justice, and entertainment.

A Focus on Interpretation and Communication

Most chapters include a section titled “Interpreting and Communicating the Results of Statistical Analyses.” These sections include advice on how to best communicate the results of a statistical analysis and also consider how to interpret statistical summaries found in journals and other published sources. Subsections titled “A Word to the Wise” reminds readers of things that must be considered in order to ensure that statistical methods are employed in reasonable and appropriate ways.

Consistent with Recommendations for the Introductory Statistics Course Endorsed by the American Statistical Association

In 2016, the American Statistical Association updated the report “Guidelines in Assessment and Instruction for Statistics Education (GAISE) College Report 2016,” which included the following six recommendations for the introductory statistics course:

1. Teach statistical thinking.
2. Focus on conceptual understanding.
3. Integrate real data with a context and purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

Introduction to Statistics and Data Analysis, Sixth Edition, is consistent with these recommendations and supports the GAISE guidelines in the following ways:

1. Teach statistical thinking.

Statistical thinking and statistical literacy are promoted throughout the text in the many examples and exercises that are drawn from the popular press. In addition, a focus on the role of variability, consistent use of context, and an emphasis on interpreting and communicating results in context work together to help students develop skills in statistical thinking.

2. Focus on conceptual understanding.

Nearly all exercises in *Introduction to Statistics and Data Analysis*, Sixth Edition, are multipart and ask students to go beyond just computation. They focus on interpretation and communication, not just in the chapter sections specifically devoted to this topic, but throughout the text. The examples and explanations are designed to promote conceptual understanding. Hands-on activities in each chapter are also constructed to strengthen conceptual understanding. Which brings us to . . .

3. Integrate real data with a context and a purpose.

The examples and exercises from *Introduction to Statistics and Data Analysis*, Sixth Edition, are context driven, and the reference sources include the popular press as well as journal articles.

4. Foster active learning.

While this recommendation speaks more to pedagogy and classroom practice, *Introduction to Statistics and Data Analysis*, Sixth Edition, provides more than 30 hands-on activities in the text and additional activities in the accompanying instructor resources that can be used in class or assigned to be completed outside of class.

5. Use technology to explore concepts and analyze data.

The computer brings incredible statistical power to the desktop of every investigator. The wide availability of statistical computer packages such as Minitab, JMP, and SPSS, and the graphical capabilities of the modern microcomputer have transformed both the teaching and learning of statistics. To highlight the role of the computer in contemporary statistics, we have included sample output throughout the book. In addition, numerous exercises contain data that can easily be analyzed using statistical software. However, access to a particular statistical package is not assumed. Technology manuals for specific software packages and for the graphing calculator are available in the online materials that accompany this text. The sixth edition of *Introduction to Statistics and Data Analysis* also includes a number of Shiny web apps that can be used to illustrate statistical concepts and to implement the simulation-based inference methods covered in new optional sections.

The appearance of handheld calculators with significant statistical and graphing capability has also changed statistics instruction in classrooms where access to computers is still limited. There is not universal or even wide agreement about the proper role for the graphing calculator in college statistics classes, where access to a computer is more common. At the same time, for tens of thousands of students in

Advanced Placement Statistics in our high schools, the graphing calculator is the only dependable access to statistical technology.

This text allows the instructor to balance the use of computers and calculators in a manner consistent with his or her philosophy. As with computer packages, our exposition avoids assuming the use of a particular calculator and presents the calculator capabilities in a generic format. For those using a TI graphing calculator, there is a technology manual available in the online materials that accompany this text.

6. Use assessments to improve and evaluate student learning.

Assessment materials in the form of a test bank, quizzes, and chapter exams are available in the instructor resources that accompany this text. The items in the test bank reflect the data-in-context philosophy of the text's exercises and examples.

Advanced Placement® Statistics

We have designed this book with a particular eye toward the syllabus of the Advanced Placement® Statistics course and the needs of high school teachers and students. Concerns expressed and questions asked in teacher workshops and on the AP® Statistics teacher community have strongly influenced our explanation of certain topics, especially in the area of experimental design and probability. We have taken great care to provide precise definitions and clear examples of concepts that Advanced Placement® Statistics instructors have acknowledged as difficult for their students. We have also expanded the variety of examples and exercises, recognizing the diverse potential futures envisioned by very capable students who have not yet focused on a college major. The AP® edition of this text also contains a collection of multiple choice and free response questions that can be used to help students review for the AP® Statistics exam.

Topic Coverage

Our book can be used in courses as short as one quarter or as long as one year in duration. Particularly in shorter courses, an instructor will need to be selective in deciding which topics to include and which to set aside. The book divides naturally into four major sections: collecting data and descriptive methods (Chapters 1–5), probability (Chapters 6–8), the basic one- and two-sample inferential techniques (Chapters 9–12), and more advanced inferential methodology (Chapters 13–16).

We include an early chapter (Chapter 5) on descriptive methods for bivariate numerical data. This early exposure introduces questions and issues that should stimulate student interest in the subject. It is also advantageous for those teaching courses in which time constraints preclude covering advanced inferential material. However, this chapter can easily be postponed until the basics of inference have been covered, and then combined with Chapter 13 for a unified treatment of regression and correlation.

With the possible exception of Chapter 5, Chapters 1 through 10 should be covered in order. We anticipate that most instructors will then continue with two-sample inference (Chapter 11) and methods for categorical data analysis (Chapter 12), although regression could be covered before either of these topics. Optional portions of Chapter 14 (multiple regression) and Chapter 15 (analysis of variance) and Chapter 16 (nonparametric methods) are included in the online materials that accompany this text.

A Note on Probability

The content of the probability chapters is consistent with the Advanced Placement® Statistics course description. It includes both a traditional treatment of probability and probability distributions at an introductory level, as well as a section on the use of simulation as a tool for estimating probabilities. For those who prefer a more informal treatment of probability, see the text *Statistics: Learning from Data*, Second Edition, by Roxy Peck and Tom Short.

In This Edition

Look for the following in the Sixth Edition:

- **NEW Updated Examples and Exercises.** In our continuing effort to keep things interesting and relevant, the sixth edition contains many updated examples and exercises that use data from recent journal articles, newspapers, and web posts, on topics of interest to students.
- **NEW Sections on Randomization-Based Inference Methods.** Research indicates that randomization-based instruction in statistical inference may help learners to better understand the concepts of confidence and significance. The sixth edition includes new optional sections on randomization-based inference methods. These methods are also particularly useful in that they provide an alternative method of analysis that can be used when the conditions required for normal distribution-based inference are not met. The inference chapters (Chapters 9–11) now contain new optional sections on randomization-based inference that include bootstrap methods for simulation-based confidence intervals and randomization-based tests of hypotheses. These new sections are accompanied by online Shiny apps, which can be used to construct bootstrap confidence intervals and to carry out randomization tests.
- **Helpful hints in exercises.** To help students who might be having trouble getting started, hints have been added to many of the exercises directing students to relevant examples in the text.
- **Margin notes to highlight the importance of context and the process of data analysis.** Margin notations appear in appropriate places in the examples. These include *Understanding the context*, *Consider the data*, *Formulate a plan*, *Do the work*, and *Interpret the results*. These notes are designed to increase student awareness of the steps in the data analysis process.
- **Activities at the end of each chapter.** These activities can be used as a chapter capstone or can be integrated at appropriate places as the chapter material is covered in class.
- **Advanced topics** that are often omitted in a one-quarter or one-semester course, such as survey design (Section 2.6), logistic regression (Section 5.6), inference based on the estimated regression line (Sections 13.4 and 13.5), inference and variable selection methods in multiple regression (Sections 14.3 and 14.4), analysis of variance for randomized block and two-factor designs (Sections 15.3 and 15.4), and distribution-free procedures (Chapter 16) **are available in the online materials that accompany this text**.
- **Updated materials for instructors** are included on the Instructor Companion Site. In addition to the usual instructor supplements such as a complete solutions manual and a test bank, the website contains examples that can be incorporated into classroom presentations and cross-references to resources such as Fathom, *Workshop Statistics*, and *Against All Odds*. Of particular interest to those teaching Advanced Placement Statistics, the website also includes additional data analysis questions of the type encountered on the free response portion of the Advanced Placement exam, as well as a collection of model responses.

Instructor and Student Resources



MindTap™

Available via WebAssign is MindTap™ Reader, Cengage Learning's next-generation eBook. MindTap Reader provides robust opportunities for students to annotate, take notes, navigate, and interact with the text (e.g., ReadSpeaker). Annotations captured in MindTap Reader are automatically tied to the Notepad app, where they can be viewed chronologically and in a cogent, linear fashion. Instructors also can edit the text and assets in the Reader, as well as add videos or URLs.



WebAssign

WebAssign for Peck/Short/Olsen's *Introduction to Statistics and Data Analysis*, 6th Edition, is a flexible and fully customizable online instructional solution that puts powerful tools in the hands of instructors, empowering you to deploy assignments, instantly assess individual student and class performance, and help your students master the course concepts. With WebAssign's powerful digital platform and Introduction to Probability and Statistics specific content, you can tailor your course with a wide range of assignment settings, add your own questions and content, and access student and course analytics and communication tools. Learn more at www.webassign.com.



Access to JMP is free with the purchase of a new book.

JMP Statistical Software

JMP is a statistics software for Windows and Macintosh computers from SAS, the market leader in analytics software and services for industry. JMP Student Edition is a streamlined, easy-to-use version that provides all the statistical analysis and graphics covered in this textbook. Once data is imported, students will find that most procedures require just two or three mouse clicks. JMP can import data from a variety of formats, including Excel and other statistical packages, and you can easily copy and paste graphs and output into documents.

JMP also provides an interface to explore data visually and interactively, which will help your students develop a healthy relationship with their data, work more efficiently with data, and tackle difficult statistical problems more easily. Because its output provides both statistics and graphs together, the student will better see and understand the application of concepts covered in this book as well. JMP Student Edition also contains some unique platforms for student projects, such as mapping and scripting. JMP functions in the same way on both Windows and Mac platforms and instructions contained with this book apply to both platforms.

Access to this software is available for free with new copies of the book and available for purchase standalone at Cengage.com or www.jmp.com/getse. Find out more at www.jmp.com.

Web Apps

A collection of easy to use web apps is available at statistics.cengage.com/PSO6e/Apps.html. This collection includes apps that support new sections on bootstrap confidence intervals and randomization tests, as well as apps that help students visualize the meaning of confidence level and to understand the concept of sampling variability.

Student Resources

Digital



To access additional course materials and companion resources, please visit www.cengage.com. At the Cengage.com home page, search for the ISBN of your title (from the back cover of your book) using the search box at the top of the page. This will take you to the product page where free companion resources can be found.

- Complete step-by-step instructions for JMP, TI-84 Graphing Calculators, Excel, Minitab, and SPSS.
- Data sets in Excel, ASCII-comma, ASCII-tab, JMP, Minitab, R, SAS, and SPSS file formats indicated by the ● icon throughout the text.
- Applets used in the Activities found in the text.

Print

Student Solutions Manual (ISBN: 978-1-337-79417-6): The Student Solutions Manual, prepared by Stephen Miller, contains fully worked-out solutions to all of the odd-numbered exercises in the text, giving students a way to check their answers and ensure that they took the correct steps to arrive at an answer.

Instructor Resources

Print

Annotated Instructor’s Edition (ISBN: 978-1-337-79415-2): The Annotated Instructor’s Edition contains answers for all exercises, including those not found in the answer section of the student edition. There also are suggested assignments and teaching tips for each section in the book, along with an annotated table of contents with comments written by Roxy Peck.

AP® Teacher’s Resource Binder: The Teacher’s Resource Binder, prepared by Chris Olsen, is full of wonderful resources for both college professors and AP® Statistics teachers. These include

- Additional examples from published sources (with references), classified by chapter in the text. These examples can be used to enrich your classroom discussions.
- Model responses—examples of responses that can serve as a model for work that would be likely to receive a high mark on the AP® exam.
- A collection of data explorations written by Chris Olsen that can be used throughout the year to help students prepare for the types of questions that they may encounter on the investigative task on the AP® Statistics Exam.
- Advice to AP® Statistics teachers on preparing students for the AP® Exam, written by Brian Kotz.
- Activity worksheets, prepared by Carol Marchetti, that can be duplicated and used in class.
- A list of additional resources for activities, videos, and computer demonstrations, cross-referenced by chapter.
- A test bank that includes assessment items, quizzes, and chapter exams written by Chris Olsen, Josh Tabor, and Peter Flanagan-Hyde.

Online

- **Instructor Companion Site:** Everything you need for your course in one place! This collection of book-specific lecture and class tools is available online via www.cengage.com/login. Access and download PowerPoint presentations, images, instructor’s manual, and more.
- **Cengage Learning Testing Powered by Cognero (ISBN: 978-1-337-79423-7)** is a flexible, online system that allows you to author, edit, and manage test bank content, create multiple test versions in an instant, and deliver tests from your LMS, your classroom or wherever you want. This is available online via www.cengage.com/login.
- **Complete Solutions Manual** This manual contains solutions to all exercises from the text, including Chapter Review Exercises and Cumulative Review Exercises. This manual can be found on the Instructor Companion Site.

Acknowledgments

We are grateful for the thoughtful feedback from the following reviewers that has helped to shape this text over the last three editions:

Reviewers for the Sixth Edition

Weizhong Tian, Eastern New Mexico University
Greg Perkins, Hartnell College
Bambi Jones, Lake Land College
Paul Holmes, University of Georgia
Christina Cornejo, Erie Community College

Carl Brezovec, Franklin Pierce University
David Manley, Rowan University
John Racquet, University at Albany, Excelsior College
Charles Conrad, Volunteer State Community College
Zhongming Huang, Midland University
Chad Bemis, Pierce College Fort Steilacoom

Reviewers for the Fifth, Fourth, Third, and Second Editions

Arun K. Agarwal, Jacob Amidon, Holly Ashton, Barb Barnet, Eddie Bevilacqua, Piotr Bialas, Kelly Black, Jim Bohan, Pat Buchanan, Gabriel Chandler, Andy Chang, Jerry Chen, Richard Chilcoat, Mary Christman, Marvin Creech, Kathleen Dale, Ron Degged, Hemangini Deshmukh, Ann Evans, Guangxiong Fang, Sharon B. Finger, Donna Flint, Steven Garren, Mark Glickman, Rick Gmina, Debra Hall, Tyler Haynes, Sonja Hensler, Trish Hutchinson, John Imbrie, Bessie Kirkwood, Jeff Kollath, Christopher Lacke, Austin Lampros, Michael Leitner, Zia Mahmood, Art Mark, Pam Martin, David Mathiason, Bob Mattson, Kendra Mhoon, C. Mark Miller, Megan Mocko, Paul Myers, Kane Nashimoto, Helen Noble, Douglas Noe, Broderick Oluyede, Elaine Paris, Shelly Ray Parsons, Deanna Payton, Judy Pennington-Price, Michael Phelan, Alan Polansky, Mamunur Rashid, Leah Rathbun, Michael Ratliff, David Rauth, Kevin J. Reeves, Lawrence D. Ries, Hazel Shedd, Robb Sinn, Greg Sliwa, Angela Stabley, Jeffery D. Sykes, Yolanda Tra, Joe Ward, Nathan Wetzel, Mark Wilson, Yong Yu, and Toshiyuki Yuasa, Cathleen Zucco-Teveloff.

We would also like to express our thanks and gratitude to those whose support made this sixth edition possible:

- Cassie Van Der Laan, Product Manager
- Brendan Killion and Abby DeVeuve, Content Managers
- Elinor Gregory, Learning Designer
- Lori Hazzard, our senior project manager at MPS Limited
- Stephen Miller for his work in creating new student and instructor solutions manuals to accompany the text
- A. Palanisamy, for checking the accuracy of examples and solutions
- Carolyn Crockett and Molly Taylor, our former editors at Cengage, for their support on the previous editions of this book

And, as always, we thank our families, friends, and colleagues for their continued support.

Roxy Peck

Tom Short

Chris Olsen

Introduction to Statistics and Data Analysis

1

The Role of Statistics and the Data Analysis Process



ESB Professional/Shutterstock.com

Statistics is the scientific discipline that provides methods that help us make sense of data. Statistical methods offer a set of powerful tools for gaining insight into the world around us. The use of statistical analyses in fields such as business, medicine, agriculture, social science, natural science, and engineering has led to increased recognition that statistical literacy should be part of a well-rounded education.

The field of statistics helps us to make intelligent judgments and informed decisions in the presence of uncertainty and variability. In this chapter, we consider the role of variability in statistical settings, introduce some basic terminology, and look at some simple graphical displays for summarizing data.

LEARNING OBJECTIVES

Students will understand:

- The steps in the data analysis process.

Students will be able to:

- Distinguish between a population and a sample.
- Distinguish between categorical, discrete numerical, and continuous numerical data.
- Construct a frequency distribution and a bar chart, and describe the distribution of a categorical variable.
- Construct a dotplot and describe the distribution of a numerical variable.

SECTION 1.1 Why Study Statistics?

There is an old saying that “without data, you are just another person with an opinion.” While anecdotes and coincidences may make for interesting stories, you wouldn’t want to make important decisions on the basis of anecdotes alone. For example, just because a friend of a friend ate 16 apricots and then experienced relief from joint pain doesn’t mean that this is all you need to know to help one of your parents choose a treatment for arthritis! Before recommending apricots, you would definitely want to consider relevant data—that is, data that would allow you to investigate the effectiveness of apricots as a treatment for arthritis.

It is challenging to function in today’s world without a basic understanding of statistics. For example, here are just a few headlines from articles that draw conclusions based on data that appeared in a single newspaper on one day.

How many people does it take to build an airplane? The article [“Boeing Delivers Records” \(The Wall Street Journal, January 10, 2018\)](#) looked at the ways in which Boeing has been able to increase airplane production and introduce new airplane models. The article states that “Boeing has boosted output by two-thirds over the past seven years but cut the average number of employees needed to build each plane.” This statement was supported by a graph that showed the number of employees per jet airplane produced over time. In 2017, this number was at its lowest, with 94 employees per jet produced.

[“New Venture Fund Targets Autos” \(The Wall Street Journal, January 10, 2018\)](#) is the title of an article that looked at funding for innovation in the auto industry. Venture capital is money invested in start-up companies exploring new technologies. The article included a graph of data that shows how the amount of venture capital funding for the automotive industry has been increasing over time, noting that this funding nearly tripled in 2017 compared with 2016. The hope is that this will increase the pace of introduction of new technologies in new cars.

The article [“Companies Take to the Sky in Race to Deliver on Time” \(The Wall Street Journal, January 10, 2018\)](#) notes that growth in online shopping and a healthy economy is creating an increased demand for shipping by air. Two graphs of data are included with the article. One shows the yearly change in air cargo volume for the years 2013 to 2017, noting increases in each of these years and a particularly large increase in 2017. The second graph shows how the cost of air cargo shipping has increased since 2015. Based on these trends, Amazon has started its own airline to handle the shipping of Amazon orders and is converting older passenger jets for cargo use.

The article [“Saudis Target Religious Extremism” \(The Wall Street Journal, January 10, 2018\)](#) reported on public reaction to Saudi Arabia’s decision to lift a ban that prohibits women from driving. The article includes graphs summarizing data collected in a survey of 500 Saudis. The people surveyed were asked if they were pleased with the decision to lift the ban on women driving. The graphs showed that 74% of the women surveyed and 55% of the men surveyed said that they were pleased with the decision. The graphs also showed that the percentage of men who were not sure if they were pleased or if they were not pleased was greater than this percentage for women (17% of the men and 11% of the women). These data enable the Saudi government to assess support for social change.

As people approach retirement age, many are finding that they are not prepared financially. The article [“37% of Gen X Can’t Afford to Retire, Poll Finds” \(The Wall Street Journal, January 10, 2018\)](#) summarized the results of a survey of 828 people born between 1965 and the late 1970s (known as “Generation X”) and 990 people born between 1945 and 1964 (the “Baby Boomers”). They found that while 47% of Baby Boomers expect that they will be very secure in retirement, only 33% of Gen Xers expect that they will be very secure. They also found that 37% of Gen Xers would like to stop working someday, but fear that they will not be able to afford to, and that 49% are worried about running out of money once they leave the workforce. These findings have implications for those who provide social services to older Americans.

To be an informed consumer of reports such as those described above, you must be able to do the following:

1. Extract information from tables, charts, and graphs.
2. Follow numerical arguments.
3. Understand the basics of how data should be gathered, summarized, and analyzed in order to draw statistical conclusions.

Your statistics course will help prepare you to perform these tasks.

Studying statistics will also enable you to collect data in a sensible way and then use the data to answer questions of interest. In addition, studying statistics will allow you to critically evaluate the work of others by providing you with the tools you need to make informed judgments.

Throughout your personal and professional life, you will need to understand and use data to make decisions. To do this, you must be able to

1. Decide whether existing data are adequate or whether additional information is required.
2. If necessary, collect more information in a reasonable and thoughtful way.
3. Summarize the available data in a useful and informative manner.
4. Analyze the available data.
5. Draw conclusions, make decisions, and assess the risk of an incorrect decision.

These are the steps in the data analysis process. These steps are considered in more detail in Section 1.3.

We hope that this textbook will help you to understand the logic behind statistical reasoning, prepare you to apply statistical methods appropriately, and enable you to recognize when statistical arguments are faulty.

SECTION 1.2 The Nature and Role of Variability

Statistical methods allow us to collect, describe, analyze, and draw conclusions from data. If we lived in a world where all measurements were identical for every individual, these tasks would be simple. Imagine a population consisting of all students at a particular university. Suppose that *every* student is enrolled in the same number of courses, spent exactly the same amount of money on textbooks this semester, and favors increasing student fees to support expanding library services. For this population, there is *no* variability in number of courses, amount spent on books, or student opinion on the fee increase. A researcher studying students from this population in order to draw conclusions about these three variables would have a particularly easy task. It would not matter how many students the researcher studied or how the students were selected. In fact, the researcher could collect information on number of courses, amount spent on books, and opinion on the fee increase by just stopping the next student who happened to walk by. Because there is no variability in the population, this one individual would provide complete and accurate information about the population. The researcher could draw conclusions with no risk of error.

The situation just described is obviously unrealistic. Populations with no variability are rare. We need to understand variability to be able to collect, describe, analyze, and draw conclusions from data in a sensible way.

Examples 1.1 and 1.2 illustrate how describing and understanding variability are important.

Example 1.1 If the Shoe Fits

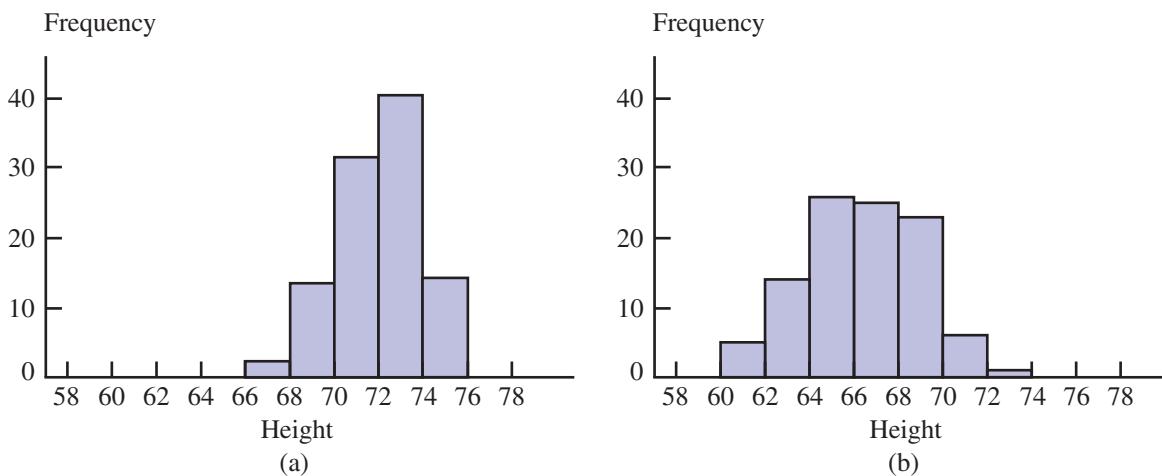
Understand the context ➤

The graphs in Figure 1.1 are examples of a type of graph called a histogram. (The construction and interpretation of histograms is discussed in Chapter 3.) Figure 1.1(a) shows the distribution of the heights of female basketball players who played at a particular university between 2005 and 2013. Each bar in the histogram represents a particular range of player heights. The height of each bar in the graph indicates how many players' heights were in the corresponding range. For example, 40 basketball players had heights between 72 inches and 74 inches, while only 2 players had heights between 66 inches and 68 inches. Figure 1.1(b) shows the distribution of heights for members of the women's gymnastics team. Both histograms are based on the heights of 100 women.

Consider the data ➤

FIGURE 1.1

Histograms of heights (in inches) of female athletes: (a) 100 basketball players; (b) 100 gymnasts.



Interpret the results ➤

The first histogram shows that the heights of female basketball players varied, with most heights falling between 68 inches and 76 inches. In the second histogram we see that the heights of female gymnasts also varied, with most heights in the range of 60 inches to 72 inches. It is also clear that there is more variation in the heights of the gymnasts than in the heights of the basketball players, because the gymnast histogram spreads out more about its center than does the basketball histogram.

Now suppose that a tall woman (5 feet 11 inches) tells us she is looking for her sister who is practicing with her team at the gym. Should we direct her to where the basketball team is practicing or to where the gymnastics team is practicing? What reasoning could we use to decide? If we found a pair of size 6 shoes left in the locker room, should we first try to return them by checking with members of the basketball team or the gymnastics team?

We would probably send the woman looking for her sister to the basketball practice and we would probably try to return the shoes to a gymnastics team member. Reaching these conclusions requires statistical reasoning that combines knowledge of the relationship between heights of siblings and between shoe size and height with the information about the distributions of heights presented in Figure 1.1. We might have reasoned that heights of siblings tend to be similar and that a height as great as 5 feet 11 inches, although not impossible, would be unusual for a gymnast. On the other hand, a height as tall as 5 feet 11 inches would be common for a basketball player.

Similarly, we might have reasoned that tall people tend to have bigger feet and that short people tend to have smaller feet. The shoes found were a small size, so it is more likely that they belong to a gymnast than to a basketball player, because small heights are common for gymnasts and unusual for basketball players.

Example 1.2 Monitoring Water Quality

Understand the context ➤

As part of its regular water quality monitoring efforts, an environmental control board selects five containers of water from a particular well each day. The concentration of contaminants in parts per million (ppm) is measured for each of the five containers, and then the average of the five measurements is calculated. The histogram in Figure 1.2 summarizes the average contamination values for 200 days.



David Clasey/Photodisc/Getty Images

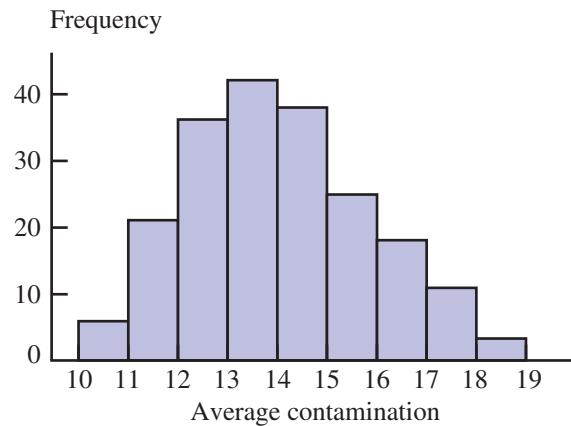
Consider the data ➤

Now suppose that a chemical spill has occurred at a manufacturing plant 1 mile from the well. It is not known whether a spill of this nature would contaminate groundwater in the area of the spill and, if so, whether a spill this distance from the well would affect the quality of well water.

One month after the spill, five containers of water are collected from the well, and the average contamination is 15.5 ppm. Considering the variability before the spill, should we interpret this as evidence that the well water was affected by the spill? What if the calculated average was 17.4 ppm? 22.0 ppm? How is the reasoning related to the histogram in Figure 1.2?

FIGURE 1.2

Average contamination concentration (in ppm) measured each day for 200 days.



Interpret the results ➤

Before the spill, the average contaminant concentration varied from day to day. An average of 15.5 ppm would not have been an unusual value, so seeing an average of 15.5 ppm after the spill isn't necessarily an indication that contamination has increased. On the other hand, an average as large as 17.4 ppm is less common, and an average as large as 22.0 ppm is not at all typical of the pre-spill values. In this case, we would probably conclude that the well contamination level has increased.

In these two examples, reaching a conclusion required an understanding of variability. Understanding variability allows us to distinguish between common and unusual values. The ability to recognize unusual values in the presence of variability is an important aspect of most statistical procedures. It also enables us to quantify the chance of being incorrect when a conclusion is based on data. These concepts will be developed further in later chapters.

SECTION 1.3 Statistics and the Data Analysis Process

Statistical studies are undertaken to answer questions about our world. Is a new flu vaccine effective in preventing illness? Is the use of bicycle helmets on the rise? Are injuries that result from bicycle accidents less severe for riders who wear helmets than for those who do not? Data collection and analysis allow researchers to answer questions like these.

The data analysis process can be viewed as a sequence of steps that lead from planning to data collection to making informed conclusions based on the resulting data. The process can be organized into the six steps described below.

The Data Analysis Process

- 1. Understanding the nature of the problem.** Effective data analysis requires an understanding of the research problem. We must know the goal of the research and what questions we hope to answer. It is important to have a clear direction before gathering data to ensure that we will be able to answer the questions of interest using the data collected.
- 2. Deciding what to measure and how to measure it.** The next step in the process is deciding what information is needed to answer the questions of interest. In some cases, the choice is obvious. For example, in a study of the relationship between the weight of a Division I football player and position played, we would need to collect data on player weight and position. In other cases the choice of information is not as straightforward. For example, in a study of the relationship between preferred learning style and intelligence, how should we define learning style and measure it? What measure of intelligence should we use? It is important to carefully define the variables to be studied and to develop appropriate methods for determining their values.
- 3. Data collection.** The data collection step is very important. The researcher must first decide whether an existing data source is adequate or whether new data must be collected. If a decision is made to use existing data, it is important to understand how the data were collected and for what purpose, so that any resulting limitations are also fully understood. If new data are to be collected, a careful plan must be developed, because the type of analysis that is appropriate and the conclusions that can be drawn depend on how the data are collected.
- 4. Data summarization and preliminary analysis.** After the data are collected, the next step is usually a preliminary analysis that includes summarizing the data graphically and numerically. This initial analysis provides insight into important characteristics of the data and provides guidance in selecting appropriate methods for further analysis.
- 5. Formal data analysis.** The data analysis step requires the researcher to select appropriate statistical methods. Much of this textbook is devoted to methods that can be used to carry out this step.
- 6. Interpretation of results.** Several questions should be addressed in this final step. Some examples are: What can we learn from the data? What conclusions can be drawn from the analysis? How can our results guide future research? The interpretation step often leads to the formulation of new research questions. These new questions lead back to the first step. In this way, good data analysis is often an iterative process.

To illustrate these steps, consider the following example. The admissions director at a large university might be interested in learning why some applicants who were accepted for the fall 2019 term failed to enroll at the university. The **population** of interest to the director consists of all accepted applicants who did not enroll in the fall 2019 term. Because this population is large and it may be difficult to contact all the individuals, the director might decide to collect data from only 300 selected students. These 300 students constitute a **sample**.

DEFINITIONS

Population: The entire collection of individuals or objects about which information is desired is called the **population** of interest.

Sample: A **sample** is a subset of the population, selected for study.

Deciding how to select the 300 students and what data should be collected from each student are steps 2 and 3 in the data analysis process. Step 4 in the process involves

organizing and summarizing data. Methods for organizing and summarizing data, such as the use of tables, graphs, and numerical summaries, make up the branch of statistics called **descriptive statistics**. A second major branch of statistics, **inferential statistics**, involves generalizing from a sample to the population from which it was selected. When we generalize in this way, we run the risk of an incorrect conclusion, because a conclusion about the population is based on incomplete information. An important aspect in the development of inferential techniques involves quantifying the chance of an incorrect conclusion.

DEFINITIONS

Descriptive statistics: The branch of statistics that includes methods for organizing and summarizing data.

Inferential statistics: The branch of statistics that involves generalizing from a sample to the population from which the sample was selected and assessing the reliability of such generalizations.

Example 1.3 illustrates the steps in the data analysis process.

Example 1.3 Chew More, Eat Less?

Understand the context ➤

The article “[Increasing the Number of Chews before Swallowing Reduces Meal Size in Normal-Weight, Overweight, and Obese Adults](#)” (*Journal of the Academy of Nutrition and Dietetics* [2014]: 926–931) describes a study that investigated whether chewing each bite of food more before swallowing would result in people eating less. Participants in the study were adults between the ages of 18 and 45 years. At the beginning of the study, each participant was observed as they each ate five pizza rolls, and the number of chews made before swallowing was observed in order to determine a baseline for that participant.

Participants were then invited back for a second session on a different day. They were asked to eat their usual breakfast on that day and to not eat anything after breakfast. At the second session, all participants were provided with a platter of pizza rolls and were told to eat until they were comfortably full. They were also told they could request more pizza rolls if they wanted more. Each participant was also told how many times to chew each pizza roll before swallowing. Then, each participant was assigned to one of three groups. The participants in group 1 were given a number of chews equal to their baseline. The participants in group 2 were given a number of chews that was 150% of (one and a half times as large as) their baseline. The participants in group 3 were assigned a number of chews that was 200% of (twice as large as) their baseline.

Interpret the results ➤

After analyzing data from this study, the researcher concluded that people ate about 10% less when they increased the number of chews by 50% (group 2) and about 15% less when they doubled the number of chews.

This study illustrates the nature of the data analysis process. A clearly defined research question and an appropriate choice of how to measure the variables of interest (the number of chews and how much people ate) preceded the data collection. Assuming that a reasonable method was used to collect the data (we will see how this can be evaluated in Chapter 2) and that appropriate methods of analysis were employed, the investigators reached the conclusion that increasing the number of chews before swallowing results in people tending to eat less.

EXERCISES 1.1 - 1.11

● Data set available online

- 1.1 Give brief definitions of the terms *descriptive statistics* and *inferential statistics*.
- 1.2 Give brief definitions of the terms *population* and *sample*.

- 1.3 The following conclusion from a study appeared in the article “[Smartphone Nation](#)” (*AARP Bulletin*, September 2009): “If you love your smart phone, you are not alone. Half of all boomers sleep with their cell phone within arm’s length. Two of three

people age 50 to 64 use a cell phone to take photos, according to a 2010 Pew Research Center report.” Are the given proportions (half and two of three) population values, or were they calculated from a sample?

- 1.4** Based on a study of 2121 children between the ages of 1 and 4, researchers at the Medical College of Wisconsin concluded that there was an association between iron deficiency and the length of time that a child is bottle-fed (*Milwaukee Journal Sentinel*, November 26, 2005). Describe the sample and the population of interest for this study.
- 1.5** The student senate at a university with 15,000 students is interested in the proportion of students who favor a change in the grading system to allow for plus and minus grades (for example, B+, B, B–, rather than just B). Two hundred students are interviewed to determine their attitude toward this proposed change.
- What is the population of interest?
 - What group of students constitutes the sample in this problem?
- 1.6** The National Retail Federation used data from a survey of 7439 adult Americans to estimate the percent who planned to spend more on holiday shopping in 2017 than they spent in 2016. They estimated that while 24% of adult Americans planned to spend more, for those age 16 to 24, the percentage was 46% (“Almost Half of Younger Consumers Plan to Spend More During the Holidays,” nrf.com/media/press-releases/almost-half-of-younger-consumers-plan-spend-more-during-the-holidays, retrieved February 5, 2018). Are the estimates given calculated using data from a sample or for the entire population?
- 1.7** The supervisors of a rural county are interested in the proportion of property owners who support the construction of a sewer system. Because it is too costly to contact all 7000 property owners, a survey of 500 owners is undertaken. Describe the population and sample for this problem.
- 1.8** A consumer group conducts crash tests of new model cars. To determine the severity of damage to 2019 Toyota Camrys resulting from a 10-mph crash into a concrete wall, the research group tests six cars of this type and assesses the amount of damage. Describe the population and the sample for this problem.
- 1.9** A building contractor has a chance to buy an odd lot of 5000 used bricks at an auction. She is

interested in determining the proportion of bricks in the lot that are cracked and therefore unusable for her current project, but she does not have enough time to inspect all 5000 bricks. Instead, she checks 100 bricks to determine which ones are cracked. Describe the population and the sample for this problem.

- 1.10** The article “**Brain Shunt Tested to Treat Alzheimer’s**” (*San Francisco Chronicle*, October 23, 2002) summarizes the findings of a study that appeared in the journal *Neurology*. Doctors at Stanford Medical Center were interested in determining whether a new surgical approach to treating Alzheimer’s disease results in improved memory functioning. The surgical procedure involves implanting a thin tube, called a shunt, which is designed to drain toxins from the fluid-filled space that cushions the brain. Eleven patients had shunts implanted and were followed for a year, receiving quarterly tests of memory function. Another sample of Alzheimer’s patients was used as a comparison group. Those in the comparison group received the standard care for Alzheimer’s disease. After analyzing the data from this study, the investigators concluded that the “results suggested the treated patients essentially held their own in the cognitive tests while the patients in the control group steadily declined. However, the study was too small to produce conclusive statistical evidence.”
- What were the researchers trying to learn? What questions motivated their research?
 - Do you think that the study was conducted in a reasonable way? What additional information would you want in order to evaluate this study?
- 1.11** In a study of whether taking a garlic supplement reduces the risk of getting a cold, participants were assigned to either a garlic supplement group or to a group that did not take a garlic supplement (“**Garlic for the Common Cold**,” *Cochrane Database of Systematic Reviews*, 2009). Based on the study, it was concluded that the proportion of people taking a garlic supplement who get a cold is lower than the proportion of those not taking a garlic supplement who get a cold.
- What were the researchers trying to learn? What questions motivated their research?
 - Do you think that the study was conducted in a reasonable way? What additional information would you want in order to evaluate this study?

SECTION 1.4 Types of Data and Some Simple Graphical Displays

Every discipline has its own particular way of using common words, and statistics is no exception. You will recognize some of the terminology from previous math and science courses, but much of the language of statistics will be new to you. In this section, you will learn some of the terminology used to describe data.

Types of Data

The individuals or objects in any particular population might possess many characteristics that could be studied. Consider a group of students currently enrolled in a statistics class. One characteristic of the students in the population is the brand of calculator they use (Casio, Hewlett-Packard, Sharp, Texas Instruments, and so on). Another characteristic is the number of textbooks purchased that semester, and yet another is the distance from the college to each student's home. A **variable** is any characteristic whose value may change from one individual or object to another. For example, *calculator brand* is a variable, and so are *number of textbooks purchased* and *distance to the college*. **Data** result from making observations either on a single variable or on two or more variables at the same time.

DEFINITIONS

Variable: A characteristic whose value may change from one observation to another.

Data: A collection of observations on one or more variables.

A **univariate data set** consists of observations on a single variable made on individuals in a sample or population. There are two types of univariate data sets: **categorical** (sometimes also called qualitative) and **numerical** (sometimes also called quantitative). In the previous example, *calculator brand* is a categorical variable, because each student's response to the query, "What brand of calculator do you use?" is a category. The collection of responses from all these students forms a categorical data set. The other two variables, *number of textbooks purchased* and *distance to the college*, are both numerical in nature. Determining the values of a numerical variable (by counting or measuring) results in a numerical data set.

DEFINITIONS

Univariate data set: A data set consisting of observations on a single characteristic.

Categorical data set: A univariate data set is **categorical** (or **qualitative**) if the individual observations are categorical responses.

Numerical data set: A univariate data set is **numerical** (or **quantitative**) if each observation is a number.

Example 1.4 College Choice Do-Over?

Understand the context ➤

The Higher Education Research Institute at UCLA surveys over 20,000 college seniors each year. One question on the survey asks seniors the following question: If you could make your college choice over, would you still choose to enroll at your current college? Possible responses are definitely yes (DY), probably yes (PY), probably no (PN), and definitely no (DN). Responses for 20 students were:

DY PN DN DY PY PY PN PY PY DY
DY PY DY DY PY PY DY DY PN DY

Consider the data ➤

(These data are just a small subset of the data from the survey.) Because the response to the question about college choice is categorical, this is a univariate categorical data set.

In Example 1.4, the data set consists of observations on a single variable (college choice response), so this is a univariate data set. In some studies, we are interested in two different characteristics. For example, both height (in inches) and weight (in pounds) might be recorded for each person on a basketball team. The resulting data set consists of pairs of numbers, such as (74, 185). This is called a **bivariate data set**. **Multivariate data** result from recording a value for each of two or more attributes (so bivariate data are a special case of multivariate data). For example, multivariate data would result from determining height, weight, pulse rate, and blood pressure for each person on a basketball team. Example 1.5 illustrates a bivariate data set.

Example 1.5 How Safe Are College Campuses?

Understand the context ➤

- Consider the accompanying data on violent crime on public college campuses in Florida during 2016 (fbi.gov, retrieved February 6, 2018).

Consider the data ➤

University/College	Student Enrollment	Number of Violent Crimes Reported in 2016
Florida A&M University	9,928	6
Florida Atlantic University	30,380	6
Florida Gulf Coast University	14,833	3
Florida International University	49,782	15
Florida South Western State College	15,709	0
Florida State University, Tallahassee	40,830	20
New College of Florida	861	1
Pensacola State College	9,840	0
Santa Fe College	14,767	0
Tallahassee Community College	12,445	2
University of Central Florida	62,953	14
University of Florida	50,645	10
University of North Florida	15,675	3
University of South Florida, St. Petersburg	4,739	0
University of South Florida, Tampa	42,067	8
University of West Florida	12,763	10

Here two variables—*student enrollment* and *number of violent crimes reported*—were recorded for each of the 16 schools. Because this data set consists of values of two variables for each school, it is a bivariate data set. Each of the two variables considered here is numerical (rather than categorical).

Two Types of Numerical Data

There are two different types of numerical data: **discrete** and **continuous**. Consider a number line (Figure 1.3) for locating values of a numerical variable. Each possible number (2, 3.125, 8.12976, etc.) corresponds to exactly one point on the number line.

Suppose that the variable of interest is the number of courses in which a student is enrolled. If no student is enrolled in more than eight courses, the possible values are 1, 2, 3, 4, 5, 6, 7, and 8. These values are identified in Figure 1.4(a) by the dots at the points marked 1, 2, 3, 4, 5, 6, 7, and 8. These possible values are isolated from one another on the number line. We can place an interval around any possible value that is small enough that no other possible value is included in the interval. On the other hand, the line segment in Figure 1.4(b) identifies a plausible set of possible values for the time (in seconds) it takes for the first kernel in a bag of microwave popcorn to pop. Here the possible values make up an entire interval on the number line, and no possible value is isolated from other possible values.

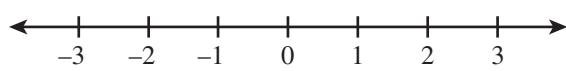


FIGURE 1.3

A number line.

- Data set available online

FIGURE 1.4

Possible values of a variable:
 (a) number of courses;
 (b) popping time (in seconds).



DEFINITIONS

Discrete numerical variable: A numerical variable results in **discrete** data if the possible values of the variable correspond to isolated points on the number line.

Continuous numerical variable: A numerical variable results in **continuous** data if the set of possible values forms an entire interval on the number line.

Discrete data usually arise when observations are determined by counting (for example, the number of roommates a student has or the number of petals on a flower).

Example 1.6 Do U Txt?

- The number of text messages sent on a particular day is recorded for each of 12 students. The resulting data set is

23 0 14 13 15 0 60 82 0 40 41 22

Possible values for the variable *number of text messages sent* are 0, 1, 2, 3, These are isolated points on the number line, so this data set consists of discrete numerical data.

Suppose that instead of the number of text messages sent, the *time spent texting* had been recorded. Even though time spent may have been reported rounded to the nearest minute, the actual time spent could have been 6 minutes, 6.2 minutes, 6.28 minutes, or any other value in an entire interval. So, recording values of *time spent texting* would result in continuous data.

In general, data are continuous when observations involve making measurements, as opposed to counting. In practice, measuring instruments do not have infinite accuracy, so possible measured values, strictly speaking, do not form an entire interval on the number line. However, any number in the interval *could* be a value of the variable. The distinction between discrete and continuous data will be important in our discussion of probability models in Chapter 6.

Frequency Distributions and Bar Charts for Categorical Data

An appropriate graphical or tabular display of data can be an effective way to summarize and communicate information. When a data set is categorical, a common way to present the data is in the form of a table, called a **frequency distribution**.

DEFINITIONS

Frequency distribution for categorical data: A table that displays the possible categories along with the associated frequencies and/or relative frequencies.

Frequency: The **frequency** for a particular category is the number of times the category appears in the data set.

Relative frequency: The **relative frequency** for a particular category is calculated as

$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations in the data set}}$$

The relative frequency for a particular category is the proportion of the observations that belong to that category.

Relative frequency distribution: A frequency distribution that includes relative frequencies.

Example 1.7 | Motorcycle Helmets—Can You See Those Ears?

Understand the context ➤

The U.S. Department of Transportation establishes standards for motorcycle helmets. To ensure a certain degree of safety, helmets should reach the bottom of the motorcyclist's ears. The report **"Motorcycle Helmet Use in 2014—Overall Results" (National Highway Traffic Safety Administration, January 2015)** summarized data collected in 2014 by observing 806 motorcyclists nationwide at selected roadway locations. Each time a motorcyclist passed by, the observer noted whether the rider was wearing no helmet, a noncompliant helmet, or a compliant helmet. The following coding was used:

N = no helmet

NH = noncompliant helmet

CH = compliant helmet

Consider the data ➤

A few of the observations were

CH N CH NH N CH CH CH N N

There were also 796 additional observations, which we didn't reproduce here. In total, there were 250 riders who wore no helmet, 40 who wore a noncompliant helmet, and 516 who wore a compliant helmet.

The corresponding frequency distribution is given in Table 1.1.

Do the work ➤

TABLE 1.1 Frequency Distribution for Helmet Use

Helmet Use Category	Frequency	Relative Frequency
No helmet	250	0.310 ← 250/806
Noncompliant helmet	40	0.050 ← 40/806
Compliant helmet	516	0.640
	806	1.000 ← Total number of observations

Should total 1, but in some cases may be slightly off due to rounding

Interpret the results ➤

From the frequency distribution, we can see that a large percentage of riders (31%) were not wearing a helmet, but most of those who wore a helmet were wearing one that met the Department of Transportation safety standard.

A frequency distribution summarizes a data set in a table. It is also common to display categorical data graphically. A bar chart is one of the most widely used types of graphical displays for categorical data.

Bar Charts

A **bar chart** is a graph of a frequency distribution for categorical data. Each category in the frequency distribution is represented by a bar or rectangle, and the picture is constructed in such a way that the *area* of each bar is proportional to the corresponding frequency or relative frequency.

Bar Charts

When to Use Categorical data.

How to Construct

1. Draw a horizontal axis, and write the category names or labels below the line at equally spaced intervals.
2. Draw a vertical axis, and label the scale using either frequency or relative frequency.

(continued)

3. Place a rectangular bar above each category label. The height is determined by the category's frequency or relative frequency, and all bars should have the same width. With the same width, both the height and the area of the bar are proportional to frequency and relative frequency.

What to Look For

- Frequently and infrequently occurring categories.

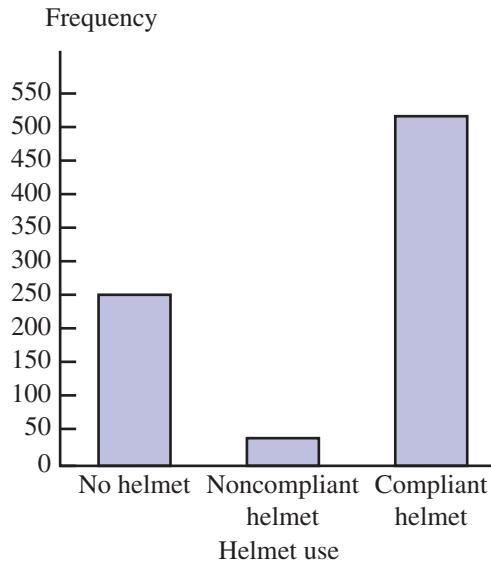
Example 1.8 Revisiting Motorcycle Helmets

Understand the context ➤

Example 1.7 used data on helmet use from a sample of 806 motorcyclists to construct a frequency distribution (Table 1.1). Figure 1.5 shows the bar chart corresponding to this frequency distribution.

FIGURE 1.5

Bar chart of helmet use.



Interpret the results ➤

The bar chart provides a visual representation of the information in the frequency distribution. From the bar chart, it is easy to see that the compliant helmet use category occurred most often in the data set. The bar for compliant helmets is about twice as tall (and therefore has twice the area) as the bar for no helmet because approximately twice as many motorcyclists wore compliant helmets than wore no helmet.

Dotplots for Numerical Data

A dotplot is a simple way to display numerical data when the data set is reasonably small. Each observation is represented by a dot above the location corresponding to its value on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence and these dots are stacked vertically.

Dotplots

When to Use Small numerical data sets.

How to Construct

1. Draw a horizontal line and mark it with an appropriate measurement scale.
2. Locate each value in the data set along the measurement scale, and represent it by a dot. If there are two or more observations with the same value, stack the dots vertically.

(continued)

What to Look for

Dotplots convey information about:

- A representative or typical value in the data set.
- The extent to which the data values vary.
- The shape of the distribution of values along the number line.
- The presence of unusual values in the data set.

Example 1.9 Making It to Graduation . . .

Understand the context ➤

- The article “[Keeping Score When It Counts: Graduation Success and Academic Progress Rates for the 2016 NCAA Men’s Division I Basketball Tournament Teams](#)” ([The Institute for Diversity and Ethics in Sport, University of Central Florida, March 2016](#)) compared graduation rates of basketball players to those of all student athletes for the universities and colleges that sent teams to the 2016 Division I playoffs. The graduation rates in the accompanying table are the percentage of athletes who started college in 2005, 2006, 2007, and 2008 who graduated within 6 years.

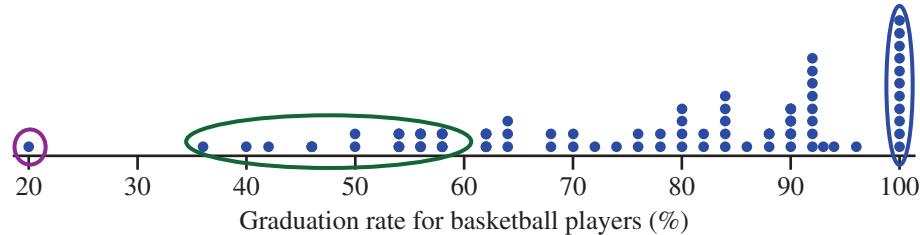
Consider the data ➤

Minitab, a computer software package for statistical analysis, was used to construct a dotplot of the graduation rates for basketball players (see Figure 1.6). From this dotplot, we see that basketball graduation rates varied a great deal from school to school, ranging from a low of 20% to a high of 100%.

We can also see that the graduation rates seem to cluster in several groups, denoted by the colored ovals that have been added to the dotplot. There are 11 schools with graduation rates of 100% (excellent!). The majority of schools are in the large cluster with graduation rates from about 62% to about 96%. And then there is a group of 12 schools with low graduation rates for basketball players and one school with an unusually low graduation rate of 20%.

FIGURE 1.6

Minitab dotplot of graduation rates for basketball players.



School	ALL	BB	Difference	School	ALL	BB	Difference
1	79	93	-14	17	82	78	4
2	88	91	-3	18	91	70	21
3	88	100	-12	19	84	85	-1
4	71	73	-2	20	92	88	4
5	74	57	17	21	90	70	20
6	98	100	-2	22	83	80	3
7	98	100	-2	23	60	42	18
8	71	77	-6	24	60	39	21
9	69	53	16	25	81	91	-10
10	97	88	9	26	90	55	35
11	67	58	9	27	85	82	3
12	87	67	20	28	78	54	24
13	90	92	-2	29	79	92	-13
14	80	75	5	30	78	80	-2
15	87	63	24	31	83	92	-9
16	87	100	-13	32	78	80	-2

● Data set available online

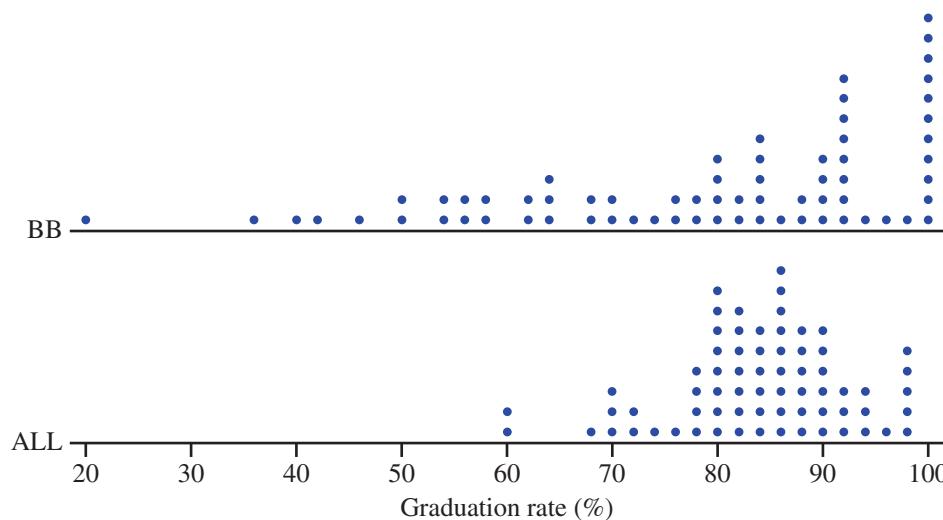
(Continued)

School	ALL	BB	Difference	School	ALL	BB	Difference
33	79	55	24	51	80	50	30
34	79	36	43	52	82	62	20
35	86	83	3	53	81	82	-1
36	85	20	65	54	70	50	20
37	95	100	-5	55	85	100	-15
38	78	62	16	56	87	83	4
39	89	100	-11	57	83	90	-7
40	84	100	-16	58	86	64	22
41	81	90	-9	59	92	91	1
42	85	91	-6	60	85	67	18
43	89	93	-4	61	93	83	10
44	89	89	0	62	94	100	-6
45	85	90	-5	63	76	83	-7
46	85	80	5	64	69	100	-31
47	81	46	35	65	82	83	-1
48	80	75	5	66	80	63	17
49	98	100	-2	67	94	91	3
50	84	71	13	68	98	95	3

Figure 1.7 shows two dotplots of graduation rates—one for basketball players and one for all student athletes. There are some striking differences that are easy to see when the data is displayed in this way. The graduation rates for all student athletes tend to be higher and to vary less from school to school than the graduation rates for only basketball players.

FIGURE 1.7

Minitab dotplot of graduation rates for basketball players and for all athletes.



The dotplots in Figure 1.7 are informative, but we can do even better. The data given here are an example of *paired data*. Each basketball graduation rate is paired with a graduation rate for all student athletes from the same school. When data are paired in this way, it is more informative to look at differences—in this case, the difference between the graduation rate for all student athletes and for basketball players for each school. These differences (All – Basketball) are also shown in the data table.

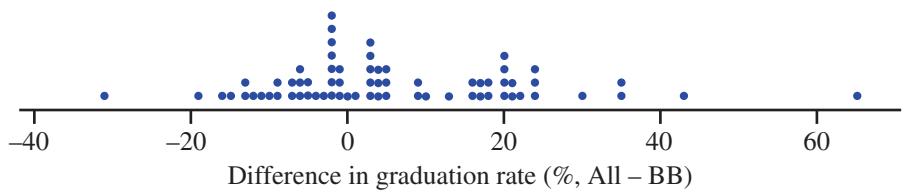
Interpret the results ➤

Figure 1.8 gives a dotplot of the differences. Notice that one difference is equal to 0. This corresponds to a school for which the basketball graduation rate is equal to the graduation rate of all student athletes. There are 30 schools for which the difference is negative. Negative differences correspond to schools that have a higher graduation rate for basketball players than for all student athletes.

The most interesting feature of the difference dotplot is the variability in the positive differences. Positive differences correspond to schools that have a lower graduation rate for basketball players. The positive differences range from 1% all the way up to 65%. There were five schools that had a graduation rate for all athletes that was 30 percentage points or more greater than the graduation rate for basketball players. (In case you were wondering, these schools were University of Oregon with a difference of 30%, Syracuse University and University of North Carolina Wilmington with differences of 35%, University of Cincinnati with a difference of 43%, and University of Connecticut with a difference of 65%).

FIGURE 1.8

Dotplot of graduation rate differences (all athletes – basketball players).



EXERCISES 1.12 - 1.32

● Data set available online

- 1.12** Classify each of the following variables as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.
- Number of students in a class of 35 who turn in a term paper before the due date
 - Gender of the next baby born at a particular hospital
 - Amount of fluid (in ounces) dispensed by a machine used to fill bottles with soda pop
 - Thickness of the gelatin coating of a vitamin E capsule
 - Birth order classification (only child, firstborn, middle child, lastborn) of a math major
- 1.13** Classify each of the following variables as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.
- Brand of computer purchased by a customer
 - State of birth for someone born in the United States
 - Price of a textbook
 - Concentration of a contaminant (micrograms per cubic centimeter) in a water sample
 - Zip code (Think carefully about this one.)
 - Actual weight of coffee in a 1-pound can, labeled as containing 1 pound
- 1.14** For the following numerical variables, state whether each is discrete or continuous.
- The number of insufficient-funds checks received by a grocery store during a given month
 - The amount by which a 1-pound package of ground beef decreases in weight (because of moisture loss) before purchase
- 1.15** For the following numerical variables, state whether each is discrete or continuous.
- The length of a 1-year-old rattlesnake
 - The altitude of a location in California selected randomly by throwing a dart at a map of the state
 - The distance from the left edge at which a 12-inch plastic ruler snaps when bent sufficiently to break
 - The price per gallon paid by the next customer to buy gas at a particular station
- 1.16** For each of the following situations, give a set of possible data values that might arise from making the observations described.
- The manufacturer for each of the next 10 automobiles to pass through a given intersection is noted.
 - The grade point average for each of the 15 seniors in a statistics class is determined.
 - The number of gas pumps in use at each of 20 gas stations at a particular time is determined.
 - The actual net weight of each of 12 bags of fertilizer having a labeled weight of 50 pounds is determined.
 - Fifteen different radio stations are monitored during a 1-hour period, and the amount of time devoted to commercials is determined for each.

- 1.17** In a survey of 100 people who had recently purchased motorcycles, data on the following variables were recorded:

Sex of purchaser
 Brand of motorcycle purchased
 Number of previous motorcycles owned by purchaser
 Telephone area code of purchaser
 Weight of motorcycle as equipped at purchase

- Which of these variables are categorical?
- Which of these variables are discrete numerical?
- Which type of graphical display would be an appropriate choice for summarizing the sex data, a bar chart or a dotplot?
- Which type of graphical display would be an appropriate choice for summarizing the weight data, a bar chart or a dotplot?

- 1.18** The Gallup report “[More Americans Say Real Estate Is Best Long-Term Investment](#)” ([gallup.com, April 20, 2016, retrieved April 15, 2017](#)) included data from a poll of 1015 adults. The responses to the question “What do you think is the best long-term investment?” are summarized in the given relative frequency distribution.

Response	Relative Frequency
Real Estate	0.35
Stocks & Mutual Funds	0.22
Gold	0.17
Savings	0.15
Bonds	0.07
Other	0.04

- Use this information to construct a bar chart for the response data.
- Write a few sentences commenting on the distribution of responses to the questions posed.

- 1.19** An article in the [New Times San Luis Obispo \(February 4, 2016\)](#) reported the accompanying concussion rates for different high school sports. The given data are concussion rates per 10,000 athletes participating in high school sports in 2012.

Sport	Concussion Rate (Concussions per 10,000 athletes)
Football	11.2
Lacrosse (Boys)	6.9
Lacrosse (Girls)	5.2
Wrestling	6.2
Basketball (Girls)	5.6
Basketball (Boys)	2.8
Soccer (Girls)	6.7

(continued)

Sport	Concussion Rate (Concussions per 10,000 athletes)
Soccer (Boys)	4.2
Field Hockey	4.2
Volleyball	2.4
Softball	1.6
Baseball	1.2

- Construct a dotplot of the concussion rate data.
- In addition to the three girls’ sports indicated in the table (lacrosse, basketball, and soccer), the reported concussion rates for field hockey, volleyball, and softball are also for girls. Locate the points on the dotplot that correspond to concussion rates for girls’ sports and highlight them in a different color. Based on the dotplot, would you say that the concussion rates for girls’ sports tend to be lower than or higher than for boys’ sports? Explain.

- 1.20** ● Box Office Mojo ([boxofficemojo.com](#)) tracks movie ticket sales. Ticket sales (in millions of dollars) for each of the top 20 movies in 2014 and 2015 are shown in the accompanying table.

Movie	2014
	2014 Sales (millions of dollars)
American Sniper	350.1
The Hunger Games: Mockingjay—Part 1	337.1
Guardians of the Galaxy	333.1
Captain America: The Winter Soldier	259.8
The LEGO Movie	257.8
The Hobbit: The Desolation of Smaug	255.1
Transformers: Age of Extinction	245.4
Maleficent	241.4
X-Men: Days of Future Past	233.9
Big Hero 6	222.5
Dawn of the Planet of the Apes	208.5
The Amazing Spider-Man 2	202.9
Godzilla	200.7
22 Jump Street	191.7
Teenage Mutant Ninja Turtles	191.2
Interstellar	188.0
How to Train Your Dragon 2	177.0
Gone Girl	167.8
Divergent	150.9
Neighbors	150.2

(continued)

2015	
Movie	2015 Sales (millions of dollars)
Star Wars: The Force Awakens	936.7
Jurassic World	652.3
Avengers: Age of Ultron	459.0
Inside Out	356.5
Furious 7	353.0
Minions	336.0
The Hunger Games: Mockingjay—Part 2	281.7
The Martian	228.4
Cinderella	201.2
Spectre	200.1
Mission Impossible—Rogue Nation	195.0
Pitch Perfect 2	184.3
The Revenant	183.6
Ant-Man	180.2
Home	177.4
Hotel Transylvania 2	169.7
Fifty Shades of Grey	166.2
The SpongeBob Movie: Sponge Out of Water	163.0
Straight Outta Compton	161.2
San Andreas	155.2

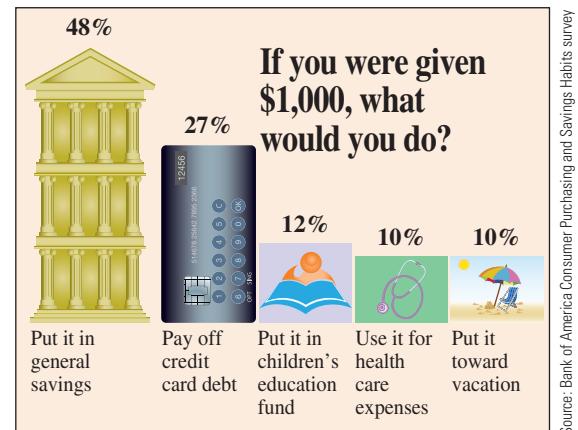
- 1.21** The report “With Their Whole Lives Ahead of Them” (publicagenda.org/files/theirwhole_livesaheadofthem.pdf, retrieved February 6, 2018) includes data from a survey of 200 students who started college but did not complete a degree. Each of these students was asked, “How much have you thought about going back to school?” The accompanying frequency distribution summarizes the responses to this question.

Response	Frequency
A lot of thought	130
Some thought	48
No thought	18
Don’t know	4

- a.** Summarize the response data using a bar chart.
b. Write a few sentences commenting on the distribution of the responses.

- 1.22** The following display is a graph similar to one that appeared in *USA TODAY* (June 29, 2009). This graph is meant to be a bar graph of responses to the question shown in the graph.

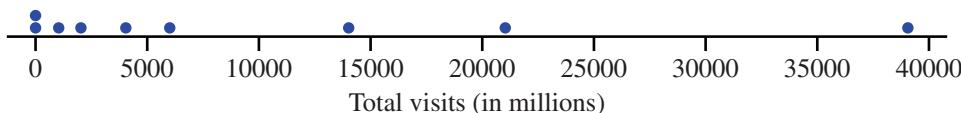
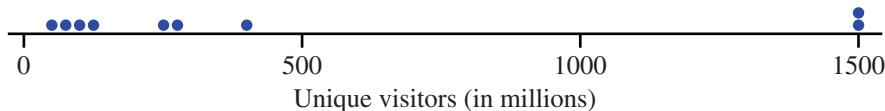
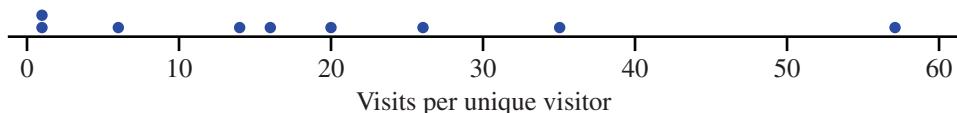
- Is response to the question a categorical or numerical variable?
- Explain why a bar chart rather than a dotplot was used to display the response data.
- There must have been an error made in constructing this graph. How can you tell that the graph is not a correct representation of the response data?



- 1.23** The accompanying table gives the total number visits and the number of unique visitors for some popular social networking sites in the United States for the month of July 2017. The number of unique visitors data are taken from the online article “[Top 15 Most Popular Social Networking Sites](http://ism.org/images/files/Social-media-platforms-from-Engage-to-Succeed-webinar.pdf)” (ism.org/images/files/Social-media-platforms-from-Engage-to-Succeed-webinar.pdf, retrieved February 7, 2018). The total number of visits were estimated using data from semrush.com (retrieved February 7, 2018). The data on total visits and unique visitors were used to compute the values in the final column of the data table, in which

$$\text{visits per unique visitor} = \frac{\text{total visits}}{\text{number of unique visitors}}$$

Site	Unique Visitors (in millions)	Total Visits (in millions)	Visits per Unique Visitor
Facebook	1,500	20,730	13.8
YouTube	1,499	39,000	26.0
Twitter	400	13,860	34.7
Instagram	275	4,320	15.7
LinkedIn	250	1,398	5.6
reddit	125	2,445	19.6
Pinterest	105	5,970	56.9
flickr	80	84	1.1
meetup	42	45	1.1

Figure for Exercise 1.23a**Figure for Exercise 1.23b****Figure for Exercise 1.23c**

- A dotplot of the total visits data is shown at the top of the page. What are the most obvious features of the dotplot? What does it tell you about the online social networking sites?
- A dotplot for the number of unique visitors is shown at the top of the page. In what way is this dotplot different from the dotplot for total visits in Part (a)? What does this tell you about the online social networking sites?
- A dotplot for the visits per unique visitor data is shown at the top of the page. What new information about the online social networks is provided by this dotplot?

- 1.24** Heal the Bay is an environmental organization that releases an annual beach report card based on water quality ([Heal the Bay Beach Report Card, beachreportcard.org](http://HealtheBayBeachReportCard.beachreportcard.org), retrieved May 7, 2016). The grades for 20 beaches in three counties in Washington (Whatcom, Snohomish, and Island counties) during dry weather were:

C B A+ A+ A+ A+ A+ A+ A+ A+
A+ C A+ A+ A+ C A F F B

Summarize the dry weather grades by constructing a relative frequency distribution and a bar chart.

- 1.25** The report referenced in the previous exercise also gave wet weather grades for the same beaches:

A+ A+ A+ A+ A+ A+ A+ F F F
A+ A+ F A+ A+ F A+ A+ F A+

- Construct a bar chart for the wet weather grades.
- Do the bar charts from Part (a) and from the previous exercise support the statement that beach water quality tends to be better in dry weather conditions? Explain.

- 1.26** ● The U.S. Department of Health and Human Services reported the estimated percentage of households with only wireless phone service (no landline) in 2014 for the 50 states and the District

of Columbia (cdc.gov/nchs/data/nhis/earlyrelease/wireless_state_201602.pdf, retrieved February 8, 2018). In the accompanying data table, each state was also classified into one of three geographical regions—West (W), Middle states (M), and East (E).

Wireless %	Region	State
43.4	M	AL
39.7	W	AK
49.4	W	AZ
56.2	M	AR
42.8	W	CA
50.5	W	CO
26.7	E	CT
29.4	E	DE
49.7	E	DC
47.6	E	FL
45.9	E	GA
38.3	W	HI
56.1	W	ID
45.7	M	IL
47.7	M	IN
50.7	M	IA
51.6	M	KS
47.1	M	KY
40.9	M	LA
40.8	E	ME
36.2	E	MD
31.5	E	MA
47.8	M	MI
43.1	M	MN
55.1	M	MS
51.5	M	MO
41.0	W	MT
46.5	M	NE
48.4	W	NV
43.6	M	ND
31.2	E	NH
25.1	E	NJ

(continued)

Wireless %	Region	State
47.0	W	NM
31.1	E	NY
42.9	E	NC
45.8	E	OH
50.4	M	OK
47.0	W	OR
30.0	E	PA
34.6	E	RI
49.5	E	SC
41.4	M	SD
46.6	M	TN
54.6	M	TX
52.2	W	UT
41.1	E	VA
37.2	E	VT
48.3	W	WA
37.2	E	WV
46.6	M	WI
51.8	W	WY

- a. Display the data graphically in a way that makes it possible to compare wireless percent for the three geographical regions.
- b. Does the graphical display in Part (a) reveal any striking differences in wireless percent for the three geographical regions or are the distributions of wireless percent observations similar for the three regions?
- 1.27 ● Example 1.5 gave the accompanying data on violent crime on public college campuses in Florida during 2016 ([fbi.gov, retrieved February 6, 2018](#)):

University/College	Student Enrollment	Number of Violent Crimes Reported in 2016
Florida A&M University	9,928	6
Florida Atlantic University	30,380	6
Florida Gulf Coast University	14,833	3
Florida International University	49,782	15
Florida South Western State College	15,709	0
Florida State University, Tallahassee	40,830	20
New College of Florida	861	1
Pensacola State College	9,840	0
Santa Fe College	14,767	0
Tallahassee Community College	12,445	2
University of Central Florida	62,953	14

(continued)

University/College	Student Enrollment	Number of Violent Crimes Reported in 2016
University of Florida	50,645	10
University of North Florida	15,675	3
University of South Florida, St. Petersburg	4,739	0
University of South Florida, Tampa	42,067	8
University of West Florida	12,763	10

- a. Construct a dotplot using the 16 observations on number of violent crimes reported. Which schools stand out from the rest?
- b. One of the Florida schools has only 861 students and a few of the schools are quite a bit larger than the rest. Because of this, it might make more sense to consider a crime rate by calculating the number of violent crimes reported per 1000 students. For example, for Florida A&M University the violent crime rate would be

$$\frac{6}{9928}(1000) = (0.001)(1000) = 1.0$$

Calculate the violent crime rate for the other 15 schools and then use those values to construct a dotplot. Do the same schools stand out as unusual in this dotplot?

- c. Based on your answers from Parts (a) and (b), write a couple of sentences commenting on violent crimes reported at Florida universities and colleges in 2016.

- 1.28 ● The article “[Fliers Trapped on Tarmac Push for Rules on Release](#)” (*USA TODAY*, July 28, 2009) gave the following data for 17 airlines on number of flights that were delayed on the tarmac for at least 3 hours for the period from October 2008 to May 2009:

Airline	Number of Delays	Rate per 10,000 Flights
ExpressJet	93	4.9
Continental	72	4.1
Delta	81	2.8
Comair	29	2.7
American Eagle	44	1.6
US Airways	46	1.6
JetBlue	18	1.4
American	48	1.3
Northwest	24	1.2

(continued)

Airline	Number of Delays	Rate per 10,000 Flights
Mesa	17	1.1
United	29	1.1
Frontier	5	0.9
SkyWest	29	0.8
Pinnacle	13	0.7
Atlantic Southeast	11	0.6
AirTran	7	0.4
Southwest	11	0.1

The graph at the bottom of the page shows two dotplots: one displays the number of delays data, and one displays the rate per 10,000 flights data.

- a. If you were going to rank airlines based on flights delayed on the tarmac for at least 3 hours, would you use the *total number of flights* data or the *rate per 10,000 flights* data? Explain the reason for your choice.
- b. Write a short paragraph that could be used as part of a newspaper article on flight delays that could accompany the dotplot of the *rate per 10,000 flights* data.
- 1.29 The report “[Trends in Community Colleges](#)” (collegeboard.com/trends April 2016, trends.collegeboard.org/sites/default/files/trends-in-community-colleges-research-brief.pdf, retrieved February 8, 2018) included the accompanying information on student debt for students graduating with an AA degree from a public community college in 2012.

Debt	Relative Frequency
None	0.59
Less than \$10,000	0.20
Between \$10,000 and \$20,000	0.12
More than \$20,000	0.09

- a. Use the given information to construct a bar chart.
- b. Write a few sentences commenting on student debt for public community college graduates.

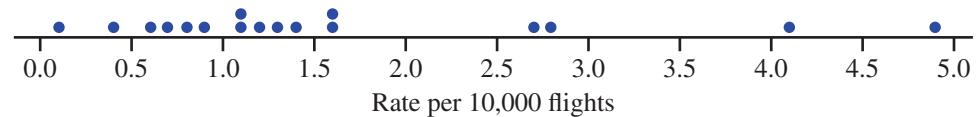
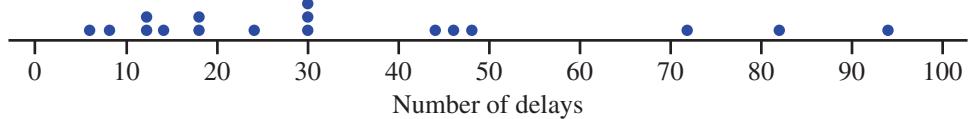
- 1.30 The article “[Where College Students Buy Textbooks](#)” ([USA TODAY](#), October 14, 2010) gave data on where students purchased books. The accompanying frequency table summarizes data from a sample of 1152 full-time college students.

Where Books Purchased	Frequency
Campus bookstore	576
Campus bookstore web site	48
Online bookstore other than campus bookstore	240
Off-campus bookstore	168
Rented textbooks	36
Purchased mostly eBooks	12
Didn't buy any textbooks	72

- a. Construct a bar chart to summarize the data distribution.
- b. Write a few sentences commenting on where students are buying textbooks.

- 1.31 The report [2013 International Bedroom Poll: Summary of Findings](#) describes a survey of 251 adult Americans conducted by the National Sleep Foundation (sleepfoundation.org/sites/default/files/RPT495a.pdf, retrieved April 15, 2017). Participants in the survey were asked how often they change the sheets on their bed and were asked to respond with one of the following categories: more than once a week, once a week, every other week, every three weeks, or less often than every three weeks. For this group, 10% responded more than once a week, 53% responded once a week, 26% responded every other week, 5% responded every three weeks, and 6% responded less often than every three weeks.
- a. Use the given information to make a relative frequency distribution for the responses to the question.
- b. Summarize the given information by constructing a bar chart.
- 1.32 In the United States, movies are rated by the Motion Picture Association of America (MPAA). The accompanying table gives the MPAA rating of the 25 top money-making movies of 2015

Figure For Exercise 1.28



(data from boxofficemojo.com, retrieved October 10, 2016).

Movie (Ordered from high to low based on amount of money made)	Rating
Star Wars: The Force Awakens	PG-13
Jurassic World	PG-13
Avengers: Age of Ultron	PG-13
Inside Out	PG
Furious 7	PG-13
Minions	PG
The Hunger Games: Mockingjay—Part 2	PG-13
The Martian	PG-13
Cinderella	PG
Spectre	PG-13
Mission Impossible: Rogue Nation	PG-13
Pitch Perfect 2	PG-13
The Revenant	R

(continued)

Movie (Ordered from high to low based on amount of money made)	Rating
Ant-Man	PG-13
Home	PG
Hotel Transylvania 2	PG
Fifty Shades of Grey	R
The SpongeBob Movie: Sponge Out of Water	PG
Straight Outta Compton	R
San Andreas	PG-13
Mad Max: Fury Road	R
Daddy's Home	PG-13
The Divergent Series: Insurgent	PG-13
Peanuts Movie	G
Kingsman: The Secret Service	R

Use the given information to construct a bar chart of the ratings for the top 25 movies of 2015. Write a few sentences describing the ratings distribution.

CHAPTER ACTIVITIES

ACTIVITY 1.1 HEAD SIZES: UNDERSTANDING VARIABILITY

Materials needed: Each team will need a measuring tape. For this activity, you will work in teams of 6 to 10 people.

1. Designate a team leader for your team by choosing the person on your team who celebrated his or her last birthday most recently.
2. The team leader should measure and record the head size (measured as the circumference at the widest part of the forehead) of each of the other members of his or her team.
3. Record the head sizes for the individuals on your team as measured by the team leader.
4. Next, each individual on the team should measure the head size of the team leader. Do not share your measurement with the other team members until all team members have measured the team leader's head size.
5. After all team members have measured the team leader's head, record the different team leader head size measurements obtained by the individuals on your team.
6. Using the data from Step 3, construct a dotplot of the team leader's measurements of team head sizes. Then, using the same scale, construct a separate dotplot of the different measurements of the team leader's head size (from Step 5).

Now use the available information to answer the following questions:

7. Do you think the team leader's head size changed in between measurements? If not, explain why the measurements of the team leader's head size are not all the same.
8. Which data set was more variable—head size measurements of the different individuals on your team or the different measurements of the team leader's head size? Explain the basis for your choice.
9. Consider the following scheme (you don't actually have to carry this out): Suppose that a group of 10 people measured head sizes by first assigning each person in the group a number between 1 and 10. Then person 1 measured person 2's head size, person 2 measured person 3's head size, and so on, with person 10 finally measuring person 1's head size. Do you think that the resulting head size measurements would be more variable, less variable, or show about the same amount of variability as a set of 10 measurements resulting from a single individual measuring the head size of all 10 people in the group? Explain.

ACTIVITY 1.2 ESTIMATING SHAPE SIZES

1. Construct an activity sheet that consists of a table that has 6 columns and 10 rows. Label the columns of the table with the following six headings: (1) Shape, (2) Estimated Size, (3) Actual Size, (4) Difference (Estimated – Actual), (5) Absolute Difference, and (6) Squared Difference. Enter the numbers from 1 to 10 in the “Shape” column.
2. Next you will be visually estimating the sizes of the shapes in Figure 1.9. Size will be described as the number of squares of this size



that would fit in the shape. For example, the shape



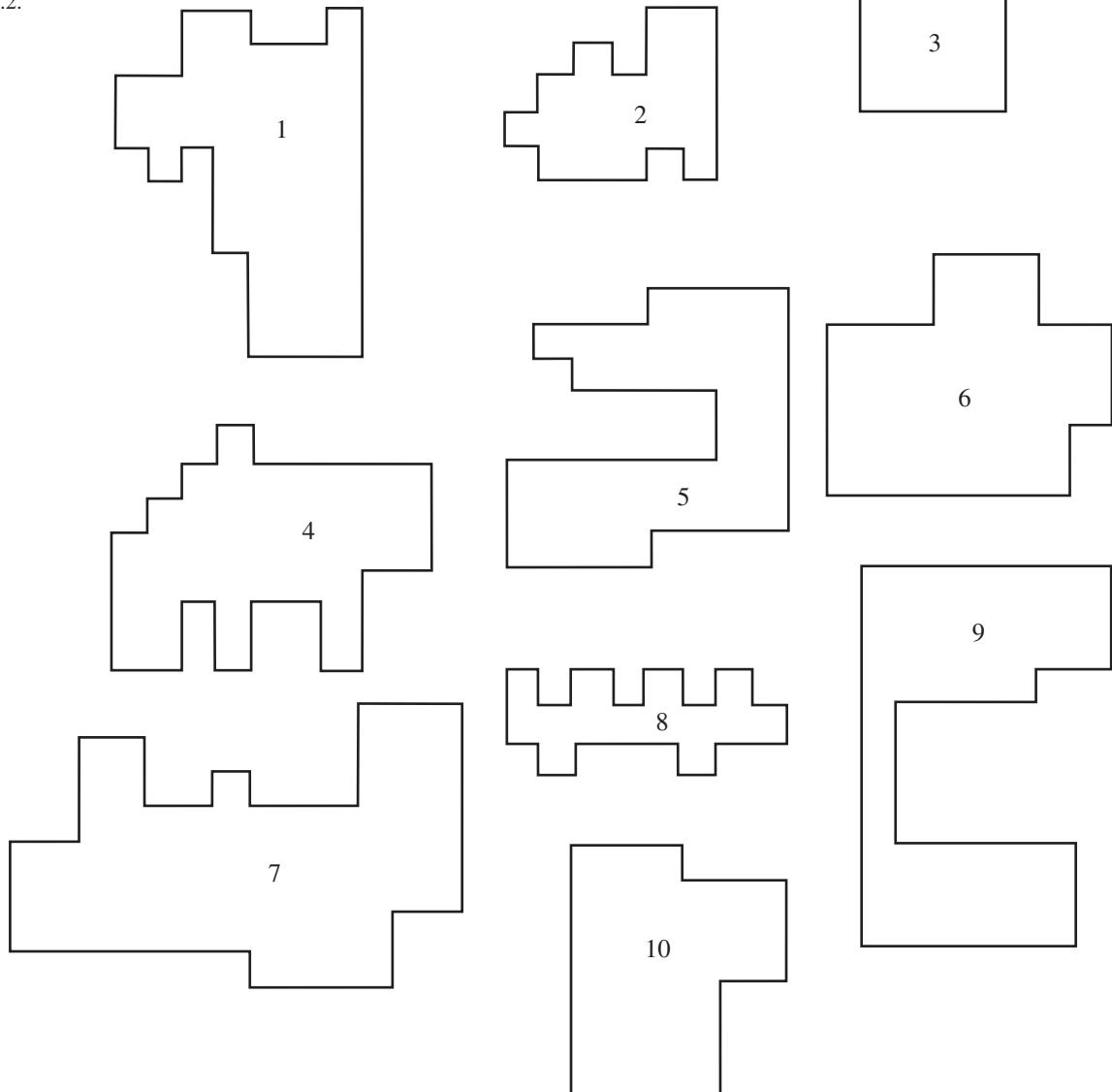
would be size 3, as illustrated by



You should now quickly *visually* estimate the sizes of the shapes in Figure 1.9. *Do not* draw on the figure—these are to be quick visual estimates. Record your estimates in the “Estimated Size” column of the activity sheet.

3. Your instructor will provide the actual sizes for the 10 shapes, which should be entered into the “Actual Size” column of the activity sheet. Now complete the “Difference” column by subtracting the actual value from your estimate for each of the 10 shapes.

FIGURE 1.9
Shapes for Activity 1.2.



4. What would cause a difference to be negative? What would cause a difference to be positive?
 5. Would the sum of the differences tell you if the estimates and actual values were in close agreement? Does a sum of 0 for the differences indicate that all the estimates were equal to the actual value? Explain.
 6. Compare your estimates with those of another person in the class by comparing the sum of the absolute values of the differences between estimates and corresponding actual values. Who was better at estimating shape sizes? How can you tell?
 7. Use the last column of the activity sheet to record the squared differences (for example, if the difference
- for shape 1 was -3 , the squared difference would be $(-3)^2 = 9$. Explain why the sum of the squared differences can also be used to assess how accurate your shape estimates were.
8. For this step, work with three or four other students from your class. For each of the 10 shapes, form a new size estimate by computing the average of the size estimates for that shape made by the individuals in your group. Is this new set of estimates more accurate than your own individual estimates were? How can you tell?
 9. Does your answer from Step 8 surprise you? Explain why or why not.

ACTIVITY 1.3 A MEANINGFUL PARAGRAPH

Write a meaningful paragraph that includes the following six terms: **sample**, **population**, **descriptive statistics**, **bar chart**, **numerical variable**, and **dotplot**.

A “meaningful paragraph” is a coherent piece of writing in an appropriate context that uses all of the listed words. The paragraph should show that you understand

the meanings of the terms and their relationships to one another. A sequence of sentences that just defines the terms is *not* a meaningful paragraph. When choosing a context, think carefully about the terms you need to use. Choosing a good context will make writing a meaningful paragraph easier.

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
Population	The entire collection of individuals or measurements about which information is desired.	Continuous numerical data	Possible values form an entire interval along the number line.
Sample	A part of the population selected for study.	Univariate, bivariate, and multivariate data	Each observation consists of one (univariate), two (bivariate), or two or more (multivariate) responses or values.
Descriptive statistics	Numerical, graphical, and tabular methods for organizing and summarizing data.	Frequency distribution for categorical data	A table that displays frequencies, and sometimes relative frequencies, for each of the possible values of a categorical variable.
Inferential statistics	Methods for generalizing from a sample to a population.	Bar chart	A graph of a frequency distribution for a categorical data set. Each category is represented by a bar, and the area of the bar is proportional to the corresponding frequency or relative frequency.
Categorical data	Individual observations are categorical responses (nonnumerical).	Dotplot	A graph of numerical data in which each observation is represented by a dot on or above a horizontal measurement scale.
Numerical data	Individual observations are numerical (quantitative) in nature.		
Discrete numerical data	Possible values are isolated points along the number line.		

CHAPTER REVIEW Exercises 1.33 - 1.38

- 1.33** • The report “[Testing the Waters 2009](#)” ([nrdc.org](#)) included information on the water quality at the 82 most popular swimming beaches in California. Thirty-eight of these beaches are in Los Angeles

• Data set available online

County. For each beach, water quality was tested weekly and the data below are the percent of the tests in 2008 that failed to meet water quality standards.

Los Angeles County

32	4	6	4	4	7	4	27	19	23
19	13	11	19	9	11	16	23	19	16
33	12	29	3	11	6	22	18	31	43
17	26	17	20	10	6	14	11		

Other Counties

0	0	0	2	3	7	5	11	5	7
15	8	1	5	0	5	4	1	0	1
1	0	2	7	0	2	2	3	5	3
0	8	8	8	0	0	17	4	3	7
10	40	3							

- a. Construct a dotplot of the percent of tests failing to meet water quality standards for the Los Angeles County beaches. Write a few sentences describing any interesting features of the dotplot.
- b. Construct a dotplot of the percent of tests failing to meet water quality standards for the beaches in other counties. Write a few sentences describing any interesting features of the dotplot.
- c. Based on the two dotplots from Parts (a) and (b), describe how the percent of tests that fail to meet water quality standards for beaches in Los Angeles County differs from those of other counties.
- 1.34** The U.S. Department of Education reported that 14% of adults were classified as being below a basic literacy level, 29% were classified as being at a basic literacy level, 44% were classified as being at an intermediate literacy level, and 13% were classified as being at a proficient level ([2003 National Assessment of Adult Literacy](#)).
- a. Is the variable *literacy level* categorical or numerical?
- b. Construct a bar chart to display the given data on literacy level.
- c. Would it be appropriate to display the given information using a dotplot? Explain why or why not.
- 1.35** The Computer Assisted Assessment Center at the University of Luton published a report titled [“Technical Review of Plagiarism Detection Software.”](#) The authors of this report asked faculty at academic institutions about the extent to which they agreed with the statement “Plagiarism is a significant problem in academic institutions.” The responses are summarized in the accompanying table. Construct a bar chart for these data.

Response	Frequency
Strongly disagree	5
Disagree	48
Not sure	90
Agree	140
Strongly agree	39

- 1.36** The article [“Just How Safe Is That Jet?” \(USA TODAY, March 13, 2000\)](#) gave the following relative frequency distribution that summarizes data on the type of violation for fines imposed on airlines by the Federal Aviation Administration:

Type of Violation	Relative Frequency
Security	0.43
Maintenance	0.39
Flight operations	0.06
Hazardous materials	0.03
Other	0.09

- a. Use this information to construct a bar chart for type of violation.
- b. Write a sentence or two commenting on the relative occurrence of the various types of violation.
- 1.37** Each year, [U.S. News and World Report](#) publishes a ranking of U.S. business schools. The following data give the acceptance rates (percentage of applicants admitted) for the best 25 programs in a recent survey:

16.3 12.0 25.1 20.3 31.9 20.7 30.1 19.5 36.2
46.9 25.8 36.7 33.8 24.2 21.5 35.1 37.6 23.9
17.0 38.4 31.2 43.8 28.9 31.4 48.9

- a. Construct a dotplot.
- b. Comment on the interesting features of the plot.

- 1.38** Many adolescent boys aspire to be professional athletes. The paper [“Why Adolescent Boys Dream of Becoming Professional Athletes” \(Psychological Reports \[1999\]:1075–1085\)](#) examined some of the reasons. Each boy in a sample of teenage boys was asked the following question: “Previous studies have shown that more teenage boys say that they are considering becoming professional athletes than any other occupation. In your opinion, why do these boys want to become professional athletes?” The resulting data are shown in the following table:

Response	Frequency
Fame and celebrity	94
Money	56
Attract women	29
Like sports	27
Easy life	24
Don’t need an education	19
Other	19

Construct a bar chart to display these data.

TECHNOLOGY NOTES

Bar Charts

JMP

- Enter the raw data into a column (**Note:** To open a new data table, click **File** then select **New** then **Data Table**)
- Click **Graph** and select **Chart**
- Click and drag the column name containing your stored data from the box under **Select Columns** to the box next to **Categories, X, Levels**
- Click **OK**

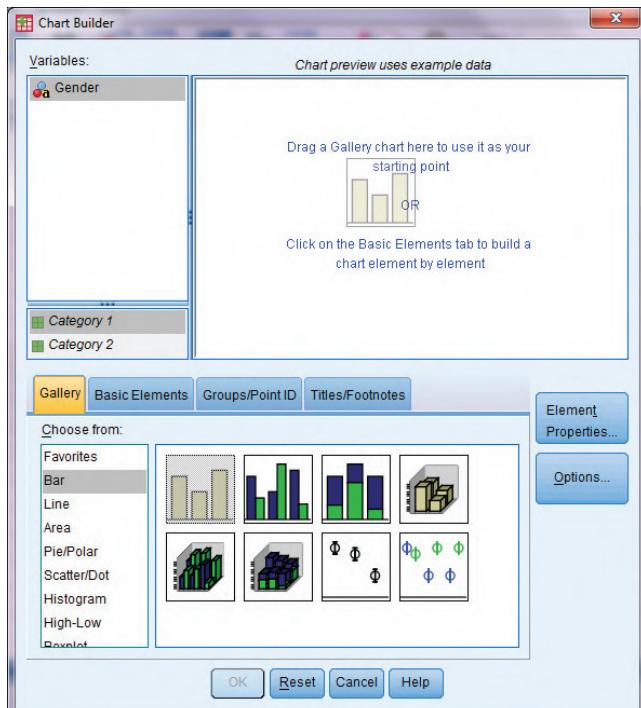
Minitab

- Enter the raw data into C1
- Select **Graph** and choose **Bar Chart...**
- Highlight Simple
- Click **OK**
- Double click C1 to add it to the **Categorical Variables** box
- Click **OK**

Note: You may add or format titles, axis titles, legends, etc., by clicking on the **Labels...** button prior to performing step 6 above.

SPSS

- Enter the raw data into a column
- Select **Graph** and choose **Chart Builder...**
- Under **Choose from** highlight **Bar**
- Click and drag the first bar chart in the right box (Simple Bar) to the Chart preview area

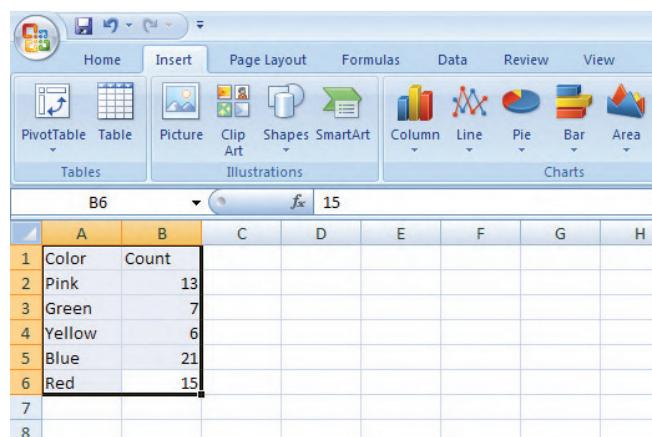


- Click and drag the variable from the **Variables** box to the **X-Axis?** box in the chart preview area
- Click **OK**

Note: By default, SPSS displays the count of each category in a bar chart. To display the percentage for each category, follow the above instructions to Step 5. Then, in the Element Properties dialog box, select Percentage from the drop-down box under Statistic. Click **Apply** then click **OK** on the Chart Builder dialog box.

Excel 2007

- Enter the category names into column A (you may input the title for the variable in cell A1)
- Enter the count or percent for each category into column B (you may put the title "Count" or "Percentage" in cell B1)
- Select all data (including column titles if used)
- Click on the **Insert** Ribbon



- Choose **Column** and select the first chart under 2-D Column (Clustered Column)
- The chart will appear on the same worksheet as your data

Note: You may add or format titles, axis titles, legends, and so on, by right-clicking on the appropriate piece of the chart.

Note: Using the Bar Option on the Insert Ribbon will produce a horizontal bar chart.

TI-83/84

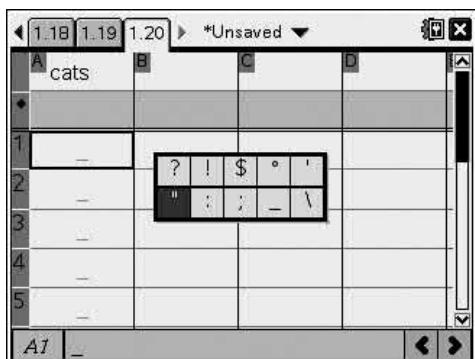
The TI-83/84 does not have the functionality to produce bar charts.

TI-Nspire

- Enter the category names into a list (to access data lists select the spreadsheet option and press **enter**)

Note: In order to correctly enter category names, you will input them as text. Begin by pressing **?!** and then select " and press **enter**. Now type the category name and press **enter**.

Note: Be sure to title the list by selecting the top row of the column and typing a title.



2. Enter the counts for each category into a second list (be sure to title this list as well!)
3. Press the **menu** key and navigate to **3:Data** and then **5:Frequency Plot** and select this option
4. For **Data List** select the list containing your category names from the drop-down menu
5. For **Frequency List** select the list containing the counts for each category from the drop-down menu
6. Press **OK**

Minitab

1. Input the raw data into C1
2. Select **Graph** and choose **Dotplot**
3. Highlight Simple under One Y
4. Click **OK**
5. Double click C1 to add it to the **Graph Variables** box
6. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels...** button prior to performing Step 6 above.

SPSS

1. Input the raw data into a column
2. Select **Graph** and choose **Chart Builder...**
3. Under **Choose from** highlight Scatter/Dot
4. Click and drag the second option in the second row (Simple Dot Plot) to the Chart preview area
5. Click and drag the variable name from the **Variables** box into the **X-Axis?** box
6. Click **OK**

Excel 2007

Excel 2007 does not have the functionality to create dotplots.

TI-83/84

The TI-83/84 does not have the functionality to create dotplots.

TI-Nspire

1. Enter the data into a data list (to access data lists select the spreadsheet option and press **enter**)

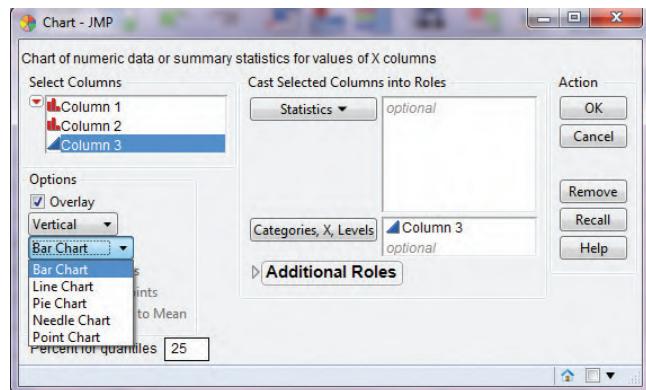
Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key then select **3:Data** then select **6:QuickGraph** and press **enter** (a dotplot will appear)

Dotplots

JMP

1. Input the raw data into a column
2. Click **Graph** and select **Chart**
3. Click and drag the column name containing the data from the box under **Select columns** to the box next to **Categories, X, Levels**
4. Select **Point Chart** from the second drop-down box in the **Options** section



5. Click **OK**

2

Collecting Data Sensibly



Kwame Zikomo/Purestock/SuperStock

Data and conclusions from data are everywhere—in newspapers, magazines, online resources, and professional publications. But should you believe what you read? For example, should you supplement your diet with black currant oil to stop hair loss? Will playing solitaire for 20 minutes each day help you feel less tired? If you eat proteins before carbohydrates when you eat a meal, will it lower your blood sugar? Should you donate blood twice a year to lower your risk of heart disease? These are just four recommendations out of many that appear in one issue of *Woman's World* (April 4, 2016), a magazine with more than 1.3 million readers. In fact, if you followed all of the recommendations in that issue, you would also be loading up on prickly pear oil, hot chocolate, ginger tea, bread, bananas, sweet potatoes, bell peppers, tomatoes, and onions! The magazine suggests that these claims are based on research studies, but how reliable are the studies? Are the conclusions drawn reasonable, and do they apply to you?

A primary goal of statistical studies is to collect data that can be used to make informed decisions. It should come as no surprise that the ability to make good decisions depends on the quality of the information available.

Both the type of analysis that is appropriate and the conclusions that can be drawn depend on how the data are collected. In this chapter, we first consider two types of statistical studies and then focus on two widely used methods of data collection: sampling and experimentation.

LEARNING OBJECTIVES

Students will understand:

- That the types of conclusions that can be drawn from data depend on the way the data were collected.
- That bias may be present when data are collected from a sample.
- Why random selection is an important component of a sampling plan.
- Why random assignment is important when collecting data in an experiment.
- The purposes of a control group and blinding in an experiment.

Students will be able to:

- Distinguish between an observational study and an experiment.
- Distinguish between selection bias, measurement or response bias, and nonresponse bias.
- Select a simple random sample from a given population.
- Distinguish between simple random sampling, stratified random sampling, cluster sampling, systematic sampling, and convenience sampling.

- Describe a procedure for randomly assigning subjects to treatments in an experiment.
- Design a completely randomized experiment.
- Design a randomized block experiment.

SECTION 2.1 Statistical Studies: Observation and Experimentation

On September 25, 2009, results from a study of the relationship between spanking and IQ were reported by a number of different news media. Some of the headlines that appeared that day were:

- “Spanking lowers a child’s IQ” (*Los Angeles Times*)
- “Do you spank? Studies indicate it could lower your kid’s IQ” (*SciGuy, Houston Chronicle*)
- “Spanking can lower IQ” (*NBC4i, Columbus, Ohio*)
- “Smacking hits kids’ IQ” (*newscientist.com*)

In the study that these headlines refer to, the investigators followed 806 kids age 2 to 4 and 704 kids age 5 to 9 for 4 years. IQ was measured at the beginning of the study and again 4 years later. The researchers found that at the end of the study, the average IQ of kids in the younger group who were not spanked was 5 points higher than that of kids who were spanked. For the older group, the average IQ of kids who were not spanked was 2.8 points higher.

These headlines all imply that spanking was the cause of the observed difference in IQ. Is this conclusion reasonable? The answer depends in a critical way on the study design. After considering some important aspects of study design, we’ll return to these headlines and consider whether they are appropriate.

Observation and Experimentation

Data collection is an important step in the data analysis process. When we set out to collect information, it is important to keep in mind the questions we hope to answer on the basis of the resulting data. Sometimes we are interested in answering questions about characteristics of a single existing population or in comparing two or more well-defined populations. To accomplish this, we select a sample from each population under consideration and use the sample information to gain insight into characteristics of those populations.

For example, an ecologist might be interested in estimating the average shell thickness of bald eagle eggs. A social scientist studying a rural community may want to determine whether age and attitude toward abortion are related. These are examples of studies that are *observational* in nature. In these studies, we want to observe characteristics of members of an existing population or of several populations, and then use the resulting information to draw conclusions. In **observational studies**, it is important to obtain samples that are representative of the corresponding populations.

Sometimes the questions we are trying to answer deal with the effect of certain explanatory variables on some response and cannot be answered using data from an observational study. These questions are often of the form, “What happens when . . . ?” or, “What is the effect of . . . ?” For example, an educator may wonder what would happen to test scores if the required lab time for a chemistry course was increased from 3 hours to 6 hours per week. To answer such questions, an experiment is carried out to collect relevant data. The value of some response variable (test score in the chemistry example) is recorded under different experimental conditions (3-hour lab and 6-hour lab). In an **experiment**, one or more explanatory variables, also sometimes called factors, define the experimental conditions.

The type of conclusion that can be drawn from a statistical study depends on how the study was conducted. Both observational studies and experiments can be used to compare groups, but in an experiment the researcher controls who is in which group, whereas this is not the case in an observational study. This seemingly small difference is critical when it comes to drawing conclusions based on data from the study.

DEFINITIONS

Observational study: A study in which characteristics of a sample selected from one or more existing populations are observed. The goal of an observational study is usually to draw conclusions about the corresponding population or about differences between two or more populations. In a well-designed observational study, the sample is selected in a way that is designed to produce a sample that is representative of the population.

Experiment: A study that investigates how a response variable behaves when one or more explanatory variables, also called factors, are manipulated. The usual goal of an experiment is to determine the effect of the explanatory variables (factors) on the response variable. In a well-designed experiment, the composition of the groups that will be exposed to different experimental conditions is determined by random assignment.

A well-designed experiment can result in data that provide evidence for a cause-and-effect relationship. This is an important difference between an observational study and an experiment. In an observational study, it is not possible to draw clear cause-and-effect conclusions because we cannot rule out the possibility that the observed effect is due to some variable other than the explanatory variable being studied. Such variables are called **confounding variables**.

DEFINITION

Confounding variable: A variable that is related to both how the experimental groups were formed and the response variable of interest.

Consider the role of confounding variables in the following two studies:

- A health-related blog summarized a study of the relationship between exposure to loud noise and heart disease (articles.mercola.com/sites/articles/archive/2015/11/11/loud-noise-exposure-increases-heart-disease-risk.aspx, November 11, 2015, retrieved September 25, 2016). The title of the blog entry was “**Long-Term Exposure to Loud Noise Raises Your Risk of Heart Disease.**” This title suggests a cause-and-effect relationship. But the study referenced in the blog entry (“**Exposure to Loud Noise, Bilateral High-Frequency Hearing Loss and Coronary Heart Disease,**” *Occupational & Environmental Medicine*, OnlineFirst, September 15, 2015, oem.bmjjournals.org/content/early/2015/09/15/oemed-2014-102778, retrieved September 25, 2016) was an observational study that compared a group of people with hearing loss to a group of people without hearing loss. The researchers who conducted the study found that the percentage of people in the hearing loss group who had heart disease was greater than the percentage for the group that did not have hearing loss. Because this was an observational study, it is possible that the difference in percentages might be explained by other factors. For example, the two groups might have differed with respect to age, exercise habits, or general health, all of which might be alternative explanations for the observed difference. These potential confounding variables make it difficult to justify a cause-and-effect conclusion.
- Studies have shown that people over age 65 who get a flu shot are less likely to die from a flu-related illness during the following year than those who do not get a flu shot. However, recent research has shown that people over age 65 who get flu shots are also less likely to die from *any* cause during the following year than those who don’t get flu shots (*International Journal of Epidemiology*, December 21, 2005). This has led to the speculation that those over age 65 who get flu shots are healthier as a group than those who do not get flu shots. If this is the case, observational studies that compare two groups—those who get flu shots and those who do not—may

overestimate the effectiveness of the flu vaccine because general health differs in the two groups. General health is a possible confounding variable in such studies.

Each of the studies described above illustrates why potential confounding variables make it unreasonable to draw cause-and-effect conclusions from observational studies.

Let's return to the study on spanking and IQ described at the beginning of this section. Is this study an observational study or an experiment? Two groups were compared (children who were spanked and children who were not spanked), but the researchers did not randomly assign children to the spanking or no-spanking groups. The study is observational, and so cause-and-effect conclusions such as "spanking lowers IQ" are not justified based on the observed data. What we can say is that there is evidence that, as a group, children who are spanked tend to have a lower IQ than children who are not spanked. What we cannot say is that spanking is the *cause* of the lower average IQ. It is possible that other variables—such as home or school environment, socio-economic status, or parents' education—are related to both IQ and whether or not a child was spanked. These are examples of possible confounding variables.

Fortunately, not everyone made the same mistake as the writers of the headlines given earlier in this section. Some examples of headlines that got it right are:

["Lower IQ's measured in spanked children" \(world-science.net\)](#)

["Children who get spanked have lower IQs" \(livescience.com\)](#)

["Research suggests an association between spanking and lower IQ in children" \(CBSnews.com\)](#)

Drawing Conclusions from Statistical Studies

In this section, two different types of conclusions have been described. One type involves generalizing from what we have seen in a sample to some larger population, and the other involves reaching a cause-and-effect conclusion about the effect of an explanatory variable on a response. When is it reasonable to draw such conclusions? The answer depends on the way that the data were collected. Table 2.1 summarizes the types of conclusions that can be made with different study designs.

As can be seen from Table 2.1, it is important to think carefully about the objectives of a statistical study before planning how the data will be collected. Both observational studies and experiments must be carefully designed if the resulting data are to be useful. The common sampling procedures used in observational studies are considered in Section 2.2. In Sections 2.3 and 2.4, we consider experimentation and explore what constitutes good practice in the design of simple experiments.

Table 2.1 Drawing Conclusions from Statistical Studies

Study Description	Reasonable to Generalize Conclusions about Group Characteristics to the Population?	Reasonable to Draw Cause-and-Effect Conclusion?
Observational study with sample selected at random from population of interest	Yes	No
Observational study based on convenience or voluntary response sample (poorly designed sampling plan)	No	No
Experiment with groups formed by random assignment of individuals or objects to experimental conditions		
• Individuals or objects used in study are volunteers or not randomly selected from some population of interest	No	Yes
• Individuals or objects used in study are randomly selected from some population of interest	Yes	Yes
Experiment with groups not formed by random assignment to experimental conditions (poorly designed experiment)	No	No

EXERCISES 2.1 - 2.12

- 2.1** The article “[How Dangerous Is a Day in the Hospital?](#)” (*Medical Care* [2011]: 1068–1075) describes a study to determine if the risk of an infection is related to the length of a hospital stay. The researchers looked at a large number of hospitalized patients and compared the proportion who got an infection for two groups of patients—those who were hospitalized overnight and those who were hospitalized for more than one night. Indicate whether the study is an observational study or an experiment. Give a brief explanation for your choice.
- 2.2** The authors of the paper “[Fudging the Numbers: Distributing Chocolate Influences Student Evaluations of an Undergraduate Course](#)” (*Teaching in Psychology* [2007]: 245–247) carried out a study to see if events unrelated to an undergraduate course could affect student evaluations. Students enrolled in statistics courses taught by the same instructor participated in the study. All students attended the same lectures and one of six discussion sections that met once a week. At the end of the course, the researchers chose three of the discussion sections to be the “chocolate group.” Students in these three sections were offered chocolate prior to having them fill out course evaluations. Students in the other three sections were not offered chocolate.
- The researchers concluded that “Overall, students offered chocolate gave more positive evaluations than students not offered chocolate.” Indicate whether the study is an observational study or an experiment. Give a brief explanation for your choice.
- 2.3** The article “[Why We Fall for This](#)” (*AARP Magazine*, May/June 2011) described a study in which a business professor divided his class into two groups. He showed students a mug and then asked students in one of the groups how much they would pay for the mug. Students in the other group were asked how much they would sell the mug for if it belonged to them. Surprisingly, the average value assigned to the mug was quite different for the two groups! Indicate whether the study is an observational study or an experiment. Give a brief explanation for your choice.
- 2.4** The article “[Adolescents Living the 24/7 Lifestyle: Effects of Caffeine and Technology on Sleep Duration and Daytime Functioning](#)” (*Pediatrics* [2009]: e1005–e1010) describes a study in which researchers investigated whether there is a relationship between amount of sleep and caffeine consumption. They found that teenagers who usually get less than 8 hours of sleep on school nights were more likely to report falling asleep during school and to consume more caffeine on average than teenagers who usually get 8 to 10 hours of sleep on school nights.
- a. Is the study described an observational study or an experiment?
- b. Is it reasonable to conclude that getting less than 8 hours of sleep on school nights causes teenagers to fall asleep during school and to consume more caffeine, on average? Explain. (Hint: Look at Table 2.1.)
- 2.5** The article “[Acupuncture for Bad Backs: Even Sham Therapy Works](#)” (*Time*, May 12, 2009) summarized a study conducted by researchers in Seattle. In this study, 638 adults with back pain were randomly assigned to one of four groups. People in group 1 received the usual care for back pain. People in group 2 received acupuncture at a set of points tailored specifically for each individual. People in group 3 received acupuncture at a standard set of points typically used in the treatment of back pain. Those in group 4 received fake acupuncture—they were poked with a toothpick at the same set of points chosen for the people in group 3!
- Two conclusions from the study were:
- (1) patients receiving real or fake acupuncture experienced a greater reduction in pain than those receiving usual care; and (2) there was no significant difference in pain reduction for those who received acupuncture and those who received fake acupuncture toothpick pokes.
- a. Is this study an observational study or an experiment? Explain.
- b. Is it reasonable to conclude that receiving either real or fake acupuncture was the cause of the observed reduction in pain in those groups compared to the usual care group? What aspect of this study supports your answer? (Hint: Look at Table 2.1.)
- 2.6** The article “[Display of Health Risk Behaviors on MySpace by Adolescents](#)” (*Archives of Pediatrics and Adolescent Medicine* [2009]: 27–34) described a study in which researchers looked at a random sample of 500 publicly accessible MySpace web profiles posted by 18-year-olds. The content of each profile was analyzed. One of the conclusions reported was that displaying sport or hobby involvement was associated with decreased references to risky behavior (sexual references or references to substance abuse or violence).

- a. Is the study described an observational study or an experiment?
- b. Is it reasonable to generalize the stated conclusion to all 18-year-olds with a publicly accessible MySpace web profile? What aspect of the study supports your answer?
- c. Not all MySpace users have a publicly accessible profile. Is it reasonable to generalize the stated conclusion to all 18-year-old MySpace users? Explain.
- d. Is it reasonable to generalize the stated conclusion to all MySpace users with a publicly accessible profile? Explain.
- 2.7** The article “[Popping Cork Sound Makes Wine Taste Better](#)” ([decanter.com/wine-news/popping-cork-sound-makes-wine-taste-better-experiment-377364](#), retrieved February 11, 2018) describes a study in which 140 people were assigned to one of two groups. Those in one group heard the noise made by a bottle of wine being uncorked before tasting a glass of wine. Those in the other group heard the noise made by a person releasing a screw cap before tasting a glass of wine. Both groups tasted the same wine. It was reported that the average rating was higher for the group that heard the cork popping than the average rating for the group that heard the screw cap being released.
- a. Is the study described an observational study or an experiment?
- b. Can a case be made for the researcher’s conclusion that hearing a cork pop rather than a screw cap being released was the cause for the higher rating? Explain.
- 2.8** “[Fruit Juice May Be Fueling Pudgy Preschoolers, Study Says](#)” is the title of an article that appeared in the [San Luis Obispo Tribune](#) (February 27, 2005). This article describes a study that found that for 3- and 4-year-olds, drinking something sweet once or twice a day doubled the risk of being seriously overweight one year later. The authors of the study state
- Total energy may be a confounder if consumption of sweet drinks is a marker for other dietary factors associated with overweight ([Pediatrics](#), November 2005).
- Give an example of a dietary factor that might be one of the potentially confounding variables the study authors are worried about.
- 2.9** The article “[Americans are ‘Getting the Wrong Idea’ on Alcohol and Health](#)” (Associated Press,

[April 19, 2005](#)) reported that observational studies in recent years that have concluded that moderate drinking is associated with a reduction in the risk of heart disease may be misleading. The article refers to a study conducted by the Centers for Disease Control and Prevention that showed that moderate drinkers, as a group, tended to be better educated, wealthier, and more active than nondrinkers.

Explain why the existence of these potentially confounding variables prevents drawing the conclusion that moderate drinking is the cause of reduced risk of heart disease.

- 2.10** Based on a survey conducted on the eDiets.com web site, investigators concluded that women who regularly watched *Oprah* were only one-seventh as likely to crave fattening foods as those who watched other daytime talk shows ([San Luis Obispo Tribune](#), October 14, 2000).
- a. Is it reasonable to conclude that watching *Oprah* causes a decrease in cravings for fattening foods? Explain.
- b. Is it reasonable to generalize the results of this survey to all women in the United States? To all women who watch daytime talk shows? Explain why or why not.
- 2.11** A survey of adult Americans who are Internet users carried out in 2016 found that 79% were Facebook users ([“Social Media Update 2016,” Pew Research Center](#), November 11, 2016).
- a. What condition on how the data were collected would make the generalization from the sample to the population of all adult American Internet users reasonable?
- b. Would it be reasonable to generalize from the sample and say that 79% of all adult Americans use Facebook? Explain.
- 2.12** Does sitting for long periods of time hurt your heart? The article “[Why Sitting May Be Bad for Your Heart](#)” ([The New York Times](#), December 20, 2017) describes a study of 1700 people who were participants in the Dallas Heart Study. The study found that the people who sat for long periods of time tended to have higher levels of troponin in their blood. Troponin is a protein that is released when the heart muscle has been damaged.
- The article states that “Of course, this was an observational study and can show only that sitting is linked to high troponin, not that it causes troponins to rise.” Do you agree with this statement? Explain.

SECTION 2.2 Sampling

Many studies are conducted to learn about a population. In this case, it is important that the sample selected for the study is representative of the population. To be reasonably sure of this, we must think carefully about how the sample is selected.

It is sometimes tempting to take the easy way out and gather data in a haphazard way. But if a sample is chosen on the basis of convenience alone, it is not possible to interpret the resulting data with confidence. For example, it might be easy to use the students in your statistics class as a sample of students at your university. However, not all majors include a statistics course in their curriculum, and most students take statistics in their freshman or sophomore year. When we attempt to generalize from this convenience sample, it is not clear how these factors (and others that we might not be aware of) affect any conclusions based on information from the sample.

There is no way to tell just by looking at a sample whether it is representative of the population from which it was selected. Our only assurance comes from the method used to select the sample.

There are many reasons for selecting a sample rather than obtaining information from an entire population (a **census**). Sometimes the process of measuring a characteristic of interest damages the item that is being measured. This is the case with measuring the lifetime of flashlight batteries (which results in dead batteries) or measuring the sugar content of oranges (which results in oranges that cannot be sold). It would be foolish to study the entire population in situations like these. But the most common reason for selecting a sample is limited resources. Restrictions on available time or money usually make it impossible to collect data from an entire population.

Bias in Sampling

Bias in sampling is the tendency for samples to differ from the corresponding population in some systematic way. Bias can result from the way in which the sample is selected or from the way in which information is obtained once the sample has been chosen. The most common types of bias encountered in sampling situations are selection bias, measurement or response bias, and nonresponse bias.

Selection bias (sometimes also called undercoverage) is introduced when the way the sample is selected systematically excludes some part of the population of interest. For example, a researcher may wish to generalize from the results of a study to the population consisting of all residents of a particular city, but the method of selecting individuals may tend to exclude the homeless or those without telephones.

If those who are excluded from the sampling process differ in some systematic way from those who are included, the sample is virtually guaranteed to be unrepresentative of the population. If this difference between the included and the excluded occurs on a variable that is important to the study, conclusions based on the sample data may not be valid for the population of interest.

Selection bias also occurs if only volunteers or self-selected individuals are used in a study, because those who choose to participate (for example, in a call-in telephone poll) may differ from those who choose not to participate.

Measurement or response bias occurs when the method of observation tends to produce values that systematically differ from the true value in some way. This might happen if an improperly calibrated scale is used to weigh items or if questions on a survey are worded in a way that tends to influence the response.

For example, David Wilson in his blog for *The Huffington Post* (“Fox News Poll: What Does Bad Question Wording Say About Their Polling Data?,” September 21, 2010, huffingtonpost.com/david-c-wilson/fox-news-poll-what-does-b_b_734101.html, retrieved September 25, 2016) writes:

One of the main sources of error in interpreting poll results is poor question wording. In surveys questions are supposed to be accurate measures of one's attitudes, opinions, beliefs, and behaviors; just like the scales in bathrooms accurately measure one's weight. They should at least be well written and easily understood by respondents, and they should not be biased toward a particular viewpoint.

He goes on to give several examples of "bad" questions, including the following question, which was part of a Fox News poll conducted in September 2010:

Do you agree or disagree with the following statement: the federal government has gotten totally out of control and threatens our basic liberties unless we clean house and commit to drastic change?

Wilson points out that this question makes a number of assertions, such as the government is out of control and basic liberties are threatened. If someone responded "agree" to this question, it would be unclear exactly which assertion they were in agreement with. In addition, words like *control* and *threaten* may lead people to respond in a particular way.

Other things that might contribute to response bias are the appearance or behavior of the person asking the question, the group or organization conducting the study, and the tendency for people not to be completely honest when asked about illegal behavior or unpopular beliefs.

Although the terms *measurement bias* and *response bias* are often used interchangeably, the term *measurement bias* is usually used to describe systematic deviation from the true value as a result of a faulty measurement instrument. Response bias is typically used to describe systematic deviations from the true value when people provide answers to survey questions.

Nonresponse bias occurs when responses are not obtained from all individuals selected for inclusion in the sample. As with selection bias, nonresponse bias can distort results if those who respond differ in important ways from those who do not respond. Although some level of nonresponse is unavoidable in most surveys, the biasing effect on the resulting sample is lowest when the response rate is high. To minimize nonresponse bias, it is important that a serious effort be made to follow up with individuals who do not respond to an initial request for information.

The nonresponse rate for surveys or opinion polls varies dramatically, depending on how the data are collected. Surveys are commonly conducted by mail, by phone, and by personal interview. Mail surveys are inexpensive but often have high nonresponse rates. Telephone surveys can also be inexpensive and can be implemented quickly, but they work well only for short surveys and they can also have high nonresponse rates. Personal interviews are generally expensive but tend to have better response rates. Some of the many challenges of conducting surveys are discussed in Section 2.6 (available online).

Types of Bias

Selection Bias

Tendency for samples to differ from the population as a result of systematic exclusion of some part of the population.

Measurement or Response Bias

Tendency for samples to differ from the population because the method of observation tends to produce values that differ from the true values.

Nonresponse Bias

Tendency for samples to differ from the population because data are not obtained from all individuals selected for inclusion in the sample.

It is important to understand that bias is introduced by the way a sample is selected or by the way the data are collected from the sample. Increasing the size of the sample, although possibly desirable for other reasons, does not reduce bias if the method of selecting the sample is flawed or if the nonresponse rate remains high.

Potential sources of bias are illustrated in the following examples.

Example 2.1 Are Cell Phone Users Different?

Understand the context ➤

Many surveys are conducted by telephone and participants are often selected from phone directories that include only landline telephones. For many years, it was thought that this was not a serious problem because most cell phone users also had a landline phone and so they still had a chance of being included in the survey. But the number of people with only cell phones is growing. This trend is a concern for survey organizations.

The article “[Omitting Cell Phone Users May Affect Polls](#)” (*Associated Press, September 25, 2008*) described a study that examined whether people who only have a cell phone are different from those who have landline phones. One finding from the study was that for people under the age of 30 with only a cell phone, 28% were Republicans compared to 36% of landline users. This suggests that researchers who use telephone surveys need to worry about how selection bias might influence the ability to generalize the results of a survey if only landlines are used.

Example 2.2 Think before You Order That Burger!

Understand the context ➤

The article “[What People Buy from Fast-Food Restaurants: Caloric Content and Menu Item Selection](#)” (*Obesity [2009]: 1369–1374*) reported that the average number of calories consumed at lunch in New York City fast-food restaurants was 827. The researchers selected 267 fast-food locations at random. The paper states that at each of these locations “adult customers were approached as they entered the restaurant and asked to provide their food receipt when exiting and to complete a brief survey.”

Approaching customers as they entered the restaurant and before they ordered may have influenced what they purchased. This introduces the potential for response bias. In addition, some people chose not to participate when approached. If those who chose not to participate differed from those who did participate, the researchers also need to be concerned about nonresponse bias. Both of these potential sources of bias limit the researchers’ ability to generalize conclusions based on data from this study.

Random Sampling

Many methods introduced in this text are based on random selection. The most straightforward sampling method is called simple random sampling. A **simple random sample** is a sample chosen using a method that ensures that each different possible sample of the desired size has an equal chance of being the one chosen.

For example, suppose that we want a simple random sample of 10 employees chosen from all those who work at a large design firm. For the sample to be a simple random sample, the method used to select the sample must ensure that each of the many different subsets of 10 employees are equally likely to be selected. A sample taken from only full-time employees would not be a simple random sample of *all* employees, because someone who works part-time has no chance of being selected. Although a simple random sample may, by chance, include only full-time employees, it must be selected in such a way that each possible sample, and therefore *every* employee, has the same chance of being selected.

It is the selection process, not the final sample, which determines whether a sample is a simple random sample.

The letter n is used to denote sample size. It is the number of individuals or objects in the sample. For the design firm scenario just described, $n = 10$ because 10 employees were to be selected.

DEFINITION

Simple random sample of size n : A sample that is selected from a population in a way that ensures that every different possible sample of size n has the same chance of being selected.

The definition of a simple random sample implies that every individual member of the population has an equal chance of being selected. *However, the fact that every individual has an equal chance of selection, by itself, is not enough to guarantee that the sample is a simple random sample.*

For example, suppose that a class is made up of 100 students, 60 of whom are female. A researcher decides to select 6 of the female students by writing all 60 names on slips of paper, mixing the slips, and then picking 6. She then selects 4 male students from the class using a similar procedure. Even though every student in the class has an equal chance of being included in the sample (6 of 60 females are selected and 4 of 40 males are chosen), the resulting sample is *not* a simple random sample because not all different possible samples of 10 students from the class have the same chance of selection. Many possible samples of 10 students—for example, a sample of 7 females and 3 males or a sample of all females—have no chance of being selected. The sample selection method described here is not necessarily a bad choice (in fact, it is an example of stratified sampling, to be discussed in more detail shortly). But it does not produce a simple random sample. When this is the case, it is sometimes necessary to use different methods when generalizing results from the sample to the population. For this reason, the sampling method is an important consideration when a method is chosen for analyzing data.

Selecting a Simple Random Sample

A number of different methods can be used to select a simple random sample. One way is to put the name or number of each member of the population on different but identical slips of paper. The process of thoroughly mixing the slips and then selecting n slips yields a random sample of size n . This method is easy to understand, but it has obvious drawbacks. The mixing must be adequate, and producing the necessary slips of paper can be extremely tedious, even for relatively small populations.

A commonly used method for selecting a random sample is to first create a list, called a **sampling frame**, of the objects or individuals in the population. Each item on the list is identified by a number. A table of random digits or a random number generator can then be used to select the sample. A random number generator is an algorithm that produces a sequence of numbers that satisfies properties associated with the notion of randomness. Most statistics software packages and many calculators include a random number generator. A small table of random digits can be found in Appendix A, Table 1.

For example, suppose a list containing the names of the 427 customers who purchased a new car during 2018 at a large dealership is available. The owner of the dealership wants to interview a sample of these customers to learn about customer satisfaction. She plans to select a simple random sample of 20 customers. Because it would be tedious to write all 427 names on slips of paper, random numbers can be used to select the sample. To do this, we can use three-digit numbers, starting with 001 and ending with 427, to represent the individuals on the list.

The random digits from rows 6 and 7 of Appendix A, Table 1 are shown here:

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

We can use blocks of three digits from this list (underlined in the lists above) to identify the individuals who should be included in the sample. The first block of three digits is 093, so the 93rd person on the list will be included in the sample. The next five blocks of three digits (876, 799, 562, 565, and 842) do not correspond to anyone on the list, so we ignore them. The next block that corresponds to a person on the list is 410, so that person is included in the sample. This process would continue until 20 people have been selected for the sample. We would ignore any three-digit repeats since any particular person should only be selected once for the sample.

Another way to select the sample would be to use computer software or a graphing calculator to generate 20 random numbers. For example, Minitab produced the following numbers when 20 random numbers between 1 and 427 were requested.

289	67	29	26	205	214	422	31	233	98
10	203	346	186	232	410	43	293	25	371

These numbers could be used to determine which 20 customers to include in the sample.

When selecting a random sample, researchers can choose to do the sampling with or without replacement. **Sampling with replacement** means that after each item is selected for the sample, the item is “replaced” back into the population and may therefore be selected again at a later stage. In practice, sampling with replacement is rarely used. Instead, the more common method is to not allow the same item to be included in the sample more than once. After being included in the sample, an individual or object would not be considered for further selection. This is called **sampling without replacement**.

DEFINITIONS

Sampling without replacement: Once an individual from the population is selected for inclusion in the sample, it may not be selected again in the sampling process. A sample selected without replacement includes n different individuals from the population.

Sampling with replacement: After an individual from the population is selected for inclusion in the sample and the corresponding data are recorded, the individual is placed back in the population and can be selected again in the sampling process. A sample selected with replacement might include any particular individual from the population more than once.

Although these two forms of sampling are different, when the sample size n is small relative to the population size, as is often the case, there is little practical difference between them. In practice, the two methods can be viewed as equivalent if the sample size is less than 10% of the population size.

Example 2.3 Selecting a Random Sample of Glass Soda Bottles

Understand the context ➤

Breaking strength is an important characteristic of glass soda bottles. Suppose that we want to measure the breaking strength of each bottle in a random sample of size $n = 3$ selected from four crates containing a total of 100 bottles (the population). Each crate contains five rows of five bottles each. We can identify each bottle with a number from 1 to 100 by numbering across the rows in each crate, starting with the top row of crate 1, as pictured:



BananaStock/Alamy Stock Photo

Crate 1				
1	2	3	4	5
6	...			
				...

Crate 2				
26	27	28	...	
			...	

Crate 4				
76	77	...		
		...		100

Formulate a plan ➤

Using a random number generator from a calculator or statistical software package, we could generate three random numbers between 1 and 100 to determine which bottles would be included in the sample. This might result in bottles 15 (row 3 column 5 of crate 1), 89 (row 3 column 4 of crate 4), and 60 (row 2 column 5 of crate 3) being selected.

The goal of random sampling is to produce a sample that is likely to be representative of the population. Although random sampling does not *guarantee* that the sample will be representative, it does allow us to assess the risk of an unrepresentative sample. It is the ability to quantify this risk that will enable us to generalize with confidence from a random sample to the corresponding population.

An Important Note Concerning Sample Size

It is a common misconception that if the size of a sample is relatively small compared to the population size, the sample cannot possibly accurately reflect the population. Critics of polls often make statements such as, “There are 14.6 million registered voters in California. How can a sample of 1000 registered voters possibly reflect public opinion when only about 1 in every 14,000 people is included in the sample?” These critics do not understand the power of random selection!

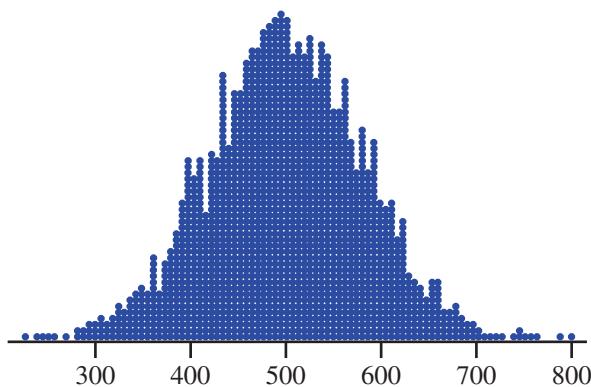
Consider a population consisting of 5000 applicants to a state university, and suppose that we are interested in math SAT scores for this population. A dotplot of the values in this population is shown in Figure 2.1(a). Figure 2.1(b) shows dotplots of the math SAT scores for individuals in five different random samples from the population, ranging in sample size from $n = 50$ to $n = 1000$.

Notice that each of the samples tend to reflect the distribution of scores in the population. If we were interested in using the sample to estimate the population average or to say something about the variability in math SAT scores, even the smallest of the samples pictured ($n = 50$) would provide reliable information.

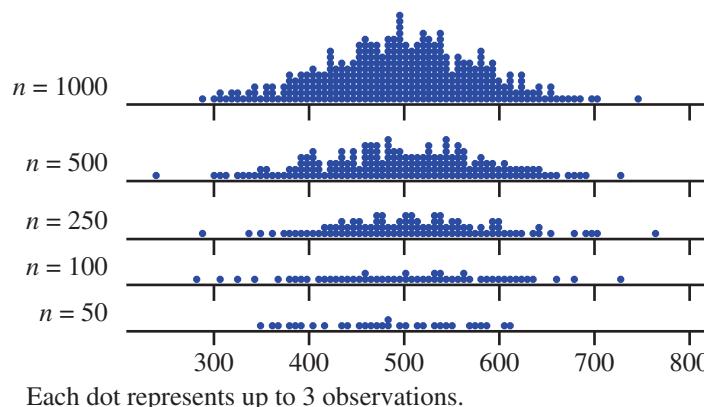
Although it is possible to obtain a simple random sample that does not do a reasonable job of representing the population, this is likely to happen only when the sample size is very small. Unless the population itself is small, this risk does not depend on what fraction of the population is sampled. The random selection process allows us to be confident that the resulting sample adequately reflects the population, even when the sample consists of only a small fraction of the population.

FIGURE 2.1

- (a) Dotplot of math SAT scores for the entire population.
- (b) Dotplots of math SAT scores for random samples of sizes 50, 100, 250, 500, and 1000.



(a)



(b)

Other Sampling Methods

Simple random sampling provides researchers with a sampling method that is objective and free of selection bias. In some settings, however, alternative sampling methods may be less costly, easier to implement, and sometimes even more accurate.

Stratified Random Sampling

When the entire population can be divided into a set of nonoverlapping subgroups, a method known as **stratified sampling** is often easier to implement and more cost-effective than simple random sampling. In stratified random sampling, separate simple random samples are independently selected from each subgroup.

For example, to estimate the average cost of malpractice insurance, a researcher might find it convenient to view the population of all doctors practicing in a particular city as being made up of four groups: (1) surgeons, (2) internists and family practitioners, (3) obstetricians, and (4) a group that includes all other areas of specialization. Rather than taking a simple random sample from the population of all doctors, the researcher could take four separate simple random samples—one from the group of surgeons, another from the internists and family practitioners, and so on. These four samples would provide information about the four subgroups as well as information about the overall population of doctors.

When the population is divided in this way, the subgroups are called **strata** and each individual subgroup is called a stratum (the singular of strata). Stratified sampling selects a separate simple random sample from each stratum. Stratified sampling can be used instead of simple random sampling if it is important to obtain information about characteristics of the individual strata as well as of the entire population, although a stratified sample is not required to do this—subgroup estimates can also be obtained by using an appropriate subset of data from a simple random sample.

The real advantage of stratified sampling is that it often allows us to make more accurate inferences about a population than does simple random sampling. In general, it is much easier to produce relatively accurate estimates of characteristics of a homogeneous group than of a heterogeneous group. For example, even with a small sample, it is possible to obtain an accurate estimate of the average grade point average (GPA) of students graduating with high honors from a university. The individual GPAs of these students are all quite similar (a homogeneous group), and even a sample of three or four individuals from this group should be representative. On the other hand, producing a reasonably accurate estimate of the average GPA of *all* seniors at the university, a much more diverse group of GPAs, is a more difficult task. This means that if a varied population can be divided into strata, with each stratum being much more homogeneous than the population with respect to the characteristic of interest, then a stratified random sample can produce more accurate estimates of population characteristics than a simple random sample of the same size.

Cluster Sampling

Sometimes it is easier to select groups of individuals from a population than it is to select individuals themselves. **Cluster sampling** involves dividing the population of interest into nonoverlapping subgroups, called **clusters**. Clusters are then selected at random, and then *all* individuals in the selected clusters are included in the sample.

For example, suppose that a large urban high school has 600 senior students, all of whom are enrolled in a first period homeroom. There are 24 senior homerooms, each with approximately 25 students. If school administrators wanted to select a sample of about 75 seniors to participate in an evaluation of the college and career placement advising available to students, they might find it much easier to select three of the senior homerooms at random and then include all the students in the selected homerooms in the sample. Then a survey could be administered to all students in the selected homerooms at the same time—certainly easier to implement than randomly selecting 75 individual seniors and then administering the survey to these students.

Because whole clusters are selected, the ideal situation for cluster sampling is when each cluster mirrors the characteristics of the population. When this is the case, a small number of clusters results in a sample that is representative of the population. If it is not reasonable

to think that the variability present in the population is reflected in each cluster, as is often the case when the cluster sizes are small, then it becomes important that a large number of clusters are included in the sample.

Be careful not to confuse clustering and stratification. Even though both of these sampling strategies involve dividing the population into subgroups, both the way in which the subgroups are sampled and the best strategy for creating the subgroups are different.

In stratified sampling, we sample from every subgroup, whereas in cluster sampling, we include only selected whole clusters in the sample. Because of this difference, to increase the chance of obtaining a sample that is representative of the population, we want to create homogeneous groups for strata and heterogeneous (reflecting the variability in the population) groups for clusters.

Systematic Sampling

Systematic sampling is a method that can be used when it is possible to view the population of interest as consisting of a list or some other sequential arrangement. A value k is specified (for example, $k = 50$ or $k = 200$). Then one of the first k individuals is selected at random, after which every k th individual in the sequence is included in the sample. A sample selected in this way is called a **1 in k systematic sample**.

For example, a sample of faculty members at a university might be selected from the faculty phone directory. One of the first $k = 20$ faculty members listed could be selected at random, and then every 20th faculty member after that on the list would also be included in the sample. This would result in a 1 in 20 systematic sample.

The value of k for a 1 in k systematic sample is generally chosen to achieve a desired sample size. For example, in the faculty directory scenario just described, if there were 900 faculty members at the university, the 1 in 20 systematic sample described would result in a sample size of 45. If a sample size of 100 was desired, a 1 in 9 systematic sample could be used (because $900/100 = 9$).

As long as there are no repeating patterns in the population sequence, systematic sampling works reasonably well. However, if there are repeating patterns, systematic sampling can result in an unrepresentative sample. For example, suppose that workers at the entry station of a state park have recorded the number of visitors to the park each day for the past 10 years. In a 1 in 70 systematic sample of days from this list, we would pick one of the first 70 days at random and then every 70th day after that. But if the first day selected happened to be a Wednesday, *every* day selected in the entire sample would also be a Wednesday (because there are 7 days a week and 70 is a multiple of 7). It is unlikely that such a sample would be representative of the entire collection of days. The number of visitors is likely to be higher on weekend days, and no Saturdays or Sundays would be included in the sample.

Convenience Sampling: Don't Go There!

It is often tempting to resort to **convenience sampling**—that is, using an easily available or convenient group to form a sample. This is a recipe for disaster! Results from such samples are rarely informative, and it is a mistake to try to generalize from a convenience sample to any larger population.

One common form of convenience sampling is sometimes called **voluntary response sampling**. Such samples rely entirely on individuals who volunteer to be a part of the sample, often by responding to an advertisement, calling a publicized telephone number to register an opinion, or logging on to an Internet site to complete a survey. It is extremely unlikely that individuals participating in such voluntary response surveys are representative of any larger population of interest.

EXERCISES 2.13 - 2.34

- 2.13** A New York psychologist recommends that if you feel the need to check your e-mail in the middle of a movie or if you sleep with your cell phone next to your bed, it might be time to “power off” (*AARP Bulletin, September 2010*). Suppose that you want to learn about the proportion of students at your college who would feel the need to check e-mail during the middle of a movie and that you have access to a list of all students enrolled at your college. Describe how you would use this list to select a simple random sample of 100 students.
- 2.14** As part of a curriculum review, a psychology department would like to select a simple random sample of 20 of last year’s 140 graduates to obtain information on how graduates perceived the value of the curriculum. Describe two different methods that might be used to select the sample.
- 2.15** A petition with 500 signatures is submitted to a university’s student council. The council president would like to determine the proportion of those who signed the petition who are actually registered students at the university. There is not enough time to check all 500 names with the registrar, so the council president decides to select a simple random sample of 30 signatures. Describe how this might be done.
- 2.16** The article “*Bicyclists and Other Cyclists*” (*Annals of Emergency Medicine [2010]: 426*) reported that in 2008, there were 716 bicyclists killed on public roadways in the United States, and that the average age of the cyclists killed was 41 years. These figures were based on an analysis of the records of all traffic-related deaths of bicyclists on U.S. public roadways (this information is kept by the National Highway Traffic Safety Administration).
- a. Does the group of 716 bicycle fatalities represent a census or a sample of the 2008 bicycle fatalities?
- b. If the population of interest is 2008 bicycle traffic fatalities, is the given average age of 41 years a number that describes a sample or a number that describes the population?
- 2.17** The article “*Teenage Physical Activity Reduces Risk of Cognitive Impairment in Later Life*” (*Journal of the American Geriatrics Society [2010]*) describes a study of more than 9000 women from Maryland, Minnesota, Oregon, and Pennsylvania. The women were asked about their physical activity as teenagers and at ages 30 and 50. A press release about this study (wiley.com/WileyCDA/PressRelease/pressReleaseld-77637.html retrieved February 13, 2018) generalized the results of this study to all

American women. In the press release, the researcher who conducted the study is quoted as saying

Our study shows that women who are regularly physically active at any age have lower risk of cognitive impairment than those who are inactive but that being physically active at teenage is most important in preventing cognitive impairment.

Answer the following four questions for this observational study. (Hint: Reviewing Examples 2.1 and 2.2 might be helpful.)

- a. What is the population of interest?
 - b. Was the sample selected in a reasonable way?
 - c. Is the sample likely to be representative of the population of interest?
 - d. Are there any obvious sources of bias?
- 2.18** For each of the situations described, state whether the sampling procedure is simple random sampling, stratified random sampling, cluster sampling, systematic sampling, or convenience sampling.
- a. All first-year students at a university are enrolled in one of 30 sections of a seminar course. To select a sample of freshmen at this university, a researcher selects four sections of the seminar course at random from the 30 sections and all students in the four selected sections are included in the sample.
 - b. To obtain a sample of students, faculty, and staff at a university, a researcher randomly selects 50 faculty members from a list of faculty, 100 students from a list of students, and 30 staff members from a list of staff.
 - c. A university researcher obtains a sample of students at his university by using the 85 students enrolled in his Psychology 101 class.
 - d. To obtain a sample of the seniors at a particular high school, a researcher writes the name of each senior on a slip of paper, places the slips in a box and mixes them, and then selects 10 slips. The students whose names are on the selected slips of paper are included in the sample.
 - e. To obtain a sample of those attending a basketball game, a researcher selects the 24th person through the door. Then, every 50th person after that is also included in the sample.

- 2.19** Of the 6500 students enrolled at a community college, 3000 are part time and the other 3500 are full time. The college can provide a list of students that is sorted so that all full-time students are listed first, followed by the part-time students.

- a. Describe a procedure for selecting a stratified random sample that uses full-time and part-time students as the two strata and that includes 10 students from each stratum.
- b. Does every student at this community college have the same chance of being selected for inclusion in this stratified random sample? Explain.
- 2.20** Briefly explain why it is advisable to avoid the use of convenience samples.
- 2.21** A sample of pages from this book is to be selected, and the number of words on each page in the sample will be determined. For the purposes of this exercise, equations are not counted as words and a number is counted as a word only if it is spelled out—that is, *ten* is counted as a word, but *10* is not.
- Describe a sampling procedure that would result in a simple random sample of pages from this book.
 - Describe a sampling procedure that would result in a stratified random sample. Explain why you chose the specific strata used in your sampling plan.
 - Describe a sampling procedure that would result in a systematic sample.
 - Describe a sampling procedure that would result in a cluster sample.
- 2.22** Using the process you gave in Part (a) of Exercise 2.21, select a simple random sample of at least 20 pages, and record the number of words on each of the selected pages. Construct a dotplot of the resulting sample values, and write a sentence or two commenting on what it reveals about the number of words on a page.
- 2.23** Using the process you gave in Part (b) of Exercise 2.21, select a stratified random sample that includes a total of at least 20 selected pages, and record the number of words on each of the selected pages. Construct a dotplot of the resulting sample values, and write a sentence or two commenting on what it reveals about the number of words on a page.
- 2.24** The chairman of a California ballot initiative campaign to add “none of the above” to the list of ballot options in all candidate races was quite critical of a Field poll that showed his measure trailing by 10 percentage points. The poll was based on a random sample of 1000 registered voters in California. He is quoted by the [Associated Press \(January 30, 2000\)](#) as saying, “Field’s sample in that poll equates to one out of 17,505 voters,” and he added that this was so dishonest that Field should get out of the polling business! If you worked on the Field poll, how would you respond to this criticism? (Hint: See discussion of sample size on page 39.)
- 2.25** The authors of the paper [“Digital Inequality: Differences in Young Adults’ Use of the Internet”](#) ([Communication Research \[2008\]: 602–621](#)) were

interested in determining if people with higher levels of education use the Internet in different ways than those who do not have as much formal education. To answer this question, they used data from a national telephone survey. Approximately 1300 households were selected for the survey, and 270 of them completed the interview. What type of bias should the researchers be concerned about and why? (Hint: See the box on page 35 that contains the definitions of different types of bias.)

- 2.26** The 2013 National Study of Substance Use Habits of College Student-Athletes surveyed student athletes at NCAA member colleges and universities. The passage below is from the survey website ([ncaa.org/about/resources/research/about-survey](#), retrieved February 13, 2018).

All NCAA member institutions are asked to participate. The sampling plan achieves an appropriate representation of all NCAA student-athletes while minimizing burden to institutions by asking that all student-athletes on no more than three teams be surveyed on any campus. The teams surveyed are determined by a computer-generated random draw.

The survey is administered to the selected teams in a classroom setting, and no identifying information about the athletes or the college is collected. The web site also states

It is important to note that even with measures to ensure anonymity, self-reported data of this kind can be problematic due to the sensitive nature of the issues. Therefore, absolute levels of use might be underestimated in a study such as this.

- Was this sample a simple random sample, a stratified sample, a cluster sample, a systematic sample, or a convenience sample? Explain.
- Give two reasons why an estimate of the proportion of students who reported using illegal drugs based on data from this survey should not be generalized to all U.S. college students.

- 2.27** The paper [“Deception and Design: The Impact of Communication Technology on Lying Behavior”](#) ([Computer-Human Interaction \[2009\]: 130–136](#)) describes an investigation into whether lying is less common in face-to-face communication than in other forms of communication such as phone conversations or e-mail. Participants in this study were 30 students in an upper-division communications course at Cornell University who received course credit for participation. Participants were asked to record all of their social interactions for a week, making note of any lies told.

Based on data from these records, the authors of the paper concluded that students lie more often in phone conversations than in face-to-face

conversations and more often in face-to-face conversations than in e-mail.

Discuss the limitations of this study, commenting on the way the sample was selected and potential sources of bias.

- 2.28** The authors of the paper “[Playing by the Rules: Parental Mediation of Video Game Play](#)” (*Journal of Family Issues* [2015]: 1–24) used data from a sample of parents to investigate the ways in which parents monitor their children’s use of video games. The sample of parents consisted of 427 people who responded to a survey conducted on an Amazon-operated marketplace where people can complete surveys in exchange for compensation.

- a. Do you think that the sample of parents was selected in a way that makes it reasonable to think it is representative of the population of all parents?
- b. Is it reasonable to generalize conclusions based on data from this survey to all parents? Explain why or why not.

- 2.29** Participants in a study of honesty in online dating profiles were recruited through print and online advertisements in the *Village Voice*, one of New York City’s most prominent weekly newspapers, and on Craigslist New York City (“[The Truth About Lying in Online Dating Profiles](#),” *Computer-Human Interaction* [2007]: 1–4). The actual height, weight, and age of the participants were compared to what appeared in their online dating profiles. The resulting data was then used to draw conclusions about how common deception was in online dating profiles.

What concerns do you have about generalizing conclusions based on data from this study to the population of all people who have an online dating profile? Be sure to address at least two concerns and give the reasons for your concern.

- 2.30** The article “[Credit Card Activity for College Students](#)” (wallethub.com/edu/credit-card-statistics-for-college-students/25535/, retrieved February 13, 2018) estimated that in 2013, 62% of undergraduates with credit cards pay them off each month and that the average outstanding balance on undergraduates’ credit cards is \$650. These estimates were based on an online survey of 800 college students. What additional information would you want in order to decide if it is reasonable to generalize the reported estimates to the population of all undergraduate students?

- 2.31** The financial aid advisor of a university plans to use a stratified random sample to estimate the average amount of money that students spend on textbooks each term. For each of the following proposed stratification schemes, discuss whether it would be worthwhile to

stratify the university students in this manner. (Hint: Remember that it is desirable to create strata that are homogeneous.)

- a. Strata corresponding to class standing (freshman, sophomore, junior, senior, graduate student)
- b. Strata corresponding to field of study, using the following categories: engineering, architecture, business, other
- c. Strata corresponding to the first letter of the last name: A–E, F–K, etc.

- 2.32** Suppose that you were asked to help design a survey of adult city residents in order to estimate the proportion who would support a sales tax increase. The plan is to use a stratified random sample, and three stratification schemes have been proposed.

Scheme 1: Stratify adult residents into four strata based on the first letter of their last name (A–G, H–N, O–T, U–Z).

Scheme 2: Stratify adult residents into three strata: college students, nonstudents who work full time, nonstudents who do not work full time.

Scheme 3: Stratify adult residents into five strata by randomly assigning residents into one of the five strata.

Which of the three stratification schemes would be best in this situation? Explain.

- 2.33** The article “[High Levels of Mercury Are Found in Californians](#)” (*Los Angeles Times*, February 9, 2006) describes a study in which hair samples were tested for mercury. The hair samples were obtained from more than 6000 people who voluntarily sent hair samples to researchers at Greenpeace and The Sierra Club. The researchers found that nearly one-third of those tested had mercury levels that exceeded the concentration thought to be safe. Is it reasonable to generalize this result to the larger population of U.S. adults? Explain why or why not.

- 2.34** Whether or not to continue a Mardi Gras Parade through downtown San Luis Obispo, California, is a hotly debated topic. The parade is popular with students and many residents, but some celebrations have led to complaints and a call to eliminate the parade. The local newspaper conducted online and telephone surveys of its readers and was surprised by the results. The online survey site received more than 400 responses, with more than 60% favoring continuing the parade, while the telephone response line received more than 120 calls, with more than 90% favoring banning the parade (*San Luis Obispo Tribune*, March 3, 2004). What factors may have contributed to these very different results?

SECTION 2.3 Simple Comparative Experiments

Sometimes the questions we are trying to answer are about the effect of certain explanatory variables on some response. Such questions are often of the form, “What happens when . . . ?” or “What is the effect of . . . ?” For example, an industrial engineer may be considering two different workstation designs and might want to know whether the choice of design affects work performance. A medical researcher may want to determine how a proposed treatment for a disease compares to a standard treatment. Experiments provide a way to collect data to answer questions like these.

DEFINITIONS

Experiment: A study in which one or more explanatory variables are manipulated in order to observe the effect on a response variable.

Explanatory variables: Those variables that have values that are controlled by the experimenter. Explanatory variables are also called **factors**.

Response variable: A variable that is thought to be related to the explanatory variables in an experiment. It is measured as part of the experiment, but it is not controlled by the experimenter.

Experimental condition: Any particular combination of values for the explanatory variables. Experimental conditions are also called **treatments**.

Suppose we are interested in determining the effect of room temperature on performance on a first-year calculus exam. In this case, the explanatory variable is room temperature (it can be manipulated by the experimenter). The response variable is exam performance (the variable that is not controlled by the experimenter and that will be measured).

In general, we can identify the explanatory variables and the response variable easily if we can describe the purpose of the experiment in the following terms:

The purpose is to assess the effect of _____ on _____.
explanatory response
variable variable

Let's return to the example of an experiment to assess the effect of room temperature on exam performance. We might decide to use two room temperature settings, 65° and 75°. This would result in an experiment with two experimental conditions (or equivalently, two treatments) corresponding to the two temperature settings.

Suppose that there are 10 sections of first-semester calculus that have agreed to participate in our study. We might design an experiment in this way: Set the room temperature (in degrees Fahrenheit) to 65° in five of the rooms and to 75° in the other five rooms on test day, and then compare the exam scores for the 65° group and the 75° group. Suppose that the average exam score for the students in the 65° group was noticeably higher than the average for the 75° group. Could we conclude that the increased temperature resulted in a lower average score?

Based on the information given, the answer is no because many other factors might be related to exam score. Were the sections at different times of the day? Did they have different instructors? Different textbooks? Did the sections differ with respect to the abilities of the students? Any of these other factors could provide a plausible explanation (having nothing to do with room temperature) for why the average test score was different for the two groups. It is not possible to separate the effect of temperature from the effects of these other variables. As a consequence, simply setting the room temperatures as described makes for a poorly designed experiment.

A well-designed experiment requires more than just manipulating the explanatory variables. The design must also eliminate other possible explanations for any observed differences in the response variable.

The goal is to design an experiment that will allow us to determine the effects of the explanatory variables on the chosen response variable. To do this, we must take into consideration any **extraneous variables** that, although not of interest in the current study, might also affect the response variable.

DEFINITION

Extraneous variable: A variable that is not one of the explanatory variables in the study but is thought to affect the response variable.

A well-designed experiment copes with the potential effects of extraneous variables by using **random assignment** to experimental conditions and sometimes also by incorporating direct control and/or blocking into the design of the experiment. Each of these strategies—random assignment, direct control, and blocking—is described in the paragraphs that follow.

A researcher can **directly control** some extraneous variables. In the calculus test example, the textbook used is an extraneous variable because part of the differences in test results might be attributed to this variable. We could control this variable directly, by requiring that all sections use the same textbook. Then any observed differences in test scores between temperature groups could not be explained by the use of different textbooks. The extraneous variable *time of day* might also be directly controlled in this way by having all sections meet at the same time.

The effects of some extraneous variables can be filtered out by a process known as **blocking**. Extraneous variables that are addressed through blocking are called *blocking variables*. An investigator using blocking creates groups (called blocks) that are similar with respect to blocking variables. Then all treatments are tried in each block. In our example, we might use *instructor* as a blocking variable. If five instructors are each teaching two sections of calculus, we would make sure that for each instructor, one section was part of the 65° group and the other section was part of the 75° group. With this design, if we see a difference in exam scores for the two temperature groups, the extraneous variable *instructor* can be ruled out as a possible explanation, because all five instructors' students were present in each temperature group. (Had we controlled the instructor variable by choosing to have only one instructor, that would be an example of direct control. Of course we can't directly control both time of day and instructor.)

If one instructor taught all the 65° sections and another taught all the 75° sections, we would be unable to distinguish the effect of temperature from the effect of the instructor. In this situation, the two variables (temperature and instructor) are said to be **confounded**.

Two variables are **confounded** if their effects on the response variable cannot be distinguished from one another.

If an extraneous variable is confounded with the explanatory variables (which define the treatments), it is not possible to draw a clear conclusion about the effect of the treatment on the response. Both direct control and blocking are effective in ensuring that the controlled variables and blocking variables are not confounded with the variables that define the treatments.

We can directly control some extraneous variables by holding them constant, and we can use blocking to create groups that are similar to essentially filter out the effect of other extraneous variables. But what about variables, such as student ability in our calculus test example, which cannot be controlled by the experimenter and which would be difficult to use as blocking variables? These extraneous variables are handled by the use of **random assignment** to experimental groups.

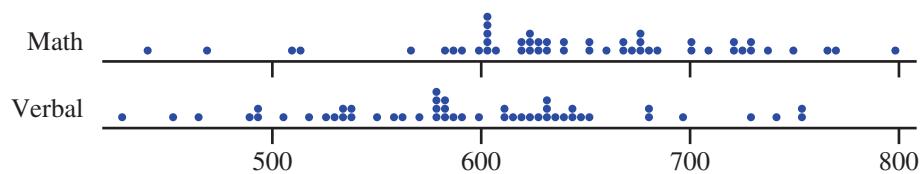
Random assignment ensures that our experiment does not systematically favor one experimental condition over any other and attempts to create experimental groups that are as much alike as possible. For example, if the students requesting calculus could be assigned to one of the ten available sections using a random mechanism, we would expect the resulting groups to be similar with respect to student ability as well as with respect to other extraneous variables that are not directly controlled or used as a basis for blocking.

Notice that random assignment in an experiment is different from random selection of subjects. The ideal situation would be to have both random selection of subjects and random assignment of subjects to experimental conditions, as this would allow conclusions from the experiment to be generalized to a larger population.

For many experiments the random selection of subjects is not possible. As long as subjects are assigned at random to experimental conditions, it is still possible to assess treatment effects.

To get a sense of how random assignment tends to create similar groups, suppose that 50 college freshmen are available to participate as subjects in an experiment to investigate whether completing an online review of course material before an exam improves exam performance. The 50 subjects vary quite a bit with respect to achievement, which is reflected in their math and verbal SAT scores, as shown in Figure 2.2.

FIGURE 2.2
Dotplots of math and verbal SAT scores for 50 freshmen.



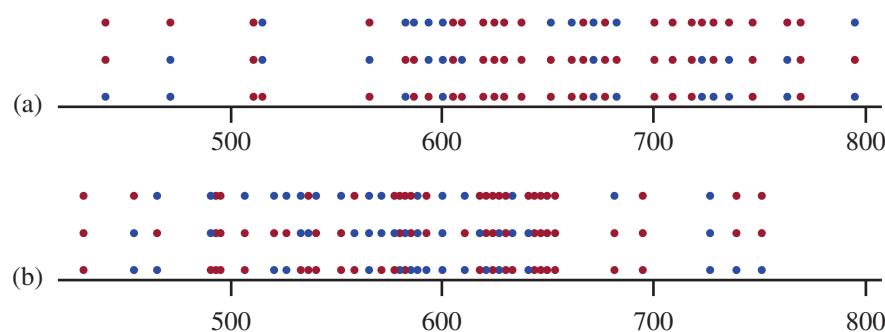
If these 50 students are to be assigned to the two experimental groups (one that will complete the online review and one that will not), we want to make sure that the assignment of students to groups does not favor one group over the other by tending to assign the higher achieving students to one group and the lower achieving students to the other.

Creating groups of students with similar achievement levels in a way that considers both verbal and math SAT scores simultaneously would be difficult, so we rely on random assignment. Figure 2.3(a) shows the math SAT scores of the students assigned to each of the two experimental groups (one shown in red and one shown in blue) for each of three different random assignments of students to groups. Figure 2.3(b) shows the verbal SAT scores for the two experimental groups for each of the same three random assignments.

Notice that each of the three random assignments produced groups that are similar with respect to *both* verbal and math SAT scores. So, if any of these three assignments were used and the two groups differed on exam performance, we could rule out differences in math or verbal SAT scores as possible competing explanations for the difference.

FIGURE 2.3

Dotplots for three different random assignments to two groups, one shown in red and one shown in blue:
 (a) math SAT score;
 (b) verbal SAT score.



Not only will random assignment tend to create groups that are similar with respect to verbal and math SAT scores, but it will also tend to even out the groups with respect to other extraneous variables.

As long as the number of subjects is not too small, we can rely on random assignment to produce comparable experimental groups. This is the reason that random assignment is a part of all well-designed experiments.

Not all experiments require the use of human subjects. For example, a researcher interested in comparing the effect of three different gasoline additives on gas mileage might conduct an experiment using a single car with an empty tank. One gallon of gas with one of the additives will be put in the tank, and the car will be driven along a standard route at a constant speed until it runs out of gas. The total distance traveled on the gallon of gas could then be recorded. This could be repeated a number of times—10, for example—with each additive.

The experiment just described can be viewed as consisting of a sequence of trials. Because a number of extraneous variables (such as variations in environmental conditions like wind speed or humidity and small variations in the condition of the car) might have an effect on gas mileage, it would not be a good idea to use additive 1 for the first 10 trials, additive 2 for the next 10 trials, and so on. A better approach would be to randomly assign additive 1 to 10 of the 30 planned trials, and then randomly assign additive 2 to 10 of the remaining 20 trials. The resulting plan for carrying out the experiment might look as follows:

Trial	1	2	3	4	5	6	7	...	30
Additive	2	2	3	3	2	1	2	...	1

When an experiment can be viewed as a sequence of trials, random assignment involves the random assignment of treatments to trials. *Remember that random assignment—either of subjects to treatments or of treatments to trials—is a critical component of a good experiment.*

Random assignment can be effective only if the number of subjects or observations in each experimental condition (treatment) is large enough for each experimental group to reliably reflect variability in the overall group. For example, if there were only 20 students requesting calculus, it is unlikely that we would get equivalent groups for comparison, even with random assignment to the ten sections. **Replication** is the design strategy of making multiple observations for each experimental condition. Together, replication and random assignment allow the researcher to be reasonably confident of comparable experimental groups.

Principles of Experimental Design

Random Assignment

Random assignment (of subjects to treatments or of treatments to trials) to ensure that the experiment does not systematically favor one experimental condition (treatment) over another.

Blocking

Using extraneous variables to create groups (blocks) that are similar. All experimental conditions (treatments) are then tried in each block.

Direct Control

Holding extraneous variables constant so that their effects are not confounded with those of the experimental conditions (treatments).

Replication

Ensuring that there is an adequate number of observations for each experimental condition.

To illustrate the design of a simple experiment, consider the dilemma of Anna, a waitress in a local restaurant. She would like to increase the amount of her tips, and her strategy is simple: She will write “Thank you” on the back of some of the checks before giving them to the patrons and on others she will write nothing. She plans to calculate the percentage of the tip as her measure of success (for example, a 15% tip is common). She will compare the average percentage of the tips calculated from checks with and without the handwritten “Thank you.” If writing “Thank you” does not produce higher tips, she may try a different strategy.

Anna wonders whether writing “Thank you” on the customers’ bills will have an effect on the amount of her tip. We refer to the writing of “Thank you” and the not writing of “Thank you” as **treatments** (the two experimental conditions to be compared in the experiment). The two treatments together are the possible values of the **explanatory variable**. The tipping percentage is the **response variable**. The idea behind this terminology is that the tipping percentage is a *response* to the treatments *writing “Thank you” or not writing “Thank you.”*

Anna’s experiment may be thought of as an attempt to explain the variability in the response variable in terms of its presumed cause, the variability in the explanatory variable. That is, as she manipulates the explanatory variable, she expects the response by her customers to vary. Anna has a good start, but now she must consider the four fundamental design principles.

Replication. Anna cannot run a successful experiment by gathering tipping information on only one person for each treatment. There is no reason to believe that any single tipping incident is representative of what would happen in other incidents, and therefore it would be impossible to evaluate the two treatments with only two subjects. To interpret the effects of a particular treatment, she must **replicate** each treatment in the experiment.

Blocking. Suppose that Anna works on both Thursdays and Fridays. Because day of the week might affect tipping behavior, Anna should block on day of the week and make sure that observations for both treatments are made on each of the two days.

Direct Control and Random Assignment. There are a number of extraneous variables that might have an effect on the size of tip. Some restaurant patrons will be seated near the window with a nice view. Some will have to wait for a table, whereas others may be seated immediately. Some may be on a fixed income and cannot afford a large tip. Some of these variables can be directly controlled. For example, Anna may choose to use only window tables in her experiment, thus eliminating table location as a potential confounding variable. Other variables, such as length of wait and customer income, cannot be easily controlled. As a result, it is important that Anna use random assignment to decide which

of the window tables will be in the “Thank you” group and which will be in the “No thank you” group. She might do this by flipping a coin as she prepares the check for each window table. If the coin lands with the head side up, she could write “Thank you” on the bill, omitting the “Thank you” when a tail is observed.

The accompanying box summarizes how experimental designs deal with extraneous variables.

Taking Extraneous Variables into Account

Extraneous variables are variables other than the explanatory variables in an experiment that may also have an effect on the response variable. There are several strategies for dealing with extraneous variables in order to avoid confounding.

Extraneous variables that we know about and choose to incorporate into the experimental design:

Strategies

Direct control—holds extraneous variables fixed so that they can't affect the response variable

Blocking—allows for valid comparisons because each treatment is tried in each block

Sometimes extraneous variables are unknown to the experimenter, or might be impossible or very expensive to control.

Extraneous variables that are not incorporated into the experimental design through direct control or blocking:

Strategy

Random assignment

Extraneous variables that are not incorporated into the design of the experiment are sometimes called **lurking variables**.

A Note on Random Assignment

There are several strategies that can be used to perform random assignment of subjects to treatments or treatments to trials. Two common strategies are:

- Write the name of each subject or a unique number that corresponds to a subject on a slip of paper. Place all of the slips in a container and mix well. Then draw out the desired number of slips to determine those that will be assigned to the first treatment group. This process of drawing slips of paper then continues until all treatment groups have been determined.
- Assign each subject a unique number from 1 to n , where n represents the total number of subjects. Use a random number generator or table of random numbers to obtain numbers that will identify which subjects will be assigned to the first treatment group. This process would be repeated, ignoring any random numbers generated that correspond to subjects that have already been assigned to a treatment group, until all treatment groups have been formed.

The two strategies above work well and can be used for experiments in which the desired number of subjects in each treatment group has been predetermined.

Another strategy that is sometimes employed is to use a random mechanism (such as tossing a coin or rolling a die) to determine which treatment will be assigned to a particular subject. For example, in an experiment with two treatments, you might toss a coin to determine if the first subject is assigned to treatment 1 or treatment 2. This could continue for each subject—if the coin lands H, the subject is assigned to treatment 1, and if the coin lands T, the subject is assigned to treatment 2. This strategy is fine, but may result

in treatment groups of unequal size. For example, in an experiment with 100 subjects, 53 might be assigned to treatment 1 and 47 to treatment 2. If this is acceptable, the coin flip strategy is a reasonable way to assign subjects to treatments.

But suppose you want to ensure that there is an equal number of subjects in each treatment group. Is it acceptable to use the coin flip strategy until one treatment group is complete and then just assign all of the remaining subjects to groups that are not yet full? The answer to this question is that it is probably not acceptable. For example, suppose a list of 20 subjects is in order by age from youngest to oldest and that we want to form two treatment groups each consisting of 10 subjects. Tossing a coin to make the assignments might result in the following (based on using the first row of random digits in Appendix A, Table 1, with an even number representing H and an odd number representing T):

Subject	Random Number	Coin Toss Equivalent	Treatment Group
1	4	H	1
2	5	T	2
3	1	T	2
4	8	H	1
5	5	T	2
6	0	H	1
7	3	T	2
8	3	T	2
9	7	T	2
10	1	T	2
11	2	H	1
12	8	H	1
13	4	H	1
14	5	T	2
15	1	T	2
16			1
17	Treatment group 2 filled. Assign all others to treatment group 1.		1
18			1
19			1
20			1

If the list of subjects was ordered by age, treatment group 1 would end up with a disproportionate number of older people. This strategy usually results in one treatment group drawing disproportionately from the end of the list. This means that the only time the strategy of assigning at random until groups fill up and then assigning the remaining subjects to the group that is not full is reasonable is if you can be sure that the list is in random order with respect to all variables that might be related to the response variable. Because of this, it is best to avoid this strategy. Activity 2.5 investigates potential difficulties with this type of strategy.

On the other hand, if the number of subjects is large, it may not be important that every treatment group has exactly the same number of subjects. If this is the case, it is reasonable to use a coin flip strategy (or other strategies of this type) that does not involve stopping assignment of subjects to a group that becomes full.

Evaluating an Experimental Design

The experimental design principles (see page 49) provide a framework for thinking about an experimental design, as illustrated in the following examples.

Example 2.4 Put Away That Laptop!

The article “[The Pen Is Mightier Than the Keyboard](#)” (*Psychological Science* [2014]: 1–10) describes several experiments designed to investigate whether the method that students use to take notes has an effect on learning. In one of the experiments described in the article, 67 students from Princeton University were assigned to one of two groups. Both groups were shown a video of a TED Talk ([seeded.com/talks](#)) and asked to take notes during the talk using their usual note-taking strategy. One group watched the talk in a room equipped with laptops that were not connected to the Internet, and they were asked to use the laptops to take notes. The other group watched the talk in a room where they were given notebooks and asked to take their notes by hand.

When the talk was finished, students were engaged in other tasks for 30 minutes and then were asked to take a test on the material from the TED talk. The test included both fact-recall questions and conceptual-application questions.

The researchers found that the two groups performed equally well on the fact-recall questions, but that on the conceptual-application questions, the group that used laptops to take notes performed significantly worse than the group that took notes by hand.

If we assume that the researcher randomly assigned the subjects to the two groups, then this study is an experiment that compares two treatments (laptop used to take notes and handwritten notes). The responses measured were the scores on the fact-recall questions part of the test and the scores on the conceptual-application questions part of the test. The experiment uses replication (many subjects in each treatment group) and random assignment to control for extraneous variables that might affect the response.

Example 2.5 Morality in the Morning

The article “[The Morning Morality Effect: The Influence of Time of Day on Unethical Behavior](#)” (*Psychological Science* [2014]: 95–102) describes four studies that investigated whether people are more ethical in the morning than in the afternoon. In one of the studies (Experiment 1), volunteers selected either a morning or an afternoon session. During the session, participants were paid to complete a task that presented opportunities to increase the amount earned by purposely providing incorrect responses. The researchers concluded that participants in the afternoon sessions cheated in order to increase the amount they would be paid significantly more often than those in the morning session.

Can you spot an obvious flaw in this study? You probably noticed that the researchers did not randomly assign participants to the two experimental conditions (morning time and afternoon time). This flaw was noted in the paper, and the authors stated

An important limitation of the two previous experiments was that participants self-selected a morning or afternoon session. It is possible that unethical people, in general, are more likely to sign up for afternoon sessions than ethical people are; if true, this would provide an alternative explanation for our previous findings.

To address this limitation, two other experiments were also carried out. In one of these experiments, 70 volunteers were randomly assigned to a morning or afternoon session. In these sessions, participants were shown 20 sets each containing 12 numbers with three digits, such as 6.38. For each of these 20 sets, participants were instructed to look for a pair of numbers that added to 10 (such as 3.86 and 6.14). Each set of numbers was shown for 15 seconds, and then participants indicated whether or not they had found a pair of numbers that added to 10. Participants did not have to specify what those numbers were, and they were told that they would earn money for each pair that they found. Although the participants did not know this, only 10 of the 20 sets of numbers actually contained two numbers that added to 10. This allowed the researchers to assess whether people were cheating on the task by saying they had found the numbers in order to increase the amount they earned. The researchers found that those in the afternoon session reported finding significantly more pairs that added to 10 than the participants in the morning session. This led researchers to conclude that people are more honest in the morning.

Many experiments compare a group that receives a particular treatment to a **control group** that receives no treatment.

Example 2.6 Chilling Newborns? Then You Need a Control Group . . .

Researchers for the National Institute of Child Health and Human Development studied 208 infants whose brains were temporarily deprived of oxygen as a result of complications at birth (*The New England Journal of Medicine*, October 13, 2005). These babies were subjects in an experiment to determine if reducing body temperature for three days after birth improved their chances of surviving without brain damage. The experiment was summarized in a paper that stated “infants were randomly assigned to usual care (control group) or whole-body cooling.” Including a control group in the experiment provided a basis for comparison of death and disability rates for the proposed cooling treatment and those for usual care.

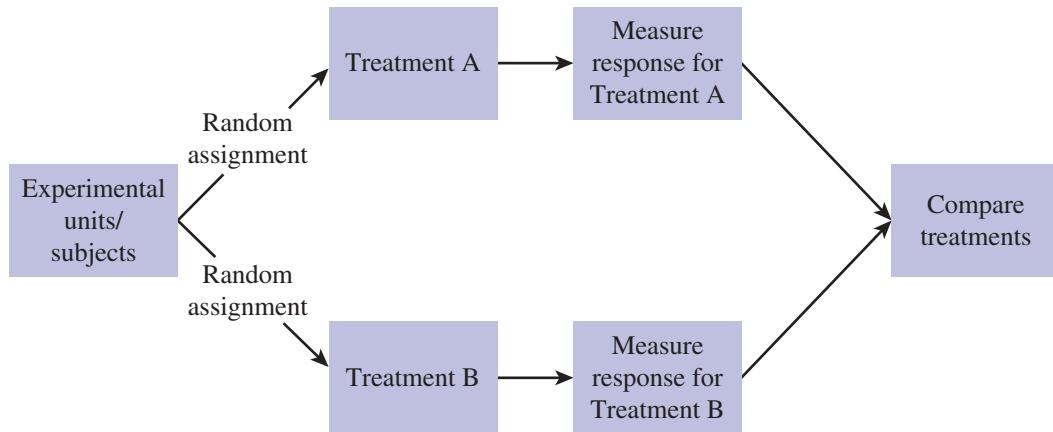
Some extraneous variables that might also affect death and disability rates, such as the duration of oxygen deprivation, could not be directly controlled, so to ensure that the experiment did not unintentionally favor one experimental condition over the other, random assignment of the infants to the two groups was critical. Because this was a well-designed experiment, the researchers were able to use the resulting data and statistical methods that you will see in Chapter 11 to conclude that cooling did reduce the risk of death and disability for infants deprived of oxygen at birth.

Visualizing the Underlying Structure of Some Common Experimental Designs

Simple diagrams are sometimes used to highlight important features of experimental designs. The structure of an experiment that is based on random assignment of experimental units (the units to which treatments are assigned, usually subjects or trials) to one of two treatments is displayed in Figure 2.4. The diagram can be easily adapted for an experiment with more than two treatments. In any particular setting, we would also want to customize the diagram by indicating what the treatments are and what response will be measured. This is illustrated in Example 2.7.

FIGURE 2.4

Diagram of an experiment with random assignment of experimental units to two treatments.



Example 2.7 A Helping Hand

Understand the context ➤

Can moving their hands help children learn math? This is the question investigated by the authors of the paper “*Gesturing Gives Children New Ideas about Math*” (*Psychological Science* [2009]: 267–272). An experiment was conducted to compare two different methods for teaching children how to solve math problems of the form $3 + 2 + 8 = \underline{\hspace{1cm}} + 8$. One method involved having students point to the $3 + 2$ on the left side of the equal sign with one hand and then point to the blank on the right side of the

equal sign before filling in the blank to complete the equation. The other method did not involve using these hand gestures.

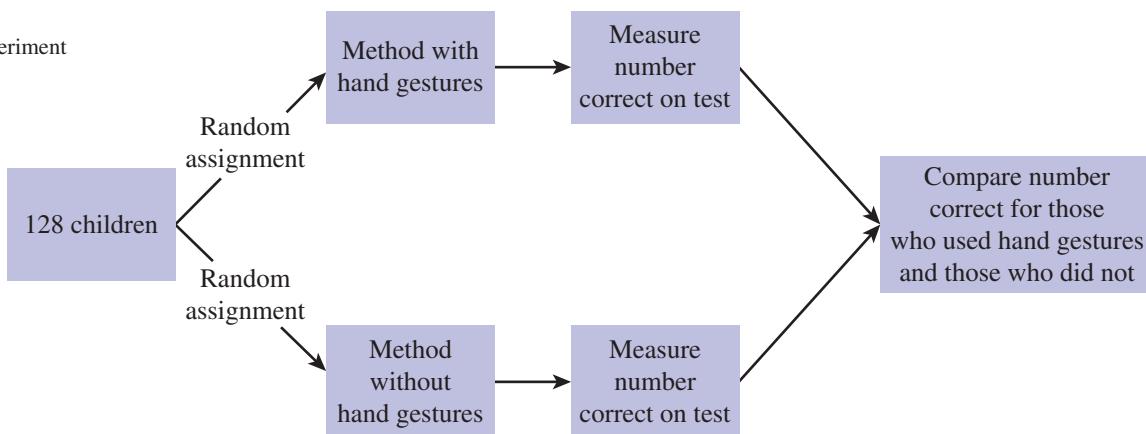
The paper states that the study used children ages 9 and 10 who were given a pretest containing six problems of the type described above. Only children who answered all six questions incorrectly became subjects in the experiment. There were a total of 128 subjects.

To compare the two methods, the 128 children were assigned at random to the two experimental conditions. Children assigned to one experimental condition were taught the method that used hand gestures and children assigned to the other experimental condition were taught a similar strategy that did not involve using hand gestures. Each child then took a test with six problems and the number correct was determined for each child.

The researchers used the resulting data to reach the conclusion that the average number correct for children who used the method that incorporated hand gestures was significantly higher than the average number correct for children who were taught the method that did not use hand gestures.

FIGURE 2.5

Diagram for the experiment of Example 2.7.

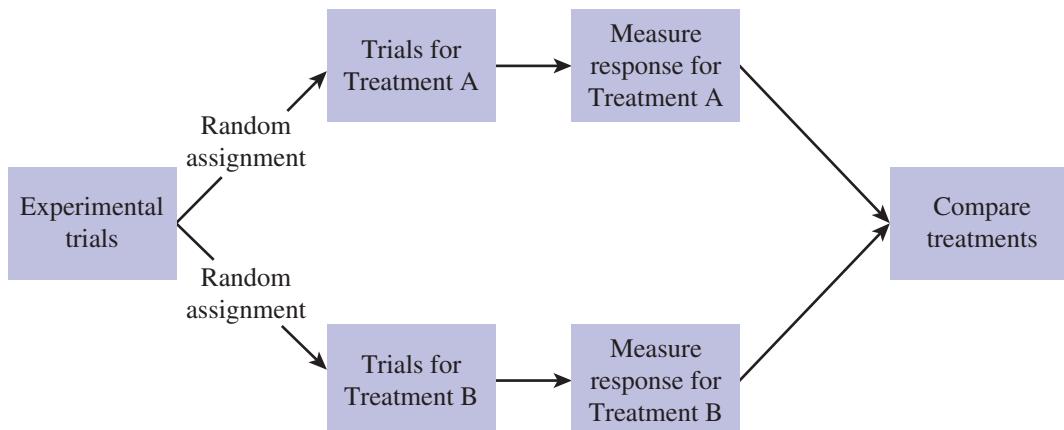


The basic structure of this experiment can be diagrammed as shown in Figure 2.5. This type of diagram provides a nice summary of the experiment, but notice that several important characteristics of the experiment are not captured in the diagram. For example, the diagram does not show that some extraneous variables were considered by the researchers and directly controlled. In this example, both age and prior math knowledge were directly controlled by using only children who were 9 and 10 years old and who were not able to solve any of the questions on the pretest correctly. *This means that while a diagram of an experiment may be a useful tool, it usually cannot stand alone in describing an experimental design.*

Some experiments consist of a sequence of trials, and treatments are assigned at random to the trials. The diagram in Figure 2.6 illustrates the underlying structure of such an experiment. Example 2.8 shows how this diagram can be customized to describe a particular experiment.

FIGURE 2.6

Diagram of an experiment with random assignment of treatments to trials.



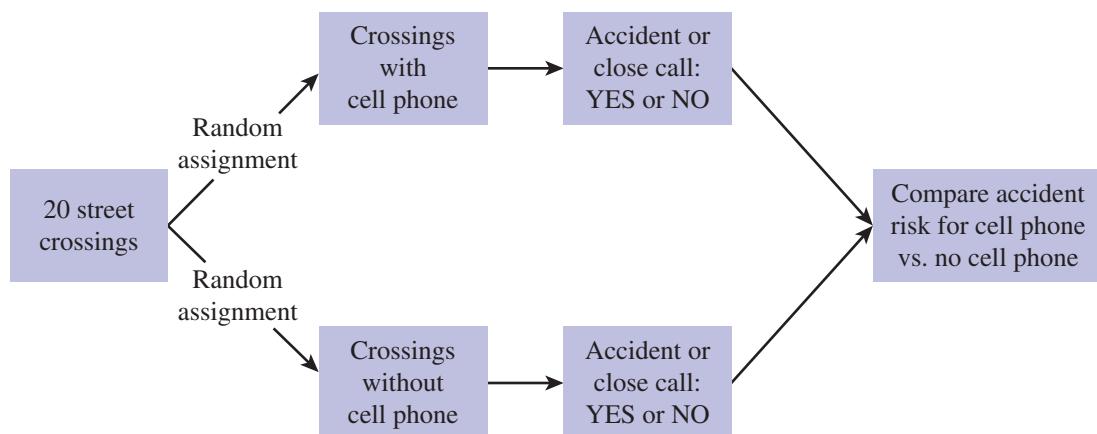
Example 2.8 Distracted? Watch Out for Those Cars!

Understand the context ➤

The paper “Effect of Cell Phone Distraction on Pediatric Pedestrian Injury Risk” (*Pediatrics* [2009]: e179–e185) describes an experiment to investigate whether pedestrians who are talking on a cell phone are at greater risk of an accident when crossing the street than when not talking on a cell phone. No children were harmed in this experiment—a virtual interactive pedestrian environment was used! One possible way of conducting such an experiment would be to have a person cross 20 streets in this virtual environment. The person would talk on a cell phone for some crossings and would not use the cell phone for others.

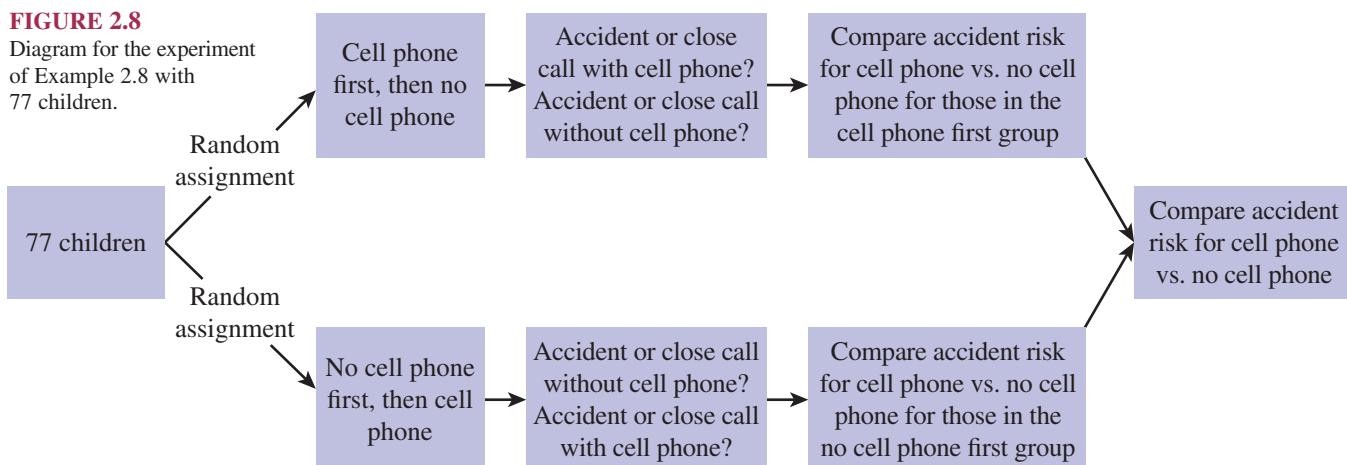
It would be important to randomly assign the two treatments (talking on the phone, not talking on the phone) to the 20 trials (the 20 simulated street crossings). This would result in a design that did not favor one treatment over the other because the pedestrian became more careful with experience or more tired and, therefore, more easily distracted over time. The basic structure of this experiment is diagrammed in Figure 2.7.

FIGURE 2.7
Diagram for the experiment of Example 2.8 with random assignment to trials.



The actual experiment conducted by the authors of the paper was a bit more sophisticated than the one just described. In this experiment, 77 children age 10 and 11 each performed simulated crossings with and without a cell phone. Random assignment was used to decide which children would cross first with the cell phone followed by no cell phone and which children could cross first with no cell phone. The structure of this experiment is diagrammed in Figure 2.8.

FIGURE 2.8
Diagram for the experiment of Example 2.8 with 77 children.



As was the case in Example 2.7, notice that while the diagram is informative, by itself, it does not capture all of the important aspects of the design. In particular, it does not capture the direct control of age (only children age 10 and 11 were used as subjects in the experiment).

Experimental designs in which experimental units are assigned at random to treatments or in which treatments are assigned at random to trials (like those of the experiments in Examples 2.7 and 2.8) are called **completely randomized designs**.

Diagrams are also useful for highlighting the structure of experiments that use blocking. This is illustrated in Example 2.9.

Example 2.9 A Helping Hand Revisited

Understand the context ➤

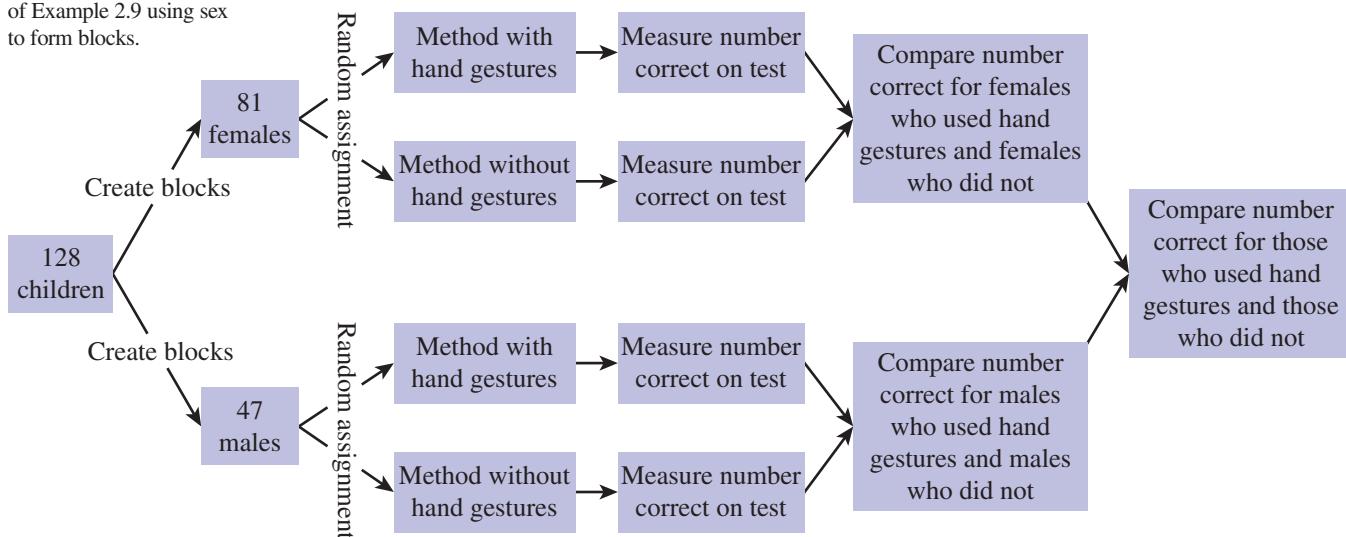
Let's return to the experiment described in Example 2.7. Take a minute to go back and re-read that example. The experiment described in Example 2.7, a completely randomized design with 128 subjects, was used to compare two different methods for teaching kids how to solve a particular type of math problem. Age and prior math knowledge were extraneous variables that the researchers thought might be related to performance on the math test given at the end of the lesson, so the researchers chose to directly control these variables. The 128 children were assigned at random to the two experimental conditions (treatments). The researchers relied on random assignment to create treatment groups that would be roughly equivalent with respect to other extraneous variables.

But suppose that we were worried that sex might also be related to performance on the math test. One possibility would be to use direct control of sex—that is, we might use only males or only females as subjects in the experiment. Then if we saw a difference in test performance for the two teaching methods, it could not be due to one experimental group containing more males and fewer females than the other group. The downside to this strategy is that if we use only males in the experiment, there is no basis for also generalizing any conclusions from the experiment to females.

Another strategy for dealing with extraneous variables is to incorporate blocking into the design. In the case of sex, we could create two blocks, one consisting of females and one consisting of males. Then, once the blocks are formed, we would randomly assign the females to the two treatments and randomly assign the males to the two treatments. In the actual study, the group of 128 children included 81 females and 47 males. A diagram that shows the structure of an experiment that includes blocking using sex is shown in Figure 2.9.

FIGURE 2.9

Diagram for the experiment of Example 2.9 using sex to form blocks.



When blocking is used, the design is called a **randomized block design**. Notice that one difference between the diagram that describes the experiment in which blocking is used (Figure 2.9) and the diagram of the original experiment (Figure 2.5) is the point where the random assignment occurs. *When blocking is incorporated in an experiment, the random assignment to treatments occurs after the blocks have been formed and is done separately for each block.*

Before proceeding with an experiment, you should be able to give a satisfactory answer to each of the following 10 questions.

1. What is the research question that data from the experiment will be used to answer?
2. What is the response variable?
3. How will the values of the response variable be determined?
4. What are the explanatory variables for the experiment?
5. For each explanatory variable, how many different values are there, and what are these values?
6. What are the treatments for the experiment?
7. What extraneous variables might influence the response?
8. How does the design incorporate random assignment of subjects to treatments (or treatments to subjects) or random assignment of treatments to trials?
9. For each extraneous variable listed in Question 7, does the design protect against its potential influence on the response through blocking, direct control, or random assignment?
10. Will you be able to answer the research question using the data collected in this experiment?

EXERCISES 2.35 - 2.51

2.35 The head of the quality control department at a printing company would like to carry out an experiment to determine which of three different glues results in the greatest binding strength. Although they are not of interest in the current study, other factors thought to affect binding strength are the number of pages in the book and whether the book is being bound as a paperback or a hardback. (Hint: See box on page 45.)

- a. What is the response variable in this experiment?
- b. What explanatory variable will determine the experimental conditions?
- c. What two extraneous variables are mentioned in the problem description? Can you think of any other extraneous variables that should be considered?

2.36 A study of college students showed a temporary gain of up to 9 IQ points after listening to a Mozart piano sonata. This conclusion, dubbed the Mozart effect, has since been criticized by a number of researchers who have been unable to confirm the result in similar studies. Suppose that you wanted to see whether there is a Mozart effect for students at your school. (Hint: See Examples 2.4 and 2.5.)

- a. Describe how you might design an experiment for this purpose.
- b. Does your experimental design include direct control of any extraneous variables? Explain.
- c. Does your experimental design use blocking? Explain why you did or did not include blocking in your design.
- d. What role does random assignment play in your design?

2.37 According to the article “**Rubbing Hands Together Under Warm Air Dryers Can Counteract Bacteria Reduction**” (*Infectious Disease News*, September 22, 2010) washing your hands isn’t enough—good “hand hygiene” also includes drying hands thoroughly. The article described an experiment to compare bacteria reduction for three different hand-drying methods. In this experiment, subjects handled uncooked chicken for 45 seconds, then washed their hands with a single squirt of soap for 60 seconds, and then used one of the three hand-drying methods. After completely drying their hands, the bacteria count on their hands was measured.

Suppose you want to carry out a similar experiment and that you have 30 subjects who are willing to participate. Describe a method for randomly assigning each of the 30 subjects to one of the hand-drying methods. (Hint: See A Note on Random Assignment on page 50.)

2.38 The following is from an article titled “**After the Workout, Got Chocolate Milk?**” that appeared in the *Chicago Tribune* (January 18, 2005):

Researchers at Indiana University at Bloomington have found that chocolate milk effectively helps athletes recover from an intense workout. They had nine cyclists bike, rest four hours, then bike again, three separate times. After each workout, the cyclists downed chocolate milk or energy drinks Gatorade or Endurox (two to three glasses per hour); then, in the second workout of each set, they cycled to exhaustion. When they drank chocolate milk, the amount of time they could cycle until they were exhausted was similar to

when they drank Gatorade and longer than when they drank Endurox.

The article is not explicit about this, but in order for this to have been a well-designed experiment, it must have incorporated random assignment. Briefly explain where the researcher would have needed to use random assignment in order for the stated conclusion to be valid.

- 2.39** The report “[Comparative Study of Two Computer Mouse Designs](#)” ([Cornell Human Factors Laboratory Technical Report RP7992](#)) included the following description of the subjects used in an experiment:

Twenty-four Cornell University students and staff (12 males and 12 females) volunteered to participate in the study. Three groups of 4 men and 4 women were selected by their stature to represent the 5th percentile (female 152.1 ± 0.3 cm, male 164.1 ± 0.4 cm), 50th percentile (female 162.4 ± 0.1 cm, male 174.1 ± 0.7 cm), and 95th percentile (female 171.9 ± 0.2 cm, male 185.7 ± 0.6 cm) ranges . . . All subjects reported using their right hand to operate a computer mouse.

This experimental design incorporated direct control and blocking.

- a. Is the potential effect of the extraneous variable stature (height) addressed by blocking or direct control?
- b. Whether the right or left hand is used to operate the mouse was considered to be an extraneous variable. Is the potential effect of this variable addressed by blocking or direct control?

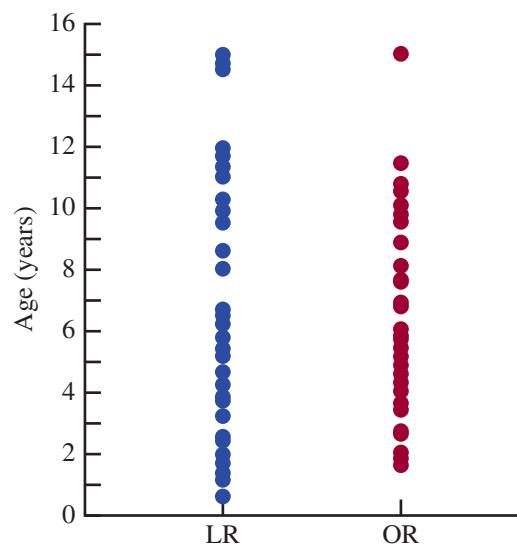
- 2.40** The Institute of Psychiatry at Kings College London found that dealing with infomania (information overload) has a temporary, but significant derogatory effect on IQ ([Discover, November 2005](#)). Researchers divided volunteers into two groups. Each subject took an IQ test. One group had to check e-mail and respond to instant messages while taking the test, and the second group took the test without any distraction. The distracted group had an average score that was 10 points lower than the average for the control group.

Explain why it is important that the researchers created the two experimental groups in this study by using random assignment.

- 2.41** In an experiment to compare two different surgical procedures for hernia repair (“[A Single-Blinded, Randomized Comparison of Laparoscopic Versus Open Hernia Repair in Children](#),” [Pediatrics \[2009\]: 332–336](#)), 89 children were assigned at random to one of the two surgical methods. The researchers relied on the random assignment of subjects to

treatments to create comparable groups with respect to extraneous variables that they did not control. One such extraneous variable was age.

After random assignment to treatments, the researchers looked at the age distribution of the children in each of the two experimental groups (laparoscopic repair [LR] and open repair [OR]). The accompanying figure is similar to one in the paper.



Based on this figure, has the random assignment of subjects to experimental groups been successful in creating groups that are similar with respect to the ages of the children in the groups? Explain.

- 2.42** In many digital environments, users are allowed to choose how they are represented visually online. Does how people are represented online affect online behavior? This question was examined by the authors of the paper “[The Proteus Effect: The Effect of Transformed Self-Representation on Behavior](#)” ([Human Communication Research \[2007\]: 271–290](#)). Participants were randomly assigned either an attractive avatar (a graphical image that represents a person) to represent them or an unattractive avatar.
- a. The researchers concluded that when interacting with a person of the opposite gender in an online virtual environment, those assigned an attractive avatar moved significantly closer to the other person than those who had been assigned an unattractive avatar. This difference was attributed to the attractiveness of the avatar. Explain why the researchers would not have been able to reach this conclusion if participants had been allowed to choose one of the two avatars (attractive, unattractive) to represent them online.
 - b. Construct a diagram to represent the underlying structure of this experiment.

- 2.43** Does playing action video games provide more than just entertainment? The authors of the paper “Action-Video-Game Experience Alters the Spatial Resolution of Vision” (*Psychological Science* [2007]: 88–94) concluded that spatial resolution, an important aspect of vision, is improved by playing action video games. They based this conclusion on data from an experiment in which 32 volunteers who had not played action video games were “equally and randomly divided between the experimental and control groups.” Subjects in each group played a video game for 30 hours over a period of 6 weeks. Those in the experimental group played Unreal Tournament 2004, an action video game. Those in the control group played the game Tetris, a game that does not require the user to process multiple objects at once.

Explain why the random assignment to the two groups is an important aspect of this experiment.

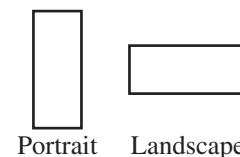
- 2.44** Construct a diagram to represent the note-taking experiment of Example 2.4.
- 2.45** Construct a diagram to represent the gasoline additive experiment described on page 48.
- 2.46** An advertisement for a sweatshirt that appeared in *SkyMall Magazine* (a catalog distributed by some airlines) stated the following: “This is not your ordinary hoody! Why? Fact: Research shows that written words on containers of water can influence the water’s structure for better or worse depending on the nature and intent of the word. Fact: The human body is 70% water. What if positive words were printed on the inside of your clothing?”

The reference to the “fact” that written words on containers of water can influence the water’s structure appears to be based on the work of Dr. Masaru Emoto who typed words on paper, pasted the words on bottles of water, and observed how the water reacted to the words by seeing what kind of crystals were formed in the water. He describes several of his experiments in his self-published book, *The Message from Water*.

If you were going to interview Dr. Emoto, what questions would you want to ask him about the design of his experiment?

- 2.47** The paper “Turning to Learn: Screen Orientation and Reasoning from Small Devices” (*Computers in Human Behavior* [2011]: 793–797) describes a study that investigated whether cell phones with small screens are useful for gathering information. The researchers wondered if the ability to reason using information read on a small screen was affected by the screen orientation. The researchers assigned 33 undergraduate students who were enrolled in a psychology course at a large public university to one of two groups at random.

One group read material that was displayed on a small screen in portrait orientation, and the other group read material on the same size screen but turned to display the information in landscape orientation (see figure below).



The researchers found that performance on a reasoning test that was based on the displayed material was better for the group that read material in the landscape orientation.

- Is the described study an observational study or an experiment?
- Did the study use random selection from some population?
- Did the study use random assignment to experimental groups?

- 2.48** Consider the study described in the previous exercise.

- Is the conclusion—that reasoning using information displayed on a small screen is improved by turning the screen to landscape orientation—appropriate, given the study design? Explain.
- Is it reasonable to generalize the conclusions from this study to some larger population? If so, what population?

- 2.49** The Pew Research Center conducted a study of gender bias. The report “Men or Women: Who is the Better Leader? A Paradox in Public Attitudes” (pewsocialtrends.org, August 28, 2008) describes how the study was conducted:

In the experiment, two separate random samples of more than 1000 registered voters were asked to read a profile sent to them online of a hypothetical candidate for U.S. Congress in their district. One random sample of 1161 respondents read a profile of Ann Clark, described as a lawyer, a churchgoer, a member of the local Chamber of Commerce, an environmentalist and a member of the same party as the survey respondent. They were then asked what they liked and didn’t like about her, whether they considered her qualified and whether they were inclined to vote for her. There was no indication that this was a survey about gender or gender bias. A second random sample of 1139 registered voters was asked to read a profile of Andrew Clark, who—except for his gender—was identical in every way to Ann Clark. These respondents were then asked the same questions.

- What are the two treatments in this experiment?
- What are the response variables in this experiment?

- 2.50** Consider the study described in the previous exercise. Explain why “taking two separate random samples” has the same benefits as random assignment to the two treatments in this experiment.
- 2.51** Red wine contains flavonol, an antioxidant thought to have beneficial health effects. But to have an effect, the antioxidant must be absorbed into the blood. The article *“Red Wine is a Poor Source of Bioavailable Flavonols in Men”* (*The Journal of Nutrition* [2001]: 745–748) describes a study to investigate three sources of dietary flavonol—red wine, yellow onions, and black tea—to determine

the effect of source on absorption. The article included the following statement:

We recruited subjects via posters and local newspapers. To ensure that subjects could tolerate the alcohol in the wine, we only allowed men with a consumption of at least seven drinks per week to participate . . . Throughout the study, the subjects consumed a diet that was low in flavonols.

- What are the three treatments in this experiment?
- What is the response variable?
- What are three extraneous variables that the researchers chose to control in the experiment?

SECTION 2.4 More on Experimental Design

The previous section covered basic principles for designing simple comparative experiments—control, blocking, random assignment, and replication. The goal of an experimental design is to provide a method of data collection that

- minimizes the effect of extraneous sources of variability in the response so that any differences in response for various experimental conditions can be more easily assessed and
- creates experimental groups that are similar with respect to extraneous variables that cannot be controlled either directly or through blocking.

In this section, we look at some additional things to consider when planning an experiment.

Use of a Control Group

If the purpose of an experiment is to determine whether some treatment has an effect, it is important to include an experimental group that does not receive the treatment. Such a group is called a **control group**. The use of a control group allows the experimenter to assess how the response variable behaves when the treatment is not used. This provides a baseline against which the treatment groups can be compared to determine whether the treatment had an effect.

Example 2.10 Comparing Gasoline Additives

Suppose that an engineer wants to know whether a gasoline additive increases fuel efficiency (miles per gallon). The experiment might use a single car (to eliminate car-to-car variability) and a sequence of trials in which 1 gallon of gas is put in an empty tank, the car is driven around a racetrack at a constant speed, and the distance traveled on the gallon of gas is recorded.

To determine whether the additive increases gas mileage, it would be necessary to include a control group of trials in which distance traveled was measured when gasoline without the additive was used. The trials would be assigned *at random* to one of the two experimental conditions (additive or no additive).

Even though this experiment consists of a sequence of trials all with the same car, random assignment of trials to experimental conditions is still important because there might be extraneous variables that could affect fuel efficiency. For example, temperature or other environmental conditions might change over the sequence of trials, or the physical condition of the car might change slightly from one trial to another. Random assignment of experimental conditions to trials will tend to even out the effects of these extraneous variables.



Stockbyte/Getty Images

Although we usually think of a control group as one that receives no treatment, in experiments designed to compare a new treatment to an existing standard treatment, the term control group is sometimes also used to describe the group that receives the current standard treatment.

Not all experiments require the use of a control group. Many experiments are designed to compare two or more conditions—for example, an experiment to determine how oven temperature affects the cooking time of a particular type of cake. However, sometimes a control group is included even when the ultimate goal is to compare two or more different treatments. This is because an experiment with two treatments and no control group might allow us to determine whether there is a difference between the two treatments and even to assess the magnitude of the difference if one exists, but it would not allow us to assess the individual effect of either treatment. For example, without a control group, we might be able to say that there is no difference in the increase in mileage for two different gasoline additives, but we would not be able to tell if this was because both additives increased gas mileage by a similar amount or because neither additive had any effect on gas mileage.

Use of a Placebo

In experiments that use human subjects, use of a control group may not be enough to determine whether a treatment really does have an effect. People sometimes respond merely to the power of suggestion! For example, suppose a study designed to determine whether a particular herbal supplement is effective in promoting weight loss uses an experimental group that takes the herbal supplement and a control group that takes nothing. It is possible that those who take the herbal supplement and believe that they are taking something that will help them to lose weight may be more motivated and may unconsciously change their eating behavior or activity level, resulting in weight loss.

Although there is debate about the degree to which people respond, many studies have shown that people sometimes respond to treatments with no active ingredients and that they often report that such “treatments” relieve pain or reduce symptoms. So, if researchers want to determine whether a treatment really has an effect, comparing a treatment group to a control group may not be enough. To address the problem, many experiments use what is called a **placebo**.

DEFINITION

Placebo: Something that is identical (in appearance, taste, feel, etc.) to the treatment received by the treatment group, except that it contains no active ingredients.

For example, in the herbal supplement experiment, rather than using a control group that received *no* treatment, the researchers might want to include a placebo group. Individuals in the placebo group would take a pill that looked just like the herbal supplement but did not contain the herb or any other active ingredient. As long as the subjects did not know whether they were taking the herb or the placebo, the placebo group would provide a better basis for comparison and would allow the researchers to determine whether the herbal supplement had any real effect over and above the “placebo effect.”

Single-Blind and Double-Blind Experiments

Because people often have their own personal beliefs about the effectiveness of various treatments, it is sometimes desirable to conduct experiments in a way that subjects do not know what treatment they are receiving. For example, in an experiment comparing four different doses of a medication for relief of headache pain, someone who knows that he is receiving the medication at its highest dose may be subconsciously influenced to report a greater degree of headache pain reduction. By ensuring that subjects are not aware of which treatment they receive, we can prevent the subjects’ personal perceptions from influencing the response.

An experiment in which subjects do not know what treatment they have received is described as **single-blind**. Of course, not all experiments can be made single-blind. For example, in an experiment to compare the effect of two different types of exercise on blood pressure, it is not possible for participants to be unaware of whether they are in the swimming group or the jogging group! However, when it is possible, “blinding” the subjects in an experiment is generally a good strategy.

In some experiments, someone other than the subject is responsible for measuring the response. To ensure that the person measuring the response does not let personal beliefs influence the way in which the response is recorded, the researchers should make sure that the measurer does not know which treatment was given to any particular individual. For example, in a medical experiment to determine whether a new vaccine reduces the risk of getting the flu, doctors must decide whether a particular individual who is not feeling well actually has the flu or some other unrelated illness. If the doctor knew that a participant with flu-like symptoms had received the new flu vaccine, she might be less likely to determine that the participant had the flu and more likely to interpret the symptoms as being the result of some other illness.

We have now considered two ways in which blinding might occur in an experiment. One involves blinding the subjects, and the other involves blinding the individuals who measure the response. If subjects do not know which treatment was received *and* those measuring the response do not know which treatment was given to which subject, the experiment is described as **double-blind**. If only one of the two types of blinding is present, the experiment is **single-blind**.

DEFINITIONS

Double-blind experiment: An experiment in which neither the subjects nor the individuals who measure the response know which treatment was received.

Single-blind experiment: An experiment in which the subjects do not know which treatment was received but the individuals measuring the response do know which treatment was received, or one in which the subjects do know which treatment was received but the individuals measuring the response do not know which treatment was received.

Experimental Units and Replication

An **experimental unit** is the smallest unit to which a treatment is applied. In the language of experimental design, treatments are assigned at random to experimental units, and replication means that each treatment is applied to more than one experimental unit.

Replication is necessary for random assignment to be an effective way to create similar experimental groups and to get a sense of the variability in the values of the response for individuals who receive the same treatment. As we will see in Chapters 9–15, this enables us to use statistical methods to decide whether differences in the responses in different treatment groups can be attributed to the treatment received or whether they can be explained by chance variation (the natural variability seen in the responses to a single treatment).

Be careful when designing an experiment to ensure that there is replication. For example, suppose that children in two third-grade classes are available to participate in an experiment to compare two different methods for teaching arithmetic. It might at first seem reasonable to select one class at random to use one method and then assign the other method to the remaining class. But what are the experimental units here? If treatments are randomly assigned to classes, classes are the experimental units. Because only one class is assigned to each treatment, this is an experiment with no replication, even though there are many children in each class. We would *not* be able to determine whether there was

a difference between the two methods based on data from this experiment, because we would have only one observation per treatment.

One last note on replication: Do not confuse replication in an experimental design with replicating an experiment. Replicating an experiment means conducting a new experiment using the same experimental design as a previous experiment. This is a way of confirming conclusions based on a previous experiment, but it does not eliminate the need for replication in each of the individual experiments themselves.

Using Volunteers as Subjects in an Experiment

Although the use of volunteers in a study that involves collecting data through sampling is never a good idea, it is a common practice to use volunteers as subjects in an experiment. Even though the use of volunteers limits the researcher's ability to generalize to a larger population, random assignment of the volunteers to treatments should result in comparable groups, and so treatment effects can still be assessed.

EXERCISES 2.52 - 2.63

- 2.52** Explain why some studies include both a control group and a placebo treatment. What additional comparisons are possible if both a control group and a placebo group are included?
- 2.53** Explain why blinding is a reasonable strategy in many experiments.
- 2.54** Give an example of an experiment for each of the following:
- Single-blind experiment with the subjects blinded
 - Single-blind experiment with the individuals measuring the response blinded
 - Double-blind experiment
 - An experiment for which it is not possible to blind the subjects
- 2.55** The article “[Study Points to Benefits of Knee Replacement Surgery Over Therapy Alone](#)” (*The New York Times*, October 21, 2015) describes a study to compare two treatments for people with knee pain. In the study, 50 people with arthritis received knee replacement surgery followed by a program of exercise. Another 50 people with arthritis did not have surgery but received the same program of exercise. After 1 year, 85% of the people who had surgery and 68% of the people who did not have surgery reported pain relief.
- Why would it be important to determine if the researchers randomly assigned the people participating in the study to one of the two groups?
 - Explain why you think that the researchers did not include a control group (a group that did not receive surgery and also did not receive an exercise program) in this study.
- 2.56** In an experiment to compare two different surgical procedures for hernia repair ([“A Single-Blinded, Randomized Comparison of Laparoscopic Versus Open Hernia Repair in Children,” Pediatrics \[2009\]: 332–336](#)), 89 children were assigned at random to one of the two surgical methods. The methods studied were laparoscopic repair and open repair. In laparoscopic repair, three small incisions are made and the surgeon works through these incisions with the aid of a small camera that is inserted through one of the incisions. In the open repair, a larger incision is used to open the abdomen. One of the response variables in this study was the amount of medication that was given after the surgery for the control of pain and nausea. The paper states “For postoperative pain, rescue fentanyl (1 µg/kg) and for nausea, ondansetron (0.1 mg/kg) were given as judged necessary by the attending nurse blinded to the operative approach.”
- Why do you think it was important that the nurse who administered the medications did not know which type of surgery was performed?
 - Explain why it was not possible for this experiment to be double-blind.
- 2.57** The article “[Placebos Are Getting More Effective. Drug Makers Are Desperate to Know Why.](#)” (*Wired Magazine*, August 8, 2009) states that “according to research, the color of a tablet can boost the effectiveness even of genuine meds—or help convince a patient that a placebo is a potent remedy.” Describe how you would design an experiment to investigate if adding color to Tylenol tablets would result in greater perceived pain relief. Be sure to address how you would select subjects, how you would measure pain relief, what colors you would use, and whether or not you would include a control group in your experiment.

- 2.58** The article “[Yes That Miley Cyrus Biography Helps Learning](#)” (*The Globe and Mail*, August 5, 2010) describes an experiment investigating whether providing summer reading books to low-income children would affect school performance. Subjects in the experiment were 1330 children randomly selected from first and second graders at low-income schools in Florida. A group of 852 of these children were selected at random from the group of 1330 participants to be in the “book” group. The other 478 children were assigned to the control group.

Children in the book group were invited to a book fair in the spring to choose any 12 reading books which they could then take home. Children in the control group were not given any reading books but were given some activity and puzzle books. This process was repeated each year for 3 years until the children reached third and fourth grade. The researchers then compared reading test scores of the two groups.

- Do you think that randomly selecting 852 of the 1330 children to be in the book group is equivalent to random assignment of the children to the two experimental groups? Explain.
- Explain the purpose of including a control group in this experiment.

- 2.59** Suppose that the researchers who carried out the experiment described in the previous exercise thought that sex might be a potentially confounding variable. If 700 of the children participating in the experiment were females and 630 were males, describe how blocking could be incorporated into the experiment. Be specific about how you would assign the children to treatment groups.

- 2.60** The article “[Doctor Dogs Diagnose Cancer by Sniffing It Out](#)” (*Knight Ridder Newspapers*, January 9, 2006) reports the results of an experiment described in the journal *Integrative Cancer Therapies*. In this experiment, dogs were trained to distinguish between people with breast and lung cancer and people without cancer by sniffing exhaled breath. Dogs were trained to lie down if they detected cancer in a breath sample. After training, dogs’ ability to detect cancer was tested using breath samples from people whose breath had not been used in training the dogs. The paper states “The researchers blinded both the dog handlers and the experimental observers to the identity of the breath samples.” Explain why this blinding is an important aspect of the design of this experiment.

- 2.61** Pismo Beach, California, has an annual clam festival that includes a clam chowder contest. Judges rate clam chowders from local restaurants, and the judging is done in such a way that the judges are not aware of which chowder is from which restaurant. One year, much to the dismay of the seafood restaurants on the waterfront, Denny’s chowder was declared the winner! (When asked what the ingredients were, the cook at Denny’s said he wasn’t sure—he just had to add the right amount of nondairy creamer to the soup stock that he got from Denny’s distribution center!)

- Do you think that Denny’s chowder would have won the contest if the judging had not been “blind”? Explain.
- Although this was not an experiment, your answer to Part (a) helps to explain why those measuring the response in an experiment are often blinded. Using your answer in Part (a), explain why experiments are often blinded in this way.

- 2.62** The [San Luis Obispo Tribune](#) (May 7, 2002) reported that “a new analysis has found that in the majority of trials conducted by drug companies in recent decades, sugar pills have done as well as—or better than—antidepressants.” What effect is being described here? What does this imply about the design of experiments with a goal of evaluating the effectiveness of a new medication?

- 2.63** The article “[A Debate in the Dentist’s Chair](#)” (*San Luis Obispo Tribune*, January 28, 2000) described an ongoing debate over whether newer resin fillings are a better alternative to the more traditional silver amalgam fillings. Because amalgam fillings contain mercury, there is concern that they could be mildly toxic and prove to be a health risk to those with some types of immune and kidney disorders. One experiment described in the article used sheep as subjects and reported that sheep treated with amalgam fillings had impaired kidney function.

- In the experiment, a control group of sheep that received no fillings was used but there was no placebo group. Explain why it is not necessary to have a placebo group in this experiment.
- The experiment compared only an amalgam filling treatment group to a control group. What would be the benefit of also including a resin filling treatment group in the experiment?
- Why do you think the experimenters used sheep rather than human subjects?

SECTION 2.5 Interpreting and Communicating the Results of Statistical Analyses

Statistical studies are conducted to answer questions about characteristics of some population of interest or about the effect of some treatment. These questions are answered using data, and how the data are obtained determines the quality of information available and the type of conclusions that can be drawn. As a consequence, when describing a study you have conducted (or when evaluating a published study), it is important to consider how the data were collected.

The description of the data collection process should make it clear whether the study is an observational study or an experiment. For observational studies, some of the issues that should be addressed are:

1. What is the population of interest? What is the sampled population? Are these two populations the same? If the sampled population is only a subset of the population of interest, **undercoverage** limits our ability to generalize to the population of interest. For example, if the population of interest is all students at a particular university, but the sample is selected from only those students who choose to list their phone number in the campus directory, undercoverage may be a problem. We would need to think carefully about whether it is reasonable to consider the sample as representative of the population of all students at the university.

Oversampling results when the sampled population is actually larger than the population of interest. This would be the case if we were interested in the population of all high schools that offer Advanced Placement (AP) Statistics but sampled from a list of all schools that offered an AP class in any subject. Both undercoverage and oversampling can be problematic.

2. How were the individuals or objects in the sample actually selected? A description of the sampling method helps the reader to make judgments about whether the sample can reasonably be viewed as representative of the population of interest.
3. What are potential sources of bias, and is it likely that any of these will have a substantial effect on the observed results? When describing an observational study, you should acknowledge that you are aware of potential sources of bias and explain any steps that were taken to minimize their effect. For example, in a mail survey, nonresponse can be a problem, but the sampling plan may seek to minimize its effect by offering incentives for participation and by following up one or more times with those who do not respond to the first request.

A common misperception is that increasing the sample size is a way to reduce bias in observational studies, but this is not the case. For example, if measurement bias is present, as in the case of a scale that is not correctly calibrated and tends to weigh too high, taking 1000 measurements rather than 100 measurements cannot correct for the fact that the measured weights will be too large. Similarly, a larger sample size cannot compensate for response bias introduced by a poorly worded question.

For experiments, some of the issues that should be addressed are:

1. What is the role of random assignment? All good experiments use random assignment as a means of coping with the effects of potentially confounding variables that cannot easily be directly controlled. When describing an experimental design, you should be clear about how random assignment (subjects to treatments, treatments to subjects, or treatments to trials) was incorporated into the design.
2. Were any extraneous variables directly controlled by holding them at fixed values throughout the experiment? If so, which ones and at which values?
3. Was blocking used? If so, how were the blocks created? If an experiment uses blocking to create groups of homogeneous experimental units, the criteria used to create the blocks should be described and an explanation of why these blocks were used should be given. For example, we might say something like “Subjects

were divided into two blocks—those who exercise regularly and those who do not exercise regularly—because it was believed that exercise status might affect the responses to the diets.”

Because each treatment appears at least once in each block, the block size must be at least as large as the number of treatments. Ideally, the block sizes should be equal to the number of treatments, because this presumably would allow the experimenter to create small groups of extremely homogeneous experimental units. For example, in an experiment to compare two methods for teaching calculus to first-year college students, we may want to block on previous mathematics knowledge by using math SAT scores. If 100 students are available as subjects for this experiment, rather than creating two large groups (above-average math SAT score and below-average math SAT score), we might want to create 50 blocks of two students each, the first consisting of the two students with the highest math SAT scores, the second containing the two students with the next highest scores, and so on. We would then select one student in each block at random and assign that student to teaching method 1. The other student in the block would be assigned to teaching method 2.

A Word to the Wise: Cautions and Limitations

It is a mistake to begin collecting data before thinking carefully about research objectives and developing a plan. A poorly designed plan for data collection may result in data that do not enable the researcher to answer key questions of interest or to generalize conclusions based on the data to the desired populations of interest.

Clearly defining the objectives at the outset enables the investigator to determine whether an experiment or an observational study is the best way to proceed. Watch out for the following *inappropriate* actions:

1. Drawing a cause-and-effect conclusion from an observational study. Don’t do this, and don’t believe it when others do it!
2. Generalizing results of an experiment that uses volunteers as subjects to a larger population. This is not sensible without a convincing argument that the group of volunteers can reasonably be considered to be representative of the population.
3. Generalizing conclusions based on data from a sample to some population of interest. This is sometimes a sensible thing to do, but on other occasions it is not reasonable. Generalizing from a sample to a population is justified only when there is reason to believe that the sample is likely to be representative of the population. This would be the case if the sample was a random sample from the population and there were no major potential sources of bias. If the sample was not selected at random or if potential sources of bias were present, these issues would have to be addressed before a judgment could be made regarding the appropriateness of generalizing the study results.

For example, the [Associated Press \(January 25, 2003\)](#) reported on the high cost of housing in California. The median home price was given for each of the 10 counties in California with the highest home prices. Although these 10 counties are a sample of the counties in California, they were not randomly selected and (because they are the 10 counties with the highest home prices) it would not be reasonable to generalize to all California counties based on data from this sample.

4. Generalizing conclusions based on an observational study that used voluntary response or convenience sampling to a larger population. This is almost never reasonable.

EXERCISES 2.64 - 2.68

- 2.64** The following paragraph appeared in *USA TODAY* (August 6, 2009):

Cement doesn't hold up to scrutiny

A common treatment that uses medical cement to fix cracks in the spinal bones of elderly people worked no better than a sham treatment, the first rigorous studies of a popular procedure reveal. Pain and disability were virtually the same up to six months later, whether patients had a real treatment or a fake one, shows the research in today's *New England Journal of Medicine*. Tens of thousands of Americans each year are treated with bone cement, especially older women with osteoporosis. The researchers said it is yet another example of a procedure coming into wide use before proven safe and effective. Medicare pays \$1500 to \$2100 for the outpatient procedure.

The paper referenced in this paragraph is "A Randomized Trial of Vertebroplasty for Painful Osteoporotic Vertebral Fractures" (*New England Journal of Medicine* [2009]: 557–568). Obtain a copy of this paper through your university library or your instructor. Read the following sections of the paper: the abstract on page 557; the study design section on page 558; the participants section on pages 558–559; the outcome assessment section on pages 559–560; and the discussion section that begins on page 564.

The summary of this study that appeared in *USA TODAY* consisted of just one paragraph. If the newspaper had allowed four paragraphs, other important aspects of the study could have been included. Write a four-paragraph summary that the paper could have used. Remember—you are writing for the *USA TODAY* audience, not for the readers of *The New England Journal of Medicine*!

- 2.65** The article "Effects of Too Much TV Can Be Undone" (*USA TODAY*, October 1, 2007) included the following paragraph:

Researchers at Johns Hopkins Bloomberg School of Public Health report that it's not only how many hours children spend in front of the TV, but at what age they watch that matters. They analyzed data from a national survey in which parents of 2707 children were interviewed first when the children were 30–33 months old and again when they were $5\frac{1}{2}$, about their TV viewing and their behavior.

- Is the study described an observational study or an experiment?
- The article says that data from a sample of 2707 parents were used in the study. What other information about the sample would you want in order to evaluate the study?

- 2.66** The actual paper referred to in the *USA TODAY* article described in the previous exercise was "Children's Television Exposure and Behavioral and Social Outcomes at 5.5 years: Does Timing of Exposure Matter?" (*Pediatrics* [2007]: 762–769). The paper describes the sample as follows:

The study sample included 2707 children whose mothers completed telephone interviews at both 30 to 33 months and 5.5 years and reported television exposure at both time points. Of those completing both interviewers, 41 children (1%) were excluded because of missing data on television exposure at one or both time points. Compared with those enrolled in the HS clinical trial, parents in the study sample were disproportionately older, white, more educated, and married.

- The "HS clinical trial" referred to in the excerpt from the paper was a nationally representative sample used in the Healthy Steps for Young Children national evaluation. Based on the above description of the study sample, do you think that it is reasonable to regard the sample as representative of parents of all children at age 5.5 years? Explain.
- The *USA TODAY* article also includes the following summary paragraph:

The study did not examine what the children watched and can't show TV was the cause of later problems, but it does "tell parents that even if kids are watching TV early in life, and they stop, it could reduce the risk for behavioral and social problems later," Mistry says.

What potentially confounding variable is identified in this passage?

- The passage in Part (b) says that the study cannot show that TV was the cause of later problems. Is the quote from Kamila Mistry (one of the study authors) in the passage consistent with the statement about cause? Explain.

- 2.67** An article titled "I Said, Not While You Study: Science Suggests Kids Can't Study and Groove at the Same Time" appeared in the *Washington Post* (September 5, 2006). This provides an example of a reporter summarizing the result of a scientific study in a way that is designed to make it accessible to the newspaper's readers. You can find the newspaper article online by searching on the title or by going to washingtonpost.com/wp-dyn/content/article/2006/09/03/AR2006090300592.html. The study referenced in the newspaper article was published in the *Proceedings of the National Academies of Science* and can be found at pnas.org/content/103/31/11778.full.

Read the newspaper article and then take a look at the published paper. Comment on whether you think that the author was successful in communicating the findings of the study to the intended audience.

- 2.68** The short article “*Developing Science-Based Food and Nutrition Information*” (*Journal of the American*

Dietetic Association [2001]: 1144–1145) includes some guidelines for evaluating a research paper. Obtain a copy of this paper through your university library or your instructor. Read this article and make a list of questions that can be used to evaluate a research study.

CHAPTER ACTIVITIES

ACTIVITY 2.1 FACEBOOK FRIENDING

Background: The article “*Professors Prefer Face Time to Facebook*” appeared in the student newspaper at Cal Poly, San Luis Obispo (*Mustang Daily, August 27, 2009*). The article examines how professors and students felt about using Facebook as a means of faculty-student communication. The student who wrote this article got mixed opinions when she interviewed students to ask whether they wanted to become Facebook friends with their professors. Two student comments included in the article were

“I think the younger the professor is, the more you can relate to them and the less awkward it would be if you were to become friends on Facebook. The older the professor, you just would have to wonder, ‘Why are they friending me?’”

and

“I think becoming friends with professors on Facebook is really awkward. I don’t want them being able to see into my personal life, and frankly, I am not really interested in what my professors do in their free time.”

Even if the students interviewed had expressed a consistent opinion, it would still be unreasonable to think this represented general student opinion on this issue because only four students were interviewed and it is not clear from the article how these students were selected.

In this activity, you will work with a partner to develop a plan to assess student opinion about being Facebook friends with professors at your school.

1. Suppose you will select a sample of 50 students at your school to participate in a survey. Write one or more questions that you would ask each student in the sample.
2. Discuss with your partner whether you think it would be easy or difficult to obtain a simple random sample of 50 students at your school and to obtain the desired information from all the students selected for the sample. Write a summary of your discussion.
3. With your partner, decide how you might go about selecting a sample of 50 students from your school that reasonably could be considered representative of the population of interest even if it may not be a simple random sample. Write a brief description of your sampling plan, and point out the aspects of your plan that you think make it reasonable to argue that it will be representative.
4. Explain your plan to another pair of students. Ask them to critique your plan. Write a brief summary of the comments you received. Now reverse roles, and provide a critique of the plan devised by the other pair.
5. Based on the feedback you received in Step 4, would you modify your original sampling plan? If not, explain why this is not necessary. If so, describe how the plan would be modified.

ACTIVITY 2.2 AN EXPERIMENT TO TEST FOR THE STROOP EFFECT

Background: In 1935, John Stroop published the results of his research into how people respond when presented with conflicting signals. Stroop noted that most people are able to read words quickly and that they cannot easily ignore them and focus on other attributes of a printed word, such as text color. For example, consider the following list of words:

green blue red blue yellow red

It is easy to quickly read this list of words. It is also easy to read the words even if the words are printed in

color, and even if the text color is different from the color of the word. For example, people can read the words in a list of words that are printed in different colors like the list shown here

green blue red blue yellow red

as quickly as they can read the list that isn’t printed in color.

However, Stroop found that if people are asked to name the text colors of the words in the list (red, yellow, blue, green, red, green), it takes them longer. Psychologists

believe that this is because the reader has to inhibit a natural response (reading the word) and produce a different response (naming the color of the text).

If Stroop is correct, people should be able to name colors more quickly if they do not have to inhibit the word response, as would be the case if they were shown colored rectangles like those shown below.



1. Design an experiment to compare times to identify colors when they appear as text to times to identify colors when there is no need to inhibit a word

response. Indicate how random assignment is incorporated into your design. What is your response variable? How will you measure it? How many subjects will you use in your experiment, and how will they be chosen?

2. When you are satisfied with your experimental design, carry out the experiment. You will need to construct your list of colored words and a corresponding list of colored bars to use in the experiment. You will also need to think about how you will implement the random assignment scheme.
3. Summarize the resulting data in a brief report that explains whether your findings are consistent with the Stroop effect.

ACTIVITY 2.3 McDONALD'S AND THE NEXT 100 BILLION BURGERS

Background: The article “Potential Effects of the Next 100 Billion Hamburgers Sold by McDonald’s” (*American Journal of Preventative Medicine* [2005]: 379–381) estimated that 992.25 million pounds of saturated fat would be consumed as McDonald’s sells its next 100 billion hamburgers. This estimate was based on the assumption that the average weight of a burger sold would be 2.4 ounces. This is the average of the weight of a regular hamburger (1.6 ounces) and a Big Mac (3.2 ounces). The authors took this approach because

McDonald’s does not publish sales and profits of individual items. Thus, it is not possible to estimate how many of McDonald’s first 100 billion beef burgers sold were 1.6 ounce hamburgers, 3.2 ounce Big Macs (introduced in 1968), 4.0 ounce Quarter Pounders (introduced in 1973), or other sandwiches.

This activity can be completed as an individual or as a team. Your instructor will specify which approach (individual or team) you should use.

1. The authors of the article believe that the use of 2.4 ounces as the average size of a burger sold at McDonald’s is “conservative,” which would result in the estimate of 992.25 million pounds of saturated fat being lower than the actual amount that would be consumed. Explain why the authors’ belief might be justified.
2. Do you think it would be possible to collect data that could lead to an estimate of the average burger size that would be better than 2.4 ounces? If so, explain how you would recommend collecting such data. If not, explain why you think it is not possible.

ACTIVITY 2.4 VIDEO GAMES AND PAIN MANAGEMENT

Background: Video games have been used for pain management by doctors and therapists who believe that the attention required to play a video game can distract the player and thereby decrease the sensation of pain. The paper “Video Games and Health” (*British Medical Journal* [2005]:122–123) states

However, there has been no long term follow-up and no robust randomized controlled trials of such interventions. Whether patients eventually tire of such games is also unclear. Furthermore, it is not known whether any distracting effect depends simply on concentrating on an interactive task or whether the content of games is also an important factor as there have been no controlled trials comparing video games with other distractors. Further research should examine factors within games such as novelty, users’ preferences, and relative levels of challenge and should compare video games with other potentially distracting activities.

1. Working with a partner, select one of the areas of potential research suggested in the passage from the paper and formulate a specific question that could be addressed by performing an experiment.
2. Propose an experiment that would provide data to address the question from Step 1. Be specific about how subjects might be selected, what the experimental conditions (treatments) would be, and what response would be measured.
3. At the end of Section 2.3 there are 10 questions that can be used to evaluate an experimental design. Answer these 10 questions for the design proposed in Step 2.
4. After evaluating your proposed design, are there any changes you would like to make to your design? Explain.

ACTIVITY 2.5 BE CAREFUL WITH RANDOM ASSIGNMENT!

When individuals climb to high altitudes, a condition known as acute mountain sickness (AMS) may occur. AMS is brought about by a combination of reduced air pressure and lower oxygen concentration that occurs at high altitudes. Two standard treatments for AMS are a medication, acetazolamide (which stimulates breathing and reduces mild symptoms) and the use of portable hyperbaric chambers.

With increasing numbers of younger inexperienced mountaineers, it is important to re-evaluate these treatments for the 12- to 14-year age group. An experimental plan under consideration is to study the first 18 youngsters diagnosed with AMS at a high altitude park ranger station whose parents consent to participation in the experiment. Equal numbers of each treatment are desired and the researchers are considering the following strategy for random assignment of treatments: Assign the treatments using a coin flip until one treatment has been assigned nine times; then assign the other treatment to the remaining subjects.

The table below presents data on the first 18 young climbers whose parents consented to participation in the experiment.

Order	Gender	Age
1	male	12.90
2	female	13.34
3	male	12.39
4	male	13.95
5	male	13.63
6	male	13.62
7	female	12.55
8	female	13.54
9	male	12.34
10	female	13.74
11	female	13.78
12	male	14.05
13	female	14.22
14	female	13.91
15	male	14.39
16	female	13.54
17	female	13.85
18	male	14.11

- Describe how you would implement a strategy equivalent to the one proposed by the researchers. Your plan should assign the treatments M (medicine) and H (hyperbaric chamber) to these climbers as they appear at the ranger station.
- Implement your strategy in Step (1), assigning treatments to climbers 1–18.
- Looking at which climbers were assigned to each of the two groups, do you feel that this method worked well? Why or why not?
- Calculate the proportion of females in the medicine group. How does this proportion compare to the proportion of females in the entire group of 18 subjects?
- Construct two dotplots—one of the ages of those assigned to the medicine treatment and one of the ages of those assigned to the hyperbaric chamber treatment. Are the age distributions for the two groups similar?
- Calculate the average age of those assigned to the medicine group. How does it compare to the average age for the other treatment group?
- Record the proportion of females in the medicine group, the average age of those assigned to the medicine group, and the average age of those assigned to the hyperbaric chamber group obtained by each student in your class.
- Using the values from Step (7), construct a dotplot of each of the following: the proportion of females in the medicine group, the average age of those assigned to the medicine group, and the average age of those assigned to the hyperbaric chamber group.
- Using the results of the previous steps, evaluate the success of this random assignment strategy. Write a short paragraph explaining to the researchers whether or not they should use the proposed strategy for random assignment and why.

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
Observational study	A study that observes characteristics of an existing population.	Stratified sampling	Dividing a population into subgroups (strata) and then taking a separate random sample from each stratum.
Simple random sample	A sample selected in a way that gives every different sample of size n an equal chance of being selected.	Cluster sampling	Dividing a population into subgroups (clusters) and forming a sample by randomly selecting clusters and including all individuals or objects in the selected clusters in the sample.

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
1 in k systematic sampling	Forming a sample from an ordered arrangement of a population by choosing a starting point at random from the first k individuals on the list and then selecting every k th individual thereafter.	Blocking	Using extraneous variables to create groups that are similar with respect to those variables and then assigning treatments at random within each block, thereby filtering out the effect of the blocking variables.
Confounding variable	A variable that is related both to group membership and to the response variable.	Random assignment	Assigning experimental units to treatments or treatments to trials at random.
Measurement or response bias	The tendency for samples to differ from the population because the method of observation tends to produce values that differ from the true value.	Replication	A strategy for ensuring that there is an adequate number of observations for each experimental treatment.
Selection bias	The tendency for samples to differ from the population because of systematic exclusion of some part of the population.	Placebo treatment	A treatment that resembles the other treatments in an experiment in all apparent ways but that has no active ingredients.
Nonresponse bias	The tendency for samples to differ from the population because measurements are not obtained from all individuals selected for inclusion in the sample.	Control group	A group that receives no treatment.
Experiment	A procedure for investigating the effect of <i>experimental conditions</i> (treatments) on a <i>response variable</i> .	Single-blind experiment	An experiment in which the subjects do not know which treatment they received but the individuals measuring the response do know which treatment was received, or an experiment in which the subjects do know which treatment they received but the individuals measuring the response do not know which treatment was received.
Treatments	The experimental conditions imposed by the experimenter.	Double-blind experiment	An experiment in which neither the subjects nor the individuals who measure the response know which treatment was received.
Extraneous variable	A variable that is not an explanatory variable in the study but is thought to affect the response variable.		
Direct control	Holding extraneous variables constant so that their effects are not confounded with those of the experimental conditions.		

CHAPTER REVIEW Exercises 2.69 - 2.78

2.69 A mortgage lender routinely places advertisements in a local newspaper. The advertisements are of three different types: one focusing on low interest rates, one featuring low fees for first-time buyers, and one appealing to people who may want to refinance their homes. The lender would like to determine which advertisement format is most successful in attracting customers to call for more information.

- Describe an experiment that would provide the information needed to make this determination. Be sure to consider extraneous variables, such as the day of the week that the advertisement appears in the paper, the section of the paper in which the advertisement appears, or daily fluctuations in the interest rate.
- What role does random assignment play in your design?

2.70 The article “Rethinking Calcium Supplements” (*US Airways Magazine, October 2010*) describes a study investigating whether taking calcium supplements increases the risk of heart attack. Consider the following four study descriptions. For each study, answer the following five questions:

- Question 1: Is the described study an observational study or an experiment?
- Question 2: Did the study use random selection from some population?
- Question 3: Did the study use random assignment to experimental groups?
- Question 4: Based on the study description, is it reasonable to conclude that taking calcium supplements is the cause of the increased risk of heart attack?

Question 5: Is it reasonable to generalize conclusions from this study to some larger population? If so, what population?

Study 1: Every heart attack patient and every patient admitted for an illness other than a heart attack during the month of December at a large urban hospital was asked if he or she took calcium supplements. The researchers found that the proportion of heart attack patients who took calcium supplements was significantly higher than the proportion of patients admitted for other illnesses who took calcium supplements.

Study 2: Two hundred people were randomly selected from a list of all people living in Minneapolis who receive Social Security. Each person in the sample was asked whether or not they took calcium supplements. These people were followed for 5 years, and whether or not they had a heart attack during the 5-year period was noted. The researchers found that the proportion of heart attack victims in the group taking calcium supplements was significantly higher than the proportion of heart attack victims in the group not taking calcium supplements.

Study 3: Two hundred people were randomly selected from a list of all people living in Minneapolis who receive Social Security. Each person was asked to participate in a statistical study, and all agreed to participate. Those who had no previous history of heart problems were instructed to take calcium supplements. Those with a previous history of heart problems were instructed not to take calcium supplements. The participants were followed for 5 years, and whether or not they had a heart attack during the 5-year period was noted. The researchers found that the proportion of heart attack victims in the calcium supplement group was significantly higher than the proportion of heart attack victims in the no supplement group.

Study 4: Four hundred people volunteered to participate in a 10-year study. Each volunteer was assigned at random to either group 1 or group 2. Those in group 1 took a daily calcium supplement. Those in group 2 did not take a calcium supplement. The proportion who suffered a heart attack during the 10-year study period was noted for each group. The researchers found that the proportion of heart attack victims in group 1 was significantly higher than the proportion of heart attack victims in group 2.

2.71 A pollster for the Public Policy Institute of California explains how the Institute selects a sample of California adults ([“It’s about Quality, Not Quantity,” San Luis Obispo Tribune, January 21, 2000](#)):

That is done by using computer-generated random residential telephone numbers with all California prefixes, and when there are no answers, calling back repeatedly to the original numbers selected to avoid a bias against hard-to-reach people. Once a call is completed, a second random selection is made by asking for the adult in the household who had the most recent birthday. It is as important to randomize who you speak to in the household as it is to randomize the household you select. If you didn’t, you’d primarily get women and older people.

Comment on this approach to selecting a sample. How does the sampling procedure attempt to minimize certain types of bias? Are there sources of bias that may still be a concern?

2.72 The article [“I’d Like to Buy a Vowel, Drivers Say” \(USA TODAY, August 7, 2001\)](#) speculates that young people prefer automobile names that consist of just numbers and/or letters that do not form a word (such as Hyundai’s XG300, Mazda’s 626, and BMW’s 325i). The article goes on to state that Hyundai had planned to identify the car that was eventually marketed as the XG300 with the name Concerto, until they determined that consumers hated it and that they thought XG300 sounded more “technical” and deserving of a higher price. Do the students at your school feel the same way? Describe how you would go about selecting a sample to answer this question.

2.73 A study in Florida is examining whether health literacy classes and using simple medical instructions that include pictures and avoid big words and technical terms can keep Medicaid patients healthier ([San Luis Obispo Tribune, October 16, 2002](#)). Twenty-seven community health centers are participating in the study. For 2 years, half of the centers will administer standard care. The other centers will have patients attend classes and will provide special health materials that are easy to understand.

Explain why it is important for the researchers to assign the 27 centers to the two groups (standard care and classes with simple health literature) at random.

2.74 The press release [“Men Need to Man Up, According to Ball Park Brand Survey” \(PR Newswire, October 14, 2015\)](#) describes the results of a study in which 1012 U.S. men were asked a number of questions about “life’s tough conversations.” One result from this survey was summarized in a [USA TODAY Snapshot \(USA TODAY, November 6, 2015\)](#) that said that “nearly 1 in 5 men would pay someone to handle their breakup for them.”

a. Is the study described an observational study or an experiment?

- b. Give at least one reason why the conclusion that “nearly 1 in 5 men would pay someone to handle their breakup for them” may not generalize to the population of all U.S. men.

2.75 A news release from Intel, “[Intel's Security International Internet of Things Smart Home Survey Shows Many Respondents Sharing Personal Data for Money](http://newsroom.intel.com/news-releases/intel-securitys-international-internet-of-things-smart-home-survey/)” (March 30, 2016, newsroom.intel.com/news-releases/intel-securitys-international-internet-of-things-smart-home-survey/, retrieved September 25, 2016), described a survey conducted in 2015. The news release states “A total of 9,000 consumers were interviewed globally, including 2,500 from the United States, 1,000 from the United Kingdom, 1,000 from France, 1,000 from Germany, 1,000 from Brazil, 1,000 from India, 500 from Canada, 500 from Mexico and 500 from Australia.” Among the findings from the survey were that 54% of the respondents worldwide would be willing to share personal data collected from devices in their homes with companies in exchange for money.

Do you think that the study described was an observational study or an experiment?

2.76 [USA TODAY](http://www.usatoday.com/story/money/2015/08/25/kate-middleton-wins-best-shopping-buddy-contest/31030003/) (August 25, 2015) reported that “American women favor Kate Middleton as a shopping buddy over Michelle Obama by 10 percentage points.” This statement was based on a study in which 1001 adults were surveyed about their shopping preferences.

Describe any potential sources of bias that might limit the researcher’s ability to draw conclusions about American women based on the data collected in this survey.

2.77 The paper “[Effect of a Nutritional Supplement on Hair Loss in Women](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3604033/)” (*Journal of Cosmetic*

Dermatology [2015]: 76–82) describes an experiment to see if a dietary supplement consisting of Omega 3, Omega 6, and antioxidants could reduce hair loss in women with stage 1 hair loss. One hundred twenty women volunteered to participate in the study and were randomly assigned to either the supplement group or a control group. The women in the supplement group took the supplement for 6 months. Photos of the top of the head were taken of all the women at the beginning of the study and 6 months later at the end of the study. The two photos of each woman were evaluated by an independent expert who visually determined the change in hair density. The expert who determined the change in hair density did not know which of the women had taken the supplement.

- a. Evaluate this experimental design. Do you think this is a good design or a poor design, and why?
 b. If you were designing such a study, what, if anything, would you propose to do differently?

2.78 A manufacturer of clay roofing tiles would like to investigate the effect of clay type on the proportion of tiles that crack in the kiln during firing. Two different types of clay are to be considered. One hundred tiles can be placed in the kiln at any one time. Firing temperature varies slightly at different locations in the kiln, and firing temperature may also affect cracking.

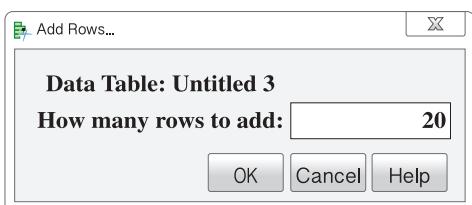
- a. Discuss the design of an experiment to collect information that could be used to decide between the two clay types.
 b. How does your proposed design deal with the extraneous variable temperature?

TECHNOLOGY NOTES

Generating random integers

JMP

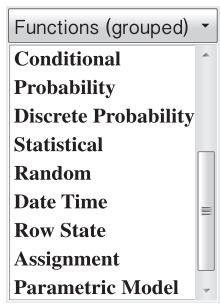
- Once JMP is initialized, select **File > New > DataTable** to create a new data table. A table will appear with “Column 1” in the top row. If you wish, double-click on “Column 1” and change this to “RandSample”.
- Click on **Rows > AddRows**. The following panel will appear:



- Decide how many random integers you wish, and type that number in the box. Then press **OK**. Note that JMP will not protect you against repeats, so you might wish to generate more numbers than you actually need. JMP will then present the column with the desired number of rows.

Rand Sample
1
2
3
4
5
6
7

4. Right-click on **RandSample** and select **Formula**.
5. Using the slider, find RANDOM in the Functions (grouped) panel.



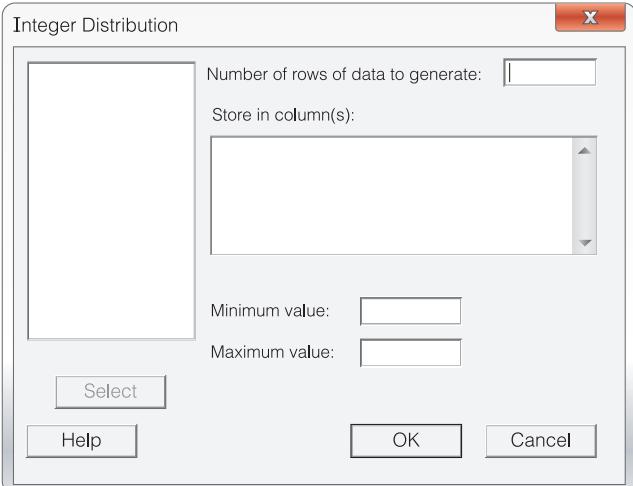
6. Click on **RANDOM**, then **Integer**. The “Formula” box will update to “RandomInteger(n1)”.
7. Replace n1 with “1<N”, where N is the maximum number desired. (If you want just 1 to 50, you can just substitute 50 for n1.) The random integers will appear in the RandSample column.

Minitab

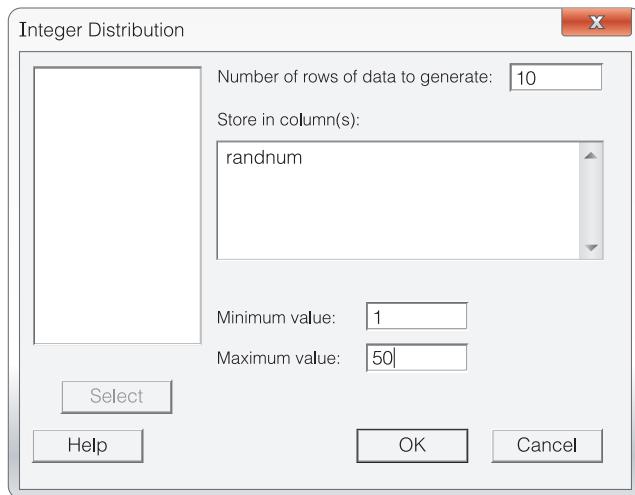
1. In the main Minitab Data Editor, at the top of the C1 variable column, enter “randnum” as the variable name.

	C1	C2
	randnum	
1		
2		
3		

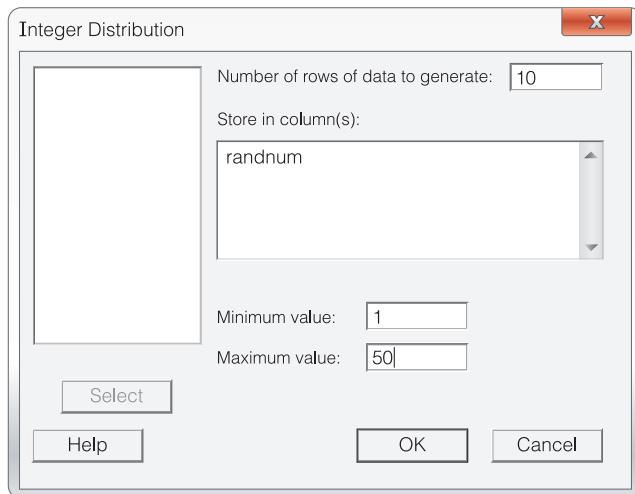
2. In succession, click on **Calc**, **RandomData**, and finally **Integer**. The Integer Distribution panel will appear.



3. Click on the **Store in column(s)** panel, and then double-click on **C1 randnum**.



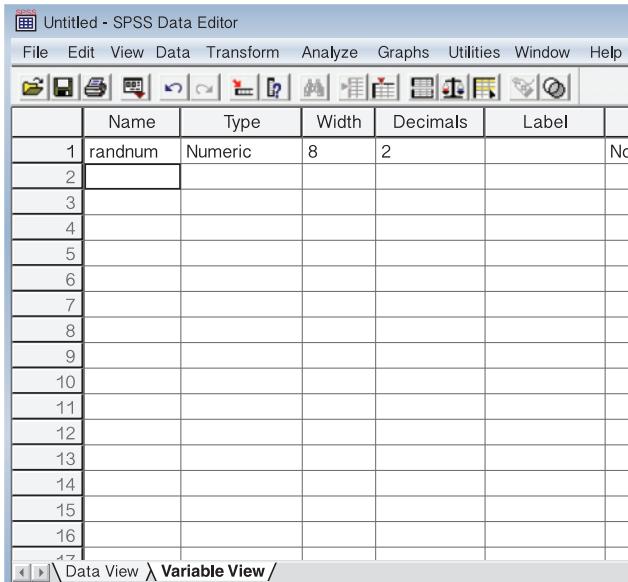
4. Enter the number of random integers desired and the minimum and maximum values.



5. Press OK. The random integers will appear in the randnum column.

SPSS

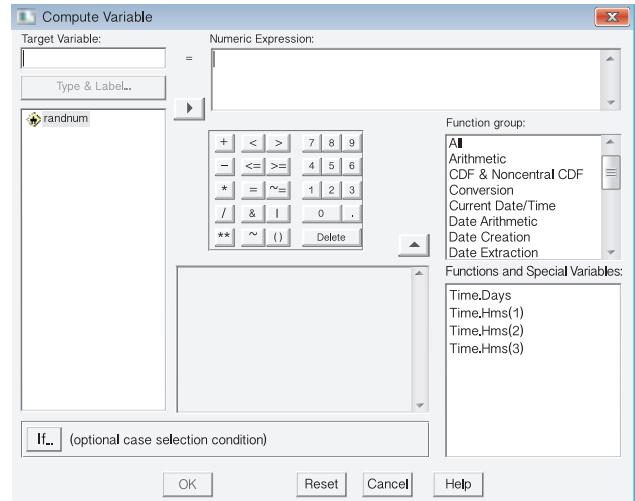
- In the SPSS Data Editor, click on the Variable View tab at the bottom of the screen. Then enter randnum in the name column and hit return.



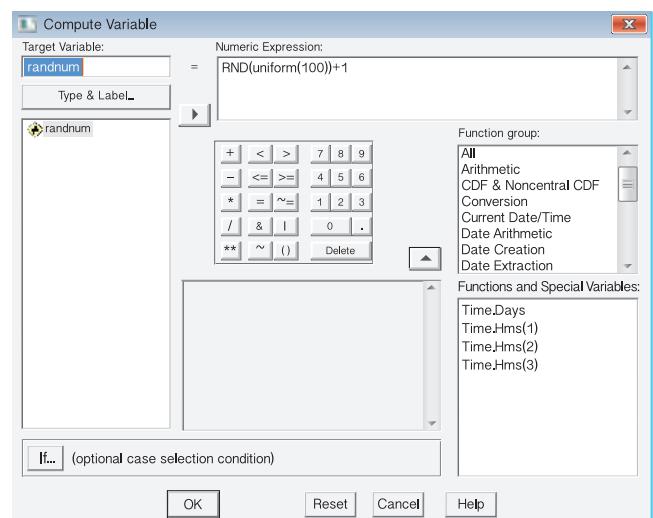
- Click on the Data View tab at the bottom of the screen. Then enter any number in the cell corresponding to how many random numbers you want to generate. For example, to generate 10 random numbers, type any number in cell 10 of the randnum column.

11 : randnum			
	randnum	var	var
1	-		
2	-		
3	-		
4	-		
5	-		
6	-		
7	-		
8	-		
9	-		
10	6.00		
11			
12			
13			
14			
15			
16			

- Choose Compute from the Transform menu.



- Type randnum in the Target Variable box. Type RND(uniform(N))+1 in the Numeric Expression box. N is a number that represents the largest integer you want to include. For example, to generate random integers between 1 and 100, you would type 100 for N.



- Click on OK and then click on OK in the box that asks if you want to change the existing variable. The requested random numbers will now appear in the randnum column of the Data Editor.

Excel 2007

- Choose a cell in Excel and enter “=RANDBETWEEN(Lo, Hi)” where Lo and Hi are the lower and upper ends of the range of values of random integers that you want. Press **ENTER**.

SUM	:	X ✓ fx	=RANDBETWEEN(1, 50)
A		B	C
1	=RANDBETWEEN(1, 50)		

2 RANDBETWEEN(bottom, top)

- Click on the small square in the lower right of the cell. Drag the square downward the number of rows equal to the number of random integers that you wish to have.

A1	:	X ✓ fx	=RANDBETWEEN(1, 50)
A		B	C
1	30		
2			
3			

- Check for any repetitions. If found, use the lower-right-corner technique to add more.

TI-83/84

Note: The TI calculator has a built-in sequence of pseudo-random numbers that has been determined by “seeding” the starting random number. Unless this number is changed, everyone in a group with a new calculator will get the same sequence of random numbers. Because of this, Step 0 is the “reseeding” step. Normally, this would only be done one time.

- At the home screen, enter an 8-digit integer. Press **STO > MATH > PROB**.

Choose “rand” and press **ENTER**.

To generate n random numbers, with no repeats, between 1 and N:

- Select **MATH > PROB > randIntNoRep**
- Supply the relevant information about the desired random integers. For example,

lower: 1

upper: 20

n: 7

Paste

Press **ENTER** after each of these lines.

- randIntNoRep(1, 50, 7)** will appear. Press **ENTER** and the random integers will be displayed.

If your calculator does not have the most recent operating system, you must supply the numbers indicated. The randInt function will not necessarily give nonrepeating integers.

- Select **MATH > PROB > randInt**
- Enter **1, 50, 7, “)”** and **ENTER**.
- randInt(1, 50, 7)** should appear on the screen. Press **ENTER** and your numbers will be displayed.

To store the random integers in a TI List, do not press **ENTER** after **randIntNoRep(1, 50, 7)** or **randInt(1, 50, 7)** in Step 3. Instead of Enter, continue with pressing **STO > 2nd > L1**. The random integers will be stored in List1.

3

Graphical Methods for Describing Data



istock.com/florintt

Most college students (and their parents) are concerned about the cost of a college education. [The National Center for Education Statistics \(nces.ed.gov\)](http://The%20National%20Center%20for%20Education%20Statistics%20(nces.ed.gov)) reported the average in state tuition and fees for 4-year public institutions in each of the 50 U.S. states for the 2015–2016 academic year. Average tuition and fees (in dollars) are given below for each state.

Several questions could be posed about these data. What is a typical value of average in-state tuition and fees for the 50 states? Are observations concentrated near the typical value, or do average tuition and fees differ quite a bit from state to state? Are there any states whose average tuition and fees are unusual compared to the rest? What proportion of the states have average tuition and fees less than \$7500? More than \$10,000?

Questions like these are most easily answered if the data can be organized in a sensible manner. In this chapter, we introduce some techniques for organizing and describing data using tables and graphs.

9,179	6,880	9,884	7,577	9,070	9,128	11,106	11,670	4,438	7,011
9,263	6,915	13,387	8,745	7,879	8,011	9,490	8,162	9,186	8,942
11,670	11,708	10,701	7,175	8,178	6,443	7,446	5,298	14,986	13,021
6,262	7,647	6,944	7,208	9,757	6,680	9,406	13,516	11,321	11,791
8,273	8,932	8,091	6,140	15,062	11,669	7,782	6,900	8,504	4,179

LEARNING OBJECTIVES

Students will understand:

- That selecting an appropriate graphical display depends on the data type (categorical or numerical) and whether or not the purpose of the display is to compare groups.
- How a graphical display of numerical data is described in terms of center, shape, and spread.
- How a scatterplot is used to investigate the relationship between two numerical variables.
- How a time series plot is used to investigate a trend over time.

Students will be able to:

- Construct and interpret graphical displays of categorical data: pie charts and segmented bar charts.

- Construct and interpret graphical displays of numerical data: stem-and-leaf displays, histograms, and relative frequency histograms.
- Construct and interpret graphical displays designed to compare groups: comparative bar charts and comparative stem-and-leaf displays.
- Construct and interpret a scatterplot of bivariate numerical data.
- Construct and interpret a time series plot.

SECTION 3.1 Displaying Categorical Data: Comparative Bar Charts and Pie Charts

Comparative Bar Charts

In Chapter 1 we saw that categorical data could be summarized in a frequency distribution and displayed graphically using a bar chart. Bar charts can also be used to give a visual comparison of two or more groups. This is accomplished by constructing two or more bar charts that use the same set of horizontal and vertical axes, as illustrated in Example 3.1.

Example 3.1 Is Education Worth the Cost?

Understand the context ➤

Do people with college degrees think that their education was worth the cost? This question was posed to 2548 adults with an associate degree and to 30,151 adults with a bachelor's degree. The data, from the Gallup report “[Two-Year Grads Satisfied with Cost of Degree](#)” ([gallup.com](#), April 11, 2016), are summarized in the accompanying table.

Consider the data ➤

Response	Frequency		Relative Frequency	
	Associate Degree Holders	Bachelor's Degree Holders	Associate Degree Holders	Bachelor's Degree Holders
Strongly disagree	178	1,508	0.07	0.05
Disagree	153	2,111	0.06	0.07
Neither agree nor disagree	408	4,221	0.16	0.14
Agree	637	8,743	0.25	0.29
Strongly agree	1,172	13,568	0.46	0.45
Total	2,548	30,151	1.00	1.00

Because the two sample sizes are very different, it is important to use relative frequencies rather than frequencies to construct the scale for the vertical axis in the comparative bar chart. The same set of steps that were used to construct a bar chart are used to construct a comparative bar chart, but in a comparative bar chart each category will have bars for all of the groups.

Interpret the results ➤

The comparative bar chart for these data is shown in Figure 3.1. Looking at the comparative bar chart, it is easy to compare the response distributions of the two groups. Notice that the response distributions are similar, indicating that both associate degree holders and bachelor's degree holders generally agreed that their education was worth the cost, with more than 70% in the Agree or Strongly agree categories.

To see why we use relative frequencies rather than frequencies to compare groups of different sizes, consider the *incorrect* comparative bar chart constructed using the frequencies rather than the relative frequencies (Figure 3.2). The incorrect comparative bar chart conveys a very different and misleading impression.

FIGURE 3.1
Comparative bar chart.

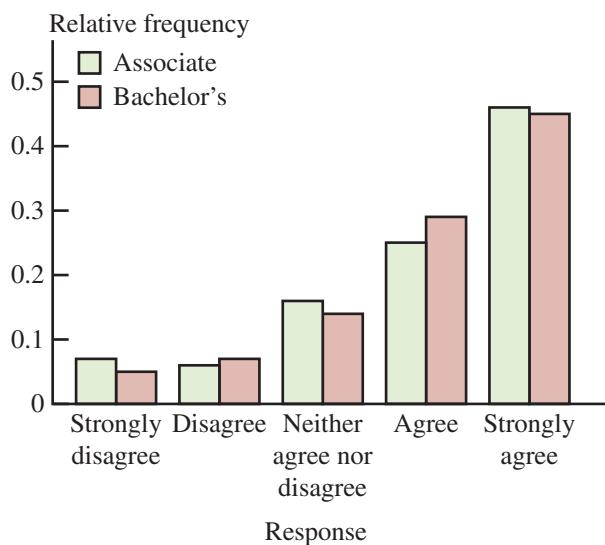
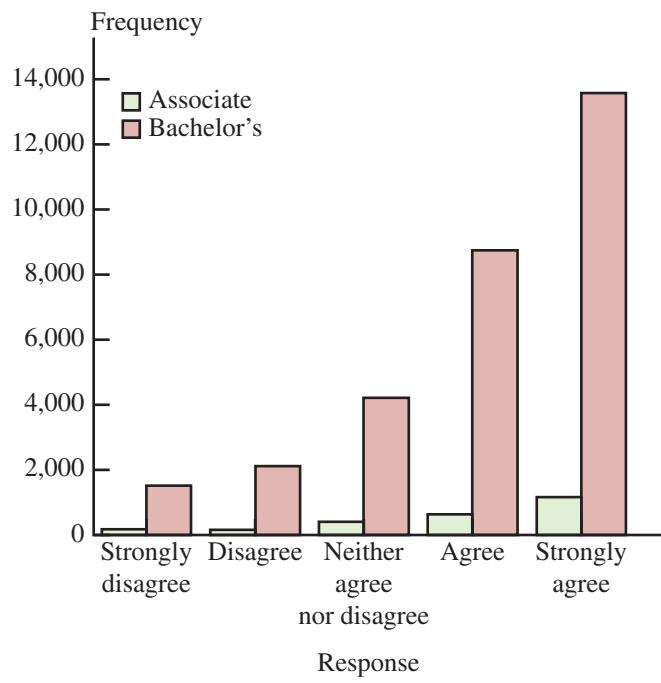


FIGURE 3.2
An **incorrect** comparative bar chart
for the data of Example 3.1.



When constructing a comparative bar chart, use relative frequency rather than frequency to construct the scale on the vertical axis so that meaningful comparisons can be made even if the sample sizes are not the same.

Pie Charts

A categorical data set can also be summarized using a pie chart. In a pie chart, a circle is used to represent the whole data set, with “slices” of the pie representing the possible categories. The size of the slice for a particular category is proportional to the corresponding frequency or relative frequency. Pie charts are most effective for summarizing data sets when there are not too many different categories.

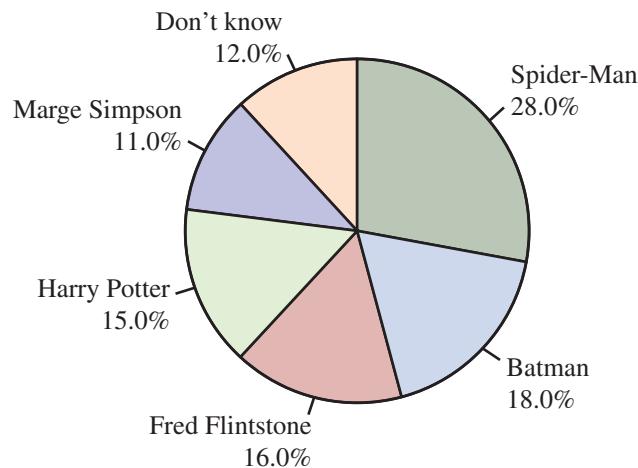
Example 3.2 Life Insurance for Cartoon Characters?

Understanding the context ➤

The article “[Fred Flintstone, Check Your Policy](#)” (*The Washington Post*, October 2, 2005) summarized the results of a survey of 1014 adults conducted by the Life and Health Insurance Foundation for Education. Each person surveyed was asked to select which of five fictional characters, Spider-Man, Batman, Fred Flintstone, Harry Potter, and Marge Simpson, he or she thought had the greatest need for life insurance. The resulting data are summarized in the pie chart of Figure 3.3.

FIGURE 3.3

Pie chart of data on which fictional character most needs life insurance.



Interpret the results ➤

The survey results were quite different from an insurance expert’s assessment. His opinion was that Fred Flintstone, a married father with a young child, was by far the one with the greatest need for life insurance. Spider-Man, unmarried with an elderly aunt, would need life insurance only if his aunt relied on him to supplement her income. Batman, a wealthy bachelor with no dependents, doesn’t need life insurance in spite of his dangerous job!

Pie Chart for Categorical Data

When to Use Categorical data with a relatively small number of possible categories. Pie charts are useful for illustrating proportions of the whole data set for various categories.

How to Construct

1. Draw a circle to represent the entire data set.
2. For each category, calculate the “slice” size. Because there are 360 degrees in a circle

$$\text{slice size} = 360 \cdot (\text{category relative frequency})$$
3. Draw a slice of appropriate size for each category. This can be tricky, so most pie charts are generated using a graphing calculator or a statistical software package.

What to Look For

- Categories that form particularly large and small proportions of the data set.

Example 3.3 Watch Those Typos

Understanding the context ➤

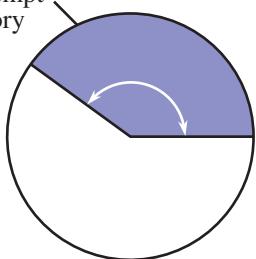
Typos on a résumé do not make a very good impression when applying for a job. Senior executives were asked how many typos in a résumé would make them not consider a job candidate ("Job Seekers Need a Keen Eye," *USA TODAY*, August 3, 2009). The resulting data are summarized in the accompanying relative frequency distribution.

Consider the data ➤

Number of Typos	Frequency	Relative Frequency
1	60	0.40
2	54	0.36
3	21	0.14
4 or more	10	0.07
Don't know	5	0.03

Do the work ➤

144 degrees,
to represent
first attempt
category



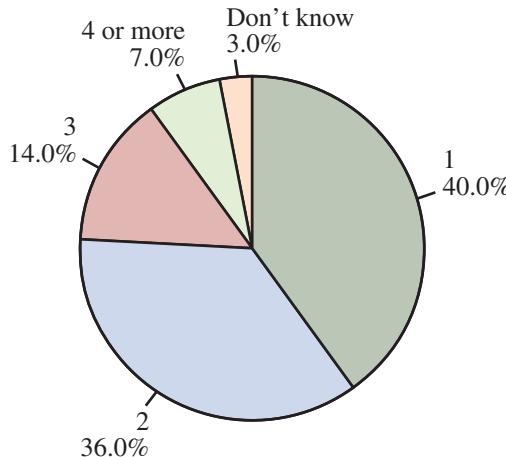
To draw a pie chart by hand, we would first compute the slice size for each category. For the one typo category, the slice size would be

$$\text{slice size} = (0.40)(360) = 144 \text{ degrees}$$

We would then draw a circle and use a protractor to mark off a slice corresponding to about 144° , as illustrated in the figure shown in the margin. Continuing to add slices in this way leads to a completed pie chart.

It is much easier to use a statistical software package to create pie charts than to construct them by hand. A pie chart for the typo data, created with the statistical software package Minitab, is shown in Figure 3.4.

FIGURE 3.4
Pie chart for the typo data of Example 3.3.



Interpret the results ➤

From the completed pie chart, it is easy to see that even one or two typos would result in many employers not considering a candidate for a job.

Pie charts can be used effectively to summarize a single categorical data set if there are not too many different categories. However, pie charts are not usually the best tool if the goal is to compare groups on the basis of a categorical variable. This is illustrated in Example 3.4.

Example 3.4 Scientists and Nonscientists Do Not See Eye-to-Eye

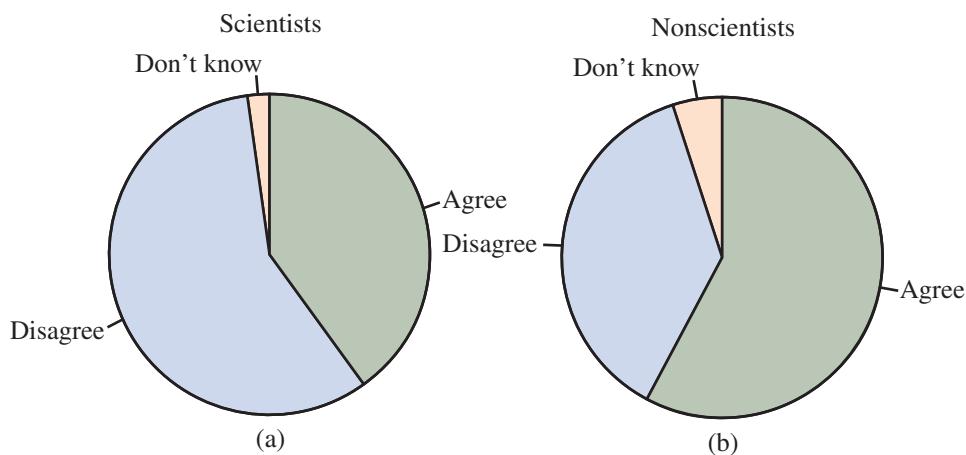
Understand the context ➤

Scientists and nonscientists were asked to indicate if they agreed or disagreed with the following statement: "When something is run by the government, it is usually inefficient"

and wasteful.” The resulting data (from “**Scientists, Public Differ in Outlooks**,” *USA TODAY*, July 10, 2009) were used to create the two pie charts in Figure 3.5.

FIGURE 3.5

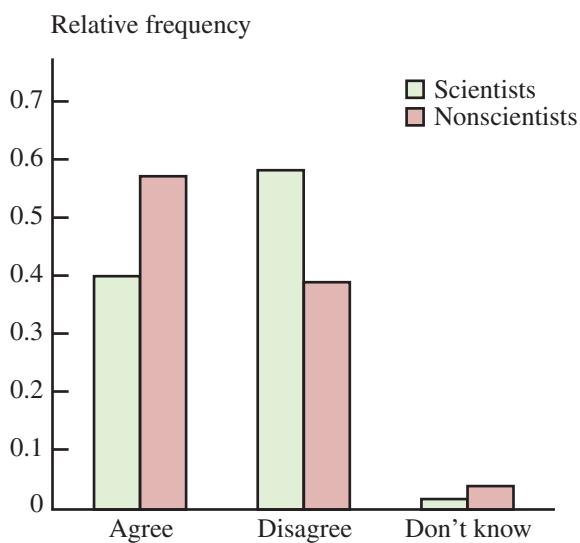
Pie charts for Example 3.4:
(a) scientist data;
(b) nonscientist data.



Although differences between scientists and nonscientists can be seen by comparing the pie charts of Figure 3.5, it can be difficult to compare category proportions using pie charts. A comparative bar chart (Figure 3.6) makes this type of comparison easier.

FIGURE 3.6

Comparative bar chart for the scientist and nonscientist data.



Interpret the results ➤

From the comparative bar chart, it is easy to see that the scientists were much less likely to agree with the statement than people who were not scientists.

A Different Type of “Pie” Chart: Segmented Bar Charts

A pie chart can be difficult to construct by hand, and the circular shape sometimes makes it difficult to compare areas for different categories, particularly when the relative frequencies for categories are similar. The **segmented bar chart** (also sometimes called a stacked bar chart) avoids these difficulties by using a rectangular bar rather than a circle to represent the entire data set. The bar is divided into segments, with different segments representing different categories. As with pie charts, the area of the segment for a particular category is proportional to the relative frequency for that category. Example 3.5 illustrates the construction of a segmented bar chart.

Example 3.5 How College Seniors Spend Their Time

Understand the context ➤

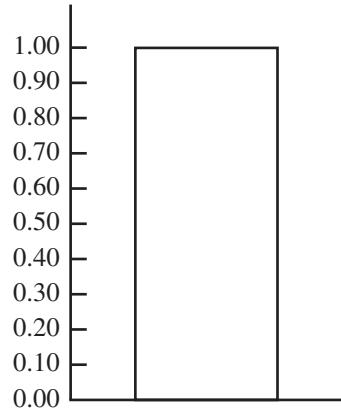
Each year, the Higher Education Research Institute conducts a survey of college seniors. In 2008, approximately 23,000 seniors participated in the survey ([“Findings from the 2008 Administration of the College Senior Survey,” Higher Education Research Institute, June 2009](#)). The accompanying relative frequency table summarizes student response to the question: “During the past year, how much time did you spend studying and doing homework in a typical week?”

Consider the data ➤

Studying/Homework	
Amount of Time	Relative Frequency
2 hours or less	0.074
3 to 5 hours	0.227
6 to 10 hours	0.285
11 to 15 hours	0.181
16 to 20 hours	0.122
Over 20 hours	0.111

Do the work ➤

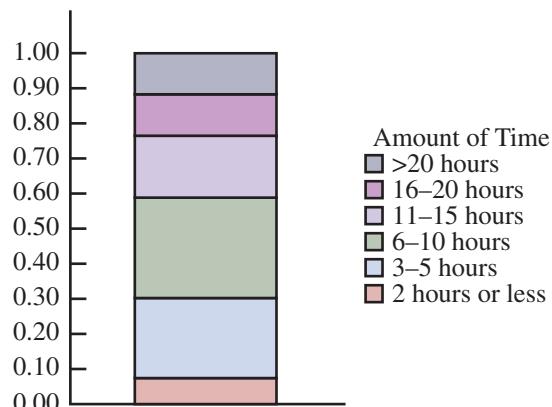
To construct a segmented bar chart for these data, first draw a bar of any fixed width and length, and then add a scale that ranges from 0 to 1, as shown.



Then divide the bar into six segments, corresponding to the six possible time categories in this example. The first segment, corresponding to the 2 hours or less category, ranges from 0 to 0.074. The second segment, corresponding to 3 to 5 hours, ranges from 0.074 to 0.301 (for a length of 0.227, the relative frequency for this category), and so on. The segmented bar chart is shown in Figure 3.7.

FIGURE 3.7

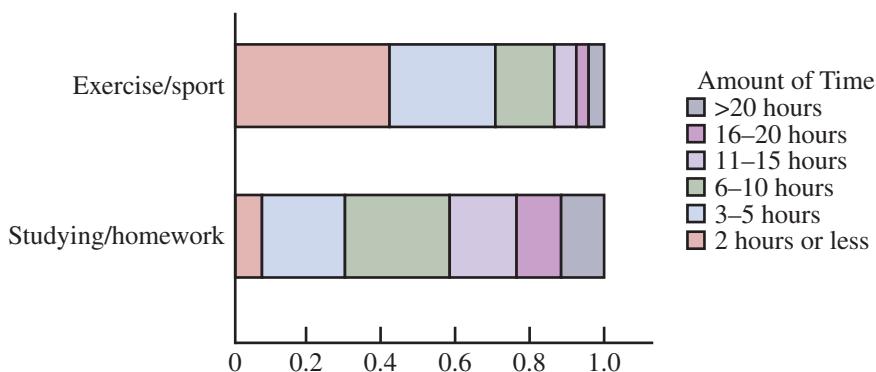
Segmented bar chart for the study time data of Example 3.5.



The same report also gave data on amount of time spent on exercise or sports in a typical week. Figure 3.8 shows horizontal segmented bar charts (segmented bar charts can be displayed either vertically or horizontally) for both time spent studying and time spent exercising. Viewing these graphs side by side makes it easy to see how students differ with respect to time spent on these two types of activities.

FIGURE 3.8

Segmented bar charts for time spent studying and time spent exercising.



Interpret the results ➤

The proportion of students who reported spending 5 hours or less per week exercising was much larger than the proportion who reported spending 5 hours or less per week studying. The proportion of students in each of the categories corresponding to 6 or more hours per week was noticeably higher for studying than for exercising.

Other Uses of Bar Charts and Pie Charts

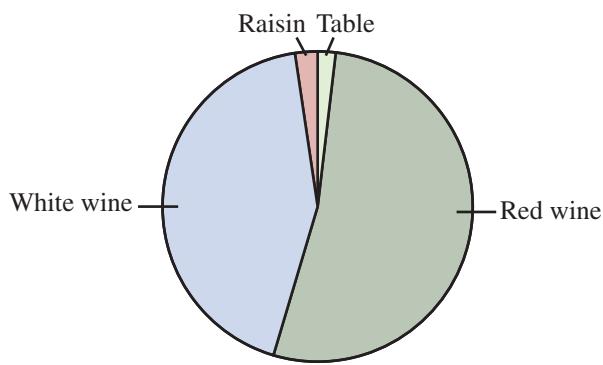
As we have seen in previous examples, bar charts and pie charts can be used to summarize categorical data sets. However, they are sometimes also used for other purposes, as illustrated in Examples 3.6 and 3.7.

Example 3.6 Grape Production

- The 2015 [Grape Crush Report for California](http://nass.usda.gov/Statistics_by_State/California/Publications/Grape_Crush/Prelim/) gave the following information on grape production for each of four different types of grapes (nass.usda.gov/Statistics_by_State/California/Publications/Grape_Crush/Prelim/, retrieved April 17, 2017):

**FIGURE 3.9**

Pie chart for grape production data.



Type of Grape	Tons Produced
Red Wine Grapes	2,037,000
White Wine Grapes	1,662,000
Raisin Grapes	92,000
Table Grapes	71,000
Total	3,862,000

Although this table is not a frequency distribution, it is common to represent information of this type graphically using a pie chart, as shown in Figure 3.9. The pie represents the total grape production, and the slices show the proportion of the total production for each of the four types of grapes.

Example 3.7 Back-to-College Spending

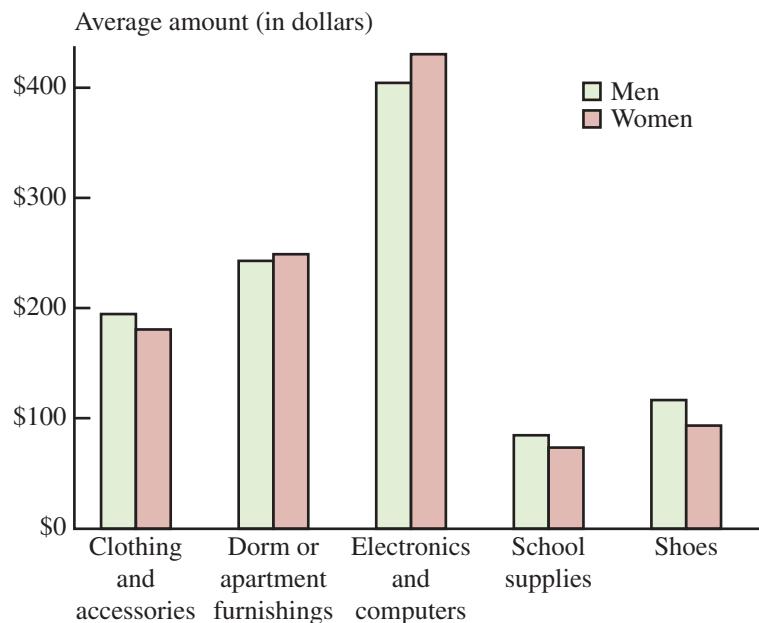
- The National Retail Federation's [2015 Back to College Survey \(nrf.com, retrieved August 1, 2016\)](http://nrf.com) asked each person in a sample of college students how much they planned to spend in various categories during the upcoming academic year. The average amounts of money (in dollars) that men and women planned to spend for five different types of purchases are shown in the accompanying table.

Type of Purchase	Average for Men	Average for Women
Clothing and accessories	\$195.03	\$180.62
Dorm or apartment furnishings	\$242.95	\$248.78
Electronics and computers	\$403.65	\$429.40
School supplies	\$83.45	\$73.03
Shoes	\$116.91	\$93.27

Even though this table is not a frequency distribution, this type of information is often represented graphically in the form of a comparative bar chart, as illustrated in Figure 3.10. From the bar chart, we can see that the average amount of money that men and women plan to spend is similar for all of the types of purchases except for electronics and computers, and shoes. The average for electronics and computers is a bit greater for women than for men and the average for shoes is a bit greater for men than for women.

FIGURE 3.10

Comparative bar chart for the back-to-college spending data of men and women.



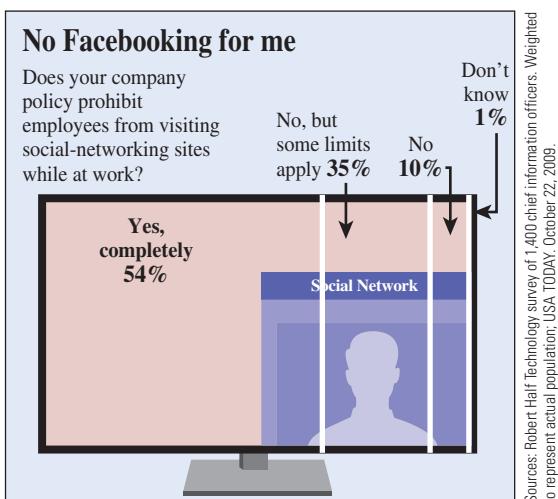
- Data set available online

EXERCISES 3.1 - 3.14

- Data set available online

- 3.1** Each person in a nationally representative sample of 1252 young adults age 23 to 28 years old was asked how they viewed their financial physique (financial health) (["2009 Young Adults & Money Survey Findings," Charles Schwab, 2009](#)). "Toned and fit" was chosen by 18% of the respondents, while 55% responded "a little bit flabby," and 27% responded "seriously out of shape." Summarize this information in a pie chart. (Hint: See Examples 3.2 and 3.3.)

- 3.2** The graphical display on the next page is similar to one that appeared in [USA TODAY \(October 22, 2009\)](#). It summarizes survey responses to a question about whether visiting social networking sites is allowed at work. Which of the graph types introduced in this section is used to display the responses? (USA TODAY frequently adds artwork and text to their graphs to try to make them look more interesting.)



- 3.3** The survey referenced in the previous exercise was conducted by Robert Half Technology. This company issued a press release ([“Whistle—But Don’t Tweet—While You Work,” roberthalftechnology.com, October 6, 2009](#)) that provided more detail than in the *USA TODAY* graph. The actual question asked was “Which of the following most closely describes your company’s policy on visiting social networking sites, such as Facebook, MySpace and Twitter, while at work?” The responses are summarized in the following table:

Response Category	Relative Frequency (expressed as percent)
Prohibited completely	54%
Permitted for business purposes only	19%
Permitted for limited personal use	16%
Permitted for any type of personal use	10%
Don’t know/no answer	1%

- a. Explain how the survey response categories and corresponding relative frequencies were used or modified to produce the graphical display in the previous exercise.
- b. Using the data in the table, construct a segmented bar chart. (Hint: See Example 3.5.)
- c. What are two other types of graphical displays that would be appropriate for summarizing these data?

- 3.4** The National Confectioners Association asked 1006 adults the following question: “Do you set aside a personal stash of Halloween candy?” Fifty-five percent of those surveyed responded no, 41% responded yes, and 4% either did not answer the question or said they did not know ([USA TODAY, October 22, 2009](#)). Use the given information to construct a pie chart.

- 3.5** College student attitudes about e-books were investigated in a survey of 1625 students. Students were asked to indicate their level of agreement with the following statement:

“I would like to be able to get all my textbooks in digital form.”

The responses are summarized in the accompanying table. ([The Chronicle of Higher Education, August 23, 2013](#))

Response	Percentage in This Category
Strongly disagree	21.3
Disagree	29.2
Agree	25.0
Strongly agree	13.2
Don’t know	11.3

- a. Construct an appropriate graphical display to summarize the information given in the table.
- b. Write a headline that would be appropriate for a newspaper article that summarized the results of this survey.

- 3.6** ● The Center for Science in the Public Interest evaluated school cafeterias in 20 school districts across the United States. Each district was assigned a numerical score on the basis of rigor of food codes, frequency of food safety inspections, access to inspection information, and the results of cafeteria inspections. Based on the score assigned, each district was also assigned one of four grades.

The scores and grades are summarized in the accompanying table, which appears in the report [“Making the Grade: An Analysis of Food Safety in School Cafeterias”](#) ([cspi.us/new/pdf/makingthegrade.pdf](#), 2007).

Jurisdiction	Overall Score (out of 100)
City of Fort Worth, TX	80
King County, WA	79
City of Houston, TX	78
Maricopa County, AZ	77
City and County of Denver, CO	75
Dekalb County, GA	73
Farmington Valley Health District, CT	72
State of Virginia	72
Fulton County, GA	68
City of Dallas, TX	67
City of Philadelphia, PA	67
City of Chicago, IL	65
City and County of San Francisco, CA	64
Montgomery County, MD	63

(continued)

Jurisdiction	Overall Score (out of 100)
Hillsborough County, FL	60
City of Minneapolis, MN	60
Dade County, FL	59
State of Rhode Island	54
District of Columbia	46
City of Hartford, CT	37

- a. Two variables are summarized in the figure, grade and overall score. Is overall score a numerical or categorical variable? Is grade (indicated by the different colors in the figure) a numerical or categorical variable?
- b. Explain how the figure is equivalent to a segmented bar chart of the grade data.
- 3.7 Using the data given in the previous exercise, construct a dotplot of the overall score data. Based on the dotplot, suggest an alternate assignment of grades (top of class, passing, etc.) to the 20 school districts. Explain the reasoning you used to make your assignment. (Hint: Dotplots were covered in Section 1.4.)
- 3.8 ● The article “**Housework around the World**” (*USA TODAY*, September 15, 2009) included the percentage of women who say their spouses never help with household chores for five different countries. Display the information in the accompanying table in a bar chart.
- | Country | Percentage |
|----------------|------------|
| Japan | 74 |
| France | 44 |
| United Kingdom | 40 |
| United States | 34 |
| Canada | 31 |
- 3.9 The article referenced in the previous exercise did not state how the author arrived at the given percentages.
- a. What are two questions that you would want to ask the author about how the data used to calculate the percentages were collected?
- b. Assuming that the data that were used to calculate these percentages were collected in a reasonable way, write a few sentences describing how the five countries differ in terms of spouses helping their wives with housework.
- 3.10 ● The authors of the report “**Findings from the 2009 Administration of the College Senior Survey**” (*Higher Education Research Institute, 2010*) asked a large number of college seniors how they would rate themselves compared to the average person of their age with respect to physical health. The

accompanying relative frequency table summarizes the responses for men and women.

Rating of Physical Health	Relative Frequency	
	Men	Women
Highest 10%	0.359	0.227
Above average	0.471	0.536
Average	0.160	0.226
Below average	0.009	0.007
Lowest 10%	0.001	0.004

- a. Construct a comparative bar chart of the responses that allows you to compare the responses of men and women.
- b. There were 8110 men and 15,260 women who responded to the survey. Explain why it is important that the comparative bar chart be constructed using the relative frequencies rather than the actual numbers of people (the frequencies) responding in each category.
- c. Write a few sentences commenting on how college seniors perceive themselves with respect to physical health and how men and women differ in their perceptions.

- 3.11 The survey on student attitude toward e-books described in Exercise 3.5 was conducted in 2011. A similar survey was also conducted in 2012 (*The Chronicle of Higher Education, August 23, 2013*). Data from 1588 students who participated in the 2012 survey are summarized in the accompanying table.

Response	Percentage in This Category
Strongly disagree	19.1
Disagree	27.5
Agree	26.3
Strongly agree	16.0
Don't know	11.1

- a. Use these data and the data from Exercise 3.5 to construct a comparative bar chart that shows the distribution of responses for the two years. (Hint: See Example 3.1.)
- b. Based on your graph from part (a), do you think there was much of a change in attitude toward e-books from 2011 to 2012?
- 3.12 During 2017, Gallup conducted a survey of adult Americans and asked the following question: What was the main reason you decided to enroll in the school or college where you completed your highest level of education? (“**Why Higher Ed?**,” *Gallup, Inc., January 2018*). The responses are summarized in the accompanying table.

Location	28%
Access/Affordability	22%
School reputation and fit	20%
Good job or career	19%
Learning and knowledge	5%
Family or social expectations	4%
Other/No response	2%

- a. Construct a pie chart to summarize these data.
- b. Construct a bar chart to summarize these data.
- c. Which of these charts—a pie chart or a bar chart—best summarizes the important information? Explain.

3.13 An article about college loans (“[New Rules Would Protect Students](#),” *USA TODAY*, June 16, 2010) reported the percentage of students who had defaulted on a student loan within 3 years of when they were scheduled to begin repayment. Information was given for public colleges, private non-profit colleges, and for-profit colleges.

Loan Status	Relative Frequency		
	Public Colleges	Private Non-profit Colleges	For-Profit Colleges
Good Standing	0.928	0.953	0.833
In Default	0.072	0.047	0.167

- a. Construct a comparative bar chart that would allow you to compare loan status for the three types of colleges.

- b. The article states “those who attended for-profit schools were more likely to default than those who attended public or private non-profit schools.” What aspect of the comparative bar chart supports this statement?

3.14 The report “[Findings From the 2014 College Senior Survey](#)” (Higher Education Research Institute, December 2014) summarizes data collected from more than 13,000 college seniors across the United States. One question in the survey asked students to rate themselves based on their critical thinking skills. For engineering majors, 60.4% rated critical thinking as “a major strength” while 39.6% did not see critical thinking as a major strength. Data were also provided for humanities majors, social science majors, biological sciences majors, and business majors, and these data are summarized in the accompanying table.

Response	Relative Frequency				
	Engineering Majors	Social Humanities Majors	Biological Science Majors	Science Majors	Business Majors
A major strength	0.604	0.513	0.470	0.461	0.391
Not a major strength	0.396	0.487	0.530	0.539	0.609

Construct a comparative bar chart and compare the responses over the five different majors.

SECTION 3.2 Displaying Numerical Data: Stem-and-Leaf Displays

A stem-and-leaf display is an effective and compact way to summarize numerical data. Each number in the data set is broken into two pieces, referred to as a stem and a leaf. The **stem** is the first part of the number and consists of the beginning digit(s). The **leaf** is the last part of the number and consists of the final digit(s). For example, the number 213 might be split into a stem of 2 and a leaf of 13 or a stem of 21 and a leaf of 3. The resulting stems and leaves are then used to construct the display.

Example 3.8 Going Wireless

Understand the context ➤

- The U.S. Department of Health and Human Services reported the estimated percentage of U.S. households with only wireless phone service (no land line) for the 50 states and the District of Columbia ([cdc.gov/nchs/data/nhis/earlyrelease/wireless_state_201602.pdf](https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless_state_201602.pdf), retrieved February 16, 2018). Data for the 18 Eastern states and the District of Columbia (DC) are given here.

Consider the data ➤

State	Wireless %	State	Wireless %	State	Wireless %
CT	26.7	FL	45.9	MD	36.2
DE	29.4	GA	45.9	MA	31.5
DC	49.7	ME	40.8	NH	31.2

● Data set available online

(continued)

State	Wireless %	State	Wireless %	State	Wireless %
NJ	25.1	PA	30.0	VT	37.2
NY	31.1	RI	34.6	WV	37.2
NC	42.9	SC	49.5		
OH	45.8	VA	41.1		

Figure 3.11 shows a stem-and-leaf display for the wireless percentage data.

FIGURE 3.11

Stem-and-leaf display of wireless percentage for Eastern states.

2	6.7, 9.4, 5.1
3	6.2, 1.5, 1.2, 1.1, 0.0, 4.6, 7.2, 7.2
4	9.7, 7.6, 5.9, 0.8, 2.9, 5.8, 9.5, 1.1

Stem: Tens

Leaves: Ones

The numbers in the vertical column on the left of the display are the **stems**. Each number to the right of the vertical line is a **leaf** corresponding to one of the observations in the data set. The legend

Stem: Tens

Leaf: Ones

tells us that the observation that had a stem of 2 and a leaf of 6.7 corresponds to a wireless percentage of 26.7 (as opposed to 2.67). Similarly, the observation with the stem of 3 and leaf of 6.2 corresponds to a wireless percentage of 36.2.

Interpret the results ➤

This display shows that for many Eastern states, the percentage of households with only wireless phone service was in the 30% to 50% range. Only three states had percentages that were less than 30%.

The leaves on each line of the display in Figure 3.11 have not been arranged in order from smallest to largest. Most statistical software packages order the leaves this way, but it is not necessary to do so to get an informative display that still shows many of the important characteristics of the data set, such as shape and variability.

Stem-and-leaf displays can be useful to get a sense of a typical value for the data set, as well as a sense of how variable the values in the data set are. It is also easy to spot data values that are unusually far from the rest of the values in the data set. Such values are called outliers. The stem-and-leaf display of the wireless percentage data (Figure 3.11) does not show any outliers.

DEFINITION

Outlier: An unusually small or large data value. A precise rule for deciding when an observation is an outlier is given in Chapter 4.

Stem-and-Leaf Displays

When to Use Numerical data sets with a small to moderate number of observations (does not work well for very large data sets)

How to Construct

1. Select one or more leading digits for the stem values. The trailing digits (or sometimes just the first one of the trailing digits) become the leaves.
2. List possible stem values in a vertical column.
3. Record the leaf for every observation beside the corresponding stem value.
4. Indicate the units for stems and leaves someplace in the display.

(continued)

What to Look For The display conveys information about

- a representative or typical value in the data set
- the extent of variability about a typical value
- the presence of gaps and outliers in the data
- the extent of symmetry in the distribution of values
- the number and location of peaks

Example 3.9 Tuition at Public Universities

Understand the context ➤

- The introduction to this chapter gave data on average in-state tuition and fees at public institutions in the year 2016 for the 50 U.S. states. The observations ranged from a low value of \$4178 to a high value of \$15,062. The data are reproduced here:

Consider the data ➤

9,179	6,880	9,884	7,577	9,070	9,128	11,106	11,670	4,438	7,011
9,263	6,915	13,387	8,745	7,879	8,011	9,490	8,162	9,186	8,942
11,670	11,708	10,701	7,175	8,178	6,443	7,446	5,298	14,986	13,021
6,262	7,647	6,944	7,208	9,757	6,680	9,406	13,516	11,321	11,791
8,273	8,932	8,091	6,140	15,062	11,669	7,782	6,900	8,504	4,179

Do the work ➤

- Notice that some of the data values are only four digits (such as 7179) and others are five digits (such as 11,670). It is easiest to proceed if we first make all of the data values five digit numbers by adding leading zeros as shown:

09,179	06,880	09,884	07,577	09,070	09,128	11,106	11,670	04,438	07,011
09,263	06,915	13,387	08,745	07,879	08,011	09,490	08,162	09,186	08,942
11,670	11,708	10,701	07,175	08,178	06,443	07,446	05,298	14,986	13,021
06,262	07,647	06,944	07,208	09,757	06,680	09,406	13,516	11,321	11,791
08,273	08,932	08,091	06,140	15,062	11,669	07,782	06,900	08,504	04,179

A natural choice for the stem is the leading two digits. This would result in a display with 12 stems (04, 05, 06, 07, 08, 09, 10, 11, 12, 13, 14, and 15). Using the first three digits of a number of the stem would result in 110 stems (041 to 150). A stem-and-leaf display with 110 stems would not be an effective summary of the data. In general, stem-and-leaf displays that use between 5 and 30 stems tend to work well.

If we choose the first two digits as the stem, the remaining three digits (the hundreds, tens, and ones) would form the leaf. For example, for the first few values in the data set, we would have

$$\begin{aligned} 09,179 &\rightarrow \text{stem} = 09, \text{leaf} = 179 \\ 06,880 &\rightarrow \text{stem} = 06, \text{leaf} = 880 \\ 09,884 &\rightarrow \text{stem} = 09, \text{leaf} = 884 \end{aligned}$$

Interpret the results ➤

The leaves have been entered in the display of Figure 3.12 in the order they are encountered in the data set. Commas are used to separate the leaves only when each leaf has two or more digits. Figure 3.12 shows that most states had average in-state tuition and fees in the \$6000 to \$10,000 range and that the typical average tuition and fees is around \$9000. Five states have average in-state tuition and fees at public four-year institutions that are quite a bit higher than most other states (the five states with the highest values were Vermont, New Hampshire, Pennsylvania, Illinois, and New Jersey).

FIGURE 3.12

Stem-and-leaf display of average tuition and fees.

04	438, 179
05	298
06	880, 915, 443, 262, 944, 680, 140, 900
07	577, 011, 879, 175, 446, 647, 208, 782
08	745, 011, 162, 942, 178, 273, 932, 091, 504
09	179, 884, 070, 128, 263, 490, 186, 757, 406
10	701
11	106, 670, 670, 708, 321, 791, 669
12	
13	387, 021, 516
14	986
	Stems: Thousands
15	062
	Leaves: Ones

An alternative display (Figure 3.13) results from dropping all but the first digit of the leaf. This is what most statistical computer packages do when generating a display. Little information about typical value, variability, or shape is lost in this truncation and the display is simpler and more compact.

FIGURE 3.13

Stem-and-leaf display of average in-state tuition and fees using truncated leaves.

04	41
05	2
06	89429619
07	50814627
08	701912905
09	180124174
10	7
11	1667376
12	
13	305
14	9
	Stems: Thousands
15	0
	Leaves: Ones

Repeated Stems to Stretch a Display

Sometimes a natural choice of stems gives a display in which too many observations are concentrated on just a few stems. A more informative picture can be obtained by dividing the leaves at any given stem into two groups: those that begin with 0, 1, 2, 3, or 4 (the “low” leaves) and those that begin with 5, 6, 7, 8, or 9 (the “high” leaves). Then each stem value is listed twice when constructing the display, once for the low leaves and once again for the high leaves. It is also possible to repeat a stem more than twice. For example, each stem might be repeated five times, once for each of the leaf groupings {0, 1}, {2, 3}, {4, 5}, {6, 7}, and {8, 9}.

Example 3.10 Median Ages in 2030

Understand the context ➤

- The accompanying data on the Census Bureau’s projected median age in 2030 for the 50 U.S. states and Washington, D.C. appeared in the article [“2030 Forecast: Mostly Gray” \(USA TODAY, April 21, 2005\)](#). The median age for a state is the age that divides the state’s residents so that half are younger than the median age and half are older than the median age.

- Data set available online

Projected Median Age														
Consider the data ➤	41.0 32.9 39.3 29.3 37.4 35.6 41.1 43.6 33.7 45.4 35.6 38.7 39.2 37.8 37.7 42.0 39.1 40.0 38.8 46.9 37.5 40.2 40.2 39.0 41.1 39.6 46.0 38.4 39.4 42.1 40.8 44.8 39.9 36.8 43.2 40.2 37.9 39.1 42.1 40.7 41.3 41.5 38.3 34.6 30.4 43.9 37.8 38.5 46.7 41.6 46.4													

The ages in the data set range from 29.3 to 46.9. Using the first two digits of each data value for the stem results in a large number of stems, while using only the first digit results in a stem-and-leaf display with only three stems.

Do the work ➤ The stem-and-leaf display using single digit stems and leaves truncated to a single digit is shown in Figure 3.14. A stem-and-leaf display that uses repeated stems is shown in Figure 3.15. Here each stem is listed twice, once for the low leaves (those beginning with 0, 1, 2, 3, 4) and once for the high leaves (those beginning with 5, 6, 7, 8, 9). This display is more informative than the one in Figure 3.14, and it is still much more compact than a display based on two-digit stems.

FIGURE 3.14

Stem-and-leaf display for the projected median age data.

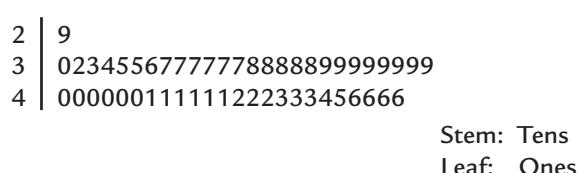
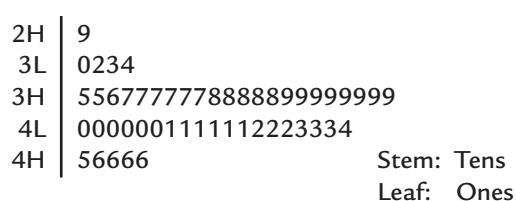


FIGURE 3.15

Stem-and-leaf display for the projected median age data using repeated stems.



Comparative Stem-and-Leaf Displays

Sometimes we want to see whether two data sets differ in some fundamental way. A comparative stem-and-leaf display, in which the leaves for one data set are listed to the right of the stem values and the leaves for the second data set are listed to the left, can provide a basis for making comparisons.

Example 3.11 Progress for Children

Understand the context ➤

- The report “[Progress for Children](#)” ([UNICEF, April 2005](#)) included the accompanying data on the percentage of primary-school-age children who were enrolled in school for 19 countries in Northern Africa and for 23 countries in Central Africa.

Consider the data ➤

Northern Africa

54.6	34.3	48.9	77.8	59.6	88.5	97.4	92.5	83.9	96.9	88.9
98.8	91.6	97.8	96.1	92.2	94.9	98.6	86.6			

Central Africa

58.3	34.6	35.5	45.4	38.6	63.8	53.9	61.9	69.9	43.0	85.0
63.4	58.4	61.9	40.9	73.9	34.8	74.4	97.4	61.0	66.7	79.6
										98.9

Formulate a plan ➤

We will construct a comparative stem-and-leaf display using the first digit of each observation as the stem and the remaining two digits as the leaf. To keep the display simple the leaves will be truncated to one digit. For example, the observation 54.6 would be processed as

$$54.6 \rightarrow \text{stem} = 5, \text{leaf} = 4 \text{ (truncated from 4.6)}$$

● Data set available online

and the observation 34.3 would be processed as

$$34.3 \rightarrow \text{stem} = 3, \text{leaf} = 4 \text{ (truncated from 4.3)}$$

The resulting comparative stem-and-leaf display is shown in Figure 3.16.

FIGURE 3.16

Comparative stem-and-leaf display for percentage of children enrolled in primary school.

Central Africa	Northern Africa	
4854	3	4
035	4	8
838	5	49
6113913	6	
943	7	76
5	8	8386
87	9	7268176248

Stem: Tens
Leaf: Ones

Interpret the results ➤

From the comparative stem-and-leaf display we can see that there is quite a bit of variability in the percentage enrolled in school for both Northern and Central African countries and that the shapes of the two data distributions are quite different. The percentage enrolled in school tends to be higher in Northern African countries than in Central African countries, although the smallest value in each of the two data sets is about the same. For Northern African countries the distribution of values has a single peak in the 90s with the number of observations declining as we move toward the stems corresponding to lower percentages enrolled in school. For Central African countries the distribution is more symmetric, with a typical value in the mid 60s.

EXERCISES 3.15 - 3.23

● Data set available online

- 3.15** ● The National Center for Health Statistics provided the data in the accompanying table in the report “National Vital Statistics Report” (January 5, 2017, cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_01.pdf, retrieved February 17, 2018). Entries in the table are the birth rates (births per 1000 of population) for the year 2015.

State	*Births	State	*Births
Alabama	12.3	Kansas	13.4
Alaska	15.3	Kentucky	12.6
Arizona	12.5	Louisiana	13.9
Arkansas	13.1	Maine	9.5
California	12.6	Maryland	12.3
Colorado	12.2	Massachusetts	10.5
Connecticut	10.0	Michigan	11.4
Delaware	11.8	Minnesota	12.7
District of Columbia	14.2	Mississippi	12.8
Florida	11.1	Missouri	12.3
Georgia	12.9	Montana	12.2
Hawaii	12.9	Nebraska	14.1
Idaho	13.8	Nevada	12.6
Illinois	12.3	New Hampshire	9.3
Indiana	12.7	New Jersey	11.5
Iowa	12.6	New Mexico	12.4

State	*Births	State	*Births
North Carolina	12.0	Tennessee	12.4
North Dakota	14.9	Texas	14.7
Ohio	12.0	Utah	16.9
Oklahoma	13.6	Vermont	9.4
Oregon	11.8	Virginia	12.3
Pennsylvania	11.0	Washington	12.4
Rhode Island	10.4	West Virginia	10.7
South Carolina	11.9	Wisconsin	11.6
South Dakota	14.4	Wyoming	13.2

*Births per 1000 of population

Construct a stem-and-leaf display using stems 9, 10, 11 . . . , 16. Comment on the interesting features of the display. (Hint: See Example 3.9.)

- 3.16** ● The paper “State-Level Cancer Mortality Attributable to Cigarette Smoking in the United States” (*JAMA Internal Medicine* [2016]: 1792–1798) included the following state estimates of the total number cancer deaths attributable to cigarette smoking in 2014.

(continued)

State	Number of People	State	Number of People
Alabama	3,183	Montana	581
Alaska	296	Nebraska	927
Arizona	3,246	Nevada	1,535
Arkansas	2,175	New Hampshire	723
California	14,689	New Jersey	4,388
Colorado	1,876	New Mexico	964
Connecticut	1,774	New York	9,296
Delaware	591	North Carolina	5,844
District of Columbia	310	North Dakota	341
Florida	12,596	Ohio	7,598
Georgia	4,816	Oklahoma	2,441
Hawaii	642	Oregon	2,143
Idaho	731	Pennsylvania	7,931
Illinois	7,114	Rhode Island	631
Indiana	4,099	South Carolina	2,962
Iowa	1,793	South Dakota	476
Kansas	1,587	Tennessee	4,613
Kentucky	3,452	Texas	10,310
Louisiana	3,044	Utah	495
Maine	927	Vermont	382
Maryland	2,900	Virginia	4,110
Massachusetts	3,565	Washington	3,298
Michigan	6,232	West Virginia	1,581
Minnesota	2,552	Wisconsin	3,081
Mississippi	1,992	Wyoming	251

- a.** Construct a stem-and-leaf display using thousands as the stems and truncating the leaves to the tens digit.
- b.** Write a few sentences describing the shape of the distribution and any unusual observations.
- c.** The four largest values were for California, Texas, Florida, and New York. Does this indicate that cancer deaths due to cigarette smoking is more of a problem in these states than elsewhere? Explain.
- d.** If you wanted to compare states on the basis of the cancer deaths due to cigarette smoking, would you use the data in the given table? If yes, explain why this would be reasonable. If no, what would you use instead as the basis for the comparison?
- 3.17** The accompanying data on seat belt use for each of the 50 U.S. states and the District of Columbia are from [“Traffic Safety Facts,” \(National Highway Traffic Safety Administration, June 2015\)](#). The observations represent the percentage of drivers wearing seat belts in a large nationwide observational survey.

95.7 88.4 87.2 74.4 97.1 82.4 85.1 91.9 93.2 88.8 97.3 93.5
 80.2 94.1 90.2 92.8 85.7 86.1 84.1 85.0 92.1 76.6 93.3 94.7
 78.3 78.8 74.0 70.0 94.0 70.4 87.6 92.1 90.6 90.6 81.0 85.0
 86.3 97.8 83.6 87.4 90.0 68.9 87.7 90.7 83.4 84.1 77.3 94.5
 87.8 84.7 79.2

- a.** The values in the data set range from 68.9% to 97.8%. Construct a stem-and-leaf display that uses repeated stems 6H, 7L, 7H, ... 9H.
 (Hint: See Example 3.10.)
- b.** Write a few sentences commenting on what the stem-and-leaf display suggests about seat belt use.

- 3.18** The previous exercise gave data on seat belt use for each of the 50 U.S. states and the District of Columbia ([“Traffic Safety Facts,” National Highway Traffic Safety Administration, June 2015](#)). The observations represent the percentage of drivers wearing seat belts in a large nationwide observational survey.

Some, but not all, states enforce seat belts laws. Below are the seat belt usage data divided into two groups—states that enforce seat belts laws and those that do not.

States with Seat Belt Law Enforcement:

95.7 88.4 74.4 97.1 85.1 91.9 93.2 88.8 97.3 93.5 94.1 90.2
 92.8 85.7 86.1 84.1 85.0 92.1 93.3 94.7 78.3 87.6 92.1 90.6
 90.6 86.3 97.8 87.4 90.0 68.9 87.7 90.7 94.5 87.8 84.7

States without Seat Belt Law Enforcement:

87.2 82.4 80.2 76.6 78.8 74.0 70.0 94.0 70.4 81.0 85.0 83.6
 68.9 83.4 84.1 77.3 79.2

- a.** Construct a comparative stem-and-leaf display using the tens digit as the stem and truncating the leaves to a single digit (the ones digit).
- b.** Write a few sentences commenting on similarities or differences in the seat belt use distributions for states with seat belt enforcement and states without seat belt enforcement.

- 3.19** ● The U.S. Department of Health and Human Services reported the estimated percentage of households with only wireless phone service (no land line) in 2014 for the 50 U.S. states and the District of Columbia ([cdc.gov/nchs/data/nhis/earlyrelease/wireless_state_201602.pdf, retrieved February 17, 2018](#)). In the accompanying data table, each state was also classified into one of three geographical regions—West (W), Middle states (M), and East (E).

Wireless			Wireless		
%	Region	State	%	Region	State
43.4	M	AL	41.0	W	MT
39.7	W	AK	46.5	M	NE
49.4	W	AZ	48.4	W	NV
56.2	M	AR	43.6	M	ND
42.8	W	CA	31.2	E	NH
50.5	W	CO	25.1	E	NJ
26.7	E	CT	47.0	W	NM
29.4	E	DE	31.1	E	NY
49.7	E	DC	42.9	E	NC
47.6	E	FL	45.8	E	OH
45.9	E	GA	50.4	M	OK
38.3	W	HI	47.0	W	OR
56.1	W	ID	30.0	E	PA
45.7	M	IL	34.6	E	RI
47.7	M	IN	45.9	E	SC
50.7	M	IA	41.4	M	SD
51.6	M	KS	46.6	M	TN
47.1	M	KY	54.6	M	TX
40.9	M	LA	52.2	W	UT
40.8	E	ME	41.1	E	VA
36.2	E	MD	37.2	E	VT
31.5	E	MA	48.3	W	WA
47.8	M	MI	37.2	E	WV
43.1	M	MN	46.6	M	WI
55.1	M	MS	51.8	W	WY
51.5	M	MO			

- a. Construct a stem-and-leaf display for the wireless percentage using the data from all 50 states and the District of Columbia. What is a typical value for this data set?
- b. Construct a comparative stem-and-leaf display for the wireless percentage of the states in the West and the states in the East. How do the distributions of wireless percentages compare for states in the East and states in the West? (Hint: See Example 3.11.)

- 3.20** The article “**Economy Low, Generosity High**” (*USA TODAY*, July 28, 2009) noted that despite a weak economy in 2008, more Americans volunteered in their communities than in previous years. Based on census data (volunteeringinamerica.gov), the top and bottom five states in terms of percentage of the population who volunteered in 2008 were identified. The top five states were Utah (43.5%), Nebraska (38.9%), Minnesota (38.4%), Alaska (38.0%), and Iowa (37.1%). The bottom five states were New York (18.5%), Nevada (18.8%), Florida (19.6%), Louisiana (20.1%), and Mississippi (20.9%).
- a. For the data set that includes the percentage who volunteered in 2008 for each of the 50 states, what is the largest value? What is the smallest value?

- b. If you were going to construct a stem-and-leaf display for the data set consisting of the percentage who volunteered in 2008 for the 50 states, what stems would you use to construct the display? Explain your choice.

- 3.21** The U.S. gasoline tax per gallon data for each of the 50 states and the District of Columbia in 2015 were obtained from the U.S. Energy Information Administration (eia.gov/tools/faqs/faq.cfm?id=10&t=10, retrieved April 17, 2017).

State	Gasoline Tax (cents per gallon)
Alabama	19.0
Alaska	9.0
Arizona	19.0
Arkansas	21.8
California	37.2
Colorado	23.3
Connecticut	25.0
Delaware	23.0
District of Columbia	23.5
Florida	30.6
Georgia	26.5
Hawaii	18.5
Idaho	33.0
Illinois	33.1
Indiana	29.0
Iowa	31.8
Kansas	25.0
Kentucky	26.0
Louisiana	20.9
Maine	31.4
Maryland	32.8
Massachusetts	26.7
Michigan	30.9
Minnesota	30.6
Mississippi	18.4
Missouri	17.3
Montana	27.8
Nebraska	27.7
Nevada	23.8
New Hampshire	23.8
New Jersey	14.6
New Mexico	18.9
New York	33.8
North Carolina	35.3
North Dakota	23.0
Ohio	28.0
Oklahoma	17.0
Oregon	30.0
Pennsylvania	51.4
Rhode Island	34.1
South Carolina	16.8
South Dakota	30.0

(continued)

State	Gasoline Tax (cents per gallon)
Tennessee	21.4
Texas	20.0
Utah	30.1
Vermont	30.5
Virginia	16.8
Washington	44.6
West Virginia	33.2
Wisconsin	32.9
Wyoming	24.0

- Construct a stem-and-leaf display of these data.
- Based on the stem-and-leaf display, what do you notice about the center and spread of the data distribution?
- Do any values in the data set stand out as unusual? If so, which states correspond to the unusual observations, and how do these values differ from the rest?

3.22 A report from [Texas Transportation Institute \(Texas A&M University System, 2005\)](#) titled “Congestion Reduction Strategies” included the accompanying data on extra travel time for peak travel time in hours per year per traveler for different-sized urban areas.

Very Large Urban Areas	Extra Hours per Year per Traveler
Los Angeles, CA	93
San Francisco, CA	72
Washington DC-VA-MD	69
Atlanta, GA	67
Houston, TX	63
Dallas, Fort Worth, TX	60
Chicago, IL-IN	58
Detroit, MI	57
Miami, FL	51
Boston, MA-NH-RI	51
New York, NY-NJ-CT	49
Phoenix, AZ	49
Philadelphia, PA-NJ-DE-MD	38

Large Urban Areas	Extra Hours per Year per Traveler
Riverside, CA	55
Orlando, FL	55
San Jose, CA	53
San Diego, CA	52
Denver, CO	51
Baltimore, MD	50
Seattle, WA	46
Tampa, FL	46
Minneapolis, St Paul, MN	43

Large Urban Areas	Extra Hours per Year per Traveler
Sacramento, CA	40
Portland, OR, WA	39
Indianapolis, IN	38
St Louis, MO-IL	35
San Antonio, TX	33
Providence, RI-MA	33
Las Vegas, NV	30
Cincinnati, OH-KY-IN	30
Columbus, OH	29
Virginia Beach, VA	26
Milwaukee, WI	23
New Orleans, LA	18
Kansas City, MO-KS	17
Pittsburgh, PA	14
Buffalo, NY	13
Oklahoma City, OK	12
Cleveland, OH	10

- Construct a comparative stem-and-leaf plot for extra travel time per traveler for the two different sizes of urban areas.
- Is the following statement consistent with the display constructed in Part (a)? Explain.

The larger the urban area, the greater the extra travel time during peak period travel.

3.23 The percentage of teens not in school or working in 2010 for the 50 states were given in the [2012 Kids Count Data Book \(aecf.org\)](#) and are shown in the following table:

State	Percentage	State	Percentage
Alabama	11%	Massachusetts	5%
Alaska	11%	Michigan	9%
Arizona	12%	Minnesota	5%
Arkansas	12%	Mississippi	13%
California	8%	Missouri	9%
Colorado	7%	Montana	9%
Connecticut	5%	Nebraska	4%
Delaware	9%	Nevada	15%
Florida	10%	New Hampshire	6%
Georgia	12%	New Jersey	8%
Hawaii	12%	New Mexico	12%
Idaho	11%	New York	8%
Illinois	8%	North Carolina	10%
Indiana	8%	North Dakota	5%
Iowa	6%	Ohio	8%
Kansas	6%	Oklahoma	9%
Kentucky	11%	Oregon	10%
Louisiana	14%	Pennsylvania	7%
Maine	7%	Rhode Island	5%
Maryland	8%	South Carolina	9%

(continued)

(continued)

State	Percentage	State	Percentage
South Dakota	8%	Virginia	7%
Tennessee	10%	Washington	8%
Texas	9%	West Virginia	14%
Utah	9%	Wisconsin	7%
Vermont	4%	Wyoming	9%

Note that the percentages range from a low of 4% to a high of 15%. In constructing a stem-and-leaf display for these data, if we regard each percentage as a two-digit number and use the first digit for the stem, then there are only two possible stems, 0 and 1. One solution is to use repeated stems. Consider a scheme

that divides the leaf range into five parts: 0 and 1, 2 and 3, 4 and 5, 6 and 7, and 8 and 9. Then, for example, stem 0 could be repeated as

- 0 with leaves 0 and 1
- 0t with leaves 2 and 3
- 0f with leaves 4 and 5
- 0s with leaves 6 and 7
- 0* with leaves 8 and 9

Construct a stem-and-leaf display for this data set that uses stems 0t, 0f, 0s, 0*, and 1, 1t, and 1f. Comment on the important features of the display. (Hint: See Example 3.10.)

SECTION 3.3 Displaying Numerical Data: Frequency Distributions and Histograms

A stem-and-leaf display is not always an effective way to summarize data, because it can be unwieldy when the data set contains a large number of observations. Frequency distributions and histograms are displays that work well for large data sets.

Frequency Distributions and Histograms for Discrete Numerical Data

Discrete numerical data often results from counting. When this is the case, each observation is a whole number. As in the case of categorical data, a frequency distribution for discrete numerical data lists each possible value (either individually or grouped into intervals), the associated frequency, and sometimes the corresponding relative frequency. Recall that relative frequency is calculated by dividing the frequency by the total number of observations in the data set.

Example 3.12 Promiscuous Queen Bees

Understand the context ➤

- Queen honey bees mate shortly after they become adults. During a mating flight, the queen usually takes multiple partners, collecting sperm that she will store and use throughout the rest of her life. The authors of the paper “[The Curious Promiscuity of Queen Honey Bees](#)” (*Annals of Zoology* [2001]: 255–265) studied the behavior of 30 queen honey bees to learn about the length of mating flights and the number of partners a queen takes during a mating flight.

The accompanying data on number of partners on one mating flight were generated to be consistent with summary values and graphs given in the paper.

Consider the data ➤

Number of Partners

12	2	4	6	6	7	8	7	8	11
8	3	5	6	7	10	1	9	7	6
9	7	5	4	7	4	6	7	8	10

The corresponding relative frequency distribution is given in Table 3.1. The smallest value in the data set is 1 and the largest is 12, so the possible values from 1 to 12 are listed in the table, along with the corresponding frequency and relative frequency.

● Data set available online

TABLE 3.1 Relative Frequency Distribution for Number of Partners

Number of Partners	Frequency	Relative Frequency
1	1	0.033 ← $\frac{1}{30} = 0.033$
2	1	0.033
3	1	0.033
4	3	0.100
5	2	0.067
6	5	0.167
7	7	0.233
8	4	0.133
9	2	0.067
10	2	0.067
11	1	0.033
12	1	0.033
Total	30	0.999 ← <i>Differs from 1 due to rounding</i>

Interpret the results ►

From the relative frequency distribution, we can see that five of the queen bees had six partners during their mating flight. The corresponding relative frequency, $\frac{5}{30} = 0.167$, tells us that the proportion of queens with six partners is 0.167, or equivalently 16.7% of the queens had six partners. Adding the relative frequencies for the values 10, 11, and 12 gives

$$0.067 + 0.033 + 0.033 = 0.133$$

indicating that 13.3% of the queens had 10 or more partners.

It is possible to create a more compact frequency distribution by grouping some of the possible values into intervals. For example, we might group together 1, 2, and 3 partners to form an interval of 1–3, with a corresponding frequency of 3. The grouping of other values in a similar way results in the relative frequency distribution shown in Table 3.2.

TABLE 3.2 Relative Frequency Distribution of Number of Partners Using Intervals

Number of Partners	Frequency	Relative Frequency
1–3	3	0.100
4–6	10	0.333
7–9	13	0.433
10–12	4	0.133

A histogram for discrete numerical data is a graph of the frequency or relative frequency distribution, and it is similar to the bar chart for categorical data. Each frequency or relative frequency is represented by a rectangle centered over the corresponding value (or range of values) and the area of the rectangle is proportional to the corresponding frequency or relative frequency.

Histogram for Discrete Numerical Data

When to Use Discrete numerical data. Works well for large data sets.

How to Construct

1. Draw a horizontal scale, and mark the possible values of the variable.
2. Draw a vertical scale, and mark it with either frequency or relative frequency.
3. Above each possible value, draw a rectangle centered at that value (so that the rectangle for 1 is centered at 1, the rectangle for 5 is centered at 5, and so on). The height of each rectangle is determined by the corresponding frequency or

(continued)

relative frequency. Often possible values are consecutive whole numbers, in which case the base width for each rectangle is 1.

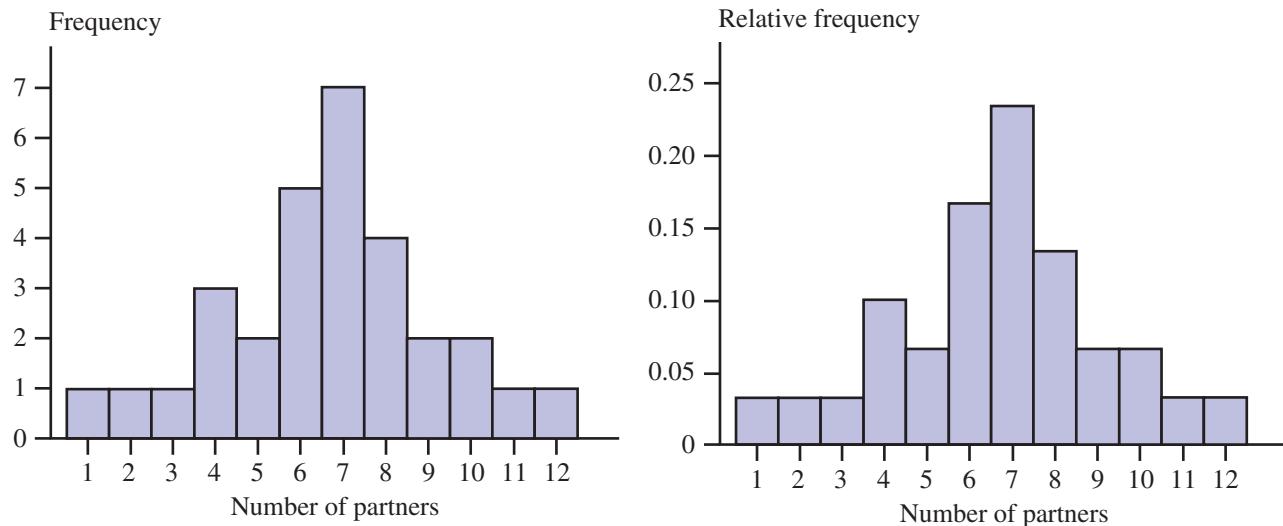
What to Look For

- Center or typical value
- Extent of variability
- General shape
- Location and number of peaks
- Presence of gaps and outliers

Example 3.13 Promiscuous Queen Bees Revisited

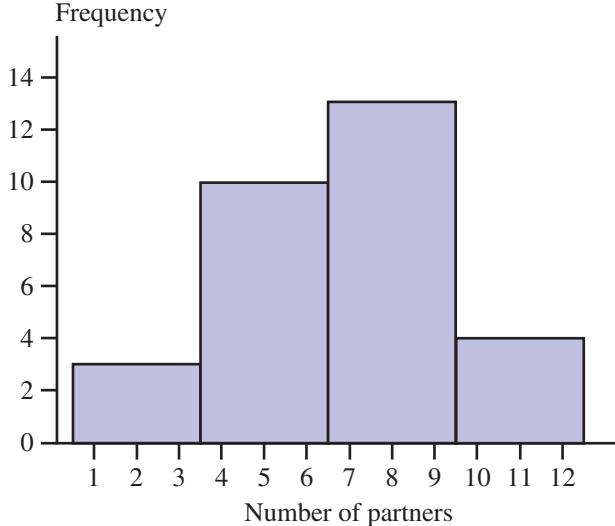
The queen bee data of Example 3.12 were summarized in a frequency distribution. The corresponding histogram is shown in Figure 3.17. Note that each rectangle in the histogram is centered over the corresponding value. When relative frequency instead of frequency is used for the vertical scale, the scale on the vertical axis is different but all essential characteristics of the graph (shape, center, variability) are unchanged.

FIGURE 3.17
Histogram and relative frequency histogram of queen bee data.



A histogram based on the grouped frequency distribution of Table 3.2 can be constructed in a similar fashion, and is shown in Figure 3.18. A rectangle represents the frequency or relative frequency for each interval. For the interval from 1 to 3, the rectangle extends from 0.5 to 3.5 so that there are no gaps between the rectangles of the histogram.

FIGURE 3.18
Histogram of queen bee data using intervals.



Sometimes a discrete numerical data set contains a large number of possible values and perhaps also has a few large or small values that are far away from most of the data. In this case, rather than forming a frequency distribution with a very long list of possible values, it is common to group the observed values into intervals or ranges. This is illustrated in Example 3.14.

Example 3.14 Math SAT Score Distribution

Understand the context ➤

Each of the 1,637,589 students who took the math portion of the SAT exam in 2016 received a score between 200 and 800. The score distribution was summarized in a frequency distribution table that appeared in the **College Board** report titled “**2016 College Bound Seniors**.” A relative frequency distribution is given in Table 3.3 and the corresponding relative frequency histogram is shown in Figure 3.19. Notice that rather than list each possible individual score value between 200 and 800, the scores are grouped into intervals (200 to 299, 300 to 399, etc.). This results in a much more compact table that still communicates the important features of the data set. Also, notice that because the data set is so large, the frequencies are also large numbers. Because of these large frequencies, it is easier to focus on the relative frequencies in our interpretation.

TABLE 3.3 Relative Frequency Distribution of Math SAT Score

Consider the data ➤

Math SAT Score	Frequency	Relative Frequency
200–299	117,067	0.071
300–399	277,969	0.170
400–499	461,793	0.282
500–599	488,370	0.298
600–699	234,465	0.143
700–800	57,925	0.035

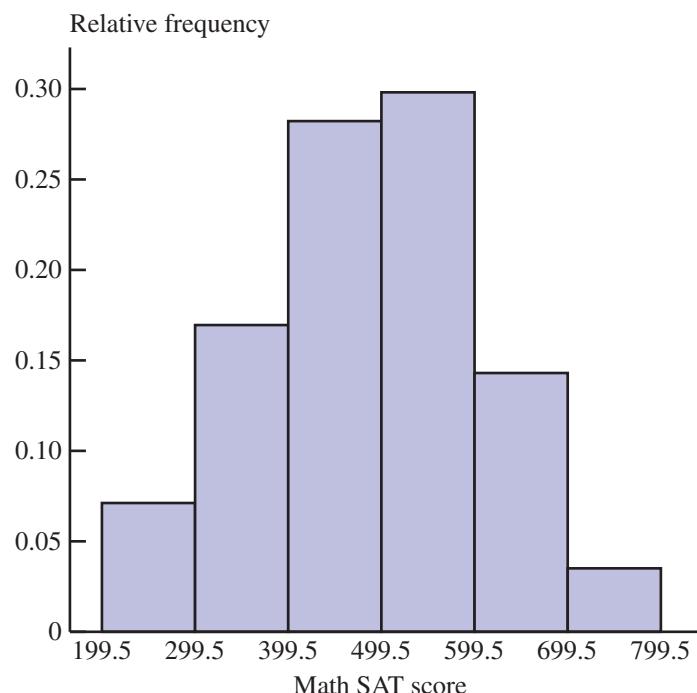
Interpret the results ➤

From the relative frequency distribution and histogram, we can see that while there is a lot of variability in individual math SAT scores, the majority were in the 400 to 600 range and a typical value for math SAT looks to be something in the high 400s.

Before leaving this example, take a second look at the relative frequency histogram of Figure 3.19. Notice that there is one rectangle for each score interval in the relative frequency distribution. For simplicity we have chosen to treat the very last interval, 700 to 800, as if it were 700 to 799 so that all of the score ranges in the frequency distribution are the same width.

FIGURE 3.19

Relative frequency histogram for the math SAT data.



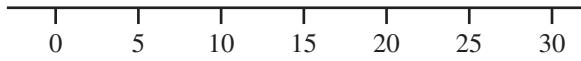
Also note that the rectangle representing the score range 400 to 499 actually extends from 399.5 to 499.5 on the score scale. This is similar to what happens in histograms for discrete numerical data where there is no grouping. For example, in Figure 3.17 the rectangle representing 2 is centered at 2 but extends from 1.5 to 2.5 on the number of partners scale.

Frequency Distributions and Histograms for Continuous Numerical Data

For continuous data, such as observations on reaction time (in seconds) or weight of airline passenger carry-on luggage (in pounds), there are no natural categories. In this case, we define our own categories. For carry-on luggage weight, we might expect weights up to about 30 pounds. One way to group the weights into 5-pound intervals is shown in Figure 3.20. Then each observed data value could be classified into one of these intervals. The intervals used are sometimes called **class intervals**. The class intervals play the same role that the categories or individual values played in frequency distributions for categorical or discrete numerical data.

FIGURE 3.20

Suitable class intervals for carry-on luggage weight data.



There is one further difficulty that needs to be addressed. Where should we place an observation such as 20, which falls on a boundary between classes? Our convention is to define intervals so that such an observation is placed in the upper rather than the lower class interval. This means that in a frequency distribution, one class might be 15 to <20 , where the symbol $<$ is a substitute for the phrase *less than*. This class interval will contain all observations that are greater than or equal to 15 and less than 20. The observation 20 would then fall in the class interval 20 to <25 .

Example 3.15 Going Away to College

Understand the context ➤

- States differ widely in the percentage of college students who attend college in their home state. The percentages of freshmen who attended college in their home state for each of the 50 states are shown here (*The Chronicle of Higher Education*, August 23, 2013). The data have been ordered from largest to smallest.

Consider the data ➤

Percentage of College Students Attending College in Home State

93	92	91	91	90	90	90	90	89	89
89	89	89	89	89	88	87	87	85	85
85	85	84	84	83	81	81	81	80	78
77	77	76	76	76	76	72	72	70	68
67	65	65	64	62	60	58	57	57	50

The smallest observation is 50% (Vermont) and the largest is 93% (Mississippi). It is reasonable to start the first class interval at 40 and let each interval have a width of 10. This gives class intervals of 50 to <60 , 60 to <70 , 70 to <80 , 80 to <90 , and 90 to <100 .

Table 3.4 displays the resulting frequency distribution, along with the relative frequencies.

TABLE 3.4 Frequency Distribution for Percentage of College Students Attending College in Home State

Class Interval	Frequency	Relative Frequency
50 to <60	4	0.08
60 to <70	7	0.14
70 to <80	10	0.20
80 to <90	21	0.42
90 to <100	8	0.16
	50	1.00

● Data set available online

Various relative frequencies can be combined to yield other interesting information. For example,

$$\begin{pmatrix} \text{proportion of states with} \\ \text{percent attending college} \\ \text{in home state less than 70} \end{pmatrix} = \begin{pmatrix} \text{proportion in} \\ 50 \text{ to } < 60 \text{ class} \end{pmatrix} + \begin{pmatrix} \text{proportion in} \\ 60 \text{ to } < 70 \text{ class} \end{pmatrix}$$

$$= 0.08 + 0.14 = 0.22 (22\%)$$

and

$$\begin{pmatrix} \text{proportion of states} \\ \text{with percent attending} \\ \text{college in home state} \\ \text{between 60 and 90} \end{pmatrix} = \begin{pmatrix} \text{proportion in} \\ 60 \text{ to } < 70 \\ \text{class} \end{pmatrix} + \begin{pmatrix} \text{proportion in} \\ 70 \text{ to } < 80 \\ \text{class} \end{pmatrix} + \begin{pmatrix} \text{proportion in} \\ 80 \text{ to } < 90 \\ \text{class} \end{pmatrix}$$

$$= 0.14 + 0.20 + 0.42 = 0.76 (76\%).$$

There are no set rules for selecting either the number of class intervals or the length of the intervals. Using a few relatively wide intervals will bunch the data, whereas using a great many relatively narrow intervals may spread the data over too many intervals, so that no interval contains more than a few observations. Neither type of distribution will give an informative picture of how data values are distributed, and interesting features of the data set may be missed. In general, with a small number of observations, relatively few intervals, perhaps between 5 and 10, should be used. With a large number of observations, a distribution based on 15 to 20 (or even more) intervals is often recommended. The quantity

$$\sqrt{\text{number of observations}}$$

is sometimes used as an estimate of an appropriate number of intervals: 5 intervals for 25 observations, 10 intervals when the number of observations is 100, and so on.

Histograms for Continuous Numerical Data

When the class intervals in a frequency distribution are all of equal width, it is easy to construct a histogram using the information in a frequency distribution.

Histogram for Continuous Numerical Data When the Class Interval Widths Are Equal

When to Use Continuous numerical data. Works well, even for large data sets.

How to Construct

1. Mark the boundaries of the class intervals on a horizontal axis.
2. Use either frequency or relative frequency on the vertical axis.
3. Draw a rectangle for each interval directly above the corresponding interval (so that the edges are at the class interval boundaries). The height of each rectangle is the corresponding frequency or relative frequency.

What to Look For

- Center or typical value
- Extent of variability
- General shape
- Location and number of peaks
- Presence of gaps and outliers

Two people making reasonable and similar choices for the number of intervals, the interval width, and the starting point of the first interval, will usually obtain histograms that are similar in terms of shape, center, and variability.

Example 3.16 Sleep Deficit and School Start Time

Understand the context ➤

The authors of the paper “[The Influence of School Time on Sleep Patterns of Children and Adolescents](#)” (*Sleep Medicine* [2016]: 33–39) were interested in determining if early school start time has an effect on the amount of sleep school-age children get. They studied students in Brazil in a region that offered both a morning school start time (with classes from 7:30 a.m. to noon) and an afternoon school start time (with classes from 1:30 p.m. to 5:30 p.m.). One variable of interest in the study was sleep deficit, which they defined as the difference in sleep duration on weekends and the sleep duration on nights with school the next day. Sleep deficit was measured in hours, so a student that typically slept 9 hours a night on weekends and only 7 hours a night on weeknights would have a sleep deficit of 2 hours. A student with a sleep deficit of –3 hours would be one who typically slept 3 hours longer on weeknights than on weekends, resulting in a negative difference.

Table 3.5 gives relative frequencies (as approximate values based on a graph that appears in the paper) for various sleep deficit intervals for students with the morning start time and for students with the afternoon start time. Relative frequencies are used because there were 538 students in the morning start group and only 101 students in the afternoon start group.

TABLE 3.5 Relative Frequency Distribution of Sleep Deficit

Consider the data ➤

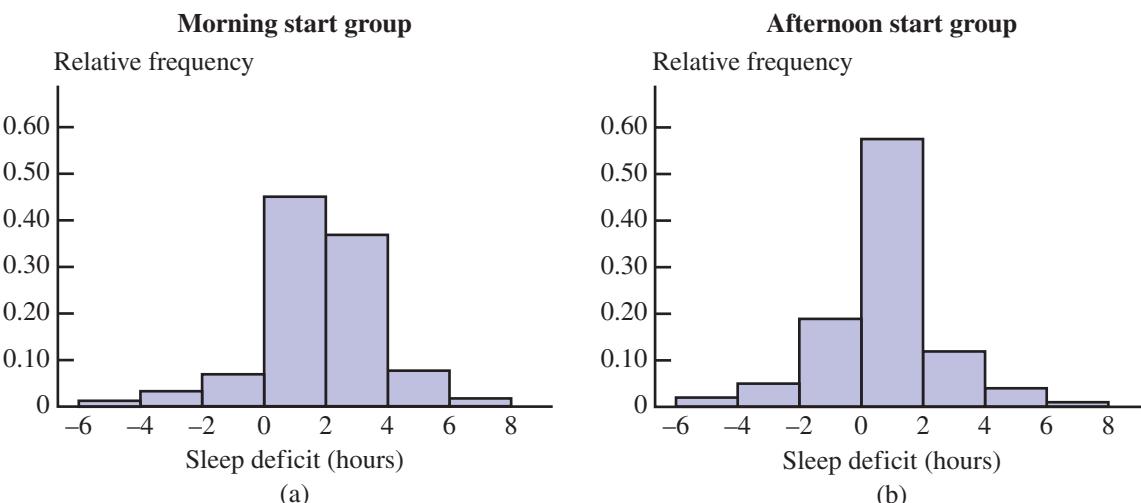
Sleep Deficit (in hours)	Morning Start Relative Frequency	Afternoon Start Relative Frequency
–6 to < –4	0.007	0.020
–4 to < –2	0.028	0.050
–2 to < 0	0.065	0.190
0 to < 2	0.442	0.570
2 to < 4	0.364	0.120
4 to < 6	0.078	0.040
6 to < 8	0.015	0.010

Figure 3.21(a) is the relative frequency histogram for the morning start time group and Figure 3.21(b) is the relative frequency histogram for afternoon start time group. Because we would like to compare the distributions of sleep deficit for the morning start and afternoon start groups, it is important to use the same scales for the two histograms.

FIGURE 3.21

Histogram of sleep deficit:

- (a) morning start;
- (b) afternoon start.



Interpret the results ➤

Notice that both histograms have a single peak, with the majority of students in both groups having positive values for sleep deficit. However, students in the afternoon school start time group tended to have smaller deficits than the students in the morning start time group, indicating that difference between sleep duration on weekends and sleep duration on weekdays tended to be smaller for students with an afternoon start time. This is one reason that many people are recommending later school start times.

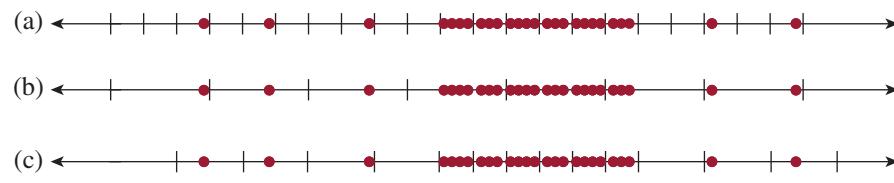
Class Intervals of Unequal Widths

Figure 3.22 shows a data set in which a large number of observations are concentrated in the middle of the data set, with only a few unusual values far below and far above the center of the data. If a frequency distribution is based on short intervals of equal width, many intervals will be required to capture all observations, and a lot of them will contain no observations, as shown in Figure 3.22(a). On the other hand, only a few wide intervals will capture all values, but then most of the observations will be grouped into just a few intervals, as shown in Figure 3.22(b). In such situations, it is best to use a combination of wide class intervals where there are few data points and shorter intervals where there are many data points, as shown in Figure 3.22(c).

FIGURE 3.22

Three choices of class intervals for a data set with outliers:

- (a) many short intervals of equal width;
- (b) a few wide intervals of equal width;
- (c) intervals of unequal width.



Constructing a Histogram for Continuous Data When Class Interval Widths Are Unequal

When class intervals are not all the same width, frequencies or relative frequencies should not be used on the vertical axis of a histogram. Instead, the height of each rectangle, called the **density** for the class interval, is given by

$$\text{density} = \frac{\text{relative frequency of class interval}}{\text{class interval width}}$$

The vertical axis is called the **density scale**.

The use of the density scale to construct the histogram ensures that the area of each rectangle in the histogram will be proportional to the corresponding relative frequency. The formula for density can also be used when class widths are equal. However, when the intervals all have equal width, the extra arithmetic required to obtain the densities is not necessary.

Example 3.17 Student Debt on the Rise

Understand the context ➤

At many U.S. colleges and universities, low-income students often need to take on a large debt burden to pay for their education. The article “[Poor Feel the Bite of Rising College Costs](#)” (*The Wall Street Journal*, February 20, 2016) classified 1319 colleges into intervals based on the median yearly debt for students from families with an annual income of \$30,000 or less. The data are summarized in the frequency distribution given in Table 3.6.

TABLE 3.6 Frequency Distribution for Median Student Debt

Consider the data ➤

Class Interval	Frequency	Relative Frequency	Interval Width	Density
\$0 to < \$10,000	166	0.126	\$10,000	0.000013
\$10,000 to < \$15,000	389	0.295	\$5,000	0.000059
\$15,000 to < \$20,000	456	0.346	\$5,000	0.000069
\$20,000 to < \$40,000	308	0.234	\$20,000	0.000012

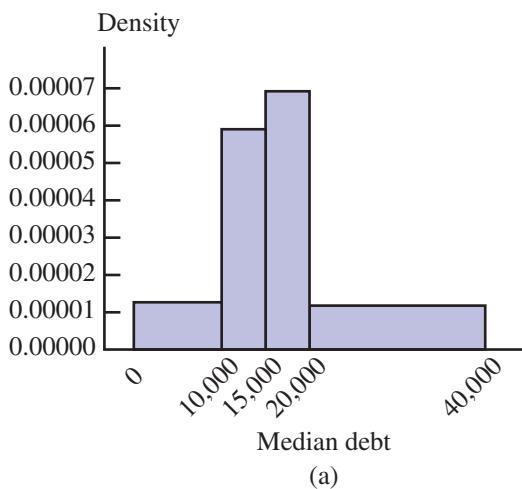
Looking at the frequency distribution, you can see that the median student debt for low-income students was between \$20,000 and \$40,000 for 308 of the colleges and was between \$15,000 and \$20,000 for 456 of the colleges. Notice also that the income intervals used in the article are not of equal width. Two intervals have a width of \$5000 (for example, the interval from \$10,000 to < \$15,000), and the other two intervals have widths of \$10,000 and \$20,000, respectively.

Figure 3.23 displays two histograms based on this frequency distribution. The histogram in Figure 3.23(a) is correctly drawn, with density used to determine the height of each bar. The histogram in Figure 3.23(b) has height equal to relative frequency and is therefore not correct. In particular, this second histogram exaggerates the proportion of colleges with low median debt and the proportion of colleges with very high median debt—the areas of the two most extreme rectangles are much too large. The eye is naturally drawn to large areas, so it is important that the areas correctly represent the relative frequencies.

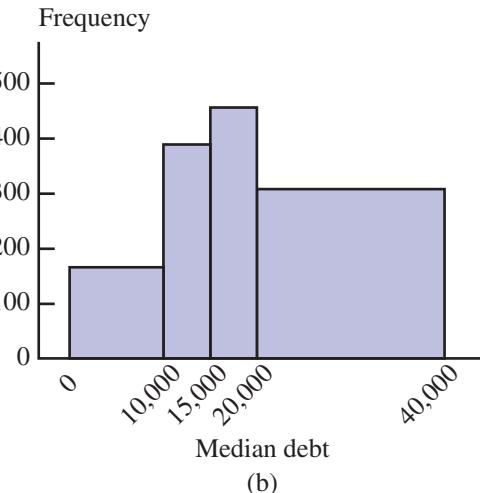
FIGURE 3.23

Histograms for median student debt:

- (a) a correct histogram
(height = density);
- (b) an incorrect histogram
(height = relative frequency).

Correct Histogram of Median Debt

(a)

Incorrect Histogram of Median Debt

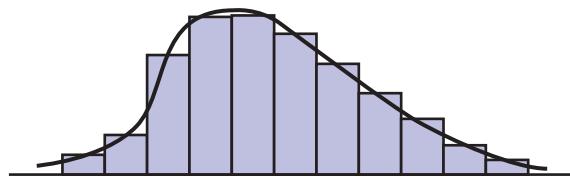
(b)

Histogram Shapes

General shape is an important characteristic of a histogram. In describing various shapes it is convenient to approximate the histogram with a smooth curve (called a *smoothed histogram*). This is illustrated in Figure 3.24.

FIGURE 3.24

Approximating a histogram with a smooth curve.



One description of general shape relates to the number of peaks, or **modes**.

DEFINITIONS

Unimodal: A histogram is **unimodal** if it has a single peak.

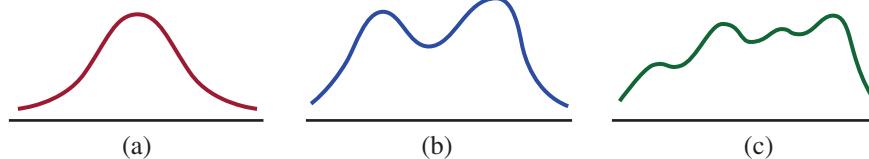
Bimodal: A histogram is **bimodal** if it has two peaks.

Multimodal: A histogram is **multimodal** if it has more than two peaks.

These shapes are illustrated in Figure 3.25.

FIGURE 3.25

Smoothed histograms with various numbers of modes:
(a) unimodal;
(b) bimodal;
(c) multimodal.



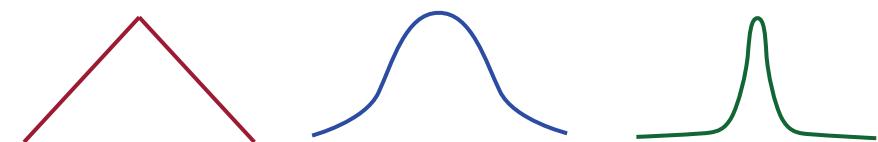
Bimodal histograms sometimes occur when the data set consists of observations on two quite different kinds of individuals or objects. For example, consider a large data set consisting of driving times for automobiles traveling between San Luis Obispo, California, and Monterey, California. This histogram would show two peaks, one for those cars that took the inland route (roughly 2.5 hours) and another for those cars traveling up the coast highway (3.5–4 hours).

However, bimodality does not automatically follow in such situations. Bimodality will occur in the histogram of the combined groups only if the centers of the two separate histograms are far apart relative to the variability in the two data sets. For example, a large data set consisting of heights of college students would probably not produce a bimodal histogram because the typical height for males (about 69 in.) and the typical height for females (about 66 in.) are not very far apart relative to the variability in the height distributions.

Unimodal histograms come in a variety of shapes. A unimodal histogram is **symmetric** if there is a vertical line of symmetry such that the part of the histogram to the left of the line is a mirror image of the part to the right. (Bimodal and multimodal histograms can also be symmetric in this way.) Several different symmetric smoothed histograms are shown in Figure 3.26.

FIGURE 3.26

Several symmetric unimodal smoothed histograms.



Proceeding to the right from the peak of a unimodal histogram, we move into what is called the **upper tail** of the histogram. Going in the opposite direction moves us into the **lower tail**.

DEFINITIONS

Skewed: A unimodal histogram that is not symmetric is described as **skewed**.

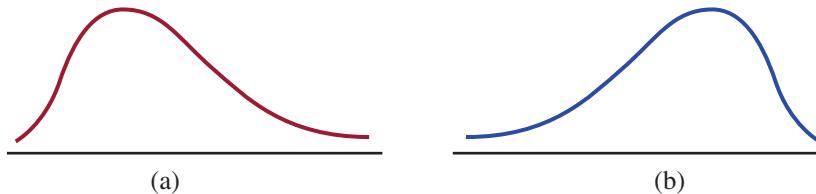
Positively skewed or right skewed: A skewed histogram in which the upper tail of the histogram stretches out much farther than the lower tail is described as **positively skewed or right skewed**.

Negatively skewed or left skewed: A skewed histogram in which the lower tail of the histogram stretches out much farther than the upper tail is described as **negatively skewed or left skewed**.

These two types of skewness are illustrated in Figure 3.27. Positive skewness is much more frequently encountered than is negative skewness. An example of positive skewness occurs in the distribution of single-family home prices in Los Angeles County. Most homes are moderately priced (at least for California), whereas the relatively few homes in Beverly Hills and Malibu have much higher price tags.

FIGURE 3.27

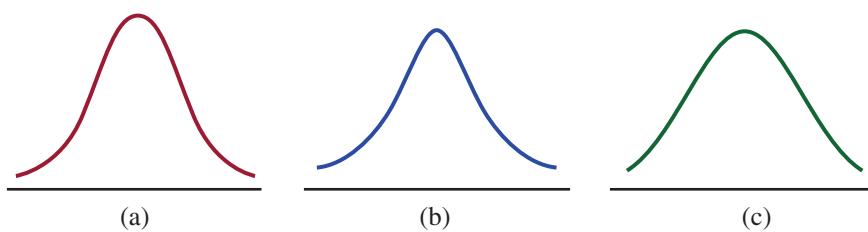
Two examples of skewed smoothed histograms:
(a) positive skew;
(b) negative skew.



One specific shape, a **normal curve**, arises more frequently than any other in statistical applications. Many histograms can be well approximated by a normal curve (for example, characteristics such as arm span and the weight of an apple). Here we briefly mention several of the most important characteristics of normal curves, postponing a more detailed discussion until Chapter 7. A normal curve is both symmetric and bell-shaped, and it looks like the curve in Figure 3.28(a). However, not all bell-shaped curves are normal. In a normal curve, starting from the top of the bell the height of the curve decreases at a well-defined rate when moving toward either tail. (This rate of decrease is specified by a certain mathematical function.)

FIGURE 3.28

Three examples of bell-shaped histograms:
(a) normal;
(b) heavy-tailed;
(c) light-tailed.



A curve with tails that do not decline as rapidly as the tails of a normal curve is called **heavy-tailed** (compared to the normal curve). Similarly, a curve with tails that decrease more rapidly than the normal tails is called **light-tailed**. Figure 3.28(b) and (c) illustrate these possibilities. The reason that we are concerned about the tails in a distribution is that many inferential procedures that work well (meaning they result in a high proportion of correct conclusions) when the population distribution is approximately normal perform poorly when the population distribution is heavy-tailed.

Do Sample Histograms Resemble Population Histograms?

Sample data are usually collected to learn about a population. Conclusions based on sample data may not be correct if the sample is unrepresentative of the population. So how similar might a histogram of sample data be to the histogram of all population values? Will the two histograms be similar in terms of center and variability? Will they have the same number of peaks, and will the peaks occur at approximately the same places?

A related issue concerns the extent to which histograms based on different samples from the same population resemble one another. If two different sample histograms can be expected to differ from one another in obvious ways, then at least one of them might differ substantially from the population histogram. If the sample differs substantially from the population, conclusions about the population based on the sample are likely to be incorrect. **Sampling variability**—the extent to which samples from the same population differ from one another and from the population—is a central idea in statistics. Example 3.18 considers sampling variability in histogram shapes.

Example 3.18 What You Should Know About Bus Drivers . . .

Understand the context ➤

- A sample of 708 bus drivers employed by public corporations was selected, and the number of traffic accidents in which each bus driver was involved during a 4-year period was determined ([“Application of Discrete Distribution Theory to the Study of Noncommunicable Events in Medical Epidemiology,” in Random Counts in Biomedical and Social Sciences, G. P. Patil, ed. \[University Park, PA: Pennsylvania State University Press, 1970\]](#)). A listing of the 708 sample observations might look like this:

3 0 6 0 0 2 1 4 1 ... 6 0 2

The frequency distribution (Table 3.7) shows that 117 of the 708 drivers had no accidents, a relative frequency of $117/708 = 0.165$ (or 16.5%). Similarly, the proportion of sampled drivers who had 1 accident is 0.222 (or 22.2%). The largest sample observation was 11.

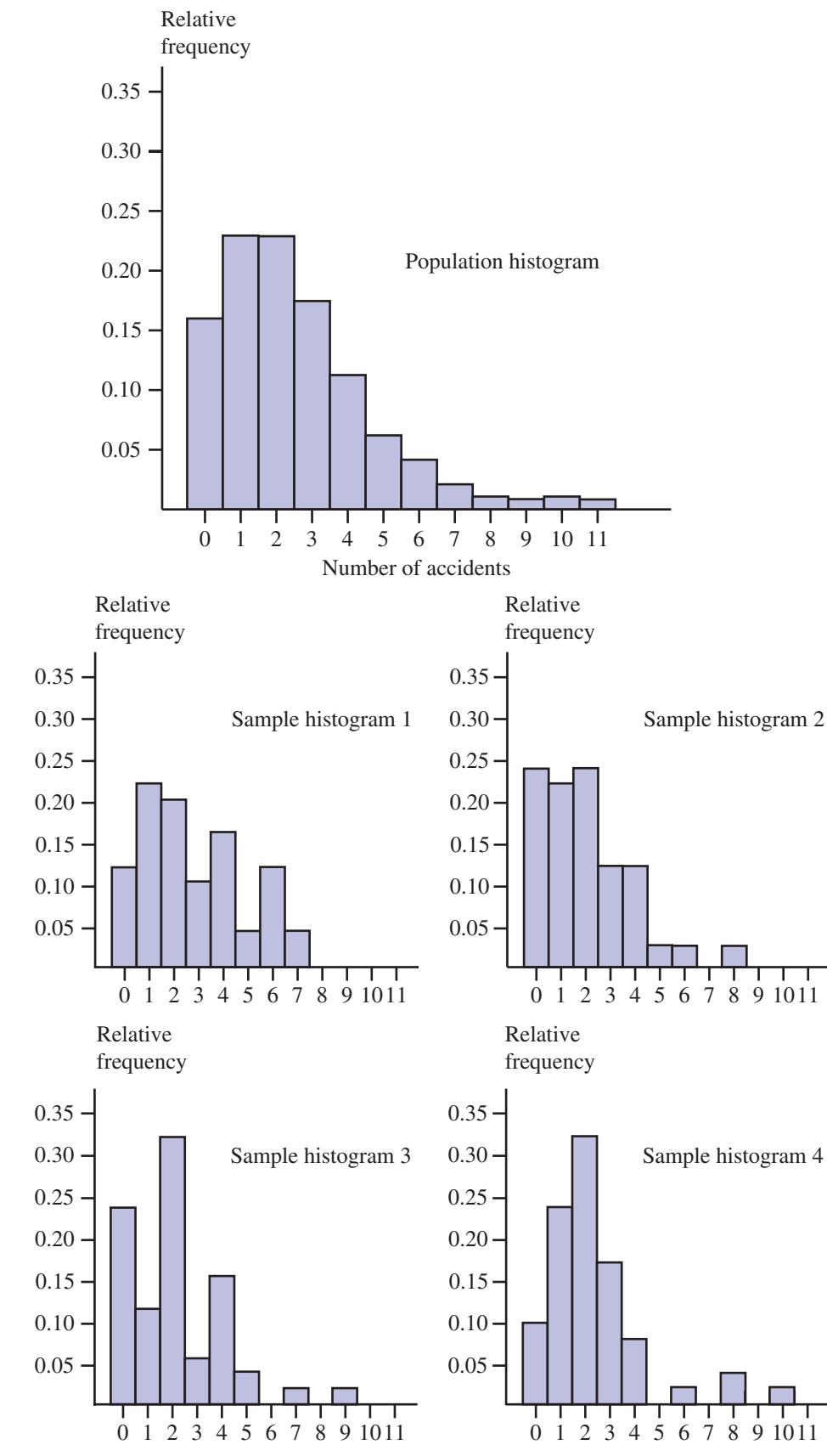
TABLE 3.7 Frequency Distribution for Number of Accidents by Bus Drivers

Number of Accidents	Frequency	Relative Frequency
0	117	0.165
1	157	0.222
2	158	0.223
3	115	0.162
4	78	0.110
5	44	0.062
6	21	0.030
7	7	0.010
8	6	0.008
9	1	0.001
10	3	0.004
11	1	0.001
	708	0.998

Although the 708 observations were actually a sample from the population of all bus drivers, we will regard the 708 observations as a population. The first histogram in Figure 3.29, represents the population histogram based on all 708 observations. The other four histograms in Figure 3.29 are based on four different random samples of 50 observations from this population. The five histograms resemble one another in a general way, but there are also some obvious differences. The population histogram rises to a peak and then declines smoothly, whereas the sample histograms tend to have more peaks, valleys, and gaps. Although the population data set contained an observation of 11, none of the four samples did. In fact, in the first two samples, the largest observations were 7 and 8, respectively. In Chapters 8–15 we will see how sampling variability can be described and taken into account when we use sample data to draw conclusions about a population.

FIGURE 3.29

Comparison of population and sample histograms for number of accidents.



Cumulative Relative Frequencies and Cumulative Relative Frequency Plots

Rather than wanting to know what proportion of the data fall in a particular class interval, we often would like to determine the proportion falling below a specified value. This is easily done when the value is a class interval boundary.

Consider the following intervals and relative frequencies for carry-on luggage weight (in lbs.) for passengers on flights between Phoenix and New York City during October 2009:

Weight	0 to 5	5 to <10	10 to <15	15 to <20	...
Relative frequency	0.05	0.10	0.18	0.25	...

Then

proportion of passengers with carry-on luggage weight less than 15 lbs.

$$\begin{aligned} &= \text{proportion in one of the first three intervals} \\ &= 0.05 + 0.10 + 0.18 \\ &= 0.33 \end{aligned}$$

Similarly,

proportion of passengers with carry-on luggage weight less than 20 lbs.

$$= 0.05 + 0.10 + 0.18 + 0.25 = 0.33 + 0.25 = 0.58$$

Each such sum of relative frequencies is called a **cumulative relative frequency**. Notice that the cumulative relative frequency 0.58 is the sum of the previous cumulative relative frequency 0.33 and the “current” relative frequency 0.25. The use of cumulative relative frequencies is illustrated in Example 3.19.

Example 3.19 Albuquerque Rainfall

Understand the context ▶

The National Climatic Data Center has been collecting weather data for many years. Annual rainfall totals for Albuquerque, New Mexico, from 1950 to 2008 (ncdc.noaa.gov/oa/climate/research/cag3/city.html) were used to construct the relative frequency distribution shown in Table 3.8. The table also contains a column of cumulative relative frequencies.

TABLE 3.8 Relative Frequency Distribution for Albuquerque Rainfall Data with Cumulative Relative Frequencies

Annual Rainfall (inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
4 to <5	3	0.052	0.052
5 to <6	6	0.103	0.155 = 0.052 + 0.103
6 to <7	5	0.086	0.241 = 0.052 + 0.103 + 0.086 or 0.155 + 0.086
7 to <8	6	0.103	0.344
8 to <9	10	0.172	0.516
9 to <10	4	0.069	0.585
10 to <11	12	0.207	0.792
11 to <12	6	0.103	0.895
12 to <13	3	0.052	0.947
13 to <14	3	0.052	0.999

The proportion of years with annual rainfall less than 10 inches is 0.585, the cumulative relative frequency for the 9 to <10 interval. What about the proportion of years with annual rainfall less than 8.5 inches? Because 8.5 is not the endpoint of one of the intervals in the frequency distribution, we can only estimate this from the information given. The value 8.5

is halfway between the endpoints of the 8 to 9 interval, so it is reasonable to estimate that half of the relative frequency of 0.172 for this interval belongs in the 8 to 8.5 range. Then

$$\left(\begin{array}{l} \text{estimate of proportion of} \\ \text{years with rainfall less} \\ \text{than 8.5 inches} \end{array} \right) = 0.052 + 0.103 + 0.086 + 0.103 + \frac{1}{2}(0.172) = 0.430$$

This proportion could also have been estimated using the cumulative relative frequencies as

$$\left(\begin{array}{l} \text{estimate of proportion of} \\ \text{years with rainfall less} \\ \text{than 8.5 inches} \end{array} \right) = 0.344 + \frac{1}{2}(0.172) = 0.430$$

Similarly, since 11.25 is one-fourth of the way between 11 and 12,

$$\left(\begin{array}{l} \text{estimate of proportion of} \\ \text{years with rainfall less} \\ \text{than 11.25 inches} \end{array} \right) = 0.792 + \frac{1}{4}(0.103) = 0.818$$

A **cumulative relative frequency plot** is a graph of the cumulative relative frequencies against the upper endpoint of the corresponding interval. The pairs

(upper endpoint of interval, cumulative relative frequency)

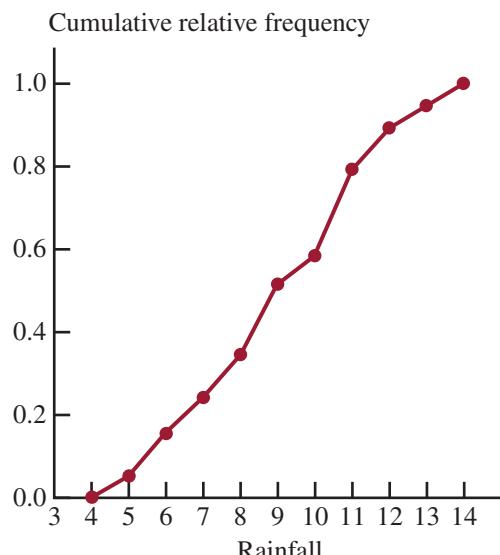
are plotted as points in two dimensions with the upper endpoint of the interval on the x axis and the cumulative relative frequency on the y axis. Then, successive points in the plot are connected by a line segment. For the rainfall data of Example 3.19, the plotted points would be

(5, 0.052)	(6, 0.155)	(7, 0.241)	(8, 0.344)	(9, 0.516)
(10, 0.585)	(11, 0.792)	(12, 0.895)	(13, 0.947)	(14, 0.999)

One additional point, the pair (lower endpoint of first interval, 0), is also included in the plot (for the rainfall data, this would be the point (4, 0)). Then, points are connected by line segments. Figure 3.30 shows the cumulative relative frequency plot for the rainfall data.

FIGURE 3.30

Cumulative relative frequency plot for the rainfall data of Example 3.19.



The cumulative relative frequency plot can be used to obtain approximate answers to questions such as

What proportion of the observations is smaller than a particular value?

and

What value separates the smallest p percent from the larger values?

For example, to determine the approximate proportion of years with annual rainfall less than 9.5 inches, we would follow a vertical line up from 9.5 on the x -axis and then read across to the y -axis to obtain the corresponding relative frequency, as illustrated in Figure 3.31(a). Approximately 0.55, or 55%, of the years had annual rainfall less than 9.5 inches.

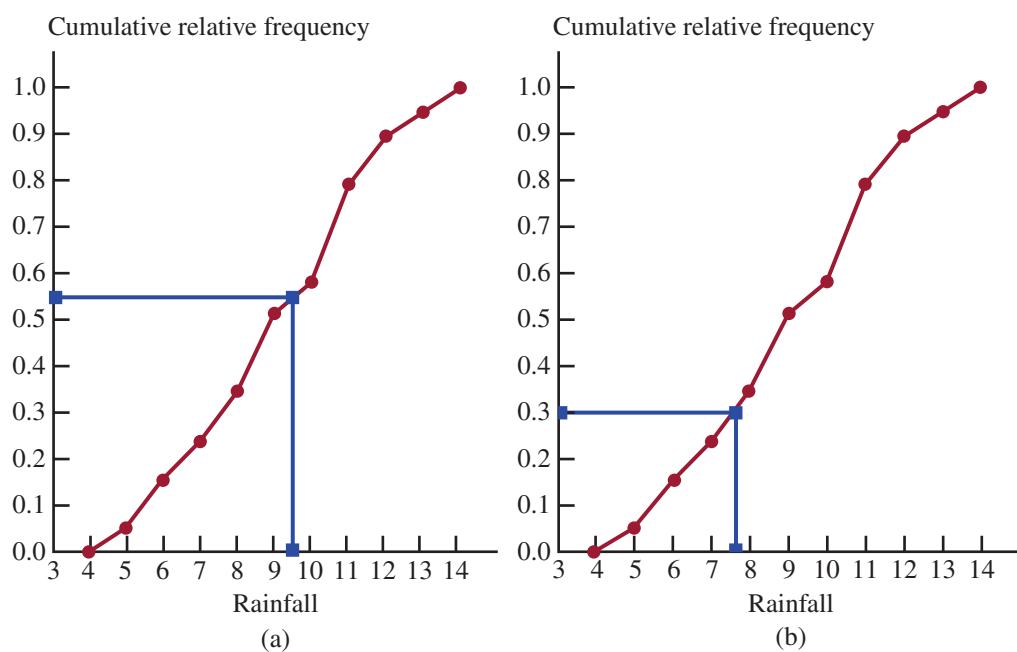


FIGURE 3.31

Using the cumulative relative frequency plot.

- (a) Determining the approximate proportion of years with annual rainfall less than 9.5 inches.
- (b) Finding the amount of rainfall that separates the 30% of years with the lowest rainfall from the 70% with higher rainfall.

Similarly, to find the approximate amount of rainfall that separates the 30% of years with the smallest annual rainfall from years with higher rainfall, start at 0.30 on the cumulative relative frequency axis and move across and then down to find the corresponding rainfall amount, as shown in Figure 3.31(b). Approximately 30% of the years had annual rainfall of less than 7.6 inches.

EXERCISES 3.24 - 3.39

● Data set available online

- 3.24** The data in the accompanying table are from the [Organization for Economic Co-operation and Development \(data.oecd.org/eduatt/population-with-tertiary-education.htm, retrieved February 18, 2018](http://Organization for Economic Co-operation and Development (data.oecd.org/eduatt/population-with-tertiary-education.htm, retrieved February 18, 2018),

Entries in the table are the percentage of 25- to 34-year-old people who have completed a 4-year college degree for 27 countries in 2016.

Country	Percentage of people age 25 to 34 with 4-year degree	Country	Percentage of people age 25 to 34 with 4-year degree
Australia	49.3	Italy	25.6
Austria	39.7	Japan	60.1
Canada	60.6	Mexico	21.8
Colombia	28.1	New Zealand	43.4
Costa Rica	28.9	Norway	48.6
Czech Republic	32.6	Poland	43.5
Denmark	45.9	Portugal	35.0
Finland	41.1	Spain	41.0
France	44.0	Sweden	47.2
Germany	30.5	Switzerland	48.8
Greece	41.0	Turkey	30.5
Hungary	30.4	United Kingdom	52.0
Iceland	43.3		
Israel	47.4	United States	47.5

- a. Construct a histogram of these data using the class intervals 20 to < 30, 30 to < 40, ..., 60 to < 70. (Hint: See Example 3.16.)
- b. Write a few sentences describing the shape, center, and variability of the distribution.
- 3.25** ● The accompanying data on annual maximum wind speed (in meters per second) in Hong Kong for each year in a 45-year period were given in an article that appeared in the journal *Renewable Energy* (March, 2007).
- | | | | | | | | | |
|------|------|------|------|------|------|------|------|------|
| 30.3 | 39.0 | 33.9 | 38.6 | 44.6 | 31.4 | 26.7 | 51.9 | 31.9 |
| 27.2 | 52.9 | 45.8 | 63.3 | 36.0 | 64.0 | 31.4 | 42.2 | 41.1 |
| 37.0 | 34.4 | 35.5 | 62.2 | 30.3 | 40.0 | 36.0 | 39.4 | 34.4 |
| 28.3 | 39.1 | 55.0 | 35.0 | 28.8 | 25.7 | 62.7 | 32.4 | 31.9 |
| 37.5 | 31.5 | 32.0 | 35.5 | 37.5 | 41.0 | 37.5 | 48.6 | 28.1 |
- a. Use the annual maximum wind speed data to construct a histogram.
- b. Is the histogram approximately symmetric, positively skewed, or negatively skewed?
- c. Would you describe the histogram as unimodal, bimodal, or multimodal?
- 3.26** ● The accompanying relative frequency table is based on data from the *2016 College Bound Seniors Report* (collegeboard.org, retrieved February 18, 2018).
- a. Construct a relative frequency histogram for SAT critical reading score for males.
- b. Construct a relative frequency histogram for SAT critical reading score for females.

- c. Based on the histograms from Parts (a) and (b), write a few sentences commenting on the similarities and differences in the distribution of SAT critical reading scores for males and females.

Score on SAT Critical Reading Exam	Relative Frequency for Males	Relative Frequency for Females
200 to < 300	0.049	0.038
300 to < 400	0.151	0.156
400 to < 500	0.308	0.332
500 to < 600	0.285	0.286
600 to < 700	0.155	0.144
700 to < 800	0.052	0.044

- 3.27** ● The data in the accompanying table represents the percentage of workers who are members of a union for each U.S. state and the District of Columbia (*AARP Bulletin*, September 2009).

State	% of Workers who Belong to a Union	State	% of Workers who Belong to a Union
Alabama	9.8	Nebraska	8.3
Alaska	23.5	Nevada	16.7
Arizona	8.8	New Hampshire	3.5
Arkansas	5.9	New Jersey	6.1
California	18.4	New Mexico	10.6
Colorado	8.0	New York	18.3
Connecticut	16.9	North Carolina	7.2
Delaware	12.2	North Dakota	24.9
District of Columbia	13.4	Ohio	14.2
Florida	6.4	Oklahoma	6.6
Georgia	3.7	Oregon	16.6
Hawaii	24.3	Pennsylvania	15.4
Idaho	7.1	Rhode Island	16.5
Illinois	16.6	South Carolina	3.9
Indiana	12.4	South Dakota	5.0
Iowa	10.6	Tennessee	5.5
Kansas	7.0	Texas	4.5
Kentucky	8.6	Utah	5.8
Louisiana	4.6	Vermont	4.1
Maine	12.3	Massachusetts	12.6
Maryland	15.7	Michigan	18.8
Michigan	18.8	Minnesota	16.1
Minnesota	16.1	Mississippi	5.3
Mississippi	5.3	Missouri	11.2
Missouri	11.2	Wisconsin	15.0
Montana	12.2	Wyoming	7.7

- a.** Construct a histogram of these data using class intervals of 0 to <5 , 5 to <10 , 10 to <15 , 15 to <20 , and 20 to <25 .
- b.** Construct a dotplot of these data. Comment on the interesting features of the plot. (Hint: Dotplots were covered in Section 1.4.)
- c.** For this data set, which is a more informative graphical display—the dotplot from Part (b) or the histogram constructed in Part (a)? Explain.
- 3.28** Construct a histogram for the data in the previous exercise using about twice as many class intervals. Use 2.5 to <5 as the first class interval. Write a few sentences that explain why this histogram does a better job of displaying this data set than the histogram in the previous exercise.
- 3.29** The following two relative frequency distributions are based on data that appeared in *The Chronicle of Higher Education* (August 23, 2013). The data are from a survey of students at four-year colleges. One relative frequency distribution is for the number of hours spent online at social network sites in a typical week. The second relative frequency distribution is for the number of hours spent playing video and computer games in a typical week.

Number of Hours on Social Networks	Relative Frequency
0 to <1	0.234
1 to <6	0.512
6 to <21	0.211
21 or more	0.044

Number of Hours Playing Video and Computer Games	Relative Frequency
0 to <1	0.621
1 to <6	0.252
6 to <21	0.108
21 or more	0.020

- a.** Construct a histogram for the social media data. For purposes of constructing the histogram, assume that none of the students in the sample spent more than 40 hours on social media in a typical week and that the last interval can be regarded as 21 to <40 . Be sure to use the density scale when constructing the histogram. (Hint: See Example 3.17.)
- b.** Construct a histogram for the video and computer game data. Use the same scale that you used for the histogram in Part (a) so that it will be easy to compare the two histograms.
- c.** Comment on the similarities and differences in the histograms from Parts (a) and (b).

- 3.30** U.S. Census data for San Luis Obispo County, California, were used to construct the following relative frequency distribution for commute time (in minutes) of working adults in 2015 (datausa.io/profile/geo/san-luis-obispo-paso-robles-ca-metro-area/#housing, retrieved February 18, 2018) and so are only approximate:

Commute Time	Relative Frequency
0 to <5	0.056
5 to <10	0.156
10 to <15	0.177
15 to <20	0.155
20 to <25	0.147
25 to <30	0.061
30 to <35	0.121
35 to <40	0.015
40 to <45	0.024
45 to <60	0.040
60 to <90	0.030
90 to <120	0.018

- a.** Notice that not all intervals in the frequency distribution are equal in width. Why do you think that unequal width intervals were used?
- b.** Construct a table that adds a density column to the given relative frequency distribution. (Hint: See Example 3.17.)
- c.** Use the densities computed in Part (b) to construct a histogram for this data set. (The web site referenced earlier actually displays an incorrectly drawn histogram based on relative frequencies rather than densities!) Write a few sentences commenting on the important features of the histogram.

- 3.31** Use the commute time data given in the previous exercise to complete the following:

- a.** Calculate the cumulative relative frequencies, and construct a cumulative relative frequency plot.
- b.** Use the cumulative relative frequency plot constructed in Part (a) to answer the following questions.
- Approximately what proportion of commute times were less than 50 minutes?
 - Approximately what proportion of commute times were greater than 22 minutes?
 - What is the approximate commute time value that separates the shortest 50% of commute times from the longest 50%?

- 3.32** The report “**Trends in College Pricing 2012**” (collegeboard.com) included the information in

the accompanying relative frequency distributions for public and for private not-for-profit four-year college students.

Tuition and Fees	Public Four-Year College Students	Private Not- for-Profit Four-Year College Students
	Proportion of Students (Relative Frequency)	Proportion of Students (Relative Frequency)
0 to < 3,000	0.009	0.000
3,000 to < 6,000	0.107	0.066
6,000 to < 9,000	0.436	0.011
9,000 to < 12,000	0.199	0.027
12,000 to < 15,000	0.124	0.032
15,000 to < 18,000	0.033	0.030
18,000 to < 21,000	0.027	0.052
21,000 to < 24,000	0.019	0.074
24,000 to < 27,000	0.015	0.103
27,000 to < 30,000	0.020	0.094
30,000 to < 33,000	0.005	0.103
33,000 to < 36,000	0.003	0.103
36,000 to < 39,000	0.002	0.066
39,000 to < 42,000	0.002	0.080
42,000 to < 45,000	0.000	0.136
45,000 to < 48,000	0.000	0.022

- a. Construct a relative frequency histogram for tuition and fees for students at public four-year colleges. Write a few sentences describing the distribution of tuition and fees, commenting on center, variability, and shape.
 - b. Construct a relative frequency histogram for tuition and fees for students at private not-for-profit four-year colleges. Use the same scale for the vertical and horizontal axes as you used for the histogram in Part (a). Write a few sentences describing the distribution of tuition and fees for students at private not-for-profit four-year colleges.
 - c. Write a few sentences describing the differences in the distributions.
- 3.33** An exam is given to students in an introductory statistics course. What is likely to be true of the shape of the histogram of scores if:
- a. the exam is quite easy?
 - b. the exam is quite difficult?
 - c. half the students in the class have had calculus, the other half have had no prior college math courses, and the exam emphasizes mathematical manipulation?
- Explain your reasoning in each case.

- 3.34** The accompanying frequency distribution summarizes data on the number of times smokers who had successfully quit smoking attempted to quit before their final successful attempt ("Demographic Variables, Smoking Variables, and Outcome Across Five Studies," *Health Psychology* [2007]: 278–287).

Number of Attempts	Frequency
0	778
1	306
2	274
3–4	221
5 or more	238

Assume that no one had made more than 10 unsuccessful attempts, so that the last entry in the frequency distribution can be regarded as 5–10 attempts. Summarize this data set using a histogram. Be careful—the class intervals are not all the same width, so a density scale should be used for the histogram. Also remember that for a discrete variable, the bar for 1 will extend from 0.5 to 1.5. Think about what this will mean for the bars for the 3–4 group and the 5–10 group.

- 3.35** Example 3.19 used annual rainfall data for Albuquerque, New Mexico, to construct a relative frequency distribution and cumulative relative frequency plot. The National Climate Data Center also gave the accompanying annual rainfall (in inches) for Medford, Oregon, from 1950 to 2008.

28.84 20.15 18.88 25.72 16.42 20.18 28.96 20.72 23.58 10.62
 20.85 19.86 23.34 19.08 29.23 18.32 21.27 18.93 15.47 20.68
 23.43 19.55 20.82 19.04 18.77 19.63 12.39 22.39 15.95 20.46
 16.05 22.08 19.44 30.38 18.79 10.89 17.25 14.95 13.86 15.30
 13.71 14.68 15.16 16.77 12.33 21.93 31.57 18.13 28.87 16.69
 18.81 15.15 18.16 19.99 19.00 23.97 21.99 17.25 14.07

- a. Construct a relative frequency distribution for the Medford rainfall data.
- b. Use the relative frequency distribution of Part (a) to construct a histogram. Describe the shape of the histogram.

- 3.36** Use the relative frequency distribution constructed in the previous exercise to answer the following questions.

- a. Construct a cumulative relative frequency plot for the Medford rainfall data.
- b. Use the cumulative relative frequency plot of Part (a) to answer the following questions:
 - i. Approximately what proportion of years had annual rainfall less than 15.5 inches?
 - ii. Approximately what proportion of years had annual rainfall less than 25 inches?
 - iii. Approximately what proportion of years had annual rainfall between 17.5 and 25 inches?

- 3.37** The authors of the paper “**Myeloma in Patients Younger than Age 50 Years Presents with More Favorable Features and Shows Better Survival**” (*Blood* [2008]: 4039–4047) studied patients who had been diagnosed with stage 2 multiple myeloma prior to the age of 50. For each patient who received high dose chemotherapy, the number of years that the patient lived after the therapy (survival time) was recorded. The cumulative relative frequencies in the accompanying table were approximated from survival graphs that appeared in the paper.

Years Survived	Cumulative Relative Frequency
0 to <2	0.10
2 to <4	0.52
4 to <6	0.54
6 to <8	0.64
8 to <10	0.68
10 to <12	0.70
12 to <14	0.72
14 to <16	1.00

- a. Use the given information to construct a cumulative relative frequency plot.
- b. Use the cumulative relative frequency plot from Part (a) to answer the following questions:
 - i. What is the approximate proportion of patients who lived fewer than 5 years after treatment?

- ii. What is the approximate proportion of patients who lived fewer than 7.5 years after treatment?
- iii. What is the approximate proportion of patients who lived more than 10 years after treatment?

- 3.38** Use the cumulative relative frequencies given in the previous exercise to complete the following:
- a. Calculate the relative frequencies for each class interval and construct a relative frequency distribution.
 - b. Summarize the survival time data using a histogram.
 - c. Based on the histogram, write a few sentences describing survival time of the stage 2 myeloma patients in this study.
 - d. What additional information would you need in order to decide if it is reasonable to generalize conclusions about survival time from the group of patients in the study to all patients younger than 50 years old who are diagnosed with multiple myeloma and who receive high dose chemotherapy?

- 3.39** Using the five class intervals 100 to 120, 120 to 140, . . . , 180 to 200, devise a frequency distribution based on 70 observations whose histogram could be described as follows:
- a. symmetric
 - b. bimodal
 - c. positively skewed
 - d. negatively skewed

SECTION 3.4 Displaying Bivariate Numerical Data

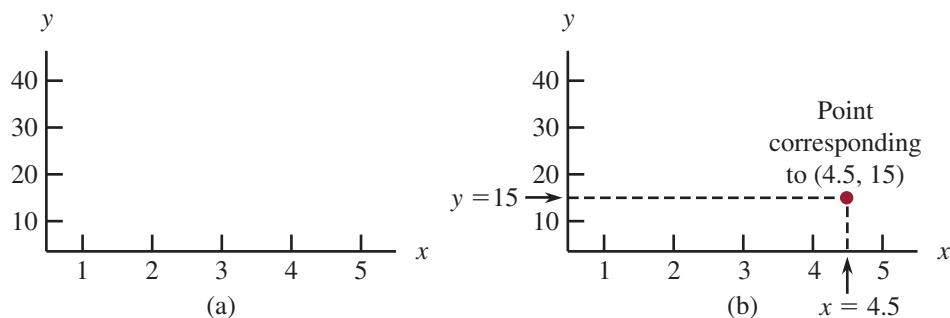
A bivariate data set consists of measurements or observations on two variables, x and y . For example, x might be the distance from a highway and y the lead content of soil at that distance. When both x and y are numerical variables, each observation consists of a pair of numbers, such as (14, 5.2) or (27.63, 18.9). The first number in a pair is the value of x , and the second number is the value of y .

An unorganized list of bivariate data provides little information about the distribution of either the x values or the y values separately and even less information about how the two variables are related to one another. Just as graphical displays can be used to summarize univariate data, they can also be used to summarize bivariate data. The most important graph based on bivariate numerical data is a **scatterplot**.

In a scatterplot each observation (pair of numbers) is represented by a point on a rectangular coordinate system, as shown in Figure 3.32(a). The horizontal axis is identified with values of x and is scaled so that any x value can be easily located. Similarly, the vertical or y -axis is marked for easy location of y values. The point corresponding to any particular (x, y) pair is placed where a vertical line from the value on the x -axis meets a horizontal line from the value on the y -axis. Figure 3.32(b) shows the point representing the observation (4.5, 15). It is above 4.5 on the horizontal axis and to the right of 15 on the vertical axis.

FIGURE 3.32

Constructing a scatterplot:
 (a) rectangular coordinate system;
 (b) point corresponding to $(4.5, 15)$.



- Data set available online

Understand the context ➤

Example 3.20 Olympic Figure Skating

- Consider the data ➤
- Do tall skaters have an advantage when it comes to earning high artistic scores in figure skating competitions? Data on $x = \text{height}$ (in inches) and $y = \text{components score}$ (based on artistry, interpretation, and presentation) in the free skate for both male and female singles skaters at the 2018 Winter Olympics are shown in the accompanying table. (Data from olympic.org/pyeongchang-2018/results/en/figure-skating/results-men-single-skating-fnl-000100-.htm and olympic.org/pyeongchang-2018/results/en/figure-skating/results-ladies-single-skating-fnl-000100-.htm, retrieved March 4, 2018.)

Name	Sex	Height (in)	Components Score	Name	Sex	Height (in)	Components Score
Yuzuru HANYU	M	67	96.62	Alina ZAGITOVA	F	61	75.03
Shoma UNO	M	63	92.72	Evgenia MEDVEDEVA	F	63	77.47
Javier FERNANDEZ	M	68	96.14	Kaetlyn OSMOND	F	64	75.65
JIN Boyang	M	67	85.76	Satoko MIYAHARA	F	60	71.24
Nathan CHEN	M	65	87.44	Carolina KOSTNER	F	67	75.65
Vincent ZHOU	M	67	79.92	Kaori SAKAMOTO	F	62	68.11
Dmitri ALIEV	M	70	85.14	CHOI Dabin	F	61	62.75
Mikhail KOLYADA	M	69	87.94	Maria SOTSKOVA	F	68	67.30
Patrick CHAN	M	66	91.86	Bradie TENNELL	F	66	62.93
Adam RIPPON	M	67	86.94	Mirai NAGASU	F	63	62.05
Alexei BYCHENKO	M	69	83.80	Karen CHEN	F	60	64.10
Keegan MESSING	M	64	85.44	Elizabet TURSYNBAEVA	F	59	58.80
Daniel SAMOHIN	M	69	81.72	KIM Hanul	F	66	54.35
Jorik HENDRICKX	M	69	82.42	Nicole RAJICOVA	F	64	56.80
CHA Junhwan	M	69	81.22	Gabrielle DALEMAN	F	60	61.75
Michal BREZINA	M	68	84.34	Loena HENDRICKX	F	63	55.99
Misha GE	M	68	86.08	Kailani CRAINE	F	63	53.95
Keiji TANAKA	M	68	81.14	Nicole SCHOTT	F	65	56.58
Deniss VASILJEVS	M	68	80.64	Mae Berenice MEITE	F	66	52.12
Brendan KERRY	M	68	77.42	Emmi PELTONEN	F	64	56.45
Matteo RIZZO	M	67	75.92	Alexia PAGANINI	F	66	50.06
Paul FENTZ	M	70	72.84	LI Xiangning	F	57	51.41
YAN Han	M	67	79.58	Ivett TOTH	F	61	50.39
Morisi KVITELASHVILI	M	71	70.66	Isadora WILLIAMS	F	61	51.05

Figure 3.33(a) gives a scatterplot of the data. Looking at the data and the scatterplot, we can see that

Interpret the results ➤

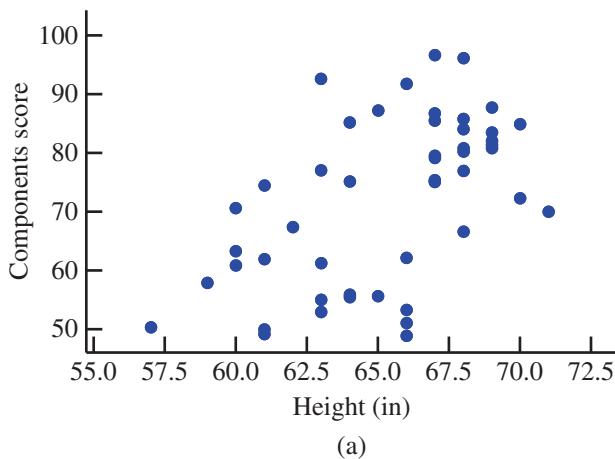
1. Several observations have identical x values but different y values (for example, $x = 67$ inches for both Yuzuru Hanyu and Adam Rippon, but Hanyu's component score was 96.62 and Rippon's component score was 86.94). This shows that the value of y is *not* determined *solely* by the value of x but by various other factors as well.
2. At any given height there is quite a bit of variability in artistic score. For example, for those skaters with height 67 inches, component scores ranged from a low of about 76 to a high of about 97.
3. There appears to be a tendency for components score to increase as height increases. However, there is not a strong linear relationship between height and components score. The points in the scatterplot do not cluster tightly around a line.

The data set used to construct the scatterplot included data for both male and female skaters. Figure 3.33(b) shows a scatterplot of the (height, components score) pairs with observations for male skaters shown in red and observations for female skaters shown in blue. Not surprisingly, the female skaters tend to be shorter than the male skaters (the observations for females tend to be concentrated toward the left side of the scatterplot). Careful examination of this plot shows that while there was a weak linear pattern in the combined (male and female) data set, there does not appear to be a relationship between height and components score for female skaters.

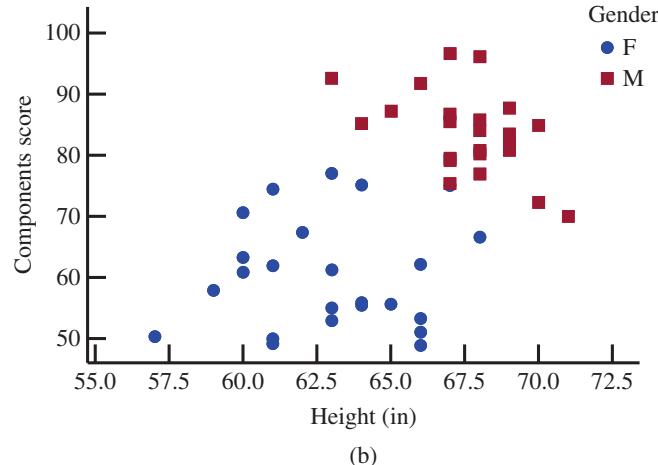
Figures 3.33(c) and (d) show separate scatterplots for the male and female skaters, respectively. It is interesting to note that it appears that for male skaters, higher components scores seem to be associated with smaller height values, but for women there does not appear to be a relationship between height and components score.

FIGURE 3.33

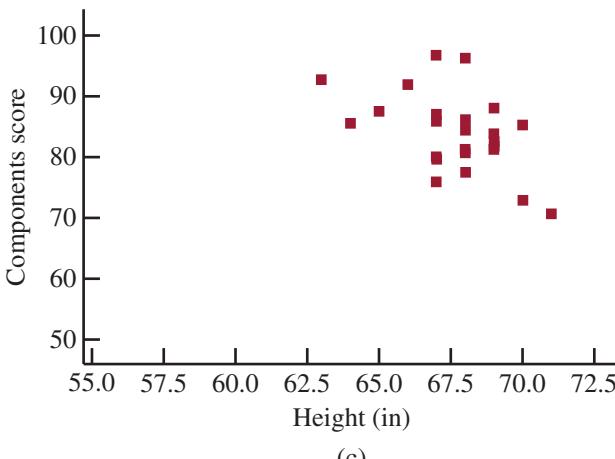
Scatterplots for the data of Example 3.20:
 (a) scatterplot of data;
 (b) scatterplot of data with observations for males and females distinguished by color;
 (c) scatterplot for male skaters;
 (d) scatterplot for female skaters.



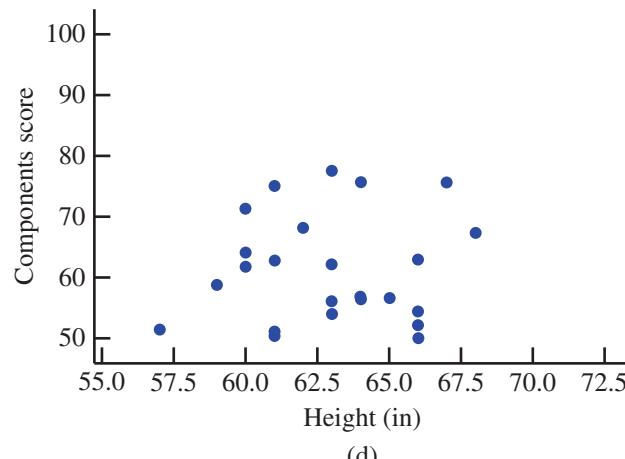
(a)



(b)



(c)



(d)

The relationship between height and components score for men is not evident in the scatterplot of the combined data, which seems to indicate that higher components scores are paired with greater heights.

The horizontal and vertical axes in the scatterplots of Figure 3.33 do not intersect at the point (0, 0). In many data sets, the values of x or of y or of both variables differ considerably from 0 relative to the ranges of the values in the data set. For example, a study of how air conditioner efficiency is related to maximum daily outdoor temperature might involve observations at temperatures of $80^\circ, 82^\circ, \dots, 98^\circ, 100^\circ$. In such cases, the plot will be more informative if the axes intersect at some point other than (0, 0) and are marked accordingly. This is illustrated in Example 3.21.

Example 3.21 Emotional Health and Work Environment

Understand the context ➤

The accompanying table provides data on a measure of emotional health and a measure of the quality of the work environment for 13 different occupations. These data are from an article titled “U.S. Teachers Love Their Lives, but Struggle in the Workplace” that appeared on the Gallup web site (gallup.com/poll/161516/teachers-love-lives-struggle-workplace.aspx?g_source, retrieved April 17, 2017).

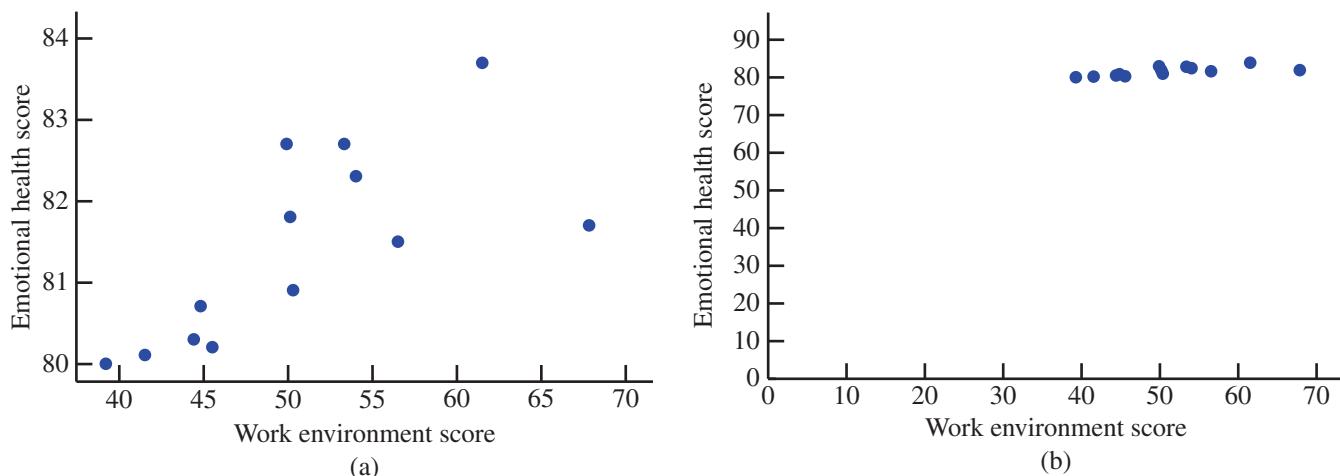
Consider the data ➤

Occupation	Emotional Health Score	Work Environment Score
Physician	83.7	61.5
Teacher	82.7	49.9
Farming, fishing, or forestry	82.7	53.3
Professional	82.3	54.0
Nurse	81.8	50.1
Business owner	81.7	67.8
Manager, executive, or official	81.5	56.5
Construction or mining	80.9	50.3
Installation or repair	80.7	44.8
Clerical or office	80.3	44.4
Sales	80.2	45.5
Manufacturing or production	80.1	41.5
Transportation	80.0	39.2

Figures 3.34(a) and (b) show two scatterplots of $x =$ Work Environment Score and $y =$ Emotional Health Score. The scatterplots were produced by the statistical computer package Minitab. In Figure 3.34(a), we let Minitab select the scale for both axes. Figure 3.34(b) was obtained by specifying that the axes would intersect at the point (0, 0). The second plot does not make effective use of space. The data points are more crowded together than in the first plot, and such crowding can make it difficult to see important features of the relationship between Emotional Health Score and Work Environment Score for these occupations. It is more difficult to spot the curvature that is visible in the first plot in a crowded plot. It is also more difficult to see unusual observations. For example, there is one observation that stands out as being different from the others in the plot in Figure 3.34(a). This observation (the one with the highest Work Environment Score of 67.8 and corresponds to business owners) has a lower Emotional Health Score than might have been expected for an occupation with such a high Work Environment Score. This is not noticeable in the plot where the axes cross at (0, 0).

Time Series Plots

Data sets often consist of measurements collected over time at regular intervals so that we can learn about change over time. For example, stock prices, sales figures, and other socio-economic indicators might be recorded on a weekly or monthly basis. A **time series plot**

**FIGURE 3.34**

Minitab scatterplots of the data in Example 3.21:
(a) scale for both axes selected by Minitab;

(b) axes intersect at (0, 0).

(sometimes also called a time plot) is a graph of data collected over time that can be used to identify trends or patterns that might be of interest.

A time series plot can be constructed by thinking of the data set as a bivariate data set, where y is the variable observed and x is the time at which the observation was made. These (x, y) pairs are plotted as a scatterplot. Consecutive observations are then connected by a line segment. This helps us to see trends over time.

Example 3.22 Exercise on the Rise?

Understand the context ➤

Gallup conducts frequent polls in which large samples of adult Americans are asked how often they exercise. The article “[So Far in 2015, More Americans Exercising Frequently](http://gallup.com/poll/184403/far-2015-americans-exercising-frequently.aspx?g_source)” (gallup.com/poll/184403/far-2015-americans-exercising-frequently.aspx?g_source, retrieved April 17, 2018) used information from these polls to estimate that during the first half of 2015, on average 52.5% of Americans exercised for 30 minutes or more at least 3 days a week. The article also provided estimates for the years 2008 to 2015, as shown in the accompanying table.

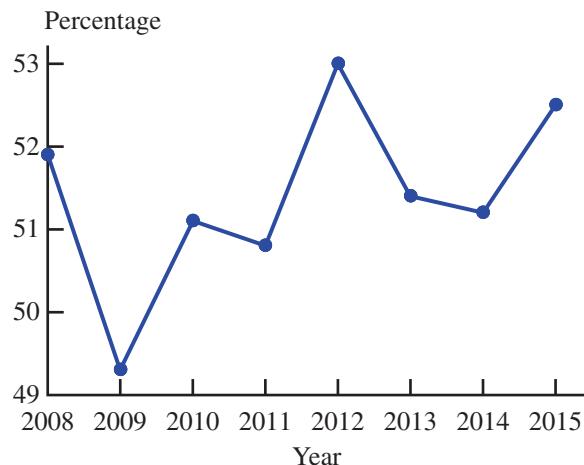
Consider the data ➤

Year	Percentage	Year	Percentage
2008	51.9	2012	53.0
2009	49.3	2013	51.4
2010	51.1	2014	51.2
2011	50.8	2015	52.5

Figure 3.35 shows a time series plot of these data. Notice that the eight (year, percentage) pairs have been plotted and that these points have been connected by line segments.

FIGURE 3.35

Time series plot of percentage of Americans who exercise for 30 minutes or more at least 3 times per week.



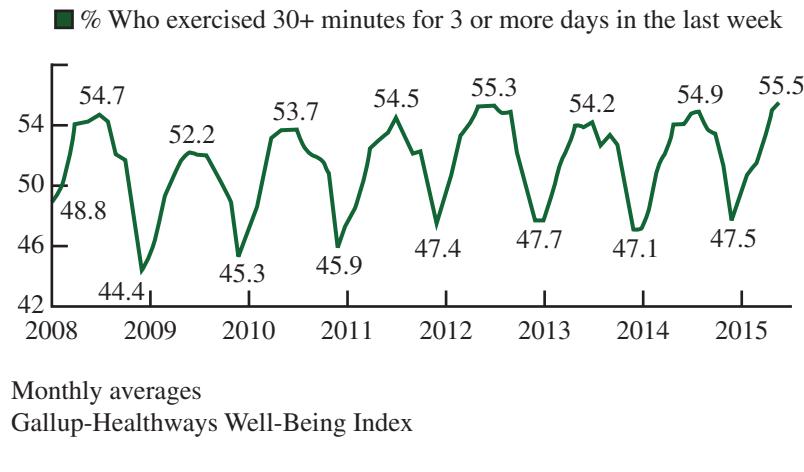
Interpret the results ➤

This makes it easier to see any trend over time. You can see from the time series plot that the percentage has not steadily increased year to year, although there does appear to be a general upward trend following the drop that occurred in 2009.

The article also included a time series plot that was based on monthly estimates of the percentage exercising 30 minutes or more at least 3 times per week. Figure 3.36 is similar to the plot that appeared in the article. In this plot, in addition to the general increasing trend from year to year, you can also see a pattern that repeats each year, with the percentage tending to increase during the first half of each year and decrease in the second half of each year.

FIGURE 3.36

Time series of percentage exercising based on monthly data.

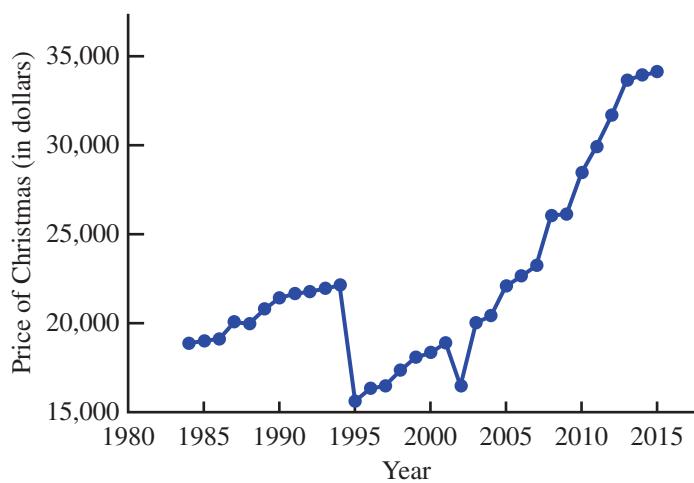
**Example 3.23 The Cost of Christmas**

Understand the context ➤

The Christmas Price Index is calculated each year by PNC Advisors, and it is a humorous look at the cost of giving all of the gifts described in the popular Christmas song “The Twelve Days of Christmas.” The year 2015 was the most costly year since the index began, with the “cost of Christmas” at \$34,131. Historical data from the PNC web site (pncchristmaspriceindex.com) were used to construct the time series plot of Figure 3.37.

FIGURE 3.37

Time series plot for the Christmas Price Index data of Example 3.23.



Interpret the results ➤

The plot shows an upward trend in the index from 1984 until 1993. There has been a clear upward trend in the index since 1995. You can visit the web site to see individual time series plots for each of the twelve gifts that are used to determine the Christmas Price Index (a partridge in a pear tree, two turtle doves, etc.). See if you can figure out what caused the dramatic decline from 1994 to 1995.

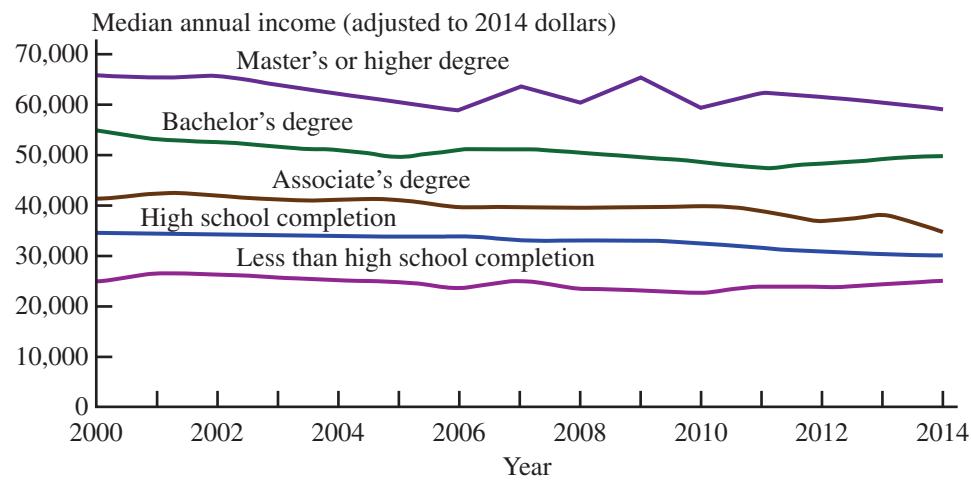
Example 3.24 Education and Income—Stay in School!

Consider the data ➤

The time series plot in Figure 3.38 is similar to one appearing on the web site of the [National Center for Education Statistics \(nces.ed.gov/programs/coe/indicator_cba.asp\)](http://nces.ed.gov/programs/coe/indicator_cba.asp), retrieved April 17, 2017). It shows the change over time in median annual earnings by education level. (The median annual earnings is the value for which half of the population earns less and half earns more. For example, in the year 2000, about half of those with bachelor's degrees earned less than \$55,000 per year, and half of those with bachelor's degrees earned more than \$55,000.) From this plot, you can see that the median annual earnings varied more from year to year for those with master's or higher degrees, but that the large gap between the median earnings for those who only completed high school and those with a college degree has remained about the same since the year 2000.

FIGURE 3.38

Time series plots of median annual earnings by education level.


EXERCISES 3.40 - 3.47

● Data set available online

- 3.40** The accompanying table gives data from a survey of new car owners conducted by J.D. Power and Associates (USA TODAY, usatoday.com, March 29, 2016). For each brand of car sold in the United States, data on a quality rating (defects per 100 cars, so lower numbers indicate higher quality) and a customer satisfaction rating (called the APEAL rating) are given in the accompanying table. The APEAL rating is a score between 0 and 1000, with higher values indicating greater satisfaction.

Brand	Quality Rating	APEAL Rating
Acura	64	814
Audi	80	858
BMW	76	849
Buick	74	792
Cadillac	58	826
Chevrolet	64	791
Chrysler	58	783
Dodge	58	790
Fiat	38	768

(continued)

Brand	Quality Rating	APEAL Rating
Ford	66	789
GMC	60	791
Honda	71	783
Hyundai	70	804
Infiniti	63	826
Jeep	43	762
Kia	72	791
Land Rover	55	853
Lexus	76	844
Lincoln	65	835
Mazda	74	790
Mercedes-Benz	67	842
MINI	68	795
Mitsubishi	51	748
Nissan	63	786
Porsche	76	882
Scion	65	779
Subaru	78	766
Toyota	72	783
Volkswagen	67	796
Volvo	69	812

- Construct a scatterplot of $x = \text{quality rating}$ and $y = \text{APEAL rating}$. (Hint: See Example 3.21.)
- Does customer satisfaction (as measured by the APEAL rating) appear to be related to car quality? Explain.

3.41 ● *Consumer Reports Health* (consumerreports.org)

gave the accompanying data on saturated fat (in grams), sodium (in mg), and calories for 36 fast-food items.

Fat	Sodium	Calories	Fat	Sodium	Calories
2	1042	268	0	520	140
5	921	303	2.5	1120	330
3	250	260	1	240	120
2	770	660	3	650	180
1	635	180	1	1620	340
6	440	290	4	660	380
4.5	490	290	3	840	300
5	1160	360	1.5	1050	490
3.5	970	300	3	1440	380
1	1120	315	9	750	560
2	350	160	1	500	230
3	450	200	1.5	1200	370
6	800	320	2.5	1200	330
3	1190	420	3	1250	330
2	1090	120	0	1040	220
5	570	290	0	760	260
3.5	1215	285	2.5	780	220
2.5	1160	390	3	500	230

- Construct a scatterplot using $y = \text{calories}$ and $x = \text{fat}$. Does it look like there is a relationship between fat and calories? Is the relationship what you expected? Explain.
- Construct a scatterplot using $y = \text{calories}$ and $x = \text{sodium}$. Write a few sentences commenting on the difference between the relationship of calories to fat and calories to sodium.
- Construct a scatterplot using $y = \text{sodium}$ and $x = \text{fat}$. Does there appear to be a relationship between fat and sodium?
- Add a vertical line at $x = 3$ and a horizontal line at $y = 900$ to the scatterplot in Part (c). This divides the scatterplot into four regions, with some of the points in the scatterplot falling into each of the four regions. Which of the four regions corresponds to healthier fast-food choices? Explain.

3.42 The report “[Wireless Substitution: Early Release of Estimates from the National Health Interview Survey](#)” ([Center for Disease Control, 2015](#)) gave the following estimates of the percentage of homes in the United States that had only wireless phone service at 6-month intervals from June 2005 to December 2008.

Date	Percent with Only Wireless Phone Service
December 2013	41.0
June 2014	44.0
December 2014	45.4
June 2015	47.4
December 2015	48.3
June 2016	49.3
December 2016	50.8

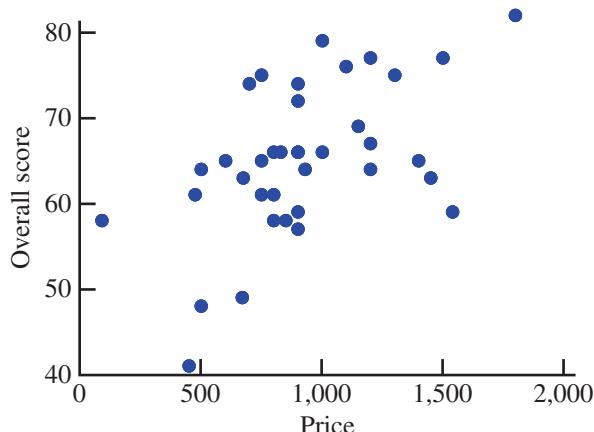
- Construct a time series plot for these data and describe the trend in the percent of homes with only wireless phone service over time. (Hint: See Example 3.23.)
- Has the percent increased at a fairly steady rate?

3.43 *Consumer Reports* rated 29 fitness trackers (such as Fitbit and Jawbone) on factors such as ease of use and accuracy of step count to obtain an overall score ([consumerreports.org, retrieved October 13, 2016](#)). The accompanying table gives price and overall score for these 29 fitness trackers.

Price	Overall Score	Price	Overall Score
260	87	220	65
150	83	60	61
200	82	100	61
200	82	100	60
150	82	150	57
150	79	100	56
130	78	70	56
100	77	100	55
176	74	120	54
180	74	100	54
125	71	125	51
120	70	100	50
100	69	100	44
90	66	22	41
200	66		

- Construct a scatterplot using $y = \text{overall score}$ and $x = \text{price}$.
- Based on the scatterplot from Part (a), does there appear to be a relationship between price and overall score? Does the scatterplot support the statement that the more expensive fitness trackers tended to receive higher overall scores?

3.44 *Consumer Reports* ([consumerreports.org](#)) rated 37 different models of laptops that were for sale in 2015. An overall score was assigned to each model based on consideration of several factors, including performance, portability, and battery life. Data on price and overall score were used to construct the following scatterplot.



Would you describe the relationship between price and overall score for these laptops as positive (overall score tends to increase as price increases) or negative (overall score tends to decrease as price increases)? Explain.

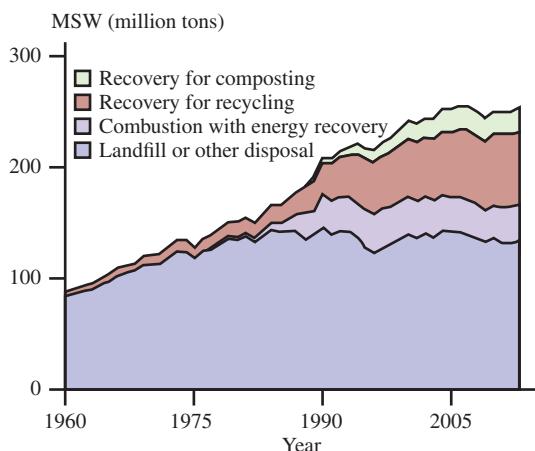
- 3.45** The Solid Waste Management section of the Environmental Protection Agency [Report on the Environment \(epa.gov/roe/\)](#), retrieved April 17, 2017) included a graph similar to accompanying graph. The report also included the following statement:

The last several decades have seen steady growth in recycling and composting, while the total amounts landfilled peaked in 1990 (145 MT) and have generally declined since then (134 MT in 2013).

Explain how the time series plot is consistent or is not consistent with the given statement.

EXHIBIT 1.

Municipal solid waste generated and managed in the U.S., 1960–2013

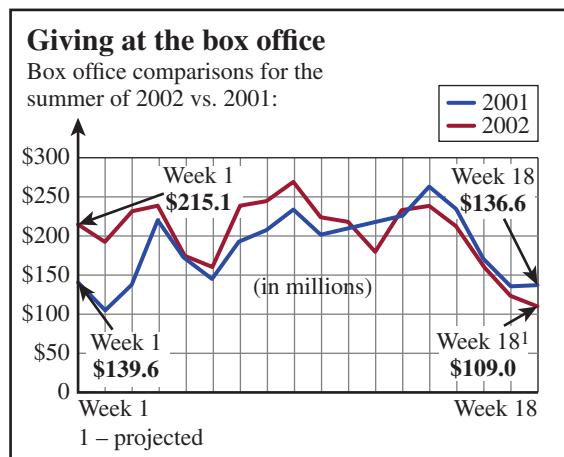


- 3.46** The report “[Daily Cigarette Use: Indicators on Children and Youth](#)” (Child Trends Data Bank, [childtrends.org/wp-content/uploads/2012/11/03_Smoking_new.pdf](#), retrieved April 17, 2017) included the accompanying data on the percentage of students who report smoking cigarettes daily, for students in grades 8, 10, and 12.

Year	Percentage of Students Who Smoke		
	Grade 8	Grade 10	Grade 12
2000	7.4	14.0	20.6
2001	5.5	12.2	19.0
2002	5.1	10.1	16.9
2003	4.5	8.9	15.8
2004	4.4	8.3	15.6
2005	4.0	7.5	13.6
2006	4.0	7.6	12.2
2007	3.0	7.2	12.3
2008	3.1	5.9	11.4
2009	2.7	6.3	11.2
2010	2.9	6.6	10.7
2011	2.4	5.5	10.3
2012	1.9	5.0	9.3
2013	1.8	4.4	8.5
2014	1.4	3.2	6.7

- Construct a time series plot for students in grade 12, and comment on any trend over time.
- Construct a time series plot that shows trends over time for each of the three grade levels. Graph each of the three time series on the same set of axes, using different colors to distinguish the different grade levels. Either label the time series in the plot or include a legend to indicate which time series corresponds to which grade level.
- Write a paragraph based on the plot from Part (b). Discuss the similarities and differences for the three grade levels.

- 3.47** The accompanying time series plot of movie box office totals (in millions of dollars) over 18 weeks of summer for both 2001 and 2002 is similar to one that appeared in [USA TODAY \(September 3, 2002\)](#):



Sources: Nielsen; USA TODAY, September 03, 2002.

Patterns that tend to repeat on a regular basis over time are called seasonal patterns. Describe any seasonal patterns that you see in the summer box office data. (Hint: Look for patterns that seem to be consistent from year to year.)

SECTION 3.5 Interpreting and Communicating the Results of Statistical Analyses

A graphical display, when used appropriately, can be a powerful tool for organizing and summarizing data. By sacrificing some of the detail of a complete listing of a data set, important features of the data distribution are more easily seen and more easily communicated to others.

Communicating the Results of Statistical Analyses

When reporting the results of a data analysis, a good place to start is with a graphical display of the data. A well-constructed graphical display is often the best way to highlight the essential characteristics of the data distribution, such as shape and variability for numerical data sets or the nature of the relationship between the two variables in a bivariate numerical data set.

For effective communication with graphical displays, some things to remember are

- Be sure to select a display that is appropriate for the given type of data.
- Be sure to include scales and labels on the axes of graphical displays.
- In comparative plots, be sure to include labels or a legend so that it is clear which parts of the display correspond to which samples or groups in the data set.
- Although it is sometimes a good idea to have axes that do not cross at $(0, 0)$ in a scatterplot, the vertical axis in a bar chart or a histogram should always start at 0 (see the cautions and limitations later in this section for more about this).
- Keep your graphs simple. A simple graphical display is much more effective than one that has a lot of extra “junk.” Most people will not spend a great deal of time studying a graphical display, so its message should be clear and straightforward.
- Keep your graphical displays honest. People tend to look quickly at graphical displays, so it is important that a graph’s first impression is an accurate and honest portrayal of the data distribution. In addition to the graphical display itself, data analysis reports usually include a brief discussion of the features of the data distribution based on the graphical display.
- For categorical data, the discussion of the graphical display might be a few sentences on the relative proportion for each category, possibly pointing out categories that were either common or rare compared to other categories.
- For numerical data sets, the discussion of the graphical display usually summarizes the information that the display provides on three characteristics of the data distribution: center or location, variability, and shape.
- For bivariate numerical data, the discussion of the scatterplot would typically focus on the nature of the relationship between the two variables used to construct the plot.
- For data collected over time, any trends or patterns in the time series plot would be described.

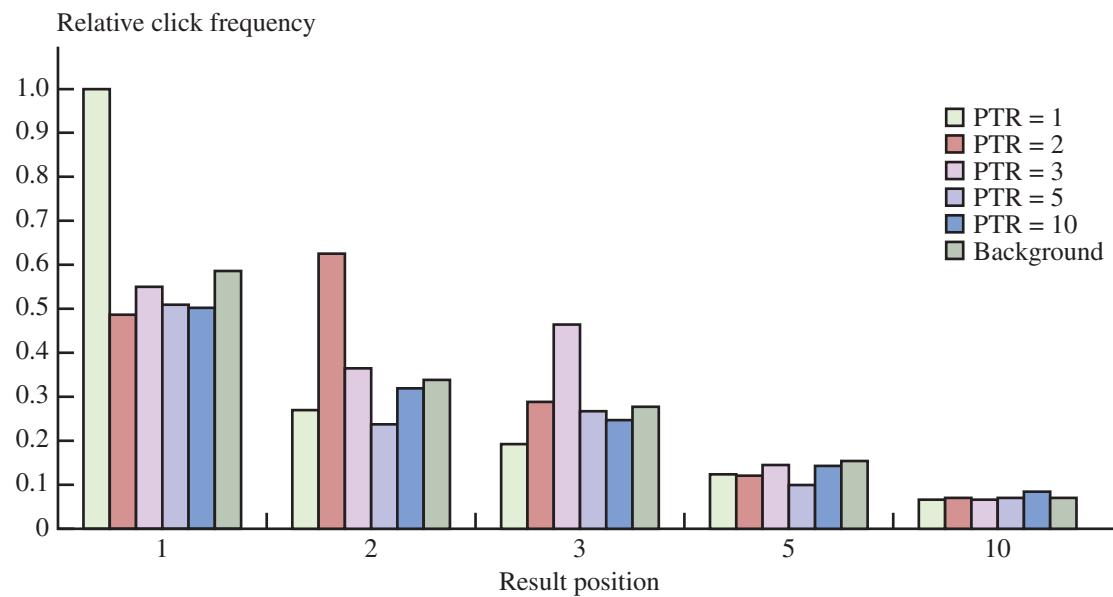
Interpreting the Results of Statistical Analyses

When someone uses a web search engine, do they rely on the ranking of the search results returned or do they first scan the results looking for the most relevant? The authors of the paper [“Learning User Interaction Models for Predicting Web Search Result Preferences” \(*Proceedings of the 29th Annual ACM Conference on Research and Development in Information Retrieval*, 2006\)](#) attempted to answer this question by observing user behavior when they varied the position of the most relevant result in the list of resources returned in response to a web search.

They concluded that people clicked more often on results near the top of the list, even when they were not relevant. They supported this conclusion with the comparative bar chart in Figure 3.39.

FIGURE 3.39

Comparative bar chart for click frequency data.



Although this comparative bar chart is a bit complicated, we can learn a great deal from this graphical display. Let's start by looking at the first group of bars. The different bars correspond to where in the list of search results the result that was considered to be most relevant was located. For example, in the legend PTR = 1 means that the most relevant result was in position 1 in the list returned. PTR = 2 means that the most relevant result was in the second position in the list returned, and so on. PTR = Background means that the most relevant result was not in the first 10 results returned.

The first group of bars shows the proportion of times users clicked on the first result returned. Notice that all users clicked on the first result when it was the most relevant, but nearly half clicked on the first result when the most relevant result was in the second position and more than half clicked on the first result when the most relevant result was even farther down the list.

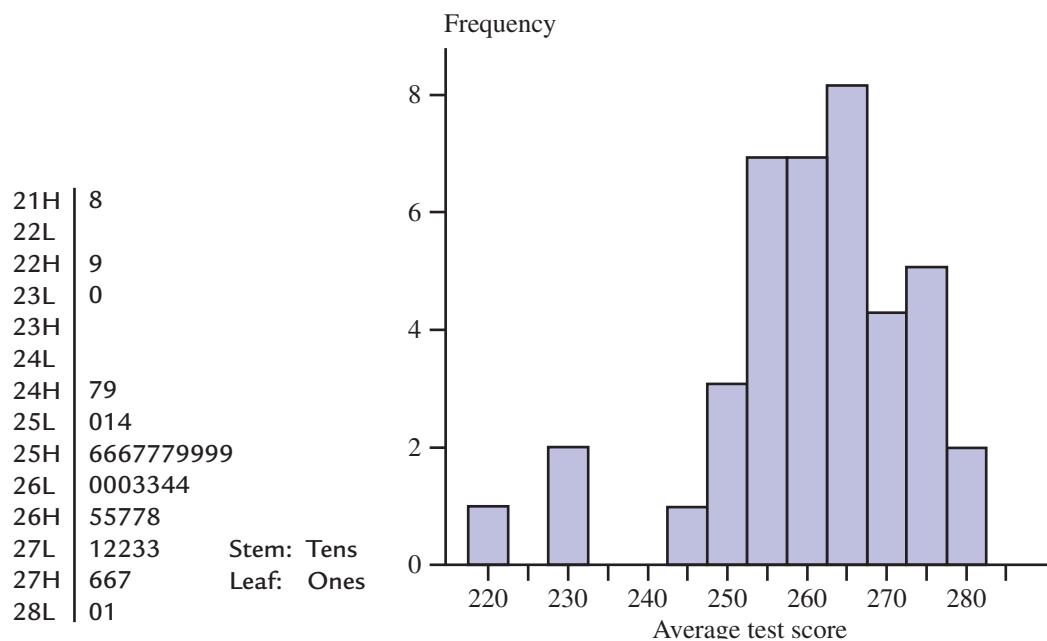
The second group of bars represents the proportion of users who clicked on the second result. Notice that the proportion who clicked on the second result was highest when the most relevant result was in that position. Stepping back to look at the entire graphical display, we see that users tended to click on the most relevant result if it was in one of the first three positions, but if it appeared after that, very few selected it. Also, if the most relevant result was in the third or a later position, users were more likely to click on the first result returned, and the likelihood of a click on the most relevant result decreased the farther down the list it appeared. To fully understand why the researchers' conclusions are justified, we need to be able to extract this kind of information from graphical displays.

The use of graphical data displays is quite common in newspapers, magazines, and journals. For example, data on test scores for a standardized math test given to eighth graders in 37 states, 2 territories (Guam and the Virgin Islands), and the District of Columbia were used to construct the stem-and-leaf display and histogram shown in Figure 3.40. Careful examination of these displays reveals the following:

1. Most of the participating states had average eighth-grade math scores between 240 and 280. We would describe the shape of this display as negatively skewed, because of the longer tail on the low end of the distribution.
2. Three of the average scores differed substantially from the others. These turn out to be 218 (Virgin Islands), 229 (District of Columbia), and 230 (Guam). These three scores could be described as outliers. It is interesting to note that the three unusual values are from the areas that are not states.
3. There do not appear to be any outliers on the high side.
4. A “typical” average math score for the 37 states would be somewhere around 260.
5. There is quite a bit of variability in average score from state to state.

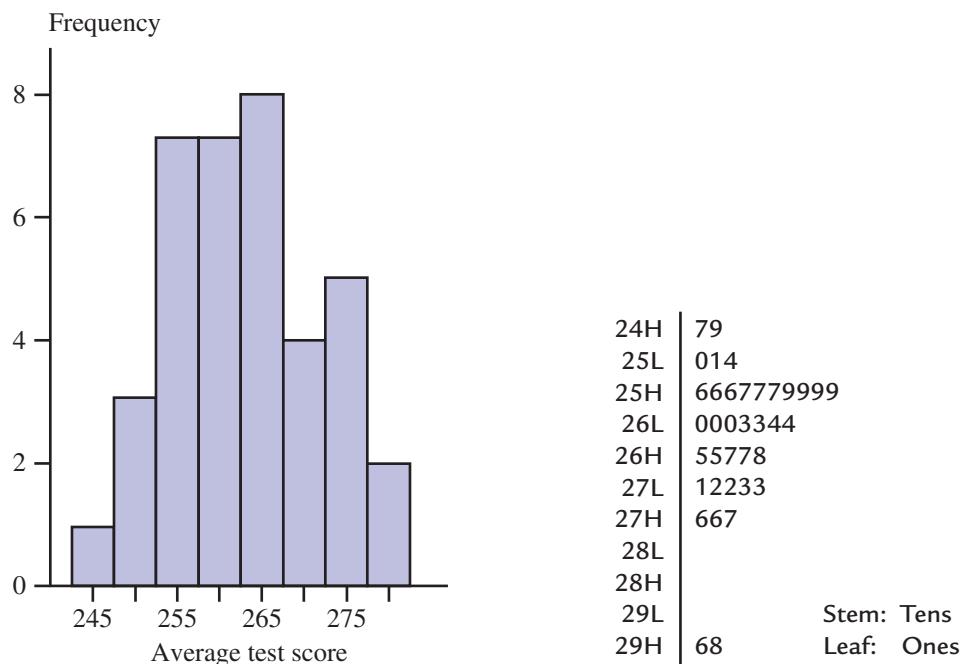
FIGURE 3.40

Stem-and-leaf display and histogram for math test scores.



How would the displays have been different if the two territories and the District of Columbia had not participated in the testing? The resulting histogram is shown in Figure 3.41. Notice that the display is now more symmetric, with no noticeable outliers. The display still reveals quite a bit of state-to-state variability in average score, and 260 still looks reasonable as a “typical” average score.

Now suppose that the two highest values among the 37 states (Montana and North Dakota) had been even higher. The stem-and-leaf display might then look like the one given in Figure 3.42. In this stem-and-leaf display, two values stand out from the main part of the display. This would catch our attention and might cause us to look carefully at these two states to determine what factors may be related to high math scores.

**FIGURE 3.41**

Histogram for the modified math score data.

FIGURE 3.42

Stem-and-leaf display for modified math score data.

What to Look for in Published Data

Here are some questions you might ask yourself when attempting to extract information from a graphical data display:

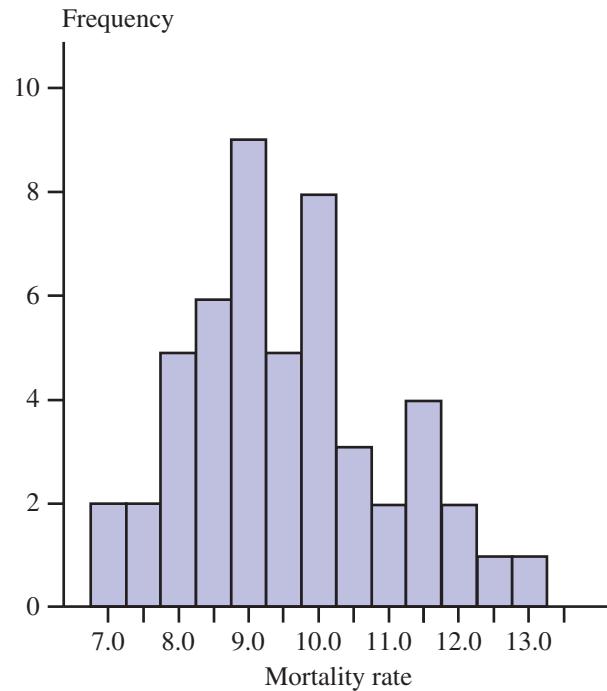
- Is the chosen display appropriate for the type of data collected?
- For graphical displays of univariate numerical data, how would you describe the shape of the distribution, and what does this say about the variable being summarized?
- Are there any outliers (noticeably unusual values) in the data set? Is there any plausible explanation for why these values differ from the rest of the data? (The presence of outliers often leads to further avenues of investigation.)
- Where do most of the data values fall? What is a typical value for the data set? What does this say about the variable being summarized?
- Is there much variability in the data values? What does this say about the variable being summarized?

Of course, you should always think carefully about how the data were collected. If the data were not gathered in a reasonable manner (based on sound sampling methods or experimental design principles), you should be cautious in formulating any conclusions based on the data.

Consider the histogram in Figure 3.43, which is based on data published by the [National Center for Health Statistics](#). The data set summarized by this histogram consisted of infant mortality rates (deaths per 1000 live births) for the 50 states in the United States. A histogram is an appropriate way of summarizing these data (although with only 50 observations, a stem-and-leaf display or dotplot would also have been reasonable).

FIGURE 3.43

Histogram of infant mortality rates.



The histogram itself is slightly positively skewed, with most mortality rates between 7.5 and 12. There is quite a bit of variability in infant mortality rate from state to state—perhaps more than we might have expected. This variability might be explained by differences in economic conditions or in access to health care. We may want to look further into these issues.

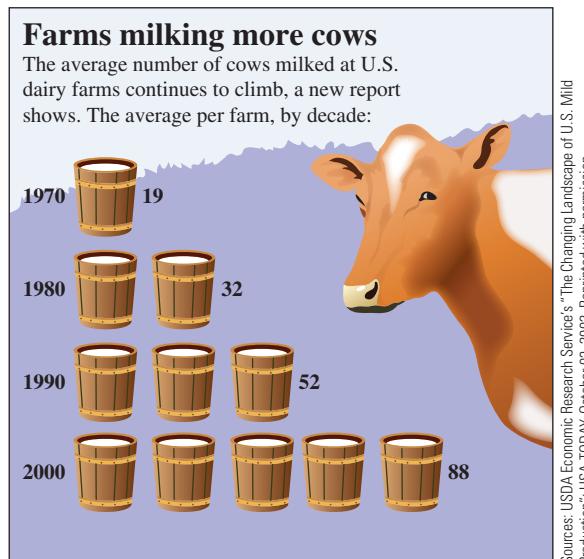
Although there are no obvious outliers, the upper tail is a little longer than the lower tail. The three largest values in the data set are 12.1 (Alabama), 12.3 (Georgia), and 12.8 (South Carolina)—all Southern states. Again, this may suggest some interesting questions that deserve further investigation.

A typical infant mortality rate would be about 9.5 deaths per 1000 live births. This represents an improvement, because researchers at the National Center for Health Statistics stated that the overall rate for 1988 was 10 deaths per 1000 live births. However, they also point out that the United States still ranked 22 out of 24 industrialized nations surveyed, with only New Zealand and Israel having higher infant mortality rates.

A Word to the Wise: Cautions and Limitations

When constructing and interpreting graphical displays, you need to keep in mind these things:

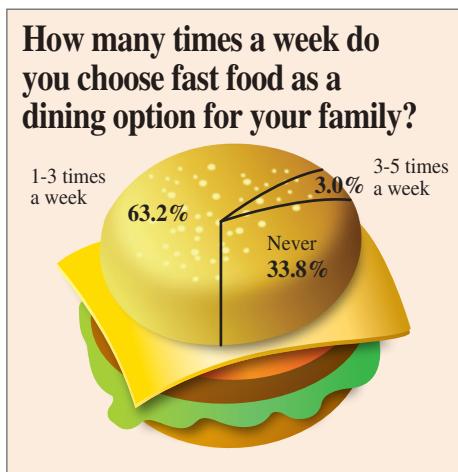
- 1. Areas should be proportional to frequency, relative frequency, or magnitude of the number being represented.** The eye is naturally drawn to large areas in graphical displays, and it is natural for the observer to make informal comparisons based on area. Correctly constructed graphical displays, such as pie charts, bar charts, and histograms, are designed so that the areas of the pie slices or the bars are proportional to frequency or relative frequency. Sometimes, in an effort to make graphical displays more interesting, designers lose sight of this important principle, and the resulting graphs are misleading. For example, consider the following graph similar to one that appeared in *USA TODAY* (October 3, 2002):



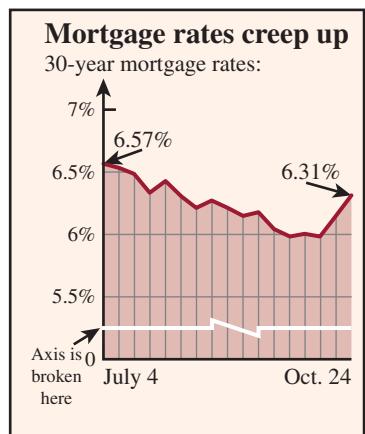
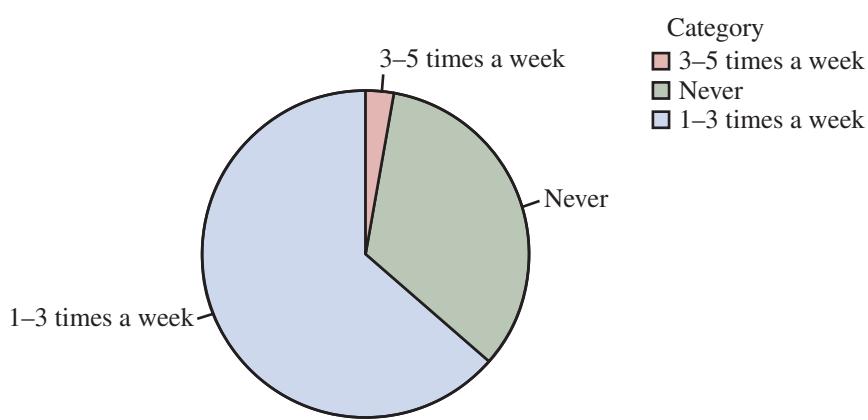
In trying to make the graph more visually interesting by replacing the bars of a bar chart with milk buckets, areas are distorted. For example, the two buckets for 1980 represent 32 cows, whereas the one bucket for 1970 represents 19 cows. This is misleading because 32 is not twice as big as 19. Other areas are distorted as well.

Another common distortion occurs when a third dimension is added to bar charts or pie charts. For example, the pie chart at the top left of the next page is similar to one that appeared in *USA TODAY* (September 17, 2009).

Adding the third dimension distorts the areas and makes it much more difficult to interpret correctly. A correctly drawn pie chart is also shown on the next page.



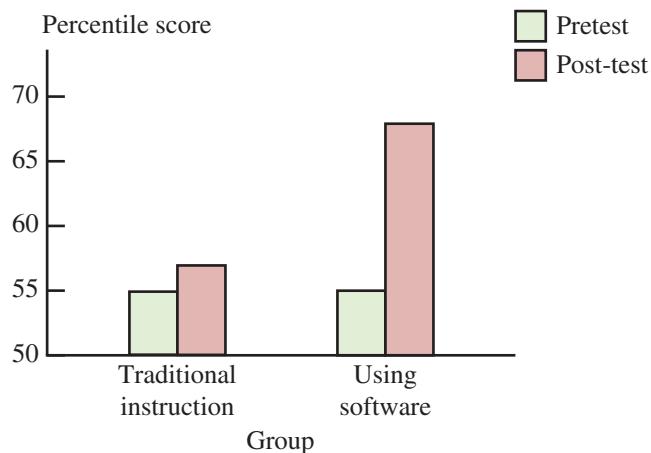
Sources: Market Day survey of 600 mothers of school age children; USA TODAY. September 17, 2009.



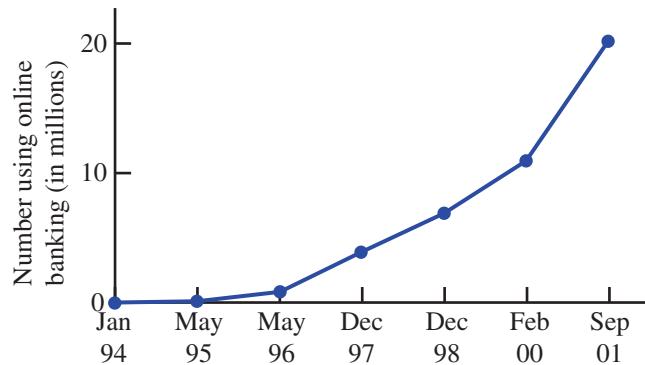
Sources: Freddie Mac; USA TODAY. October 22, 2002.

- 2. Be cautious of graphs with broken axes.** Although it is common to see scatterplots with broken axes, be extremely cautious of time series plots, bar charts, or histograms with broken axes. The use of broken axes in a scatterplot does not distort information about the nature of the relationship in the bivariate data set used to construct the display. On the other hand, in time series plots, broken axes can sometimes exaggerate the magnitude of change over time. Although it is not always inadvisable to break the vertical axis in a time series plot, it is something you should watch for, and if you see a time series plot with a broken axis, as in the accompanying time series plot of mortgage rates (similar to a graph appearing in *USA TODAY*, October 25, 2002), you should pay particular attention to the scale on the vertical axis and take extra care in interpreting the graph. Notice that the spacing between 0 and 5.5% is about the same as the spacing between 5.5% and 6%, which has the potential to be misleading.

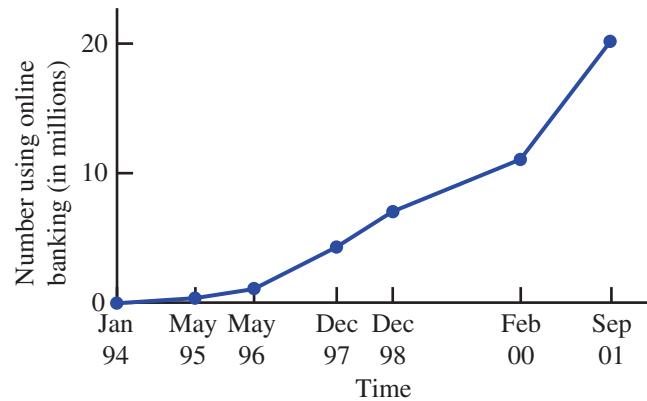
In bar charts and histograms, the vertical axis (which represents frequency, relative frequency, or density) should *never* be broken. If the vertical axis is broken in this type of graph, the resulting display will violate the “proportional area” principle and the display will be misleading. For example, the accompanying bar chart is similar to one appearing in an advertisement for a software product designed to help teachers raise student test scores. By starting the vertical axis at 50, the gain for students using the software is exaggerated. Areas of the bars are not proportional to the magnitude of the numbers represented—the area for the rectangle representing 68 is more than three times the area of the rectangle representing 55!



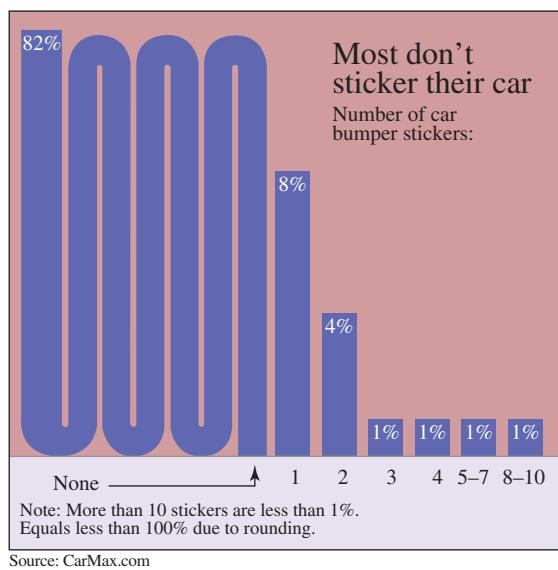
- 3. Watch out for unequal time spacing in time series plots.** If observations over time are not made at regular time intervals, special care must be taken in constructing the time series plot. Consider the accompanying time series plot, which is similar to one appearing in the *San Luis Obispo Tribune* (September 22, 2002) in an article on online banking:



Notice that the intervals between observations are irregular, yet the points in the plot are equally spaced along the time axis. This makes it difficult to make a coherent assessment of the rate of change over time. This could have been remedied by spacing the observations differently along the time axis, as shown in the following plot:



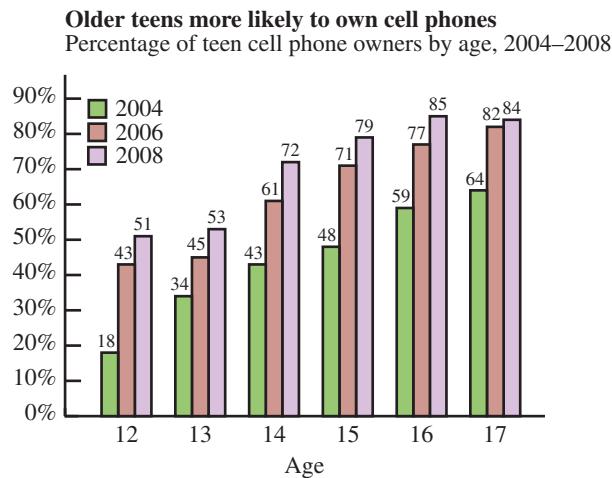
- 4. Be careful how you interpret patterns in scatterplots.** A strong pattern in a scatterplot means that the two variables tend to vary together in a predictable way, but it does not mean that there is a cause-and-effect relationship between the two variables. We will consider this point further in Chapter 5, but in the meantime, when describing patterns in scatterplots, be careful not to use wording that implies that changes in one variable *cause* changes in the other.
- 5. Make sure that a graphical display creates the right first impression.** For example, consider the graph on the next page, which is similar to one appearing in *USA TODAY* (June 25, 2002). Although this graph does not violate the proportional area principle, the way the “bar” for the “none” category is displayed makes this graph difficult to read, and a quick glance at this graph would leave the reader with an incorrect impression.



EXERCISES 3.48 - 3.53

● Data set available online

- 3.48** The accompanying comparative bar chart is similar to one in the report [“More and More Teens on Cell Phones” \(Pew Research Center, pewresearch.org, August 19, 2009\)](#).



Suppose that you plan to include this graph in an article that you are writing for your school newspaper. Write a few paragraphs that could accompany the graph. Be sure to address what the graph reveals about how teen cell phone ownership is related to age and how it has changed over time.

- 3.49** The figure at the top left of the next page is from the Fall 2008 Census Enrollment Report at Cal Poly, San Luis Obispo. It uses both a pie chart and a segmented bar chart to summarize data on ethnicity for students enrolled at the university in Fall 2008.

- Use the information in the graphical display to construct a single segmented bar chart for the ethnicity data.
- Do you think that the original graphical display or the one you created in Part (a) is more informative? Explain your choice.
- Why do you think that the original graphical display format (combination of pie chart and segmented bar chart) was chosen over a single pie chart with 7 slices?

- 3.50** The figure at the top right of the next page is similar to one that appeared in [USA TODAY \(August 5, 2008\)](#). This graph is a modified comparative bar chart. Most likely, the modifications (incorporating hands and the earth) were made to try to make a display that readers would find more interesting.

- Use the information in the *USA TODAY* graph to construct a traditional comparative bar chart.
- Explain why the modifications made in the *USA TODAY* graph may make interpretation more difficult than with the traditional comparative bar chart.

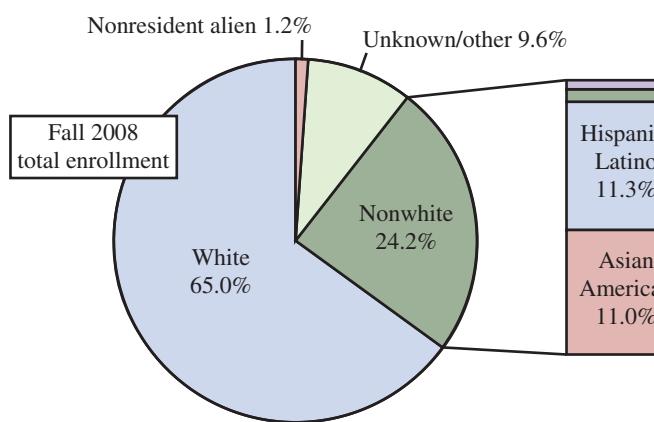


Figure for Exercise 3.49

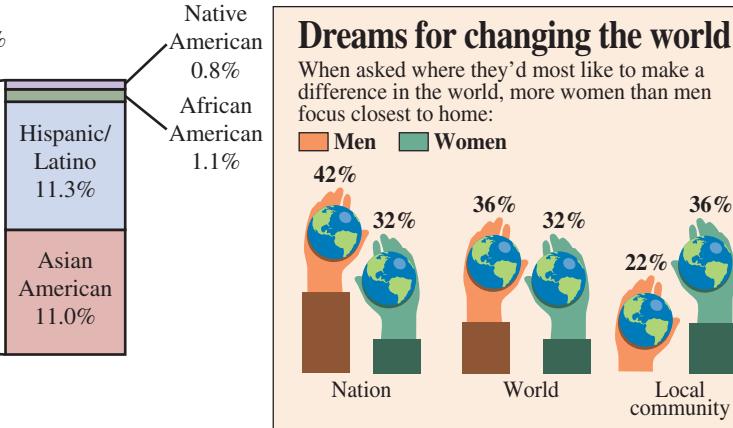
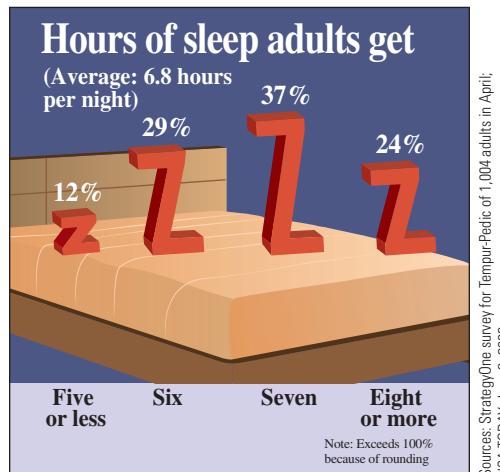


Figure for Exercise 3.50

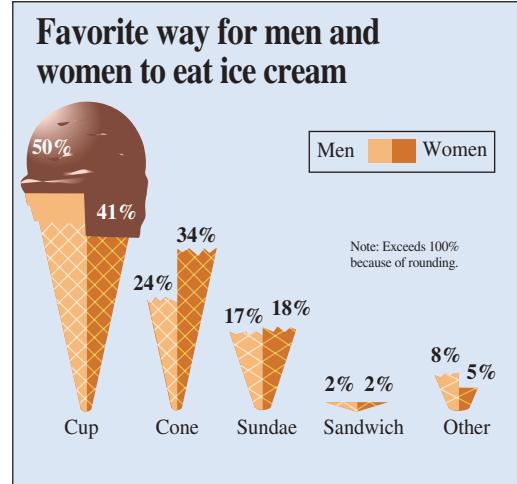
Sources: American Express survey of 1,000 adults conducted online by ICR and Authentic Response, USA TODAY, August 5, 2008.

- 3.51** The two graphical displays below are similar to ones that appeared in *USA TODAY* (June 8, 2009 and July 28, 2009). One is an appropriate representation and the other is not. For each of the two, explain why it is or is not drawn appropriately.

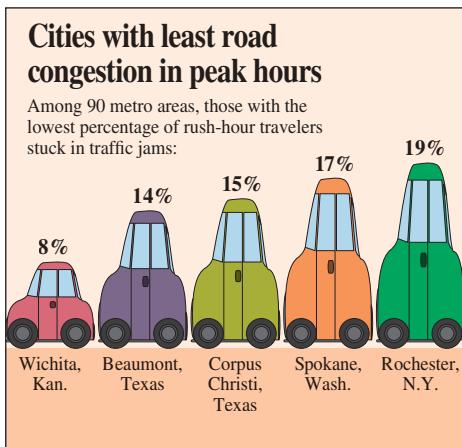


Sources: StrategyOne survey for Tempur-Pedic of 1,004 adults in April; USA TODAY, June 8, 2009.

effective summary of the data? If so, explain why. If not, explain why not and construct a display that makes it easier to compare the ice cream preferences of men and women.



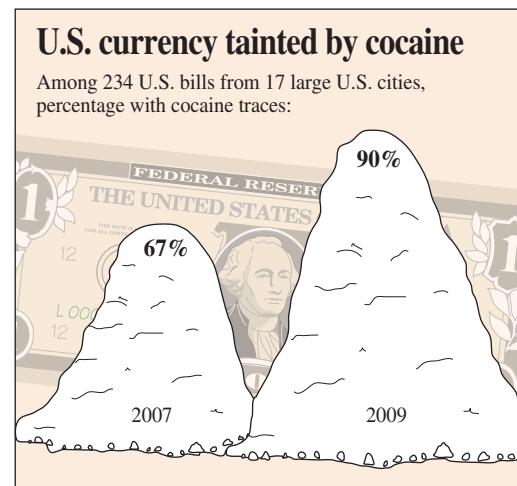
Sources: Harris Interactive survey of 2,177 adults conducted online June 8, 2015, USA TODAY, August 3, 2009.



Sources: Texas Transportation Institute; USA TODAY, July 28, 2009.

- 3.52** The following graphical display is similar to one that appeared in *USA TODAY* and is meant to be a comparative bar chart (*USA TODAY*, August 3, 2009). Do you think that this graphical display is an

- 3.53** Explain why the following graphical display (similar to one appearing in *USA TODAY*, September 17, 2009) is misleading.



Sources: American Chemical Society; USA TODAY, September 17, 2009.

CHAPTER ACTIVITIES

ACTIVITY 3.1 LOCATING STATES

Background: A newspaper article bemoaning the state of students' knowledge of geography claimed that more students could identify the island where the 2002 season of the TV show *Survivor* was filmed than could locate Vermont on a map of the United States. In this activity, you will collect data that will allow you to estimate the proportion of students who can correctly locate the states of Vermont and Nebraska.

1. Working as a class, decide how you will select a sample that you think will be representative of the students from your school.
2. Use the sampling method from Step 1 to obtain the subjects for this study. Subjects should be shown the accompanying map of the United States and asked to point out the state of Vermont. After the subject has given his or her answer, ask the subject to point out the state of Nebraska. For each subject, record whether or not Vermont was correctly identified and whether or not Nebraska was correctly identified.



3. When the data collection process is complete, summarize the resulting data in a table like the one shown here:

Response	Frequency
Correctly identified both states	
Correctly identified Vermont but not Nebraska	
Correctly identified Nebraska but not Vermont	
Did not correctly identify either state	

4. Construct a pie chart that summarizes the data in the table from Step 3.
5. What proportion of sampled students were able to correctly identify Vermont on the map?
6. What proportion of sampled students were able to correctly identify Nebraska on the map?
7. Construct a comparative bar chart that shows the proportion correct and the proportion incorrect for each of the two states considered.
8. Which state, Vermont or Nebraska, is closer to the state in which your school is located? Based on the pie chart, do you think that the students at your school were better able to identify the state that was closer than the one that was farther away? Justify your answer.
9. Write a paragraph commenting on the level of knowledge of U.S. geography demonstrated by the students participating in this study.
10. Would you be comfortable generalizing your conclusions in Step 8 to the population of students at your school? Explain why or why not.

ACTIVITY 3.2 BEAN COUNTERS!

Materials needed: A large bowl of dried beans (or marbles, plastic beads, or any other small, fairly regular objects) and a coin.

In this activity, you will investigate whether people can hold more in their right hand or in their left hand.

1. Flip a coin to determine which hand you will measure first. If the coin lands heads side up, start with the right hand. If the coin lands tails side up, start with the left hand. With the designated hand, reach into the bowl and grab as many beans as possible. Raise the hand over the

bowl and count to 4. If no beans drop during the count to 4, drop the beans onto a piece of paper and record the number of beans grabbed. If any beans drop during the count, restart the count. That is, you must hold the beans for a count of 4 without any beans falling before you can determine the number grabbed. Repeat the process with the other hand, and then record the following information: (1) right-hand number, (2) left-hand number, and (3) dominant hand (left or right, depending on whether you are left- or right-handed).

2. Create a class data set by recording the values of the three variables listed in Step 1 for each student in your class.
3. Using the class data set, construct a comparative stem-and-leaf display with the right-hand counts displayed on the right and the left-hand counts displayed on the left of the stem-and-leaf display. Comment on the interesting features of the display and include a comparison of the right-hand count and left-hand count distributions.
4. Now construct a comparative stem-and-leaf display that allows you to compare dominant-hand count to nondominant-hand count. Does the display support the

theory that dominant-hand count tends to be higher than nondominant-hand count?

5. For each observation in the data set, compute the difference

$$\text{dominant-hand count} - \text{nondominant-hand count}$$
- Construct a stem-and-leaf display of the differences. Comment on the interesting features of this display.
6. Explain why looking at the distribution of the differences (Step 5) provides more information than the comparative stem-and-leaf display (Step 4). What information is lost in the comparative display that is retained in the display of the differences?

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
Frequency distribution	A table that displays frequencies, and sometimes relative and cumulative relative frequencies, for categories (categorical data), possible values (discrete numerical data), or class intervals (continuous data).	Histogram	A graph of the information in a frequency distribution for a numerical data set. A rectangle is drawn above each possible value (discrete data) or class interval. The rectangle's area is proportional to the corresponding frequency or relative frequency.
Comparative bar chart	Two or more bar charts that use the same set of horizontal and vertical axes.	Histogram shapes	A (smoothed) histogram may be unimodal (a single peak), bimodal (two peaks), or multimodal. A unimodal histogram may be symmetric, positively skewed (a long right or upper tail), or negatively skewed. A frequently occurring shape is one that is approximately normal.
Pie chart	A graph of a frequency distribution for a categorical data set. Each category is represented by a slice of the pie, and the area of the slice is proportional to the corresponding frequency or relative frequency.	Cumulative relative frequency plot	A graph of a cumulative relative frequency distribution.
Segmented bar chart	A graph of a frequency distribution for a categorical data set. Each category is represented by a segment of the bar, and the area of the segment is proportional to the corresponding frequency or relative frequency.	Scatterplot	A graph of bivariate numerical data in which each observation (x, y) is represented as a point with respect to a horizontal x -axis and a vertical y -axis.
Stem-and-leaf display	A method of organizing numerical data in which the stem values (leading digit(s) of the observations) are listed in a column, and the leaf (trailing digit(s)) for each observation is then listed beside the corresponding stem. Sometimes stems are repeated to stretch the display.	Time series plot	A graphical display of numerical data collected over time.

CHAPTER REVIEW Exercises 3.54 - 3.65

- 3.54** Each year, *The Princeton Review* conducts surveys of high school students who are applying to college and of parents of college applicants. The report “[2016 College Hopes & Worries Survey Findings](#)”

● Data set available online

princetonreview.com/cms-content/final_cohowo2016survrpt.pdf, retrieved April 15, 2017) included a summary of how 8347 high school students responded to the question “Ideally

how far from home would you like the college you attend to be?" Students responded by choosing one of four possible distance categories. Also included was a summary of how 2087 parents of students applying to college responded to the question "How far from home would you like the college your child attends to be?" The accompanying relative frequency table summarizes the student and parent responses.

Ideal Distance	Frequency	
	Students	Parents
Less than 250 miles	2,587	1,085
250 to less than 500 miles	2,671	626
500 to less than 1,000 miles	1,753	271
At least 1,000 miles	1,336	105
Total	8,347	2,087

- a. Explain why relative frequencies should be used when constructing a comparative bar chart to compare ideal distance for students and parents.
- b. Construct a comparative bar chart for these data.
- c. Write a few sentences commenting on similarities and differences in the distributions of ideal distance for parents and students.

- 3.55** Each year the College Board publishes a profile of students taking the SAT. In the report ["2016 College Bound Seniors: Total Group Profile Report,"](#) the average SAT scores were reported for three groups defined by first language learned. Use the data in the accompanying table to construct a bar chart of the average critical reading SAT score for the three groups.

First Language Learned	Average Critical Reading SAT Score
English	508
English and another language	476
A language other than English	465

- 3.56** The report referenced in the previous exercise also gave average math SAT scores for the three language groups, as shown in the following table.

First Language Learned	Average Math SAT
English	508
English and another language	499
A language other than English	525

- a. Construct a comparative bar chart for the average critical reading and math scores for the three language groups.

- b. Write a few sentences describing the differences and similarities between the three language groups as shown in the bar chart.

- 3.57** The data in the table below are the percentage of drivers who are uninsured in each of the 50 U.S. states and the District of Columbia as reported in the article ["2015's Most and Least Risky States for Drivers' Wallets"](#) ([wallenthub.com](#)). Construct a graphical display that shows the distribution of percent uninsured drivers and write a few sentences commenting on what the display reveals about the distribution.

State	Percent Uninsured Drivers	State	Percent Uninsured Drivers
Alabama	19.6	Missouri	13.5
Alaska	13.2	Montana	14.1
Arizona	10.6	Nebraska	6.7
Arkansas	15.9	Nevada	12.2
California	14.7	New Hampshire	9.3
Colorado	16.2	New Jersey	10.3
Connecticut	8.0	New Mexico	21.6
Delaware	11.5	New York	5.3
District of Columbia	11.9	North Carolina	9.1
Florida	23.8	North Dakota	5.9
Georgia	11.7	Ohio	13.5
Hawaii	8.9	Oklahoma	25.9
Idaho	6.7	Oregon	9.0
Illinois	13.3	Pennsylvania	6.5
Indiana	14.2	Rhode Island	17.0
Iowa	9.7	South Carolina	7.7
Kansas	9.4	South Dakota	7.8
Kentucky	15.8	Tennessee	20.1
Louisiana	13.9	Texas	13.3
Maine	4.7	Utah	5.8
Maryland	12.2	Vermont	8.5
Massachusetts	3.9	Virginia	10.1
Michigan	21.0	Washington	16.1
Minnesota	10.8	West Virginia	8.4
Mississippi	22.9	Wisconsin	11.7
		Wyoming	8.7

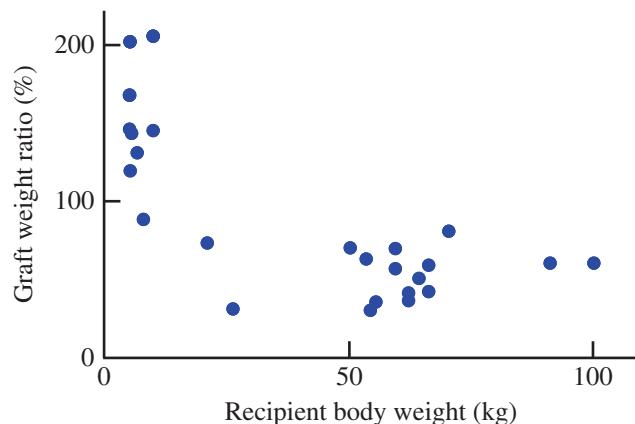
- 3.58** The following data on percentage increase in population between 2010 and 2017 for the 50 U.S. states and the District of Columbia (DC) were calculated using information from the U.S. Census Bureau ([factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=PEP_2017_PEPANNRES&src=pt](#),

(retrieved April 10, 2018). Each state is also classified as belonging to either the eastern or western part of the United States.

State	Percent Change	East/ West	State	Percent Change	East/ West
Alabama	1.8	E	Nebraska	4.7	W
Alaska	3.5	W	Nevada	9.8	W
Arizona	8.7	W	New Hampshire	1.9	E
Arkansas	2.7	E	New Jersey	2.2	E
California	5.6	W	New Mexico	1.1	W
Colorado	10.0	W	New York	2.2	E
Connecticut	0.2	E	North Carolina	6.8	E
Delaware	6.5	E	North Dakota	10.7	W
District of Columbia	12.8	E	Ohio	1.0	E
Florida	10.2	E	Oklahoma	4.4	W
Georgia	6.9	E	Oregon	7.4	W
Hawaii	4.5	W	Pennsylvania	0.7	E
Idaho	8.5	W	Rhode Island	0.6	E
Illinois	-0.3	E	South Carolina	7.7	E
Indiana	2.7	E	South Dakota	6.1	W
Iowa	3.0	E	Tennessee	5.4	E
Kansas	1.9	W	Texas	10.8	W
Kentucky	2.4	E	Utah	10.5	W
Louisiana	3.0	E	Vermont	-0.4	E
Maine	0.6	E	Virginia	5.3	E
Maryland	4.4	E	Washington	9.0	W
Massachusetts	4.3	E	West Virginia	-2.1	E
Michigan	0.9	E	Wisconsin	1.8	E
Minnesota	4.8	E	Wyoming	2.6	W
Mississippi	0.5	E			
Missouri	1.9	E			
Montana	5.7	W			

- a. Construct a stem-and-leaf display for the entire data set.
 - b. Comment on any interesting features of the display. Do any of the observations appear to be outliers?
 - c. Construct a comparative stem-and-leaf display for the Eastern and Western states. Write a few sentences comparing the two distributions.
- 3.59** Does the size of a transplanted organ matter? A study that attempted to answer this question (“Minimum Graft Size for Successful Living Donor Liver Transplantation,” *Transplantation* [1999]:1112–1116) included a graph similar to the accompanying scatterplot. “Graft weight ratio” is the weight of the transplanted liver relative to the ideal size liver for the recipient.
- a. Discuss interesting features of this scatterplot.

- b. Why do you think the overall relationship is negative?



- 3.60** The U.S. Census Bureau collects data on the percentage of households in the United States that have a computer. The accompanying table shows this percentage for the years 1984 to 2013 (census.gov/hhes/computer/publications/2012.html).

Year	Percentage of Households with a Computer
1984	8.2
1989	15.0
1993	22.9
1997	36.6
2000	51.0
2001	56.3
2003	61.8
2007	69.7
2009	74.1
2010	76.7
2011	75.6
2012	78.9
2013	83.8

- a. Construct a time series plot for these data. Be careful—the observations are not equally spaced along the *x*-axis.
 - b. Comment on any trend over time.
- 3.61** • The article “Tobacco and Alcohol Use in G-Rated Children’s Animated Films” (*Journal of the American Medical Association* [1999]: 1131–1136) reported exposure to tobacco and alcohol use in all G-rated animated films released between 1937 and 1997 by five major film studios. The researchers found that tobacco use was shown in 56% of the reviewed films. Data on the total tobacco exposure time

(in seconds) for films with tobacco use produced by Walt Disney, Inc., were as follows:

223 176 548 37 158 51 299 37 11 165
74 92 6 23 206 9

Data for 11 G-rated animated films showing tobacco use that were produced by MGM/United Artists, Warner Brothers, Universal, and Twentieth Century Fox were also given. The tobacco exposure times (in seconds) for these films was as follows:

205 162 6 1 117 5 91 155 24 55 17

- Construct a comparative stem-and-leaf display for these data.
- Comment on the interesting features of this display.

- 3.62** • The accompanying data on household expenditures on transportation for the United Kingdom appeared in “[Transport Statistics for Great Britain: 2002 Edition](#)” (*Family Spending: A Report on the Family Expenditure Survey* [The Stationery Office, 2002]). Expenditures (in pounds per week) included costs of purchasing and maintaining any vehicles owned by members of the household and any costs associated with public transportation and leisure travel.

Year	Average Transportation Expenditure	Percentage of Household Expenditures for Transportation
1990	247.20	16.2
1991	259.00	15.3
1992	271.80	15.8
1993	276.70	15.6
1994	283.60	15.1
1995	289.90	14.9
1996	309.10	15.7
1997	328.80	16.7
1998	352.20	17.0
1999	359.40	17.2
2000	385.70	16.7

- Construct time series plots of the transportation expense data and the percentage of household expense data.
- Do the time series plots of Part (a) support the statement that follows? Explain why or why not. Statement: Although actual expenditures have been increasing, the percentage of the total household expenditures that go toward transportation has remained relatively stable.

- 3.63** The National Center for Education Statistics reported the following data on the average cost per year for tuition, fees, and room and board for 4-year public institutions in the United States ([nces.ed.gov/fastfacts/display.asp?id=76](#), retrieved April 17, 2017). Construct a time series plot of these data and comment on the trend over time.

Year	Average Cost
2002	\$14,439
2003	\$15,505
2004	\$16,510
2005	\$17,451
2006	\$18,471
2007	\$19,363
2008	\$20,409
2009	\$21,126
2010	\$22,074
2011	\$23,011
2012	\$23,872
2013	\$24,706

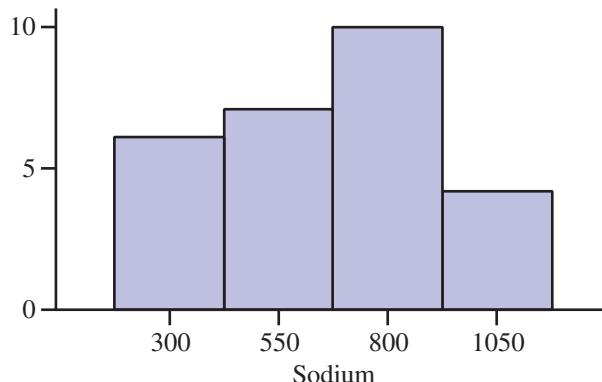
- 3.64** • Many nutritional experts have expressed concern about the high levels of sodium in prepared foods. The following data on sodium content (in milligrams) per frozen meal appeared in the article “[Comparison of ‘Light’ Frozen Meals](#)” (*Boston Globe*, April 24, 1991):

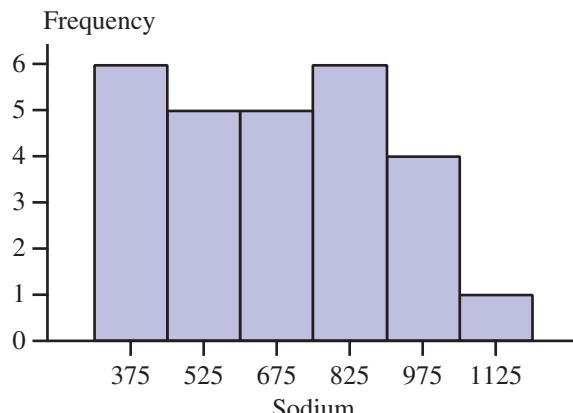
720 530 800 690 880 1050 340 810 760
300 400 680 780 390 950 520 500 630
480 940 450 990 910 420 850 390 600

Two histograms for these data are shown below and on the next page.

- Do the two histograms give different impressions about the distribution of values?
- Use each histogram to determine the approximate proportion of observations that are less than 800, and compare to the actual proportion.

Frequency





- 3.65** • Americium 241 (^{241}Am) is a radioactive material used in the manufacture of smoke detectors. The article “Retention and Dosimetry of Injected ^{241}Am in Beagles” (*Radiation Research* [1984]: 564–575) described a study in which 55 beagles were injected

with a dose of ^{241}Am (proportional to each animal’s weight). Skeletal retention of ^{241}Am (in microcuries per kilogram) was recorded for each beagle, resulting in the following data:

0.196 0.451 0.498 0.411 0.324 0.190 0.489
 0.300 0.346 0.448 0.188 0.399 0.305 0.304
 0.287 0.243 0.334 0.299 0.292 0.419 0.236
 0.315 0.447 0.585 0.291 0.186 0.393 0.419
 0.335 0.332 0.292 0.375 0.349 0.324 0.301
 0.333 0.408 0.399 0.303 0.318 0.468 0.441
 0.306 0.367 0.345 0.428 0.345 0.412 0.337
 0.353 0.357 0.320 0.354 0.361 0.329

- a. Construct a frequency distribution for these data, and draw the corresponding histogram.
 b. Write a short description of the important features of the shape of the histogram.

TECHNOLOGY NOTES

Comparative Bar Charts

JMP

1. Enter the raw data into one column
2. In a separate column enter the group information

3. Click **Graph** then select **Chart**
4. Click and drag the column containing the raw data from the box under **Select Columns** to the box next to **Categories, X, Levels**
5. Click and drag the column containing the group information from the box under **Select Columns** to the box next to **Categories, X, Labels**
6. Click **OK**

Minitab

1. Input the group information into C1
2. Input the raw data into C2 (be sure to match the response from the first column with the appropriate group)

3. Select **Graph** then choose **Bar Chart...**
4. Highlight Cluster
5. Click **OK**
6. Double click C1 to add it to the **Categorical Variables** box
7. Then double click C2 to add it to the **Categorical Variables** box
8. Click **OK**

Note: Be sure to list the grouping variable first in the Categorical Variables box.

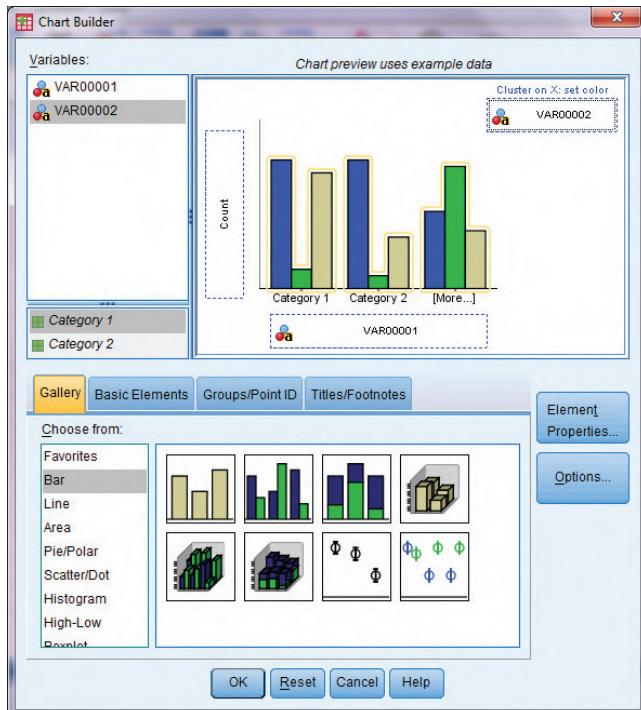
Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels...** button prior to performing Step 8 above.

SPSS

1. Enter the raw data into one column
2. Enter the group information into a second column (be sure to match the response from the first column with the appropriate group)

	VAR00001	VAR000...	var
1	pink	female	
2	blue	male	
3	red	male	
4	green	male	
5	yellow	female	
6	red	female	
7	blue	female	
8	blue	male	
9	pink	female	
10	red	male	
11	green	female	
12	yellow	male	

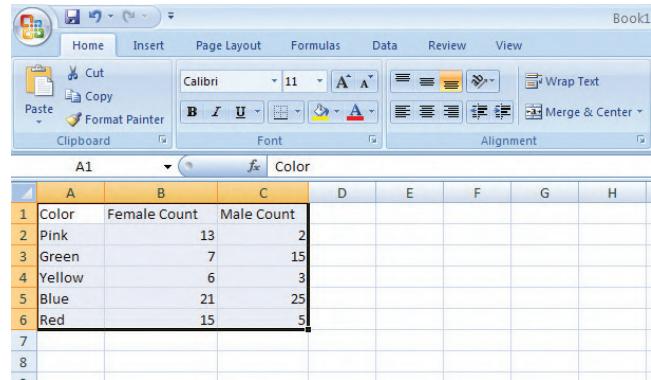
3. Select **Graph** and choose **Chart Builder...**
4. Under **Choose from** highlight Bar
5. Click and drag the second bar chart in the first column (Clustered Bar) to the Chart preview area
6. Click and drag the group variable (second column) into the **Cluster on X** box in the upper right corner of the chart preview area
7. Click and drag the data variable (first column) into the **X-Axis?** box in the chart preview area



8. Click **Ok**

Excel 2007

1. Enter the category names into column A (you may input the title for the variable in cell A1)
2. Enter the count or percent for different groups into a new column, starting with column B
3. Select all data (including column titles if used)



4. Click on the **Insert** Ribbon
5. Choose **Column** and select the first chart under 2-D Column (Clustered Column)
6. The chart will appear on the same worksheet as your data

Note: You may add or format titles, axis titles, legends, and so on, by right-clicking on the appropriate piece of the chart.

Note: Using the Bar Option on the Insert Ribbon will produce a horizontal bar chart.

TI-83/84

The TI-83/84 does not have the functionality to produce comparative bar charts.

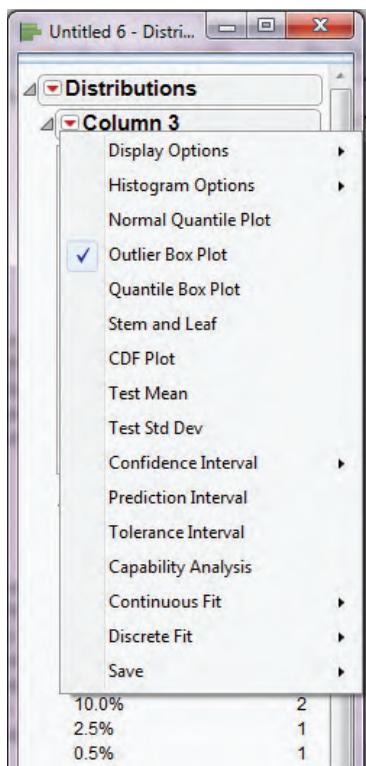
TI-Nspire

The TI-Nspire does not have the functionality to produce comparative bar charts.

Stem-and-Leaf Plots

JMP

1. Enter the raw data into a column
2. Click **Analyze** then select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to the column name



6. Select Stem and Leaf

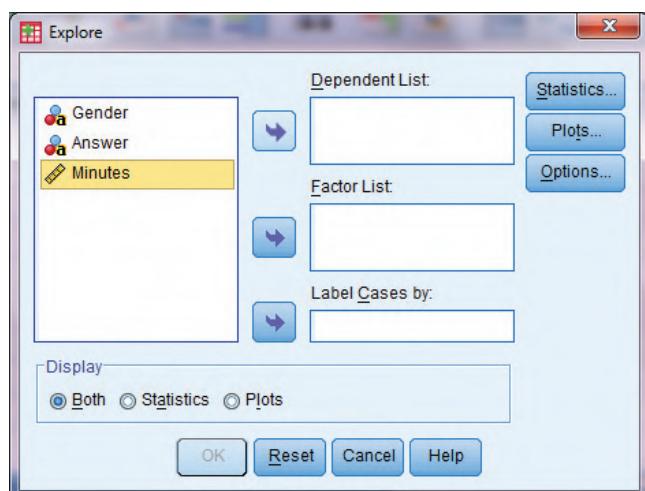
Minitab

1. Input the raw data into C1
2. Select **Graph** and choose **Stem-and-leaf...**
3. Double click C1 to add it to the **Graph Variables** box
4. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels...** button prior to performing Step 4 above.

SPSS

1. Input the raw data into a column
2. Select **Analyze** and choose **Descriptive Statistics** then **Explore**
3. Highlight the variable name from the box on the left
4. Click the arrow to the right of the Dependent List box to add the variable to this box



5. Click OK

Note: The stem-and-leaf plot is one of the plots produced along with several other descriptive statistics and plots.

Excel 2007

Excel 2007 does not have the functionality to create stem-and-leaf plots automatically.

TI-83/84

The TI-83/84 does not have the functionality to create stem-and-leaf plots.

TI-Nspire

The TI-Nspire does not have the functionality to create stem-and-leaf plots.

Histograms

JMP

1. Enter the raw data into a column
2. Click **Analyze** then select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to the column name
6. Select **Histogram Options** and click **Vertical**

Minitab

1. Input the raw data into C1
2. Select **Graph** and choose **Histogram...**
3. Highlight Simple
4. Click **OK**
5. Double click C1 to add it to the **Graph Variables** box
6. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels...** button prior to performing Step 5 above.

SPSS

1. Input the raw data into a column
2. Select **Graph** and choose **Chart Builder...**
3. Under **Choose from** highlight Histogram
4. Click and drag the first option (Simple Histogram) to the Chart preview area
5. Click and drag the variable name from the **Variables** box into the **X-Axis?** box
6. Click **OK**

Excel 2007

Note: In order to produce histograms in Excel 2007, we will use the Data Analysis Add-On. For instructions on installing this add-on, please see the note at the end of this chapter's Technology Notes.

1. Input the raw data in column A (you may use a column heading in cell A1)
2. Click on the **Data** ribbon
3. Choose the **Data Analysis** option from the **Analysis** group

4. Select **Histogram** from the Data Analysis dialog box and click **OK**
5. Click in the **Input Range:** box and then click and drag to select your data (including the column heading)
 - a. If you used a column heading, click the checkbox next to **Labels**
6. Click the checkbox next to **Chart Output**
7. Click **OK**

Note: If you have specified bins manually in another column, click in the box next to **Bin Range:** and select your bin assignments.

TI-83/84

1. Enter the data into **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **2nd** key then the **Y =** key
3. Highlight the option for **Plot1** and press **ENTER**
4. Highlight **On** and press **ENTER**
5. For **Type**, select the histogram shape and press **ENTER**

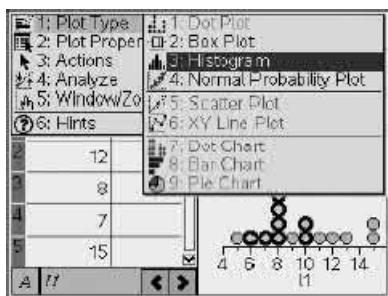


6. Press **GRAPH**

Note: If the graph window does not display appropriately, press the **WINDOW** button and reset the scales appropriately.

TI-Nspire

1. Enter the data into a data list (In order to access data lists select the spreadsheet option and press **enter**)
2. Press the **menu** key then select **3:Data** then select **6:QuickGraph** and press **enter** (a dotplot will appear)
3. Press the **menu** key and select **1:Plot Type**
4. Select **3:Histogram** and press **enter**



Scatterplots

JMP

1. Input the data for the independent variable into the first column
2. Input the data for the dependent variable into the second column

3. Click **Analyze** then select **Fit Y by X**
4. Click and drag the column name for the independent data from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the column name for the dependent data from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**

Minitab

1. Input the raw data for the independent variable into C1
2. Input the raw data for the dependent variable into C2
3. Select **Graph** and choose **Scatterplot**
4. Highlight Simple
5. Click **OK**
6. Double click on C2 to add it to the **Y Variables** column of the spreadsheet
7. Double click on C1 to add it to the **X Variables** column of the spreadsheet
8. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels...** button prior to performing Step 7 above.

SPSS

1. Input the raw data into two columns
2. Select **Graph** and choose **Chart Builder...**
3. Under **Choose from** highlight Scatter/Dot
4. Click and drag the first option (Simple Scatter) to the Chart preview area
5. Click and drag the variable name representing the independent variable to the **X-Axis?** box
6. Click and drag the variable name representing the dependent variable to the **Y-Axis?** box
7. Click **OK**

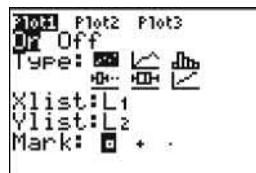
Excel 2007

1. Input the raw data into two columns (you may enter column headings)
2. Select both columns of data
3. Click on the **Insert** ribbon
4. Click **Scatter** and select the first option (Scatter with Markers Only)

Note: You may add or format titles, axis titles, legends, and so on, by right-clicking on the appropriate piece of the chart.

TI-83/84

1. Input the data for the dependent variable into L2 (In order to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Input the data for the independent variable into L1
3. Press the **2nd** key then press the **Y =** key
4. Select the **Plot1** option and press **ENTER**
5. Select **On** and press **ENTER**
6. Select the scatterplot option and press **ENTER**



- Press the **GRAPH** key

Note: If the graph window does not display appropriately, press the **WINDOW** button and reset the scales appropriately.

TI-Nspire

- Enter the data for the independent variable into a data list (In order to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

- Enter the data for the dependent variable into a separate data list
- Highlight both columns of data (by scrolling to the top of the screen to highlight one list then press shift and use the arrow key to highlight the second list)
- Press the **menu** key and select **3:Data** then select **6:QuickGraph** and press **enter**

Time Series Plots

JMP

- Input the raw data into one column
- Input the time increment data into a second column

Column 4	Column 5
1	1900
2	1901
3	1902
4	1903
5	1904
6	1905
7	1906
8	1907
9	1908
1	1909
2	1910
3	1911
4	1912

- Click **Graph** then select **Control Chart** then select **Run Chart**
- Click and drag the name of the column containing the raw data from the box under **Select Columns** to the box next to **Process**
- Click and drag the name of the column containing the time increment data from the box under **Select Columns** to the box next to **Sample Label**
- Click **OK**

Minitab

- Input the raw data into C1
- Select **Graph** and choose **Time Series Plot...**
- Highlight Simple

- Click **OK**

- Double click C1 to add it to the **Series** box
- Click the Time/Scale button
- Choose the appropriate time scale under Time Scale (Choose Calendar if you want to use Days, Years, Months, Quarters, etc.)
- Click to select one set for each variable
- In the spreadsheet, fill in the start value for the time scale
- Input the increment value into the box next to **Increment** (i.e., 1 to move by one year or one month, etc.)

- Click **OK**

- Click **OK**

Note: You may add or format titles, axis titles, legends, and so on, by clicking on the **Labels...** button prior to performing Step 12 above.

SPSS

- Input the raw data into two columns: one representing the data and one representing the time increments
- Select **Analyze** then choose **Forecasting** then **Sequence Charts...**
- Highlight the column name containing the raw data and add it to the **Variables** box
- Highlight the column name containing the time data and add it to the **Time Axis Labels** box
- Click **OK**

Excel 2007

- Input the data into two columns: one representing the data and one representing the time increments
- Select an empty cell
- Click on the **Insert** ribbon
- Click **Line** and select the first option under 2-D Line (Line)
- Right-click on the empty chart area that appears and select **Select Data...**
- Under **Legend Entries (Series)** click **Add**
- In the **Series name:** box select the column title for the data
- In the **Series values:** box select the data values
- Click **OK**
- Under **Horizontal (Category) Axis Labels** click **Edit**
- In the **Axis label range:** select the time increment data (do NOT select the column title).
- Click **OK**
- Click **OK**

TI-83/84

- Input the time increment data into **L1**
- Input the data values for each time into **L2**
- Press the **2nd** key then the **Y =** key
- Select **1:Plot1** and press **ENTER**
- Highlight **On** and press **ENTER**
- Highlight the second option in the first row and press **ENTER**
- Press **GRAPH**

Note: If the graph window does not display appropriately, press the **WINDOW** button and reset the scales appropriately.

TI-Nspire

- Enter the data for the time increments into a data list (In order to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

- Enter the data for each time increment into a separate data list
- Highlight both columns of data (by scrolling to the top of the screen to highlight one list then press shift and use the arrow key to highlight the second list)
- Press the **menu** key and select **3:Data** then select **6:QuickGraph** and press **enter**
- Press the **menu** key and select **1:Plot Type** then select **6:XY Line Plot** and press **enter**

CUMULATIVE REVIEW EXERCISES**CR3.1 - CR3.16**

● Data set available online

CR3.1 Does eating broccoli reduce the risk of prostate cancer? According to an observational study from the Fred Hutchinson Cancer Research Center (see the [CNN.com](#) web site article titled “[Broccoli, Not Pizza Sauce, Cuts Cancer Risk, Study Finds,](#)” January 5, 2000), men who ate more cruciferous vegetables (broccoli, cauliflower, brussels sprouts, and cabbage) had a lower risk of prostate cancer. This study made separate comparisons for men who ate different levels of vegetables. According to one of the investigators, “at any given level of total vegetable consumption, as the percent of cruciferous vegetables increased, the prostate cancer risk decreased.” Based on this study, is it reasonable to conclude that eating cruciferous vegetables causes a reduction in prostate cancer risk? Explain.

CR3.2 An article that appeared in [USA TODAY](#) (August 11, 1998) described a study on prayer and blood pressure. In this study, 2391 people 65 years or older, were followed for 6 years. The article stated that people who attended a religious service once a week and prayed or studied the Bible at least once a day were less likely to have high blood pressure. The researcher then concluded that “attending religious services lowers blood pressure.” The headline for this article was “Prayer Can Lower Blood Pressure.” Write a few sentences commenting on the appropriateness of the researcher’s conclusion and on the article headline.

CR3.3 Sometimes samples are composed entirely of volunteer responders. Give a brief description of the dangers of using voluntary response samples.

CR3.4 A newspaper headline stated that at a recent budget workshop, nearly three dozen people supported a sales tax increase to help deal with the city’s financial deficit ([San Luis Obispo Tribune](#), January 22, 2005). This conclusion was based on data from a survey acknowledged to be unscientific, in which 34 out of the 43 people who chose to attend the budget workshop recommended raising the sales tax. Briefly discuss

Installing Excel 2007’s Data Analysis Add-On

- Click the **Microsoft Office Button** and then Click **Excel Options**
- Click **Add-Ins**, then in the **Manage** box, select **Excel Add-ins**
- Click **Go**
- In the **Add-Ins available** box, select the **Analysis ToolPak** checkbox, and then click **OK**

Note: If this option is not listed, click **Browse** to locate it.

Note: If you get prompted that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.

- After you load the Analysis ToolPak, the **Data Analysis** command is available in the **Analysis** group on the **Data** ribbon.

why the survey was described as “unscientific” and how this might limit the conclusions that can be drawn from the survey data.

CR3.5 “More than half of California’s doctors say they are so frustrated with managed care they will quit, retire early, or leave the state within three years.” This conclusion from an article titled “[Doctors Feeling Pessimistic, Study Finds](#)” ([San Luis Obispo Tribune](#), July 15, 2001) was based on a mail survey conducted by the California Medical Association. Surveys were mailed to 19,000 California doctors, and 2000 completed surveys were returned. Describe any concerns you have regarding the conclusion drawn.

CR3.6 Based on observing more than 400 drivers in the Atlanta area, two investigators at Georgia State University concluded that people exiting parking spaces did so more slowly when a driver in another car was waiting for the space than when no one was waiting (“[Territorial Defense in Parking Lots: Retaliation Against Waiting Drivers](#),” *Journal of Applied Social Psychology* [1997]: 821-834).

- Describe how you might design an experiment to determine whether this phenomenon is true for your city.
- What is the response variable?
- What are some extraneous variables and how does your design control for them?

CR3.7 An article from the Associated Press (May 14, 2002) led with the headline “Academic Success Lowers Pregnancy Risk.” The article described an evaluation of a program that involved about 350 students at 18 Seattle schools in high crime areas. Some students took part in a program beginning in elementary school in which teachers showed children how to control their impulses, recognize the feelings of others, and get what they want without aggressive behavior. Others did not participate in the program. The study concluded that the program was effective because by the time young women in the program reached age 21, the

pregnancy rate among them was 38%, compared to 56% for the women in the experiment who did not take part in the program. Explain why this conclusion is valid only if the women in the experiment were randomly assigned to one of the two experimental groups.

CR3.8 Researchers at the University of Pennsylvania suggest that a nasal spray derived from pheromones (chemicals emitted by animals when they are trying to attract a mate) may be beneficial in relieving symptoms of premenstrual syndrome (PMS) (*Los Angeles Times*, January 17, 2003).

- Describe how you might design an experiment using 100 female volunteers who suffer from PMS to determine whether the nasal spray reduces PMS symptoms.
- Does your design from Part (a) include a placebo treatment? Why or why not?
- Does your design from Part (a) involve blinding? Is it single-blind or double-blind? Explain.

CR3.9 Students in California are required to pass an exit exam in order to graduate from high school. The pass rate for San Luis Obispo High School has been rising, as have the rates for San Luis Obispo County and the state of California (*San Luis Obispo Tribune*, August 17, 2004). The percentage of students who passed the test was as follows:

Year	District	Pass Rate
2002	San Luis Obispo High School	66%
2003		72%
2004		93%
2002	San Luis Obispo County	62%
2003		57%
2004		85%
2002	State of California	32%
2003		43%
2004		74%

- Construct a comparative bar chart that allows the change in the pass rate for each group to be compared.
- Is the change the same for each group? Comment on any difference observed.

CR3.10 A poll conducted by the **Associated Press-Ipsos** on public attitudes found that most Americans are convinced that political corruption is a major problem (*San Luis Obispo Tribune*, December 9, 2005). In the poll, 1002 adults were surveyed. Two of the questions and the summarized responses to these questions follow:

How widespread do you think corruption is in public service in America?

Hardly anyone	1%
A small number	20%
A moderate number	39%
A lot of people	28%
Almost everyone	10%
Not sure	2%

In general, which elected officials would you say are more ethical?

Democrats	36%
Republicans	33%
Both equally	10%
Neither	15%
Not sure	6%

- For each question, construct a pie chart summarizing the data.
- For each question, construct a segmented bar chart displaying the data.
- Which type of graph (pie chart or segmented bar chart) does a better job of presenting the data? Explain.

CR3.11 • The article “**Determination of Most Representative Subdivision**” (*Journal of Energy Engineering* [1993]: 43–55) gave data on various characteristics of subdivisions that could be used in deciding whether to provide electrical power using overhead lines or underground lines. Data on the variable x = total length of streets within a subdivision are as follows:

1280	5320	4390	2100	1240	3060	4770	1050
360	3330	3380	340	1000	960	1320	530
3350	540	3870	1250	2400	960	1120	2120
450	2250	2320	2400	3150	5700	5220	500
1850	2460	5850	2700	2730	1670	100	5770
3150	1890	510	240	396	1419	2109	

- Construct a stem-and-leaf display for these data using the thousands digit as the stem. Comment on the various features of the display.
- Construct a histogram using class boundaries of 0 to <1000, 1000 to <2000, and so on. How would you describe the shape of the histogram?
- What proportion of subdivisions has total length less than 2000? between 2000 and 4000?

CR3.12 • The paper “**Lessons from Pacemaker Implantations**” (*Journal of the American Medical Association* [1965]: 231–232) gave the results of a study that followed 89 heart patients who had received electronic pacemakers. The time (in months) to the first electrical malfunction of the pacemaker was recorded:

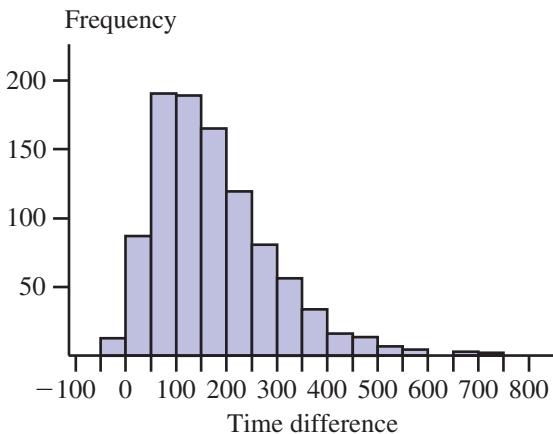
24	20	16	32	14	22	2	12	24	6	10	20
8	16	12	24	14	20	18	14	16	18	20	22
24	26	28	18	14	10	12	24	6	12	18	16
34	18	20	22	24	26	18	2	18	12	12	8
24	10	14	16	22	24	22	20	24	28	20	22
26	20	6	14	16	18	24	18	16	6	16	10
14	18	24	22	28	24	30	34	26	24	22	28
30	22	24	22	32							

- Summarize these data in the form of a frequency distribution, using class intervals of 0 to <6, 6 to <12, and so on.

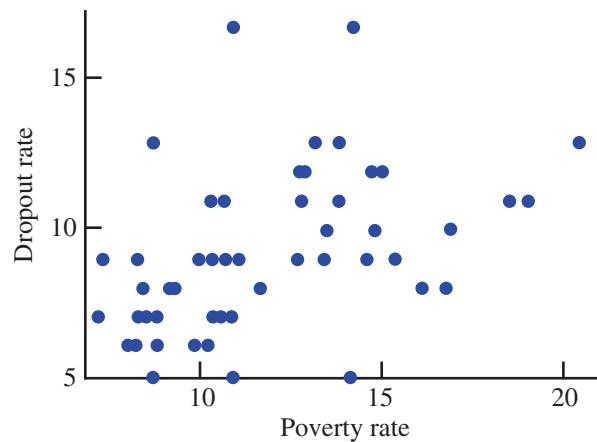
- b. Calculate the relative frequencies and cumulative relative frequencies for each class interval of the frequency distribution of Part (a).
- c. Show how the relative frequency for the class interval 12 to <18 could be obtained from the cumulative relative frequencies.
- d. Use the cumulative relative frequencies to give approximate answers to the following:
 - i. What proportion of those who participated in the study had pacemakers that did not malfunction within the first year?
 - ii. If the pacemaker must be replaced as soon as the first electrical malfunction occurs, approximately what proportion required replacement between 1 and 2 years after implantation?
- e. Construct a cumulative relative frequency plot, and use it to answer the following questions.
 - i. What is the approximate time at which 50% of the pacemakers had failed?
 - ii. What is the approximate time at which only 10% of the pacemakers initially implanted were still functioning?

CR3.13 How does the speed of a runner vary over the course of a marathon (a distance of 42.195 km)? Consider determining both the time (in seconds) to run the first 5 km and the time (in seconds) to run between the 35 km and 40 km points, and then subtracting the 5-km time from the 35–40-km time. A positive value of this difference corresponds to a runner slowing down toward the end of the race. The histogram below is based on times of runners who participated in several different Japanese marathons ([“Factors Affecting Runners’ Marathon Performance,” Chance \[Fall 1993\]: 24–30](#)).

- a. What are some interesting features of this histogram?
- b. What is a typical difference value?
- c. Roughly what proportion of the runners ran the late distance more quickly than the early distance?

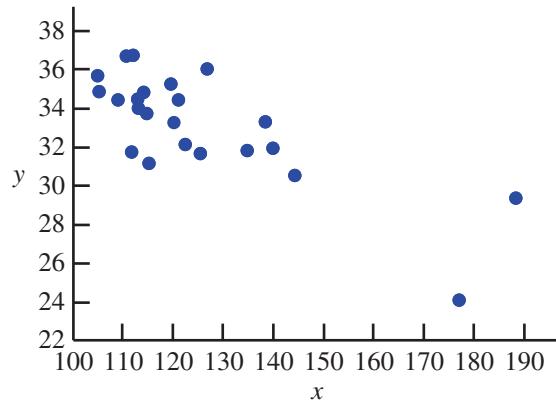


CR3.14 Data on x = poverty rate (%) and y = high school dropout rate (%) for the 50 U.S. states and the District of Columbia were used to construct the following scatterplot ([Chronicle of Higher Education, August 31, 2001](#)):



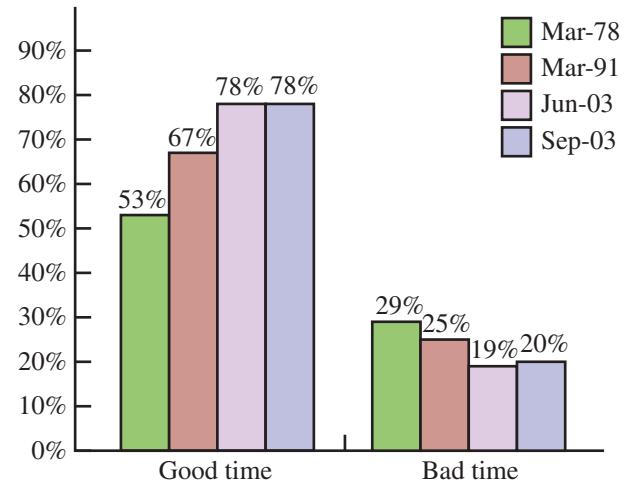
- a. Write a few sentences commenting on this scatterplot.
- b. Would you describe the relationship between poverty rate and dropout rate as positive (y tends to increase as x increases), negative (y tends to decrease as x increases), or as having no discernible relationship between x and y ?

CR3.15 One factor in the development of tennis elbow, a malady that strikes fear into the hearts of all serious players of that sport, is the impact-induced vibration of the racket-and-arm system at ball contact. It is well known that the likelihood of getting tennis elbow depends on various properties of the racket used. Consider the accompanying scatterplot of x = racket resonance frequency (in hertz) and y = sum of peak-to-peak accelerations (a characteristic of arm vibration, in meters per second per second) for $n = 23$ different rackets ([“Transfer of Tennis Racket Vibrations into the Human Forearm,” Medicine and Science in Sports and Exercise \[1992\]: 1134–1140](#)). Discuss interesting features of the data and of the scatterplot.



CR3.16 An article that appeared in *USA TODAY* (September 3, 2003) included a graph similar to the one shown here summarizing responses from polls conducted in 1978, 1991, and 2003 in which a sample of American adults were asked whether or not it was a good time or a bad time to buy a house.

- a. Construct a time series plot that shows how the percentage that thought it was a good time to buy a house has changed over time.
- b. Add a new line to the plot from Part (a) showing the percentage that thought it was a bad time to buy a house over time. Be sure to label the lines clearly.
- c. Which graph, the given bar chart or the time series plot, best shows the trend over time?



4

Numerical Methods for Describing Data



Estudi M6/Shutterstock.com

Preview Example: Just Thinking About Exercise?

Studies have shown that in some cases people overestimate the extent to which physical exercise can compensate for food consumption. When this happens, people increase food intake more than what is justified based on the exercise performed. The authors of the paper “[Just Thinking About Exercise Makes Me Serve More Food: Physical Activity and Calorie Compensation](#)” (*Appetite* [2011]: 332–335) wondered if even *thinking* about exercise would lead to increased food consumption.

They carried out an experiment in which people were offered snacks as a reward for participating in the experiment. People read a short essay and then answered a few questions about the essay. Some participants read an essay that was unrelated to exercise (the control group), some read an essay that described listening to music while taking a 30-minute walk (the fun group), and some read an essay that described strenuous exercise (the exercise group).

Participants were then given two plastic bags and invited to help themselves to two types of snacks—Chex Mix and M&Ms.

LEARNING OBJECTIVES

Students will understand:

- How the variance and standard deviation describe variability in a data set.
- The impact that outliers can have on measures of center and spread.
- The difference between a sample statistic and a population characteristic.

Students will be able to:

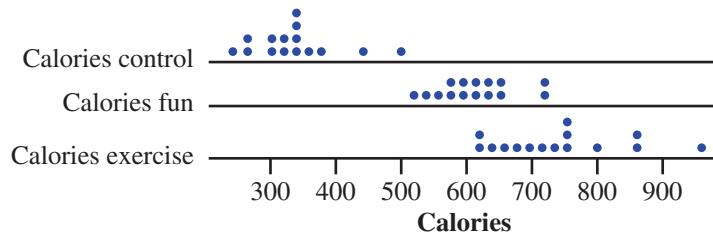
- Compute and interpret the values of the sample mean and the sample median.
- Compute and interpret the values of the sample standard deviation and the interquartile range.
- Construct and interpret a boxplot.
- Identify outliers in numerical data.
- Use Chebyshev’s Rule and the Empirical Rule to make statements about a data distribution.
- Use percentiles and z scores to describe relative standing.

After the participants served themselves, the bags were weighed so that the researchers could determine the number of calories in snacks taken.

Data on number of calories consistent with summary values in the paper were used to construct the comparative dotplot shown in Figure 4.1. From the dotplots, it is clear that the number of calories consumed differs from person to person and also tends to be quite a bit higher for those who read about exercise than for those in the control group! To further compare the three distributions, it is helpful to summarize them numerically.

FIGURE 4.1

Comparative dotplot of calories.



In this chapter, we show how to calculate numerical summary measures that describe both the center and the extent of variability in a data set.

SECTION 4.1 Describing the Center of a Data Set

When describing numerical data, it is common to report a value that is representative of the observations in the data set. Such a number describes roughly where the data are located or “centered” along the number line, and is called a measure of center. The two most widely used measures of center are the mean and the median.

The Mean

The **mean** of a numerical data set is just the familiar arithmetic average: the sum of the observations divided by the number of observations. At this point, it is helpful to introduce notation for the variable on which observations are made, for the number of observations in the data set, and for the individual observations:

x = the variable observed

n = the number of observations in the data set (the sample size)

x_1 = the first observation in the data set

x_2 = the second observation in the data set

\vdots

x_n = the n th (last) observation in the data set

For example, we might have a sample consisting of $n = 4$ observations on x = time it takes to complete an online hotel reservation (in minutes):

$$x_1 = 5.9 \quad x_2 = 7.3 \quad x_3 = 6.6 \quad x_4 = 5.7$$

Notice that the value of the subscript on x has no relationship to the magnitude of the observation. In this example, x_1 is just the first observation in the data set and not necessarily the smallest observation, and x_n is the last observation but not necessarily the largest.

The sum of x_1, x_2, \dots, x_n can be denoted by $x_1 + x_2 + \dots + x_n$, but this can also be written using summation notation. The Greek letter Σ denotes summation. In particular, Σx denotes the sum of all the x values in the data set under consideration.*

*It is also common to see Σx written as Σx_i or even as $\sum_{i=1}^n x_i$, but for simplicity we will usually omit the summation indices.

DEFINITION

Sample mean: The **sample mean** of a sample consisting of numerical observations x_1, x_2, \dots, x_n is denoted by \bar{x} , and its formula is given by

$$\bar{x} = \frac{\text{sum of all observations in the sample}}{\text{number of observations in the sample}} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\Sigma x}{n}$$

Example 4.1 Thinking About Exercise Again

Understand the context ➤

- The example in the chapter introduction described a study that investigated how thinking about exercise might affect food intake. Data on calories in snacks taken by people in each of three groups (control, read about fun walk, and read about strenuous exercise) are given in Table 4.1. Dotplots of these three data sets were given in the chapter introduction and are reproduced here as Figure 4.2.

FIGURE 4.2

Dotplot of calories.

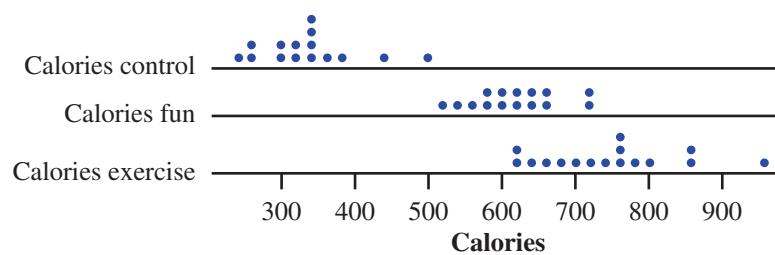


TABLE 4.1 Calorie Data

Consider the data ➤

Control Group	Fun Group	Exercise Group
340	668	626
300	585	802
329	637	768
381	588	738
331	529	751
445	597	715
320	719	670
256	612	630
252	622	862
342	553	761
332	600	648
296	658	956
357	542	671
505	711	854
242	641	703

Do the work ➤

For the control group, the sum of the sample data values is

$$\Sigma x = 340 + 300 + \dots + 242 = 5028$$

and the sample mean number of calories is

$$\bar{x} = \frac{\Sigma x}{n} = \frac{5028}{15} = 335.2$$

Interpret the results ➤

- Data set available online

The value of the sample mean describes where the number of calories for the control group is centered along the number line. It can be interpreted as a typical number of

calories for people in the control group. For the fun group, the sample mean number of calories is

$$\bar{x} = \frac{\Sigma x}{n} = \frac{9262}{15} = 617.5$$

and for the exercise group the sample mean number of calories is

$$\bar{x} = \frac{\Sigma x}{n} = \frac{11,155}{15} = 743.7$$

Notice that the sample mean number of calories for the control group was much smaller than the means for the other two groups.

The data values in Example 4.1 were all integers, yet the means were given as 335.2, 617.5 and 743.7. It is common to use more digits of decimal accuracy for the mean. This allows the value of the mean to fall between possible observable values (for example, the average number of children per family could be 1.8, whereas no single family will have 1.8 children). When calculating the value of the sample mean, it is common to round the calculated value to one more decimal place than the data. For example, if the data are integers, the value of the sample mean would be rounded to one decimal place. If the observations in the data set were numbers with one decimal place, the value of the sample mean would be rounded to two decimal places.

The sample mean \bar{x} is calculated using sample data, so it is a characteristic of that particular sample. It is customary to use Roman letters to denote sample characteristics, as we have done with \bar{x} . Characteristics of the population are usually denoted by Greek letters. One of the most important of such characteristics is the population mean.

DEFINITION

Population mean: The **population mean**, denoted by μ , is the average of all x values in the entire population.

For example, the average fuel efficiency for *all* 600,000 cars of a particular model might be $\mu = 27.52$ mpg. A sample of $n = 5$ cars might yield efficiencies of 27.3, 26.2, 28.4, 27.9, 26.5, from which we obtain $\bar{x} = 27.26$ for this particular sample (somewhat smaller than μ). However, a second sample might result in $\bar{x} = 28.52$, a third $\bar{x} = 26.85$, and so on. The value of \bar{x} varies from sample to sample, whereas there is just one value for μ .

In later chapters, we will see how the value of \bar{x} from a particular sample can be used to draw various conclusions about the value of μ . Example 4.2 illustrates how the value of \bar{x} from a particular sample can differ from the value of μ and how the value of \bar{x} differs from sample to sample.

Example 4.2 County Population Sizes

Understand the context ►

The 50 states plus the District of Columbia contain 3137 counties. Let x denote the number of residents of a county. Then there are 3137 values of the variable x in the population. The sum of these 3137 values is 231,665,106 (2018 Census Bureau estimate), so the population average value of x is

$$\mu = \frac{231,665,106}{3137} = 73,849.3 \text{ residents per county}$$

We used the Census Bureau web site to select three different random samples from this population of counties, with each sample consisting of five counties. The results appear in Table 4.2, along with the sample mean for each sample. The three \bar{x} values are different from one another because they are based on three different samples and the value of \bar{x} depends on the x values in the sample. Also notice that none of the three values comes close to

the value of the population mean, μ . If we did not know the value of μ , we might use one of these \bar{x} values as an *estimate* of μ , but in each case the estimate would be far off the mark.

TABLE 4.2 Three Samples from the Population of All U.S. Counties (x = number of residents)

Sample 1		Sample 2		Sample 3	
County	x Value	County	x Value	County	x Value
Fayette, TX	20,964	Stoddard, MO	28,509	Chattahoochee, GA	21,332
Monroe, IN	101,719	Johnston, OK	10,442	Petroleum, MT	638
Greene, NC	15,584	Sumter, AL	17,002	Armstrong, PA	77,299
Shoshone, ID	19,021	Milwaukee, WI	960,664	Schoolcraft, MI	8,625
Jasper, IN	26,570	Albany, WY	30,607	Benton, MO	12,133
$\Sigma x = 183,858$		$\Sigma x = 1,047,224$		$\Sigma x = 120,027$	
$\bar{x} = 36,771.6$		$\bar{x} = 209,444.8$		$\bar{x} = 24,005.4$	

Alternatively, we could combine the three samples into a single sample with $n = 15$ observations:

$$x_1 = 20,964, \dots, x_5 = 26,570, \dots, x_{15} = 12,133$$

$$\Sigma x = 1,351,109$$

$$\bar{x} = \frac{1,351,109}{15} = 90,073.9$$

This value is closer to the value of μ but is still not a very good estimate of the population mean. The problem here is that the population of x values exhibits a lot of variability (the largest value is $x = 7,767,421$ for Los Angeles County, California, and the smallest value is $x = 89$ for Loving County, Texas, which evidently few people love). Therefore, it is difficult for a sample of 15 observations, let alone just 5, to be reasonably representative of the population. In Chapter 9, we will see how variability is taken into account when deciding on the sample size required to accurately estimate a population mean.

One potential drawback to the mean as a measure of center for a data set is that its value can be greatly affected by the presence of even a single *outlier* (an unusually large or small observation) in the data set.

Example 4.3 Number of Visits to a Class Web Site

Understand the context ➤

- Forty students were enrolled in a section of a general education course in statistical reasoning during one fall quarter at Cal Poly, San Luis Obispo. The instructor made course materials, grades, and lecture notes available to students on a class web site, and course management software kept track of how often each student accessed any of the web pages on the class site.

One month after the course began, the instructor requested a report that indicated how many times each student had accessed a web page on the class site. The 40 observations were:

Consider the data ➤

20	37	4	20	0	84	14	36	5	331	19	0
0	22	3	13	14	36	4	0	18	8	0	26
4	0	5	23	19	7	12	8	13	16	21	7
13	12	8	42								

Interpret the results ➤

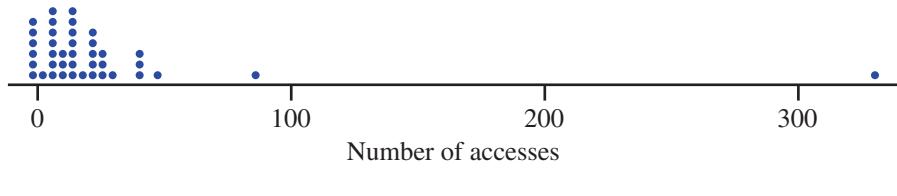
- The sample mean for this data set is $\bar{x} = 23.10$. Figure 4.3 is a Minitab dotplot of the data. Many would argue that 23.10 is not a very representative value for this sample, because 23.10 is larger than most of the observations in the data set. Notice that only 7 of 40 observations,

● Data set available online

or 17.5%, are larger than 23.10. The two outlying values of 84 and 331 (no, that was *not* a typo!) have a substantial impact on the value of \bar{x} .

FIGURE 4.3

A Minitab dotplot of the data in Example 4.3.



We now turn our attention to a measure of center that is not as sensitive to unusual values in a data set—the median.

The Median

The median strip of a highway divides the highway in half, and the median of a numerical data set does the same thing for a data set. Once the data values have been listed in order from smallest to largest, the **median** is the middle value in the list, and it divides the list into two equal parts.

Depending on whether the sample size n is even or odd, the process for determining the median is slightly different. When n is an odd number (say, 5), the sample median is the single middle value. But when n is even (say, 6), there are two middle values in the ordered list, and we average these two middle values to obtain the sample median.

DEFINITION

Sample median: The **sample median** is obtained by first ordering the n observations from smallest to largest (with any repeated values included, so that every sample observation appears in the ordered list). Then

$$\text{sample median} = \begin{cases} \text{the single middle value if } n \text{ is odd} \\ \text{the average of the middle two values if } n \text{ is even} \end{cases}$$

Example 4.4 Web Site Data Revised

The sample size for the web site access data of Example 4.3 was $n = 40$, an even number. The median is the average of the 20th and 21st values (the middle two) in the ordered list of the data. Arranging the data in order from smallest to largest produces the following ordered list (with the two middle values highlighted):

0	0	0	0	0	0	3	4	4	4	5	5
7	7	8	8	8	12	12	13	13	13	14	14
16	18	19	19	20	20	21	22	23	26	36	36
37	42	84	331								

The median can now be determined by averaging the two middle values:

Do the work ➤
$$\text{median} = \frac{13 + 13}{2} = 13$$

Interpret the results ➤

Looking at the dotplot (Figure 4.3), we see that this value appears to be a more typical value for the data set than the sample mean of 23.10.

The sample mean can be sensitive to even a single value that lies far above or below the rest of the data. The value of the mean is pulled out toward such outlying values. The median, on the other hand, is quite *insensitive* to outliers. For example, the largest sample

observation (331) in Example 4.4 can be increased by any amount without changing the value of the median. Similarly, an increase in the second or third largest observations does not affect the median, nor would a decrease in several of the smallest observations.

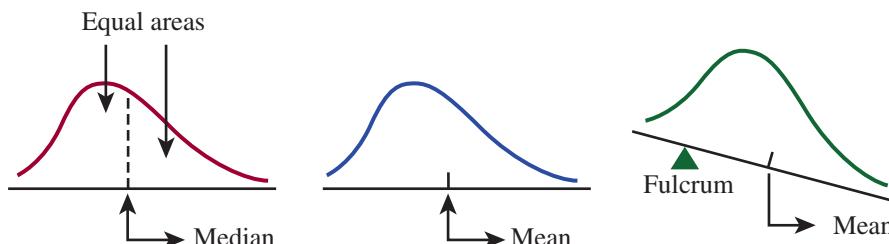
This stability of the median is what justifies its use as a measure of center in some situations. For example, the article “[Average Credit Card Debt in America: 2017 Facts and Figures](http://valuepenguin.com/average-credit-card-debt)” (valuepenguin.com/average-credit-card-debt, retrieved February 21, 2018) reported that the mean credit card debt for American households in 2017 was \$5700, whereas the median credit card debt was only \$2300. In this case, the small percentage of households with unusually high credit card debt results in a mean that may not be representative of a typical household’s credit card debt.

Comparing the Mean and the Median

Figure 4.4 shows several smoothed histograms that might represent either a distribution of sample values or a population distribution. The median is the value on the measurement axis that separates the smoothed histogram into two parts, with half (50%) of the area under each part of the curve. The mean is a bit harder to visualize. If the histogram were balanced on a triangle (a fulcrum), it would tilt unless the triangle was positioned at the mean. The mean is the balance point for the distribution.

FIGURE 4.4

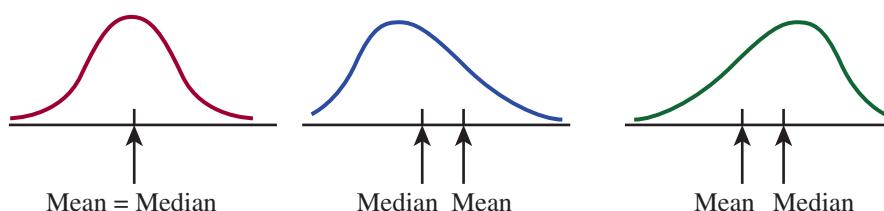
The mean and the median.



When the histogram is symmetric, the point of symmetry is both the dividing point for equal areas and the balance point, and the mean and the median are equal. However, when the histogram is unimodal (single-peaked) with a longer upper tail (positively skewed), the outlying values in the upper tail pull the mean up, so it generally lies above the median. For example, an unusually high exam score raises the mean but does not affect the median. Similarly, when a unimodal histogram is negatively skewed, the mean is generally smaller than the median (see Figure 4.5).

FIGURE 4.5

Relationship between the mean and the median.



Example 4.5 NBA Salaries

Understand the context ➤

- The web site [HoopsHype \(hoopshype.com/salaries/golden_state_warriors/\)](http://hoopshype.com/salaries/golden_state_warriors/) publishes salaries of NBA players. Salaries for the players of the Golden State Warriors in 2018 were

Consider the data ➤

Player	2018 Salary
Stephen Curry	34,682,550
Kevin Durant	25,000,000
Klay Thompson	17,826,150
Draymond Green	16,400,000
Andre Iguodala	14,814,815
Shaun Livingston	7,692,308

• Data set available online

(continued)

Player	2018 Salary
Nick Young	5,192,000
Zaza Pachulia	3,477,600
David West	2,328,652
JaVale McGee	2,116,955
Omri Casspi	2,106,470
Kevon Looney	1,471,382
Damien Jones	1,312,611
Patrick McCaw	1,312,611
Jason Thompson	893,333
Jordan Bell	815,615

A Minitab dotplot of these data is shown in Figure 4.6. Because the data distribution is positively skewed and there are outliers, we would expect the mean to be greater than the median.

For this data set, the mean is

$$\bar{x} = \frac{34,682,550 + \dots + 815,615}{16} = 8,590,190.8$$

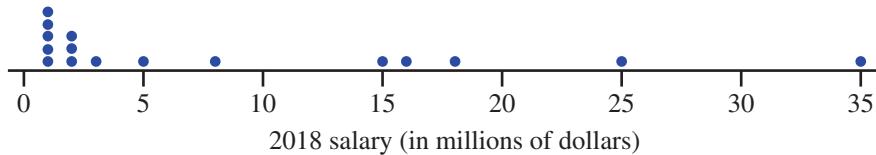
The median is the average of the middle two observations in the ordered list of salaries:

$$\text{median} = \frac{3,477,600 + 2,328,652}{2} = 2,903,126.0$$

Notice that the median is smaller than the mean, and it is probably a better description of a typical salary for this group.

FIGURE 4.6

Minitab dotplots for NBA salary data for the Golden State Warriors.



Categorical Data

The natural numerical summary quantities for a categorical data set are the relative frequencies for the different possible categories. Each relative frequency is the proportion (fraction) of responses in the corresponding category. Often there are only two possible responses (a dichotomy)—for example, male or female, does or does not have a driver’s license, did or did not vote in the last election. It is convenient in such situations to label one of the two possible responses S (for success) and the other F (for failure). As long as further analysis is consistent with the labeling, it does not matter which category is assigned the S label. When the data set is a sample, the fraction of S’s in the sample is called the **sample proportion of successes**.

DEFINITION

Sample proportion of successes: The **sample proportion of successes**, denoted by \hat{p} , is

$$\hat{p} = \text{sample proportion of successes} = \frac{\text{number of S's in the sample}}{n}$$

where S is the label used for the response designated as success.

Example 4.6 Can You Hear Me Now?

Understand the context ➤

It is not uncommon for a cell phone user to complain about the quality of his or her service provider. Suppose that each person in a sample of $n = 15$ cell phone users is asked if he or she is satisfied with the cell phone service. Each response is classified as S (satisfied) or F (not satisfied). The resulting data are

Consider the data ➤

S	F	S	S	S	F	F	S	S	F
S	S	S	F	F					



Getty Images

Do the work ➤

This sample contains nine S's and $n = 15$, so

$$\hat{p} = \frac{9}{15} = 0.60$$

Interpret the results ➤

This means that 60% of the sample responses are S's. Of those surveyed, 60% are satisfied with their cell phone service.

The letter p is used to denote the **population proportion of S's**.* We will see later how the value of \hat{p} from a particular sample can be used to draw conclusions about the population proportion p .

*Note that this is one situation in which we will not use a Greek letter to denote a population characteristic. Some statistics books use the symbol π for the population proportion and p for the sample proportion. We will not use π in this context so that there is no confusion with the mathematical constant $\pi = 3.14\dots$

EXERCISES 4.1 - 4.16

● Data set available online

- 4.1 The following are the prices (in dollars) of the six all-terrain truck tires rated most highly by *Consumer Reports* in 2018 (consumerreports.org, retrieved February 22, 2018):

159.00 199.00 157.00 127.65 123.99 126.00

- a. Calculate the values of the mean and median.
- b. Why are these values so different?
- c. Which of the two—mean or median—appears to be better as a description of a typical value for this data set? (Hint: See Example 4.5.)

- 4.2 ● The article “Caffeine Content of Drinks” (caffeinoinformer.com/the-caffeine-database, retrieved February 22, 2018) gave the following data on caffeine concentration (mg/ounce) for eight top-selling energy drinks:

Energy Drink	Caffeine Concentration (mg/oz)
Rockstar	10.0
Full Throttle	10.0
No Fear	11.4
Amp	8.9
SoBe Adrenaline Rush	9.5
Tab Energy	9.1

- a. What is the value of the mean caffeine concentration for this set of top-selling energy drinks?
- b. Coca-Cola has 2.9 mg/ounce of caffeine and Pepsi Cola has 3.2 mg/ounce of caffeine. Write a sentence explaining how the caffeine concentration of top-selling energy drinks compares to that of these colas.

- 4.3 ● *Consumer Reports Health* (consumerreports.org/health) reported the accompanying caffeine concentration (mg/cup) for 12 brands of coffee:

Energy Drink	Caffeine Concentration (mg/oz)
Red Bull	9.5
Monster	10.0

(continued)

Coffee Brand	Caffeine Concentration (mg/cup)
Eight O'Clock	140
Caribou	195
Kickapoo	155
Starbucks	115
Bucks Country Coffee Co.	195
Archer Farms	180
Gloria Jean's Coffees	110
Chock Full o'Nuts	110
Peet's Coffee	130
Maxwell House	55
Folgers	60
Millstone	60

Use at least one measure of center to compare caffeine concentration for coffee with that of the energy drinks of the previous exercise. (Note: 1 cup = 8 ounces)

- 4.4** ● Consumer Reports Health (consumerreports.org/health) reported the sodium content (mg) per 2 tablespoon serving for each of 11 different peanut butters:

120 50 140 120 150 150 150 65
170 250 110

- a. Display these data using a dotplot. Comment on any unusual features of the plot.
 - b. Calculate the mean and median sodium content for the peanut butters in this sample.
 - c. The values of the mean and the median for this data set are similar. What aspect of the distribution of sodium content—as pictured in the dotplot from Part (a)—provides an explanation for why the values of the mean and median are similar? (Hint: See the discussion of Figure 4.4.)
- 4.5** The article “[The Wedding Industry's Pricey Little Secret](#)” (June 12, 2013, slate.com) stated that the widely reported average wedding cost is grossly misleading. The article reports that in 2012, the average wedding cost was \$27,427 and the median cost was \$18,086.
- a. What does the large difference between the mean cost and the median cost tell you about the distribution of wedding costs in 2012?
 - b. Do you agree with the statement that the average wedding cost is misleading? Explain why or why not.
 - c. The article also states “the proportion of couples who spent the ‘average’ or more was actually a minority.” Do you agree with this statement? Explain why or why not using the reported values of the mean and median wedding cost.

- 4.6** The state of California defines family income groups in terms of median county income as follows:
- Extremely low income: below 30% of county median income
 - Very Low income: between 30% and 50% of county median income
 - Low income: between 50% and 80% of county median income
 - Moderate income: between 80% and 120% of county median income
- For San Luis Obispo county, the median household income in 2015 was \$60,691 (slohealthcounts.org/indicators/index/view?indicatorId=315&localeId=277, retrieved March 18, 2018).
- a. Interpret the value of the median household income in 2015 for San Luis Obispo County.
 - b. Each of the following statements is incorrect. For each statement, use the given information to explain why it is incorrect.

Statement 1: 30% of the households in San Luis Obispo County would be classified as extremely low income.

Statement 2: More than 50% of the households in San Luis Obispo County would be classified as extremely low income or very low income.

Statement 3: There cannot be any households in San Luis Obispo County that would be classified as having an income that was greater than those in the moderate-income category.

- 4.7** The report “[State of the News Media 2015](#)” (Pew Research Center, April 29, 2015) published the accompanying circulation numbers for 15 news magazines (such as *Time* and *The New Yorker*) for 2014:

3,284,012	1,469,223	1,214,590	1,046,977	993,043
931,228	905,755	843,914	783,353	574,370
483,360	412,062	147,808	119,297	41,518

Explain why the average may not be the best measure of a typical value for this data set.

- 4.8** Each student in a sample of 20 seniors at a particular university was asked if he or she was registered to vote. With R denoting registered and N denoting not registered, the sample data are:

R	R	N	N	N	R	N	R	N	R
N	R	N	R	R	R	N	R	R	R

- a. If being registered to vote is considered a “success,” what is the value of the proportion of successes for this sample?

- b. When would it be reasonable to generalize from this sample to the population of all seniors at this university?
- 4.9** ● The U.S. Department of Transportation reported the number of speed-related crash fatalities for the 15 states that had the highest numbers of these fatalities in 2012 (*Traffic Safety Facts 2012 Data, Speeding*, May 2014).

State	Speed-Related Traffic Fatalities
Texas	1,247
California	916
Pennsylvania	614
North Carolina	440
Illinois	387
Florida	361
New York	360
Ohio	356
Missouri	326
South Carolina	316
Arizona	297
Alabama	272
Virginia	271
Michigan	250
Oklahoma	218

- a. Calculate the mean number of speed-related fatalities for these 15 states.
- b. Calculate the median number of speed-related fatalities for these 15 states.
- c. Explain why it is not reasonable to generalize from this sample of 15 states to the other 35 states.
- 4.10** The ministry of **Health and Long-Term Care in Ontario, Canada**, publishes information on its web site (health.gov.on.ca) on the time that patients must wait for various medical procedures. For two cardiac procedures completed in fall of 2005, the following information was provided:

Number of Completed Procedures	Median Wait Time (days)	Mean Wait Time (days)	90% Completed Within (days)
Angioplasty	847	14	39
Bypass surgery	539	13	42

The median wait time for angioplasty is greater than the median wait time for bypass surgery but the mean wait time is shorter for angioplasty than for bypass surgery. What does this suggest about the distribution of wait times for these two procedures?

- 4.11** Houses in California are expensive, especially on the Central Coast where the air is clear, the ocean is blue, and the scenery is stunning. The median home price in San Luis Obispo County reached a new high in July 2004, soaring to \$452,272 from \$387,120 in March 2004 (*San Luis Obispo Tribune, April 28, 2004*).

The article included two quotes from people attempting to explain why the median price had increased. Richard Watkins, chairman of the Central Coast Regional Multiple Listing Services was quoted as saying, “There have been some fairly expensive houses selling, which pulls the median up.”

Robert Kleinhenz, deputy chief economist for the California Association of Realtors explained the volatility of house prices by stating: “Fewer sales means a relatively small number of very high or very low home prices can more easily skew medians.”

Are either of these statements correct? For each statement that is incorrect, explain why it is incorrect and propose a new wording that would correct any errors in the statement.

- 4.12** Consider the following statement: More than 65% of the residents of Los Angeles earn less than the average wage for that city. Could this statement be correct? If so, how? If not, why not?

- 4.13** A sample consisting of four pieces of luggage was selected from among the luggage checked at an airline counter, yielding the following data on x = weight (in pounds):

$$x_1 = 33.5, \quad x_2 = 27.3, \quad x_3 = 36.7, \quad x_4 = 30.5$$

Suppose that one more piece is selected and denote its weight by x_5 . Find a value of x_5 such that \bar{x} = sample median.

- 4.14** Suppose that 10 patients with meningitis received treatment with large doses of penicillin. Three days later, temperatures were recorded, and the treatment was considered successful if there had been a reduction in a patient’s temperature. Denoting success by S and failure by F, the 10 observations are

S S F S S S F F S S

- a. What is the value of the sample proportion of successes?
- b. Replace each S with a 1 and each F with a 0. Then calculate \bar{x} for this numerically coded sample. How does \bar{x} compare to \hat{p} ?
- c. Suppose that it is decided to include 15 more patients in the study. How many of these would have to be S’s to give $\hat{p} = 0.80$ for the entire sample of 25 patients?

- 4.15** A study of the lifetime (in hours) for a certain brand of light bulb involved putting 10 light bulbs into operation and observing them for 1000 hours. Eight of the light bulbs failed during that period, and those lifetimes were recorded. The lifetimes of the two light bulbs still functioning after 1000 hours were recorded as 1000+. The resulting sample observations were

480 790 1000+ 350 920 860 570 1000+
170 290

Which of the measures of center discussed in this section can be calculated, and what are the values of those measures?

- 4.16** An instructor has graded 19 exam papers submitted by students in a class of 20 students, and the average so far is 70. (The maximum possible score is 100.) How high would the score on the last paper have to be to raise the class average by 1 point? By 2 points?

SECTION 4.2 Describing Variability in a Data Set

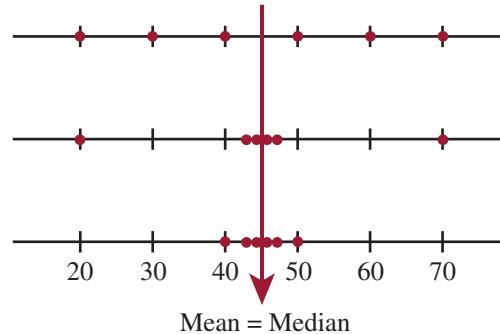
Reporting a measure of center gives only partial information about a data set. It is also important to describe how much the observations differ from one another. The three different samples displayed in Figure 4.7 all have mean = median = 45. There is a lot of variability in the first sample compared to the third sample. The second sample shows less variability than the first and more variability than the third. Most of the variability in the second sample is due to the two extreme values being so far from the center.

FIGURE 4.7

Three samples with the same center and different amounts of variability.

Sample

1. 20, 40, 50, 30, 60, 70
2. 47, 43, 44, 46, 20, 70
3. 44, 43, 40, 50, 47, 46



The simplest numerical measure of variability is the range.

DEFINITION

Range: The **range** of a data set is defined as Range = Largest Value – Smallest Value

In general, more variability will result in a larger range. However, variability is a characteristic of the entire data set, and each observation contributes to variability. The first two samples plotted in Figure 4.7 both have a range of $70 - 20 = 50$, but there is less variability in the second sample. Because it is calculated using only the largest and smallest values in the data set, the range is not usually the best measure of variability.

Deviations from the Mean

The most widely used measures of variability describe the extent to which the sample observations deviate from the sample mean \bar{x} . Subtracting \bar{x} from each observation gives a set of **deviations from the mean**.

DEFINITION

Deviations from the mean: The n deviations from the sample mean are the differences

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \dots, (x_n - \bar{x})$$

Notice that a deviation will be positive if the corresponding x value is greater than \bar{x} and negative if the x value is less than \bar{x} .

Example 4.7 The Big Mac Index

Understand the context ➤

- McDonald's fast-food restaurants are now found in many countries around the world. But the cost of a Big Mac varies from country to country. Table 4.3 shows data on the cost of a Big Mac for 12 European Union countries (converted to U.S. dollars) taken from the article “[The Big Mac Index 2015](#)” ([bigmacindex.org](#), January 22, 2015, retrieved April 18, 2017).

TABLE 4.3 Big Mac Prices for 12 European Union Countries

Consider the data ➤

Country	2015 Big Mac Price in U.S. Dollars
Austria	3.93
Belgium	4.29
Estonia	3.36
Finland	4.75
France	4.52
Germany	4.25
Greece	3.53
Ireland	4.04
Italy	4.46
Netherlands	4.00
Portugal	3.48
Spain	4.23

Notice that there is quite a bit of variability in the Big Mac prices.

For this data set, $\Sigma x = 48.84$ and $\bar{x} = \$4.07$. Table 4.4 displays the data along with the corresponding deviations, formed by subtracting $\bar{x} = 4.07$ from each observation. Six of the deviations are positive because six of the observations are greater than \bar{x} . The negative deviations correspond to observations that are less than \bar{x} . Some of the deviations are quite large in magnitude (0.68 and -0.71 , for example), indicating observations that are far from the sample mean.

Table 4.4 Deviations from the Mean for the Big Mac Data

2015 Big Mac Price in U.S. Dollars	Deviation from the Mean ($x - \bar{x}$)
3.93	-0.14
4.29	0.22
3.36	-0.71
4.75	0.68
4.52	0.45
4.25	0.18
3.53	-0.54
4.04	-0.03
4.46	0.39
4.00	-0.07
3.48	-0.59
4.23	0.16

In general, the greater the amount of variability in the sample data, the larger the magnitudes (ignoring the signs) of the deviations. We now consider how to combine the deviations into a single numerical measure of variability. A first thought might be to calculate the average deviation, by adding the deviations together and then dividing by n . This does not work, though, because when we calculate the sum of the deviations, denoted by $\Sigma(x - \bar{x})$, negative and positive deviations offset each other. In fact, except for small differences due to rounding, the sum of the deviations from the mean is always equal to 0. Notice that the value of the sum of the 12 deviations in Example 4.7 is $\Sigma(x - \bar{x}) = 0$.

Except for the effects of rounding in calculating the deviations, it is always true that

$$\Sigma(x - \bar{x}) = 0$$

Since this sum is zero, the average deviation is always zero and so it cannot be used as a measure of variability.

The Variance and Standard Deviation

One way to prevent negative and positive deviations from offsetting one another is to square them before combining. Then deviations with opposite signs but with the same magnitude, such as +2 and -2, make identical contributions to a measure of variability. The squared deviations are $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, \dots, (x_n - \bar{x})^2$ and the sum is

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 = \Sigma(x - \bar{x})^2$$

Dividing this sum by the sample size n gives the average squared deviation. Although this seems to be a reasonable measure of variability, we use a divisor slightly smaller than n . (The reason for this will be explained later in this section and in Chapter 9.)

DEFINITIONS

Sample variance: The **sample variance**, denoted by s^2 , is the sum of squared deviations from the mean divided by $n - 1$. That is,

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

Sample standard deviation: The **sample standard deviation** is the positive square root of the sample variance and is denoted by s .

A large amount of variability in a sample is indicated by a relatively large value of s^2 or s , whereas a value of s^2 or s close to zero indicates a smaller amount of variability. People find the standard deviation to be a more natural measure of variability than the variance because the standard deviation is expressed in the same units as the original data values. For example, if the observations in the data set are the prices of a Big Mac in dollars, the mean and the deviations from the mean are also in dollars. But when the deviations are squared and then combined to calculate the value of the variance, the units are dollars squared—not something that is familiar to most people! This makes it difficult to interpret the value of the variance and to decide whether the variance is large or small. Taking the square root of the variance to obtain the standard deviation results in a measure of variability that is in the same units as the original data values, making it easier to interpret.

Example 4.8 | Big Mac Revisited

Let's continue using the Big Mac data and the calculated deviations from the mean given in Example 4.7 to calculate the sample variance and standard deviation. Table 4.5 shows

the observations, deviations from the mean, and squared deviations. Adding the squared deviations to calculate the values of s^2 and s gives

$$\Sigma(x - \bar{x})^2 = 2.0926$$

and

$$s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1} = \frac{2.0926}{12 - 1} = \frac{2.0926}{11} = 0.1902$$

$$s = \sqrt{0.1902} = 0.436$$

Table 4.5 Deviations and Squared Deviations for the Big Mac Data

2015 Big Mac Price in U.S. Dollars	Deviation from the Mean ($x - \bar{x}$)	Squared Deviation from the Mean ($x - \bar{x}$) ²
3.93	-0.14	0.0196
4.29	0.22	0.0484
3.36	-0.71	0.5041
4.75	0.68	0.4624
4.52	0.45	0.2025
4.25	0.18	0.0324
3.53	-0.54	0.2916
4.04	-0.03	0.0009
4.46	0.39	0.1521
4.00	-0.07	0.0049
3.48	-0.59	0.3481
4.23	0.16	0.0256
		$\Sigma(x - \bar{x})^2 = 2.0926$

The calculation of s^2 can be a bit tedious, especially if the sample size is large. Fortunately, many calculators and computer software packages compute the variance and standard deviation. One commonly used statistical computer package is Minitab. The output resulting from using the Minitab Describe command with the Big Mac data follows. Minitab gives a variety of numerical descriptive measures, including the mean, the median, and the standard deviation.

Descriptive Statistics: Big Mac Price in Dollars										
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	
Big Mac Price in Dollars	12	0	4.070	0.126	0.436	3.360	3.630	4.135	4.417	
Variable	Maximum									
Big Mac Price in Dollars	4.750									

The standard deviation can be informally interpreted as the size of a “typical” or “representative” deviation from the mean. In Example 4.8, a typical deviation from \bar{x} is about 0.436. Some observations are closer to \bar{x} than 0.436 and others are farther away.

We calculated $s = 0.436$ in Example 4.8 without saying whether this value indicated a large or a small amount of variability. At this point, it is better to use s for comparative purposes than for an absolute assessment of variability. If Big Mac prices for a larger group of countries resulted in a standard deviation of $s = 1.153$ (this is the standard deviation for all 56 countries for which 2015 Big Mac data was available) then we would conclude that our original sample has much less variability than the data set consisting of all 56 countries.

There are measures of variability for the entire population that are analogous to s^2 and s for a sample. These measures are called the **population variance** and the **population standard deviation** and are denoted by σ^2 and σ , respectively. (We again use lowercase Greek letters for population characteristics.)

Notation

s^2	sample variance
σ^2	population variance
s	sample standard deviation
σ	population standard deviation

In many statistical procedures, we would like to use the value of σ , but unfortunately it is not usually known. Therefore, we must estimate σ using a value calculated from the sample. The divisor $(n - 1)$ is used in calculating s^2 rather than n because, on average, the resulting value tends to be a bit closer to the value of σ^2 . We will say more about this in Chapter 9.

An alternative rationale for using $(n - 1)$ is based on the property $\Sigma(x - \bar{x}) = 0$. Suppose that $(n = 5)$ and that four of the deviations are

$$(x_1 - \bar{x}) = -4 \quad (x_2 - \bar{x}) = 6 \quad (x_3 - \bar{x}) = 1 \quad (x_5 - \bar{x}) = -8$$

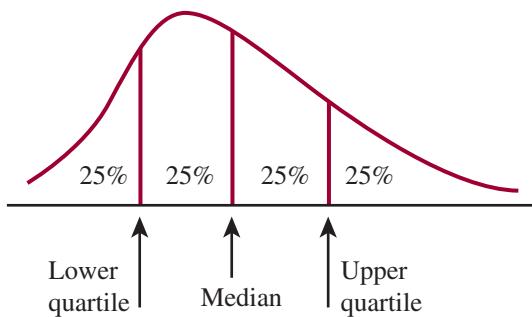
Then, because the sum of these four deviations is -5 , the remaining deviation must be $(x_4 - \bar{x}) = 5$ (so that the sum of all five deviations is zero). More generally, once any $(n - 1)$ of the deviations are known, the value of the remaining deviation is determined. The n deviations actually contain only $(n - 1)$ independent pieces of information about variability. Statisticians express this by saying that s^2 and s are based on $(n - 1)$ *degrees of freedom* (df).

The Interquartile Range

As with \bar{x} , the value of s can be greatly affected by the presence of even a single unusually small or large observation. The **interquartile range** is a measure of variability that is resistant to the effects of outliers. It is based on quantities called quartiles. The **lower quartile** separates the smallest 25% of the data set from the greatest 75%, and the **upper quartile** separates the greatest 25% from the smallest 75%. The middle quartile is the median, and it separates the lower 50% from the upper 50%. Figure 4.8 illustrates the locations of these quartiles for a smoothed histogram.

FIGURE 4.8

The quartiles for a smoothed histogram.



The quartiles for sample data are obtained by dividing the n ordered observations into a lower half and an upper half. If n is odd, the median is excluded from both halves. The upper and lower quartiles are then the medians of the two halves. (Note: The median is only temporarily excluded for the purpose of calculating quartiles. It is not deleted from the data set.)

Lower quartile: Median of the lower half of the sample

Upper quartile: Median of the upper half of the sample

(If n is odd, the median of the entire sample is excluded from both halves when calculating quartiles.)

Interquartile range (iqr): A measure of variability that is not as sensitive to the presence of outliers as the standard deviation. The iqr is calculated as

$$\text{iqr} = \text{upper quartile} - \text{lower quartile}$$

*There are several other sensible ways to define quartiles. Some calculators and software packages use alternative definitions.

The resistant nature of the interquartile range follows from the fact that up to 25% of the smallest sample observations and up to 25% of the largest sample observations can be made more extreme without affecting the value of the interquartile range.

Example 4.9 Beating That High Score

Understand the context ➤

The authors of the paper “[Striatal Volume Predicts Level of Video Game Skill Acquisition](#)” (*Cerebral Cortex [2010]: 2522–2530*) studied a number of factors that affect performance on a complex video game. One of the factors investigated was practice strategy. Forty college students who all reported that they played video games less than 3 hours per week over the 2 years prior to the study and who had never played the game Space Fortress were assigned at random to one of two practice strategies. In the Space Fortress game, points are awarded for control, velocity, and speed.

Each person completed 20 two-hour practice sessions. Those in one group, the fixed priority group, were told to work on improving their total score at each practice session. Those in the other group, the variable priority group, were told to focus on a particular aspect of the game, such as improving speed score in each practice session, and the focus changed from one practice session to another. The investigators were interested in whether practice strategy makes a difference. They measured the improvement in total score from the first day of the study to the day when the last practice session was completed.

Improvement scores (approximate values read from a graph that appears in the paper) for the 20 people in the variable priority practice strategy group are given here.

Consider the data ➤

1200	1300	2300	3200	3300	3800	4000	4100	4300	4800
5500	5700	5700	5800	6000	6300	6800	6800	6900	7700

The median is the average of the middle two observations, so

$$\text{median} = \frac{4800 + 5500}{2} = 5150$$

The lower half of the data set is

1200 1300 2300 3200 3300 3800 4000 4100 4300 4800

so the lower quartile is the median of the lower half

$$\text{lower quartile} = \frac{3300 + 3800}{2} = 3550$$

The upper half of the data set is

5500 5700 5700 5800 6000 6300 6800 6800 6900 7700

so the upper quartile is the median of the upper half

$$\text{upper quartile} = \frac{6000 + 6300}{2} = 6150$$

$$\text{lower quartile} = 3550$$

$$\text{upper quartile} = 6150$$

$$\text{iqr} = 6150 - 3550 = 2660$$

The sample mean and standard deviation for this data set are 4775 and 1867, respectively. If we were to change the two largest values from 6900 and 7700 to 10,900 and 11,700 (so that they still remain the two largest values), the median and interquartile range would not be affected, but the mean and the standard deviation would change to 7500 and 2390, respectively. The value of the interquartile range is not affected by a few extreme values in the data set.

EXERCISES 4.17 - 4.33

● Data set available online

- 4.17** ● The following data are costs (in cents) per ounce for nine different brands of sliced Swiss cheese (consumerreports.org):

29 62 37 41 70 82 47 52 49

- Calculate the variance and standard deviation for this data set. (Hint: See Example 4.8.)
- If a very expensive cheese with a cost per slice of \$1.50 (150 cents) was added to the data set, how would the values of the mean and standard deviation change?

- 4.18** ● Cost per serving (in cents) for six high-fiber cereals rated very good and for nine high-fiber cereals rated good by *Consumer Reports* are shown below. Write a few sentences describing how these two data sets differ with respect to center and variability. Use summary statistics to support your statements.

Cereals Rated Very Good

46 49 62 41 19 77

Cereals Rated Good

71 30 53 53 67 43 48 28 54

- 4.19** ● Combining the cost-per-serving data for high-fiber cereals rated very good and those rated good from the previous exercise gives the following data set:

46 49 62 41 19 77 71 30
53 53 67 43 48 28 54

- Calculate the quartiles and the interquartile range for this combined data set. (Hint: See Example 4.9.)
- Calculate the interquartile range for just the cereals rated good. Is this value greater than, less than, or about equal to the interquartile range computed in Part (a)?

- 4.20** ● The paper “Caffeinated Energy Drinks—A Growing Problem” (*Drug and Alcohol Dependence* [2009]: 1–10) gave the accompanying data on caffeine per ounce for eight top-selling energy drinks and for 11 high-caffeine energy drinks:

Top-Selling Energy Drinks

9.6 10.0 10.0 9.0 10.9 8.9 9.5 9.1

High-Caffeine Energy Drinks

21.0	25.0	15.0	21.5	35.7	15.0
33.3	11.9	16.3	31.3	30.0	

The mean caffeine per ounce is clearly higher for the high-caffeine energy drinks, but which of the two groups of energy drinks (top-selling or high-caffeine) is the most variable with respect to caffeine per ounce? Justify your choice based on a measure of variability.

- 4.21** The accompanying data are consistent with summary statistics that appeared in the paper “Shape of Glass and Amount of Alcohol Poured: Comparative Study of Effect of Practice and Concentration” (*British Medical Journal* [2005]: 1512–1514). Data represent the actual amount poured (in ml) into a tall, slender glass for individuals who were asked to pour 44.3 ml (1.5 ounces). Calculate and interpret the values of the mean and standard deviation.

44.0 49.6 62.3 28.4 39.1 39.8 60.5 73.0 57.5
56.5 65.0 56.2 57.7 73.5 66.4 32.7 40.4 21.4

- 4.22** The paper referenced in the previous exercise also gave data on the actual amount poured (in ml) into a short, wide glass for individuals who were asked to pour 44.3 ml (1.5 ounces).

89.2 68.6 32.7 37.4 39.6 46.8 66.1 79.2 66.3
52.1 47.3 64.4 53.7 63.2 46.4 63.0 92.4 57.8

- Calculate and interpret the values of the mean and standard deviation.
- What do the values of the mean amount poured in the short, wide glass and the mean calculated in the previous exercise suggest about the shape of glasses used?

- 4.23** The price (in dollars) of the eight smart phones that were rated highest by *Consumer Reports* in 2018 (consumerreports.org, retrieved February 23, 2018) were

730 850 830 800 700 1000 950 520

- a. Calculate the values of the variance and the standard deviation.
- b. The standard deviation is quite large. What does that tell you about the prices of these highly rated smart phones?
- 4.24** In addition to the prices of the highly rated smart phones given in the previous exercise, *Consumer Reports* also gave the prices of the seven smart phones that received the lowest ratings. Those prices (in dollars) were
- | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|
| 110 | 130 | 100 | 145 | 200 | 180 | 135 |
|-----|-----|-----|-----|-----|-----|-----|
- Comment on how the highest rated smart phones and the lowest rated smart phones differ with respect to price and price variability.
- 4.25** In an experiment to assess the effect of listening to audiobooks while driving, participants were asked to drive down a straight road in a driving simulator. The accompanying data on time (in milliseconds) to react when a pedestrian walked into the street for 10 drivers listening to an audiobook are consistent with summary statistics and graphs that appeared in the paper “Good Distractions: Testing the Effect of Listening to an Audiobook on Driving Performance in Simple and Complex Road Environments” (*Accident Analysis and Prevention* [2018]: 202–209). Calculate the variance and the standard deviation for this data set.
- | | | | | | | | | | |
|------|------|------|-----|-----|------|------|------|-----|------|
| 1014 | 1009 | 1052 | 982 | 931 | 1020 | 1069 | 1011 | 860 | 1108 |
|------|------|------|-----|-----|------|------|------|-----|------|
- 4.26** The paper referenced in the previous exercise also gave summary statistics and graphs for the reaction time of drivers who were not listening to audiobooks. Data on reaction time (in milliseconds) consistent with those summary statistics for 10 drivers not listening to audiobooks are given here.
- | | | | | | | | | | |
|-----|-----|------|-----|------|------|------|-----|-----|-----|
| 961 | 904 | 1010 | 976 | 1018 | 1041 | 1004 | 981 | 995 | 991 |
|-----|-----|------|-----|------|------|------|-----|-----|-----|
- a. Use the data given in this exercise and the data given in the previous exercise to construct dot-plots that would allow you to compare the reaction times for the two groups.
- b. Based on the dot plots, do you think that the standard deviation of the reaction times for people who are not listening to audiobooks would be less than, about the same as, or greater than the standard deviation that you calculated in the previous exercise for drivers who were listening to audiobooks? Explain your thinking.
- c. Calculate the standard deviation of the reaction times for the drivers who were not listening to audiobooks. Is the value of this standard deviation consistent with your answer in Part (b)?
- d. Describe how the distributions of reaction time differ for drivers who are listening to audiobooks and those who are not.

- 4.27** ● The accompanying data on number of minutes used for cell phone calls in 1 month was generated to be consistent with summary statistics published in a report of a marketing study of San Diego residents (*TeleTruth, March 2009*):

189	0	189	177	106	201	0	212	0	306
0	0	59	224	0	189	142	83	71	165
236	0	142	236	130					

- a. Calculate the values of the quartiles and the interquartile range for this data set.
- b. Explain why the lower quartile is equal to the minimum value for this data set. Will this be the case for every data set? Explain.

- 4.28** Give two sets of five numbers that have the same mean but different standard deviations, and give two sets of five numbers that have the same standard deviation but different means.

- 4.29** Morningstar is an investment research firm that publishes some online educational materials. The materials for an online course called “[Looking at Historical Risk](http://news.morningstar.com/classroom2/course.asp?docId=2927&page=2&CN=com)” (news.morningstar.com/classroom2/course.asp?docId=2927&page=2&CN=com, retrieved August 3, 2016) included the following paragraph referring to annual return (in percent) for investment funds:

Using standard deviation as a measure of risk can have its drawbacks. It’s possible to own a fund with a low standard deviation and still lose money. In reality, that’s rare. Funds with modest standard deviations tend to lose less money over short time frames than those with high standard deviations. For example, the one-year average standard deviation among ultrashort-term bond funds, which are among the lowest-risk funds around (other than money market funds), is a mere 0.64%.

- a. Explain why the standard deviation of percent return is a reasonable measure of unpredictability and why a smaller standard deviation for the percent return of an investment fund means less risk.
- b. Explain how a fund with a small standard deviation can still lose money. (Hint: Think about the average percent return.)

- 4.30** ● The [U.S. Department of Transportation](http://www.safercar.gov/crashstats/) reported the data in the accompanying table on the number of speed-related crash fatalities during holiday periods for the years from 1994 to 2003 (*Traffic Safety Facts, July 20, 2005*).

- a. Calculate the standard deviation for the New Year’s Day data.
- b. Without calculating the standard deviation of the Memorial Day data, explain whether the standard deviation for the Memorial Day data would be larger or smaller than the standard deviation of the New Year’s Day data.

Data For Exercise 4.30

Holiday Period	Speed-Related Fatalities									
	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
New Year's Day	141	142	178	72	219	138	171	134	210	70
Memorial Day	193	178	185	197	138	183	156	190	188	181
July 4th	178	219	202	179	169	176	219	64	234	184
Labor Day	183	188	166	179	162	171	180	138	202	189
Thanksgiving	212	198	218	210	205	168	187	217	210	202
Christmas	152	129	66	183	134	193	155	210	60	198

- c. Memorial Day and Labor Day are holidays that always occur on Monday and Thanksgiving always occurs on a Thursday, whereas New Year's Day, July 4th, and Christmas do not always fall on the same day of the week every year. Based on the given data, is there more or less variability in the speed-related crash fatality numbers from year to year for same day of the week holiday periods than for holidays that can occur on different days of the week? Support your answer with appropriate measures of variability.

- 4.31** The Ministry of Health and Long-Term Care in Ontario, Canada, publishes information on the time that patients must wait for various medical procedures on its web site (health.gov.on.ca). For two cardiac procedures completed in fall of 2005, the following information was provided:

Procedure	Number of Completed Procedures	Median Wait Time (days)	Mean Wait Time (days)	90% Completed Within (days)
Angioplasty	847	14	18	39
Bypass surgery	539	13	19	42

- a. Which of the following must be true for the lower quartile of the data set consisting of the 847 wait times for angioplasty?
- i. The lower quartile is less than 14.
 - ii. The lower quartile is between 14 and 18.
 - iii. The lower quartile is between 14 and 39.
 - iv. The lower quartile is greater than 39.
- b. Which of the following must be true for the upper quartile of the data set consisting of the 539 wait times for bypass surgery?
- i. The upper quartile is less than 13.
 - ii. The upper quartile is between 13 and 19.
 - iii. The upper quartile is between 13 and 42.
 - iv. The upper quartile is greater than 42.
- c. Which of the following must be true for the number of days for which only 5% of the bypass surgery wait times would be longer?

- i. It is less than 13.
- ii. It is between 13 and 19.
- iii. It is between 13 and 42.
- iv. It is greater than 42.

- 4.32** In 1997, a woman sued a computer keyboard manufacturer, charging that her repetitive stress injuries were caused by the keyboard (*Genessey v. Digital Equipment Corporation*). The jury awarded about \$3.5 million for pain and suffering, but the court then set aside that award as being unreasonable compensation. In making this determination, the court identified a “normative” group of 27 similar cases and specified a reasonable award as one within 2 standard deviations of the mean of the awards in the 27 cases.

The 27 award amounts were (in thousands of dollars)

37	60	75	115	135	140	149	150
238	290	340	410	600	750	750	750
1050	1100	1139	1150	1200	1200	1250	1576
1700	1825	2000					

What is the maximum possible amount that could be awarded under the “2-standard deviations rule”?

- 4.33** The standard deviation alone does not measure relative variation. For example, a standard deviation of \$1 would be considered large if it is describing the variability from store to store in the price of an ice cube tray. On the other hand, a standard deviation of \$1 would be considered small if it is describing store-to-store variability in the price of a particular brand of freezer.

A quantity designed to give a relative measure of variability is the *coefficient of variation*. Denoted by CV, the coefficient of variation expresses the standard deviation as a percentage of the mean. It is

defined by the formula $CV = 100\left(\frac{s}{\bar{x}}\right)$.

Consider two samples. Sample 1 gives the actual weight (in ounces) of the contents of cans of pet food labeled as having a net weight of 8 ounces. Sample 2 gives the actual weight (in pounds) of the contents of

bags of dry pet food labeled as having a net weight of 50 pounds. The weights for the two samples are

Sample 1	8.3	7.1	7.6	8.1	7.6
	8.3	8.2	7.7	7.7	7.5
Sample 2	52.3	50.6	52.1	48.4	48.8
	47.0	50.4	50.3	48.7	48.2

- a. For each of the given samples, calculate the mean and the standard deviation.
- b. Calculate the coefficient of variation for each sample. Do the results surprise you? Why or why not?

SECTION 4.3 Summarizing a Data Set: Boxplots

In Sections 4.1 and 4.2, we looked at ways to describe the center and variability of a data set using numerical measures. It would be nice to have a method of summarizing data that provides more information than just reporting a measure of center and spread and yet less detail than a stem-and-leaf display or histogram. A **boxplot** is one way to do this. A boxplot is compact, yet it provides information about the center, variability, and symmetry or skewness of the data. We will consider two types of boxplots: the skeletal boxplot and the modified boxplot.

Construction of a Skeletal Boxplot

1. Draw a horizontal (or vertical) measurement scale.
2. Construct a rectangular box with a left (or lower) edge at the lower quartile and a right (or upper) edge at the upper quartile. This makes the box width equal to the iqr.
3. Draw a vertical (or horizontal) line segment inside the box at the location of the median.
4. Extend horizontal (or vertical) line segments, called whiskers, from each end of the box to the smallest and largest observations in the data set.

Example 4.10 Revisiting Improvement Score Data

Let's reconsider the data from Example 4.9 on improvement score for people in the variable priority practice group. The ordered observations are

Ordered Data

Lower Half:

1200 1300 2300 3200 3300 3800 4000 4100 4300 4800

Median = 5150

Upper Half:

5500 5700 5700 5800 6000 6300 6800 6800 6900 7700

To construct a boxplot of these data, we need the following information: the smallest observation, the lower quartile, the median, the upper quartile, and the largest observation. This collection of summary measures is often referred to as a **five-number summary**. For this data set we have

smallest observation = 1200

lower quartile = median of the lower half = 3550

median = average of the 10th and 11th observations in the ordered list = 5150

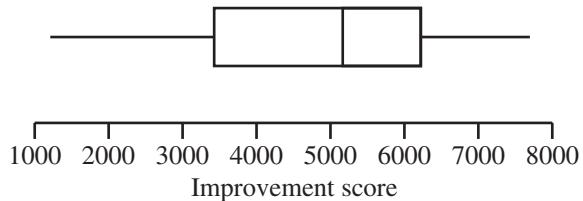
upper quartile = median of the upper half = 6150

largest observation = 7700

Figure 4.9 shows the corresponding boxplot. The median line is somewhat closer to the upper edge of the box than to the lower edge, suggesting a concentration of values in the upper part of the middle half. The lower whisker is longer than the upper whisker, suggesting that the distribution of improvement scores for the variable practice strategy group is not symmetric.

FIGURE 4.9

Skeletal boxplot for the improvement score data of Example 4.10.



Boxplots are often used to compare groups. For example, we could compare video game improvement scores for the two different practice strategies described in Example 4.9 using a **comparative boxplot**.

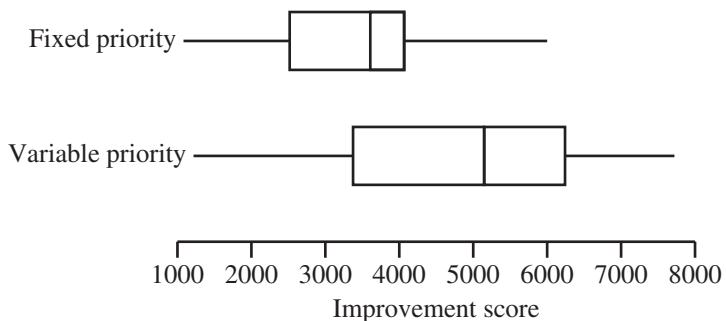
DEFINITION

Comparative boxplot: Two or more boxplots drawn using the same numerical scale.

The video game improvement scores for the two different strategy groups were used to construct the comparative boxplot shown in Figure 4.10.

FIGURE 4.10

Comparative boxplots for improvement scores of fixed priority practice group and variable priority practice group.



From the comparative boxplot, we can see that both data distributions are approximately symmetric. The improvement scores tend to be higher for the variable priority practice group than for the fixed priority practice group. However, there is more consistency in the improvement scores for the fixed priority practice group. Improvement scores in the variable priority practice group are more spread out, indicating more variability in improvement scores for this group.

The sequence of steps used to construct a skeletal boxplot is easily modified to include information about outliers.

DEFINITION

Outlier: An observation that is more than $1.5(iqr)$ away from the nearest quartile (the nearest end of the box).

An outlier is an **extreme outlier** if it is more than $3(iqr)$ from the nearest quartile and it is a **mild outlier** otherwise.

A **modified boxplot** represents mild outliers by solid circles and extreme outliers by open circles, and the whiskers extend on each end to the most extreme observations that are *not* outliers.

Construction of a Modified Boxplot

1. Draw a horizontal (or vertical) measurement scale.
2. Construct a rectangular box with a left (or lower) edge at the lower quartile and right (or upper) edge at the upper quartile. This makes the box width equal to the iqr.
3. Draw a vertical (or horizontal) line segment inside the box at the location of the median.
4. Determine if there are any mild or extreme outliers in the data set.
5. Draw whiskers that extend from each end of the box to the most extreme observation that is *not* an outlier.
6. Draw a solid circle to mark the location of any mild outliers in the data set.
7. Draw an open circle to mark the location of any extreme outliers in the data set.

Example 4.11 Golden Rectangles

Understand the context ➤

- The accompanying data came from an anthropological study of rectangular shapes (*Lowie's Selected Papers in Anthropology, Cora Dubios, ed. [Berkeley, CA: University of California Press, 1960]: 137–142*). Observations were made on the variable x = width/length for a sample of $n = 20$ beaded rectangles used in Shoshoni Indian leather handicrafts:

Consider the data ➤

0.553	0.570	0.576	0.601	0.606	0.606	0.609	0.611	0.615	0.628
0.654	0.662	0.668	0.670	0.672	0.690	0.693	0.749	0.844	0.933

The quantities needed for constructing the modified boxplot have been calculated and are shown here:

Do the work ➤

median = 0.641	iqr = $0.681 - 0.606 = 0.075$
lower quartile = 0.606	$1.5(\text{iqr}) = 0.1125$
upper quartile = 0.681	$3(\text{iqr}) = 0.225$

Then,

$$\begin{aligned} (\text{upper quartile}) + 1.5(\text{iqr}) &= 0.681 + 0.1125 = 0.7935 \\ (\text{lower quartile}) - 1.5(\text{iqr}) &= 0.606 - 0.1125 = 0.4935 \end{aligned}$$

This means that 0.844 and 0.933 are both outliers on the upper end (because they are greater than 0.7935). There are no outliers on the lower end (because no observations are less than 0.4935). Because

$$(\text{upper quartile}) + 3(\text{iqr}) = 0.681 + 0.225 = 0.906$$

0.933 is an extreme outlier and 0.844 is only a mild outlier. The upper whisker extends to the largest observation that is not an outlier, which is 0.749. The lower whisker extends to the smallest observation that is not an outlier, which is 0.553. The boxplot is shown in

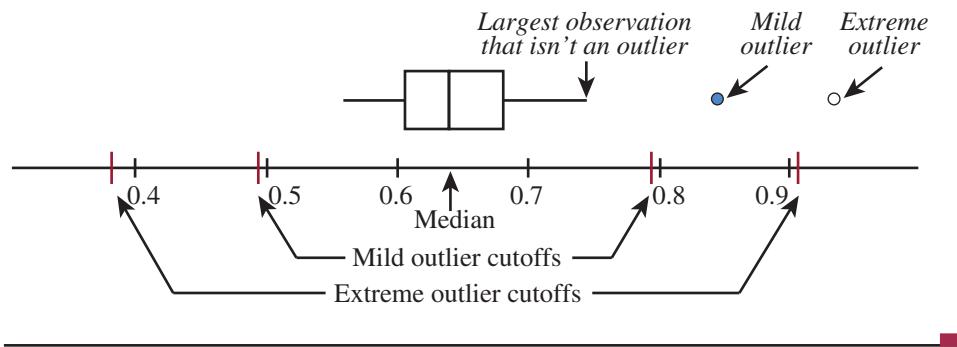
Interpret the results ➤

Figure 4.11. The median line is not at the center of the box, so there is a slight asymmetry in the middle half of the data. However, the most striking feature is the presence of the two outliers. These two x values considerably exceed the “golden ratio” of 0.618, used since antiquity as an aesthetic standard for rectangles.

● Data set available online

FIGURE 4.11

Boxplot for the rectangle data in Example 4.11.



Example 4.12 Another Look at Big Mac Prices

Understand the context ➤

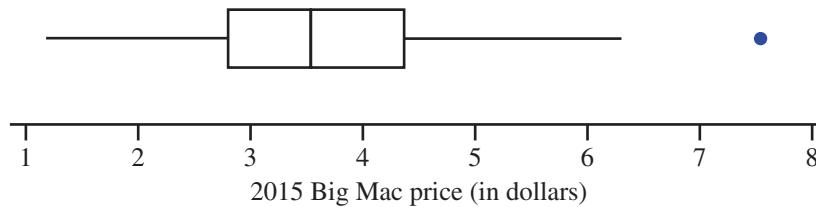
Big Mac prices in U.S. dollars for 55 different countries in 2015 were given in the article “[The Big Mac Index](#)” first introduced in Example 4.7. The 55 Big Mac prices were:

Consider the data ➤	3.25	4.32	5.21	4.37	4.64	3.35	2.77	3.34	4.01	2.92
	5.38	2.30	2.43	3.17	1.89	2.24	4.45	3.14	2.11	3.35
	4.49	6.30	2.98	3.32	3.67	2.48	1.36	2.93	3.53	2.22
	3.78	2.65	4.97	7.54	2.51	3.04	3.96	3.54	1.20	4.79
	4.63	2.53	2.81	3.93	4.29	3.36	4.75	4.52	4.25	3.53
	4.04	4.46	4.00	3.48	4.23					

Figure 4.12 shows a Minitab boxplot for the Big Mac price data. Note that the upper whisker is longer than the lower whisker and that there is one outlier on the high end (Switzerland with a Big Mac price of \$7.54).

FIGURE 4.12

Minitab boxplot of the Big Mac price data of Example 4.12.



Notice that Minitab does not distinguish between mild outliers and extreme outliers in the boxplot. For the Big Mac price data,

Do the work ➤

$$\begin{aligned} \text{lower quartile} &= 2.81 \\ \text{upper quartile} &= 4.37 \\ \text{iqr} &= 4.37 - 2.81 = 1.56 \end{aligned}$$

Then

$$\begin{aligned} 1.5(\text{iqr}) &= 2.34 \\ 3(\text{iqr}) &= 4.68 \end{aligned}$$

We can calculate outlier boundaries as follows:

$$\begin{aligned} \text{upper quartile} + 1.5(\text{iqr}) &= 4.37 + 2.34 = 6.71 \\ \text{upper quartile} + 3(\text{iqr}) &= 4.37 + 4.68 = 9.05 \end{aligned}$$

The observation for Switzerland (7.54) is a mild outlier because it is greater than 6.71 (the upper quartile + 1.5(iqr)) but less than 9.05 (the upper quartile + 3(iqr)). There are no extreme outliers in this data set.

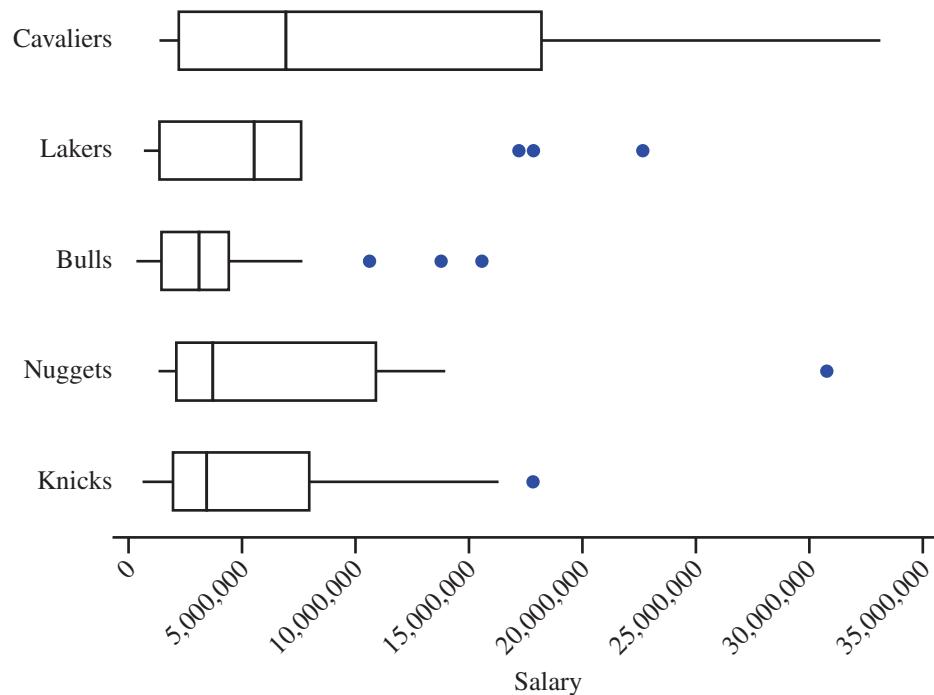
Example 4.13 NBA Salaries Revisited

Understand the context ▶

The 2017–2018 salaries of NBA players published on the web site hoopshype.com were used to construct the comparative boxplot of the salary data for five teams shown in Figure 4.13.

FIGURE 4.13

Comparative boxplot of salaries for five NBA teams.



Interpret the results ▶

The comparative boxplot reveals some interesting similarities and differences in the salary distributions of the five teams. The minimum salary was lower for the Bulls and the Knicks and highest for the Cavaliers. The median salary was about the same for the Nuggets and the Knicks. The median salaries for both the Lakers and the Cavaliers were higher than the upper quartile for the Bulls, which was around \$5 million. This means that more than half the players on the Lakers and the Cavaliers made more than \$5 million, but only a quarter of the players on the Bulls had salaries that were this high. Notice also that there were outliers in the salary distributions of four of the teams, meaning that there were a few players on each of these teams that had salaries that were quite a bit higher than the other players on their team. Even though the highest salary was for a player on the Cavaliers, this salary was not an outlier because the salaries in the upper half were quite spread out, indicating more variability in the Cavalier salaries than in the salaries of the other four teams. Notice that the boxplot for the Laker salaries looks like it is missing the line segment that would extend from the upper quartile to the smallest observation that was not an outlier. This is because those observations were so close to the upper quartile that the line segment cannot be distinguished from the end of the box.

EXERCISES 4.34 - 4.41

● Data set available online

- 4.34** Based on a large national sample of working adults, the [U.S. Census Bureau](#) reports the following information on travel time to work for those who do not work at home:
- lower quartile = 7 minutes
median = 18 minutes
upper quartile = 31 minutes

Also given was the mean travel time, which was reported as 22.4 minutes.

- a. Is the travel time distribution more likely to be approximately symmetric, positively skewed, or negatively skewed? Explain your reasoning based on the given summary quantities.

- b. Suppose that the minimum travel time was 1 minute and that the maximum travel time in the sample was 205 minutes. Construct a skeletal boxplot for the travel time data. (Hint: See Example 4.10.)
- c. Were there any mild or extreme outliers in the data set? How can you tell? (Hint: See Example 4.11.)
- 4.35** The report “[Most Licensed Drivers Age 85+: States](#)” ([bloomberg.com/graphics/best-and-worst/#most-licensed-drivers-age-85-plus-states](#), retrieved April 20, 2017) gives the percentage of drivers in each state and the District of Columbia in 2011 who were over 85 years of age.

State	Percentage of Drivers Over Age 85	State	Percentage of Drivers Over Age 85
Alaska	0.72	Ohio	1.62
Nevada	0.96	West Virginia	1.64
New Mexico	0.98	Illinois	1.66
Utah	1.05	Michigan	1.69
Georgia	1.11	Colorado	1.72
Hawaii	1.13	Delaware	1.72
Texas	1.13	Oklahoma	1.72
California	1.17	Arkansas	1.75
Montana	1.20	Iowa	1.78
Virginia	1.22	Massachusetts	1.78
Arizona	1.25	Kansas	1.81
Mississippi	1.26	New Jersey	1.85
Maryland	1.28	Oregon	1.90
Kentucky	1.30	Louisiana	1.92
Idaho	1.33	North Dakota	1.96
Washington	1.33	Rhode Island	1.97
Indiana	1.34	Florida	1.98
North Carolina	1.37	Pennsylvania	1.98
South Carolina	1.37	South Dakota	1.99
District of Columbia	1.43	Nebraska	2.02
Tennessee	1.48	New York	2.09
Wyoming	1.48	Minnesota	2.11
Missouri	1.49	Vermont	2.22
New Hampshire	1.50	Alabama	2.23
Wisconsin	1.61	Maine	2.34
		Connecticut	5.10

- a. Find the values of the median, the lower quartile, and the upper quartile.
- b. The largest value in the data set is 5.10% (Connecticut). Is this state an outlier? (Hint: See Example 4.11.)
- c. Construct a modified boxplot for this data set and comment on the interesting features of the plot. How would you describe the shape of the distribution if you don't consider the outlier?

- 4.36** Data on the gasoline tax per gallon (in cents) in 2015 for the 50 U.S. states and the District of Columbia are shown below ([eia.gov/tools/faqs/faq.cfm?id=10&t=10](#), retrieved September 1, 2016).

State	Gasoline Tax (cents per gallon)	State	Gasoline Tax (cents per gallon)
Alabama	19.0	Montana	27.8
Alaska	9.0	Nebraska	27.7
Arizona	19.0	Nevada	23.8
Arkansas	21.8	New Hampshire	23.8
California	37.2	New Jersey	14.6
Colorado	23.3	New Mexico	18.9
Connecticut	25.0	New York	33.8
Delaware	23.0	North Carolina	35.3
District of Columbia	23.5	North Dakota	23.0
Florida	30.6	Ohio	28.0
Georgia	26.5	Oklahoma	17.0
Hawaii	18.5	Oregon	30.0
Idaho	33.0	Pennsylvania	51.4
Illinois	33.1	Rhode Island	34.1
Indiana	29.0	South Carolina	16.8
Iowa	31.8	South Dakota	30.0
Kansas	25.0	Tennessee	21.4
Kentucky	26.0	Texas	20.0
Louisiana	20.9	Utah	30.1
Maine	31.4	Vermont	30.5
Maryland	32.8	Virginia	16.8
Massachusetts	26.7	Washington	44.6
Michigan	30.9	West Virginia	33.2
Minnesota	30.6	Wisconsin	32.9
Mississippi	18.4	Wyoming	24.0
Missouri	17.3		

- a. The smallest value in the data set is 9.0 (Alaska) and the largest value is 51.4 (Pennsylvania). Are these values outliers? Explain. (Hint: See Example 4.11.)
- b. Construct a boxplot of the data set and comment on the interesting features of the plot.

- 4.37** The U.S. Department of Health and Human Services reported the estimated percentage of U.S. households with only wireless phone service (no landline) in 2014 for the 50 states and the District of Columbia ([cdc.gov/nchs/data/nhis/earlyrelease/wireless_state_201602.pdf](#), retrieved April 20, 2017). In the accompanying data table, each state was also classified into one of three geographical regions—West (W), Middle states (M), and East (E).

Wireless			Wireless		
%	Region	State	%	Region	State
43.4	M	AL	41.0	W	MT
39.7	W	AK	46.5	M	NE
49.4	W	AZ	48.4	W	NV
56.2	M	AR	43.6	M	ND
42.8	W	CA	31.2	E	NH
50.5	W	CO	25.1	E	NJ
26.7	E	CT	47.0	W	NM
29.4	E	DE	31.1	E	NY
49.7	E	DC	42.9	E	NC
47.6	E	FL	45.8	E	OH
45.9	E	GA	50.4	M	OK
38.3	W	HI	47.0	W	OR
56.1	W	ID	30.0	E	PA
45.7	M	IL	34.6	E	RI
47.7	M	IN	49.5	E	SC
50.7	M	IA	41.4	M	SD
51.6	M	KS	46.6	M	TN
47.1	M	KY	54.6	M	TX
40.9	M	LA	52.2	W	UT
40.8	E	ME	41.1	E	VA
36.2	E	MD	37.2	E	VT
31.5	E	MA	48.3	W	WA
47.8	M	MI	37.2	E	WV
43.1	M	MN	46.6	M	WI
55.1	M	MS	51.8	W	WY
51.5	M	MO			

- 4.38** • Fiber content (in grams per serving) for 18 high fiber cereals (consumerreports.com) are shown below.
- Fiber Content**
- | | | | | | | | | |
|----|----|----|----|---|----|----|----|---|
| 7 | 10 | 10 | 7 | 8 | 7 | 12 | 12 | 8 |
| 13 | 10 | 8 | 12 | 7 | 14 | 7 | 8 | 8 |
- a. Find the median, quartiles, and interquartile range for the fiber content data set.
 b. Explain why the minimum value for the fiber content data set and the lower quartile for the fiber content data set are equal.

- 4.39** In addition to the fiber contents given in the previous exercise, sugar content (in grams per serving) for the same 18 high-fiber cereals were also given (consumerreports.com).

11	6	14	13	0	18	9	10	19
6	10	17	10	10	0	9	5	11

- a. Calculate the median, quartiles, and interquartile range for the sugar content data.

- b. Are there any outliers in the sugar content data?

- 4.40** Use the fiber content and sugar content data given in the previous two exercises to construct a comparative boxplot. Comment on the differences and similarities in the fiber and sugar content distributions.

- 4.41** The article “The Best—and Worst—Places to be a Working Woman” (*The Economist*, Graphic Detail for March 3, 2016) reported values of what it calls the glass-ceiling index, which is designed to rate countries based on women’s chances of equal treatment at work. The index weights factors that include participation of women in higher education, participation in the workforce by women, pay, child-care cost, and maternity benefits. The best possible value for this index is 100. Data for 29 countries are shown in the accompanying table.

Country	Glass-Ceiling Index
Iceland	82.8
Norway	79.3
Sweden	79.0
Finland	73.8
Hungary	70.4
Poland	70.0
France	68.3
Denmark	66.6
New Zealand	63.8
Belgium	63.3
Canada	62.3
Portugal	60.8
Spain	59.1
Israel	58.3
Slovakia	57.8
Austria	57.1
Germany	57.0
Australia	56.2
United States	55.9
Czech Republic	55.1
Italy	53.7
Netherlands	51.8
Greece	50.8
Great Britain	50.5
Ireland	48.4
Switzerland	40.6
Japan	28.8
Turkey	27.2
South Korea	25.0

- a. Are there outliers in this data set? If so, which observations are outliers?
 b. Draw a modified boxplot for this data set.
 c. The article points out that Nordic countries (Iceland, Sweden, Norway, and Finland) come out on top on this index. Where are the values for the Nordic countries located in terms of the boxplot?

SECTION 4.4**Interpreting Center and Variability: Chebyshev's Rule, the Empirical Rule, and z Scores**

The mean and standard deviation can be combined to make informative statements about how the values in a data set are distributed and about the relative position of a particular value in a data set. To do this, it is useful to be able to describe how far away a particular observation is from the mean in terms of the standard deviation. For example, we might say that an observation is 2 standard deviations above the mean or that an observation is 1.3 standard deviations below the mean.

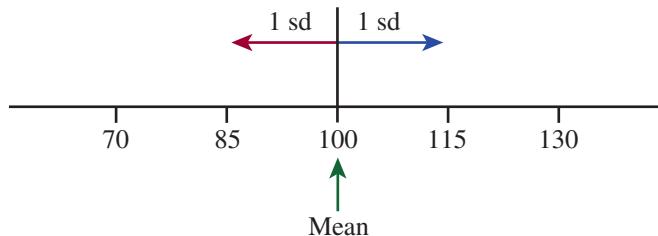
Example 4.14 Standardized Test Scores

Consider a data set of scores on a standardized test with a mean and standard deviation of 100 and 15, respectively. We can make the following statements:

1. Because $100 - 15 = 85$, we say that a score of 85 is “1 standard deviation *below* the mean.” Similarly, $100 + 15 = 115$ is “1 standard deviation *above* the mean” (see Figure 4.14).

FIGURE 4.14

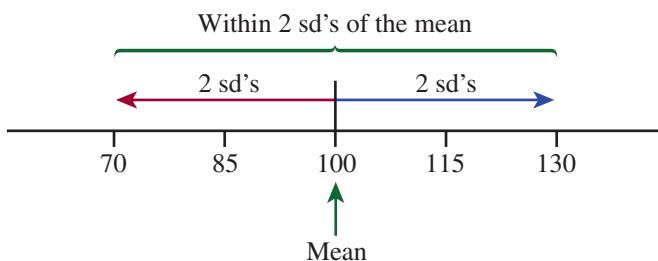
Values within 1 standard deviation of the mean (Example 4.14).



2. Because 2 times the standard deviation is $2(15) = 30$, scores between 70 and 130 are those *within* 2 standard deviations of the mean (see Figure 4.15).

FIGURE 4.15

Values within 2 standard deviations of the mean (Example 4.14).



3. Because $100 + (3)(15) = 145$, scores above 145 are greater than the mean by more than 3 standard deviations.

Sometimes in published articles, the mean and standard deviation are reported, but a graphical display of the data is not given. However, using a result called Chebyshev's Rule, it is possible to get a sense of the distribution of data values based only on knowing the mean and the standard deviation.

Chebyshev's Rule

Consider any number k , where $k \geq 1$. The percentage of observations that are within k standard deviations of the mean is at least $100\left(1 - \frac{1}{k^2}\right)\%$.

Substituting selected values of k into Chebyshev's Rule gives the following results.

Number of Standard Deviations, k	$1 - \frac{1}{k^2}$	Percentage Within k Standard Deviations of the Mean
2	$1 - \frac{1}{4} = 0.75$	at least 75%
3	$1 - \frac{1}{9} = 0.89$	at least 89%
4	$1 - \frac{1}{16} = 0.94$	at least 94%
4.472	$1 - \frac{1}{20} = 0.95$	at least 95%
5	$1 - \frac{1}{25} = 0.96$	at least 96%
10	$1 - \frac{1}{100} = 0.99$	at least 99%

Example 4.15 Child Care for Preschool Kids

Understand the context ➤

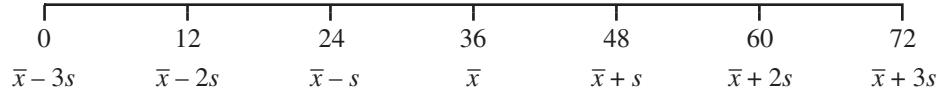
The report “Who's Minding the Kids? Child Care Arrangements, Spring 2011” (census.gov/library/publications/2013/demo/p70-135.pdf, retrieved February 25, 2018) examined various modes of care for preschool children. For a sample of families with one preschool child, it was reported that the mean child care time per week was 36 hours. Suppose that the standard deviation of child care times was 12 hours. Figure 4.16 displays values that are 1, 2, and 3 standard deviations from the mean.



Ariel Skellie/Blend Images/Getty Images

FIGURE 4.16

Measurement scale for child care time (Example 4.15).



Chebyshev's Rule allows us to assert the following:

- At least 75% of the sample observations must be between 12 and 60 hours (within 2 standard deviations of the mean).
- Because at least 89% of the observations must be between 0 and 72, at most 11% are outside this interval. Time cannot be negative, so we conclude that at most 11% of the observations exceed 72.
- The values 18 and 54 are 1.5 standard deviations to either side of \bar{x} , so using $k = 1.5$ in Chebyshev's Rule implies that at least 55.6% of the observations must be between these two values. This means that at most 44.4% of the observations are less than 18—not at most 22.2%, because the distribution of values may not be symmetric.

Because Chebyshev's Rule can be used with any data set (distribution), whether symmetric or skewed, we must be careful when making statements about the proportion above

a particular value, below a particular value, or inside or outside an interval that is not centered at the mean. The rule must be used in a conservative fashion. There is another aspect of this conservatism. The rule states that *at least* 75% of the observations are within 2 standard deviations of the mean, but for many data sets substantially more than 75% of the values satisfy this condition. The same sort of understatement is also frequently encountered for other values of k (numbers of standard deviations).

Example 4.16 IQ Scores

Understand the context ➤

Figure 4.17 gives a stem-and-leaf display of IQ scores of 112 children in one of the early studies that used the Stanford revision of the Binet–Simon intelligence scale (*The Intelligence of School Children, L. M. Terman [Boston: Houghton-Mifflin, 1919]*).

Summary quantities include

$$\bar{x} = 104.5 \quad s = 16.3 \quad 2s = 32.6 \quad 3s = 48.9$$

FIGURE 4.17

Stem-and-leaf display of IQ scores used in Example 4.16.

6	1
7	25679
8	0000124555668
9	0000112333446666778889
10	000112222233566677778899999
11	00001122333344444477899
12	01111123445669
13	006
14	26
15	2

Stem: Tens
Leaf: Ones

In Figure 4.17, all of the observations that are within two standard deviations of the mean are shown in blue. Table 4.6 shows how Chebyshev's Rule can sometimes considerably underestimate actual percentages.

TABLE 4.6 Summarizing the Distribution of IQ Scores

k = Number of sd's	$\bar{x} \pm ks$	Chebyshev	Actual
2	71.9 to 137.1	at least 75%	96% (108)
2.5	63.7 to 145.3	at least 84%	97% (109) ← the blue leaves in Figure 4.17
3	55.6 to 153.4	at least 89%	100% (112)

Empirical Rule

The fact that statements based on Chebyshev's Rule are frequently conservative suggests that we should look for rules that are less conservative and more precise. One useful rule is the **Empirical Rule**, which can be used whenever the distribution of data values is reasonably well described by a normal curve (distributions that are “mound” shaped).

The Empirical Rule

If the histogram of values in a data set can be reasonably well approximated by a normal curve, then

Approximately 68% of the observations are within 1 standard deviation of the mean.

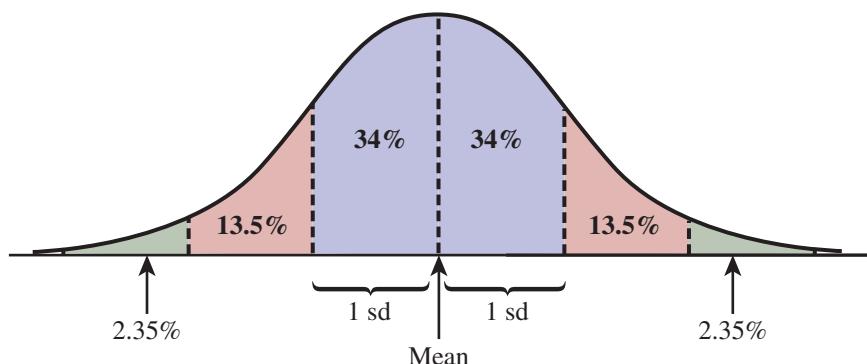
Approximately 95% of the observations are within 2 standard deviations of the mean.

Approximately 99.7% of the observations are within 3 standard deviations of the mean.

The Empirical Rule makes “approximately” instead of “at least” statements, and the percentages for $k = 1, 2$, and 3 standard deviations are much higher than those of Chebyshev’s Rule. Figure 4.18 illustrates the percentages given by the Empirical Rule. In contrast to Chebyshev’s Rule, dividing the “within” percentages in half is permissible, because a normal curve is symmetric.

FIGURE 4.18

Approximate percentages implied by the Empirical Rule.



Example 4.17 Heights of Mothers and the Empirical Rule

Understand the context ➤

One of the earliest articles to argue for the wide applicability of the normal distribution was “[On the Laws of Inheritance in Man. I. Inheritance of Physical Characters](#)” (*Biometrika* [1903]: 375–462). One of the data sets discussed in the article consisted of 1052 measurements of the heights of mothers. The mean and standard deviation were

$$\bar{x} = 62.484 \text{ in.} \quad s = 2.390 \text{ in.}$$

The data distribution was described as approximately normal. Table 4.7 contrasts actual percentages with those obtained from Chebyshev’s Rule and the Empirical Rule.



istock.com/PeopleImages

TABLE 4.7 Summarizing the Distribution of Mothers’ Heights

Number of sd's	Interval	Actual	Empirical Rule	Chebyshev Rule
1	60.094 to 64.874	72.1%	Approximately 68%	At least 0%
2	57.704 to 67.264	96.2%	Approximately 95%	At least 75%
3	55.314 to 69.654	99.2%	Approximately 99.7%	At least 89%

Clearly, the Empirical Rule is much more successful and informative in this case than Chebyshev’s Rule.

As we will see in Chapter 7, it is possible to make statements analogous to those of the Empirical Rule for values other than $k = 1, 2$, or 3 standard deviations. For now, notice that it is unusual to see an observation from a data distribution that is approximately normal that is farther than 2 standard deviations from the mean (only 5%). Also, it is very surprising to see one that is more than 3 standard deviations away. If you encountered a mother whose height was 72 inches, you might reasonably conclude that she was not part of the population described by the data set in Example 4.17.

Measures of Relative Standing

When you obtain your score after taking a test, you probably want to know how it compares to the scores of others who have taken the test. Is your score above or below the mean, and by how much? Does your score place you among the top 5% of those who took the test or only among the top 25%? Questions of this sort are answered by finding ways

to measure the position of a particular value in a data set relative to all values in the set. One measure of relative standing is a **z score**.

DEFINITION

z score: The **z score** corresponding to a particular value is

$$z \text{ score} = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

The z score tells us how many standard deviations the value is from the mean. It is positive or negative depending on whether the value is greater than or less than the mean.

The process of subtracting the mean and then dividing by the standard deviation is sometimes referred to as standardizing. A z score is one example of what is called a standardized score.

Example 4.18 Relatively Speaking, Which Is the Better Offer?

Understand the context ➤

Suppose that two graduating seniors, one a marketing major and one an accounting major, are comparing job offers. The accounting major has an offer for \$55,000 per year, and the marketing student has an offer for \$53,000 per year. Summary information about the distribution of offers follows:

Consider the data ➤

Accounting: mean = 56,000 standard deviation = 1500
 Marketing: mean = 52,500 standard deviation = 1000

Then,

$$\text{accounting } z \text{ score} = \frac{55,000 - 56,000}{1500} = -0.67$$

Do the work ➤

(so \$55,000 is 0.67 standard deviation below the mean), but

$$\text{marketing } z \text{ score} = \frac{53,000 - 52,500}{1000} = 0.5$$

Interpret the results ➤

Relative to the appropriate data sets, the marketing offer is actually more attractive than the accounting offer.

The z score is particularly useful when the distribution of observations is approximately normal. In this case, using the Empirical Rule, a z score outside the interval from -2 to $+2$ occurs for about 5% of all observations. A z score outside the interval from -3 to $+3$ occurs only about 0.3% of the time.

Percentiles

A particular observation can be located even more precisely by giving the percentage of the data that fall at or below that observation. For example, suppose 95% of all test scores are at or below 650 and only 5% are above 650. Then 650 is the *95th percentile* of the data set (or of the distribution of scores). Similarly, if 10% of all scores are at or below 400 and 90% are above 400, the value 400 is the *10th percentile*.

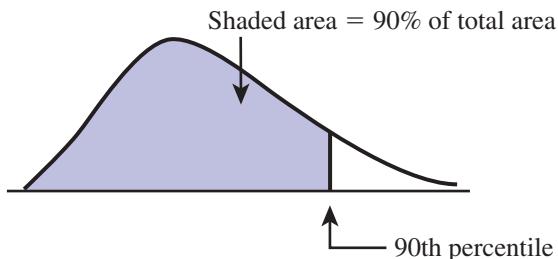
DEFINITION

Percentile: For any particular number r between 0 and 100, the **r th percentile** is a value such that r percent of the observations in the data set fall at or below that value.

Figure 4.19 illustrates the 90th percentile. We have already met several percentiles in disguise. The median is the 50th percentile, and the lower and upper quartiles are the 25th and 75th percentiles, respectively.

FIGURE 4.19

Ninetieth percentile for a smoothed histogram.



Example 4.19 Head Circumference at Birth

Understand the context ➤

In addition to weight and length, head circumference is another measure of health in newborn babies. [The National Center for Health Statistics](#) reports the following summary values for head circumference (in cm) at birth for boys (approximate values read from graphs on the Centers for Disease Control and Prevention web site at cdc.gov/growthcharts/data/set1clinical/cj41l019.pdf, retrieved April 20, 2017):

Consider the data ➤

Percentile	5	10	25	50	75	90	95
Head Circumference (cm)	32.2	33.2	34.5	35.8	37.0	38.2	38.6

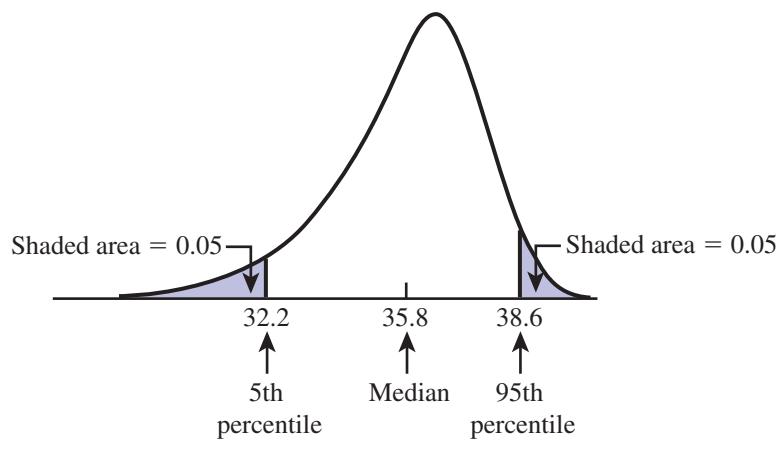
Interpret the results ➤

Interpreting these percentiles, we know that half of newborn boys have head circumferences of less than 35.8 cm, because 35.8 is the 50th percentile (the median). The middle 50% of newborn boys have head circumferences between 34.5 cm and 37.0 cm, with about 25% of the head circumferences less than 34.5 cm and about 25% greater than 37.0 cm.

We can tell that the head circumference distribution for newborn boys is not symmetric, because the 5th percentile is 3.6 cm below the median, whereas the 95th percentile is only 2.8 cm above the median. This suggests that the bottom part of the distribution stretches out more than the top part of the distribution. This would be consistent with a distribution that is negatively skewed, as shown in Figure 4.20.

FIGURE 4.20

Negatively skewed distribution.



EXERCISES 4.42 - 4.58

● Data set available online

- 4.42** The average playing time of music albums in a large collection is 35 minutes, and the standard deviation is 5 minutes.

- a. What value is 1 standard deviation above the mean? 1 standard deviation below the mean? What values are 2 standard deviations

- away from the mean? (Hint: See Example 4.14.)
- b. Without assuming anything about the distribution of times, at least what percentage of the times are between 25 and 45 minutes? (Hint: See Example 4.15.)
- c. Without assuming anything about the distribution of times, what can be said about the percentage of times that are either less than 20 minutes or greater than 50 minutes?
- d. Assuming that the distribution of times is approximately normal, about what percentage of times are between 25 and 45 minutes? less than 20 minutes or greater than 50 minutes? less than 20 minutes?
- 4.43** In a study investigating the effect of car speed on accident severity, 5000 reports of fatal automobile accidents were examined, and the vehicle speed at impact was recorded for each one. For these 5000 accidents, the average speed was 42 mph and the standard deviation was 15 mph. A histogram revealed that the vehicle speed at impact distribution was approximately normal.
- a. Approximately what proportion of these vehicle speeds were between 27 and 57 mph? (Hint: See Example 4.17.)
- b. Approximately what proportion of these vehicle speeds exceeded 57 mph?
- 4.44** The U.S. Census Bureau (2000 census) reported the following relative frequency distribution for travel time to work for a large sample of adults who did not work at home:
- | Travel Time (minutes) | Relative Frequency |
|-----------------------|--------------------|
| 0 to <5 | 0.04 |
| 5 to <10 | 0.13 |
| 10 to <15 | 0.16 |
| 15 to <20 | 0.17 |
| 20 to <25 | 0.14 |
| 25 to <30 | 0.05 |
| 30 to <35 | 0.12 |
| 35 to <40 | 0.03 |
| 40 to <45 | 0.03 |
| 45 to <60 | 0.06 |
| 60 to <90 | 0.05 |
| 90 or more | 0.02 |
- a. Draw the histogram for the travel time distribution. In constructing the histogram, assume that the last interval in the relative frequency distribution (90 or more) ends at 200; so the last interval is 90 to <200. Be sure to use the density scale to determine the heights of the bars in the histogram because not all the intervals have the same width. (Hint: Histograms were covered in Chapter 3.)
- b. Describe the interesting features of the histogram from Part (a), including center, shape, and variability.
- c. Based on the histogram from Part (a), would it be appropriate to use the Empirical Rule to make statements about the travel time distribution? Explain why or why not.
- 4.45** For the travel time distribution given in the previous exercise, the approximate mean and standard deviation for the travel time distribution are 27 minutes and 24 minutes, respectively. Based on this mean and standard deviation and the fact that travel time cannot be negative, explain why the travel time distribution could not be well approximated by a normal curve.
- 4.46** Use the information given in the previous two exercises and Chebyshev's Rule to complete this exercise.
- a. Make a statement about
- the percentage of travel times that were between 0 and 75 minutes
 - the percentage of travel times that were between 0 and 54 minutes
- b. How well do the statements in Part (a) based on Chebyshev's Rule agree with the actual percentages for the travel time distribution? (Hint: You can estimate the actual percentages from the relative frequency distribution given in Exercise 4.44.)
- 4.47** Mobile homes are tightly constructed for energy conservation. This can lead to a buildup of indoor pollutants. The paper "A Survey of Nitrogen Dioxide Levels Inside Mobile Homes" (*Journal of the Air Pollution Control Association* [1988]: 647–651) discussed various aspects of NO_2 concentration in these structures.
- a. For one sample of mobile homes in the Los Angeles area, the mean NO_2 concentration in kitchens during the summer was 36.92 ppb, and the standard deviation was 11.34. Making no assumptions about the shape of the NO_2 distribution, what can be said about the percentage of observations between 14.24 and 59.60?
- b. Inside what interval can you be sure that at least 89% of the concentration observations will lie?
- c. For a sample of mobile homes that were not in Los Angeles, the average kitchen NO_2 concentration during the winter was 24.76 ppb, and the standard deviation was 17.20. Do these values suggest that the histogram of sample observations did not closely resemble a normal curve? (Hint: What is $\bar{x} - 2s$?)
- 4.48** The article "Impact of Berkeley Excise Tax on Sugar-Sweetened Beverage Consumption" (*American Journal of Public Health* [2016]: 1865–1871) estimated that the mean number of times that adults in Berkeley, California, drank

regular soda per day to be 0.34. The standard deviation for the number of times per day was estimated to be 0.86. Would you use the Empirical Rule to approximate the proportion of adults who drink regular soda more than 1.20 times per day on average (i.e., the proportion of adults in Berkeley whose value exceeds the mean by more than 1 standard deviation)? Explain your reasoning.

- 4.49** A student took two national aptitude tests. The national mean and standard deviation were 475 and 100, respectively, for the first test and 30 and 8, respectively, for the second test. The student scored 625 on the first test and 45 on the second test. Use z scores to determine on which exam the student performed better relative to the other test takers. (Hint: See Example 4.18.)

- 4.50** Suppose that your younger sister is applying for entrance to college and has taken the SAT. She scored at the 83rd percentile on the verbal section of the test and at the 94th percentile on the math section of the test. Because you have been studying statistics, she asks you for an interpretation of these values. What would you tell her? (Hint: See Example 4.19.)

- 4.51** The report “Who Borrows Most? Bachelor’s Degree Recipients with High Levels of Student Debt” (trends.collegeboard.org/content/who-borrows-most-bachelors-degree-recipients-high-levels-student-debt-april-2010, retrieved April 20, 2017) reported the following percentiles for amount of student debt for those graduating with a bachelor’s degree in 2010:

$$\begin{array}{ll} \text{10th percentile} = \$0 & \text{25th percentile} = \$0 \\ \text{50th percentile} = \$11,000 & \text{75th percentile} = \$24,600 \\ \text{90th percentile} = \$39,300 & \end{array}$$

For each of these percentiles, write a sentence interpreting the value of the percentile.

- 4.52** The paper “Study of the Flying Ability of *Rhynchophorus ferrugineus* Adults Using a Computer-Monitored Mill” (*Bulletin of Entomological Research* [2014]: 462–467) summarized data from a study of red palm weevils, a pest that is a threat to palm trees. The following frequency distribution from the paper was constructed using the longest flight (in meters) observed for 132 weevils.

Longest Flight (meters)	Frequency (Number of Weevils)
0 to < 100	71
100 to < 2,000	32
2,000 to < 5,000	18
5,000 to < 10,000	8
10,000 or more	3

Estimate the approximate values of the following percentiles:

- a. 54th
- b. 80th
- c. 92nd

- 4.53** Suppose that the manufacturer of a scale claims that its scale weighs items up to 110 pounds and provides accuracy to within 0.25 ounce. Suppose that a 50-ounce weight was repeatedly weighed on this scale and the weight readings recorded. The mean value was 49.5 ounces, and the standard deviation was 0.1. What can be said about the percentage of the time that the scale actually showed a weight that was within 0.25 ounce of the true value of 50 ounces? (Hint: Use Chebyshev’s Rule.)

- 4.54** Suppose that your statistics professor returned your first midterm exam with only a z score written on it. She also told you that a histogram of the scores was approximately normal. How would you interpret each of the following z scores?
- a. 2.2
 - b. 0.4
 - c. 1.8
 - d. 1.0
 - e. 0

- 4.55** The paper “Answer Changing in Multiple Choice Assessment: Change that Answer When in Doubt—and Spread the Word” (*BMC Medical Education* [2007]: 28–32) reported that for a group of 72 students, the average number of responses changed from the correct answer to an incorrect answer on a test containing 78 multiple-choice items was 0.9. The corresponding standard deviation was reported to be 1.0. Based on this mean and standard deviation, what can you tell about the shape of the distribution of the variable *number of answers changed from right to wrong*? What can you say about the number of students who changed at least three answers from correct to incorrect?

- 4.56** Suppose that the average reading speed of students completing a speed-reading course is 450 words per minute (wpm). If the standard deviation is 70 wpm, find the z score associated with each of the following reading speeds.
- a. 320 wpm
 - b. 475 wpm
 - c. 420 wpm
 - d. 610 wpm

- 4.57** ● The following data values are 2014 per capita operating expenditures on public libraries for each of the 50 U.S. states and the District of Columbia (imls.gov/research-evaluation/data-collection/public-libraries-survey/explore-pls-data, retrieved April 20, 2017):

16.48 16.66 16.76 18.77 18.87 19.95 20.43 21.92 24.21 25.05 25.77
 26.01 26.07 26.50 26.55 26.60 26.80 28.25 30.70 31.69 31.77 33.23
 33.33 33.51 33.74 34.59 35.53 37.09 37.25 37.37 37.61 38.19 38.47
 41.34 42.65 43.02 44.16 45.02 48.22 49.39 50.21 51.81 53.39 54.01
 54.97 55.88 56.36 58.52 59.62 59.95 67.44

- a. Summarize this data set with a frequency distribution. Construct the corresponding histogram.
- b. Use the histogram in Part (a) to find approximate values of the following percentiles:
- i. 50th
 - ii. 70th
 - iii. 10th
 - iv. 90th
 - v. 40th

- 4.58** The accompanying table gives the mean and standard deviation of reaction times (in seconds) for each of two different stimuli:

	Stimulus 1	Stimulus 2
Mean	6.0	3.6
Standard deviation	1.2	0.8

If your reaction time is 4.2 seconds for the first stimulus and 1.8 seconds for the second stimulus, to which stimulus are you reacting relatively more quickly (compared with other individuals)? (Hint: See Example 4.18.)

SECTION 4.5 Interpreting and Communicating the Results of Statistical Analyses

As was the case with the graphical displays of Chapter 3, the summary statistics introduced in this chapter help us better understand the variables under study. If we have collected data on the amount of money students spend on textbooks at a particular university, most likely we did so because we wanted to learn about the distribution of this variable (amount spent on textbooks) for the population of interest (in this case, students at the university). Numerical measures of center and variability and boxplots provide information, and they also allow us to communicate to others what we have learned from the data.

Communicating the Results of Statistical Analyses

When reporting the results of a data analysis, it is common to begin with descriptive information about the variables of interest. It is always a good idea to start with a graphical display of the data, and, as we saw in Chapter 3, graphical displays of numerical data are usually described in terms of center, variability, and shape. The numerical measures of this chapter can help you to be more specific in describing the center and variability of a data set.

When describing center and variability, you must first decide which measures to use. Common choices are to use either the sample mean and standard deviation or the sample median and interquartile range (and maybe even a boxplot). Because the mean and standard deviation can be sensitive to extreme values in the data set, they are best used when the distribution shape is approximately symmetric and when there are few outliers. If the data set is noticeably skewed or if there are outliers, the median and iqr are generally used to describe center and variability.

Interpreting the Results of Statistical Analyses

It is relatively rare to find raw data in published sources. Typically, only a few numerical summary quantities are reported. We must be able to interpret these values and understand what they tell us about the underlying data set.

For example, a university conducted an investigation of the amount of time required to process an application for admission once an applicant had entered his or her information online. One of the individuals who performs this task was asked to record starting time and completion time for 50 randomly selected applications. The resulting times (in minutes) were summarized using the mean, median, and standard deviation:

$$\begin{aligned}\bar{x} &= 27.8 \\ \text{median} &= 27.4 \\ s &= 2.14\end{aligned}$$

What do these summary values tell us about processing times? The average time required was 27.8 minutes. The value of the standard deviation suggests that there was some variability in the times and that there were some processing times that differed quite a bit from the mean time. The median tells us that half of the applications required less than 27.4 minutes. The fact that the mean exceeds the median suggests that some unusually large values in the data set may have affected the value of the mean. This last conjecture is confirmed by the stem-and-leaf display of the data given in Figure 4.21.

FIGURE 4.21

Stem-and-leaf display of application processing times.

24	8	
25	02345679	
26	0001234566779	
27	223556688	
28	23334	
29	002	
30	011168	
31	134	
32	2	Stem: Ones
33		Leaf: Tenths
34	3	

The administrators conducting the study looked at the outlier of 34.3 minutes and at the other relatively large values in the data set. They found that the five largest values came from applications that were entered before lunch. After talking with the individual who entered the data, the administrators speculated that morning entry times might differ from afternoon entry times because there tended to be more distractions and interruptions (phone calls, etc.) during the morning hours, when the admissions office tended to be busier.

When morning and afternoon entry times were separated, the following summary statistics resulted:

$$\begin{array}{llll} \text{Morning (based on } n = 20 \text{ applications):} & \bar{x} = 29.1 & \text{median} = 28.7 & s = 2.329 \\ \text{Afternoon (based on } n = 30 \text{ applications):} & \bar{x} = 27.0 & \text{median} = 26.7 & s = 1.529 \end{array}$$

Clearly, the mean processing time is higher for applications processed in the morning. The individual processing times also differ more from one another in the mornings than in the afternoons (because the standard deviation for morning processing times, 2.329, is about 1.5 times as large as 1.529, the standard deviation for afternoon processing times).

What to Look for in Published Data

Here are a few questions to ask yourself when you interpret numerical summary measures.

- Is the chosen summary measure appropriate for the type of data collected? In particular, watch for inappropriate use of the mean and standard deviation with categorical data that has simply been coded numerically.
- If both the mean and the median are reported, how do the two values compare? What does this suggest about the distribution of values in the data set? If only the mean or the median was used, was an appropriate measure selected?
- Is the standard deviation large or small? Is the value consistent with your expectations regarding variability? What does the value of the standard deviation tell you about the variable being summarized?
- Can anything of interest be said about the values in the data set by applying Chebyshev's Rule or the Empirical Rule?

For example, consider a study that investigated whether people tend to spend more money when they are paying with a credit card than when they are paying with cash. The authors of the paper “[Monopoly Money: The Effect of Payment Coupling and Form on Spending Behavior](#)” (*Journal of Experimental Psychology: Applied* [2008]: 213–225)

randomly assigned each of 114 volunteers to one of two experimental groups. Participants were given a menu for a new restaurant that showed nine menu items. They were then asked to estimate the amount they would be willing to pay for each item. A price index was computed for each participant by averaging the nine prices assigned.

The difference between the two experimental groups was that the menu viewed by one group showed a credit card logo at the bottom of the menu while there was no credit card logo on the menu that those in the other group viewed. The following passage appeared in the results section of the paper:

On average, participants were willing to pay more when the credit card logo was present ($M = \$4.53$, $SD = 1.15$) than when it was absent ($M = \$4.11$, $SD = 1.06$). Thus, even though consumers were not explicitly informed which payment mode they would be using, the mere presence of a credit card logo increased the price that they were willing to pay.

The price index data distribution was also described as mound shaped with no outliers for each of the two groups. Because price index (the average of the prices that a participant assigned to the nine menu items) is a numerical variable, the mean and standard deviation are reasonable measures for summarizing center and variability in the data set.

Although the mean for the credit-card-logo group is higher than the mean for the no-logo group, the two standard deviations are similar, indicating similar variability in price index from person to person for the two groups.

Because the distribution of price index values was mound shaped for each of the two groups, we can use the Empirical Rule to tell us a bit more about the distribution. For example, for those in the group who viewed the menu with a credit card logo, approximately 95% of the price index values would have been between

$$4.53 - 2(1.15) = 4.53 - 2.3 = 2.23$$

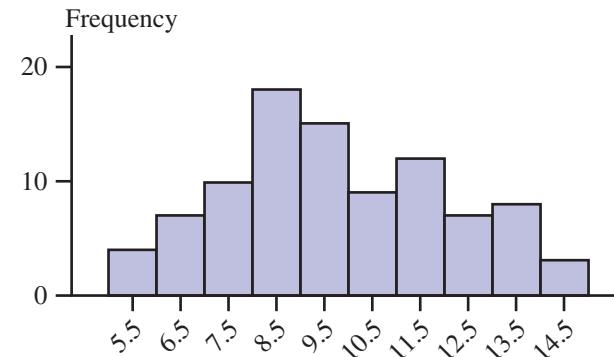
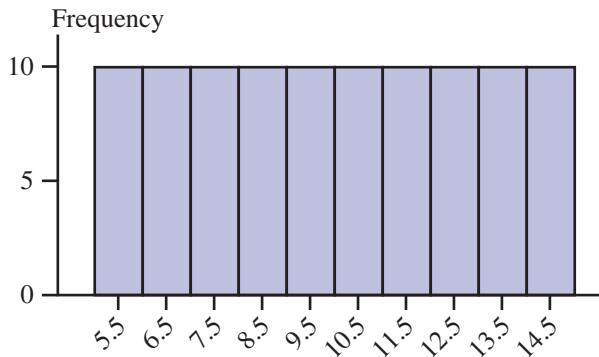
and

$$4.53 + 2(1.15) = 4.53 + 2.30 = 6.83.$$

A Word to the Wise: Cautions and Limitations

When calculating or interpreting numerical descriptive measures, you need to keep in mind the following:

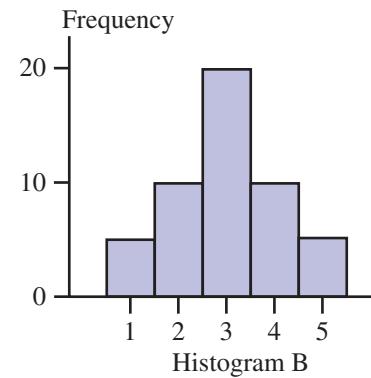
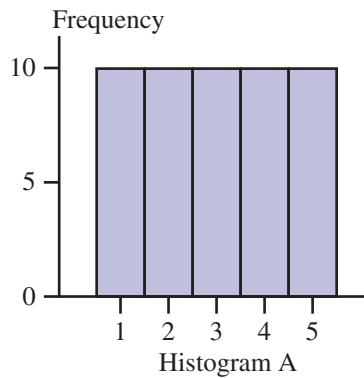
1. Measures of center don't tell all. Although measures of center, such as the mean and the median, do give us a sense of what might be considered a typical value for a variable, this is only one characteristic of a data set. Without additional information about variability and distribution shape, we don't really know much about the behavior of the variable.
2. Data distributions with different shapes can have the same mean and standard deviation. For example, consider the following two histograms:



Both histograms summarize data sets that have a mean of 10 and a standard deviation of 2, yet they have different shapes.

3. Both the mean and the standard deviation are sensitive to extreme values in a data set, especially if the sample size is small. If a data distribution is skewed or if the data set has outliers, the median and the interquartile range may be a better choice for describing center and variability.
4. Measures of center and variability describe the values of the variable studied, not the frequencies in a frequency distribution or the heights of the bars in a histogram. For example, consider the following two frequency distributions and histograms:

Frequency Distribution A		Frequency Distribution B	
Value	Frequency	Value	Frequency
1	10	1	5
2	10	2	10
3	10	3	20
4	10	4	10
5	10	5	5



There is more variability in the data summarized by Frequency Distribution and Histogram A than in the data summarized by Frequency Distribution and Histogram B. This is because the values of the variable described by Histogram and Frequency Distribution B are more concentrated near the mean than are the values for the variable described by Histogram and Frequency Distribution A. Don't be misled by the fact that there is no variability in the frequencies in Frequency Distribution A or the heights of the bars in Histogram A.

5. Be careful with boxplots based on small sample sizes. Boxplots convey information about center, variability, and shape, but when the sample size is small, you should be hesitant to overinterpret shape information. It is really not possible to decide whether a data distribution is symmetric or skewed if only a small sample of observations from the distribution is available.
6. Not all distributions are normal (or even approximately normal). Be cautious in applying the Empirical Rule in situations in which you are not convinced that the data distribution is at least approximately normal. Using the Empirical Rule in such situations can lead to incorrect statements.
7. Watch out for outliers! Unusual observations in a data set often provide important information about the variable under study, so it is important to consider outliers in addition to describing what is typical. Outliers can also be problematic—both because the values of some descriptive measures are influenced by outliers and because some methods for drawing conclusions from data may not be appropriate if the data set has outliers.

EXERCISES 4.59 - 4.60

- 4.59** The authors of the paper “Delayed Time to Defibrillation after In-Hospital Cardiac Arrest” (*New England Journal of Medicine* [2008]: 9–16) described a study of how survival is related to the length of time it takes from the time of a heart attack to the administration of defibrillation therapy. The following is a statement from the paper:

We identified 6789 patients from 369 hospitals who had in-hospital cardiac arrest due to ventricular fibrillation (69.7%) or pulseless ventricular tachycardia (30.3%). Overall, the median time to defibrillation was 1 minute (interquartile range [was] 3 minutes).

Data from the paper on time to defibrillation (in minutes) for these 6789 patients was used to produce the Minitab output and boxplot at the bottom of the page.

- Why is there no lower whisker in the given boxplot?
- How is it possible for the median, the lower quartile, and the minimum value in the data set to all be equal? (Note—this is why you do not see a median line in the box part of the boxplot.)
- The authors of the paper considered a time to defibrillation of greater than 2 minutes as unacceptable. Based on the given boxplot and summary statistics, is it possible that the percentage of patients having an unacceptable time to defibrillation is greater than 50%? Greater than 25%? Less than 25%? Explain.
- Is the outlier shown at 7 a mild outlier or an extreme outlier?

- 4.60** The paper “Portable Social Groups: Willingness to Communicate, Interpersonal Communication Gratifications, and Cell Phone Use among Young Adults” (*International Journal of Mobile Communications* [2007]: 139–156) describes a study of young adult cell phone use patterns.

- Comment on the following quote from the paper. Do you agree with the authors?

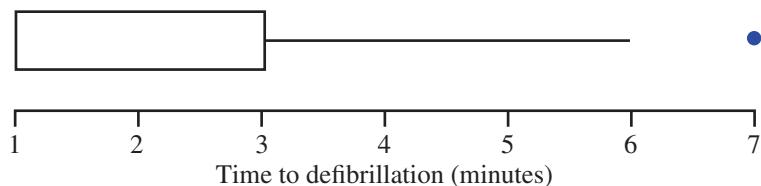
Seven sections of an Introduction to Mass Communication course at a large southern university were surveyed in the spring and fall of 2003. The sample was chosen because it offered an excellent representation of the population under study—young adults.

- Below is another quote from the paper. In this quote, the author reports the mean number of minutes of cell phone use per week for those who participated in the survey. What additional information would have been provided about cell phone use behavior if the author had also reported the standard deviation?

Based on respondent estimates, users spent an average of 629 minutes (about 10.5 hours) per week using their cell phone on or off line for any reason.

Descriptive Statistics: Time to Defibrillation

Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Time	6789	2.3737	2.0713	1.0000	1.0000	1.0000	3.0000	7.0000



CHAPTER ACTIVITIES

ACTIVITY 4.1 COLLECTING AND SUMMARIZING NUMERICAL DATA

In this activity, you will work in groups to collect data that will provide information about how many hours per week, on average, students at your school spend engaged in a particular activity. You will use the sampling plan designed in Activity 2.1 to collect the data.

1. With your group, pick one of the following activities to be the focus of your study:
 - i. Surfing the web
 - ii. Studying or doing homework
 - iii. Watching TV
 - iv. Exercising
 - v. Sleeping

or you may choose a different activity, *subject to the approval of your instructor*.

2. Use the plan developed in Activity 2.1 to collect data on the variable you have chosen for your study.
3. Summarize the resulting data using both numerical and graphical summaries. Be sure to address both center and variability.
4. Write a short article for your school paper summarizing your findings regarding student behavior. Your article should include both numerical and graphical summaries.

ACTIVITY 4.2 AIRLINE PASSENGER WEIGHTS

The article “Airlines Should Weigh Passengers, Bags, NTSB Says” (*USA TODAY*, February 27, 2004) states that the National Transportation Safety Board recommended that airlines weigh passengers and their bags to prevent overloaded planes from attempting to take off. This recommendation was the result of an investigation into the crash of a small commuter plane in 2003, which determined that too much weight contributed to the crash.

Rather than weighing passengers, airlines currently use estimates of average passenger and luggage weights. After the 2003 accident, this estimate was increased by 10 pounds for passengers and 5 pounds for luggage. Although an airplane can fly if it is somewhat overweight if all systems are working properly, if one of the plane’s

engines fails an overweight plane becomes difficult for the pilot to control.

Assuming that the new estimate of the average passenger weight is accurate, discuss the following questions with a partner and then write a paragraph that answers these questions.

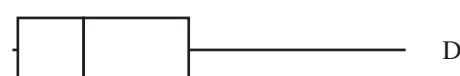
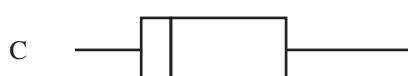
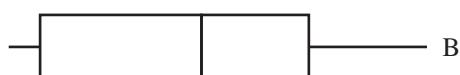
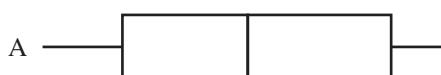
1. What role does variability in passenger weights play in creating a potentially dangerous situation for an airline?
2. Would an airline have a lower risk of a potentially dangerous situation if the variability in passenger weight is large or if it is small?

ACTIVITY 4.3 BOXPLOT SHAPES

In this activity, you will investigate the relationship between boxplot shapes and the corresponding five-number summary. The accompanying figure shows four boxplots, labeled A–D. Also given are 4 five-number summaries, labeled I–IV. Match each five-number summary to the appropriate boxplot. Note that scales are not included on the boxplots, so you will have to think about what the five-number summary implies about characteristics of the boxplot.

Five-Number Summaries

	I	II	III	IV
Minimum	40	4	0.0	10
Lower quartile	45	8	0.1	34
Median	71	16	0.9	44
Upper quartile	88	25	2.2	82
Maximum	106	30	5.1	132



SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
x_1, x_2, \dots, x_n	Notation for sample data consisting of observations on a variable x , where n is the sample size.	Chebyshev's Rule	This rule states that for any number $k \geq 1$, at least $100\left(1 - \frac{1}{k^2}\right)\%$ of the observations in any data set are within k standard deviations of the mean. It is typically conservative in that the actual percentages are often considerably greater than the stated lower bound.
Sample mean, \bar{x}	The most frequently used measure of center of a sample. It can be very sensitive to the presence of even a single outlier (unusually large or small observation).	Empirical Rule	This rule gives the approximate percentage of observations within 1 standard deviation (68%), 2 standard deviations (95%), and 3 standard deviations (99.7%) of the mean when the data distribution is well approximated by a normal curve.
Population mean, μ	The average x value in the entire population.	z score	This quantity gives the distance between an observation and the mean expressed as a certain number of standard deviations. It is positive (negative) if the observation is greater than (less than) the mean.
Sample median	The middle value in the ordered list of sample observations. (For n even, the median is the average of the two middle values.) It is very insensitive to outliers.	rth percentile	The value such that $r\%$ of the observations in the data set fall at or below that value.
Deviations from the mean: $x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$	Quantities used to assess variability in a sample. Except for rounding effects, $\sum(x - \bar{x}) = 0$.	Five-number summary	A summary of a data set that includes the minimum, lower quartile, median, upper quartile, and maximum.
The sample variance $s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$ and standard deviation $s = \sqrt{s^2}$	The most frequently used measures of variability for sample data.	Boxplot	A picture that conveys information about the most important features of a numerical data set: center, variability, extent of skewness, and presence of outliers.
The population variance σ^2 and standard deviation σ	Measures of variability for the entire population.		
Quartiles and the interquartile range	The lower quartile separates the smallest 25% of the data from the remaining 75%, and the upper quartile separates the largest 25% from the smallest 75%. The interquartile range (iqr), a measure of variability less sensitive to outliers than s , is the difference between the upper and lower quartiles.		

CHAPTER REVIEW Exercises 4.61 - 4.70

- 4.61** Acrylamide (a possible cancer-causing substance) forms in high-carbohydrate foods cooked at high temperatures and acrylamide levels can vary widely even within the same type of food. An article appearing in the journal **Food Chemistry** (March 2014, pages 204–211) included the following acrylamide content (in nanograms/gram) for five brands of biscuits:

345 292 334 276 248

- a. Calculate the mean acrylamide level and the five deviations from the mean.
- b. Verify that, except for the effect of rounding, the sum of the deviations from the mean is equal to 0

● Data set available online

for this data set. (If you rounded the sample mean or the deviations, your sum may not be exactly zero, but it should be close to zero if you have calculated the deviations correctly.)

- c. Calculate the variance and standard deviation for this data set.

4.62 ● The technical report **“Ozone Season Emissions by State”** (U.S. Environmental Protection Agency) reported the following sulphur dioxide emissions (in tons) for the 48 states in the continental U.S. states (epa.gov/airmarkets/progress/reports/emissions_reductions_so2.html#figure2, retrieved April 20, 2018):

State	SO ₂ Emissions	State	SO ₂ Emissions
AL	106,155	NC	48,154
AR	73,578	ND	55,203
AZ	23,689	NE	65,824
CA	227	NH	3,167
CO	38,287	NJ	2,432
CT	1,107	NM	17,735
DE	2,240	NV	7,427
FL	89,065	NY	17,797
GA	80,949	OH	281,986
IA	76,844	OK	74,425
ID	7	OR	14,004
IL	135,866	PA	252,078
IN	268,217	RI	16
KS	30,021	SC	26,779
KY	188,115	SD	15,342
LA	80,133	TN	56,405
MA	10,841	TX	365,507
MD	25,117	UT	21,144
ME	873	VA	38,778
MI	194,390	VT	2
MN	24,366	WA	2,859
MO	141,430	WI	62,434
MS	77,486	WV	86,201
MT	16,216	WY	40,671

- a.** Use these data to construct a boxplot that shows outliers.
- b.** Write a few sentences describing the important characteristics of the boxplot.
- 4.63** Because some homes have selling prices that are much higher than most, the median price is usually used to describe a “typical” home price for a given location. The three accompanying quotes are all from the *San Luis Obispo Tribune*, but each gives a different interpretation of the median price of a home in San Luis Obispo County. Comment on each of these statements. (Look carefully. At least one of the statements is incorrect.)
- a.** “So we have gone from 23% to 27% of county residents who can afford the median priced home at \$278,380 in SLO County. That means that half of the homes in this county cost less than \$278,380 and half cost more.” (*October 11, 2001*)
- b.** “The county’s median price rose to \$285,170 in the fourth quarter, a 9.6% increase from the same period a year ago, the report said. (The median represents the midpoint of a range.)” (*February 13, 2002*)
- c.** “‘Your median is going to creep up above \$300,000 if there is nothing available below \$300,000,’ Walker said.” (*February 26, 2002*)

4.64 ● Although bats are not known for their eyesight, they are able to locate prey (mainly insects) by emitting high-pitched sounds and listening for echoes. A paper appearing in *Animal Behaviour* (“The Echolocation of Flying Insects by Bats” [1960]: 141–154) gave the following distances (in centimeters) at which a bat first detected a nearby insect:

62 23 27 56 52 34 42 40 68 45 83

- a.** Calculate the sample mean distance at which the bat first detects an insect.
- b.** Calculate the sample variance and standard deviation for this data set. Interpret these values.

4.65 For the data in the previous exercise, subtract 10 from each sample observation. For the new set of values, calculate the mean and the deviations from the mean. How do these deviations compare to the deviations from the mean for the original sample? How does s^2 for the new values compare to s^2 for the old values? In general, what effect does subtracting (or adding) the same number to each observation have on s^2 and s ? Explain.

4.66 For the data of Exercise 4.64, multiply each data value by 10. How does s for the new values compare to s for the original values? More generally, what happens to s if each observation is multiplied by the same positive constant c ?

4.67 The Bloomberg web site included the data in the accompanying table on the number of movies made by 25 *Saturday Night Live* cast members as of 2014 (bloomberg.com/graphics/best-and-worst/#top-grossing-saturday-night-live-alumni, retrieved April 20, 2017). Also given was the top grossing movie made by each and the gross income for that movie adjusted for inflation.

Star	Number of movies	Top-grossing movie	Inflation-adjusted gross of top movie (millions of dollars)
Eddie Murphy	38	Shrek 2	550
Dan Aykroyd	56	Ghostbusters	514
Robert Downey, Jr.	57	The Avengers	646
Ben Stiller	50	Meet the Fockers	352
Bill Murray	46	Ghostbusters	514
Adam Sandler	36	The Waterboy	236
Chris Rock	33	Beverly Hills Cop II	322
Will Ferrell	40	Austin Powers: The Spy Who Shagged Me	293

(continued)

Star	Number of movies	Top-grossing movie	Inflation-adjusted gross of top movie (millions of dollars)
Mike Myers	16	Shrek 2	550
Joan Cusack	48	Toy Story 3	453
Bill Hader	29	Monsters University	274
Rob Schneider	29	Home Alone 2: Lost in New York	294
Brian Doyle-Murray	32	As Good As It Gets	220
Laurie Metcalf	28	Toy Story 3	453
Kristen Wiig	29	Despicable Me 2	376
Molly Shannon	33	How the Grinch Stole Christmas	359
David Koechner	45	Austin Powers: The Spy Who Shagged Me	293
Jon Lovitz	34	Big	231
Amy Poehler	23	Shrek the Third	370
Chevy Chase	32	National Lampoon's Vacation	147
Billy Crystal	26	Monsters University	274
Harry Shearer	27	The Simpsons Movie	210
Randy Quaid	38	Independence Day	464
David Spade	25	Grown Ups	177
Martin Short	28	Madagascar 3: Europe	224

Construct a boxplot for the number of movies data and comment on what the boxplot tells you about the distribution of the number of movies data.

4.68 Refer to the data given in the previous exercise.

- c. Are there any outliers in the inflation-adjusted gross movie income data? If so, which data values are outliers?

- d. Construct a boxplot for the inflation-adjusted gross movie income data.
e. For the inflation-adjusted gross movie income data, the mean is \$351.8 million and the median is \$322.0 million. What characteristic of the boxplot explains why the mean is greater than the median for this data set?

4.69 ● Age at diagnosis for each of 20 patients under treatment for meningitis was given in the paper “Penicillin in the Treatment of Meningitis” (*Journal of the American Medical Association* [1984]: 1870–1874). The ages (in years) were as follows:

18 18 25 19 23 20 69 18 21 18 20 18
18 20 18 19 28 17 18 18

- a. Calculate the values of the sample mean and the standard deviation.
b. Compute the upper quartile, the lower quartile, and the interquartile range.
c. Are there any mild or extreme outliers present in this data set?
d. Construct the boxplot for this data set.

4.70 Suppose that the distribution of scores on an exam can be described by a normal curve with mean 100. The 16th percentile of this distribution is 80.

- a. What is the 84th percentile?
b. What is the approximate value of the standard deviation of exam scores?
c. What z score is associated with an exam score of 90?
d. What percentile corresponds to an exam score of 140?
e. Do you think there were many scores below 40? Explain.

TECHNOLOGY NOTES

Mean

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**

Note: These commands also produce the following statistics: standard deviation, minimum, quartile 1, median, quartile 3, and maximum.

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**

3. Double-click C1 to add it to the **Variables** list
4. Click **OK**

Note: These commands also produce the following statistics: standard deviation, minimum, quartile 1, median, quartile 3, and maximum.

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Explore...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent List** box
5. Click **OK**

Note: These commands also produce the following statistics: median, variance, standard deviation, minimum, maximum, interquartile range, and several plots.

Excel 2007

1. Input the raw data into the first column
2. Click on the **Data** ribbon and select **Data Analysis**

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 3 for instructions on installing this add-on.

3. Select **Descriptive Statistics** from the dialog box
4. Click **OK**
5. Click in the box next to **Input Range:** and select the data (if you used and selected column titles, check the box next to **Labels in First Row**)
6. Check the box next to **Summary Statistics**
7. Click **OK**

Note: These commands also produce the following statistics: standard error, median, mode, standard deviation, sample variance, range, minimum, and maximum.

Note: You can also find the mean using the Excel function average.

TI-83/84

1. Input the raw data into **L1** (To access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **STAT** key
3. Use the arrows to highlight **CALC**
4. Highlight **1-Var Stats** and press **ENTER**
5. Press the **2nd** key and then the **1** key
6. Press **ENTER**

Note: You may need to scroll to view all of the statistics. This procedure also produces the standard deviation, minimum, Q1, median, Q3, maximum.

TI-Nspire

1. Enter the data into a data list (To access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then select **1:Stat Calculations** then **1:One-Variable Statistics...**
3. Press **OK**
4. For **X1 List**, select the title for the column containing your data from the drop-down menu
5. Press **OK**

Note: You may need to scroll to view all of the statistics. This procedure also produces the standard deviation, minimum, Q1, median, Q3, maximum.

Median

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**

Note: These commands also produce the following statistics: mean, standard deviation, minimum, quartile 1, quartile 3, and maximum.

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**
3. Double-click C1 to add it to the **Variables** list
4. Click **OK**

Note: These commands also produce the following statistics: mean, standard deviation, minimum, quartile 1, quartile 3, maximum.

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Explore...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent List** box
5. Click **OK**

Note: These commands also produce the following statistics: mean, variance, standard deviation, minimum, maximum, interquartile range, and several plots.

Excel 2007

1. Input the raw data into the first column
2. Click on the **Data** ribbon and select **Data Analysis**

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 3 for instructions on installing this add-on.

3. Select **Descriptive Statistics** from the dialog box
4. Click **OK**
5. Click in the box next to **Input Range:** and select the data (if you used and selected column titles, check the box next to **Labels in First Row**)

6. Check the box next to **Summary Statistics**
7. Click **OK**

Note: These commands also produce the following statistics: mean, standard error, mode, standard deviation, sample variance, range, minimum, and maximum.

Note: You can also find the median using the Excel function **median**.

TI-83/84

1. Input the raw data into **L1** (To access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **STAT** key
3. Use the arrows to highlight **CALC**
4. Highlight **1-Var Stats** and press **ENTER**
5. Press the **2nd** key and then the **1** key
6. Press **ENTER**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, standard deviation, minimum, Q1, Q3, maximum.

TI-Nspire

1. Enter the data into a data list (To access data lists select the spreadsheet option and press **enter**)
- Note:** Be sure to title the list by selecting the top row of the column and typing a title.
2. Press the **menu** key and select **4:Statistics** then select **1:Stat Calculations** then **1:One-Variable Statistics...**
3. Press **OK**
4. For **X1 List**, select the title for the column containing your data from the drop-down menu
5. Press **OK**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, standard deviation, minimum, Q1, Q3, maximum.

Variance

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to the column name
6. Click **Display Options** then select **More Moments**

Note: These commands also produce the following statistics: standard deviation, minimum, quartile 1, median, quartile 3, and maximum.

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**
3. Double-click C1 to add it to the **Variables** list

4. Click the **Statistics** button
5. Check the box next to **Variance**
6. Click **OK**
7. Click **OK**

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Explore...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent List** box
5. Click **OK**

Note: These commands also produce the following statistics: mean, median, standard deviation, minimum, maximum, interquartile range, and several plots.

Excel 2007

1. Input the raw data into the first column
2. Click on the **Data** ribbon and select **Data Analysis**
- Note:** If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 3 for instructions on installing this add-on.
3. Select **Descriptive Statistics** from the dialog box
4. Click **OK**
5. Click in the box next to **Input Range:** and select the data (if you used and selected column titles, check the box next to **Labels in First Row**)
6. Check the box next to **Summary Statistics**
7. Click **OK**

Note: These commands also produce the following statistics: mean, standard error, median, mode, standard deviation, range, minimum, and maximum.

Note: You can also find the variance using the Excel function **var**.

TI-83/84

The TI-83/84 does not automatically produce the variance; however, this can be determined by finding the standard deviation and squaring it.

TI-Nspire

The TI-Nspire does not automatically produce the variance; however, this can be determined by finding the standard deviation and squaring it.

Standard Deviation

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**

Note: These commands also produce the following statistics: mean, minimum, quartile 1, median, quartile 3, and maximum.

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**
3. Double-click C1 to add it to the **Variables** list
4. Click **OK**

Note: These commands also produce the following statistics: standard deviation, minimum, quartile 1, median, quartile 3, maximum.

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Explore...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent Listbox**
5. Click **OK**

Note: These commands also produce the following statistics: mean, median, variance, minimum, maximum, interquartile range, and several plots.

Excel 2007

1. Input the raw data into the first column
2. Click on the **Data** ribbon and select **Data Analysis**
3. Select **Descriptive Statistics** from the dialog box
4. Click **OK**
5. Click in the box next to **Input Range:** and select the data (if you used and selected column titles, check the box next to **Labels in First Row**)
6. Check the box next to **Summary Statistics**
7. Click **OK**

Note: These commands also produce the following statistics: mean, standard error, median, mode, sample variance, range, minimum, and maximum.

Note: You can also find the standard deviation using the Excel function **sd**.

TI-83/84

1. Input the raw data into **L1** (To access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **STAT** key
3. Use the arrows to highlight **CALC**
4. Highlight **1-Var Stats** and press **ENTER**
5. Press the **2nd** key and then the **1** key
6. Press **ENTER**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, minimum, Q1, median, Q3, maximum.

TI-Nspire

1. Enter the data into a data list (To access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then select **1:Stat Calculations** then **1:One-Variable Statistics...**
3. Press **OK**
4. For **X1 List**, select the title for the column containing your data from the drop-down menu
5. Press **OK**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, minimum, Q1, median, Q3, maximum.

Quartiles**JMP**

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**

Note: These commands also produce the following statistics: mean, standard deviation, minimum, median, and maximum.

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**
3. Double-click C1 to add it to the **Variables** list
4. Click **OK**

Note: These commands also produce the following statistics: standard deviation, minimum, quartile 1, median, quartile 3, maximum.

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Frequencies...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent List box**
5. Click the **Statistics** button
6. Check the box next to Quartiles
7. Click **Continue**
8. Click **OK**

Excel 2007

1. Input the raw data into the first column
2. Select the cell where you would like to place the first quartile results
3. Click the **Formulas** Ribbon
4. Click **Insert Function**
5. Select the category **Statistical** from the drop-down menu
6. In the **Select a function:** box click **Quartile**
7. Click **OK**
8. Click in the box next to **Array** and select the data
9. Click in the box next to **Quart** and type 1
10. Click **OK**

Note: To find the third quartile, type 3 into the box next to Quart in Step 9.

TI-83/84

1. Input the raw data into **L1** (To access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **STAT** key
3. Use the arrows to highlight **CALC**
4. Highlight **1-Var Stats** and press **ENTER**
5. Press the **2nd** key and then the **1** key
6. Press **ENTER**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, standard deviation, minimum, median, maximum.

TI-Nspire

1. Enter the data into a data list (To access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then select **1:Stat Calculations** then **1:One-Variable Statistics...**
3. Press **OK**
4. For **X1 List**, select the title for the column containing your data from the drop-down menu
5. Press **OK**

Note: You may need to scroll to view all of the statistics. This procedure also produces the mean, standard deviation, minimum, median, maximum.

IQR

JMP

JMP does not have the functionality to produce the iqr automatically.

Minitab

1. Input the raw data into C1
2. Select **Stat** and choose **Basic Statistics** then choose **Display Descriptive Statistics...**
3. Double-click C1 to add it to the **Variables** list
4. Click the **Statistics** button
5. Check the box next to **Interquartile range**
6. Click **OK**
7. Click **OK**

SPSS

1. Input the raw data into the first column
2. Select **Analyze** and choose **Descriptive Statistics** then choose **Explore...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Dependent List** box
5. Click **OK**

Note: These commands also produce the following statistics: mean, median, variance, standard deviation, minimum, maximum, and several plots.

Excel 2007

1. Use the steps under the **Quartiles** section to find both the first and third quartiles
2. Click on an empty cell where you would like the result for iqr to appear
3. Type = into the cell
4. Click on the cell containing the third quartile
5. Type –
6. Click on the cell containing the first quartile
7. Press **Enter**

TI-83/84

The TI-83/84 does not have the functionality to produce the iqr automatically.

TI-Nspire

The TI-Nspire does not have the functionality to produce the iqr automatically.

Boxplot

JMP

1. Input the raw data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**

Minitab

1. Input the raw data into C1
2. Select **Graph** then choose **Boxplot...**
3. Highlight Simple under One Y
4. Click **OK**
5. Double-click C1 to add it to the **Graph Variables** box
6. Click **OK**

Note: You may add or format titles, axis titles, legends, and so on by clicking on the **Labels...** button prior to performing Step 6 above.

SPSS

1. Enter the raw data the first column
2. Select **Graph** and choose **Chart Builder...**
3. Under **Choose from** highlight Boxplot
4. Click and drag the first boxplot (Simple Boxplot) to the Chart preview area
5. Click and drag the data variable into the **Y-Axis?** box in the chart preview area
6. Click **OK**

Note: Boxplots are also produced when summary statistics such as mean, median, standard deviation, and so on are produced.

Excel 2007

Excel 2007 does not have the functionality to create boxplots.

TI-83/84

1. Input the raw data into **L1** (To access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **2nd** key and then press the **Y =** key
3. Select **Plot1** and press **ENTER**
4. Highlight **On** and press **ENTER**
5. Highlight the graph option in the second row, second column, and press **ENTER**
6. Press **GRAPH**

Note: If the graph window does not display appropriately, press the **WINDOW** button and reset the scales appropriately.

TI-Nspire

1. Enter the data into a data list (To access data list select the spreadsheet option and press **enter**)
- Note:** Be sure to title the list by selecting the top row of the column and typing a title.
2. Press **menu** and select **3:Data** then select **6:QuickGraph**
3. Press **menu** and select **1:Plot Type** then select **2:Box Plot** and press **enter**

Side-by-side Boxplots**JMP**

1. Enter the raw data for both groups into a column
2. Enter the group information into a second column

Column 2	Column 3
F	8
F	3
F	6
F	3
F	9
F	2
F	3
F	4
F	8
M	3
M	5
M	4
M	6
M	7
M	2

3. Click **Analyze** then select **Fit Y by X**
4. Click and drag the column name containing the raw data from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the column name containing the group information from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Oneway Analysis of...**
8. Click **Quantiles**

Minitab

1. Input the raw data for the first group into C1
2. Input the raw data for the second group into C2
3. Continue to input data for each group into a separate column
4. Select **Graph** then choose **Boxplot...**
5. Highlight Simple under Multiple Y's
6. Click **OK**
7. Double-click the column names for each column to be graphed to add it to the **Graph Variables** box
8. Click **OK**

SPSS

1. Enter the raw data into the first column
2. Enter the group data into the second column

	VAR00001	VAR00002
1	3	male
2	6	female
3	5	male
4	7	female
5	5	male
6	8	female
7	2	male
8	3	female
9	6	male
10	5	female

3. Select **Graph** and choose **Chart Builder...**
4. Under **Choose from** highlight Boxplot
5. Click and drag the first boxplot (Simple Boxplot) to the Chart preview area
6. Click and drag the data variable into the **Y-Axis?** box in the chart preview area
7. Click and drag the group variable into the **X-Axis?** box in the chart preview area
8. Click **OK**

Excel 2007

Excel 2007 does not have the functionality to create side-by-side boxplots.

TI-83/84

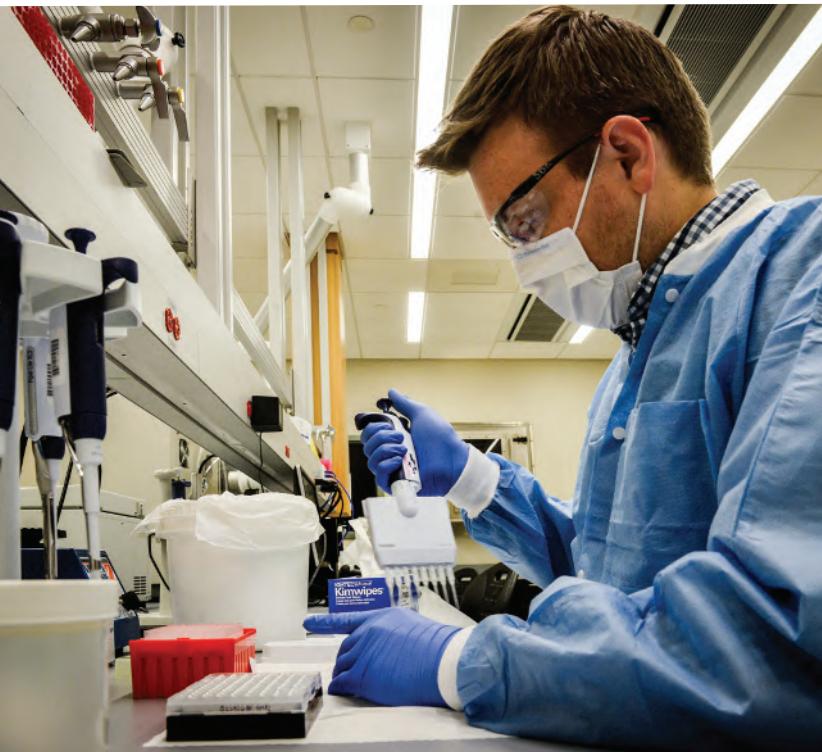
The TI-83/84 does not have the functionality to create side-by-side boxplots.

TI-Nspire

The TI-Nspire does not have the functionality to create side-by-side boxplots.

5

Summarizing Bivariate Data



The Washington Post/Getty Images

Forensic scientists must often estimate the age of an unidentified crime victim. Prior to 2010, this was usually done by analyzing teeth and bones, and the resulting estimates were not very reliable. A groundbreaking study described in the paper “[Estimating Human Age from T-Cell DNA Rearrangements](#)” (*Current Biology* [2010]) examined the relationship between age and a measure based on a blood test. Age and the blood test measure were recorded for 195 people ranging in age from a few weeks to 80 years. By studying this relationship, the investigators hoped to be able to estimate the age of a crime victim from a blood sample.

In this chapter, we will see how a line that summarizes the relationship between two numerical variables can be used to make predictions, and what can be said about the accuracy of those predictions.

LEARNING OBJECTIVES

Students will understand:

- That the relationship between two numerical variables can be described in terms of form, strength, and direction.
- The difference between a statistical relationship and a causal relationship (the difference between correlation and causation).
- How a line can be used to describe the relationship between two numerical variables.
- The meaning of least-squares in the context of fitting a least-squares line.
- Why it is risky to use the least-squares line to make predictions outside the range of the data.
- Why it is important to consider both the standard deviation about the least-squares line and the value of r^2 when assessing the usefulness of the least-squares line.
- The role of a residual plot in assessing whether a line is an appropriate way to describe the relationship between two numerical variables.

Students will be able to:

- Informally describe the form, direction, and strength of a linear relationship based on a scatterplot.
- Calculate and interpret the value of the correlation coefficient.

- Find the equation of the least-squares line and interpret the slope and intercept in context.
- Use the least-squares line to make predictions.
- Calculate and interpret the value of the standard deviation about the least-squares line, s_e .
- Calculate and interpret the value of r^2 , the coefficient of determination.
- Construct a residual plot and use it to assess whether using a line to describe the relationship between two numerical variables is appropriate.
- Identify outliers and potentially influential observations in a linear regression context.

SECTION 5.1 Correlation

Sometimes we might be interested in how two or more variables are related to one another. For example, an environmental researcher might want to know how the lead content of soil varies with distance from a major highway. Educational psychologists might wonder how vocabulary size is related to age. College admissions officers, who must try to predict whether an applicant will succeed in college, might be interested in the relationship between college grade point average and high school grade point average or score on an entrance exam.

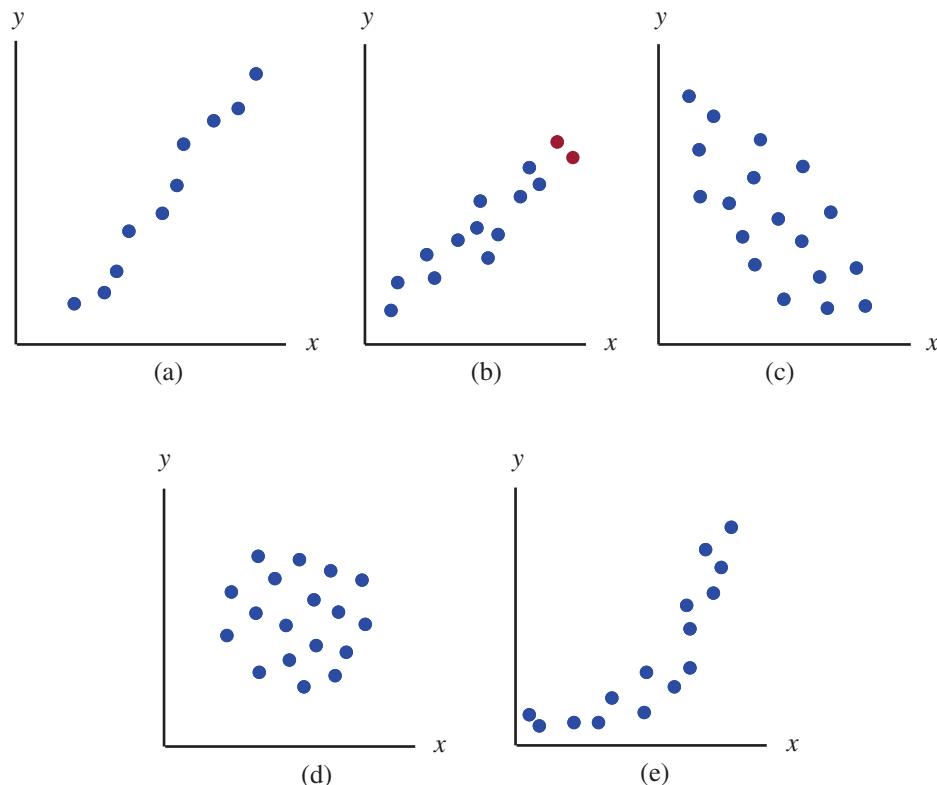
We can get a visual impression of how strongly two numerical variables are related by looking at a scatterplot. It is also possible to describe the strength of a relationship numerically. The **correlation coefficient** (from *co-* and *relation*) is a numerical assessment of the strength of the linear relationship between the x and y values in a bivariate data set consisting of (x, y) pairs.

Figure 5.1 displays several scatterplots that show different relationships between the x and y values. The plot in Figure 5.1(a) suggests a strong *positive relationship* between x and y . For every pair of points in the plot, the one with the larger x value also has the larger y value.

FIGURE 5.1

Scatterplots illustrating various types of relationships:

- positive linear relationship;
- positive linear relationship;
- negative linear relationship;
- no relationship;
- curved relationship.



The plot in Figure 5.1(b) shows a strong tendency for y to increase as x does, but there are a few exceptions. For example, the x and y values of the two points with the largest x values (shown in a different color) go in opposite directions (for this pair of points, x increases but y decreases in value). Even so, a plot like this still indicates a fairly strong positive relationship.

Figure 5.1(c) suggests that x and y are *negatively related*. As x increases, y tends to decrease. The negative relationship in this plot is not as strong as the positive relationship in Figure 5.1(b), although both plots show a well-defined linear pattern.

The plot of Figure 5.1(d) indicates that there is not a strong relationship between x and y . There is no tendency for y either to increase or to decrease as x increases. Finally, as illustrated in Figure 5.1(e), a scatterplot can show evidence of a strong relationship that is curved rather than linear.

The Sample Correlation Coefficient

The **sample correlation coefficient** measures the strength of a linear relationship between two numerical variables. The correlation coefficient is calculated using z scores. Consider replacing each x value by the corresponding z score, z_x (by subtracting \bar{x} and then dividing by s_x) and similarly replacing each y value by its z score. Notice that x values that are larger than \bar{x} will have positive z scores and those smaller than \bar{x} will have negative z scores. Also y values larger than \bar{y} will have positive z scores and those smaller will have negative z scores. The value of the correlation coefficient is based on the sum of the products of z_x and z_y for each observation in the bivariate data set. This can be written as $\sum z_x z_y$.

To see how this works, let's look at some scatterplots. The scatterplot in Figure 5.2(a) indicates a strong positive relationship. A vertical line through \bar{x} and a horizontal line through \bar{y} divide the plot into four regions. In Region I, both x and y are greater than their mean values, so the z score for x and the z score for y are both positive numbers. It follows that $z_x z_y$ is positive. The product of the z scores is also positive for any point in Region III, because both z scores are negative in Region III and multiplying two negative numbers results in a positive number. In each of the other two regions, one z score is positive and the other is negative, so $z_x z_y$ is negative. But because the points generally fall in Regions I and III, the products of z scores tend to be positive. This means that the *sum* of the products will be a relatively large positive number.

Similar reasoning for the data displayed in Figure 5.2(b), which exhibits a strong negative relationship, implies that $\sum z_x z_y$ will be a relatively large (in magnitude) negative number. When there is no strong relationship, as in Figure 5.2(c), positive and negative products tend to offset one another, producing a value of $\sum z_x z_y$ that is close to zero.

For these reasons, $\sum z_x z_y$ can be used as a measure of the strength of the linear relationship between x and y . It can be a large positive number, a large negative number, or a number close to 0, depending on whether there is a strong positive, a strong negative, or no linear relationship.

The sample correlation coefficient, denoted by r , is calculated by dividing $\sum z_x z_y$ by $(n - 1)$.

DEFINITION

Sample correlation coefficient: A measure of the strength and direction of a linear relationship between two numerical variables. Denoted by r , the sample correlation coefficient is given by

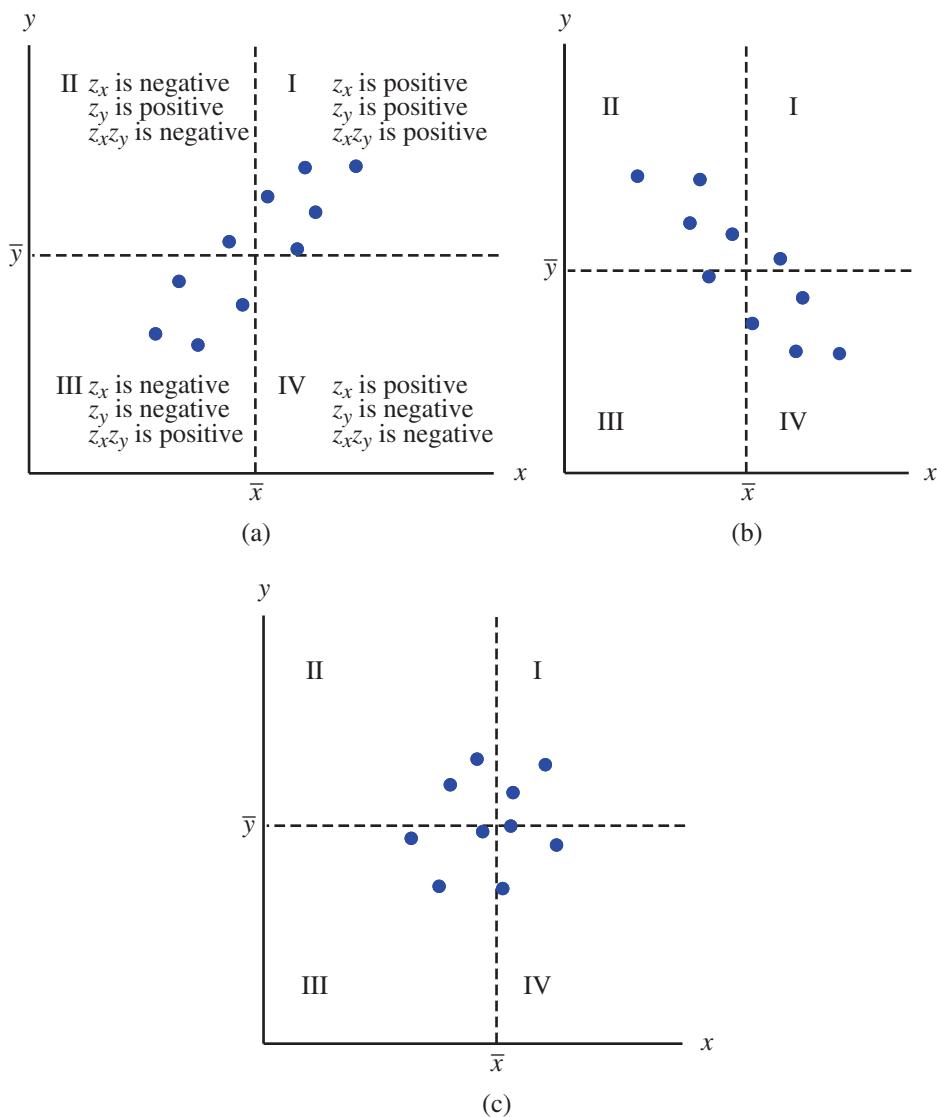
$$r = \frac{\sum z_x z_y}{n - 1}$$

Hand calculation of the correlation coefficient is quite tedious. Fortunately, all statistical software packages and most graphing calculators can compute the value of r once the x and y values have been input.

FIGURE 5.2

Viewing a scatterplot according to the signs of z_x and z_y :

- a positive relationship;
- a negative relationship;
- no strong relationship.



Example 5.1 A Face You Can Trust

Understand the context ➤

The article “How to Tell If a Guy Is Trustworthy” (*LiveScience*, March 8, 2010) described an interesting research study published in the journal *Psychological Science* (“Valid Facial Cues to Cooperation and Trust: Male Facial Width and Trustworthiness,” *Psychological Science* [2010]: 349–354). This study investigated whether there is a relationship between facial characteristics and peoples’ assessment of trustworthiness. Sixty-two students were each told that they would play a series of two-person games in which they could earn real money. In each game, the student could choose between two options:

1. They could end the game immediately and a total of \$10 would be paid out, with each player receiving \$5.
or
2. They could “trust” the other player. In this case, \$3 would be added to the money available, but the other player could decide to either split the money fairly with \$6.50 to each player or could decide to keep \$10 and only give \$3 to the first player.

Each student was shown a picture of the person he or she would be playing with. The student then indicated whether they would end the game immediately and take \$5 or trust the other player to do the right thing in hopes of increasing the payout to \$6.50. This process was repeated for a series of games, each with a photo of a different second player. For each

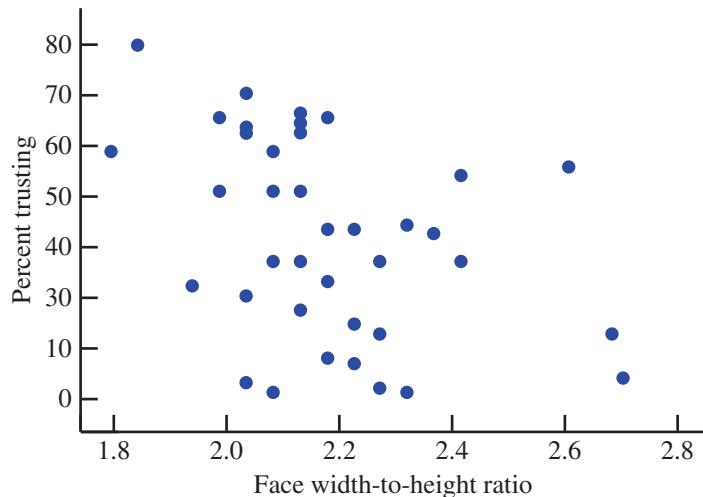
photo, the researchers recorded the width-to-height ratio of the face in the photo and the percentage of the students who chose the trust option when shown that photo. A representative subset of the data (from a graph that appeared in the paper) is given here:

Face Width-to-Height Ratio	Percentage Choosing Trust Option	Face Width-to-Height Ratio	Percentage Choosing Trust Option	Face Width-to-Height Ratio	Percentage Choosing Trust Option	Face Width-to-Height Ratio	Percentage Choosing Trust Option
1.75	58	2.05	8	2.15	42	2.30	43
1.80	80	2.05	35	2.15	31	2.30	8
1.90	30	2.05	50	2.15	15	2.35	41
1.95	50	2.05	58	2.15	65	2.40	53
1.95	65	2.10	64	2.20	42	2.40	35
2.00	10	2.10	25	2.20	22	2.60	55
2.00	28	2.10	35	2.20	14	2.68	20
2.00	62	2.10	50	2.25	9	2.70	11
2.00	70	2.10	66	2.25	20		
2.00	63	2.10	62	2.25	35		

The first observation (1.75, 58) indicates that 58% of the students shown the picture of a person with a width-to-height ratio of 1.75 chose the trust option.

A scatterplot of these data is shown in Figure 5.3.

FIGURE 5.3
Scatterplot for the data of Example 5.1.



JMP was used to compute the value of the correlation coefficient, with the following result:

Correlation						
Variable	Mean	Std Dev	Correlation	Signif. Prob	Number	
Percentage Choosing Trust Option	40.26316	20.58691	-0.39169	0.0150*	38	
Face Width-to-Height Ratio	2.153421	0.209598				

The value of the correlation coefficient (-0.39169) is found in the JMP output under the heading “Correlation.” Rounded to two decimal places, $r = -0.39$.

To calculate the value of the correlation coefficient by hand, we can use x to denote the face width-to-height ratio and y to denote the percentage choosing the trust option. For the given data,

$$\bar{x} = 2.153 \quad s_x = 0.210 \quad \bar{y} = 40.26 \quad s_y = 20.59$$

Do the work ➤ Begin by calculating z scores for each (x, y) pair in the data set. For example, the first observation is $(1.75, 58)$. The corresponding z scores are

$$z_x = \frac{1.75 - 2.153}{0.210} = -1.92 \quad z_y = \frac{58 - 40.26}{20.59} = 0.86$$

The following table shows the z scores and the product $z_x z_y$ for each observation:

x	y	z_x	z_y	$z_x z_y$	x	y	z_x	z_y	$z_x z_y$	x	y	z_x	z_y	$z_x z_y$
1.75	58	-1.92	0.86	-1.65	2.05	58	-0.49	0.86	-0.42	2.20	14	0.22	-1.28	-0.28
1.80	80	-1.68	1.93	-3.24	2.10	64	-0.25	1.15	-0.29	2.25	9	0.46	-1.52	-0.70
1.90	30	-1.2	-0.50	0.60	2.10	25	-0.25	-0.74	0.19	2.25	20	0.46	-0.98	-0.45
1.95	50	-0.97	0.47	-0.46	2.10	35	-0.25	-0.26	0.07	2.25	35	0.46	-0.26	-0.12
1.95	65	-0.97	1.20	-1.16	2.10	50	-0.25	0.47	-0.12	2.30	43	0.70	0.13	0.09
2.00	10	-0.73	-1.47	1.07	2.10	66	-0.25	1.25	-0.31	2.30	8	0.70	-1.57	-1.10
2.00	28	-0.73	-0.60	0.44	2.10	62	-0.25	1.06	-0.27	2.35	41	0.94	0.04	0.04
2.00	62	-0.73	1.06	-0.77	2.15	42	-0.01	0.08	0.00	2.40	53	1.18	0.62	0.73
2.00	70	-0.73	1.44	-1.05	2.15	31	-0.01	-0.45	0.00	2.40	35	1.18	-0.26	-0.31
2.00	63	-0.73	1.10	-0.80	2.15	15	-0.01	-1.23	0.01	2.60	55	2.13	0.72	1.53
2.05	8	-0.49	-1.57	0.77	2.15	65	-0.01	1.20	-0.01	2.68	20	2.51	-0.98	-2.46
2.05	35	-0.49	-0.26	0.13	2.20	42	0.22	0.08	0.02	2.70	11	2.60	-1.42	-3.69
2.05	50	-0.49	0.47	-0.23	2.20	22	0.22	-0.89	-0.20					

$\Sigma z_x z_y = -14.4$ and $n = 38$, so

$$r = \frac{\Sigma z_x z_y}{n - 1} = \frac{-14.4}{37} = -0.39$$

Interpret the results ➤ Based on the scatterplot and the properties of the correlation coefficient presented in the discussion that follows this example, we conclude that there is a weak negative linear relationship between face width-to-height ratio and the percentage of people who chose to trust the face when playing the game. The relationship is negative, indicating that those with larger face width-to-height ratios (wider faces) tended to be trusted less. Based on this observation, the researchers concluded “It is clear that male facial width ratio is a cue to male trustworthiness and that it predicts trust placed in male faces.”

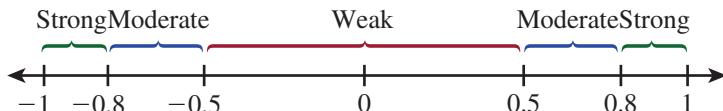
An interesting side note: In another study where subjects were randomly assigned to the player 1 and player 2 roles, the researchers found that those with larger face width-to-height ratios *were* less trustworthy! They found a positive correlation between player 2 face width-to-height ratio and the percentage of the time that player 2 decided to keep \$10 and only give \$3 to player 1.

Properties of r

1. The value of r is between -1 and $+1$. A value near the upper limit, $+1$, indicates a strong positive linear relationship, whereas a value of r close to the lower limit, -1 , suggests a strong negative linear relationship. Figure 5.4 shows a useful way to describe the strength of relationship based on r . It may seem surprising that a value of r as small as -0.5 or as large as 0.5 is in the weak category. An explanation for this is given later in the chapter.

FIGURE 5.4

Describing the strength of a linear relationship.



2. A correlation coefficient of $r = 1$ occurs only when all the points in a scatterplot of the data lie exactly on a straight line that slopes upward. Similarly, $r = -1$ only when all the points lie exactly on a downward-sloping line. $r = 1$ or $r = -1$ indicates a perfect linear relationship between x and y in the sample data.
3. The value of r is a measure of the extent to which x and y are linearly related. r measures the extent to which the points in the scatterplot fall close to a straight line. A value of r close to 0 does not necessarily mean that there is no relationship between x and y . It is possible that there could still be a strong relationship that is not linear.
4. The value of r does not depend on the unit of measurement for either variable. For example, if x is height, the corresponding z score is the same whether height is expressed in inches, meters, or miles, and the value of the correlation coefficient is not affected. The correlation coefficient measures the inherent strength of the linear relationship between two numerical variables.
5. The value of r does not depend on which of the two variables is considered x . This means that if we had let x = percentage choosing trust option and y = face width-to-height ratio in Example 5.1, the same value, $r = -0.39$, would have resulted.

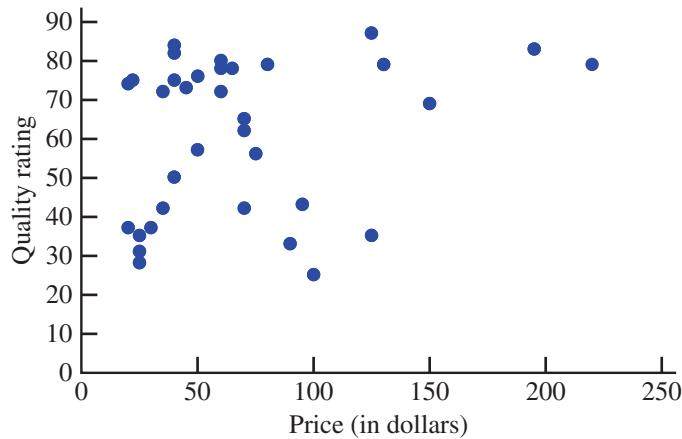
Example 5.2 Does It Pay to Pay More for a Bike Helmet?

Understand the context ➤

- Are more expensive bike helmets safer than less expensive ones? Data on x = price and y = quality rating for 35 different brands of bike helmets were used to construct the scatterplot in Figure 5.5. The data are from the *Consumer Reports* web site (consumerreports.org/products/bike-helmets/ratings-overview/, retrieved August 28, 2016). Quality rating was a number from 0 (the worst possible rating) to 100, and was determined based on factors that included how well the helmet absorbed the force of an impact, the strength of the helmet, ventilation, and ease of use. Figure 5.5 shows a scatterplot of the data.

FIGURE 5.5

Minitab scatterplot for the bike helmet data of Example 5.2.



From the scatterplot, it appears that there is only a weak positive relationship between price and quality rating. The correlation coefficient, obtained using Minitab, is

Correlations: Price, Quality Rating

• Data set available online

Correlation of Price and Quality Rating = 0.221

- Interpret the results ➤ A correlation coefficient of $r = 0.221$ indicates that although there is a weak tendency for higher quality ratings to be associated with higher priced helmets, the relationship is not very strong. In fact, some inexpensive helmets had high quality ratings.

Example 5.3 Age and Marathon Times

Understand the context ➤

- Data from the New York Marathon web site (results.nyrr.org/event/M2017/finishers, retrieved February 28, 2018) were used to calculate the average finishing time (in minutes) for the top 10 finishers by age group for female participants in the 2017 New York City marathon.

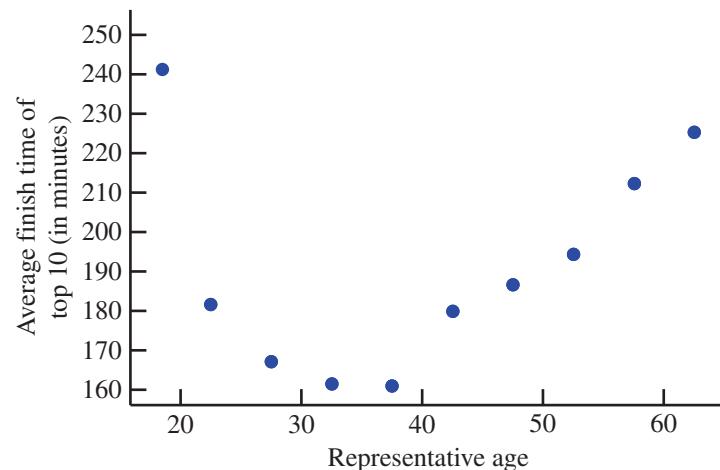
Consider the data ➤

Age Group	Representative Age	Average Finish Time of Top 10 (in minutes)
18–19	18.5	241.1
20–24	22.5	181.6
25–29	27.5	167.1
30–34	32.5	161.4
35–39	37.5	160.9
40–44	42.5	179.8
45–49	47.5	186.5
50–54	52.5	194.2
55–59	57.5	212.2
60–64	62.5	225.3

The scatterplot of average finish time versus representative age is shown in Figure 5.6.

FIGURE 5.6

Scatterplot of y = Average finish time and x = Representative age for the data of Example 5.3.



Using Minitab to compute the correlation coefficient between age and average finish time results in the following:

Correlations: Age, Average Finish Time

Correlation of Representative Age and Average Finish Time = 0.218

Interpret the results ➤

This example shows the importance of interpreting r as a measure of the strength of a *linear* relationship. Here, the value of r is not large, but there is a strong nonlinear relationship between age and average finish time. This is an important point—we should not conclude that there is no relationship whatsoever simply because the value of r is small in absolute value. Be sure to look at the scatterplot of the data before concluding that there is no relationship between two variables based on a correlation coefficient with a value near 0.

The Population Correlation Coefficient

The sample correlation coefficient r measures how strongly the x and y values in a *sample* of pairs are linearly related to one another. There is an analogous measure of how strongly x and y are related in the entire population of pairs from which the sample was selected. It is called the **population correlation coefficient** and is denoted by ρ . (Notice again the use of a Greek letter for a population characteristic and a Roman letter for a sample characteristic.) It is unusual to calculate ρ from an entire population of pairs, but it is important to know that ρ satisfies properties paralleling those of r :

1. ρ is a number between -1 and $+1$ that does not depend on the unit of measurement for either x or y , or on which variable is labeled x and which is labeled y .
2. $\rho = +1$ or -1 if and only if all (x, y) pairs in the population lie exactly on a straight line, so ρ measures the extent to which there is a linear relationship in the population.

In Chapter 13, we will see how the sample correlation coefficient r can be used to draw conclusions about the value of the population correlation coefficient ρ .

Correlation and Causation

A value of r close to 1 indicates that the larger values of one variable tend to be associated with the larger values of the other variable. This is different from saying that a large value of one variable *causes* the value of the other variable to be large. Correlation measures the extent of association, but *association does not imply causation*.

It frequently happens that two variables are highly correlated not because one is causally related to the other but because they are both strongly related to a third variable. Among all elementary school children, the relationship between the number of cavities in a child's teeth and the size of his or her vocabulary is strong and positive. Yet no one advocates eating foods that result in more cavities to increase vocabulary size (or working to decrease vocabulary size to protect against cavities). Number of cavities and vocabulary size are both strongly related to age, so older children tend to have higher values of both variables than do younger ones.

In the ABCNews.com series “Who’s Counting?” (February 1, 2001), John Paulos reminded readers that correlation does not imply causation and gave the following example: Consumption of hot chocolate is negatively correlated with crime rate (high values of hot chocolate consumption tend to be paired with lower crime rates), but both are responses to cold weather.

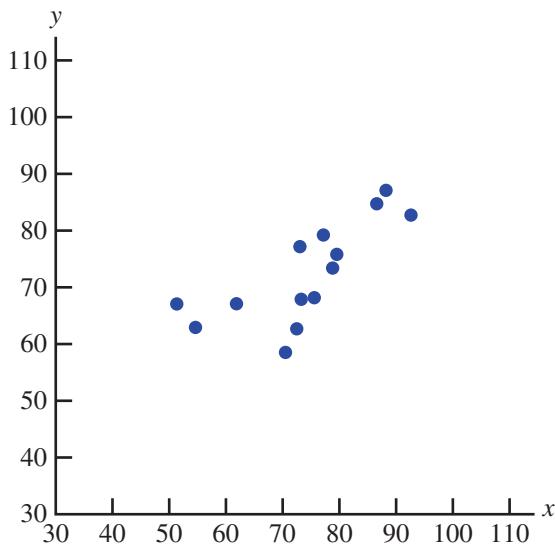
Scientific experiments can frequently make a strong case for causality by carefully controlling the values of all variables that might be related to the ones under study. Then, if y is observed to change in a “smooth” way as the experimenter changes the value of x , a plausible explanation would be that there is a causal relationship between x and y . In the absence of such control and the ability to manipulate values of one variable, we must admit the possibility that an unidentified underlying third variable is influencing both variables under investigation. A high correlation in many uncontrolled studies carried out in different settings can also marshal support for causality—as in the case of cigarette smoking and cancer—but establishing causality is an elusive task.

EXERCISES 5.1 - 5.16

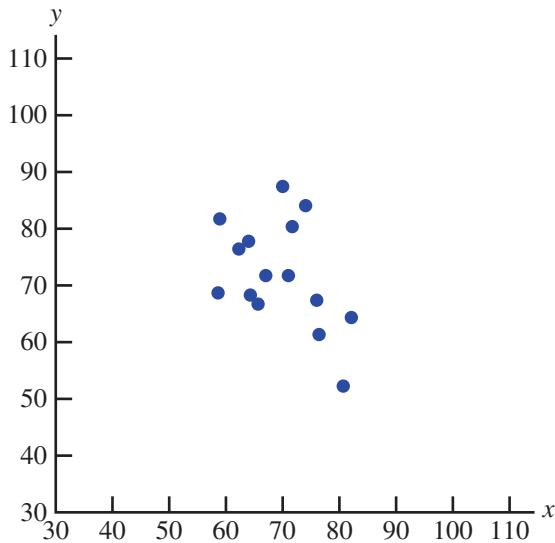
● Data set available online

- 5.1** For each of the scatterplots shown, answer the following questions:
- a. Does there appear to be a relationship between x and y ?
 - b. If so, does the relationship appear to be linear?
 - c. If so, would you describe the linear relationship as positive or negative?

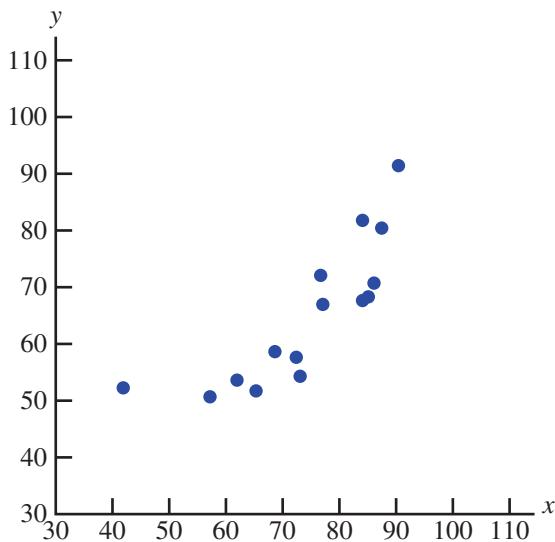
Scatterplot 1:



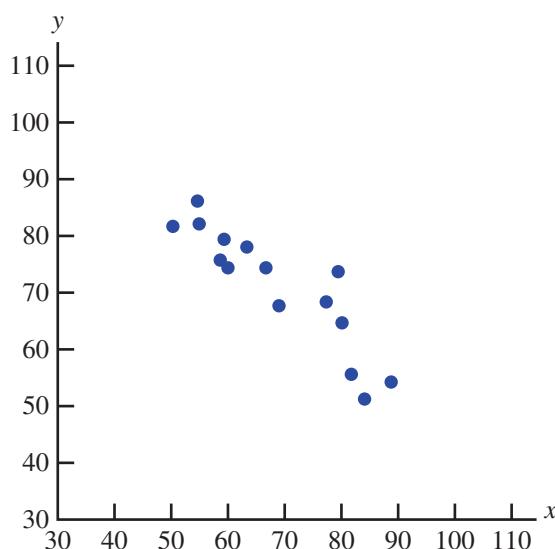
Scatterplot 2:



Scatterplot 3:



Scatterplot 4:



- 5.2** For each of the following pairs of variables, indicate whether you would expect a positive correlation, a negative correlation, or a correlation close to 0. Explain your choice.

- Maximum daily temperature and cooling costs
- Interest rate and number of loan applications
- Amount of fertilizer used per acre and crop yield
(Hint: As the amount of fertilizer is increased, yield tends to increase for a while but then tends to start decreasing.)

- 5.3** For each of the following pairs of variables, indicate whether you would expect a positive correlation, a negative correlation, or a correlation close to 0. Explain your choice.

- Incomes of husbands and wives when both have full-time jobs
- Height and IQ
- Height and shoe size

- 5.4** For each of the following pairs of variables, indicate whether you would expect a positive correlation, a negative correlation, or a correlation close to 0. Explain your choice.

- Score on the math section of the SAT exam and score on the verbal section of the same test
- Time spent on homework and time spent watching television during the same day by elementary school children

- 5.5** Is the following statement correct? Explain why or why not. (Hint: See Example 5.3.)

A correlation coefficient of 0 implies that no relationship exists between the two variables under study.

5.6 Draw a scatterplot for which $r = 1$.

5.7 Draw a scatterplot for which $r = -1$.

5.8 Each year J.D. Power and Associates surveys new car owners 90 days after they purchase their cars. This data is used to rate auto brands (such as Toyota and Ford) on quality and customer satisfaction. [USA TODAY \(usatoday.com, March 29, 2016\)](#) reported a quality rating and a satisfaction score for all 33 brands sold in the United States.

Brand	Quality Rating	APEAL Rating
Acura	64	814
Audi	80	858
BMW	76	849
Buick	74	792
Cadillac	58	826
Chevrolet	64	791
Chrysler	58	783
Dodge	58	790
Fiat	38	768
Ford	66	789
GMC	60	791
Honda	71	783
Hyundai	70	804
Infiniti	63	826
Jeep	43	762
Kia	72	791
Land Rover	55	853
Lexus	76	844
Lincoln	65	835
Mazda	74	790
Mercedes-Benz	67	842
MINI	68	795
Mitsubishi	51	748
Nissan	63	786
Porsche	76	882
Scion	65	779
Subaru	78	766
Toyota	72	783
Volkswagen	67	796
Volvo	69	812

a. Construct a scatterplot of $y =$ Satisfaction rating versus $x =$ Quality rating. How would you describe the relationship between x and y ?

b. Calculate and interpret the value of the correlation coefficient.

5.9 ● The accompanying data are $x =$ Cost (cents per serving) and $y =$ Fiber content (grams per serving) for 18 high-fiber cereals rated by [Consumer Reports \(consumerreports.org/health\)](#).

Cost per Serving	Fiber per Serving	Cost per Serving	Fiber per Serving
33	7	53	13
46	10	53	10
49	10	67	8
62	7	43	12
41	8	48	7
19	7	28	14
77	12	54	7
71	12	27	8
30	8	58	8

a. Calculate and interpret the value of the correlation coefficient for this data set. (Hint: See Example 5.1.)

b. The serving size differed for the different cereals, with serving sizes varying from $\frac{1}{2}$ cup to $1\frac{1}{4}$ cups. Converting price and fiber content to “per cup” rather than “per serving” results in the accompanying data. Is the correlation coefficient for the per cup data greater than or less than the correlation coefficient for the per serving data?

Cost per Cup	Fiber per Cup	Cost per Cup	Fiber per Cup
9.3	44.0	13.0	53.0
10.0	46.0	10.0	53.0
10.0	49.0	8.0	67.0
7.0	62.0	12.0	43.0
6.4	32.8	7.0	48.0
7.0	19.0	28.0	56.0
12.0	77.0	7.0	54.0
9.6	56.8	16.0	54.0
8.0	30.0	10.7	77.3

5.10 The authors of the paper [“Flat-footedness Is Not a Disadvantage for Athletic Performance in Children Aged 11 to 15 Years” \(Pediatrics \[2009\]: e386–e392\)](#) studied the relationship between $y =$ Arch height and scores on a number of different motor ability tests for 218 children. They reported the following correlation coefficients:

Correlation between Test Score and Arch Height	
Motor Ability Test	
Height of counter movement jump	-0.02
Hopping: average height	-0.10
Hopping: average power	-0.09
Balance, closed eyes, one leg	0.04
Toe flexion	0.05

- a. Interpret the value of the correlation coefficient between average hopping height and arch height. What does the fact that the correlation coefficient is negative say about the relationship? Do higher arch heights tend to be paired with higher or lower average hopping heights?
- b. The title of the paper suggests that having a small value for arch height (flat-footedness) is not a disadvantage when it comes to motor skills. Do the given correlation coefficients support this conclusion? Explain.

5.11 The paper “[The Relationship Between Cell Phone Use, Academic Performance, Anxiety, and Satisfaction with Life in College Students](#)” (*Computers in Human Behavior* [2014]: 343–350) described a study of cell phone use among undergraduate college students at a public university. The paper reported that the value of the correlation coefficient between x = Cell phone use (measured as total amount of time in hours spent using a cell phone on a typical day) and y = GPA (cumulative GPA determined from university records) was $r = -0.203$.

- a. Interpret the given value of the correlation coefficient. Does the value of the correlation coefficient suggest that students who use a cell phone for more hours per day tend to have higher GPAs or lower GPAs?
- b. The study also investigated the correlation between texting (measured as the total number of texts sent and texts received per day) and GPA. The direction of the relationship between texting and GPA was the same as the direction of the relationship between cell phone use and GPA, but the relationship between texting and GPA was not as strong. Which of the following possible values for the correlation coefficient between texting and GPA could have been the one observed?

$$r = -0.30 \quad r = -0.10 \quad r = 0.10 \quad r = 0.30$$

- c. The paper included the following statement: “Participants filled in two blanks—one for texts sent and one for texts received. These two texting items were nearly perfectly correlated.” Do you think that the value of the correlation coefficient for texts sent and texts received was close to -1 , close to 0 , or close to $+1$? Explain your reasoning.

5.12 ● Data from the [U.S. Federal Reserve Board](#) (federalreserve.gov/releases/housedebt/, retrieved April 21, 2017) on consumer debt (as a percentage of personal income) and mortgage debt (also as a percentage of personal income) for the 10 years from 2006 to 2015 are shown in the following table:

Consumer Debt	Household Debt
6.73	5.97
7.07	5.97
6.96	5.87
6.69	5.60
6.18	5.16
5.58	5.06
5.14	4.97
4.95	5.22
4.73	5.31
4.59	5.45

- a. What is the value of the correlation coefficient for this data set?
- b. Is it reasonable to conclude in this case that there is no strong relationship between the variables (linear or otherwise)? Use a graphical display to support your answer.
- 5.13** The article “[\\$115K! The 13 Best Paying U.S. Companies](#)” (*USA TODAY*, August 11, 2015) gave the following data on median worker pay (in thousands of dollars) and the 1-year percent change in stock price for the 13 highest paying companies in the United States.

Company	Median Worker Pay	Percent Change in Stock Price
Jupiter Networks	134.7	21.7
Netflix	132.2	92.2
Equinix	125.0	33.9
Altera	122.8	52.2
Visa	122.5	41.9
Yahoo	121.2	2.8
Xilinx	121.2	5.2
VeriSign	119.0	30.1
Microsoft	118.0	8.0
Broadcom	118.0	37.3
F5 Networks	117.6	15.9
Adobe Systems	117.4	22.0
eBay	115.1	-44.7

- a. Construct a scatterplot for these data.
- b. Calculate and interpret the value of the correlation coefficient.
- c. The article states that companies that pay more are seeing a payoff in their stock performance. Is this conclusion justified based on these data? Explain.
- d. Is it reasonable to generalize conclusions based on these data to the population of all companies in the United States? Explain why or why not.

- 5.14** It may seem odd, but one of the ways biologists can tell how old a lobster is involves measuring the concentration of a pigment called neurolipofuscin in the eyestalk of a lobster. (We are not making this up!) The authors of the paper “[Neurolipofuscin Is](#)

a Measure of Age in *Panulirus argus*, the Caribbean Spiny Lobster, in Florida" (*Biological Bulletin* [2007]: 55–66) wondered if it was sufficient to measure the pigment in just one eye stalk, which would be the case if there is a strong relationship between the concentration in the right and left eyestalks.

Pigment concentration (as a percentage of tissue sample) was measured in both eyestalks for 39 lobsters, resulting in the following summary quantities (based on data read from a graph that appeared in the paper):

$$\begin{array}{lll} n = 39 & \Sigma x = 88.8 & \Sigma y = 86.1 \\ \Sigma xy = 281.1 & \Sigma x^2 = 288.0 & \Sigma y^2 = 286.6 \end{array}$$

An alternative formula for calculating the correlation coefficient that is based on raw data and is algebraically equivalent to the one given in the text is

$$r = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} \sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}}$$

Use this formula to calculate the value of the correlation coefficient, and interpret this value.

- 5.15** An auction house released a list of 25 recently sold paintings. Eight artists were represented in these sales. The sale price of each painting also appears on the list. Would the correlation coefficient be an appropriate way to summarize the relationship between artist (x) and sale price (y)? Why or why not?

- 5.16** A sample of automobiles traversing a certain stretch of highway is selected. Each one travels at roughly a constant rate of speed, although speed does vary from auto to auto. Let x = speed and y = time needed to traverse this segment of highway. Would the sample correlation coefficient be closest to 0.9, 0.3, -0.3, or -0.9? Explain.

SECTION 5.2 Linear Regression: Fitting a Line to Bivariate Data

The objective of *regression analysis* is to use information about one variable, x , to predict the value of a second variable, y . For example, we might want to predict y = Product sales during a given period when the amount spent on advertising is x = \$10,000. The two variables in a regression analysis play different roles: y is called the **dependent** or **response variable**, and x is referred to as the **independent, predictor, or explanatory variable**.

DEFINITIONS

Dependent variable: In a bivariate data set, the variable whose value we would like to predict. The dependent variable is denoted by y . The dependent variable is also sometimes called the **response variable**.

Independent variable: In a bivariate data set, the variable that will be used to make a prediction of the dependent variable. The independent variable is denoted by x . The independent variable is also sometimes called the **predictor variable** or the **explanatory variable**.

Scatterplots frequently exhibit a linear pattern. When this is the case, it makes sense to summarize the relationship between the variables using a line. Before seeing how this is done, let's review some elementary facts about lines and linear relationships.

The equation of a line is $y = a + bx$. A particular line is specified by choosing values of a and b . For example, one line is $y = 10 + 2x$; another is $y = 100 - 5x$. If we choose some x values and calculate $y = a + bx$ for each value, the points in the plot of the resulting (x, y) pairs will fall exactly on a straight line.

DEFINITIONS

The equation of a line is

$$y = a + bx$$

↑ Intercept
 ↓ slope

Slope: The value of b , called the **slope** of the line, is the amount by which y increases when x increases by 1 unit.

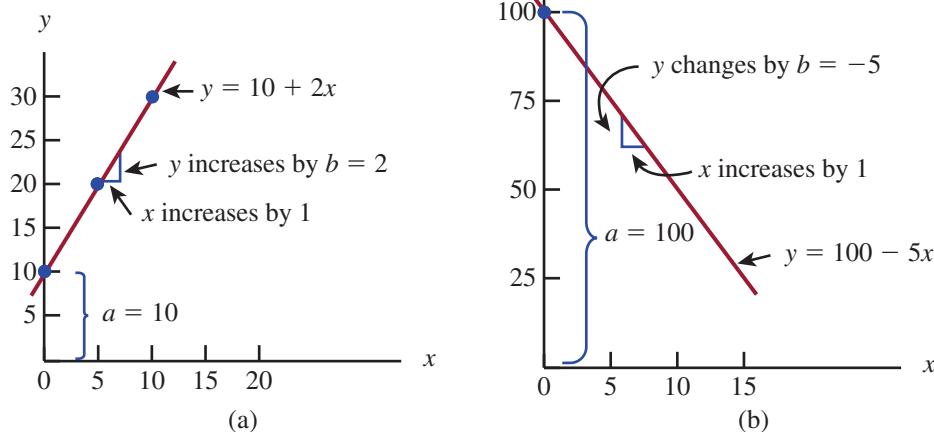
Intercept: The value of a , called the **intercept** (or sometimes the **y -intercept** or **vertical intercept**) of the line, is the height of the line above the value $x = 0$.

The line $y = 10 + 2x$ has slope $b = 2$, so each 1-unit increase in x is paired with an increase of 2 in y . When $x = 0$, $y = 10$, so the height of the line where it crosses the vertical axis (where $x = 0$) is 10. This is illustrated in Figure 5.7(a). The slope of the line $y = 100 - 5x$ is -5 , so y increases by -5 (or equivalently, decreases by 5) when x increases by 1. The height of the line above $x = 0$ is $a = 100$. This line is pictured in Figure 5.7(b).

FIGURE 5.7

Graphs of two lines:

- (a) slope $b = 2$, intercept $a = 10$;
- (b) slope $b = -5$, intercept $a = 100$.



It is easy to draw the graph of a line. Choose any two x values and substitute them into the equation of the line to obtain the two corresponding y values. Then plot the resulting two (x, y) pairs as two points. The line passes through these points. For the equation $y = 10 + 2x$, substituting $x = 5$ yields $y = 20$, and using $x = 10$ gives $y = 30$. The resulting two points are then $(5, 20)$ and $(10, 30)$. The line in Figure 5.7(a) passes through these points.

Fitting a Straight Line: The Principle of Least Squares

Figure 5.8 shows a scatterplot that also includes two lines. Line II is a better fit to the data than Line I is. Vertical deviations from the line can be used to measure the extent to which a particular line provides a good fit to data.

Line II in Figure 5.8 has equation $y = 10 + 2x$. The third and fourth points from the left in the scatterplot are $(15, 44)$ and $(20, 45)$. For these two points, the vertical deviations from this line are

$$\begin{aligned}
 \text{3rd deviation} &= y_3 - \text{height of the line above } x_3 \\
 &= 44 - [10 + 2(15)] \\
 &= 4
 \end{aligned}$$

and

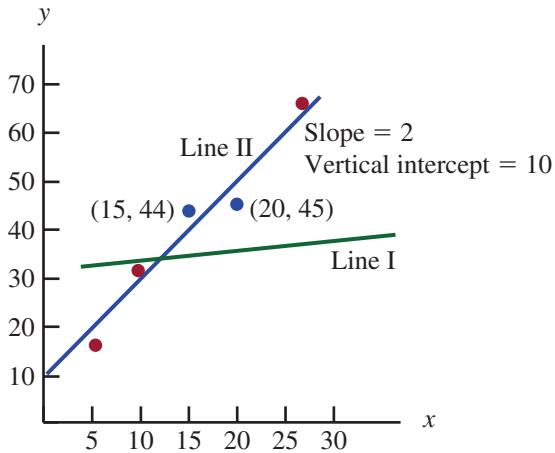
$$4\text{th deviation} = 45 - [10 + 2(20)] = -5$$

A positive vertical deviation results from a point that lies above the line, and a negative deviation results from a point that lies below the line.

A particular line is said to be a good fit to the data if the deviations from the line are small in magnitude. Line I in Figure 5.8 fits poorly, because all deviations from that line are larger in magnitude (some are much larger) than the corresponding deviations from Line II.

FIGURE 5.8

Line I gives a poor fit and Line II gives a good fit to the data.



To assess the overall fit of a line, we need a way to combine all of the deviations into a single measure of fit. One way is to square the deviations (to obtain nonnegative numbers) and to then add up the squared deviations.

DEFINITIONS

Sum of squared deviations: The most widely used measure of the goodness of fit of a line $y = a + bx$ to bivariate data $(x_1, y_1), \dots, (x_n, y_n)$ is the **sum of the squared deviations** about the line

$$\Sigma[y - (a + bx)]^2 = [y_1 - (a + bx_1)]^2 + [y_2 - (a + bx_2)]^2 + \dots + [y_n - (a + bx_n)]^2$$

Least-squares line: The line that minimizes the sum of squared deviations. The least-squares line is also sometimes called the **sample regression line**.

Fortunately, the equation of the least-squares line can be obtained without having to calculate deviations from any particular line. The accompanying box gives formulas for the slope and intercept of the least-squares line.

The slope of the least-squares line is

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

and the y intercept is

$$a = \bar{y} - b\bar{x}$$

We write the equation of the least-squares line as

$$\hat{y} = a + bx$$

where the $\hat{\cdot}$ above y indicates that \hat{y} (read as y -hat) is the prediction of y that results from substituting a particular x value into the equation.

Statistical software packages and many calculators can compute the slope and intercept of the least-squares line. If the slope and intercept are to be calculated by hand, the following calculating formula can be used to reduce the amount of time required to perform the calculations.

Calculating Formula for the Slope of the Least-Squares Line

$$b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

Example 5.4 Pomegranate Juice and Tumor Growth

Understand the context ➤

- Pomegranate, a fruit native to Persia, has been used in the folk medicines of many cultures to treat various ailments. Researchers have studied pomegranate's antioxidant properties to see if it might have any beneficial effects in the treatment of cancer. One study, described in the paper **"Pomegranate Fruit Juice for Chemoprevention and Chemotherapy of Prostate Cancer"** (*Proceedings of the National Academy of Sciences* [October 11, 2005]: 14813–14818), investigated whether pomegranate fruit extract (PFE) was effective in slowing the growth of prostate cancer tumors.

In this study, 24 mice were injected with cancer cells. The mice were then randomly assigned to one of three treatment groups. One group of eight mice received normal drinking water. The second group of eight mice received drinking water supplemented with 0.1% PFE and the third group received drinking water supplemented with 0.2% PFE. The average tumor volume for the mice in each group was recorded at several points in time.

The accompanying data on y = Average tumor volume (in mm^3) and x = Number of days after injection of cancer cells for the mice that received plain drinking water were approximated from a graph that appeared in the paper:

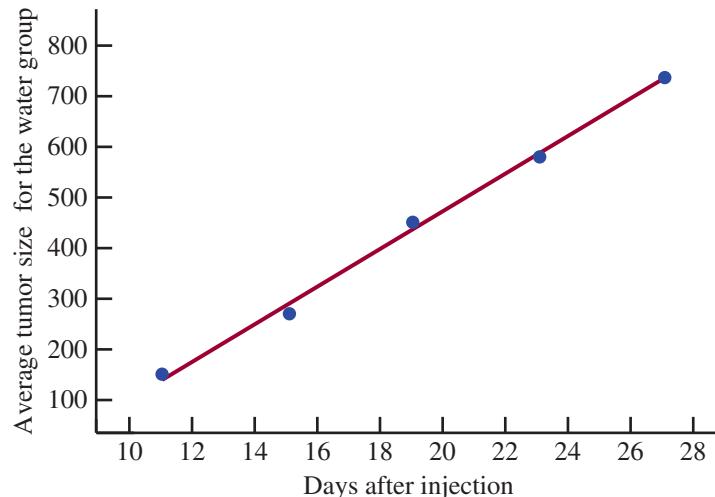
Consider the data ➤

x	11	15	19	23	27
y	150	270	450	580	740

A scatterplot of these data (Figure 5.9) shows that the relationship between number of days after injection of cancer cells and average tumor volume could reasonably be described by a line.

FIGURE 5.9

Scatterplot for the plain drinking water data of Example 5.4.



● Data set available online

Calculating the values of the slope and intercept of the least-squares line can be very tedious. This is where a statistics computer package or a graphing calculator can really be helpful. Minitab was used to find the equation of the least-squares line, and the resulting output is shown here:

Regression Analysis: y versus x

The regression equation is

$$y = -270 + 37.3x$$

Predictor	Coef	SE Coef	T	P
Constant	-269.75	23.42	-11.52	0.001
x	37.250	1.181	31.53	0.000

From the Minitab output, the equation of the least-squares line is

$$\hat{y} = -270 + 37.3x$$

We could also use the values given in the Coef column for the slope and intercept. These values have more decimal accuracy than those in the given equation.

The values of the slope and intercept of the least-squares line can also be calculated by hand. The summary quantities necessary to calculate these values are

$$\begin{aligned}\Sigma x &= 95 & \Sigma x^2 &= 1965 & \Sigma xy &= 47,570 \\ \Sigma y &= 2190 & \Sigma y^2 &= 1,181,900\end{aligned}$$

From these quantities,

Do the work ► $\bar{x} = 19$ $\bar{y} = 438$

$$b = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{47,570 - \frac{(95)(2190)}{5}}{1965 - \frac{(95)^2}{5}} = \frac{5960}{160} = 37.25$$

and

$$a = \bar{y} - b\bar{x} = 438 - (37.25)(19) = -269.75$$

The equation of the least-squares line is then

$$\hat{y} = -269.75 + 37.25x$$

This line is also shown on the scatterplot of Figure 5.9.

If we wanted to predict average tumor volume 20 days after injection of cancer cells, we could use the y value of the point on the least-squares line above $x = 20$:

$$\hat{y} = -269.75 + 37.25(20) = 475.25$$

Predicted average tumor volume for other numbers of days after injection of cancer cells could be calculated in a similar way.

Sometimes you need to be cautious when making predictions. For example, in the context of Example 5.4, the least-squares line should not be used to predict average tumor volume for times much outside the range 11 to 27 days (the range of x values in the data set) because we do not know whether the linear pattern observed in the scatterplot continues outside this range. This is sometimes referred to as the **danger of extrapolation**.

We can see that using the least-squares line to predict average tumor volume for fewer than 10 days after injection of cancer cells can lead to nonsensical predictions. For example, if the number of days after injection is five, the predicted average tumor volume is negative:

$$\hat{y} = -269.75 + 37.25(5) = -83.5$$

Because it is impossible for average tumor volume to be negative, this is a clear indication that the pattern observed for x values in the 11 to 27 range does not continue outside that range. However, the least-squares line can still be a useful tool for making predictions for x values within the 11- to 27-day range.

USE CAUTION—The Danger of Extrapolation

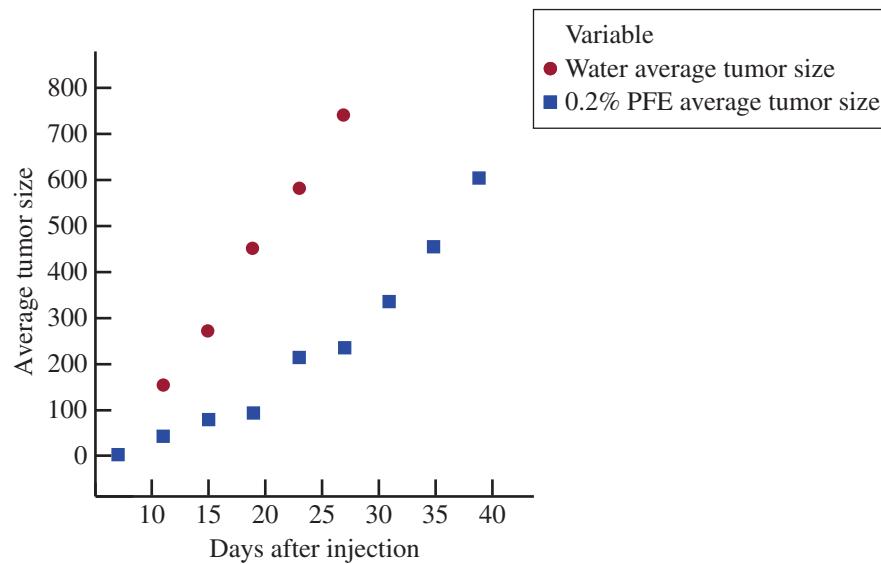
The least-squares line should not be used to make predictions outside the range of the x values in the data set because there is no evidence that the linear relationship continues outside this range.

Example 5.5 More on Pomegranate Juice and Tumor Growth

Figure 5.10 shows a scatterplot for average tumor volume versus number of days after injection of cancer cells for both the group of mice that drank only water and the group that drank water supplemented by 0.2% PFE. Notice that the tumor growth seems to be much slower for the mice that drank water supplemented with PFE. For the 0.2% PFE group, the relationship between average tumor volume and number of days after injection of cancer cells appears to be curved rather than linear. We will see in Section 5.4 how a curve (rather than a straight line) can be used to describe this relationship.

FIGURE 5.10

Scatterplot of average tumor volume versus number of days after injection of cancer cells for the water group and the 0.2% PFE group.



Example 5.6 Revisiting a Face You Can Trust

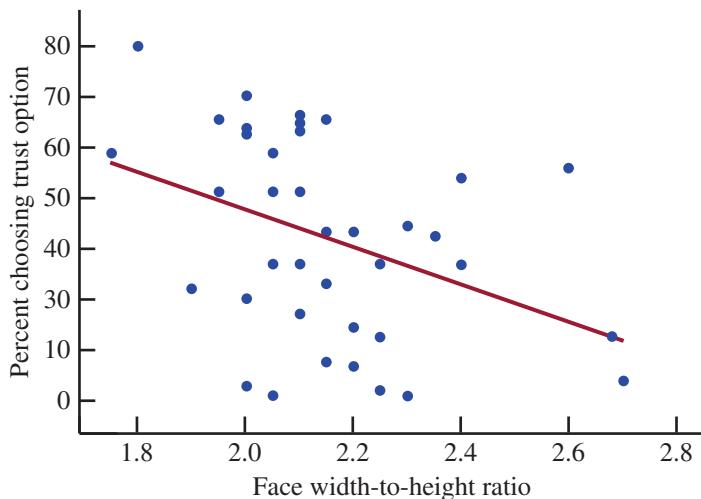
Understand the context ➤

Data on x = Face-to-height ratio and y = Percentage choosing the trust option in a game were given in Example 5.1. In that example, we saw that the correlation coefficient was -0.392 , indicating a weak negative linear relationship. This linear relationship can be summarized using the least-squares line, as shown in Figure 5.11.

Minitab was used to find the equation of the least-squares line, and Figure 5.12 shows part of the resulting output. Instead of x and y , the variable labels Face-to-width ratio and Percentage choosing trust option are used. The equation at the top is the equation of the least-squares line. In the rectangular table just below the equation, the first row gives information about the intercept, a , and the second row gives information about the slope, b . In particular, the coefficient column labeled “Coef” contains the values of a and b using more digits than in the rounded values that appear in the equation.

FIGURE 5.11

Scatterplot and least-squares line for the data of Example 5.6.

**FIGURE 5.12**

Partial Minitab output for Example 5.6.

The regression equation is					
Percentage choosing trust option = 123 - 38.5 Face width-to-height ratio					
Predictor	Coef	SE Coef	T	P	
Constant	123.11	32.58	3.78	0.001	
Face width-to-height ratio	-38.47	15.06	-2.55	0.015	

→ *Equation $\hat{y} = a + bx$*
 → *Value of a*
 → *Value of b*

The least-squares line should not be used to predict the percentage choosing the trust option for faces with width-to-height ratios such as $x = 1.00$ or $x = 4.00$. These x values are well outside the range of the data, and we do not know if the linear relationship continues outside the observed range.

Regression

The least-squares line is often called the **sample regression line**. This terminology comes from the relationship between the least-squares line and the correlation coefficient. To understand this relationship, we first need alternative expressions for the slope b and the equation of the line itself. With s_x and s_y denoting the sample standard deviations of the x 's and y 's, respectively, a bit of algebraic manipulation gives

$$b = r \left(\frac{s_y}{s_x} \right)$$

$$\hat{y} = \bar{y} + r \left(\frac{s_y}{s_x} \right) (x - \bar{x})$$

You do not need to use these formulas in any calculations, but several of their implications are important for appreciating what the least-squares line does.

- When $x = \bar{x}$ is substituted in the equation of the line, $\hat{y} = \bar{y}$ results. This means that the least-squares line passes through the *point of averages* (\bar{x}, \bar{y}) .
- Suppose for the moment that $r = 1$, so that all points lie exactly on the line whose equation is

$$\hat{y} = \bar{y} + \frac{s_y}{s_x} (x - \bar{x})$$

Now substitute $x = \bar{x} + s_x$, which is 1 standard deviation above \bar{x} :

$$\hat{y} = \bar{y} + \frac{s_y}{s_x}(\bar{x} + s_x - \bar{x}) = \bar{y} + s_y$$

We can see that with $r = 1$, when x is 1 standard deviation above its mean, we predict that the associated y value will be 1 standard deviation above its mean. Similarly, if $x = \bar{x} - 2s_x$ (2 standard deviations below its mean), then

$$\hat{y} = \bar{y} + \frac{s_y}{s_x}(\bar{x} - 2s_x - \bar{x}) = \bar{y} - 2s_y$$

which is also 2 standard deviations below the mean.

If $r = -1$, then $x = \bar{x} + s_x$ results in $\hat{y} = \bar{y} - s_y$, so the predicted y is also 1 standard deviation from its mean but on the opposite side of \bar{y} from where x is relative to \bar{x} .

In general, if x and y are perfectly correlated, the predicted y value associated with a given x value will be the same number of standard deviations (of y) from its mean \bar{y} as x is from its mean \bar{x} .

3. Now suppose that x and y are not perfectly correlated. For example, suppose $r = 0.5$, so the least-squares line has the equation

$$\hat{y} = \bar{y} + 0.5\left(\frac{s_y}{s_x}\right)(x - \bar{x})$$

Then substituting $x = \bar{x} + s_x$ gives

$$\hat{y} = \bar{y} + 0.5\left(\frac{s_y}{s_x}\right)(\bar{x} + s_x - \bar{x}) = \bar{y} + 0.5s_y$$

For $r = 0.5$, when x lies 1 standard deviation above its mean, we predict that y will be only 0.5 standard deviation above its mean. Similarly, we can predict y when r is negative. If $r = -0.5$, then the predicted y value will be only half the number of standard deviations from \bar{y} that x is from \bar{x} but x and the predicted y will now be on opposite sides of their respective means.

Consider using the least-squares line to predict the value of y associated with an x value some specified number of standard deviations away from \bar{x} . Then the predicted y value will be only r times this number of standard deviations from \bar{y} . In terms of standard deviations, except when $r = 1$ or -1 , the predicted y will always be closer to \bar{y} than x is to \bar{x} .

Using the least-squares line for prediction results in a predicted y that is pulled back in, or regressed, toward the mean of y compared to where x is relative to the mean of x . This regression effect was first noticed by Sir Francis Galton (1822–1911), a famous biologist, when he was studying the relationship between the heights of fathers and their sons. He found that predicted heights of sons whose fathers were above average in height were also above average (because r is positive here) but not by as much as the father's height. He found a similar relationship for fathers whose heights were below average. This regression effect has led to the term **regression analysis** for the collection of methods involving the fitting of lines, curves, and more complicated functions to bivariate and multivariate data.

The alternative form of the regression (least-squares) line emphasizes that predicting y from knowledge of x is not the same problem as predicting x from knowledge of y . The slope of the least-squares line for predicting x is $r(s_x/s_y)$ rather than $r(s_y/s_x)$ and the intercepts of the lines are also almost always different. For purposes of prediction, it makes a difference whether y is regressed on x , as we have done, or x is regressed on y .

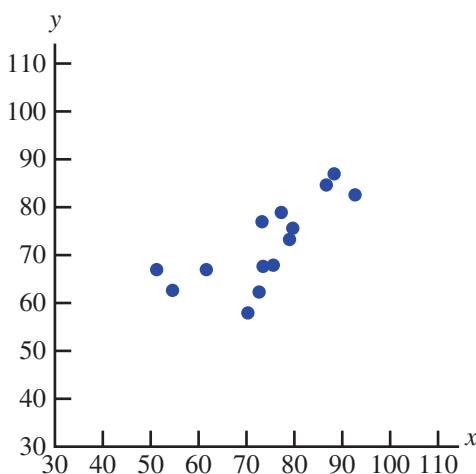
The least-squares line of y on x should not be used to predict x , because it is not the line that minimizes the sum of squared deviations in the x direction.

EXERCISES 5.17 - 5.34

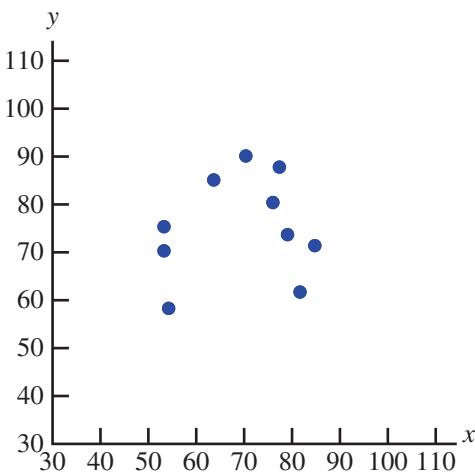
● Data set available online

- 5.17** Two scatterplots are shown below. Explain why it makes sense to use the least-squares line to summarize the relationship between x and y for one of these data sets but not the other. (Hint: See Example 5.5.)

Scatterplot 1:



Scatterplot 2:



- 5.18** The authors of the paper “Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement” (*International Journal of Nursing Studies* [2010]: 931–936) compared two different instruments for measuring a person’s ability to breathe out air. (This measurement is helpful in diagnosing various lung disorders.) The two instruments considered were a Wright peak flow meter and a mini-Wright peak flow meter. Seventeen subjects participated in the study, and for each person air flow was measured once using the Wright meter and once using the mini-Wright meter.

Subject	Mini-Wright Meter	Wright Meter	Subject	Mini-Wright Meter	Wright Meter
	Meter	Meter		Meter	Meter
1	512	494	10	445	433
2	430	395	11	432	417
3	520	516	12	626	656
4	428	434	13	260	267
5	500	476	14	477	478
6	600	557	15	259	178
7	364	413	16	350	423
8	380	442	17	451	427
9	658	650			

- a. Suppose that the Wright meter is considered to provide a better measure of air flow, but the mini-Wright meter is easier to transport and to use. If the two types of meters produce different readings but there is a strong relationship between the readings, it would be possible to use a reading from the mini-Wright meter to predict the reading that the larger Wright meter would have given. Use the given data to find an equation to predict Wright meter reading using a reading from the mini-Wright meter. (Hint: See Example 5.4.)
- b. What would you predict for the Wright meter reading for a person whose mini-Wright meter reading was 500?
- c. What would you predict for the Wright meter reading for a person whose mini-Wright meter reading was 300? (Hint: See the discussion of extrapolation that follows Example 5.4.)

- 5.19** The accompanying data are a subset of data from the report “Great Jobs, Great Lives” (*Gallup-Purdue Index 2015 Report*, gallup.com/reports/197144/gallup-purdue-index-report-2015.aspx). The values are approximate values read from a scatterplot. Students at a number of universities were asked if they agreed that their education was worth the cost. One variable in the table is the percentage of students at the university who responded *strongly agree*. The other variable in the table is the *U.S. News and World Report* ranking of the university.

Ranking	Percentage of Alumni Who Strongly Agree
28	53
29	58
30	62
37	55
45	54
47	62

(continued)

Ranking	Percentage of Alumni Who Strongly Agree
52	55
54	62
57	70
60	58
65	66
66	55
72	65
75	57
82	67
88	59
98	75

- a. Find the equation of the least-squares line that would allow you to predict the percentage of alumni who would strongly agree that their education was worth the cost, using ranking as the independent variable.
- b. Predict the percentage of alumni who would strongly agree that their education was worth the cost for a university with a ranking of 50.
- c. Explain why it would not be a good idea to use the least-squares line to predict the percentage of alumni who would strongly agree that their education was worth the cost for a university that had a ranking of 10.

5.20 The authors of the paper “[Evaluating Existing Movement Hypotheses in Linear Systems Using Larval Stream Salamanders](#)” (*Canadian Journal of Zoology* [2009]: 292–298) investigated whether water temperature was related to how far a salamander would swim and whether it would swim upstream or downstream. Data for 14 streams with different mean water temperatures where salamander larvae were released are given (approximated from a graph that appeared in the paper).

The two variables of interest are x = Mean water temperature ($^{\circ}\text{C}$) and y = Net directionality, which was defined as the difference in the relative frequency of the released salamander larvae moving upstream and the relative frequency of released salamander larvae moving downstream. A positive value of net directionality means a higher proportion were moving upstream than downstream. A negative value of net directionality means a higher proportion were moving downstream than upstream.

Mean Temperature (x)	Net Directionality (y)
11.99	0.03
12.50	-0.07
17.98	0.29
18.29	0.23
19.89	0.24
20.25	0.19
19.07	0.14
17.73	0.05
19.62	0.07

- a. Construct a scatterplot of the data. How would you describe the relationship between x and y ?
- b. Find the equation of the least-squares line describing the relationship between y = Net directionality and x = Mean water temperature.
- c. What value of net directionality would you predict for a stream that had mean water temperature of 15°C ?

5.21 The authors of the paper referenced in the previous exercise state that “when temperatures were warmer, more larvae were captured moving upstream, but when temperatures were cooler, more larvae were captured moving downstream.”

- a. Do the scatterplot and least-squares line from the previous exercise support this statement? Explain.
- b. Approximately what mean temperature would result in a prediction of the same number of salamander larvae moving upstream and downstream?

5.22 A sample of 548 ethnically diverse students from Massachusetts were followed over a 19-month period from 1995 and 1997 in a study of the relationship between TV viewing and eating habits (*Pediatrics* [2003]: 1321–1326). For each additional hour of television viewed per day, the number of fruit and vegetable servings per day was found to decrease on average by 0.14 serving.

- a. For this study, what is the dependent variable? What is the independent variable?
- b. Would the least-squares line for predicting number of servings of fruits and vegetables using number of hours spent watching TV as a predictor have a positive or negative slope? Explain.

5.23 The relationship between hospital patient-to-nurse ratio and various characteristics of job satisfaction and patient care has been the focus of a number of research studies. Suppose x = Patient-to-nurse ratio is the independent variable. For each of the following potential dependent variables, indicate whether you expect the slope of the least-squares line to be positive or negative and give a brief explanation for your choice.

Mean Temperature (x)	Net Directionality (y)
6.17	-0.08
8.06	0.25
8.62	-0.14
10.56	0.00
12.45	0.08

(continued)

- a. y = Measure of nurse's job satisfaction (higher values indicate higher satisfaction)
- b. y = Measure of patient satisfaction with hospital care (higher values indicate higher satisfaction)
- c. y = Measure of patient quality of care

- 5.24** The report "[Airline Quality Rating 2016](http://airlinequalityrating.com/reports/2016_AQR_Final.pdf)" (airlinequalityrating.com/reports/2016_AQR_Final.pdf) included the accompanying data on the on-time arrival percentage and the number of complaints filed per 100,000 passengers for U.S. airlines.

The report did not include data on the number of complaints for two of the airlines. Use the given data on the other airlines to fit the least-squares line and use it to predict the number of complaints per 100,000 passengers for Spirit and Virgin America Airlines.

Airline	On-Time Arrival Percentage	Complaints per 100,000 Passengers
Alaska	86	0.42
American	80	2.12
Delta	86	0.72
Envoy Air	74	1.59
Express Jet	78	1.01
Frontier	73	3.91
Hawaiian	88	0.89
JetBlue	76	1.17
SkyWest	80	0.84
Southwest	80	0.53
Spirit	69	Not reported
United	78	2.71
Virgin America	80	Not reported

- 5.25** Acrylamide is a chemical that is sometimes found in cooked starchy foods and which is thought to increase the risk of certain kinds of cancer. The paper "[A Statistical Regression Model for the Estimation of Acrylamide Concentrations in French Fries for Excess Lifetime Cancer Risk Assessment](#)" ([Food and Chemical Toxicology \[2012\]: 3867–3876](#)) describes a study to investigate the effect of frying time (in seconds) and acrylamide concentration (in micrograms per kilogram) in French fries. The data in the accompanying table are approximate values read from a graph that appeared in the paper.

Frying Time	Acrylamide Concentration
150	155
240	120
240	190
270	185
300	140
300	270

- a. If the goal is to learn how acrylamide concentration is related to frying time, which of these two variables is the dependent variable and which is the independent variable?

- b. Construct a scatterplot of these data. Describe any interesting features of the scatterplot.

- 5.26** Use the acrylamide data given in the previous exercise to answer the following questions.

- a. Find the equation of the least-squares line for predicting acrylamide concentration using frying time.
- b. Does the equation of the least-squares line support the conclusion that longer frying times tend to be paired with higher acrylamide concentrations? Explain.
- c. What is the predicted acrylamide concentration for a frying time of 225 seconds?
- d. Would you use the least-squares line to predict acrylamide concentration for a frying time of 500 seconds? If so, what is the predicted concentration? If not, explain why.

- 5.27** Studies have shown that people who suffer sudden cardiac arrest have a better chance of survival if a defibrillator shock is administered very soon after cardiac arrest. How is survival rate related to the time between when cardiac arrest occurs and when the defibrillator shock is delivered? This question is addressed in the paper "[Improving Survival from Sudden Cardiac Arrest: The Role of Home Defibrillators](#)" (by J. K. Stross, University of Michigan, February 2002; available at [heartstarthome.com](#)).

The accompanying data give y = Survival rate (percent) and x = Mean call-to-shock time (minutes) for a cardiac rehabilitation center (in which cardiac arrests occurred while victims were hospitalized and so the call-to-shock time tended to be short) and for four communities of different sizes:

Mean call-to-shock time, x	2	6	7	9	12
Survival rate, y	90	45	30	5	2

- a. Construct a scatterplot for these data. How would you describe the relationship between mean call-to-shock time and survival rate?
- b. Find the equation of the least-squares line.
- c. Use the least-squares line to predict survival rate for a community with a mean call-to-shock time of 10 minutes.

- 5.28** The data given in the previous exercise on x = Call-to-shock time (in minutes) and y = Survival rate (percent) were used to compute the equation of the least-squares line, which was

$$\hat{y} = 101.33 - 9.30x$$

The newspaper article “**FDA OKs Use of Home Defibrillators**” (*San Luis Obispo Tribune*, November 13, 2002) reported that “every minute spent waiting for paramedics to arrive with a defibrillator lowers the chance of survival by 10 percent.” Is this statement consistent with the given least-squares line? Explain.

- 5.29** An article on the cost of housing in California that appeared in the *San Luis Obispo Tribune* (March 30, 2001) included the following statement: “In Northern California, people from the San Francisco Bay area pushed into the Central Valley, benefiting from home prices that dropped on average \$4000 for every mile traveled east of the Bay area.” If this statement is correct, what is the slope of the least-squares line, $\hat{y} = a + bx$, where y = House price (in dollars) and x = Distance east of the Bay (in miles)? Explain.
- 5.30** The following data on sale price, size, and land-to-building ratio for 10 large industrial properties appeared in the paper “**Using Multiple Regression Analysis in Real Estate Appraisal**” (*Appraisal Journal* [2002]: 424–430):

Property	Sale Price (millions of dollars)	Size (thousands of sq. ft.)	Land-to- Building Ratio
1	10.6	2,166	2.0
2	2.6	751	3.5
3	30.5	2,422	3.6
4	1.8	224	4.7
5	20.0	3,917	1.7
6	8.0	2,866	2.3
7	10.0	1,698	3.1
8	6.7	1,046	4.8
9	5.8	1,108	7.6
10	4.5	405	17.2

- a. Calculate and interpret the value of the correlation coefficient between sale price and size.
 - b. Calculate and interpret the value of the correlation coefficient between sale price and land-to-building ratio.
 - c. If you wanted to predict sale price and you could use either size or land-to-building ratio as the basis for making predictions, which would you use? Explain.
 - d. Based on your choice in Part (c), find the equation of the least-squares line you would use for predicting y = sale price.
- 5.31** Explain why it can be dangerous to use the least-squares line to obtain predictions for x values that

are substantially larger or smaller than those contained in the sample.

- 5.32** The sales manager of a large company selected a random sample of $n = 10$ salespeople and determined for each one the values of x = Years of sales experience and y = Annual sales (in thousands of dollars). A scatterplot of the resulting (x, y) pairs showed a linear pattern.
- a. Suppose that the sample correlation coefficient is $r = 0.75$ and that the average annual sales is $\bar{y} = 100$. If a particular salesperson is 2 standard deviations above the mean in terms of experience, what would you predict for that person’s annual sales?
 - b. If a particular person whose sales experience is 1.5 standard deviations below the average experience is predicted to have an annual sales value that is 1 standard deviation below the average annual sales, what is the value of r ?
- 5.33** Explain why the slope b of the least-squares line always has the same sign (positive or negative) as the sample correlation coefficient r .
- 5.34** The California State Park System Statistical Report for the 2014/2015 Fiscal Year (parks.ca.gov/pages/795/files/14-15%20Statistical%20Report%20-%20INTERNET.pdf) gave the accompanying data on x = Amount of money collected in user fees (in thousands of dollars) and y = Operating cost (in thousands of dollars) for nine state parks in the North Coast Redwoods District.
- | User Fees (thousands of dollars) | Operating Costs (thousands of dollars) |
|----------------------------------|--|
| 17 | 100 |
| 74 | 359 |
| 811 | 3,582 |
| 380 | 1,377 |
| 33 | 241 |
| 427 | 760 |
| 500 | 1,044 |
| 734 | 2,205 |
| 760 | 1,620 |
- a. Construct a scatterplot of these data. Describe any interesting features of the scatterplot.
 - b. Find the equation of the least-squares line.
 - c. Is the slope of the least-squares line positive or negative? Is this consistent with your description of the scatterplot in Part (a)?
 - d. Based on the scatterplot in Part (a), do you think that the value of the correlation coefficient for this data set would be less than 0.5 or greater than 0.5? Explain.

SECTION 5.3 Assessing the Fit of a Line

Once we have found the equation of the least-squares line, the next step is to consider how effectively the line summarizes the relationship between x and y . Important questions to consider are

1. Is a line an appropriate way to summarize the relationship between the two variables?
2. Are there any unusual aspects of the data set that we need to consider before using the least-squares line to make predictions?
3. If we decide that it is reasonable to use the least-squares line as a basis for prediction, how accurate can we expect predictions based on the line to be?

In this section, we look at graphical and numerical methods that will allow us to answer these questions.

Most of these methods are based on the vertical deviations of the data points from the least-squares line. These vertical deviations are called residuals, and each represents the difference between an actual y value and the corresponding predicted value, \hat{y} , that would result from using the least-squares line to make a prediction.

Predicted Values and Residuals

The predicted value corresponding to the first observation in a data set is obtained by substituting that x value, x_1 , into the regression equation to obtain \hat{y}_1 , where

$$\hat{y}_1 = a + bx_1$$

The difference between the actual y value for the first observation, y_1 , and the corresponding predicted value is

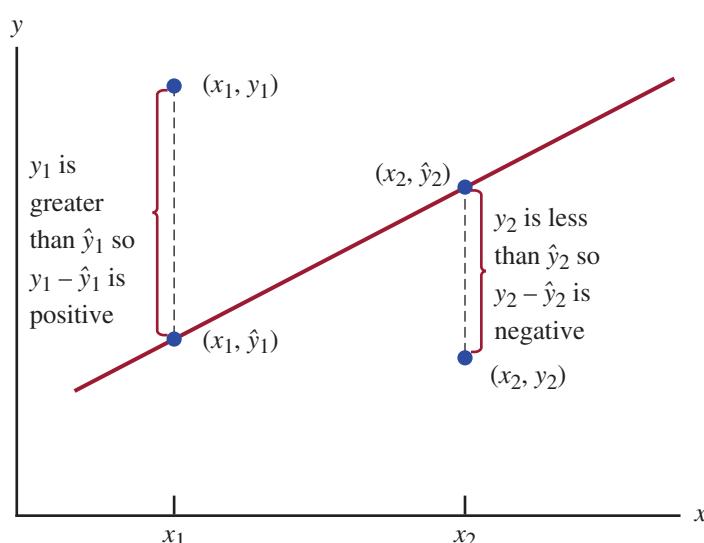
$$y_1 - \hat{y}_1$$

This difference is the **residual**, which is the vertical deviation of a point in the scatterplot from the least-squares line.

An observation falling above the line results in a positive residual, whereas a point falling below the line results in a negative residual. This is shown in Figure 5.13.

FIGURE 5.13

Positive and negative deviations from the least-squares line (residuals).



DEFINITIONS

Predicted values: The **predicted values** result from substituting each sample x value into the equation for the least-squares line. This gives

$$\begin{aligned}\hat{y}_1 &= \text{first predicted value} = a + bx_1 \\ \hat{y}_2 &= \text{second predicted value} = a + bx_2 \\ &\vdots \\ \hat{y}_n &= \text{nth predicted value} = a + bx_n\end{aligned}$$

Residuals: The **residuals** from the least-squares line are the n quantities

$$\begin{aligned}y_1 - \hat{y}_1 &= \text{first residual} \\ y_2 - \hat{y}_2 &= \text{second residual} \\ &\vdots \\ y_n - \hat{y}_n &= \text{nth residual}\end{aligned}$$

Each residual is the difference between an observed y value and the corresponding predicted y value.

Example 5.7 It May Be a Pile of Debris to You, but It Is Home to a Mouse

Understand the context ➤

- The accompanying data is a subset of data read from a scatterplot that appeared in the paper “**Small Mammal Responses to Fine Woody Debris and Forest Fuel Reduction in Southwest Oregon**” (*Journal of Wildlife Management* [2005]: 625–632). The authors of the paper were interested in how the distance a deer mouse will travel for food is related to the distance from the food to the nearest pile of fine woody debris that could provide a hiding place for the mouse. Distances were measured in meters. The data are given in Table 5.1.

TABLE 5.1 Predicted Values and Residuals for the Data of Example 5.7

Consider the data ➤

Distance from Debris (x)	Distance Traveled (y)	Predicted Distance Traveled (\hat{y})	Residual ($y - \hat{y}$)
6.94	0.00	14.76	-14.76
5.23	6.13	9.23	-3.10
5.21	11.29	9.16	2.13
7.10	14.35	15.28	-0.93
8.16	12.03	18.70	-6.67
5.50	22.72	10.10	12.62
9.19	20.11	22.04	-1.93
9.05	26.16	21.58	4.58
9.36	30.65	22.59	8.06

Minitab was used to find the equation of the least-squares line. Partial computer output follows:

Regression Analysis: Distance Traveled versus Distance to Debris

The regression equation is

Distance Traveled = -7.7 + 3.23 Distance to Debris

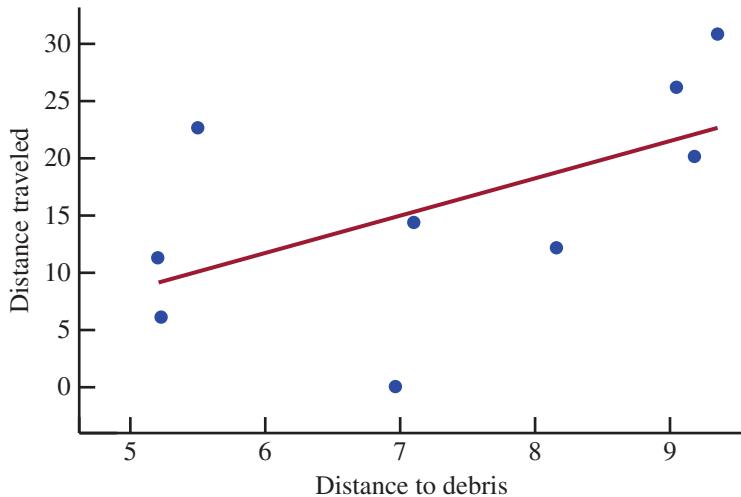
Predictor	Coef	SE Coef	T	P
Constant	-7.69	13.33	-0.58	0.582
Distance to Debris	3.234	1.782	1.82	0.112
S = 8.67071	R-Sq = 32.0%	R-Sq(adj) = 22.3%		

- Data set available online

The resulting least-squares line is $\hat{y} = -7.69 + 3.234x$.

A plot of the data that also includes the least-squares line is shown in Figure 5.14. The residuals for this data set are the signed vertical distances from the points to the line.

FIGURE 5.14
Scatterplot for the data of Example 5.7.



- Do the work ➤ For the mouse with the smallest x value (the third observation with $x_3 = 5.21$ and $y_3 = 11.29$), the corresponding predicted value and residual are

$$\text{predicted value} = \hat{y}_3 = -7.69 + 3.234(x_3) = -7.69 + 3.234(5.21) = 9.16$$

$$\text{residual} = y_3 - \hat{y}_3 = 11.29 - 9.16 = 2.13$$

The other predicted values and residuals are calculated in a similar manner and are included in Table 5.1.

Calculating the predicted values and residuals by hand can be tedious, but Minitab and other statistical software packages, as well as many graphing calculators, include them as part of the output, as shown in Figure 5.15. The predicted values and residuals can be found in the table at the bottom of the Minitab output in the columns labeled “Fit” and “Residual,” respectively.

The regression equation is

$$\text{Distance Traveled} = -7.7 + 3.23 \text{ Distance to Debris}$$

Predictor	Coeff	SE Coef	T	P
Constant	-7.69	13.33	-0.58	0.582
Distance to Debris	3.234	1.782	1.82	0.112

$$S = 8.67071 \quad R-\text{Sq} = 32.0\% \quad R-\text{Sq}(\text{adj}) = 22.3\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	247.68	247.68	3.29	0.112
Residual Error	7	526.27	75.18		
Total	8	773.95			

Obs	Distance to Debris	Distance Traveled	Fit	SE Fit	Residual	St Resid
1	6.94	0.00	14.76	2.96	-14.76	-1.81
2	5.23	6.13	9.23	4.69	-3.10	-0.42
3	5.21	11.29	9.16	4.72	2.13	0.29
4	7.10	14.35	15.28	2.91	-0.93	-0.11
5	8.16	12.03	18.70	3.27	-6.67	-0.83
6	5.50	22.72	10.10	4.32	12.62	1.68
7	9.19	20.11	22.04	4.43	-1.93	-0.26
8	9.05	26.16	21.58	4.25	4.58	0.61
9	9.36	30.65	22.59	4.67	8.06	1.10

Plotting the Residuals

A careful look at residuals can reveal many potential problems. A **residual plot** is a good place to start when assessing the appropriateness of the least-squares line.

DEFINITION

Residual plot: A scatterplot of the $(x, \text{residual})$ pairs.

Isolated points or a pattern of points in a residual plot indicate potential problems.

A desirable residual plot is one that exhibits no particular pattern, such as curvature. Curvature in the residual plot is an indication that the relationship between x and y is not linear and that a curve would be a better choice than a line for describing the relationship between x and y . Curvature is sometimes easier to see in a residual plot than in a scatterplot of y versus x , as illustrated in Example 5.8.

Example 5.8 Record Times

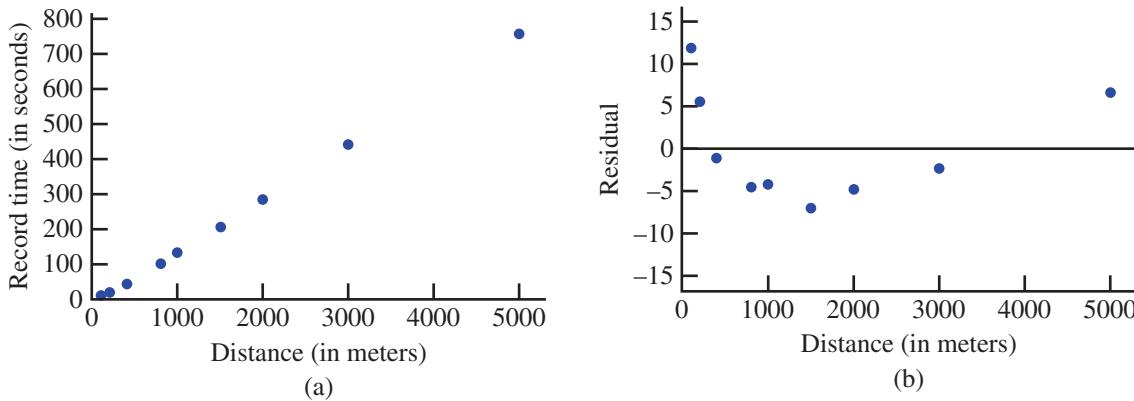
Understand the context ►

Consider the accompanying data on $x = \text{Distance}$ (in meters) and $y = \text{Record time}$ (in seconds) for men for races of various length in international track and field competitions (from *The World Almanac and Book of Facts 2016*).

Distance (in meters)	Record Time (in seconds)
100	10
200	19
400	43
800	101
1,000	132
1,500	206
2,000	285
3,000	441
5,000	757

The scatterplot, displayed in Figure 5.16(a), appears quite straight. However, even though the value of the correlation coefficient is very close to 1, when the residuals from the least-squares line are plotted (see Figure 5.16(b)), there is a definite curved pattern. Because of this, it is not accurate to say that men's record times increase linearly with distance.

FIGURE 5.16
Plots for the data of Example 5.8
(a) scatterplot;
(b) residual plot.



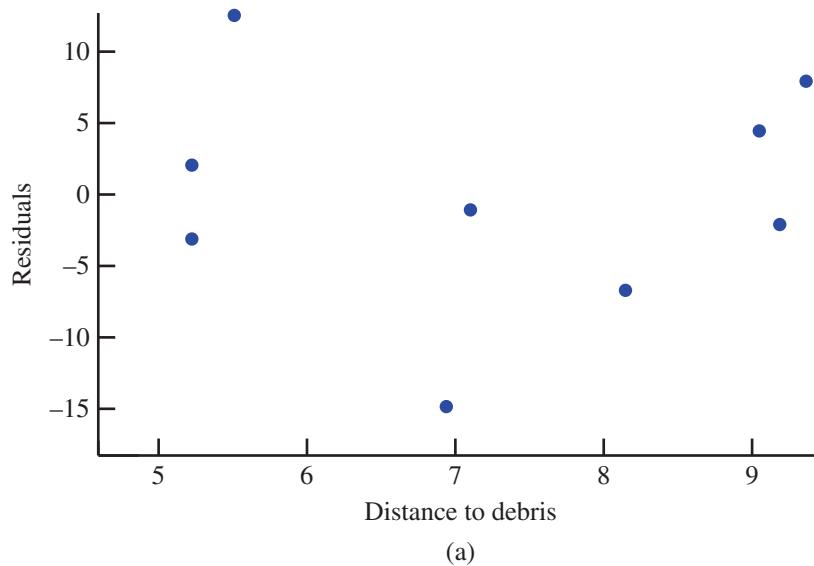
There is another common type of residual plot—one that plots the residuals versus the corresponding \hat{y} values rather than versus the x values. Because $\hat{y} = a + bx$ is simply a linear function of x , the only real difference between the two types of residual plots is the scale on the horizontal axis. The pattern of points in the residual plots will be the same, and it is this pattern of points that is important, not the scale. The two plots give equivalent information, as can be seen in Figure 5.17, which gives both plots for the data of Example 5.7.

It is also important to look for unusual values in the scatterplot or in the residual plot. A point falling far above or below the horizontal line at height 0 corresponds to a large residual, which may indicate some unusual circumstance such as a recording error, a non-standard experimental condition, or an atypical experimental subject.

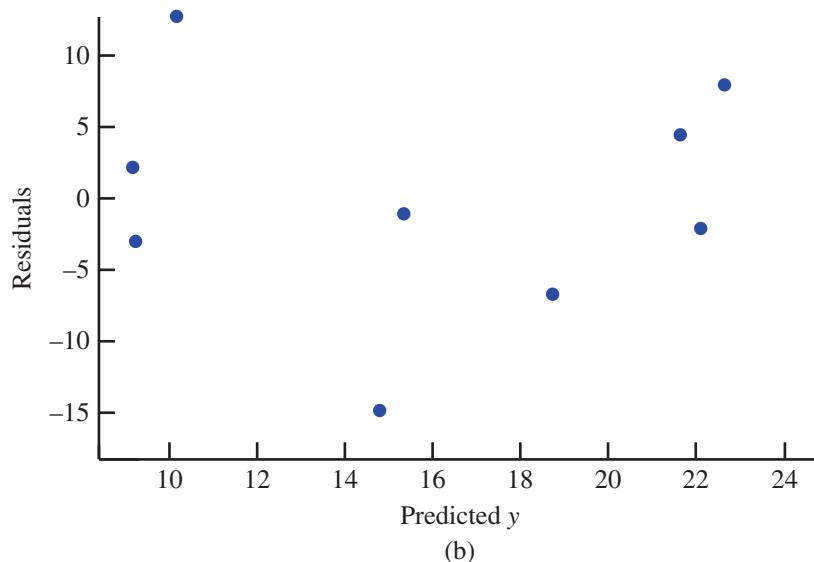
A point whose x value differs greatly from others in the data set may have exerted excessive influence in determining the fitted line. One method for assessing the impact of such an isolated point is to delete it from the data set, recompute the equation of the least-squares line, and evaluate the extent to which the equation of the line has changed.

FIGURE 5.17

Plots for the data of Example 5.7.
 (a) plot of residuals versus x ;
 (b) plot of residuals versus \hat{y} .



(a)



(b)

Example 5.9 | Older Than Your Average Bear

Understand the context ➤

- The accompanying data on $x = \text{Age}$ (in years) and $y = \text{Weight}$ (in kg) for 12 black bears appeared in the paper “**Habitat Selection by Black Bears in an Intensively Logged Boreal Forest**” (*Canadian Journal of Zoology* [2008]: 1307–1316).

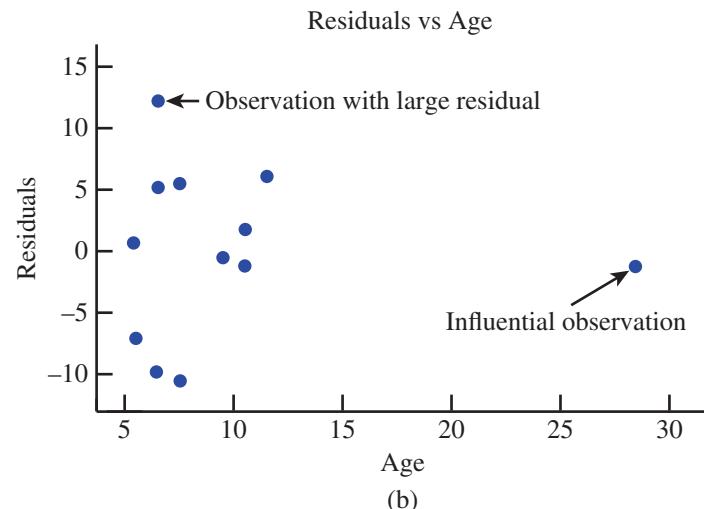
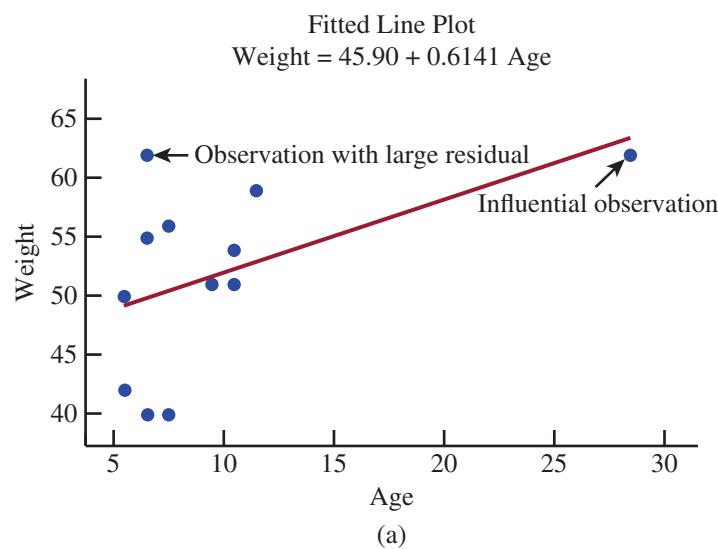
A scatterplot and residual plot are shown in Figures 5.18(a) and 5.18(b), respectively. One bear in the sample was much older than the other bears (bear 3 with an age of $x = 28.5$ years and a weight of $y = 62.00$ kg). This results in a point in the scatterplot that is far to the right of the other points in the scatterplot.

Bear	Age	Weight
1	10.5	54
2	6.5	40
3	28.5	62
4	10.5	51
5	6.5	55
6	7.5	56
7	6.5	62
8	5.5	42
9	7.5	40
10	11.5	59
11	9.5	51
12	5.5	50

FIGURE 5.18

Plots for the bear data of Example 5.9:

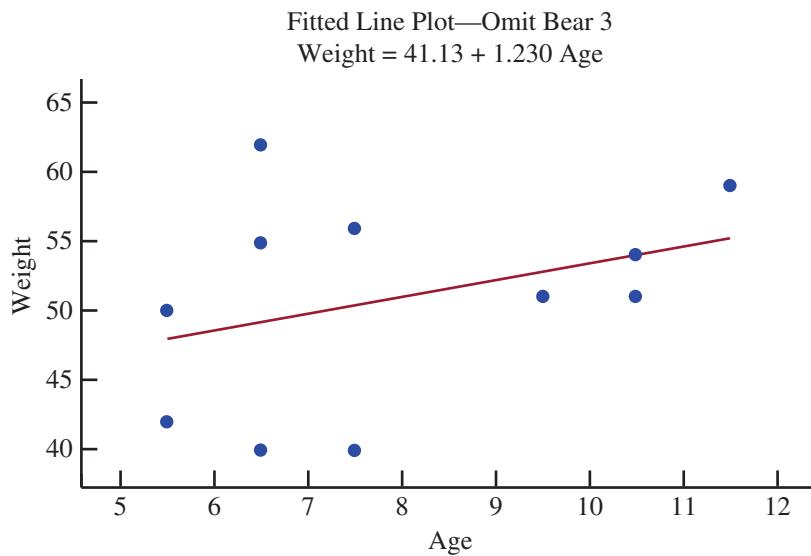
(a) scatterplot;
(b) residual plot.



Because the least-squares line minimizes the sum of squared residuals, the line is pulled toward this observation. This single observation plays a big role in determining the slope of the least-squares line, and it is therefore called an **influential observation**. Notice that this influential observation is not necessarily one with a large residual, because the least-squares line actually passes near this point. Figure 5.19 shows what happens when the influential observation is removed from the data set. Both the slope and intercept of the least-squares line are quite different from the slope and intercept of the line when this influential observation is included.

FIGURE 5.19

Scatterplot and least-squares line with bear 3 removed from data set.



Some points in a scatterplot may fall far from the least-squares line in the y direction, resulting in a large residual. These points are sometimes referred to as **outliers**. In Example 5.9, the observation with the largest residual is bear 7 with an age of $x = 6.5$ years and a weight of $y = 62.00$ kg. This observation is labeled in Figure 5.18. Even though this observation has a large residual, this observation is not very influential. The equation of the least-squares line for the data set consisting of all 12 observations is $\hat{y} = 45.90 + 0.6141x$, which is not much different from the equation that results when bear 7 is deleted from the data set ($\hat{y} = 43.81 + 0.7131x$).

DEFINITIONS

Unusual points in a bivariate data set are those that fall away from most of the other points in the scatterplot in either the x direction or the y direction.

Influential observation: An observation is potentially influential if it has an x value that is far away from the rest of the data (separated from the rest of the data in the x direction). To determine if the observation is influential, we assess whether removal of this observation has a large impact on the value of the slope or intercept of the least-squares line.

Outlier: An observation that has a large residual. Outlier observations fall far away from the least-squares line in the y direction.

Careful examination of a scatterplot and a residual plot can help us determine the appropriateness of a line for summarizing a relationship. If we decide that a line is appropriate, the next step is to think about assessing the accuracy of predictions based

on the least-squares line and deciding whether these predictions (based on the value of x) are better in general than those made without knowledge of the value of x . Two numerical measures that are helpful in this assessment are the *coefficient of determination* and the *standard deviation about the least-squares line*.

Coefficient of Determination, r^2

Suppose that we would like to predict the price of houses in a particular city. A random sample of 20 houses that are for sale is selected, and $y = \text{Price}$ and $x = \text{Size}$ (in square feet) are recorded for each house in the sample. There will be variability in house price (the houses will differ with respect to price), and it is this variability that makes accurate prediction of price a challenge. How much of the variability in house price can be explained by the fact that price is related to house size and that houses differ in size? If differences in size account for a large proportion of the variability in price, a price prediction that takes house size into account should be a big improvement over a prediction that is not based on size.

The **coefficient of determination** is a measure of the proportion of variability in the y variable that can be “explained” by a linear relationship between x and y .

DEFINITION

Coefficient of determination: The proportion of variability in y that can be attributed to an approximate linear relationship between x and y . The coefficient of determination is denoted by r^2 .

The value of r^2 is often converted to a percentage (by multiplying by 100) and interpreted as the percentage of variability in y that can be explained by an approximate linear relationship between x and y .

To understand how r^2 is calculated, we first consider variability in the y values. Variability in y can effectively be explained by an approximate straight-line relationship when the points in the scatterplot fall close to the least-squares line—that is, when the residuals are small in magnitude. A natural measure of variability about the least-squares line is the sum of the squared residuals. (Squaring before combining prevents negative and positive residuals from offsetting one another.) A second sum of squares assesses the total amount of variability in observed y values by considering how spread out the y values are from the mean y value.

DEFINITIONS

Total sum of squares: The **total sum of squares**, denoted by **SSTo**, is defined as

$$\text{SSTo} = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2 = \sum (y - \bar{y})^2$$

Residual sum of squares: The **residual sum of squares** (sometimes referred to as the error sum of squares), denoted by **SSResid**, is defined as

$$\text{SSResid} = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 = \sum (y - \hat{y})^2$$

These sums of squares can be found as part of the regression output from most standard statistical packages or can be obtained using the following formulas:

$$\text{SSTo} = \sum y^2 - \frac{\sum y^2}{n}$$

$$\text{SSResid} = \sum y^2 - a\sum y - b\sum xy$$

Example 5.10 Revisiting the Deer Mice Data

Figure 5.20 displays part of the Minitab output that results from fitting the least-squares line to the data on y = Distance traveled for food and x = Distance to nearest woody debris pile from Example 5.7. From the output,

$$\text{SSTo} = 773.95 \text{ and SSResid} = 526.27$$

Notice that SSResid is fairly large relative to SSTo.

FIGURE 5.20

Minitab output for the data of Example 5.10.

Regression Analysis: Distance Traveled versus Distance to Debris

The regression equation is

$$\text{Distance Traveled} = -7.7 + 3.23 \text{ Distance to Debris}$$

Predictor	Coeff	SE Coef	T	P
Constant	-7.69	13.33	-0.58	0.582
Distance to Debris	3.234	1.782	1.82	0.112
S = 8.67071 R-Sq = 32.0%			R-Sq(adj) = 22.3%	

Analysis of Variance

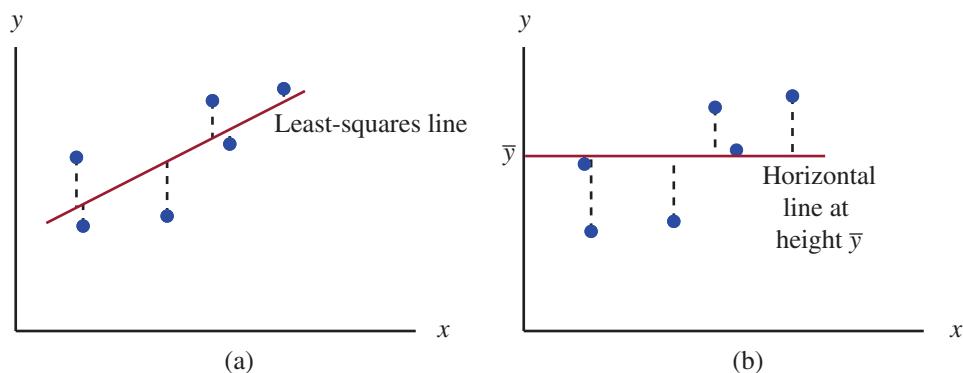
Source	DF	SS	MS	F	P
Regression	1	247.68	247.68	3.29	0.112
Residual Error	7	526.27	75.18		
Total	8	773.95			

$\xrightarrow{\text{SSTo}}$ $\xrightarrow{\text{SSResid}}$

The residual sum of squares is the sum of squared vertical deviations from the least-squares line. As Figure 5.21 illustrates, SSTo is also a sum of squared vertical deviations from a line—the horizontal line at height \bar{y} . The least-squares line is, by definition, the one having the smallest sum of squared deviations. This means that it is always the case that $\text{SSResid} \leq \text{SSTo}$. The two sums of squares are equal only when the least-squares line is the horizontal line.

FIGURE 5.21

Interpreting sums of squares:
 (a) SSResid = sum of squared vertical deviations from the least-squares line;
 (b) SSTo = sum of squared vertical deviations from the horizontal line at \bar{y} .



SSResid is sometimes referred to as a measure of unexplained variability—the amount of variability in y that cannot be attributed to the linear relationship between x and y . The more the points in the scatterplot deviate from the least-squares line, the larger the value of SSResid and the greater the amount of variability in y that cannot be explained by the approximate linear relationship. Similarly, SSTo is interpreted as a measure of total variability. The larger the value of SSTo, the greater the amount of variability in y_1, y_2, \dots, y_n .

The ratio SSResid/SSTo is the fraction or proportion of total variability that is not explained by a straight-line relationship. Subtracting this ratio from 1 results in the proportion of total variability that *is* explained:

DEFINITION

Coefficient of determination, r^2 : The coefficient of determination is defined as

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

Multiplying r^2 by 100 gives the percentage of variability in y that can be explained by the approximate linear relationship. The closer this percentage is to 100%, the more successful the linear relationship is in explaining variability in y .

Example 5.11 r^2 for the Deer Mice Data

Do the work ➤ For the data on distance traveled for food and distance to nearest debris pile from Example 5.10, we found SSTo = 773.95 and SSResid = 526.27. It follows that

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}} = 1 - \frac{526.27}{773.95} = 0.32$$

Interpret the results ➤ This means that only 32% of the observed variability in distance traveled for food can be explained by an approximate linear relationship between distance traveled for food and distance to nearest debris pile. Notice that the r^2 value can be found in the Minitab output of Figure 5.20, labeled “R-Sq.”

The symbol r was used in Section 5.1 to denote the sample correlation coefficient. It is not a coincidence that r^2 is used to represent the coefficient of determination. The notation suggests how these two quantities are related:

$$(\text{correlation coefficient})^2 = \text{coefficient of determination}$$

This means that if $r = 0.8$ or $r = -0.8$, then $r^2 = 0.64$, so 64% of the observed variability in the dependent variable can be explained by the linear relationship. Because the value of r does not depend on which variable is labeled x , the same is true of r^2 . The coefficient of determination is one of the few quantities computed in a regression analysis whose value remains the same when the roles of the dependent and independent variables are interchanged. When $r = 0.5$, we get $r^2 = 0.25$, so only 25% of the observed variability is explained by a linear relationship. This is why a value of r between -0.5 and 0.5 is not considered evidence of a strong linear relationship. Refer to Figure 5.4 on page 203 for guidance on assessing the strength of a linear relationship.

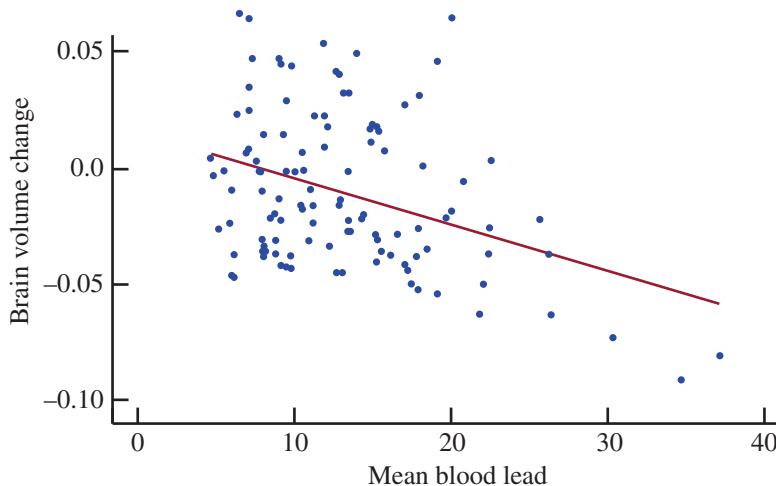
Example 5.12 Lead Exposure and Brain Volume

Understand the context ➤ The authors of the paper “Decreased Brain Volume in Adults with Childhood Lead Exposure” (*Public Library of Science Medicine* [May 27, 2008]: e112) studied the relationship between childhood environmental lead exposure and a measure of brain volume change in a particular region of the brain. Data on x = Mean childhood blood lead level ($\mu\text{g}/\text{dL}$) and y = Brain volume change (percent) read from a graph that appeared in the paper were used to produce the scatterplot in Figure 5.22. The least-squares line is also shown on the scatterplot.

Interpret the results ➤ Figure 5.23 displays part of the Minitab output that results from fitting the least-squares line to the data. Notice that although there is a slight tendency for smaller y values (corresponding to a brain volume decrease) to be paired with higher values of mean blood lead levels, the relationship is weak. The points in the plot are widely scattered around the least-squares line.

FIGURE 5.22

Scatterplot and least-squares line for the data of Example 5.12.

**FIGURE 5.23**

Minitab output for the data of Example 5.12.

Regression Analysis: Brain Volume Change versus Mean Blood Lead

The regression equation is

$$\text{Brain Volume Change} = 0.01559 - 0.001993 \text{ Mean Blood Lead}$$

S = 0.0310931 R-Sq = 13.6% R-Sq(adj) = 12.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.016941	0.0169410	17.52	0.000
Error	111	0.107313	0.0009668		
Total	112	0.124254			

From the computer output, we see that $100r^2 = 13.6\%$, so $r^2 = 0.136$. This means that differences in childhood mean blood lead level explain only 13.6% of the variability in adult brain volume change. Because the coefficient of determination is the square of the correlation coefficient, we can calculate the value of the correlation coefficient by taking the square root of r^2 . In this case, we know that the correlation coefficient will be negative (because there is a negative relationship between x and y), so we want the negative square root:

$$r = -\sqrt{0.136} = -0.369$$

Based on the values of the correlation coefficient and the coefficient of determination, we would conclude that there is a weak negative linear relationship and that childhood mean blood lead level explains only about 13.6% of the variability in adult change in brain volume.

Standard Deviation About the Least-Squares Line, s_e

The coefficient of determination measures the variability about the least-squares line *relative* to overall variability in y . A large value of r^2 does not by itself promise that the deviations from the line are small in an absolute sense. A typical observation could deviate from the line by quite a bit, but these deviations might still be small relative to overall variability in y .

Recall that in Chapter 4 the sample standard deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

was used as a measure of variability in a single sample. Roughly speaking, s is a typical amount that a sample observation deviates from the sample mean. There is a similar measure of variability that is used when a least-squares line is fit. This measure of variability is the standard deviation about the least-squares line.

DEFINITION

Standard deviation about the least-squares line: The standard deviation about the least-squares line is defined as

$$s_e = \sqrt{\frac{\text{SSResid}}{n - 2}}$$

Roughly speaking, s_e is a typical amount that an observation deviates from the least-squares line.

The standard deviation about the least-squares line is given as part of the usual regression output when statistical software or a graphing calculator is used. For the brain volume data of Example 5.12, the value of s_e can be found in the Minitab output of Figure 5.23 labeled as “S.” For the brain volume data, $s_e \approx 0.031$.

Example 5.13 Predicting Graduation Rates

Understand the context ➤

- Consider the accompanying data from 2014 on six-year graduation rate (%), instructional expenditure per full-time student (in dollars), and median SAT score for 9 primarily undergraduate public universities and colleges in the western United States with enrollments between 10,000 and 20,000 ([Source: College Results Online, The Education Trust](#)).

Consider the data ➤

	Graduation Rate	Instructional Expenditures	Median SAT
75.0	6,960	1,242	
71.5	7,274	1,114	
59.3	5,361	1,014	
56.4	5,374	1,070	
52.4	5,070	920	
48.0	5,226	888	
45.8	5,927	970	
42.7	5,600	937	
41.1	5,073	871	

Figure 5.24 displays scatterplots of graduation rate versus instructional expenditure and graduation rate versus median SAT score. The least-squares lines and the values of r^2 and s_e are also shown.

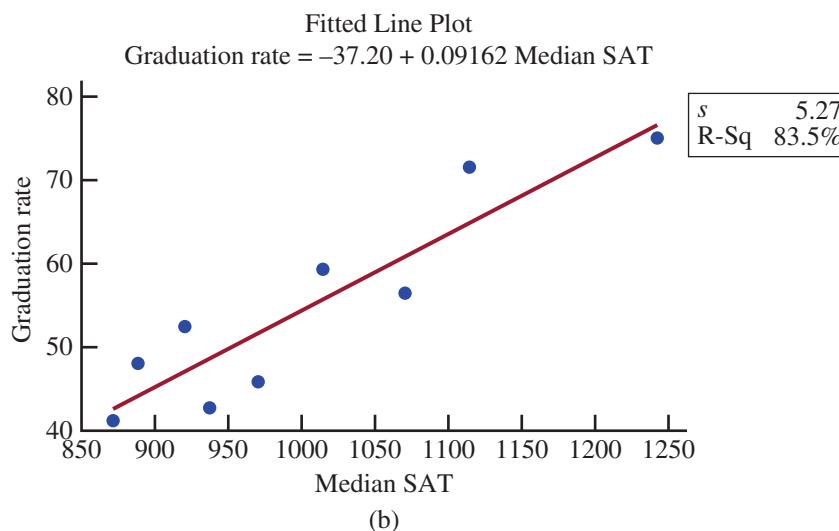
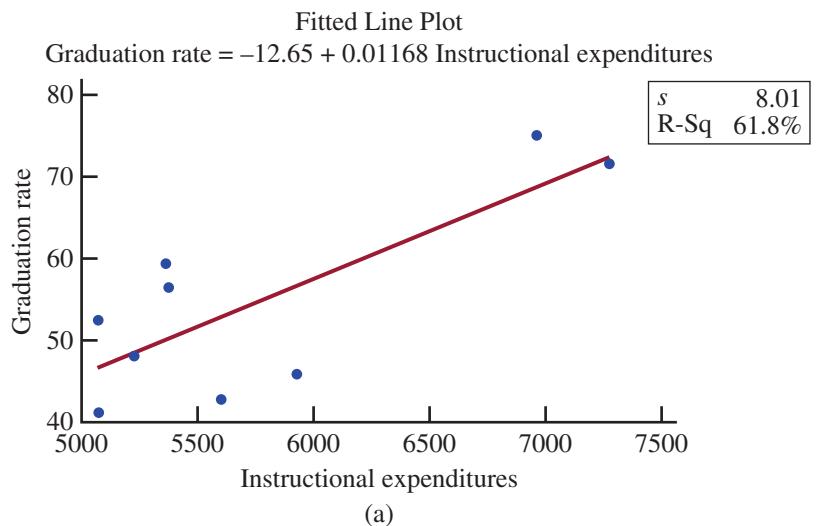
Interpret the results ➤

Notice that while there is a positive linear relationship between instructional expenditure and graduation rate, the relationship is not as strong as the relationship between median SAT score and graduation rate. The value of r^2 is 0.618 (61.8%), indicating that about 61.8% of the variability in graduation rate from university to university can be explained by differences in instructional expenditures. The standard deviation about the least-squares line is $s_e = 8.01$, which is larger than s_e for median SAT. This means that the points in the scatterplot of graduation rate versus instructional expenditure tend to fall farther from the least-squares line than is the case for the line that describes graduation rate versus median SAT.

The value of r^2 for graduation rate versus median SAT is 0.835 (83.5%) and $s_e = 5.27$, indicating that median SAT does a better job of explaining variability in graduation rates. The least-squares line that uses median SAT as a predictor would be expected to produce more accurate estimates of graduation rates than the line that uses instructional expenditure as a predictor.

FIGURE 5.24

Scatterplots for the data of Example 5.13:
 (a) graduation rate versus instructional expenditure;
 (b) graduation rate versus median SAT.



Based on the values of r^2 and s_e , median SAT would be a better choice for predicting graduation rates than instructional expenditures. It is also possible to develop a prediction equation that would incorporate both potential predictors—techniques for doing this are introduced in Chapter 14.

Also, take a second look at the scatterplot of Graduation rate versus Instructional expenditures. There are two schools that stand out in the scatterplot as potentially influential observations. The two points in the upper right-hand corner of the plot are far removed from the others in the x direction, indicating that these two universities had noticeably higher expenditures than the other seven universities. If these two data points are removed, the equation of the least-squares line changes dramatically. For this smaller data set, the equation of the least-squares line is

$$\text{Graduation rate} = 67.0 - 0.0033 \text{ Instructional expenditures}$$

and r^2 is only 2.12%. This is another reason that Median SAT would be preferred as a predictor of graduation rate.

When evaluating the usefulness of the least-squares line for making predictions, it is important to consider both the value of r^2 and the value of s_e . These two measures assess different aspects of the fit of the line. In general, we would like to have a small value for

s_e (which indicates that deviations from the line tend to be small) and a large value for r^2 (which indicates that the linear relationship explains a large proportion of the variability in the y values).

Interpreting the Values of s_e and r^2

A small value of s_e indicates that residuals tend to be small. Because a residual represents the difference between a predicted y value and an observed y value, the value of s_e tells us about the accuracy we can expect when using the least-squares line to make predictions.

A large value of r^2 indicates that a large proportion of the variability in y can be explained by the approximate linear relationship between x and y . This tells us that knowing the value of x is helpful for predicting y .

A useful least-squares line will have a reasonably small value of s_e and a reasonably large value of r^2 .

Now that we have considered all of the parts of a linear regression analysis, let's put all the parts together. The steps in a linear regression analysis are summarized in the accompanying box.

Steps in a Linear Regression Analysis

Given a bivariate numerical data set consisting of observations on a dependent variable y and an independent variable x :

- Step 1. Summarize the data graphically by constructing a scatterplot.
- Step 2. Based on the scatterplot, decide if it looks like the relationship between x and y is approximately linear. If so, proceed to the next step.
- Step 3. Find the equation of the least-squares line.
- Step 4. Construct a residual plot and look for any patterns or unusual features that may indicate that a line is not the best way to summarize the relationship between x and y . If none are found, proceed to the next step.
- Step 5. Compute the values of s_e and r^2 and interpret them in context.
- Step 6. Based on what you have learned from the residual plot and the values of s_e and r^2 , decide whether the least-squares line is useful for making predictions. If so, proceed to the last step.
- Step 7. Use the least-squares line to make predictions.

Let's return to the example from the chapter introduction and look at how following these steps can help us learn about the age of an unidentified crime victim.

Example 5.14 Revisiting Help for Crime Scene Investigators

Understand the context ➤

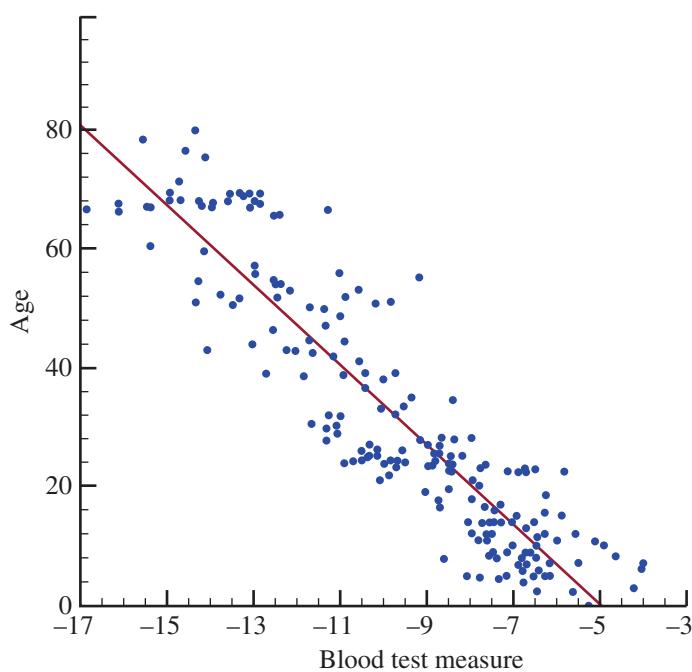
One of the tasks that forensic scientists face is estimating the age of an unidentified crime victim. Prior to 2010, this was usually done by analyzing teeth and bones, and the resulting estimates were not very reliable. In a groundbreaking study described in the paper “[Estimating Human Age from T-Cell DNA Rearrangements](#)” (*Current Biology* [2010]), scientists examined the relationship between age and a blood test measure. They recorded age and the blood test measure for 195 people ranging in age from a few weeks to 80 years.

Because the scientists were interested in predicting age using the blood test measure, the dependent variable is $y = \text{Age}$ and the independent variable is $x = \text{Blood test measure}$.

Step 1: The scientists first constructed a scatterplot of the data, which is shown in Figure 5.25.

FIGURE 5.25

Scatterplot of age versus blood test measure.



Step 2: Based on the scatterplot, it does appear that there is a reasonably strong negative linear relationship between age and the blood test measure. The scientists also reported that the correlation coefficient for this data set was $r = -0.92$, which is consistent with the strong negative linear pattern in the scatterplot.

Step 3: The scientists calculated the equation of the least-squares line to be

$$\hat{y} = -33.65 - 6.74x$$

Step 4: A residual plot constructed from these data showed a few observations with large residuals, but these observations were not far removed from the rest of the data in the x direction. These observations were not judged to be influential. Also, there were no unusual patterns in the residual plot that would suggest a nonlinear relationship between age and the blood test measure.

Step 5: The values of s_e and r^2 were

$$s_e = 8.9 \text{ and } r^2 = 0.835$$

This means that approximately 83.5% of the variability in age can be explained by the linear relationship between age and blood test measure. The value of s_e tells us that if the least-squares line is used to predict age from the blood test measure, a typical difference between the predicted age and the actual age would be about 9 years.

Step 6: Based on the residual plot, the large value of r^2 , and the relatively small value of s_e , the scientists proposed using the blood test measure and the least-squares line as a way to estimate ages of crime victims.

Step 7: To illustrate predicting age, suppose that a blood sample is taken from an unidentified crime victim and that the value of the blood test measure is determined to be -10 . The predicted age of the victim would be

$$\begin{aligned}\hat{y} &= -33.65 - 6.74(-10) \\ &= -33.65 - (-67.4) \\ &= 33.75 \text{ years}\end{aligned}$$

This is just an estimate of the actual age, and a typical prediction error is about 9 years.

EXERCISES 5.35 - 5.53

- 5.35** Does it pay to stay in school? The report *Trends in Higher Education* (The College Board, 2010) looked at the median hourly wage gain per additional year of schooling. The report states that workers with a high school diploma had a median hourly wage that was 10% higher than those who had only completed 11 years of school. Workers who had completed 1 year of college (13 years of education) had a median hourly wage that was 11% higher than that of the workers who had completed only 12 years of school. The added gain in median hourly wage for each additional year of school is shown in the accompanying table. The entry for 15 years of schooling has been intentionally omitted from the table.

Years of Schooling	Median Hourly Wage Gain for the Additional Year (percent)
12	10
13	11
14	13
16	16
17	18
18	19

- a.** Use the given data to predict the median hourly wage gain for the 15th year of schooling.
b. The actual wage gain for 15th year of schooling was 14%. How close was the actual value to the predicted wage gain percent from Part (a)?
- 5.36** The data in the accompanying table is from the paper “Six-Minute Walk Test in Children and Adolescents” (*The Journal of Pediatrics* [2007]: 395–399). Two hundred and eighty boys completed a test that measures the distance that the subject can walk on a flat, hard surface in 6 minutes. For each age group shown in the table, the median distance walked by the boys in that age group is also given.

Age Group	Representative Age (Midpoint of Age Group)	Median Six-minute Walk Distance (meters)
3–5	4.0	544.3
6–8	7.0	584.0
9–11	10.0	667.3
12–15	13.5	701.1
16–18	17.0	727.6

- a.** With x = Representative age and y = Median distance walked in 6 minutes, construct a scatterplot. Does the pattern in the scatterplot look linear?

● Data set available online

- b.** Find the equation of the least-squares line that describes the relationship between median distance walked in 6 minutes and representative age.
c. Calculate the five residuals and construct a residual plot. (Hint: See Examples 5.7 and 5.8.)
d. Are there any unusual features in the residual plot?

- 5.37** ● The paper referenced in the previous exercise also gave the 6-minute walk distances for 248 girls age 3 to 18 years. The median 6-minute walk times for girls for the five age groups were

492.4 578.3 655.8 657.6 660.9

- a.** With x = Representative age and y = Median distance walked in 6 minutes, construct a scatterplot.
b. How does the pattern in the scatterplot for girls differ from the pattern in the scatterplot for boys from the previous exercise?
c. Find the equation of the least-squares line that describes the relationship between median distance walked in 6 minutes and representative age for girls.
d. Calculate the five residuals and construct a residual plot.

- 5.38** Consider the residual plot from the previous exercise. The authors of the paper decided to use a curve rather than a straight line to describe the relationship between median distance walked in 6 minutes and age for girls. What aspect of the residual plot supports this decision? (Hint: See Example 5.8.)

- 5.39** The report “Airline Quality Rating 2016” (airlinequalityrating.com/reports/2016_AQR_Final.pdf, retrieved April 22, 2017) included the data for 13 U.S. airlines given in the table below.

Airline	On-Time Arrival Percentage	Airline Quality Rating
Alaska	86	5
American	80	10
Delta	86	3
Envoy Air	74	12
Express Jet	78	9
Frontier	73	11
Hawaiian	88	4
JetBlue	76	2
SkyWest	80	7
Southwest	80	6
Spirit	69	13
United	78	8
Virgin America	80	1

- a. With x = Airline quality rating and y = On-time arrival percentage, construct a scatterplot. Does the pattern in the scatterplot look linear?
- b. Find the equation of the least-squares line.
- c. Calculate the residuals and construct a residual plot. Are there any unusual features in the residual plot?
- 5.40** Acrylamide is a chemical that is sometimes found in cooked starchy foods and which is thought to increase the risk of certain kinds of cancer. The paper “[A Statistical Regression Model for the Estimation of Acrylamide Concentrations in French Fries for Excess Lifetime Cancer Risk Assessment](#)” (*Food and Chemical Toxicology* [2012]: 3867–3876) describes a study to investigate the effect of x = Frying time (in seconds) and y = Acrylamide concentration (in micrograms per kilogram) in French fries. The data in the accompanying table are approximate values read from a graph that appeared in the paper.
- | Frying Time | Acrylamide Concentration |
|-------------|--------------------------|
| 150 | 155 |
| 240 | 120 |
| 240 | 190 |
| 270 | 185 |
| 300 | 140 |
| 300 | 270 |
- a. Construct a scatterplot of these data.
- b. Find the equation of the least-squares line. Based on this line, what would you predict acrylamide concentration to be for a frying time of 270 seconds? What is the residual associated with the observation (270, 185)?
- 5.41** Consider the scatterplot of acrylamide concentration versus frying time from the previous exercise.
- a. Which observation is potentially influential? Explain the reason for your choice.
- b. When the potentially influential observation is deleted from the data set, the equation of the least-squares line using the remaining five observations is $\hat{y} = -44 + 0.83x$. Use this equation to predict the acrylamide concentration for a frying time of 270 seconds. How does this prediction compare to the prediction made in the previous exercise?
- 5.42** Some types of algae have the potential to cause damage to river ecosystems. The accompanying data on algae colony density (y) and rock surface area (x) for nine rivers is a subset of data that appeared in a scatterplot in a paper in the journal *Aquatic Ecology* (2010: 33–40).
- | x | 50 | 55 | 50 | 79 | 44 | 37 | 70 | 45 | 49 |
|-----|-----|----|----|----|----|-----|----|-----|----|
| y | 152 | 48 | 22 | 35 | 38 | 171 | 13 | 185 | 25 |
- a. Find the equation of the least-squares line.
- b. What is the value of r^2 for this data set? Write a sentence interpreting this value in context. (Hint: See Example 5.12.)
- c. What is the value of s_e for this data set? Write a sentence interpreting this value in context. (Hint: See Example 5.13.)
- d. Is the linear relationship between rock surface area and algae colony density positive or negative? Is it weak, moderate, or strong? Justify your answer.
- 5.43** ● The relationship between x = Total number of salmon in a creek and y = Percentage of salmon killed by bears that were transported away from the stream prior to the bear eating the salmon was examined in the paper “[Transportation of Pacific Salmon Carcasses from Streams to Riparian Forests by Bears](#)” (*Canadian Journal of Zoology* [2009]: 195–203). Data for the 10 years from 1999 to 2008 is given in the accompanying table.
- | Total Number | Percentage Transported |
|--------------|------------------------|
| 19,504 | 77.8 |
| 3,460 | 28.7 |
| 1,976 | 28.9 |
| 8,439 | 27.9 |
| 11,142 | 55.3 |
| 3,467 | 20.4 |
| 3,928 | 46.8 |
| 20,440 | 76.3 |
| 7,850 | 40.3 |
| 4,134 | 24.1 |
- a. Construct a scatterplot of the data.
- b. Does there appear to be a relationship between the total number of salmon in the stream and the percentage of salmon killed by bears that are transported away from the stream?
- c. Find the equation of the least-squares line. Draw the least-squares line on the scatterplot from Part (a).
- 5.44** The residuals from the least-squares line for the data given in the previous exercise are shown in the accompanying table.
- | Total Number | Percent Transported | Residual |
|--------------|---------------------|----------|
| 19,504 | 77.8 | 3.43 |
| 3,460 | 28.7 | 0.30 |
| 1,976 | 28.9 | 4.76 |
| 8,439 | 27.9 | -14.76 |
| 11,142 | 55.3 | 4.89 |
| 3,467 | 20.4 | -8.02 |
| 3,928 | 46.8 | 17.06 |
| 20,440 | 76.3 | -0.75 |
| 7,850 | 40.3 | -0.68 |
| 4,134 | 24.1 | -6.23 |

- The observation (3928, 46.8) has a large residual. Is this data point also an influential observation?
- The two points with unusually large x values (19,504 and 20,440) were not thought to be influential observations even though they are far removed in the x direction from the rest of the points in the scatterplot. Explain why these two points are not influential.
- Partial Minitab output resulting from fitting the least-squares line is shown here. What is the value of s_e ? Write a sentence interpreting this value.

Regression Analysis: Percent Transported versus Total Number

The regression equation is

$$\text{Percent Transported} = 18.5 + 0.00287 \text{ Total Number}$$

Predictor	Coef	SE Coef	T	P
Constant	18.483	4.813	3.84	0.005
Total Number	0.0028655	0.0004557	6.29	0.000
$S = 9.16217$	$R-\text{Sq} = 83.2\%$	$R-\text{Sq}(\text{adj}) = 81.1\%$		

- What is the value of r^2 for this data set (see Minitab output in Part (c))? Is the value of r^2 large or small? Write a sentence interpreting the value of r^2 .

- 5.45** The first Batman movie was made over 50 years ago in 1966. Over the years, Batman has been played on screen by a number of actors and even by a Lego figure in the Lego Batman movies. In the original comic books, Batman was 188 centimeters tall (about 6'2") and weighed 95 kilograms (about 210 pounds). The article “[50 Years of Batman on Film: How Has His Physique Changed?](#)” ([economist.com, March 28, 2016, retrieved April 22, 2017](#)) included the heights and weights of all the onscreen Batmen in the table below.

Batman	Height (cm)	Weight (kg)
Comic book	188	95
Lego Batman	4	4
Adam West	188	91
Michael Keaton	178	72
Val Kilmer	183	93
George Clooney	180	78
Christian Bale	183	82
Ben Affleck	193	98

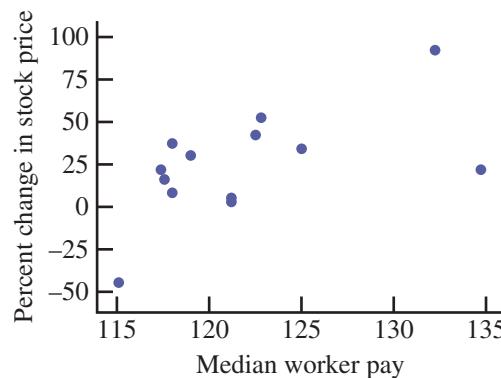
With x = Height and y = Weight, the equation of the least-squares line is $\hat{y} = 1.052 + 0.466x$.

- Calculate the residuals.
- Construct a residual plot. Are there any unusual features of the residual plot?
- The observation for Lego Batman (4, 4) is far removed from the other values in the data set (not surprising!). Is this observation influential in

determining either the slope of the least-squares line, or the intercept, or both? Justify your answer.

- 5.46** The article “[\\$115K! The 13 Best Paying U.S. Companies](#)” ([USA TODAY, August 11, 2015](#)) gave the following data on median worker pay (in thousands of dollars) and the 1-year percent change in stock price for the 13 highest paying companies in the United States. A scatterplot of these data is also shown.

Company	Median Worker Pay	Percent Change in Stock Price
Jupiter Networks	134.7	21.7
Netflix	132.2	92.2
Equinix	125.0	33.9
Altera	122.8	52.2
Visa	122.5	41.9
Yahoo	121.2	2.8
Xilinx	121.2	5.2
VeriSign	119.0	30.1
Microsoft	118.0	8.0
Broadcom	118.0	37.3
F5 Networks	117.6	15.9
Adobe Systems	117.4	22.0
eBay	115.1	-44.7



The equation of the least-squares line with y = Percent change in stock price and x = Median worker pay is $\hat{y} = -358 + 3.14x$.

- The observations for Jupiter Networks and Netflix are both far removed from the other points in the scatterplot in the x direction. Which of these observations would have the greatest impact on the equation of the least-squares line if it were to be omitted from the data set?
- Explain the difference between an influential observation and an outlier in a bivariate data set.

- 5.47** The article “[Examined Life: What Stanley H. Kaplan Taught Us About the SAT](#)” ([The New Yorker \[December 17, 2001\]: 86–92](#)) included a summary of findings regarding the use of SAT I scores, SAT II scores, and high school grade point average

(GPA) to predict first-year college GPA. The article states that “among these, SAT II scores are the best predictor, explaining 16 percent of the variance in first-year college grades. GPA was second at 15.4 percent, and SAT I was last at 13.3 percent.”

- If the data from this study were used to fit the least-squares line with $y = \text{First-year college GPA}$ and $x = \text{High school GPA}$, what would be the value of r^2 ?
- The article stated that SAT II was the best predictor of first-year college grades. Do you think that predictions based on a least-squares line with $y = \text{First-year college GPA}$ and $x = \text{SAT II score}$ would be very accurate? Explain why or why not.

- 5.48** The accompanying data are a subset of data from the report “Great Jobs, Great Lives” (Gallup-Purdue Index 2015 Report, gallup.com/reports/197144/gallup-purdue-index-report-2015.aspx, retrieved April 22, 2017). The values are approximate values read from a scatterplot. Students at a number of universities were asked if they agreed that their education was worth the cost. One variable in the table is the percentage of students at the university who responded *strongly agree*. The other variable in the table is the *U.S. News and World Report* ranking of the university.

Ranking	Percentage of Alumni Who Strongly Agree
28	53
29	58
30	62
37	55
45	54
47	62
52	55
54	62
57	70
60	58
65	66
66	55
72	65
75	57
82	67
88	59
98	75

- What is the value of r^2 for this data set? Write a sentence interpreting this value in context.
- What is the value of s_e for this data set? Write a sentence interpreting this value in context.
- Is the linear relationship between percentage of alumni who think their education was worth the cost and university ranking positive or negative? Is it weak, moderate, or strong? Justify your answer.

- 5.49** ● The article “California State Parks Closure List Due Soon” (*The Sacramento Bee*, August 30, 2009) gave the following data on $y = \text{Number of employees in fiscal year 2007–2008}$ and $x = \text{Total size of parks (in acres)}$ for the 20 state park districts in California:

Number of Employees, y	Total Park Size, x
95	39,334
95	324
102	17,315
69	8,244
67	620,231
77	43,501
81	8,625
116	31,572
51	14,276
36	21,094
96	103,289
71	130,023
76	16,068
112	3,286
43	24,089
87	6,309
131	14,502
138	62,595
80	23,666
52	35,833

- Construct a scatterplot of the data.
- Find the equation of the least-squares line.
- Do you think the least-squares line gives accurate predictions? Explain.
- Delete the observation with the largest x value from the data set and recalculate the equation of the least-squares line. Does this observation greatly affect the equation of the line?

- 5.50** ● The article referenced in the previous exercise also gave data on the percentage of operating costs covered by park revenues for the 2007–2008 fiscal year.

Number of Employees, x	Percent of Operating Cost Covered by Park Revenues, y
95	37
95	19
102	32
69	80
67	17
77	34
81	36
116	32
51	38
36	40
96	53
71	31

(continued)

Number of Employees, x	Percent of Operating Cost Covered by Park Revenues, y
76	35
112	108
43	34
87	97
131	62
138	36
80	36
52	34

- a.** Find the equation of the least-squares line relating $y = \text{Percent of operating costs covered by park revenues}$ and $x = \text{Number of employees}$.
- b.** Based on the values of r^2 and s_e , do you think that the least-squares line does a good job of describing the relationship between $y = \text{Percent of operating costs covered by park revenues}$ and $x = \text{Number of employees}$? Explain.
- c.** The graph at the bottom of the page is a scatterplot of $y = \text{Percent of operating costs covered by park revenues}$ and $x = \text{Number of employees}$. The least-squares line is also shown. Which observations are outliers? Do the observations with the largest residuals correspond to the park districts with the largest number of employees?
- 5.51** A study was carried out to investigate the relationship between the hardness of molded plastic (y , in Brinell units) and the amount of time elapsed since the plastic was molded (x , in hours). Summary quantities include $n = 15$, $\text{SSResid} = 1235.470$, and $\text{SSTo} = 25,321.368$. Calculate and interpret the value of the coefficient of determination.
- 5.52** Both r^2 and s_e are used to assess the fit of a line.
- a.** Is it possible that both r^2 and s_e could be large for a bivariate data set? Explain. (A picture might be helpful.)

- b.** Is it possible that a bivariate data set could yield values of r^2 and s_e that are both small? Explain. (Again, a picture might be helpful.)
- c.** Explain why it is desirable to have r^2 large and s_e small if the relationship between two variables x and y is to be described using a straight line.

5.53 With a bit of algebra, it can be shown that

$$\text{SSResid} = (1 - r^2) \sum (y - \bar{y})^2$$

from which it follows that

$$s_e = \sqrt{\frac{n-1}{n-2}} \sqrt{1 - r^2} s_y$$

Unless n is quite small, $(n-1)/(n-2) \approx 1$, so

$$s_e \approx \sqrt{1 - r^2} s_y$$

- a.** For what value of r is s_e as large as s_y ? What is the least-squares line in this case?
- b.** For what values of r will s_e be much smaller than s_y ?
- c.** A study by the Berkeley Institute of Human Development reported the following summary data for a sample of $n = 66$ California boys:

$$r \approx 0.80$$

At age 6, average height ≈ 46 inches, standard deviation ≈ 1.7 inches.

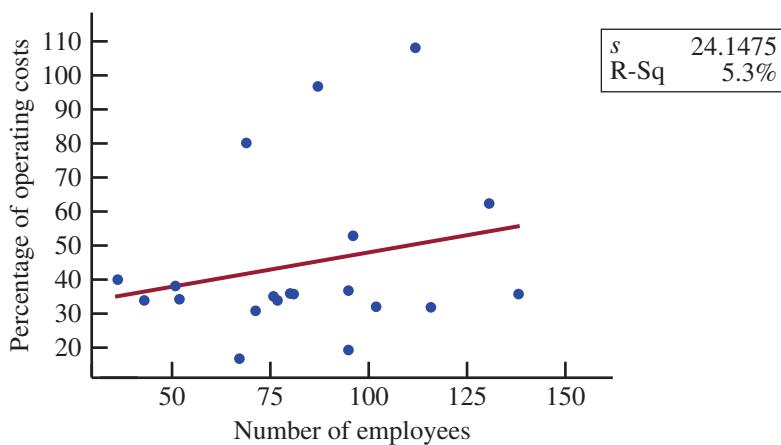
At age 18, average height ≈ 70 inches, standard deviation ≈ 2.5 inches.

What would s_e be for the least-squares line used to predict 18-year-old height from 6-year-old height?

- d.** Referring to Part (c), suppose that you wanted to predict the past value of 6-year-old height from knowledge of 18-year-old height. Find the equation for the appropriate least-squares line. What is the corresponding value of s_e ?

Figure for Exercise 5.50

Fitted Line Plot
Percentage of operating costs = $27.71 + 0.2011$ number of employees



SECTION 5.4 Nonlinear Relationships and Transformations

When the points in a scatterplot exhibit a linear pattern and the residual plot does not reveal any problems with the linear fit, the least-squares line is a sensible way to summarize the relationship between two numerical variables. But what should we do if a scatterplot or residual plot exhibits a curved pattern, indicating a more complicated relationship between x and y ? When this happens, a nonlinear function can be used to summarize the relationship between x and y . The process is similar to what is done for linear relationships, but now a curve is used instead of a line. After choosing an appropriate curve, a residual plot can be used to assess whether the curve is an appropriate way to describe the relationship. If we decide the curve is a useful summary of the relationship between x and y , we can then use the curve to make predictions.

Choosing a Nonlinear Function to Describe a Relationship

Once we have decided to use a nonlinear function to describe the relationship between two variables x and y , we need to decide what type of function should be considered. Some common functions that are used to model relationships are shown in Figure 5.26.

Example 5.15 Fishing Bears

Understand the context ➤

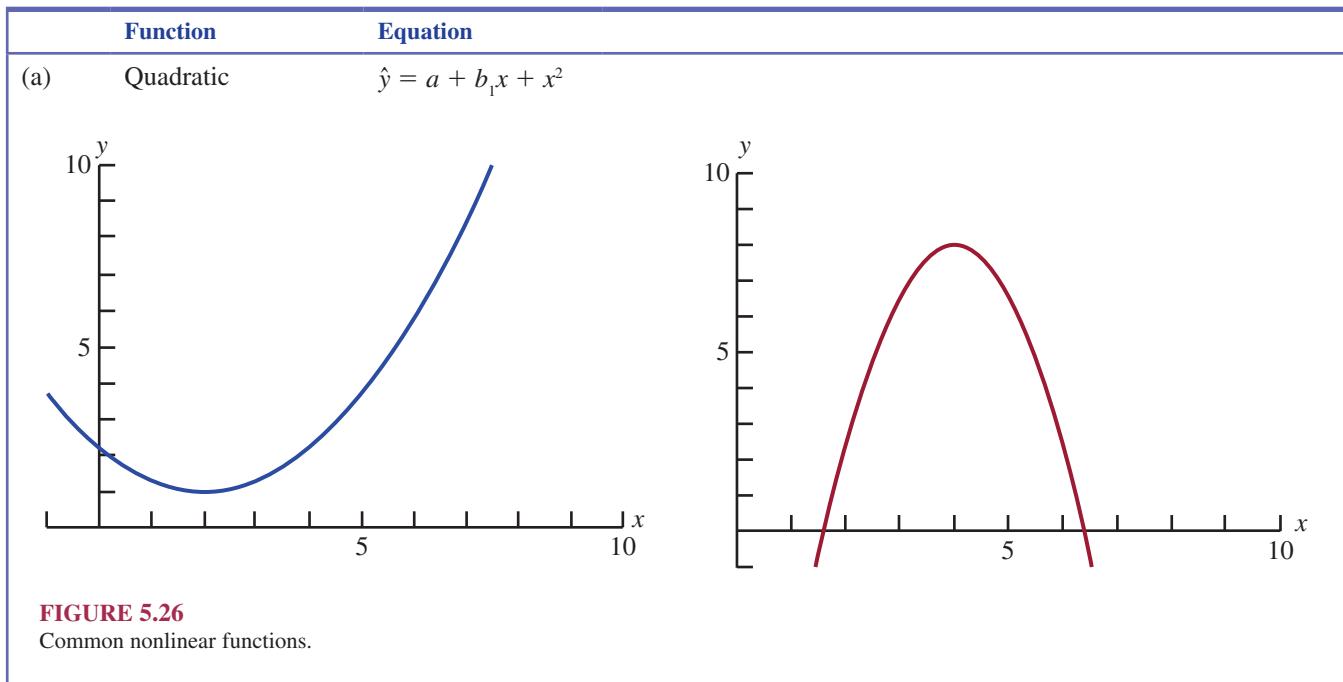
The article “Quantifying Spatiotemporal Overlap of Alaskan Brown Bears and People” (*Journal of Wildlife Management* [2005]: 810–817) describes a study of the effect of human activity on bears. The researchers wondered if sport fishing and boating might be limiting bears’ access to salmon. They collected data on

x = Number of days from the beginning of June (June 1st = 1)

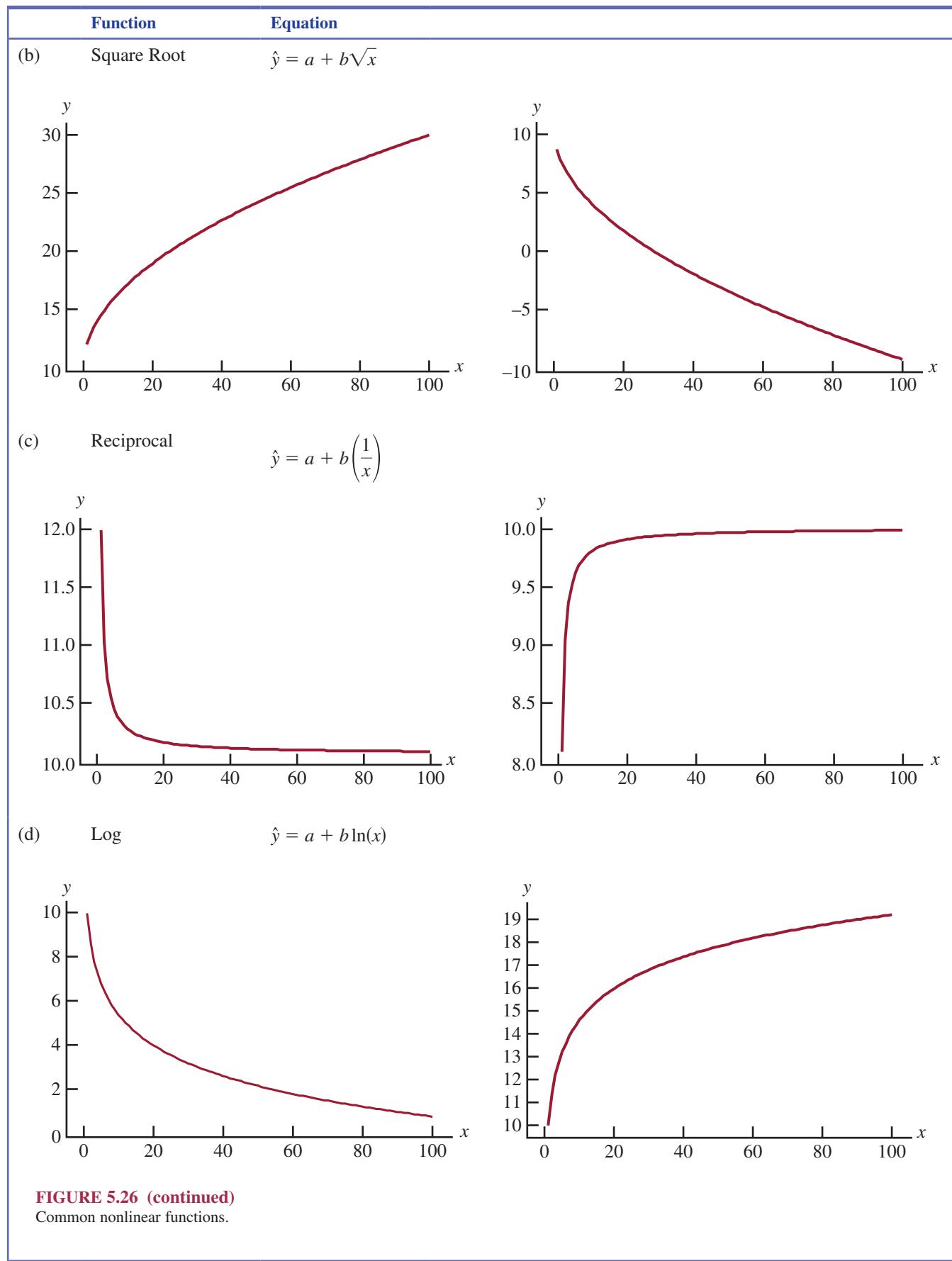
and

y = Total fishing time (in bear-hours) for the day at a particular location in Alaska

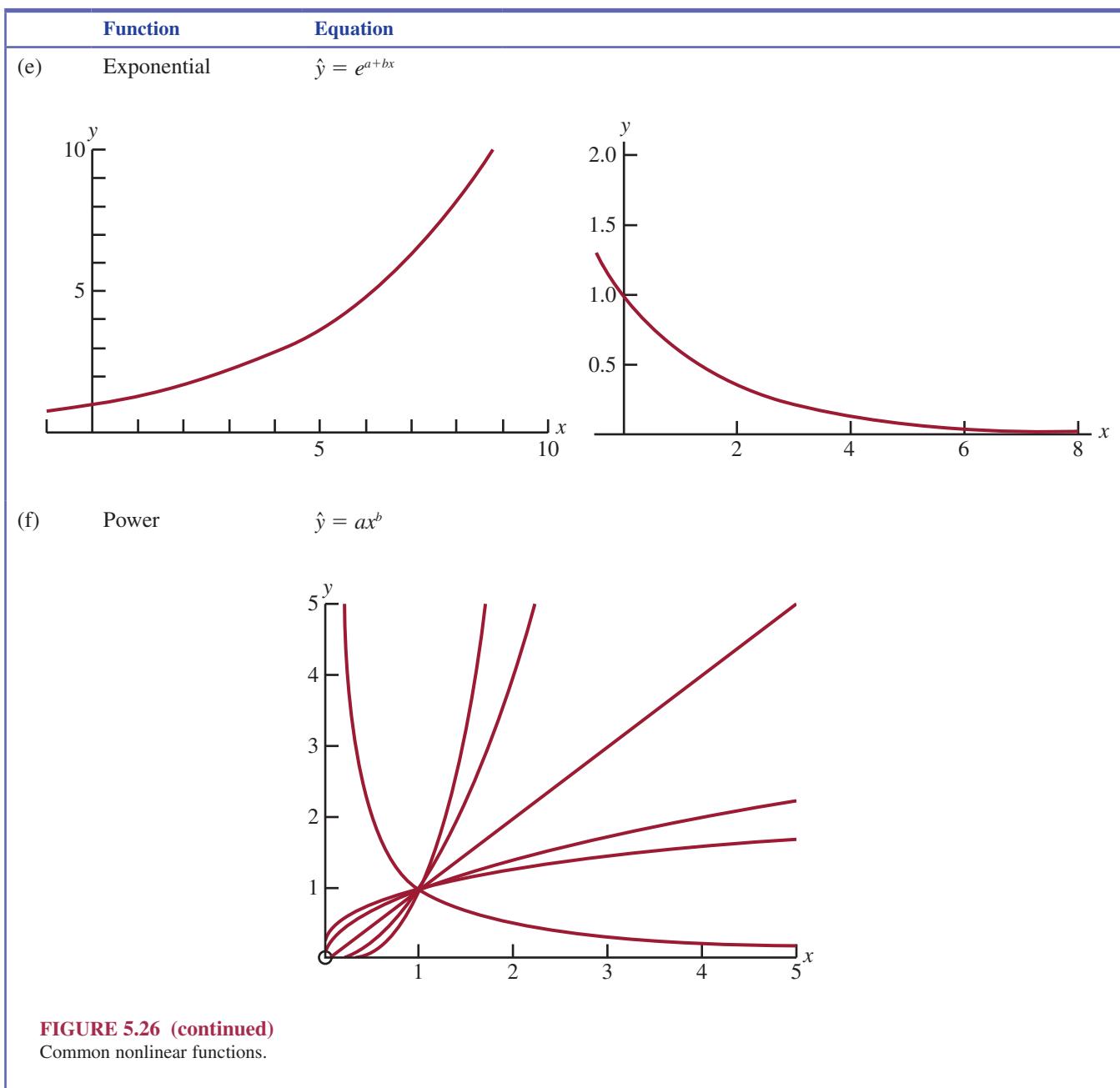
For example, the data pair (33, 45.7) corresponds to the day that is 32 days after June 1 (which is July 2). On this day, the total fishing time for bears was 45.7 bear-hours. This is the sum of the number of hours spent fishing for all bears fishing at this location on July 2.



(continued)



(continued)



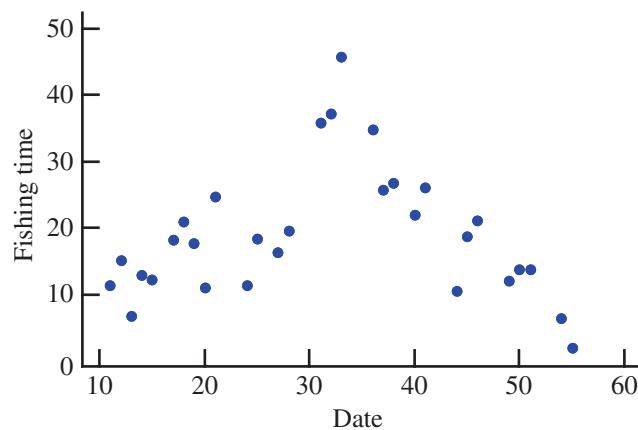
Consider the data ➤

Date (x) (June 1 = 1)	Fishing Time (bear-hours)	Date (x) (June 1 = 1)	Fishing Time (bear-hours)	Date (x) (June 1 = 1)	Fishing Time (bear-hours)
11	11.3	24	11.3	40	22.0
12	15.1	25	18.4	41	26.0
13	6.6	27	16.2	44	10.5
14	12.9	28	19.5	45	18.6
15	12.1	31	35.8	46	21.1
17	18.1	32	37.1	49	11.9
18	20.9	33	45.7	50	13.7
19	17.6	36	34.8	51	13.7
20	11.0	37	25.6	54	6.3
21	24.6	38	26.7	55	1.8

Figure 5.27 is a scatterplot of these data.

FIGURE 5.27

Scatterplot of fishing time versus date.



Formulate a plan ➤

The pattern in the scatterplot is clearly nonlinear. Because the shape of the pattern looks like a quadratic curve (see Figure 5.26(a)), a quadratic function could be used to describe this relationship. This suggests considering the following model:

$$\hat{y} = a + b_1x + b_2x^2$$

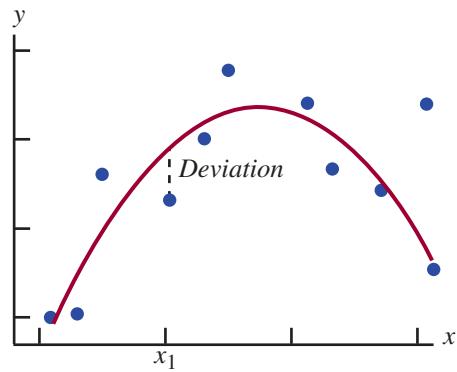
Once we have decided on a potential nonlinear model, such as quadratic or exponential, the next step is to estimate the coefficients in the model. For example, for the quadratic model $\hat{y} = a + b_1x + b_2x^2$, we would need to use the sample data to determine appropriate values for a , b_1 , and b_2 . The process used to estimate the coefficients in a nonlinear model is a bit different for quadratic models than for the other nonlinear models in Figure 5.26. Let's begin by considering quadratic models.

Quadratic Regression Models

The general form of the quadratic regression model is $\hat{y} = a + b_1x + b_2x^2$. What are the best choices for the values of a , b_1 , and b_2 ? In fitting a line to data, the principle of least-squares was used to determine the values of the slope and intercept for the “best fit” line. Least-squares can also be used to fit a quadratic function. The deviations, $y - \hat{y}$, are still represented by vertical distances in the scatterplot, but now they are vertical distances from the points to a parabola (the graph of a quadratic function) rather than to a line, as shown in Figure 5.28. Values for the coefficients in the quadratic function are then chosen to make the sum of the squared deviations as small as possible.

FIGURE 5.28

Deviation for a quadratic function.



For a quadratic regression model, the least-squares estimates of a , b_1 , and b_2 are those values that minimize the sum of squared deviations, $\sum(y - \hat{y})^2$, where $\hat{y} = a + b_1x + b_2x^2$.

For quadratic regression, a measure that is useful for assessing fit is

$$R^2 = 1 - \frac{SSResid}{SSTo}$$

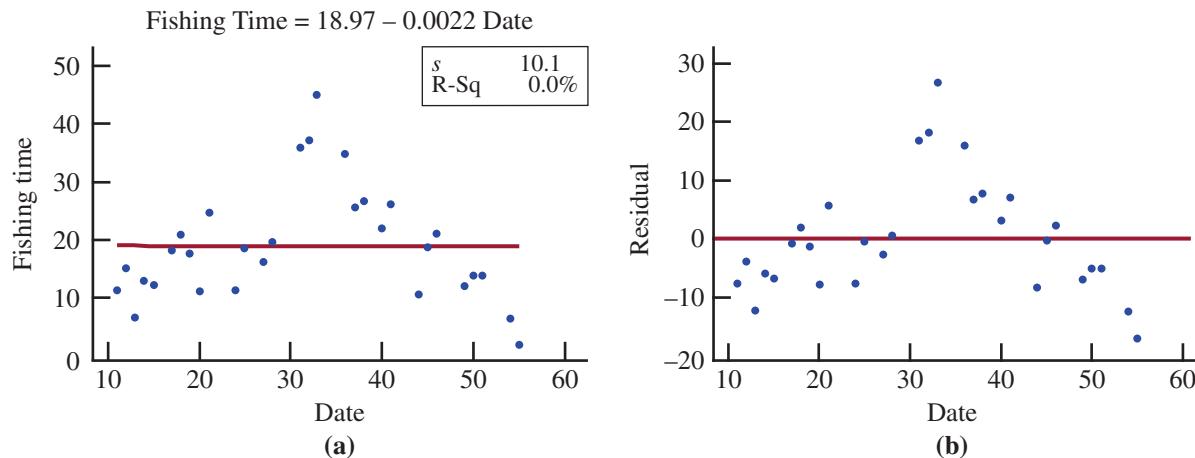
where $SSResid = \sum(y - \hat{y})^2$. The measure R^2 is defined in a way similar to r^2 for the linear regression model and is also interpreted in a similar way. The lowercase notation r^2 is used only with linear regression to emphasize the relationship between r^2 and the correlation coefficient, r , in the linear case. For nonlinear models, an uppercase R^2 is used.

The general expressions for calculating the least-squares quadratic regression estimates are somewhat complicated, so a statistical software package or a graphing calculator will be used to do the computations.

Example 5.16 Fishing Bears Revisited—Fitting a Quadratic Model

FIGURE 5.29

Plots for the bear fishing data of Examples 5.15 and 5.16:
 (a) least-squares line;
 (b) residual plot for linear model.



Do the work ➤ JMP output resulting from fitting the least-squares quadratic regression to these data is shown in Figure 5.30.

FIGURE 5.30

JMP output for the quadratic regression of Example 5.16.

Polynomial Fit Degree=2

Usage = $-20.96713 + 2.9957567 * \text{Date} - 0.0463151 * \text{Date}^2$

Summary of Fit

RSquare	0.555541
RSquare Adj	0.522618
Root Mean Square Error	6.856681
Mean of Response	18.89667
Observations (or Sum Wgts)	30

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-20.96713	7.567052	-2.77	0.0100*
Date	2.9957567	0.524223	5.71	<.0001*
Date^2	-0.046315	0.007973	-5.81	<.0001*

From the JMP output the least-squares quadratic is

$$\hat{y} = -20.967 + 2.996x - 0.046x^2$$

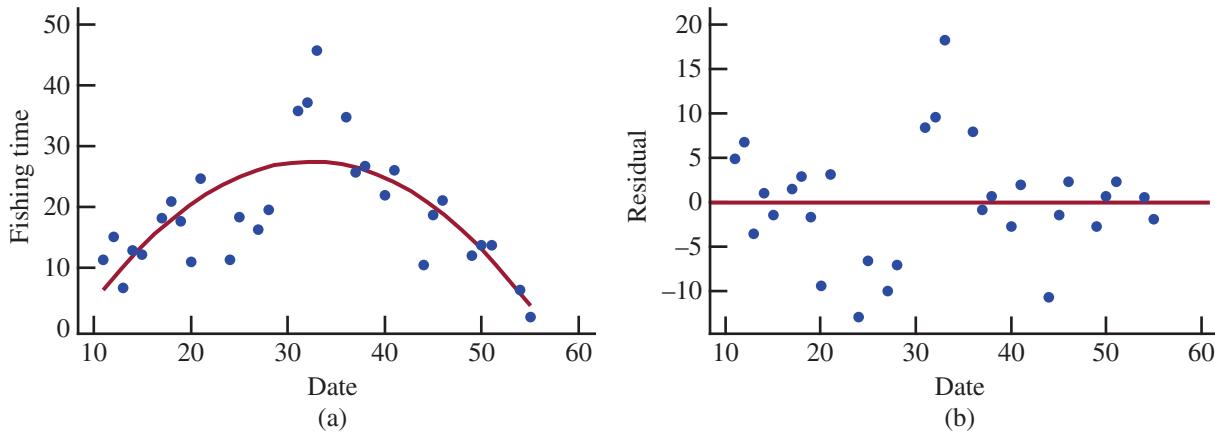
Interpret the results ➤

The curve and the corresponding residual plot for the quadratic regression are shown in Figure 5.31. Notice that there is no strong pattern in this residual plot, unlike the linear case. For the quadratic regression, $R^2 = 0.556$ (as opposed to essentially zero for the linear model). This means that 55.6% of the variability in the bear fishing time can be explained by an approximate quadratic relationship between fishing time and date.

FIGURE 5.31

Quadratic regression of Example 5.16:

- (a) scatterplot with curve;
- (b) residual plot for quadratic model.



To use the quadratic regression model to make a prediction, we substitute a value of x into the regression equation. For example, the predicted fishing time for June 18 ($x = 18$) is

$$\begin{aligned}\hat{y} &= -20.967 + 2.996x - 0.046x^2 \\ &= -20.967 + 2.996(18) - 0.046(18)^2 \\ &= -20.967 + 53.928 - 14.904 \\ &= 18.057 \text{ bear-hours}\end{aligned}$$

Other Nonlinear Regression Models: Using Transformations

Now that we have seen how to use a quadratic regression model to describe relationships that look like the curves in Figure 5.26(a), let's consider some other nonlinear models. The square root function of Figure 5.26(b) is

$$\hat{y} = a + b\sqrt{x}$$

Notice that if you define a new variable x' where

$$x' = \sqrt{x}$$

the square root model $\hat{y} = a + b\sqrt{x}$ can be written as $\hat{y} = a + bx'$, so y is a linear function of x' . This suggests that if a scatterplot of (x, y) pairs looks like the curve in Figure 5.26(b), a scatterplot of (x', y) pairs should look linear. We can then use a line to describe the relationship between y and x' —something we already know how to do. This is illustrated in the following example.

Example 5.17 River Water Velocity and Distance from Shore

Understand the context ➤

As fans of white-water rafting know, a river flows more slowly close to its banks (because of friction between the riverbank and the water). To study the nature of the relationship between water velocity and the distance from the shore, data were gathered on velocity (in

centimeters per second) of a river at different distances (in meters) from the bank. Suppose that the resulting data were as follows:

Consider the data ➤

Distance	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5
Velocity	22.00	23.18	25.48	25.25	27.15	27.83	28.49	28.18	28.50	28.63

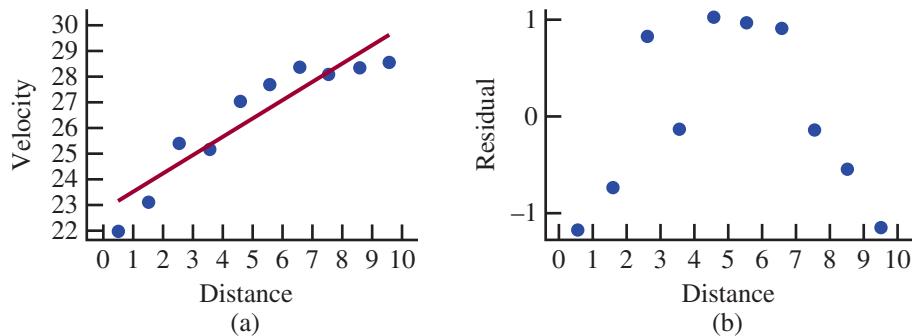
There is a nonlinear relationship between these two variables, as seen in both the scatterplot (Figure 5.32(a)) and the residual plot from a linear fit (Figure 5.32(b)).

FIGURE 5.32

Plots for the data of

Example 5.17:

(a) scatterplot of the river data;
(b) residual plot for linear model.



Formulate a plan ➤

Because the scatterplot of the (x, y) pairs looks like the curve for the square root function in Figure 5.26(b), we begin by replacing each x value by its square root:

$$x' = \sqrt{x}$$

This is called transforming the x values. The original and transformed data are shown in Table 5.2.

TABLE 5.2 Original and Transformed Data of Example 5.17

Do the work ➤

	Original Data		Transformed Data		Original Data		Transformed Data	
	x	y	x'	y	x	y	x'	y
	0.5	22.00	0.7071	22.00	5.5	27.83	2.3452	27.83
	1.5	23.18	1.2247	23.18	6.5	28.49	2.5495	28.49
	2.5	25.48	1.5811	25.48	7.5	28.18	2.7386	28.18
	3.5	25.25	1.8708	25.25	8.5	28.50	2.9155	28.50
	4.5	27.15	2.1213	27.15	9.5	28.63	3.0822	28.63

Figure 5.33(a) shows a scatterplot of y versus x' (or equivalently y versus \sqrt{x}). The pattern of points in this plot looks linear, and so you can find the least-squares line using the transformed data. The Minitab output using the transformed data is given here:

Regression Analysis

The regression equation is

```
velocity = 20.1 + 3.01 sqrt distance
Predictor      Coef    SE Coef      T      P
Constant      20.1103   0.6097  32.99  0.000
sqrt distance  3.0085   0.2726  11.03  0.000
S = 0.629242  R-Sq = 93.8%  R-Sq(adj) = 93.1%
```

The resulting regression equation is

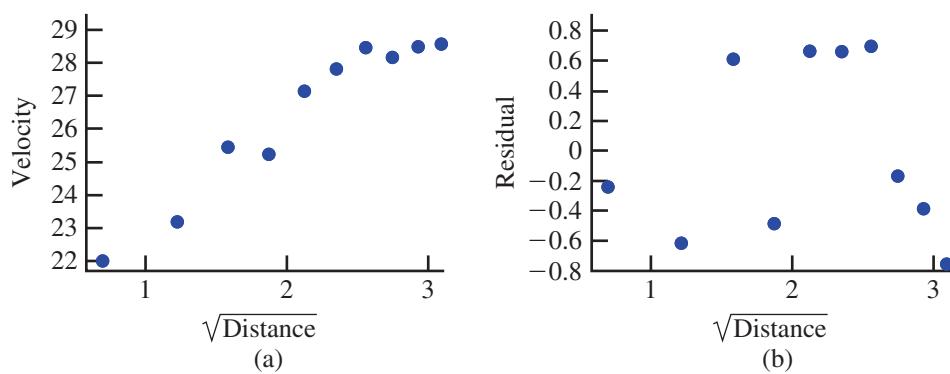
$$\hat{y} = 20.1 + 3.01x'$$

or, equivalently,

$$\hat{y} = 20.1 + 3.01\sqrt{x}$$

FIGURE 5.33

Plots for the transformed data of Example 5.17:
 (a) scatterplot of y versus x' ;
 (b) residual plot resulting from a linear fit to the transformed data.



Interpret the results ►

This model is evaluated using R^2 , s_e , and a residual plot. The residual plot (shown in Figure 5.33(b)) does not have any patterns or unusual features. The values of R^2 and s_e (see the Minitab output) indicate that a line is a reasonable way to describe the relationship between y and x' .

To predict velocity of the river at a distance of 9 meters from shore, we first calculate $x' = \sqrt{x} = \sqrt{9} = 3$ and then use the least-squares line to obtain a predicted y value:

$$\hat{y} = 20.1 + 3.01x' = 20.1 + (3.01)(3) = 29.13 \text{ cm per second}$$

A general strategy for fitting a nonlinear model is to find a way to transform the x and/or y values so that a scatterplot of the transformed data shows a linear pattern. A **transformation** (sometimes called a re-expression) involves using a simple function of a variable in place of the variable itself (like $x' = \sqrt{x}$ in Example 5.17). Common transformations involve taking square roots, logarithms, or reciprocals.

The nonlinear models (other than the quadratic model) given in Figure 5.26 (square root, reciprocal, log, exponential, and power) can all be fit by transforming variables and then fitting a line to the transformed data. It is convenient to separate these models into two types:

1. Models that involve transforming *only* the x variable (square root, reciprocal, and log).
2. Models that involve transforming the y variable (exponential and power).

Models That Involve Transforming Only x

The square root, reciprocal, and log models all have the form

$$\hat{y} = a + b(\text{some function of } x)$$

where the function of x is square root, reciprocal, or log. Notice that these models all describe a linear relationship between y and a function of x . This suggests that if the pattern in the scatterplot of (x, y) pairs looks like one of the curves in Figure 5.26(b)–(d), an appropriate transformation of the x values should result in transformed data that shows a linear pattern.

Model	Transformation
Square root	$x' = \sqrt{x}$
Reciprocal	$x' = \frac{1}{x}$
Log	$x' = \ln(x)$

Example 5.18 Electromagnetic Radiation

Understand the context ►

Is electromagnetic radiation from phone antennae associated with declining bird populations? This is one of the questions investigated by the authors of the paper “[The Urban Decline of the House Sparrow \(*Passer domesticus*\): A Possible Link with Electromagnetic](#)

Radiation" (*Electromagnetic Biology and Medicine* [2007]: 141–151). The accompanying data on x = Field strength (the strength of the electromagnetic field in volts per meter) and y = Sparrow density (sparrows per hectare) were read from a graph in the paper.

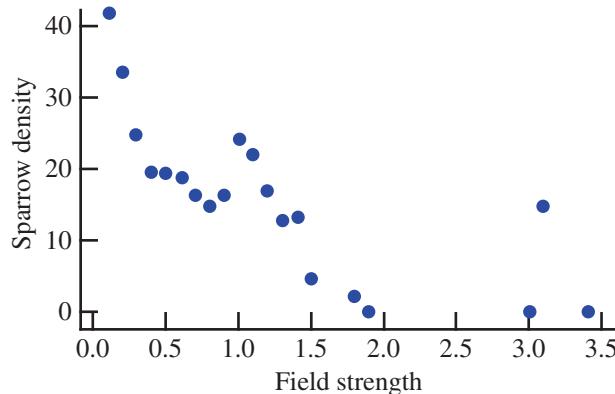
Consider the data ➤

Field Strength	Sparrow Density	Field Strength	Sparrow Density
0.11	41.71	0.90	16.29
0.20	33.60	1.20	16.97
0.29	24.74	1.30	12.83
0.40	19.50	1.41	13.17
0.50	19.42	1.50	4.64
0.61	18.74	1.80	2.11
1.01	24.23	1.90	0.00
1.10	22.04	3.01	0.00
0.70	16.29	3.10	14.69
0.80	14.69	3.41	0.00

A scatterplot of these data is shown in Figure 5.34.

FIGURE 5.34

Scatterplot of the sparrow data.



Formulate a plan ➤

Notice that there is a curved pattern in the scatterplot that has a shape that is similar to the first graph in Figure 5.26(d). This suggests that a reasonable choice for describing the relationship between sparrow density and field strength is a log model

$$\hat{y} = a + b\ln(x)$$

To fit this model, begin by transforming the x values using

$$x' = \ln(x)$$

Do the work ➤

In this example, natural logs (denoted by \ln or \log_e) are used, but in practice, either the natural log or log base 10 (denoted by \log or \log_{10}) can be used. The transformed data is given here.

ln(Field Strength)	Sparrow Density	ln(Field Strength)	Sparrow Density
-2.207	41.71	-0.105	16.29
-1.609	33.60	0.182	16.97
-1.238	24.74	0.262	12.83
-0.916	19.50	0.344	13.17
-0.693	19.42	0.405	4.64
-0.494	18.74	0.588	2.11
0.001	24.23	0.642	0.00
0.095	22.04	1.102	0.00
-0.357	16.29	1.131	14.69
-0.223	14.69	1.227	0.00

A scatterplot of the (x', y) pairs is shown in Figure 5.35. The pattern in this plot looks linear, so you can find the least-squares line using the transformed data. From the resulting Minitab output (Figure 5.36), the equation of the least-squares line is

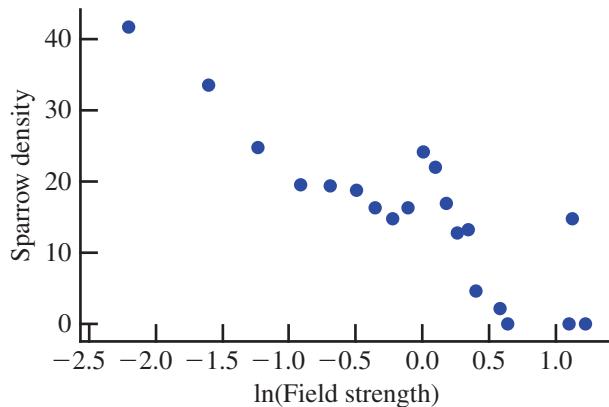
$$\hat{y} = 14.8 - 10.5x'$$

or, in terms of x

$$\hat{y} = 14.8 - 10.5 \ln(x)$$

FIGURE 5.35

Scatterplot of the transformed data of Example 5.18.

**FIGURE 5.36**

Minitab output for the log model of Example 5.18.

Regression Analysis: Sparrow Density Versus ln(Field Strength)

The regression equation is

$$\text{Sparrow Density} = 14.8 - 10.5 \ln(\text{Field Strength})$$

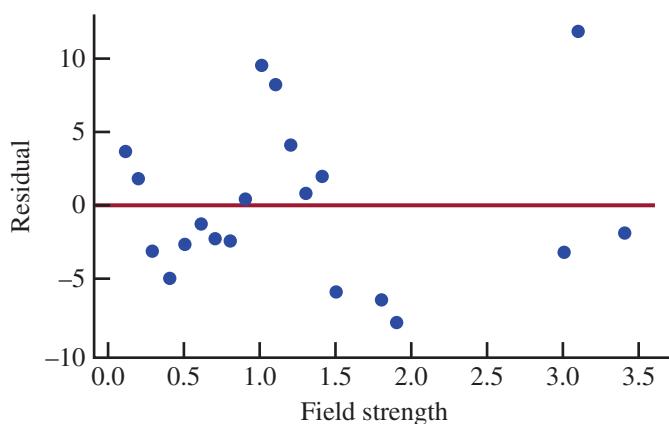
Predictor	Coef	SE Coef	T	P
Constant	14.805	1.238	11.96	0.000
ln (Field Strength)	-10.546	1.389	-7.59	0.000
S = 5.50641	R-Sq = 76.2%		R-Sq(adj) = 74.9%	

Interpret the results ➤

The value of R^2 for this model is 0.762 and $s_e = 5.5$. A residual plot from the least-squares line fit to the transformed data is shown in Figure 5.37. There are no apparent patterns or unusual features in the residual plot, so it appears that the log model is a reasonable choice for describing the relationship between sparrow density and field strength.

FIGURE 5.37

Residual plot for the log model fit to the sparrow data of Example 5.18.



This model can now be used to predict sparrow density from field strength. For example, if the field strength is 1.6 Volts per meter, we would predict that sparrow density would be

$$\begin{aligned}\hat{y} &= 14.8 - 10.5x' \\ &= 14.8 - 10.5 \ln(x) \\ &= 14.8 - 10.5 \ln(1.6) \\ &= 14.8 - 10.5 (0.470) \\ &= 9.685 \text{ sparrows per hectare}\end{aligned}$$

Models That Involve Transforming y

There are two nonlinear models in Figure 5.26 that have not yet been considered—the exponential and power models. Let's begin by considering the exponential model

$$y = e^{a+bx}$$

Using properties of logarithms, it follows that

$$\begin{aligned}\ln(y) &= \ln e^{a+bx} \\ &= a + bx\end{aligned}$$

For the exponential model, $\ln(y)$ is a linear function of x . This means that if the pattern in a scatterplot of the (x, y) pairs looks like the exponential curves of Figure 5.26(e), a scatterplot of the (x, y') pairs, where

$$y' = \ln y$$

should show a linear pattern.

Now consider the power model of Figure 5.26(f),

$$y = ax^b$$

Again using properties of logarithms,

$$\begin{aligned}\ln y &= \ln(ax^b) \\ &= \ln a + \ln x^b \\ &= \ln a + b \ln x\end{aligned}$$

If a power model is an appropriate way to describe the relationship between x and y , a linear model would describe the relationship between $x' = \ln x$ and $y' = \ln y$.

Model	Transformation
Exponential	$y' = \ln y$
Power	$x' = \ln x$

Example 5.19 Loons in Acidic Lakes

Understand the context ➤

A study of factors that affect the survival of loon chicks is described in the paper “[Does Prey Biomass or Mercury Exposure Affect Loon Chick Survival in Wisconsin?](#)” (*The Journal of Wildlife Management* [2005]: 57–67). In this study, a relationship between the pH of lake water and blood mercury level in loon chicks was observed.

The researchers thought that it is possible that the pH of the lake water could be related to the type of fish that the loons ate. The accompanying data (Table 5.3) on x = Lake pH and y = Blood mercury level (mg/g) for 37 loon chicks from different lakes in Wisconsin were read from a graph in the paper. A scatterplot is shown in Figure 5.38(a).

TABLE 5.3 Data and Transformed Data from Example 5.19

Lake pH (x)	Blood Mercury Level (y)	$y' = \ln(y)$	Lake pH (x)	Blood Mercury Level (y)	$y' = \ln(y)$	Lake pH (x)	Blood Mercury Level (y)	$y' = \ln(y)$
5.28	1.10	0.0953	6.17	0.55	-0.5978	7.03	0.12	-2.1203
5.69	0.76	-0.2744	6.22	0.43	-0.8440	7.20	0.15	-1.8971
5.56	0.74	-0.3011	6.15	0.40	-0.9163	7.89	0.11	-2.2073
5.51	0.60	-0.5108	6.05	0.33	-1.1087	7.93	0.11	-2.2073
4.90	0.48	-0.7340	6.04	0.26	-1.3471	7.99	0.09	-2.4079
5.02	0.43	-0.8440	6.24	0.18	-1.7148	7.99	0.06	-2.8134
5.02	0.29	-1.2379	6.30	0.16	-1.8326	8.30	0.09	-2.4079
5.04	0.09	-2.4079	6.80	0.45	-0.7985	8.42	0.09	-2.4079
5.30	0.10	-2.3026	6.58	0.30	-1.2040	8.42	0.04	-3.2189
5.33	0.20	-1.6094	6.65	0.28	-1.2730	8.95	0.12	-2.1203
5.64	0.28	-1.2730	7.06	0.22	-1.5141	9.49	0.14	-1.9661
5.83	0.17	-1.7720	6.99	0.21	-1.5606			
5.83	0.18	-1.7148	6.97	0.13	-2.0402			

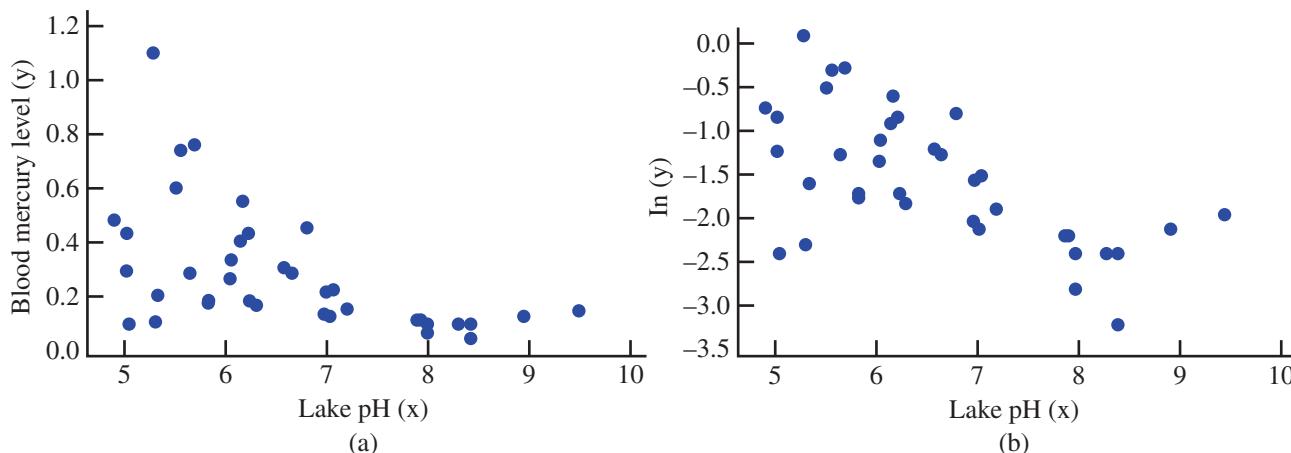
Formulate a plan ➤

The pattern in the scatterplot is typical of exponential decay and looks like the second curve in Figure 5.26(e). The change in y as x increases is much smaller for large x values than for small x values. We can see that a change of 1 in pH is associated with a much larger change in blood mercury level in the part of the plot where the x values are small than in the part of the plot where the x values are large.

FIGURE 5.38

Scatterplots for the loon data of Example 5.19:

- (a) scatterplot of original data;
- (b) scatterplot of transformed data with $y' = \ln y$.



To fit an exponential model, begin by transforming the y values using $y' = \ln y$. The transformed y values are given in Table 5.3. A scatterplot of the (x, y') pairs is shown in Figure 5.38(b). The pattern in the scatterplot of the transformed data looks linear, so it is reasonable to use the least-squares line to describe the relationship between y' and x .

The following Minitab output is the result of fitting the least-squares line to the transformed data:

Do the work ➤ The regression equation is

```
ln(y) = 1.06 - 0.396 Lake pH (x)

Predictor      Coef    SE Coef      T      P
Constant      1.0550   0.5535     1.91  0.065
Lake pH (x)   -0.39564  0.08264    -4.79  0.000
S = 0.605645  R-Sq = 39.6%  R-Sq(adj) = 37.8%
```

The resulting least-squares line is

$$\hat{y}' = 1.06 - 0.396x$$

or equivalently

$$\hat{\ln y} = 1.06 - 0.396x$$

Notice that the linear model in Example 5.19 could be used to predict values of $\ln(y)$. To translate this linear model back to the form of the exponential model, you can take antilogarithms of each side of the equation. This is illustrated in Example 5.20.

Example 5.20 Revisiting the Loon Data

For the loon data of Example 5.19, $y' = \ln y$ and the equation of the least-squares line describing the relationship between y' and x was $\hat{y}' = 1.06 - 0.396x$ or $\hat{\ln y} = 1.06 - 0.396x$. To “undo” the transformation, we can take the antilog of both sides of this equation:

$$e^{\ln y} = e^{1.06 - 0.396x}$$

Using properties of logs and exponents, $e^{\ln y} = e^{1.06 - 0.396x} = (e^{1.06})(e^{-0.396x})$. Then

$$\hat{y} = e^{1.06} e^{-0.396x} = 2.886e^{-0.396x}$$

The equation $\hat{y} = 2.886e^{-0.396x}$ can be used to predict the y value (blood mercury level) for a given x (lake pH). For example, the predicted blood mercury level when lake pH is 6 is

$$\hat{y} = 2.886e^{-0.396x} = 2.886e^{0.396(6)} = 2.886e^{-2.376} = 2.886(0.093) = 0.268$$

The process of transforming data, fitting a line to the transformed data, and then undoing the transformation to get an equation for a curved relationship between x and y usually results in a curve that provides a reasonable fit to the sample data. But when y is transformed, this process does not result in the least-squares curve for the data. For example, in Example 5.20, a transformation was used to fit the curve $\hat{y} = 2.886e^{-0.396x}$. However, there may be another equation of the form $\hat{y} = ae^{bx}$ that has a smaller sum of squared residuals for the *original* data than the one obtained using transformations. Finding the least-squares estimates for a and b for an exponential or power model is difficult. Fortunately, the curves found using transformations usually provide reasonable predictions of y .

Choosing Among Different Possible Nonlinear Models

Often there is more than one reasonable model that could be used to describe a nonlinear relationship between two variables. When this is the case, how do we choose a model? The choice often involves considering how well the model fits the data and what scientific theory suggests the form of the relationship might be. In the absence of any scientific theory, we want to choose a model that has small residuals (small s_e) and accounts for a large proportion of the variability in y (large R^2). On the other hand, if scientific theory suggests that the relationship should have a particular form, we would choose a model that is consistent with the suggested form. The following two examples illustrate these two different situations.

Example 5.21 Look for Salamanders Under Those Rocks

Understand the context ➤

The paper “The Relationship Between Rock Density and Salamander Density in a Mountain Stream” (*Herpetologica* [1987]: 357–361) describes a study of factors that influence population density of salamanders. In this study, researchers created a range of salamander habitats at different locations in a small stream in the Appalachian Mountains by using different sized rocks and pebbles. Three months later, they returned to measure salamander density. This resulted in data on

x = Rock density (rocks per 1.4 square meters)

and

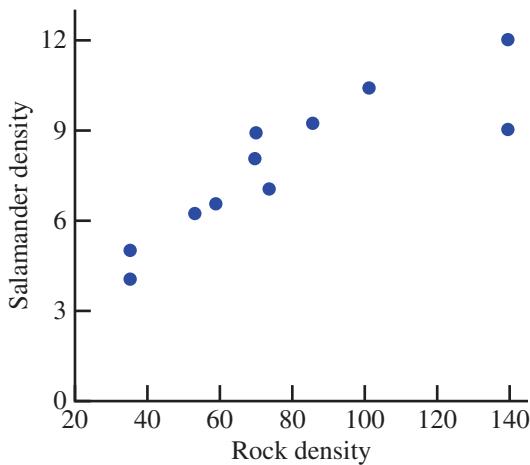
y = Salamander density (salamanders per 1.4 square meters)

Consider the data ➤

A scatterplot of these data is shown in Figure 5.39.

FIGURE 5.39

Scatterplot of salamander density versus rock density.



Formulate a plan ➤

Looking at the curves in Figure 5.26, we might think about using a square root model or a reciprocal model. The researchers chose the reciprocal regression model, $y = a + b\left(\frac{1}{x}\right)$, for two reasons: (1) it provided a good fit to the data, and (2) the reciprocal model made sense scientifically. The investigators felt that there would be an upper limit to the population density, since the streambed is a nonrenewable resource and the stream therefore had a limit on the number of salamanders that could be sustained. This limit is known as the “carrying capacity” of an environment and is estimated by the value of the intercept, a , in the reciprocal model.

Do the work ➤

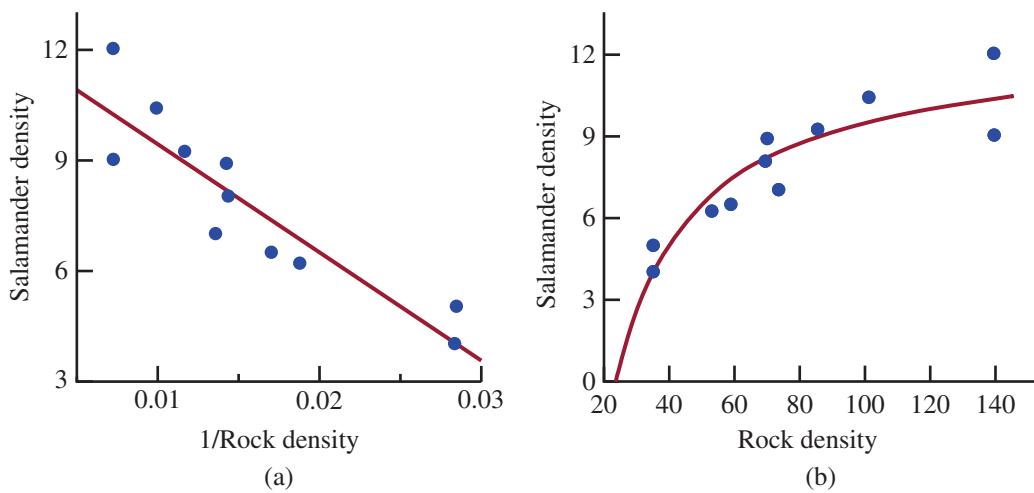
Using the transformation $x' = \frac{1}{x}$, a least-squares line was fit using the transformed data. The resulting model was

$$y = 12.37 - 292.6x' = 12.37 - 292.6\left(\frac{1}{x}\right)$$

The value of R^2 for this model is $R^2 = 0.82$. Figure 5.40(a) shows a scatterplot of the transformed data and the least-squares line. Figure 5.40(b) is a scatterplot of the original data that also shows the curve corresponding to the reciprocal model.

FIGURE 5.40

Scatterplots for the salamander data of Example 5.21:
(a) transformed data;
(b) original data.



EXAMPLE 5.22 How Old Is That Lobster?

Understand the context ➤

Can you tell how old a lobster is by its size? This question was investigated by the authors of a paper that appeared in the *Biological Bulletin (August, 2007)*. Researchers measured carapace (the exterior shell) length (in mm) of 27 laboratory-raised lobsters of known age. The data on x = Carapace length and y = Age (in years) in Table 5.4 were read from a graph in the paper.

TABLE 5.4 Original and Transformed Data for Example 5.22

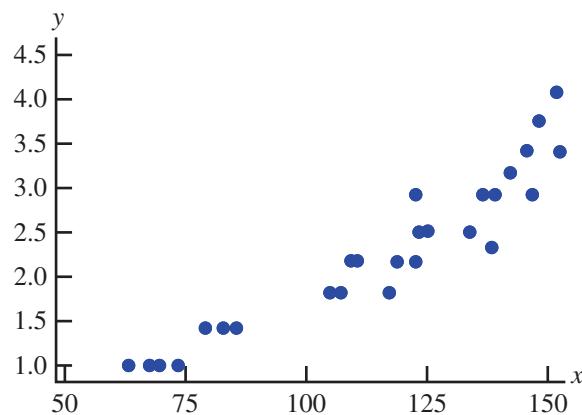
Consider the data ➤

	y	x	$\ln(x)$	$\ln(y)$		y	x	$\ln(x)$	$\ln(y)$
	1.00	63.32	4.148	0.000		2.33	138.47	4.931	0.846
	1.00	67.50	4.212	0.000		2.50	133.95	4.897	0.916
	1.00	69.58	4.242	0.000		2.51	125.25	4.830	0.920
	1.00	73.41	4.296	0.000		2.50	123.51	4.816	0.916
	1.42	79.32	4.373	0.351		2.93	146.82	4.989	1.075
	1.42	82.80	4.416	0.351		2.92	139.17	4.936	1.072
	1.42	85.59	4.450	0.351		2.92	136.73	4.918	1.072
	1.82	105.07	4.655	0.599		2.92	122.81	4.811	1.072
	1.82	107.16	4.674	0.599		3.17	142.30	4.958	1.154
	1.82	117.25	4.764	0.599		3.41	152.73	5.029	1.227
	2.18	109.24	4.694	0.779		3.42	145.78	4.982	1.230
	2.18	110.64	4.706	0.779		3.75	148.21	4.999	1.322
	2.17	118.99	4.779	0.775		4.08	152.04	5.024	1.406
	2.17	122.81	4.811	0.775					

Formulate a plan ➤

The scatterplot of the data in Figure 5.41 shows a clear curved pattern. Considering the curves in Figure 5.26, there are several models that could be considered, including the exponential model and the power model.

FIGURE 5.41
Scatterplot of age versus carapace length.



In this situation, there is no scientific theory to suggest what the form of the relationship might be. A reasonable way to proceed is to fit both models and then choose the one that provides the best fit. Let's start with the exponential model. Using the transformation

$$y' = \ln y$$

results in the transformed y' values given in Table 5.4. A scatterplot of the (x, y') pairs is shown in Figure 5.42(a). Notice that the relationship between y' and x looks

linear. Minitab was used to fit the least-squares line to the (x, y') data, resulting in the following output:

Do the work ▶ **Regression Analysis: $\ln(y)$ versus x**
The regression equation is
 $\ln(y) = -0.927 + 0.0145 x$
Predictor Coef SE Coef T P
Constant -0.92704 0.08692 -10.67 0.000
x 0.0144903 0.0007311 19.82 0.000
S = 0.105917 R-Sq = 94.0% R-Sq(adj) = 93.8%

FIGURE 5.42

Scatterplot of the transformed lobster data:
(a) x = carapace length,
 $y' = \ln(\text{age})$;
(b) $x' = \ln(\text{carapace length})$,
 $y' = \ln(\text{age})$.

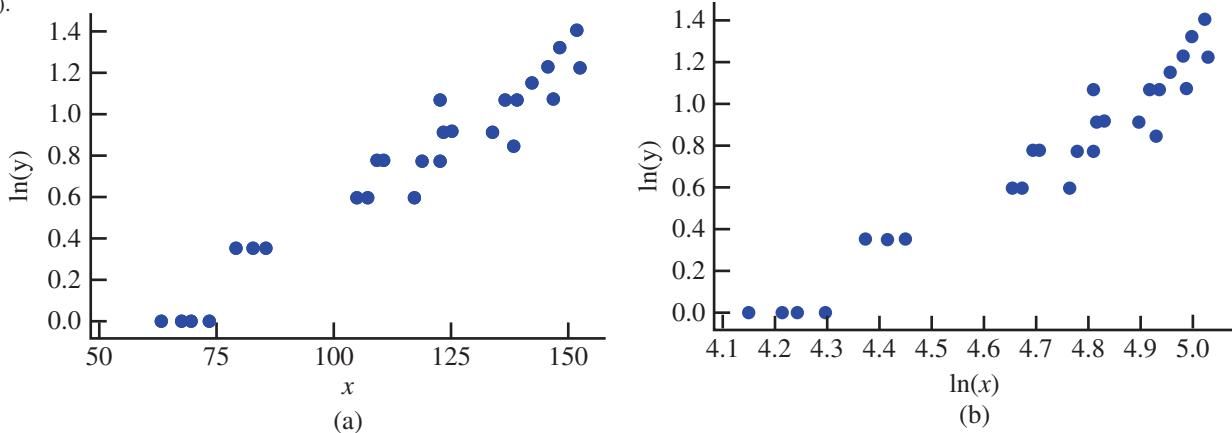


Figure 5.42(b) shows a scatterplot of the transformed data that would be used to fit a power model ($x' = \ln(x)$, $y' = \ln(y)$). The pattern in this scatterplot also looks linear. Minitab output from fitting a least-squares line using the (x', y') data is shown here:

Regression Analysis: $\ln(y)$ versus $\ln(x)$

The regression equation is
 $\ln(y) = -6.32 + 1.50 \ln(x)$
Predictor Coef SE Coef T P
Constant -6.3205 0.3687 -17.14 0.000
 $\ln(x)$ 1.49866 0.07805 19.20 0.000
S = 0.109120 R-Sq = 93.6% R-Sq(adj) = 93.4%

Notice that both the exponential model and the power model have large R^2 values. Comparing the fit of the two models results in

Model	R^2	S_e
Exponential	0.940	0.106
Power	0.936	0.109

Interpret the results ▶

Both models do a good job in describing the relationship between age and carapace length. In the absence of any scientific theory suggesting the form of the model, we might choose the exponential model because it provides a (slightly) better fit.

The exponential model is

$$\hat{y}' = -0.927 + 0.0145x$$

or equivalently

$$\ln \hat{y} = -0.927 + 0.0145x$$

or

$$\hat{y} = e^{-0.927} e^{0.0145x} = 0.396 e^{0.0145x}$$

EXERCISES 5.54 - 5.63

• Data set available online

- 5.54** The following data on x = Frying time (in seconds) and y = Moisture content (%) appeared in the paper “**Thermal and Physical Properties of Tortilla Chips as a Function of Frying Time**” (*Journal of Food Processing and Preservation* [1995]: 175–189):

x	5	10	15	20	25	30	45	60
y	16.3	9.7	8.1	4.2	3.4	2.9	1.9	1.3

- a. Construct a scatterplot of the data.
b. Would a line provide an effective summary of the relationship? Explain.

- 5.55** Use the information provided in the previous exercise to answer the following questions.

- a. Here are the values of $x' = \log(x)$ and $y' = \log(y)$:

x'	0.70	1.00	1.18	1.30	1.40	1.48	1.65	1.78
y'	1.21	0.99	0.91	0.62	0.53	0.46	0.28	0.11

Construct a scatterplot of these transformed data, and comment on the pattern.

- b. Based on the accompanying MINITAB output, does the least-squares line effectively summarize the relationship between y' and x' ?

The regression equation is

$$\text{log(moisture)} = 2.02 - 1.05 \text{ log(time)}$$

Predictor	Coef	SE Coef	T	P
Constant	2.01780	0.09584	21.05	0.000
log(time)	-1.05171	0.07091	-14.83	0.000

$S = 0.0657067$	R-Sq = 97.3%	R-Sq(adj) = 96.9%
-----------------	--------------	-------------------

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.94978	0.94978	219.99	0.000
Residual Error	6	0.02590	0.00432		
Total	7	0.97569			

- c. Use the MINITAB output to predict moisture content when frying time is 35 sec.

- 5.56** The paper “**Aspects of Food Finding by Wintering Bald Eagles**” (*The Auk* [1983]: 477–484) examined the relationship between the time that eagles spend aloft searching for food (indicated by the percentage of eagles soaring) and relative food availability. The accompanying data were taken from a scatterplot that appeared in this paper. Salmon availability is denoted by x and the percentage of eagles in the air is denoted by y .

x	0	0	0.2	0.5	0.5	1.0
y	28.2	69.0	27.0	38.5	48.4	31.1
x	1.2	1.9	2.6	3.3	4.7	6.5
y	26.9	8.2	4.6	7.4	7.0	6.8

- a. Draw a scatterplot for this data set. Would you describe the pattern in the plot as linear or curved?

- b. One possible transformation that might lead to a linear pattern involves taking the square root of both the x and y values. Construct a scatterplot using the variables \sqrt{x} and \sqrt{y} . Is a line a more reasonable choice for describing the pattern in this scatterplot than the pattern in the scatterplot in Part (a)?

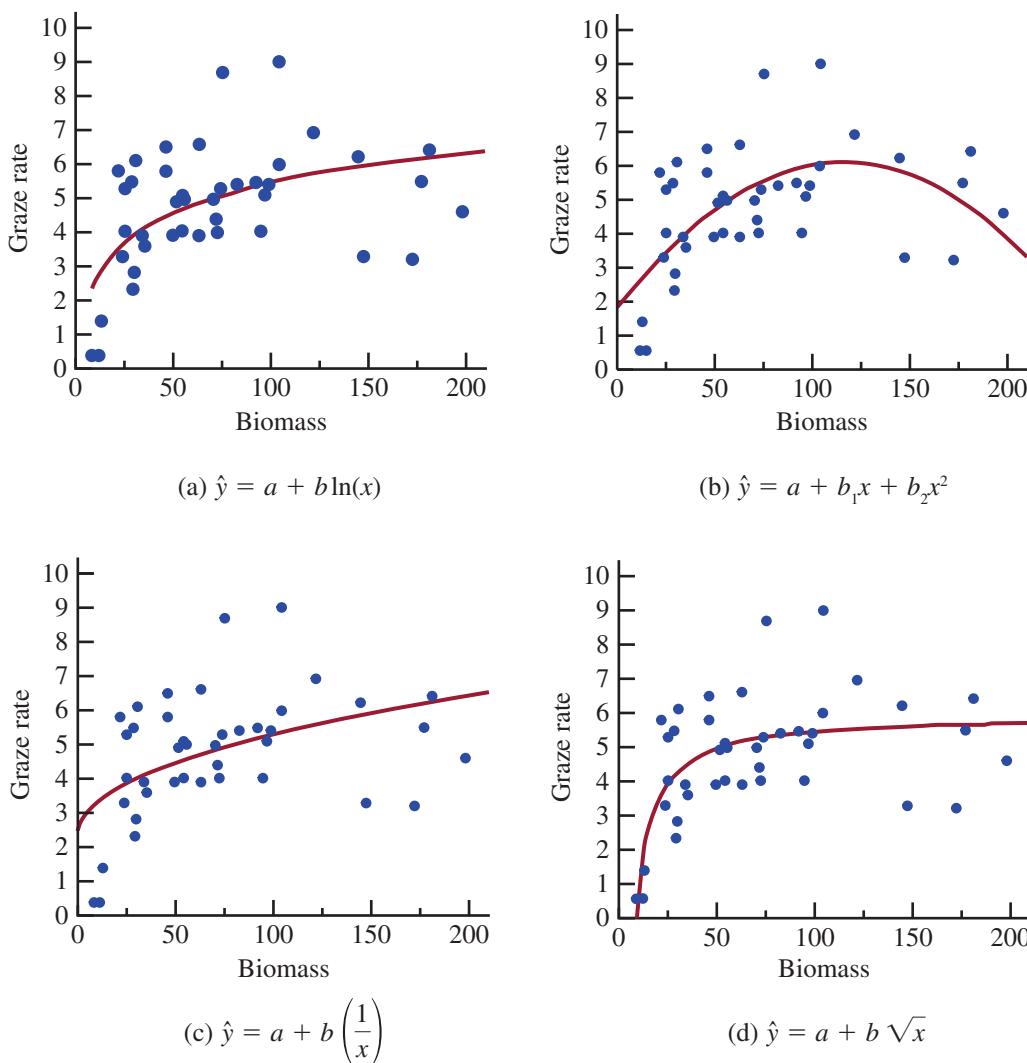
- c. After considering the scatterplot in Part (a), suggest another transformation that might be used to straighten the original plot. Use your suggestion to construct a scatterplot for your x' and/or y' . Which transformation (the one you suggested or the square root transformation from Part (b)) is preferable? Explain.

- 5.57** Food intake of grazing animals is limited by the rate grass can be chewed and swallowed, as well as the rate at which food can be digested. The authors of the paper “**What Constrains Daily Intake in Thomson’s Gazelles?**” (*Ecology* [1999]: 2338–2347) observed the grazing activity of captive Thomson’s gazelles. They recorded grazing rate (amount of grass eaten, in grams per minute) and biomass of the grazing area (food density, in grams per square meter). Scatterplots of these data with four possible functions that might be used to describe the relationship between grazing rate and biomass are shown in the figure at the top of the following page. Which of these functions would you recommend? Explain your reasoning in a few sentences. (Hint: See Example 5.21.)

- 5.58** A study, described in the paper “**Prediction of Defibrillation Success from a Single Defibrillation Threshold Measurement**” (*Circulation* [1988]: 1144–1149) investigated the relationship between defibrillation success and the energy of the defibrillation shock (expressed as a multiple of the defibrillation threshold). The accompanying data are from this study.

Energy of Shock	Success (%)
0.5	33.3
1.0	58.3
1.5	81.8
2.0	96.7
2.5	100.0

- a. Construct a scatterplot of y = Success and x = Energy of shock. Does the relationship appear to be linear or nonlinear?

Figure for Exercise 5.57

- b. Fit a least-squares line to the given data, and construct a residual plot. Does the residual plot support your conclusion in Part (a)? Explain.

- 5.59** Refer to the data given in the previous exercise.
- Consider transforming the data by leaving y unchanged and using either $x' = \sqrt{x}$ or $x'' = \ln(x)$. Which of these transformations would you recommend? Justify your choice by appealing to appropriate graphical displays.
 - Using the transformation you recommended in Part (a), find the equation of the least-squares line that describes the relationship between y and the transformed x .
 - What would you predict success to be when the energy of shock is 1.75 times the threshold level? When it is 0.8 times the threshold level?

- 5.60** The following table gives the number of heart transplants performed in the United States each year from 2006 to 2015 ([U.S. Department of Health and Human Services](http://optn.transplant.hrsa.gov/data/view-data-reports/national-data/),

optn.transplant.hrsa.gov/data/view-data-reports/national-data/, retrieved April 22, 2017):

Year	Number of Heart Transplants
1 (2006)	2,193
2	2,209
3	2,163
4	2,211
5	2,332
6	2,322
7	2,378
8	2,531
9	2,655
10 (2015)	2,804

- Construct a scatterplot of these data. Describe how the number of heart transplants has changed over time from 2006 to 2015.
- Find the equation of the least-squares line that describes the relationship between y = Number of heart transplants performed and x = Year.

- c. Calculate the 10 residuals and construct a residual plot.
- d. Are there any features of the residual plot that indicate that the relationship between year and number of heart transplants performed would be better described by curve rather than a line? Explain.
- 5.61** Refer to the heart transplant data given in the previous exercise.
- Find a transformation of x and/or y that straightens the plot. Construct a scatterplot for your transformed variables.
 - Using the transformed variables from Part (a), fit a least-squares line and use it to predict the number waiting for a heart transplant in 2016 (Year 11).
 - The prediction made in Part (b) involves prediction for an x value that is outside the range of the x values in the sample. What assumption must you be willing to make for this to be reasonable? Do you think this assumption is reasonable in this case? Would your answer be the same if the prediction had been for the year 2026 rather than 2016? Explain.
- 5.62** The paper “Population Pressure and Agricultural Intensity” (*Annals of the Association of American Geographers* [1977]: 384–396) reported a positive relationship between population density and agricultural intensity. The following data consist of measures of population density (x) and agricultural intensity (y) for 18 different subtropical locations:
- | | | | | | | |
|-----|-------|-------|------|-------|-------|-------|
| x | 1.0 | 26.0 | 1.1 | 101.0 | 14.9 | 134.7 |
| y | 9 | 7 | 6 | 50 | 5 | 100 |
| x | 3.0 | 5.7 | 7.6 | 25.0 | 143.0 | 27.5 |
| y | 7 | 14 | 14 | 10 | 50 | 14 |
| x | 103.0 | 180.0 | 49.6 | 140.6 | 140.0 | 233.0 |
| y | 50 | 150 | 10 | 67 | 100 | 100 |
- Construct a scatterplot of y versus x . Is the scatterplot compatible with the statement of positive relationship made in the paper?
- b. Construct a scatterplot that uses y and x^2 . Does this transformation result in a linear pattern in the scatterplot?
- c. Draw a scatterplot that uses $\log(y)$ and x . The $\log(y)$ values, given in order corresponding to the y values, are 0.95, 0.85, 0.78, 1.70, 0.70, 2.00, 0.85, 1.15, 1.15, 1.00, 1.70, 1.15, 1.70, 2.18, 1.00, 1.83, 2.00, and 2.00. How does this scatterplot compare with that of Part (b)?
- d. Now consider a scatterplot that uses transformations on both x and y : $\log(y)$ and x^2 . Is this effective in creating a linear pattern in the plot? Explain.
- 5.63** Determining the age of an animal can sometimes be a difficult task. One method of estimating the age of harp seals is based on the width of the pulp canal in the canine teeth. To investigate the relationship between age and the width of the pulp canal, researchers recorded age and canal width in seals of known age. The following data on x = Age (in years) and y = Canal length (in millimeters) are a portion of a larger data set that appeared in the paper “Validation of Age Estimation in the Harp Seal Using Dentinal Annuli” (*Canadian Journal of Fisheries and Aquatic Sciences* [1983]: 1430–1441):
- | | | | | | | | | |
|-----|------|------|------|------|------|------|------|------|
| x | 0.25 | 0.25 | 0.50 | 0.50 | 0.50 | 0.75 | 0.75 | 1.00 |
| y | 700 | 675 | 525 | 500 | 400 | 350 | 300 | 300 |
| x | 1.00 | 1.00 | 1.00 | 1.00 | 1.25 | 1.25 | 1.50 | 1.50 |
| y | 250 | 230 | 150 | 100 | 200 | 100 | 100 | 125 |
| x | 2.00 | 2.00 | 2.50 | 2.75 | 3.00 | 4.00 | 4.00 | 5.00 |
| y | 60 | 140 | 60 | 50 | 10 | 10 | 10 | 10 |
| x | 5.00 | 5.00 | 5.00 | 6.00 | 6.00 | | | |
| y | 15 | 10 | 10 | 15 | 10 | | | |
- Construct a scatterplot for this data set. Would you describe the relationship between age and canal length as linear? If not, suggest a transformation that might straighten the plot.

SECTION 5.5 Interpreting and Communicating the Results of Statistical Analyses

Using either the least-squares line to summarize a linear relationship or a correlation coefficient to describe the strength of a linear relationship is common in investigations that focus on more than a single variable. The methods described in this chapter are among the most widely used statistical tools. When bivariate numerical data are analyzed in journal articles and other published sources, it is common to find a scatterplot of the data and a least-squares line or a correlation coefficient.

Communicating the Results of Statistical Analyses

When reporting the results of a data analysis involving bivariate numerical data, it is important to include graphical displays as well as numerical summaries. Including a scatterplot and providing a description of what the plot reveals about the form of the relationship between the two variables under study establishes the context in which numerical summary measures, such as the correlation coefficient or the equation of the least-squares line, can be interpreted.

In general, the goal of an analysis of bivariate numerical data is to learn about the relationship between the two variables. If there is a relationship, we can describe how strong or weak the relationship is. If the goal of the study is to describe the strength of the relationship and the scatterplot shows a linear pattern, the value of the correlation coefficient or the coefficient of determination can be reported as a measure of the strength of the linear relationship.

When interpreting the value of the correlation coefficient, it is a good idea to relate the interpretation to the pattern observed in the scatterplot. This is especially important before making a statement of no relationship between two variables, because a correlation coefficient with a value near 0 does not necessarily imply that there is no relationship of any form. Similarly, a correlation coefficient near 1 or -1 , by itself, does not guarantee that the relationship is linear.

If the goal of a study is prediction, then in addition to a scatterplot and the equation of the least-squares line, how well the linear prediction model fits the data should also be addressed. At a minimum, both the values of s_e (the standard deviation about the least-squares line) and r^2 (the coefficient of determination) should be included. Including a residual plot can also provide support for the appropriateness of using a line to describe the relationship between the two variables.

What to Look for in Published Data

Here are a few things to consider when you read an article that includes an analysis of bivariate data:

- What two variables are being studied? Are they both numerical? Is a distinction made between a dependent variable and an independent variable?
- Does the article include a scatterplot of the data? If so, does there appear to be a relationship between the two variables? Can the relationship be described as linear, or is some type of nonlinear relationship a more appropriate description?
- Does the relationship between the two variables appear to be weak or strong? Is the value of a correlation coefficient reported?
- If the least-squares line is used to summarize the relationship between the dependent and independent variables, is there an assessment of how accurate predictions based on the least-squares line might be? For example, are the values of r^2 or s_e reported? How are these values interpreted, and what do they imply about the usefulness of the least-squares line?
- If a correlation coefficient is reported, is it interpreted properly? Be cautious of interpretations that claim a causal relationship.

The authors of the paper “[Recycling and Ambivalence: Quantitative and Qualitative Analyses of Household Recycling Among Young Adults](#)” (*Environment and Behavior* [2008]: 777–797) describe a study of recycling among young adults in Sweden. They considered y = Recycling behavior (a numerical measure that was based on how often six types of household waste [newspapers, glass, hard plastic, soft plastic, metal, and paper] were recycled) and x = Distance to the nearest recycling facility. They reported that “a check of a plot between the variables recycling behavior and distance to the nearest recycling facility showed an approximately linear association between the two variables.” Based on this observation, a least-squares line is an appropriate way to summarize the relationship between recycling behavior and distance to nearest recycling facility.

A related article, “**Rubbish Regression and the Census Undercount**” (*Chance* [1992]: 33), describes work done by the Garbage Project at the University of Arizona. Project researchers analyzed different categories of garbage for a number of households. They were asked by the Census Bureau to see whether any of the garbage data variables were related to household size. They reported that “the weight data for different categories of garbage were plotted on graphs against data on household size, dwelling by dwelling, and the resulting scatterplots were analyzed to see in which categories the weight showed a steady, monotonic rise relative to household size.”

The researchers determined that the strongest linear relationship appeared to be that between the amount of plastic discarded and household size. The line used to summarize this relationship was stated to be $\hat{y} = 0.2815x$, where y = Household size and x = Weight (in pounds) of plastic during a 5-week collection. Note that this line has an intercept of 0. Scientists at the Census Bureau believed that this relationship would extend to entire neighborhoods, and so the amount of plastic discarded by a neighborhood could be measured (rather than having to measure house by house) and then used to approximate the neighborhood size.

An example of the use of the correlation coefficient is found in a paper describing a study of nightingales (a species of bird known for its song). For male songbirds, both physical characteristics and the quality of the song play a role in a female’s choice of a mate. The authors of the article “**Song Repertoire Is Correlated with Body Measures and Arrival Date in Common Nightingales**” (*Animal Behaviour* [2005]: 211–217) used data from $n = 20$ nightingales to reach the conclusion that there was a positive correlation between the number of different songs in a nightingale’s repertoire and both body weight ($r = 0.53$) and wing length ($r = 0.47$), and a negative correlation between repertoire size and arrival date ($r = -0.47$). This means that heavier birds tend to know more songs, as do birds with longer wings. The authors of the paper indicated that the observed correlation between repertoire size and body characteristics was unlikely to be due solely to the age of the bird, since all nightingales in the study were more than 3 years old and prior research indicates that repertoire size does not continue to increase with age after the third year. The negative correlation between repertoire size and arrival date was interpreted as meaning that male nightingales who knew more songs tended to arrive at their breeding habitats earlier than those who knew fewer songs.

A nonlinear regression was used by the authors of the paper “**Maternal Blood Manganese Levels and Infant Birth Weight**” (*Epidemiology* [2009]: 367–373) to describe the relationship between y = Birth weight and x = Maternal blood-manganese level at delivery for 470 mother-infant pairs. The paper states:

In this cross-sectional study, there was an inverted U-shaped association between maternal blood-manganese levels at delivery and birth weight in full-term infants. This suggests that both lower and higher manganese exposures are associated with lower birth weight, although the association of higher manganese with lower weight was rather weak and imprecise. This is the first epidemiologic study to provide clear evidence of a nonlinear association between maternal manganese exposure and birth weight.

The paper goes on to suggest why the relationship may be best described by a quadratic rather than a linear equation:

One possible explanation for this effect would be oxidative stress caused by high manganese levels, leading to impairment of cellular function and growth. Manganese, like iron, is a transitional metal and can catalyze oxidative cellular reactions. Exposure to high levels of iron, a metal with overlapping chemical properties to manganese, has been associated with low birth weight.

A Word to the Wise: Cautions and Limitations

There are a number of ways to get into trouble when analyzing bivariate numerical data! Here are some of the things you need to keep in mind when conducting your own analyses or when reading reports of such analyses:

1. Correlation does not imply causation. A common media blunder is to infer a cause-and-effect relationship between two variables simply because there is a strong

correlation between them. Don't fall into this trap! A strong correlation implies only that the two variables tend to vary together in a predictable way, but there are many possible explanations for why this is occurring besides one variable causing changes in the other.

For example, the article "[Ban Cell Phones? You May as Well Ban Talking Instead](#)" (*USA TODAY*, April 27, 2000) gave data that showed a strong negative correlation between the number of cell phone subscribers and traffic fatality rates. During the years from 1985 to 1998, the number of cell phone subscribers increased from 200,000 to 60,800,000, and the number of traffic deaths per 100 million miles traveled decreased from 2.5 to 1.6 over the same period. However, based on this correlation alone, the conclusion that cell phone use improves road safety is not reasonable!

Similarly, the [Calgary Herald](#) (April 16, 2002) reported that heavy and moderate drinkers earn more than light drinkers or those who do not drink. Based on the correlation between number of drinks consumed and income, the author of the study concluded that moderate drinking "causes" higher earnings. This is obviously a misleading statement, but at least the article goes on to state that "there are many possible reasons for the correlation. It could be because better-off men simply choose to buy more alcohol. Or it might have something to do with stress: Maybe those with stressful jobs tend to earn more after controlling for age, occupation, etc., and maybe they also drink more in order to deal with the stress."

2. A correlation coefficient near 0 does not necessarily imply that there is no relationship between two variables. Before such an interpretation can be given, it is important to examine a scatterplot of the data carefully. Although it may be true that the variables are unrelated, there may in fact be a strong but nonlinear relationship.
3. The least-squares line for predicting y from x is not the same line as the least-squares line for predicting x from y . The least-squares line is, by definition, the line that has the smallest possible sum of squared deviations of points from the line *in the y direction* (it minimizes $\sum(y - \hat{y})^2$). The line that minimizes the sum of squared deviations in the y direction is not generally the same as the line that minimizes the sum of the squared deviations in the x direction. So, for example, it is not appropriate to fit a line to data using $y = \text{House price}$ and $x = \text{House size}$ and then use the resulting least-squares line $\text{Price} = a + b(\text{Size})$ to predict the size of a house by substituting in a price and then solving for size. Make sure that the dependent and independent variables are clearly identified and that the appropriate line is fit.
4. Beware of extrapolation. It is dangerous to assume that a line used to describe the relationship between x and y is valid over a wider range of x values. Using the least-squares line to make predictions outside the range of x values in the data set often leads to poor predictions.
5. Be careful in interpreting the value of the slope and intercept in the least-squares line. In particular, in many instances interpreting the intercept as the value of y that would be predicted when $x = 0$ is equivalent to extrapolating beyond the range of the x values in the data set. Interpreting the intercept should be avoided unless $x = 0$ is within the range of the data.
6. Remember that the least-squares line may be the "best" line (in that it has a smaller sum of squared deviations than any other line), but that doesn't necessarily mean that the line will produce good predictions. Be cautious of predictions based on a least-squares line without any assessment of the accuracy of predictions, such as s_e and r^2 .
7. It is not enough to look at just r^2 or just s_e when assessing a linear model. These two measures address different aspects of the fit of the line. In general, we would like to have a small value for s_e (which indicates that deviations from the line tend to be small) and a large value for r^2 (which indicates that the linear relationship explains a large proportion of the variability in the y values). It is possible to have

a small s_e combined with a small r^2 or a large r^2 combined with a large s_e . Remember to consider both values.

8. The value of the correlation coefficient as well as the values for the intercept and slope of the least-squares line can be sensitive to influential observations in the data set, particularly if the sample size is small. Because potentially influential observations are those whose x values are far away from most of the x values in the data set, it is important to look for such observations when examining the scatterplot. (Another good reason for *always* starting with a plot of the data!)
9. If the relationship between two variables is not linear, it is preferable to model the relationship using a curve rather than fitting a line to the data. A plot of the residuals from the least-squares line is particularly useful in determining whether a nonlinear model would be a more appropriate choice.

EXERCISES 5.64 - 5.67

● Data set available online

- 5.64** The “Admitted Students Highlights Report 2009” prepared by [The College Board](#) for Cal Poly San Luis Obispo summarizes responses to a survey completed by 2001 new students who enrolled at Cal Poly in fall 2008 and by 2000 students who were admitted to Cal Poly for the fall 2008 term but who enrolled at other universities.

One question in the survey presented a list of “college images” (such as career-oriented and friendly) and asked students to indicate for each image whether or not they associated that image with Cal Poly. The percentage that associated an image with Cal Poly was recorded for enrolling students and for non-enrolling students for each image. For example, 61% of enrolling students but only 46% of non-enrolling students associated the image “career-oriented” with Cal Poly.

The resulting data were used to construct a scatterplot that appeared in the report. The scatterplot is reproduced in the figure at the top of the following page.

- What do you think the two dashed lines in the scatterplot represent?
- Write a short article appropriate for a student newspaper commenting on what can be learned from this scatterplot. You can assume that the scatterplot will appear with the article.

- 5.65** The following is an excerpt from a letter to the editor written by Roger Cleary that appeared in the [San Luis Obispo Tribune](#) (September 16, 2008):

The causes of poor fuel economy have nothing to do with higher highway speeds, notwithstanding all the press hoopla, including the July 19 Miami Herald claim that “There is no question that slower speeds will save gasoline,” and the July 3 statement by Drive Smarter Challenge vehicle director Deron Lovaas that, “I’m not sure whether most people make the connection between how fast they drive and how much fuel they use.”

I decided to gather the speed facts for myself using my Chevy, which comes equipped with a fuel usage driver information center, real-time read out. At a road speed of 17.5 mph, it averages 10 mpg; at 35 mph, it averages 20 mpg; and at 65 mph, it averages 30 mpg, all testing done with engine speed standardized at 2000 rpm.

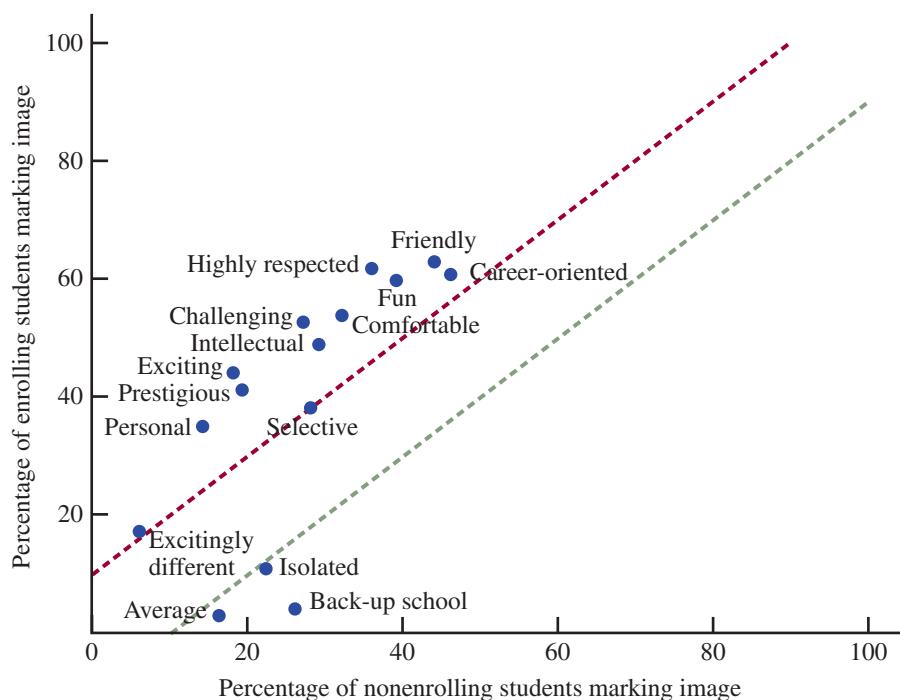
The higher the speed, the better the fuel economy. The faster you drive, the more fuel efficient you become and the more gasoline you save.

Notice that the only speeds that the letter writer provides data for are 17.5, 35, and 65 mpg. Studies of the relationship between y = Gas mileage and x = Speed have suggested that the relationship is not linear, and some have used a quadratic curve to describe the relationship between gas mileage and speed.

Write a response to Mr. Cleary that explains how his three observed data points could still be consistent with the statement that higher highway speeds lead to reduced fuel efficiency. Include a graph to support your explanation.

- 5.66** The paper “How Lead Exposure Relates to Temporal Changes in IQ, Violent Crime, and Unwed Pregnancy” ([Environmental Research](#) [2000]: 1–22) investigated whether childhood lead exposure is related to criminal behavior in young adults. Using historical data, the author paired y = Assault rate (assaults per 100,000 people) for each year from 1964 to 1997 with a measure of lead exposure (tons of gasoline lead per 1000 people) 23 years earlier. For example, the lead exposure from 1974 was paired with the assault rate from 1997.

The author chose to go back 23 years for lead exposure because the highest number of assaults are committed by people in their early twenties, and 23 years earlier would represent a time when those in this age group were infants.

Figure For Exercise 5.64

A least-squares line was used to describe the relationship between assault rate and lead exposure 23 years prior. Summary statistics given in the paper are

intercept: -24.08
slope 327.41
 r^2 0.89

Use the information provided to answer the following questions.

- What is the value of the correlation coefficient for $x = \text{Lead exposure 23 years prior}$ and $y = \text{Assault rate}$? Interpret this value. Is it reasonable to conclude that increased lead exposure is the cause of increased assault rates? Explain.
- What is the equation of the least-squares line? Use the line to predict assault rate in a year in which gasoline lead exposure 23 years prior was 0.5 tons per 1000 people.
- What proportion of year-to-year variability in assault rates can be explained by the relationship between assault rate and gasoline lead exposure 23 years earlier?
- The graph in the figure on the next page appeared in the paper. Note that this is not a scatterplot of the (x, y) pairs—it is two separate time series plots. The time scale 1941, 1942, ..., 1986 is the time scale used for the lead exposure data and the time scale 1964, 1965, ..., 2009 is used for the assault rate data. Also note that at the time the graph was constructed, assault rate data was only available through 1997. Spend a few

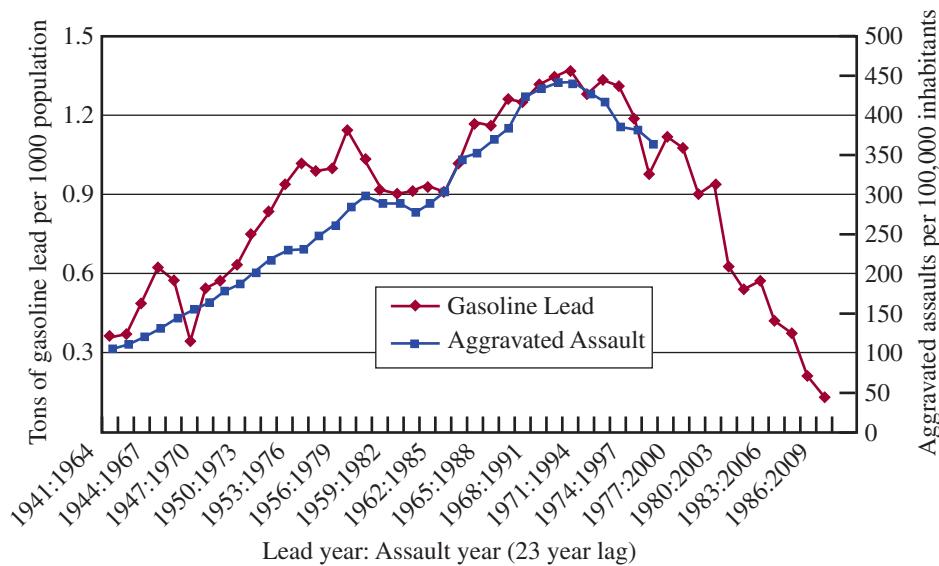
minutes thinking about the information contained in this graph and then briefly explain what aspect of this graph accounts for the reported positive correlation between assault rate and lead exposure 23 years prior.

- 5.67** The following quote is from the paper “Evaluation of the Accuracy of Different Methods Used to Estimate Weights in the Pediatric Population” (*Pediatrics* [2009]: e1045–e1051):

As expected, the model demonstrated that weight increased with age, but visual inspection of an age versus weight plot demonstrated a nonlinear relationship unless infants and children were analyzed separately. The linear coefficient for age as a predictor of weight was 6.93 in infants and 3.1 to 3.48 in children.

This quote suggests that when a scatterplot of weight versus age was constructed for all 1011 children in the study described in the paper, the relationship between $y = \text{weight}$ and $x = \text{age}$ was not linear. When the 1011 children were separated into two groups—infants (age birth to 1 year) and children (age 1 to 10 years)—and separate scatterplots were constructed, the relationship between weight and age appeared linear in each scatterplot. The slopes reported in the given quote (referred to as “the linear coefficient”) are expressed in kg/year.

Briefly explain why the relationship between weight and age in the scatterplot for the combined group would appear nonlinear.

Figure for Exercise 5.66

CHAPTER ACTIVITIES

ACTIVITY 5.1 AGE AND FLEXIBILITY

Materials needed: Yardsticks.

In this activity, you will investigate the relationship between age and a measure of flexibility. Flexibility will be measured by asking a person to bend at the waist as far as possible, extending his or her arms toward the floor. Using a yardstick, measure the distance from the floor to the fingertip closest to the floor.

- Age and the measure of flexibility just described will be measured for a group of individuals. Our goal is to determine whether there is a relationship between age and this measure of flexibility. What are two reasons why it would not be a good idea to use just the students in your class as the subjects for your study?

- Working as a class, decide on a reasonable way to collect data on the two variables of interest.
- After your class has collected appropriate data, use them to construct a scatterplot. Comment on the interesting features of the plot. Does it look like there is a relationship between age and flexibility?
- If there appears to be a relationship between age and flexibility, fit a model that is appropriate for describing the relationship.
- In the context of this activity,* write a brief description of the danger of extrapolation.

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
Scatterplot	A graph of bivariate numerical data in which each observation (x, y) is represented as a point located with respect to a horizontal x axis and a vertical y axis.	Principle of least squares	The method used to select a line that summarizes an approximate linear relationship between x and y . The least-squares line is the line that minimizes the sum of the squared errors (vertical deviations) for the points in the scatterplot.
Sample correlation coefficient $r = \frac{\sum z_x z_y}{n - 1}$	A measure of the extent to which sample x and y values are linearly related. Values close to 1 or -1 indicate a strong linear relationship.	$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$	The slope of the least-squares line.

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
$a = \bar{y} - b\bar{x}$	The intercept of the least-squares line.	Total sum of squares $SSTo = \sum(y - \bar{y})^2$	The sum of squared deviations from the sample mean is a measure of total variation in the observed y values.
Predicted (fitted) values $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$	Obtained by substituting the x value for each observation in the data set into the least-squares line; $\hat{y}_1 = a + bx_1, \dots, \hat{y}_n = a + bx_n$	Coefficient of determination $r^2 = 1 - \frac{SSResid}{SSTo}$	The proportion of variation in observed y 's that can be explained by an approximate linear relationship.
Residuals	Obtained by subtracting each predicted value from the corresponding observed y value: $y_1 - \hat{y}_1, \dots, y_n - \hat{y}_n$. These are the vertical deviations from the least-squares line.	Standard deviation about the least-squares line $s_e = \sqrt{\frac{SSResid}{n - 2}}$	The size of a “typical” deviation from the least-squares line.
Residual plot	Scatterplot of the $(x, \text{residual})$ pairs. Isolated points or a pattern of points in a residual plot are indicative of potential problems.	Transformation	A simple function of the x and/or y variable, which is then used in a regression.
Residual (error) sum of squares $SSResid = \sum(y - \hat{y})^2$	The sum of the squared residuals is a measure of y variation that cannot be attributed to an approximate linear relationship (unexplained variation).		

CHAPTER REVIEW Exercises 5.68 - 5.79

● Data set available online

- 5.68** ● The accompanying data represent x = Amount of catalyst added to accelerate a chemical reaction and y = Reaction time:

x	1	2	3	4	5
y	49	46	41	34	25

- a. Calculate the value of the correlation coefficient, r . Does the value of r suggest a strong linear relationship?
b. Construct a scatterplot. From the plot, does the word *linear* provide the most effective description of the relationship between x and y ? Explain.

- 5.69** ● The paper “A Cross-National Relationship Between Sugar Consumption and Major Depression?”

(*Depression and Anxiety* [2002]: 118–120) concluded that there was a correlation between refined sugar consumption (calories per person per day) and annual rate of major depression (cases per 100 people) based on data from six countries. The following data were read from a graph that appeared in the paper:

Country	Sugar Consumption	Depression Rate
Germany	375	5.0
Canada	390	5.2
New Zealand	480	5.7

- a. Calculate and interpret the correlation coefficient for this data set.
b. Is it reasonable to conclude that increasing sugar consumption leads to higher rates of depression? Explain.
c. Do you have any concerns about this study that would make you hesitant to generalize these conclusions to other countries?

- 5.70** ● The following data on x = Score on a measure of test anxiety and y = Exam score for a sample of $n = 9$ students are consistent with summary quantities given in the paper “Effects of Humor on Test Anxiety and Performance” (*Psychological Reports* [1999]: 1203–1212):

x	23	14	14	0	17	20	20	15	21
y	43	59	48	77	50	52	46	51	51

Higher values for x indicate higher levels of anxiety.

- a. Construct a scatterplot, and comment on the features of the plot.

Country	Sugar Consumption	Depression Rate
Korea	150	2.3
United States	300	3.0
France	350	4.4

(continued)

- b.** Does there appear to be a linear relationship between the two variables? How would you characterize the relationship?
- c.** Calculate the value of the correlation coefficient. Is the value of r consistent with your answer to Part (b)?
- d.** Is it reasonable to conclude that test anxiety caused poor exam performance? Explain.

5.71 The paper “**Effects of Canine Parvovirus (CPV) on Gray Wolves in Minnesota**” (*Journal of Wildlife Management* [1995]: 565–570) summarized a regression of y = Percentage of pups in a capture on x = Percentage of CPV prevalence among adults and pups. The equation of the least-squares line, based on $n = 10$ observations, was $\hat{y} = 62.9476 - 0.54975x$, and $r^2 = 0.57$.

- a.** One observation was (25, 70). What is the corresponding residual?
- b.** What is the value of the sample correlation coefficient?
- c.** Suppose that $SSTo = 2520.0$ (this value was not given in the paper). What is the value of s_e ?

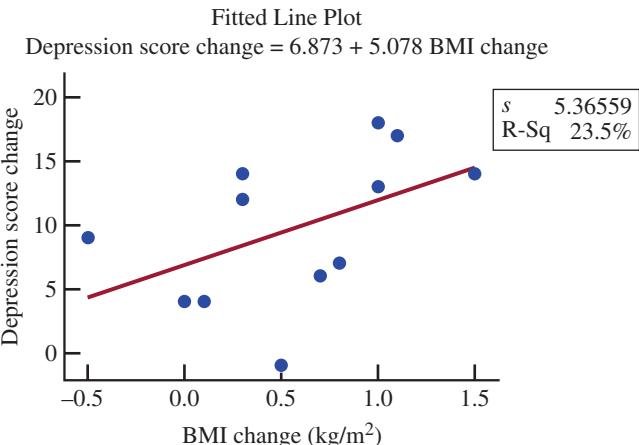
5.72 The paper “**Depression, Body Mass Index, and Chronic Obstructive Pulmonary Disease—A Holistic Approach**” (*International Journal of COPD* [2016]:239–249) gave data on change in body mass index (BMI in kilograms/meter²) and change in a measure of depression for patients suffering from depression who participated in a pulmonary rehabilitation program. The data in the table below are a subset of the data given in the paper and are approximate values read from a scatterplot in the paper.

BMI Change (kg/m ²)	Depression Score Change
0.5	-1
-0.5	9
0.0	4
0.1	4
0.7	6
0.8	7
1.0	13
1.5	14
1.1	17
1.0	18
0.3	12
0.3	14

- a.** Find the equation of the least-squares line that would allow you to predict change in depression score based on change in BMI.

- b.** Find the values of r^2 and s_e . Based on these values, do you think that the least-squares line does a good job of describing the relationship between change in depression score and change in BMI? Explain.

- c.** The following graph is a scatterplot of these data. The least-squares line is also shown. Which observations are outliers? Do the observations with the largest residuals correspond to the patients with the largest change in BMI?



5.73 The paper “**Aspects of Food Finding by Wintering Bald Eagles**” (*The Auk* [1983]: 477–484) examined the relationship between the time that eagles spend aerially searching for food (indicated by the percentage of eagles soaring) and relative food availability. The accompanying data were taken from a scatterplot that appeared in this paper. Use x to denote salmon availability and y to denote the percentage of eagles in the air.

x	0	0	0.2	0.5	0.5	1.0
y	28.2	69.0	27.0	38.5	48.4	31.1
x	1.2	1.9	2.6	3.3	4.7	6.5
y	26.9	8.2	4.6	7.4	7.0	6.8

- a.** Draw a scatterplot for this data set. Would you describe the pattern in the plot as linear or curved?
- b.** One possible transformation that might lead to a more nearly linear plot involves taking the square root of both the x and y values. Explain why this might be a reasonable transformation.
- c.** Construct a scatterplot using the variables \sqrt{x} and \sqrt{y} . Is this scatterplot more nearly linear than the scatterplot in Part (a)?
- d.** Suggest another transformation that might be used to straighten the original plot.

5.74 Data on salmon availability (x) and the percentage of eagles in the air (y) were given in the previous exercise.

- a. Calculate the correlation coefficient for these data.
- b. Because the scatterplot of the original data appeared curved, transforming both the x and y values by taking square roots was suggested. Calculate the correlation coefficient for the variables \sqrt{x} and \sqrt{y} . How does this value compare with that calculated in Part (a)? Does this indicate that the transformation was successful in straightening the plot?
- 5.75** No tortilla chip lover likes soggy chips, so it is important to find characteristics of the production process that produce chips with an appealing texture. The accompanying data on x = Frying time (in seconds) and y = Moisture content (%) appeared in the paper, “**Thermal and Physical Properties of Tortilla Chips as a Function of Frying Time**” (*Journal of Food Processing and Preservation* [1995]: 175–189):

Frying time (x): 5 10 15 20 25 30 45 60
 Moisture content (y): 16.3 9.7 8.1 4.2 3.4 2.9 1.9 1.3

- a. Construct a scatterplot of these data. Does the relationship between moisture content and frying time appear to be linear?
- b. Transform the y values using $y' = \log(y)$ and construct a scatterplot of the (x, y') pairs. Does this scatterplot look more nearly linear than the one in Part (a)?
- c. Find the equation of the least-squares line that describes the relationship between y' and x .
- d. Use the least-squares line from Part (c) to predict moisture content for a frying time of 35 minutes.
- 5.76** The article “**Reduction in Soluble Protein and Chlorophyll Contents in a Few Plants as Indicators of Automobile Exhaust Pollution**” (*International Journal of Environmental Studies* [1983]: 239–244) reported the following data on x = Distance from a highway (in meters) and y = Lead content of soil at that distance (in parts per million):

x	0.3	1	5	10	15	20
y	62.75	37.51	29.70	20.71	17.65	15.41
x	25	30	40	50	75	100
y	14.15	13.50	12.11	11.40	10.85	10.85

- a. Use a statistical computer package to construct scatterplots of y versus x , y versus $\log(x)$, $\log(y)$ versus $\log(x)$, and $\frac{1}{y}$ versus $\frac{1}{x}$.

- b. Which transformation considered in Part (a) does the best job of producing an approximately linear relationship? Use the selected transformation to predict lead content when distance is 25 m.

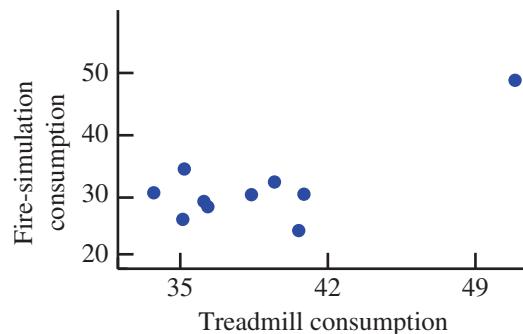
- 5.77** The following quote is from the paper “**The Weight of the Bottle as a Possible Extrinsic Clue with Which to Estimate the Price (and Quality) of the Wine? Observed Correlations**” (*Food Quality and Preference* [2012]: 41–45):

The weight of the wine bottles was positively correlated ($r = 0.12; p < 0.001$) with the price of the wines

- a. Does the value of the correlation coefficient indicate that the relationship is weak, moderate, or strong?
- b. What is the value of r^2 ? Write a sentence that gives an interpretation of this value.

- 5.78** An accurate assessment of oxygen consumption provides important information for determining energy expenditure requirements for physically demanding tasks. The paper “**Oxygen Consumption During Fire Suppression: Error of Heart Rate Estimation**” (*Ergonomics* [1991]: 1469–1474) reported on a study in which x = Oxygen consumption (in milliliters per kilogram per minute) during a treadmill test was determined for a sample of 10 firefighters. Then y = Oxygen consumption at a comparable heart rate was measured for each of the 10 individuals while they performed a fire-suppression simulation. This resulted in the following data and scatterplot:

Firefighter	1	2	3	4	5
x	51.3	34.1	41.1	36.3	36.5
y	49.3	29.5	30.6	28.2	28.0
Firefighter	6	7	8	9	10
x	35.4	35.4	38.6	40.6	39.5
y	26.3	33.9	29.4	23.5	31.6



- a. Does the scatterplot suggest an approximate linear relationship?

- b.** The investigators fit a least-squares line. The resulting Minitab output is given in the following:

The regression equation is
firecon = 211.4 + 1.09 treadcon

Predictor	Coef	Stdevt-ratio	p	
Constant	-11.37	12.46	-0.91	0.388
treadcon	1.0906	0.3181	3.43	0.009
s = 4.70	R-sq = 59.5%	R-sq(adj) = 54.4%		

Predict fire-simulation consumption when treadmill consumption is 40.

- c.** How effectively does a straight line summarize the relationship?

- d.** Delete the first observation, (51.3, 49.3), and calculate the new equation of the least-squares line and the value of r^2 . What do you conclude? (Hint: For the original data, $\Sigma x = 388.8$, $\Sigma y = 310.3$, $\Sigma x^2 = 15,338.54$, $\Sigma xy = 12,306.58$, and $\Sigma y^2 = 10,072.41$.)

- 5.79** Consider the four (x, y) pairs $(0, 0)$, $(1, 1)$, $(1, -1)$, and $(2, 0)$.

- What is the value of the sample correlation coefficient r ?
- If a fifth observation is made at the value $x = 6$, find a value of y for which $r > 0.5$.
- If a fifth observation is made at the value $x = 6$, find a value of y for which $r < 0.5$.

TECHNOLOGY NOTES

Correlation

TI-83/84

Note: Before beginning this chapter, press the **2nd** key then the **0** key and scroll down to the entry DiagnosticOn. Press **ENTER** twice. After doing this, the correlation coefficient r will appear as output with the least-squares line.

- Enter the data for the independent variable into **L1** (to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
- Input the data for the dependent variable into **L2**
- Press the **STAT** key
- Highlight **CALC** then select **LinReg(a+bx)** and press **ENTER**
- Press the **2nd** key then the **1** key
- Press ,
- Press the **2nd** key then the **2** key
- Press **ENTER**

TI-Nspire

- Enter the data for the independent variable into a data list (to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

- Enter the data for the dependent variable into a separate data list
- Press the **menu** key and select **4:Statistics** then **1:Stat Calculations** then **3:Linear Regression(mx+b)...** and press **enter**
- For **X List**: select the column with the independent variable data from the drop-down menu
- For **Y List**: select the column with the dependent variable data from the drop-down menu
- Press **OK**

Note: You may need to scroll to view the correlation coefficient in the list of output.

JMP

- Input the data for the dependent variable into one column
- Input the data for the independent variable into another column
- Click **Analyze** then select **Multivariate Methods** then select **Multivariate**
- Click and drag the column name containing the dependent variable from the box under **Select Columns** to the box next to **Y, Columns**
- Click and drag the column name containing the independent variable from the box under **Select Columns** to the box next to **Y, Columns**
- Click **OK**

Note: This produces a table of correlations. The correlation between the two variables can be found in the first row, second column.

MINITAB

- Input the data for the dependent variable into the first column
- Input the data for the independent variable into the second column
- Select **Stat** then **Basic Statistics** then **Correlation...**
- Double-click each column name in order to move it to the box under **Variables**:
- Click **OK**

SPSS

- Input the data for the dependent variable into the first column
- Input the data for the independent variable into the second column
- Select **Analyze** then choose **Correlate** then choose **Bivariate...**

4. Highlight the name of both columns by holding the **ctrl** key and clicking on each name
5. Click the arrow button to move both variables to the **Variables** box
6. Click **OK**

Note: This produces a table of correlations. The correlation between the two variables can be found in the first row, second column.

Note: The correlation can also be produced by following the steps to produce the regression equation.

Excel 2007

1. Input the data into two separate columns
2. Click the **Data** ribbon and select **Data Analysis**
3. **Note:** If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.
4. Select **Correlation** from the dialog box and click **OK**
5. Click in the box next to **Input Range:** and select BOTH columns of data (if you input and selected titles for the columns, click the box next to **Labels in First Row**)
6. Click **OK**

Note: The correlation between the variables can be found in the first column, second row of the table that is output.

Regression

TI-83/84

1. Enter the data for the independent variable into **L1** (to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Input the data for the dependent variable into **L2**
3. Press the **STAT** key
4. Highlight **CALC** then select **LinReg(a+bx)** and press **ENTER**
5. Press the **2nd** key then the **1** key
6. Press,
7. Press the **2nd** key then the **2** key
8. Press **ENTER**

TI-Nspire

1. Enter the data for the independent variable into a data list (to access data lists select the spreadsheet option and press **enter**)
- Note:** Be sure to title the list by selecting the top row of the column and typing a title.
2. Enter the data for the dependent variable into a separate data list
3. Press the **menu** key and select **4:Statistics** then **1:Stat Calculations** then **3:Linear Regression(mx+b)...** and press **enter**
4. For **X List:** select the column with the independent variable data from the drop-down menu
5. For **Y List:** select the column with the dependent variable data from the drop-down menu
6. Press **OK**

JMP

1. Enter the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Click **Analyze** then select **Fit Y by X**
4. Click and drag the column name containing the dependent data from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the column name containing the independent data from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Bivariate Fit...**
8. Click **Fit Line**

MINITAB

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Stat** then **Regression** then **Regression...**
4. Highlight the name of the column containing the dependent variable and click **Select**
5. Highlight the name of the column containing the independent variable and click **Select**
6. Click **OK**

Note: You may need to scroll up in the Session window to view the regression equation.

SPSS

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Analyze** then choose **Regression** then choose **Linear...**
4. Highlight the name of the column containing the dependent variable
5. Click the arrow button next to the **Dependent** box to move the variable to this box
6. Highlight the name of the column containing the independent variable
7. Click the arrow button next to the **Independent** box to move the variable to this box
8. Click **OK**

Note: The regression coefficients can be found in the **Coefficients** table. The intercept value can be found in the first column of the (Constant) row. The value of the slope can be found in the first column of the row labeled with the independent variable name.

Excel 2007

1. Input the data into two separate columns
2. Click the **Data** ribbon and select **Data Analysis**
3. **Note:** If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.
4. Select **Regression** from the dialog box and click **OK**
5. Click in the box next to **Y:** and select the dependent variable data

6. Click in the box next to **X:** and select the independent variable data (if you input and selected titles for BOTH columns, check the box next to **Labels**)
7. Click **OK**

Note: The regression coefficients can be found in the third table under the **Coefficients** column.

Residuals

TI 83/84

Note: When the TI 83/84 performs a regression analysis, the residuals are automatically stored in a list called "RESID." To prevent loss of the residuals, copy them into a named list.

1. Press the **2nd** key then the **List** key. **Select RESID.** (You may have to scroll down to find it.) Your screen should now show "**LRESID**."
2. Click on **STO**, then the **2nd** key, and then **L3**.
3. Click on **Enter**.

TI-Nspire

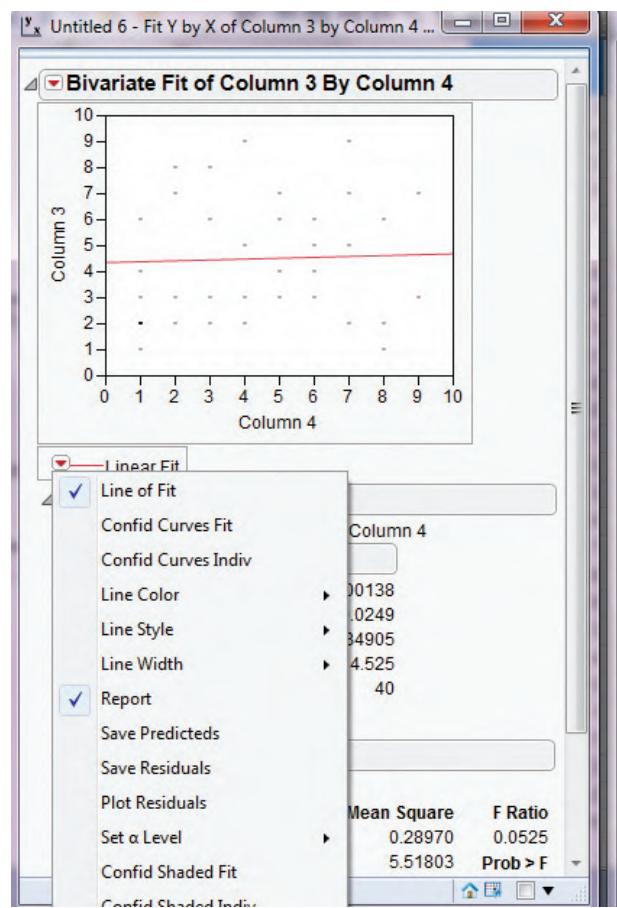
1. Enter the data for the independent variable into a data list (to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Enter the data for the dependent variable into a separate data list
3. Press the menu key and select **4:Statistics** then **1:Stat Calculations** then **3:Linear Regression(mx+b)...** and press **enter**
4. For **X List:** select the column with the independent variable data from the drop-down menu
5. For **Y List:** select the column with the dependent variable data from the drop-down menu
6. Press **OK**

JMP

1. Enter the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Click **Analyze** then select **Fit Y by X**
4. Click and drag the column name containing the dependent data from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the column name containing the independent data from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Bivariate Fit...**
8. Click **Fit Line**
9. Click the red arrow next to **Linear Fit**
10. Click **Save Residuals**



MINITAB

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Stat** then **Regression** then **Regression...**
4. Highlight the name of the column containing the dependent variable and click **Select**
5. Highlight the name of the column containing the independent variable and click **Select**
6. Click **Storage...**
7. Check the box next to **Residuals**
8. Click **OK**
9. Click **OK**

SPSS

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Analyze** then choose **Regression** then choose **Linear...**
4. Highlight the name of the column containing the dependent variable
5. Click the arrow button next to the **Dependent** box to move the variable to this box
6. Highlight the name of the column containing the independent variable
7. Click the arrow button next to the **Independent** box to move the variable to this box

8. Click on the **Save...** button
9. Click the check box next to **Unstandardized** under the **Residuals** section
10. Click **Continue**
11. Click **OK**

Note: The residuals will be saved in the SPSS worksheet in a new column.

Excel 2007

1. Input the data into two separate columns
2. Click the **Data** ribbon and select **Data Analysis**
Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.
3. Select **Regression** from the dialog box and click **OK**
4. Click in the box next to **Y:** and select the dependent variable data
5. Click in the box next to **X:** and select the independent variable data (if you input and selected titles for BOTH columns, check the box next to **Labels**)
6. Check the box next to **Residuals** under the **Residuals** section of the dialog box
7. Click **OK**

Residual Plot

TI-83/84

The TI-83/84 does not have the functionality to produce a residual plot automatically. After using Linreg(a+bx), select 2nd, Statplot. Set the first plot as a scatter with XList: L1 and YList: RESID. Select Zoom, Stats, and a residual plot is displayed.

TI-Nspire

The TI-Nspire does not have the functionality to produce a residual plot automatically.

JMP

1. Begin by saving the residuals as described in the previous section
2. Form a scatterplot of the independent variable versus the residuals using the procedures described in Chapter 2 for scatterplots

MINITAB

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column
3. Select **Stat** then **Regression** then **Regression...**
4. Highlight the name of the column containing the dependent variable and click **Select**
5. Highlight the name of the column containing the independent variable and click **Select**
6. Click **Graphs...**
7. Click the box under **Residuals versus the variables**:
8. Double-click the name of the independent variable

9. Click **OK**
10. Click **OK**

SPSS

1. Begin by saving the residuals as described in the previous section
2. Form a scatterplot of the independent variable versus the residuals using the procedures described in Chapter 2 for scatterplots

Excel 2007

1. Input the data into two separate columns
2. Click the **Data** ribbon and select **Data Analysis**
Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.
3. Select **Regression** from the dialog box and click **OK**
4. Click in the box next to **Y:** and select the dependent variable data
5. Click in the box next to **X:** and select the independent variable data (if you input and selected titles for BOTH columns, check the box next to **Labels**)
6. Check the box next to **Residuals Plots** under the **Residuals** section of the dialog box
7. Click **OK**

Transforming Variables

TI 83/84

1. Enter the data for the variable you wish to transform into **L1** (to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Click on **Stat** and select **Edit**
3. Using the arrow keys (directly above the **VARS** and **CLEAR** Keys) position the cursor at the top of **List2**. That is, the cursor should be covering the “**L2**.” In addition, “**L2=**” should appear in the bottom line

Now choose the transformation you wish to make. For example, to set the values of **L2** equal to the natural log of **L1**, do the following:

4. Click the **LN** button; the formula at the bottom of the screen will change to “**L2=ln(**”
5. Click on the **2nd** key, then the **STO** key, and finally **L1**; the formula at the bottom of the screen will change to “**L2=ln(L1)**”
6. Click on the right parenthesis key; the formula at the bottom of the screen will change to “**L2=ln(L1)**”
7. Click on **Enter**

JMP

1. Enter the pre-transformed data into **Column1**
2. Double-click at the top of **Column2**. A window with properties for **Column2** will pop up
3. In that window, click on **Column Properties**, select **Formula**, and click on **Edit Formula**

Now choose the transformation you wish to make. For example, to set the values of Column2 equal to the natural logarithm of Column1, do the following:

4. Click on Transcendental, then click on **Log**; the formula box will be updated to “**Log()**”
5. In the TableColumns panel, click on **Column1**; the formula box will be updated to “**Log(Column1)**”
6. Click **Ok** and close the column properties window
7. Respond with “Yes” to the “**Apply changes**” question

MINITAB

1. Enter the pre-transformed data into column “**C1**”
2. Click on **Calc** in the main menu, and then **Calculator**
3. Type **C2** in the “Store result in variable” box

Now choose the transformation you wish to make. For example, to set the values of **C2** equal to the natural log of **C1**:

4. Click on Natural log (log bas e) in the Functions drop-down menu and click on **Select**. The “Expression:” window will be updated to “**LN(number)**”

5. Double-click on **C1** in the variables list panel; the “Expression:” window will be updated to “**LN(C1)**”
6. Click **OK**

Automatic Nonlinear Regression

TI 83/84

1. Enter the data for the independent variable into **L1** (to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Input the data for the dependent variable into **L2**
3. Press the **STAT** key
4. Select **CALC**

Now choose the nonlinear option you wish; for example, to perform exponential regression, do the following:

5. Select **ExpReg** and press **ENTER**
6. Press the **2nd** key then the **1** key
7. Press,
8. Press the **2nd** key then the **2** key
9. Press **ENTER**

CUMULATIVE REVIEW EXERCISES

CR5.1 - CR5.15

● Data set available online

CR5.1 The article “[Rocker Shoe Put to the Test: Can it Really Walk the Walk as a Way to Get in Shape?](#)” (*USA TODAY*, October 12, 2009) describes claims made by Skechers about Shape-Ups, a shoe line introduced in 2009. These curved-sole sneakers are supposed to help you “get into shape without going to the gym” according to a Skechers advertisement. Briefly describe how you might design a study to investigate this claim. Include how you would select subjects and what variables you would measure. Is the study you designed an observational study or an experiment?

CR5.2 Data from a survey of 1046 adults age 50 and older were summarized in the [AARP Bulletin \(November 2009\)](#). The following table gives relative frequency distributions of the responses to the question, “How much do you plan to spend for holiday gifts this year?” for respondents age 50 to 64 and for respondents age 65 and older.

Construct a histogram for each of the two age groups and comment on the differences between the two age groups. (Notice that the interval widths in the relative frequency distribution are not the same, so you shouldn’t use relative frequency on the y-axis for your histograms.)

Amount Plan to Spend	Relative Frequency for Age Group 50 to 64	Relative Frequency for Age Group 65 and Older
less than \$100	0.20	0.36
\$100 to <\$200	0.13	0.11
\$200 to <\$300	0.16	0.16
\$300 to <\$400	0.12	0.10
\$400 to <\$500	0.11	0.05
\$500 to <\$1000	0.28	0.22

CR5.3 The graph in the figure at the top of the following page appeared in the report [“Testing the Waters 2009” \(Natural Resources Defense Council\)](#). Spend a few minutes looking at the graph and reading the caption that appears with the graph. Briefly explain how the graph supports the claim that discharges of polluted storm water may be responsible for increased illness levels.

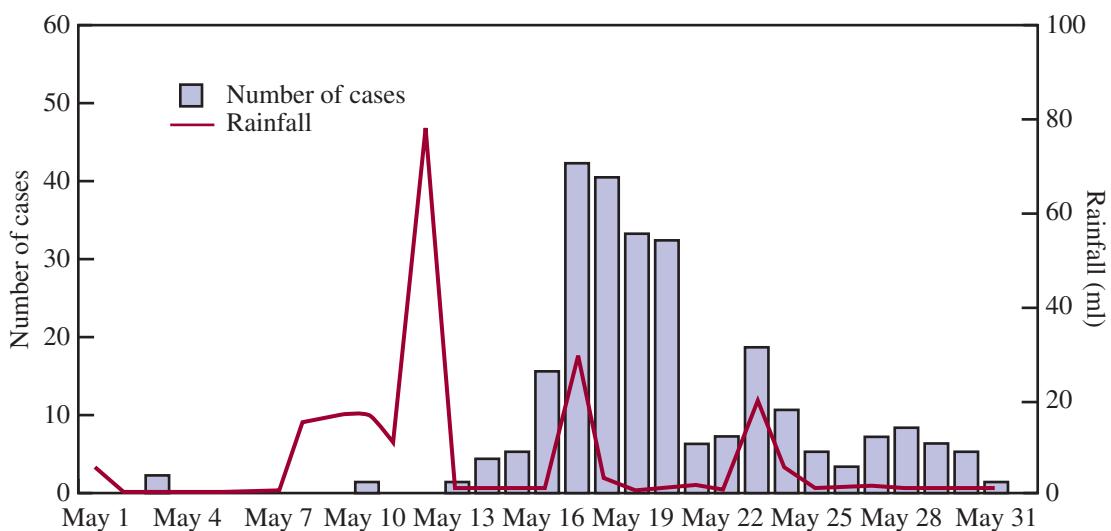
CR5.4 The cost of Internet access was examined in the report [“Home Broadband Adoption 2009” \(pewinternet.org\)](#). In 2009, the mean and median amount paid monthly for service for broadband users was reported as \$39.00 and \$38.00, respectively. For dial-up users, the mean and median amount paid monthly were \$26.60 and \$20.00, respectively. What do the values of the mean and median tell you about the shape of the distribution of monthly amount paid for broadband users? For dial-up users?

CR5.5 Each year the Harris Poll surveys Americans on a number of issues. It uses responses to several questions to calculate a “Happiness” index that measures overall happiness. The article [“Latest Happiness Index Reveals American Happiness at All-Time Low,” theharrispoll.com /health-and-life/American-Happiness_at-All-Time-Low.html, retrieved April 21, 2017](#)) included the happiness index for the 7 years between 2008 and 2016. Also included in the article were the percentages of people who responded “somewhat agree” or “strongly agree” to the following statements:

- Statement 1 (happy with life) 1: At this time, I’m generally happy with my life.
- Statement 2 (won’t benefit): I won’t get much benefit from the things that I do anytime soon.

Figure for Exercise CR5.3

Influence of heavy rainfall on occurrence of *E. coli* infections.



The graph shows the relationship between unusually heavy rainfall and the number of confirmed cases of *E. coli* infection that occurred during a massive disease outbreak in Ontario, Quebec, in May 2000. The incubation period for *E. coli* is usually 3 to 4 days, which is consistent with the lag between extreme precipitation events and surges in the number of cases.

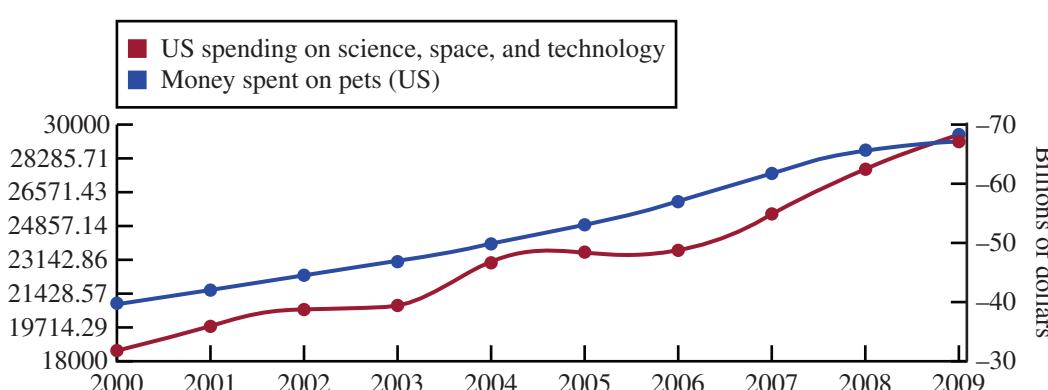
Year	Happiness Index	Happy with Life Statement (percentage somewhat or strongly agree)	Won't Benefit Statement (percentage somewhat or strongly agree)
2008	35	83	32
2009	35	81	38
2010	33	80	36
2011	33	80	38
2013	33	77	42
2015	34	82	36
2016	31	81	41

- Calculate the value of the correlation coefficient for Happiness index and the response to the Happy with life statement.
- Calculate the value of the correlation coefficient for Happiness index and the response to the Won't benefit statement.

- Is there a stronger relationship between Happiness index and the response to Statement 1 or between Happiness index and the response to Statement 2?
- Write a few sentences describing the relationships between Happiness index and the responses to the two statements.

CR5.6 The amount of money spent each year on science, space, and technology in the United States (in millions of dollars) and the amount of money spent on pets in the United States (in billions of dollars) for the years 2000 to 2009 were used to construct the graph below. (The data are from the web site tylervigen.com/spurious-correlations, accessed August 28, 2016).

Based on these time series plots, would the correlation coefficient between amount spent on science, space, and technology and the amount spent on pets be positive or negative? Weak or strong? What aspect of the time series plots support your answer?



CR5.7 Below are the data used to construct the time series plots in the previous exercise. Calculate the value of the correlation coefficient for the amount spent on science, space, and technology and the amount spent on pets. Explain how this value is consistent with your answer to the previous exercise.

Year	Amount Spent on Science, Space, and Technology (millions of dollars)	Amount Spent on Pets (billions of dollars)
2000	18,594	39.7
2001	19,753	41.9
2002	20,734	44.6
2003	20,831	48.8
2004	23,029	49.8
2005	23,597	53.1
2006	23,584	56.9
2007	25,525	61.8
2008	27,731	65.7
2009	29,449	67.1

CR5.8 In August 2009, Harris Interactive released the results of the “Great Schools” survey. In this survey, 1086 parents of children attending a public or private school were asked approximately how much they had spent on school supplies over the last school year. For this sample, the mean amount spent was \$235.20 and the median amount spent was \$150.00. What does the large difference between the mean and median tell you about this data set?

CR5.9 • The paper “Total Diet Study Statistics on Element Results” (Food and Drug Administration, April 25, 2000) gave information on sodium content for various types of foods. Twenty-six tomato catsups were analyzed. Data consistent with summary quantities given in the paper were

Sodium content (mg/kg)

12,148	10,426	10,912	9116	13,226	11,663
11,781	10,680	8457	10,788	12,605	10,591
11,040	10,815	12,962	11,644	10,047	10,478
10,108	12,353	11,778	11,092	11,673	8758
11,145	11,495				

Calculate the values of the quartiles and the interquartile range.

CR5.10 • The paper referenced in Exercise CR5.9 also gave data on sodium content (in milligrams per kilogram) of 10 chocolate puddings made from instant mix:

3099	3112	2401	2824	2682	2510	2297
3959	3068	3700				

- Calculate the mean, the standard deviation, and the interquartile range for sodium content of these chocolate puddings.
- Based on the interquartile range, is there more or less variability in sodium content for the chocolate pudding

data than for the tomato catsup data of Exercise CR5.9?

CR5.11 • A report from Texas Transportation Institute (Texas A&M University System, 2005) on congestion reduction strategies looked into the extra travel time (due to traffic congestion) for commute travel per traveler per year in hours for different urban areas. Below are the data for urban areas that had a population of over 3 million for the year 2002.

Urban Area	Extra Hours per Traveler per Year
Los Angeles	98
San Francisco	75
Washington, DC	66
Atlanta	64
Houston	65
Dallas, Fort Worth	61
Chicago	55
Detroit	54
Miami	48
Boston	53
New York	50
Phoenix	49
Philadelphia	40

- Calculate the mean and median values for extra travel hours. Based on the values of the mean and median, is the distribution of extra travel hours likely to be approximately symmetric, positively skewed, or negatively skewed?
- Construct a boxplot that shows outliers for these data and comment on any interesting features of the plot.

CR5.12 • The paper “Relationship Between Blood Lead and Blood Pressure Among Whites and African Americans” (a technical report published by Tulane University School of Public Health and Tropical Medicine, 2000) gave summary quantities for blood lead level (in micrograms per deciliter) for a sample of whites and a sample of African Americans. Data consistent with the given summary quantities follow:

Whites	8.3	0.9	2.9	5.6	5.8	5.4	1.2
	1.0	1.4	2.1	1.3	5.3	8.8	6.6
	5.2	3.0	2.9	2.7	6.7	3.2	
African	4.8	1.4	0.9	10.8	2.4	0.4	5.0
Americans	5.4	6.1	2.9	5.0	2.1	7.5	3.4
	13.8	1.4	3.5	3.3	14.8	3.7	

- Calculate the values of the mean and the median for blood lead level for the sample of African Americans.
- Which of the mean or the median is larger? What characteristic of the data set explains the relative values of the mean and the median?
- Construct a comparative boxplot for blood lead level for the two samples.
- Write a few sentences comparing the blood lead level distributions for the two samples.

CR5.13 • Cost-to-charge ratios (the percentage of the amount billed that represents the actual cost) for 11 Oregon hospitals of similar size were reported separately for inpatient and outpatient services. The data are shown in the following table.

Hospital	Cost-to-Charge Ratio	
	Inpatient	Outpatient
Blue Mountain	80	62
Curry General	76	66
Good Shepherd	75	63
Grande Ronde	62	51
Harney District	100	54
Lake District	100	75
Pioneer	88	65
St. Anthony	64	56
St. Elizabeth	50	45
Tillamook	54	48
Wallowa Memorial	83	71

- a. Does there appear to be a strong linear relationship between the cost-to-charge ratio for inpatient and outpatient services? Justify your answer based on the value of the correlation coefficient and examination of a scatterplot of the data.
- b. Are any unusual features of the data evident in the scatterplot?
- c. Suppose that the observation for Harney District was removed from the data set. Would the correlation

coefficient for the new data set be greater than or less than the one computed in Part (a)? Explain.

CR5.14 In the article “Reproductive Biology of the Aquatic Salamander *Amphiuma tridactylum* in Louisiana” (*Journal of Herpetology* [1999]: 100–105), 14 female salamanders were studied. Using regression, the researchers predicted $y = \text{Clutch size}$ (number of salamander eggs) from $x = \text{Snout-vent length}$ (in centimeters) as follows:

$$\hat{y} = -147 + 6.175x$$

For the salamanders in the study, the range of snout-vent lengths was approximately 30 to 70 cm.

- a. What is the value of the y intercept of the least-squares line?
- b. What is the value of the slope of the least-squares line? Interpret the slope in the context of this problem.
- c. Would you be reluctant to predict the clutch size when snout-vent length is 22 cm? Explain.

CR5.15 The previous exercise gave the least-squares line for predicting $y = \text{Clutch size}$ from $x = \text{Snout-vent length}$ (“Reproductive Biology of the Aquatic Salamander *Amphiuma tridactylum* in Louisiana,” *Journal of Herpetology* [1999]: 100–105). The paper also reported $r^2 = 0.7664$ and SSTo = 43,951.

- a. Interpret the value of r^2 .
- b. Find and interpret the value of s_e (the sample size was $n = 14$).

6 Probability



Doug Menuez/Forrester Images/Getty Images

LEARNING OBJECTIVES

Students will understand:

- Basic properties of probabilities.
- What it means for two events to be mutually exclusive.
- What it means for two events to be independent.
- The difference between an unconditional and a conditional probability.

Students will be able to:

- Interpret a probability in context as a long-run relative frequency of occurrence.
- Identify the sample space for a given chance experiment.
- Calculate the probability of events, including events that are the union of, the intersection of, or the complement of other events.
- Calculate conditional probabilities.
- Estimate probabilities empirically and using simulation.

We make decisions based on uncertainty every day. Should you buy an extended warranty for your new laptop? It depends on the likelihood that it will fail during the warranty period. Should you allow 45 minutes to get to your 8 a.m. class, or is 35 minutes enough? From experience, you may know that most mornings you can drive to school and park in 25 minutes or less. Most of the time, the walk from your parking space to class is 5 minutes or less. But how often will the drive to school or the walk to class take longer than you expect? When it takes longer than usual to drive to campus, is it more likely that it will also take longer to walk to class? less likely? Or are the driving and walking times unrelated?

Some questions involving uncertainty are more serious: If an artificial heart has four key parts, how likely is each one to fail? How likely is it that at least one will fail? Probability, the systematic study of uncertainty, can be used to answer questions like these.

SECTION 6.1 Chance Experiments and Events

The basic ideas and terminology of probability are most easily introduced in situations that are both familiar and reasonably simple. Because of this, some of the initial examples involve such elementary activities as tossing a coin, selecting cards from a deck, and rolling a die. However, after considering a few of these simple examples, we will move on to more interesting and realistic situations.

Chance Experiments

When a single coin is tossed, there are two possible outcomes. The coin can land with its heads side up or its tails side up. The selection of a single card from a well-mixed standard deck of 52 playing cards may result in the ace of spades, the five of diamonds, or any one of the other 50 possibilities. Situations such as these are referred to as **chance experiments**.

DEFINITION

Chance experiment: The process of making an observation when there is uncertainty about which of two or more possible outcomes will result.

When the term chance experiment is used in a probability setting, we mean something different from what was meant by the term experiment in Chapter 2 (where experiments were a type of statistical study that investigates the effect of two or more treatments on a response). For example, in an opinion poll or survey, there is uncertainty about whether an individual selected at random from the population of interest supports a ballot initiative, and when a die is rolled, there is uncertainty about which face will land on the top face. Both of these situations fit the definition of a *chance* experiment.

Consider a chance experiment to investigate whether men or women are more likely to choose a hybrid car over a traditional internal combustion engine car when purchasing a Honda Civic at a particular car dealership. The Honda Civic is available with either a hybrid or a traditional engine. In this chance experiment, a customer will be selected at random from those who purchased a Honda Civic. The type of vehicle purchased (hybrid or traditional) will be determined and the customer's gender will be recorded. Before the customer is selected the outcome of this chance experiment is unknown to us. We do know, however, what the possible outcomes are. This set of possible outcomes is called the **sample space**.

DEFINITION

Sample space: The collection of all possible outcomes of a chance experiment.

The sample space of a chance experiment can be represented in many ways. One representation is a simple list of all the possible outcomes. For the car-purchase chance experiment described above, the possible outcomes are:

1. A male who bought a hybrid engine
2. A female who bought a hybrid engine
3. A male who bought a traditional engine
4. A female who bought a traditional engine

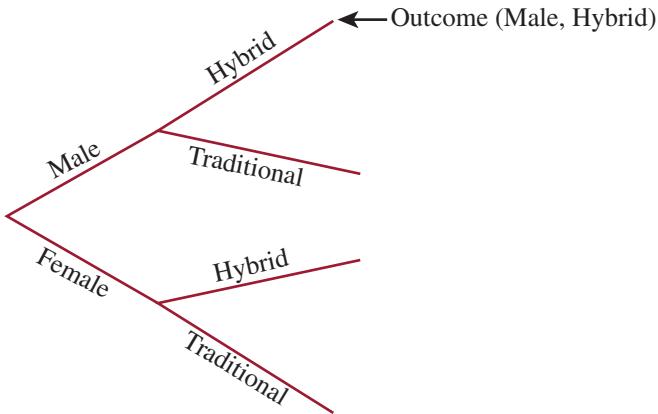
We can also use set notation and ordered pairs. A male who bought a hybrid engine is represented as (male, hybrid). The sample space is then

$$\text{sample space} = \left\{ (\text{male, hybrid}), (\text{female, hybrid}), (\text{male, traditional}), (\text{female, traditional}) \right\}$$

Another useful representation of the sample space is a tree diagram. A tree diagram (shown in Figure 6.1) for the outcomes of the car-purchase chance experiment has two sets of branches corresponding to the two pieces of information that were gathered. To identify any particular outcome in the sample space, traverse the tree by first selecting a branch corresponding to the sex of the selected customer and then a branch identified with an engine type.

FIGURE 6.1

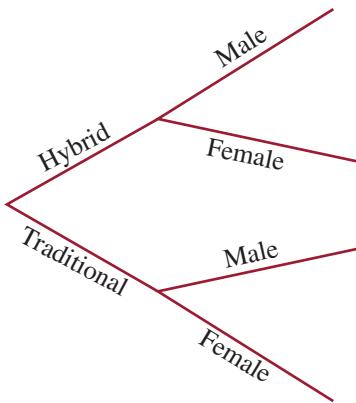
Tree diagram for the car purchase example.



In the tree diagram of Figure 6.1, there is no particular reason for having the sex of the customer as the first generation branch and the type of engine as the second generation branch. Some chance experiments involve making observations in a particular order, in which case the order of the branches in the tree does matter. In this example, however, it would be acceptable to represent the sample space with the tree diagram shown in Figure 6.2.

FIGURE 6.2

Another tree diagram for the car purchase example.



As we have seen, the sample space for a chance experiment can be represented in several ways, but the representations all have one thing in common: *Every* possible outcome of the chance experiment is included in the representation.

Events

In the car-purchase chance experiment, we might be interested in which particular outcome will result. Or we might focus on a group of outcomes that involve the purchase of a hybrid—the group consisting of (male, hybrid) and (female, hybrid). When we form a collection of one or more individual outcomes, we are creating what is known as an *event*.

DEFINITIONS

Event: Any collection of outcomes from the sample space of a chance experiment.

Simple event: An event consisting of exactly one outcome.

We usually represent an event by an uppercase letter, such as A , B , C , and so on. Sometimes the same letter with different numerical subscripts, such as E_1 , E_2 , E_3 , ... may also be used to represent events.

Example 6.1 Car Preferences

Reconsider the situation in which a person who purchased a Honda Civic was categorized by sex (M or F) and type of engine purchased (H = hybrid, T = traditional). Using this notation, one possible representation of the sample space is

$$\text{sample space} = \{\text{MH, FH, MT, FT}\}$$

Because there are four outcomes, there are four simple events:

$$E_1 = \text{MH} \quad E_2 = \text{FH} \quad E_3 = \text{MT} \quad E_4 = \text{FT}$$

One event of interest might be the event consisting of all outcomes for which a hybrid was purchased. The event *hybrid* is

$$\text{hybrid} = \{\text{MH, FH}\}$$

Another event is the event that the purchaser is male,

$$\text{male} = \{\text{MH, MT}\}$$

Example 6.2 Video Game

Suppose you believe that after losing to an opponent in a video game, a player is more likely to lose the next game. You carry out a chance experiment that consists of watching two consecutive games for a particular player and observing whether the player won, tied, or lost each of the two games. In this case (using W, T, and L to represent win, tie, and loss, respectively), the sample space can be represented as

$$\text{sample space} = \{\text{WW, WT, WL, TW, TT, TL, LW, LT, LL}\}$$

The event *player loses exactly one of the two games*, denoted by L_1 , can then be defined as

$$L_1 = \{\text{WL, TL, LW, LT}\}$$

Only one outcome (one simple event) occurs when a chance experiment is performed. We say that a given event occurs whenever one of the outcomes making up the event occurs. For example, if the outcome in the car purchase example is MH, then the simple event *male* who bought a *hybrid* has occurred, and so has the event *hybrid purchased*. The event *hybrid purchased* is not a simple event because it is made up of more than one possible outcome.

Forming New Events

Once some events have been specified, there are several ways they can be used to create new events.

DEFINITIONS

Let A and B denote two events.

Not A: The event that consists of all outcomes in the sample space that are not in event A . *Not A* is called the *complement* of A and is also sometimes denoted by A^c , A' , or \bar{A} .

A or B: The event that consists of all outcomes in the sample space that are in at least one of the two events (in A or in B or in both). A or B is called the *union* of the two events and is also sometimes denoted by $A \cup B$.

A and B: The event that consists of all outcomes in the sample space that are in both the event A and the event B . A and B is called the *intersection* of the two events and is also sometimes denoted by $A \cap B$.

Example 6.3 Turning Directions



A traffic engineer has been asked to consider whether a stop sign at the bottom of a freeway off-ramp should be replaced by a traffic light. To help in this decision, she plans to observe traffic patterns for this off-ramp. Suppose she were to record the turning direction (L = left or R = right) of each of three successive vehicles. This is a chance experiment and the sample space contains eight outcomes:

all 3 cars turn left

the 1st car turns left and the next 2 turn right

{LLL, RLL, LRL, LLR, RRL, RLR, LRR, RRR}

Each of these outcomes determines a simple event. Other events include

A = event that exactly one of the cars turns right = {RLL, LRL, LLR}

B = event that at most one of the cars turns right = {LLL, RLL, LRL, LLR}

C = event that all cars turn in the same direction = {LLL, RRR}

Some other events that can be formed from the events just defined are

$\text{not } C = C^c$ = event that not all cars turn in the same direction
= {RLL, LRL, LLR, RRL, RLR, LRR}

$A \text{ or } C = A \cup C$ = event that exactly one of the cars turns right or all cars turn in the same direction
= {RLL, LRL, LLR, LLL, RRR}

$B \text{ and } C = B \cap C$ = event that at most one car turns right and all cars turn in the same direction
= {LLL}

Example 6.4 More on Video Games

In Example 6.2, in addition to the event that a video game player loses exactly one of the two games, $L_1 = \{\text{WL}, \text{TL}, \text{LW}, \text{LT}\}$, we could also define events corresponding to $L_0 = \text{neither game is lost}$ and $L_2 = \text{both games are lost}$. Then $L_0 = \{\text{WW}, \text{WT}, \text{TW}, \text{TT}\}$ and $L_2 = \{\text{LL}\}$.

Some other events that can be formed from those just defined include

L_2^c = event that at most one game was lost
= {WW, WT, WL, TW, TT, TL, LW, LT}

$L_1 \cup L_2$ = event that at least one game was lost
= {WL, TL, LW, LT, LL}

$L_1 \cap L_2$ = event that exactly one game was lost and two games were lost
= the empty set (there are no outcomes that are in both L_1 and L_2)

Sometimes two events have no common outcomes, as was the case for events L_1 and L_2 in the video game example. Such events are described using special terminology.

DEFINITION

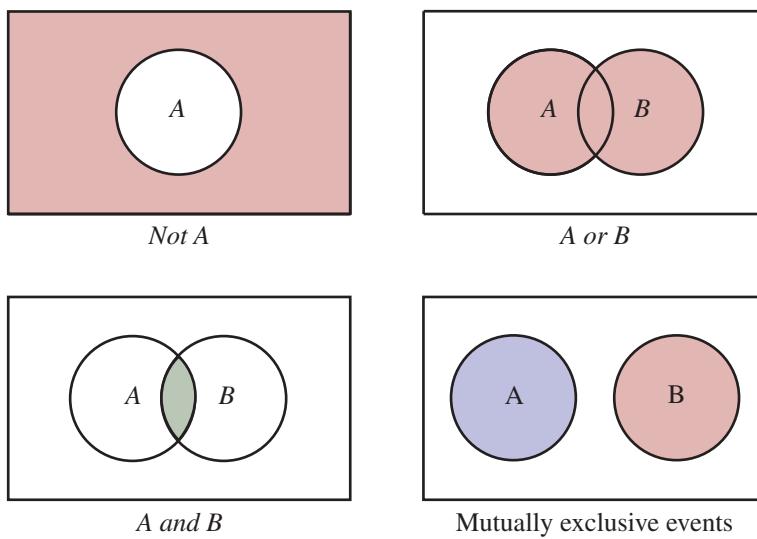
Mutually exclusive: Two events are **mutually exclusive** if they have no outcomes in common. The term **disjoint** is also sometimes used to describe events that have no outcomes in common.

Saying that two events are mutually exclusive means that they can't both occur when the chance experiment is performed once.

It is sometimes useful to draw an informal picture of events to visualize relationships. In a **Venn diagram**, the collection of all possible outcomes is typically shown as the interior of a rectangle. Other events are then identified by specified regions inside this rectangle. Figure 6.3 illustrates several Venn diagrams.

FIGURE 6.3

Venn diagrams.



The use of the *or* and *and* operations can be extended to form new events from more than two events, as described in the following box and illustrated in Figure 6.4.

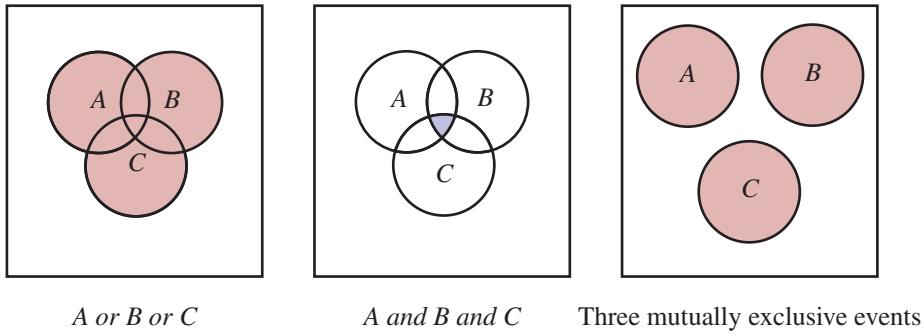
Let A_1, A_2, \dots, A_k denote k events.

1. The event A_1 or A_2 or ... or A_k consists of all outcomes in at least one of the individual events A_1, A_2, \dots, A_k .
2. The event A_1 and A_2 and ... and A_k consists of all outcomes that are in every one of the individual events A_1, A_2, \dots, A_k .

k events are mutually exclusive if no two of them have any common outcomes.

FIGURE 6.4

Venn diagrams with three events.



A or B or C

A and B and C

Three mutually exclusive events

Example 6.5 Asking Questions

A seminar class has four students. The instructor has an unusual way of asking questions. Four slips of paper numbered 1, 2, 3, and 4 are placed in a box. The instructor determines the student to whom any particular question will be addressed by selecting one of these four slips at random. Suppose that one question is to be posed during each of the next two class meetings. One possible outcome could be represented as (3, 1)—the first question is addressed to Student 3 and the second question to Student 1. There are 15 other possibilities. Consider the following events:

the event that the same student is asked both questions

$$\longrightarrow A = \{(1,1), (2,2), (3,3), (4,4)\}$$

the event that Student 1 is asked at least one of the two questions

$$\begin{aligned} \longrightarrow B &= \{(1,1), (1,2), (1,3), (1,4), (2,1), (3,1), (4,1)\} \\ C &= \{(3,1), (2,2), (1,3)\} \\ D &= \{(3,3), (3,4), (4,3)\} \end{aligned}$$

$$\begin{aligned}E &= \{(1,1), (1,3), (2,2), (3,1), (4,2), (3,3), (2,4), (4,4)\} \\F &= \{(1,1), (1,2), (2,1)\}\end{aligned}$$

Then

$$A \text{ or } C \text{ or } D = \{(1,1), (2,2), (3,3), (4,4), (3,1), (1,3), (3,4), (4,3)\}$$

The outcome $(3,1)$ is contained in each of the events B , C , and E , as is the outcome $(1,3)$. These are the only two common outcomes in all three events, so

$$B \text{ and } C \text{ and } E = \{(3,1), (1,3)\}$$

The events C , D , and F are mutually exclusive because no outcome in any one of these events is contained in either of the other two events.

EXERCISES 6.1 - 6.16

● Data set available online

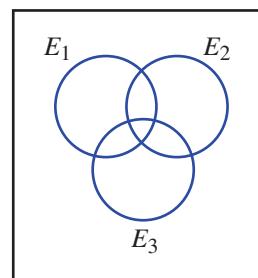
- 6.1** Define the term *chance experiment*, and give an example of a chance experiment with four possible outcomes.
- 6.2** Define the term *sample space*, and then give the sample space for the chance experiment you described in Exercise 6.1.
- 6.3** Consider the chance experiment in which the type of transmission—automatic (A) or manual (M)—is recorded for each of the next two cars purchased from a certain dealer.
- What is the set of all possible outcomes (the sample space)? (Hint: See Example 6.2.)
 - Display the possible outcomes in a tree diagram.
- 6.4** Refer to the chance experiment described in the previous exercise and the sample space for that experiment.
- List the outcomes in each of the following events. Which of these events are simple events? (Hint: See Example 6.3.)
 - the event that at least one car has an automatic transmission
 - the event that exactly one car has an automatic transmission
 - the event that neither car has an automatic transmission - What outcomes are in the event B and C ? In the event B or C ?
- 6.5** A tennis shop sells five different brands of rackets, each of which comes in either a midsized version or an oversize version. Consider the chance experiment in which brand and size are noted for the next racket purchased. One possible outcome is Head midsized, and another is Prince oversize. Possible outcomes correspond to cells in the following table:

	Head	Prince	Slazenger	Wimbledon	Wilson
Midsized					
Oversize					

- Let A denote the event that an oversize racket is purchased. List the outcomes in A .
 - Let B denote the event that the name of the brand purchased begins with a W. List the outcomes in B .
 - List the outcomes in the event *not* B .
- 6.6** Refer to the chance experiment described in the previous exercise.
- Head, Prince, and Wilson are U.S. companies and the others are not. Let C denote the event that the racket purchased is made by a U.S. company. List the outcomes in the event B or C . (Hint: See Example 6.4.)
 - List outcomes in B and C .
 - Display the possible outcomes on a tree diagram, with a first-generation branch for each brand.
- 6.7** A new model of laptop computer can be ordered with one of three screen sizes (10 inches, 12 inches, 15 inches) and one of four hard drive sizes (50 GB, 100 GB, 150 GB, and 200 GB). Consider the chance experiment in which a laptop order is selected and the screen size and hard drive size are recorded.
- Display possible outcomes using a tree diagram.
 - Let A be the event that the order is for a laptop with a screen size of 12 inches or smaller. Let B be the event that the order is for a laptop with a hard drive size of at most 100 GB. What outcomes are in
 - A^C ?
 - $A \cup B$?
 - $A \cap B$?

- c.** Let C denote the event that the order is for a laptop with a 200 GB hard drive. Are A and C mutually exclusive events? Are B and C mutually exclusive?
- 6.8** A college library has four copies of a certain book. The copies are numbered 1, 2, 3, and 4. Two of these books are selected at random. The first selected book is placed on 2-hour reserve, and the second book can be checked out overnight.
- Construct a tree diagram to display the 12 outcomes in the sample space.
 - Let A denote the event that at least one of the books selected is an even-numbered copy. What outcomes are in A ?
 - Suppose that copies 1 and 2 are hardcover books, whereas copies 3 and 4 are softcover books. Let B denote the event that exactly one of the copies selected is a hardcover book. What outcomes are contained in B ?
- 6.9** A library has five copies of a certain textbook on reserve of which two copies (1 and 2) are hardcover books and the other three (3, 4, and 5) are softcover books. A student examines these books in random order, stopping only when a softcover book has been selected.
- Display the possible outcomes in a tree diagram.
 - What outcomes are contained in the event A , that exactly one book is examined before the chance experiment terminates?
 - What outcomes are contained in the event C , that the chance experiment terminates with the examination of book 5?
- 6.10** Suppose that, starting at a certain time, batteries coming off an assembly line are examined one by one to see whether they are defective (let D = defective and N = not defective). The chance experiment terminates as soon as a nondefective battery is obtained.
- Give five possible outcomes for this chance experiment.
 - What can be said about the number of outcomes in the sample space?
 - What outcomes are in the event E , that the number of batteries examined is an even number?
- 6.11** Refer to the previous exercise and now suppose that the chance experiment terminates only when two nondefective batteries have been obtained.
- Let A denote the event that at most three batteries must be examined before the chance experiment terminates. What outcomes are contained in A ?
 - Let B be the event that exactly four batteries must be examined before the chance experiment terminates. What outcomes are in B ?
- c.** What can be said about the number of possible outcomes for this chance experiment?
- 6.12** A family consisting of three people— P_1 , P_2 , and P_3 —belongs to a medical clinic that always has a physician at each of stations 1, 2, and 3. During a certain week, each member of the family visits the clinic exactly once and is randomly assigned to a station. One experimental outcome is $(1, 2, 1)$, which means that P_1 is assigned to station 1, P_2 to station 2, and P_3 to station 1.
- List the 27 possible outcomes. (Hint: First list the nine outcomes in which P_1 goes to station 1, then the nine in which P_1 goes to station 2, and finally the nine in which P_1 goes to station 3; a tree diagram might help.)
 - List all outcomes in the event A , that all three people go to the same station.
 - List all outcomes in the event B , that all three people go to different stations.
 - List all outcomes in the event C , that no one goes to station 2.
- 6.13** Using the outcomes for the chance experiment described in the previous exercise, identify outcomes in each of the following events:
- B^c
 - C^c
 - $A \cup B$
 - $A \cap B$
 - $A \cap C$
- 6.14** An engineering construction firm is currently working on power plants at three different sites. Define events E_1 , E_2 , and E_3 as follows:
- E_1 = the plant at Site 1 is completed by the contract date
- E_2 = the plant at Site 2 is completed by the contract date
- E_3 = the plant at Site 3 is completed by the contract date

The following Venn diagram pictures the relationships among these events:



Shade the region in the Venn diagram corresponding to each of the following events. Redraw the Venn diagram for each part of the problem. (Hint: See the discussion of Venn diagrams.)

- At least one plant is completed by the contract date.

- b. All plants are completed by the contract date.
 - c. None of the plants are completed by the contract date.
- 6.15** For the events described in the previous exercise, shade the region in the Venn diagram corresponding to each of the following events. Redraw the Venn diagram for each part.
- a. Only the plant at Site 1 is completed by the contract date.
 - b. Exactly one of the three plants is completed by the contract date.
- c. Either the plant at Site 1 or both of the other two plants are completed by the contract date.
- 6.16** Consider a Venn diagram picturing two events A and B that are not mutually exclusive.
- a. Shade the event $(A \cup B)^c$. On a separate Venn diagram shade the event $A^c \cap B^c$. How are these two events related?
 - b. Shade the event $(A \cap B)^c$. On a separate Venn diagram shade the event $A^c \cup B^c$. How are these two events related? (Note: These two relationships together are called DeMorgan's laws.)

SECTION 6.2 Definition of Probability

Reasoning about games of chance has a long history. Archeologists have found evidence that Egyptians used a small bone in mammals, the astragalus, as a sort of four-sided die as early as 3500 b.c. Games of chance were common in Greek and Roman times and during the Renaissance of Western Europe. From this early work related to games of chance, new interpretations of probability have evolved, including an empirical approach preferred by many statisticians today. We begin our discussion of probability with a look at some of the different approaches to probability: the classical, relative frequency, and subjective approaches.

Classical Approach to Probability

Early mathematicians' development of the theory of probability was primarily in the context of gambling and games of chance. A characteristic common to most games of chance is a physical device used to produce outcomes. For example, in children's games, dice or a "spinner" might be used to create chance outcomes.

Dice are constructed so that the physical characteristics of each face are alike except for the number of dots. This ensures that the different outcomes for an individual die are equally likely. If this is the case, the probability of getting a five when a single six-sided die is rolled is one-sixth (1 chance in 6). In general, if there are N *equally likely* outcomes in the sample space for a chance experiment (such as rolling a die), the probability of each of the outcomes is $1/N$.

Classical Approach to Probability for Equally Likely Outcomes

When the outcomes in the sample space of a chance experiment are equally likely, the **probability of an event E** , denoted by $P(E)$, is the ratio of the number of outcomes favorable to E to the total number of outcomes in the sample space:

$$P(E) = \frac{\text{Number of outcomes favorable to } E}{\text{Number of outcomes in the sample space}}$$

According to this definition, the calculation of a probability consists of counting the number of outcomes that make up an event, counting the number of outcomes in the sample space, and then dividing.

Chance experiments that involve tossing fair coins, rolling fair dice, or selecting cards from a well-mixed deck of playing cards have equally likely outcomes. For example, if a fair die is rolled once, each outcome (simple event) has probability $1/6$. With E denoting

the event that the number rolled is even, $P(E) = 3/6$. This is just the number of outcomes in E (2, 4, and 6) divided by the total number of possible outcomes.

Example 6.6 Calling the Toss

On some football teams, the honor of calling the toss at the beginning of a football game is determined by random selection. Suppose that this week a member of the offense will call the toss. There are 5 linemen on the 11-player offense. If we define the event L as the event that a lineman is selected to call the toss, 5 of the 11 possible outcomes are included in L . The probability that a lineman will be selected is then

$$P(L) = \frac{5}{11}$$

Example 6.7 Math Contest

Four students (Adam, Betina, Carlos, and Debra) submitted correct solutions to a math contest that had two prizes. The contest rules specify that if more than two correct responses are submitted, the winners will be selected at random from those submitting correct responses. We can use (A, B) to denote the outcome that Adam and Betina are the two selected. The other possible outcomes can be denoted in a similar way. Then the sample space for the chance experiment of selecting the two winners from the four correct responses is

$$\{(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)\}$$

Because the winners are selected at random, the six outcomes are equally likely and the probability of each individual outcome is $1/6$.

Suppose that E denotes the event that both selected winners are the same sex. Then

$$E = \{(A, C), (B, D)\}$$

Because E contains two outcomes, $P(E) = 2/6 = 0.333$.

If F denotes the event that at least one of the selected winners is female, then F consists of all outcomes except (A, C) and $P(F) = 5/6 = 0.833$.

The classical approach to probability works well for chance experiments that have a finite number of outcomes that are equally likely. However, many chance experiments do not have equally likely outcomes. For example, consider the chance experiment of selecting a student from those enrolled at a school and observing whether the student is a freshman, sophomore, junior, or senior. If there are more seniors than sophomores at the school, then the four possible outcomes for this chance experiment are not equally likely. In this situation, it would be a mistake to think that the probability of each outcome is $1/4$. To calculate or estimate probabilities in situations where outcomes are not equally likely, a different approach is needed.

Relative Frequency Approach to Probability

Early scientific investigators were aware that chance experiments do not always give the same results when repeated. However, even though it is not possible to predict the outcome of a chance experiment in any particular instance, there is a dependable and stable regularity when a chance experiment is repeated many times. This became the basis for fair wagering in games of chance.

For example, suppose that two friends, Chris and Tom, meet to play a game. A fair coin is flipped. If it lands heads up, Chris pays Tom \$1; otherwise, Tom pays the same amount to Chris. After many repetitions, the proportion of the time that Chris wins will be close to one-half, and the two friends will have simply enjoyed the pleasure of each other's company for a few hours. This happy circumstance occurs because *in the long run* (i.e., over many repetitions) the proportion of heads is 0.5. In the long run, half the time Chris wins, and half the time Tom wins.

When any given chance experiment is performed, some events are relatively likely to occur, whereas others are not as likely to occur. For a specified event E , we want to assign a probability that gives information about how likely it is that E will occur. In the relative frequency approach to probability, the probability describes how frequently E occurs when the chance experiment is performed many times.

Example 6.8 Tossing a Coin

Frequently, we hear a coin described as fair, or we are told that there is a 50% chance of a coin landing heads up. What is the meaning of the expression *fair*? It cannot refer to the result of a single toss, because a single toss cannot result in both a head and a tail. Might “fairness” and “50%” refer to 10 successive tosses yielding exactly five heads and five tails? Not really, because it is easy to imagine a fair coin landing heads up on only 3 or 4 of the 10 tosses.

Consider the chance experiment of tossing a coin just once. Define the simple events

$$\begin{aligned} H &= \text{event that the coin lands with its heads side facing up} \\ T &= \text{event that the coin lands with its tails side facing up} \end{aligned}$$

Now suppose that we take a fair coin and begin to toss it over and over. After each toss, we calculate the relative frequency of heads observed so far. This calculation gives the value of the ratio

$$\frac{\text{number of times event } H \text{ occurs}}{\text{number of tosses}}$$

The results of the first 10 tosses might be as follows:

Toss	1	2	3	4	5	6	7	8	9	10
Cumulative number of H 's	0	1	2	3	3	3	4	5	5	5
Relative frequency of H 's	0	0.5	0.667	0.75	0.6	0.5	0.571	0.625	0.556	0.5

Figure 6.5 illustrates how the relative frequency of heads fluctuates during one sequence of 50 tosses. As the number of tosses increases, the relative frequency of heads does not continue to fluctuate wildly but instead stabilizes and approaches some fixed number (the *limiting value*). This stabilization is illustrated for a sequence of 1000 tosses in Figure 6.6.

FIGURE 6.5

Relative frequency of heads in the first 50 of a long series of coin tosses.

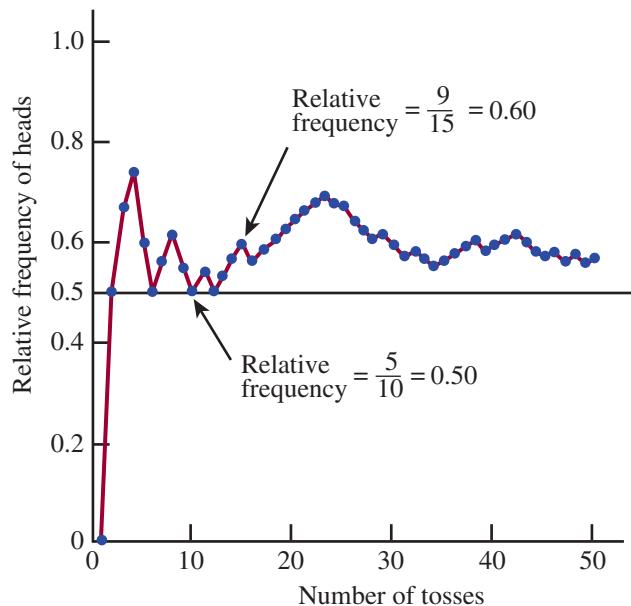
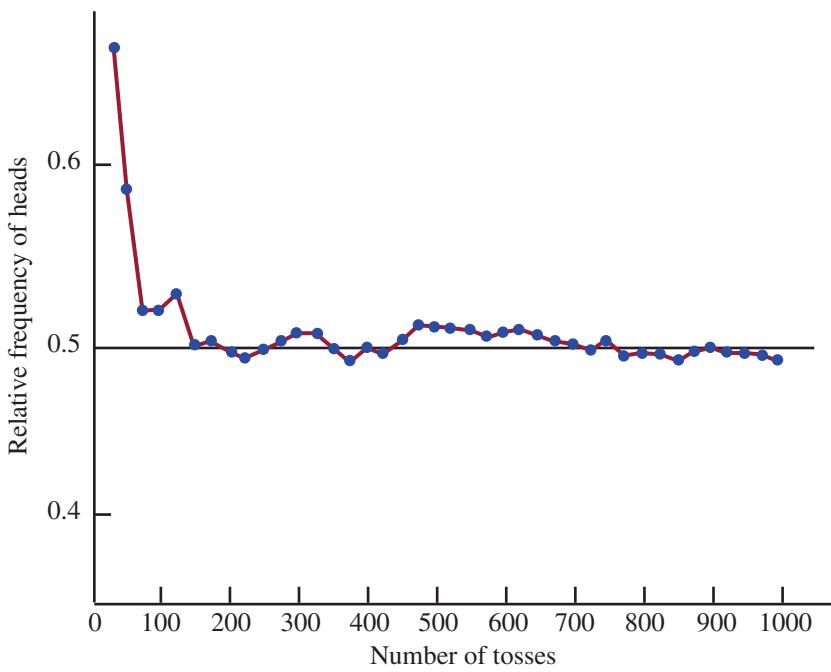


FIGURE 6.6

Stabilization of the relative frequency of heads in coin tossing.



It may seem natural to you that the proportion of H 's would get closer and closer to the “real” probability of 0.5. That an empirically observed proportion could behave like this in real life also seemed reasonable to James Bernoulli in the early 18th century. He focused his considerable mathematical power on this topic and was able to prove mathematically what we now know as the law of large numbers.*

Law of Large Numbers

As the number of repetitions of a chance experiment increases, the chance that the relative frequency of occurrence for an event will differ from the true probability of the event by more than any small number approaches 0.

The law of large numbers tells us that, as the number of repetitions of a chance experiment increases, the proportion of the time an event E occurs gets close to and stays close to the real probability of E occurring *even if the value of this probability is not known*. This means that we can observe the outcomes of repetitions of a chance experiment and then use the observed outcomes to estimate probabilities.

Relative Frequency Approach To Probability

The **probability of an event E** , denoted by $P(E)$, is defined to be the value approached by the relative frequency of occurrence of E when a chance experiment is performed many times. If the number of times the chance experiment is performed is quite large,

$$P(E) \approx \frac{\text{Number of times } E \text{ occurs}}{\text{Number of times the experiment is performed}}$$

The relative frequency definition of probability depends on being able to repeat a chance experiment under identical conditions. Suppose that we perform a chance experiment that consists of flipping a cap from a 20-ounce bottle of soda and noting whether the cap lands with the open side up or down. The crucial difference from the previous coin-tossing

*Technically, this should be referred to more specifically as the weak law of large numbers for Bernoulli trials. After Bernoulli's proof, mathematical statisticians proved more general laws of large numbers.

chance experiment is that there is no particular reason to believe the cap is equally likely to land top up or top down.

If we assume that the chance experiment can be repeated under similar conditions (which seems reasonable), then we can flip the cap many times and calculate the relative frequency of the event *top up* (the proportion of the time the event has occurred so far):

$$\frac{\text{number of times the event } \textit{top up} \text{ occurs}}{\text{number of flips}}$$

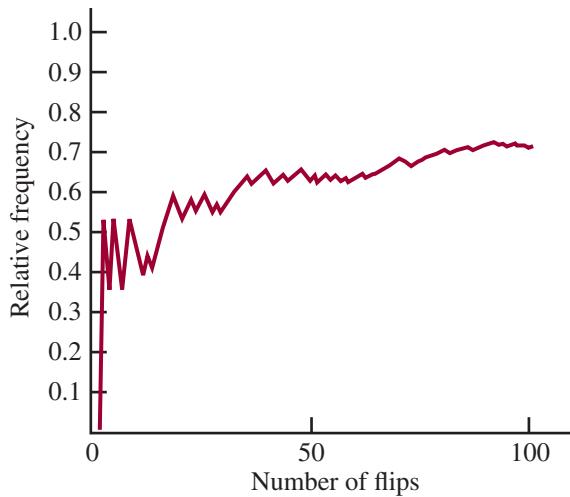
The results of the first 10 flips, with U indicating top up and D indicating top down, might be:

Flip	1	2	3	4	5	6	7	8	9	10
Outcome	U	U	D	U	D	D	U	D	U	U
Cumulative										
number of <i>ups</i>	1	2	2	3	3	3	4	4	5	6
Relative frequency of <i>ups</i>	1.0	1.0	0.67	0.75	0.6	0.5	0.57	0.5	0.56	0.6

Figure 6.7 illustrates how the relative frequency of the event *top up* fluctuates during a sequence of 100 flips. Based on these results and faith in the law of large numbers, it is reasonable to think that the probability of the cap landing top up is about 0.7.

FIGURE 6.7

Stabilization of the relative frequency of a bottle cap landing top up.



The relative frequency approach to probability is based on observation. From repeated observation, we can obtain stable relative frequencies that will, in the long run, provide good estimates of the probabilities of different events. The relative frequency interpretation of probability is intuitive and widely used.

Subjective Approach to Probability

A third approach to probability is based on subjective judgments. In this view, a probability is interpreted as a personal measure of the strength of belief that a particular event will occur. A probability of 1 represents a belief that the event will certainly occur. A probability of 0 represents a belief that the event will certainly not occur—that it is impossible. Other probabilities are placed somewhere between 0 and 1, based on the strength of one's beliefs.

For example, an airline passenger might report her subjective assessment of the probability of being denied a seat as a result of overbooking on a particular flight as 0.01. Because this probability is close to 0, she believes it is very unlikely that she will be denied a seat.

The subjective interpretation of probability presents some difficulties. Different people may assign different probabilities to the same outcome, because each could have a

different subjective belief. Although subjective probabilities are useful in studying and analyzing decision making, they are of limited use because they are personal and not generally replicable by others.

How Are Probabilities Determined?

Probability statements are frequently encountered in newspapers and magazines (and, of course, in this textbook). What is the basis for these probability statements? In most cases, research allows reasonable estimates based on observation and analysis. In general, if a probability is stated, it is based on one of the following approaches:

1. *The classical approach:* This approach is appropriate only for modeling chance experiments with equally likely outcomes.
2. *The subjective approach:* In this case, the probability represents an individual's judgment based on facts combined with personal evaluation of other information.
3. *The relative frequency approach:* An estimate is based on observing the outcomes of a chance experiment.

When you see probabilities in this book and in other published sources, you should consider how the probabilities were produced. Be particularly cautious in the case of subjective probabilities.

SECTION 6.3 Basic Properties of Probability

Estimating probabilities by observing outcomes and then calculating a relative frequency is not always practical. The most obvious problem is time. If calculating probabilities even for simple chance experiments were to require hundreds or thousands of observations, everyone would avoid this task! An alternative approach that simplifies things in some situations is to study fundamental properties of probability that can be used to find probabilities of complex events. These fundamental properties are given in the accompanying box. A discussion of each property follows.

Fundamental Properties of Probability

1. For any event E , $0 \leq P(E) \leq 1$.
2. If S is the sample space for a chance experiment, $P(S) = 1$.
3. If two events E and F are mutually exclusive, then $P(E \text{ or } F) = P(E) + P(F)$.
4. For any event E , $P(E) + P(\text{not } E) = 1$. It follows that

$$P(\text{not } E) = 1 - P(E)$$

and

$$P(E) = 1 - P(\text{not } E).$$

Property 1: For any event E , $0 \leq P(E) \leq 1$

To understand the first property of probability, recall the previous bottle cap chance experiment. You may remember that we were keeping track of the number of flips landing with the top of the bottle cap facing up. Suppose that we have flipped the cap N times. The number of times that the cap landed with the top facing up must be between 0 (it never happened) and N (it always happened).

Then the relative frequency of top up must be between two numbers:

$$\frac{0}{N} \leq \text{relative frequency} \leq \frac{N}{N}$$

This means that the relative frequency must always be greater than or equal to 0 and less than or equal to 1. As N increases, the long-run value of the relative frequency, which is the probability, must also lie between 0 and 1.

Property 2: If S is the sample space for a chance experiment, $P(S) = 1$

The probability of any event is the proportion of time an outcome in the event will occur in the long run. Because the sample space consists of all possible outcomes for a chance experiment, in the long run an outcome that is in S must occur 100% of the time. This means that $P(S) = 1$.

Property 3: If two events E and F are mutually exclusive, then

$$P(E \text{ or } F) = P(E) + P(F)$$

This is one of the most important properties of probability because it provides a method for calculating probabilities if the number of possible outcomes (simple events) in the sample space is finite. Any event is just a collection of outcomes from the sample space—some outcomes are in the event and others are not. Because simple events are mutually exclusive, any event can be viewed as a union of mutually exclusive events.

For example, suppose that a chance experiment has a sample space that consists of six outcomes, O_1, O_2, \dots, O_6 . Then $P(O_1)$ can be interpreted as the long run proportion of times that the outcome O_1 will occur. The probabilities of the other outcomes can be interpreted in a similar way. Suppose that an event E is made up of outcomes O_1, O_2 , and O_4 . Then E will occur whenever O_1, O_2 , or O_4 occurs. Because O_1, O_2 , and O_4 are mutually exclusive, the long-run proportion of time that E will occur is just the sum of the proportion of time that each of the outcomes O_1, O_2 , and O_4 will occur.

Because it is always possible to express any event as a collection of mutually exclusive simple events in this way, we can find the probability of an event by finding the probabilities of the simple events (outcomes) that make up the event and then adding. If the probabilities of all the simple events are known, it is then easy to calculate the probability of more complex events.

Property 4: For any event E , $P(E) + P(\text{not } E) = 1$. It follows that

$$P(\text{not } E) = 1 - P(E) \text{ and } P(E) = 1 - P(\text{not } E)$$

Property 4 follows from Properties 2 and 3. Property 2 tells us that $P(E)$ is the sum of the probabilities of the simple events that are in E and that $P(\text{not } E)$ is the sum of the probabilities for simple events that are in $\text{not } E$. Every simple event in the sample space is in either E or $\text{not } E$. We know from Property 3 that the sum of the probabilities of all the simple events is 1. It follows that $P(E) + P(\text{not } E)$ must equal 1.

Property 4 implies that $P(E) = 1 - P(\text{not } E)$. Knowing this is surprisingly useful. There are many situations where it is easier to calculate $P(\text{not } E)$ than it is to calculate $P(E)$ directly.

Example 6.9 Preventable Deaths

Understand the context ➤

The authors of the paper “[Preventable Deaths Due to Problems in Care in English Acute Hospitals](#)” (*BMJ Quality & Safety* [2012]: 737–745) carried out a study of hospital deaths. They estimated that 5.2% of the deaths that occurred in English hospitals were preventable. They attributed the preventable deaths to the following types of errors: (1) poor patient monitoring; (2) diagnostic errors; (3) inadequate drug or fluid management; and (4) other. It was estimated that for the preventable deaths

31% were due to poor patient monitoring

30% were due to diagnostic errors

21% were due to inadequate drug or fluid management

18% were due to other errors

Consider the chance experiment that consists of observing the type of error for a randomly selected preventable death. The simple events for this experiment can be represented by O_1, O_2, O_3 , and O_4 :

O_1 = the simple event that the death was due to poor patient monitoring

O_2 = the simple event that the death was due to a diagnostic error

O_3 = the simple event that the death was due to inadequate drug or fluid management
 O_4 = the simple event that the death was due to other errors

The following table displays the probabilities for the simple events based on the estimates given in the paper.

Simple Event (Outcome)	O_1	O_2	O_3	O_4
Probability	0.31	0.30	0.21	0.18

Let's consider the event E , the event that *a randomly selected preventable death is due to poor patient monitoring, a diagnostic error, or inadequate drug or fluid control*. This event consists of the outcomes O_1 , O_2 , and O_3 . It follows that

$$\begin{aligned} P(E) &= P(O_1 \cup O_2 \cup O_3) \\ &= P(O_1) + P(O_2) + P(O_3) \\ &= 0.31 + 0.30 + 0.21 \\ &= 0.82 \end{aligned}$$

That is, in the long run, 82% of all preventable deaths are due to poor patient monitoring, diagnostic errors, or inadequate drug or fluid monitoring. Another way to calculate this probability is to notice that the event *not E* consists only of the outcome O_4 . Another way to calculate $P(E)$ is

$$P(E) = 1 - P(\text{not } E) = 1 - P(O_4) = 1 - 0.18 = 0.82$$

Having established the fundamental properties of probability, we now present some probability rules that can be used to evaluate probabilities in some situations. An important aspect of each rule is the set of conditions necessary in order to use the rule. A common error for beginning statistics students is to believe that the rules can be used in *any* probability calculation. This is *not* the case. You must be careful to use the rules only after verifying that any necessary conditions are met.

Equally Likely Outcomes

The first probability rule that we consider applies only when the simple events in the sample space are equally likely. Chance experiments involving tossing fair coins, rolling fair dice, or selecting cards from a well-mixed deck of playing cards are examples of chance experiments that have equally likely outcomes. If a fair die is rolled once, each possible outcome (1, 2, 3, 4, 5, and 6) has probability 1/6. With E denoting the event that the outcome is an even number,

$$P(E) = P(2 \text{ or } 4 \text{ or } 6) = P(2) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}$$

This is just the ratio of the number of outcomes in E to the total number of possible outcomes. The following box presents the generalization of this result.

Calculating Probabilities When Simple Events in the Sample Space Are Equally Likely

Consider a chance experiment that can result in any one of N possible outcomes. Denote the corresponding simple events by O_1, O_2, \dots, O_N . If these simple events are equally likely to occur, then

1. $P(O_1) = \frac{1}{N}, P(O_2) = \frac{1}{N}, \dots, P(O_N) = \frac{1}{N}$

2. For any event E ,

$$P(E) = \frac{\text{number of simple events in } E}{N}$$

Addition Rule for Mutually Exclusive Events

We have seen previously that the probability of an event can be calculated by adding together probabilities of the simple events that correspond to the outcomes making up the event. This addition process can also be used to calculate the probability of the union of two events that are mutually exclusive.

The Addition Rule for Mutually Exclusive Events

Suppose E and F are mutually exclusive events. Then

$$P(E \text{ or } F) = P(E \cup F) = P(E) + P(F)$$

This property of probability is known as the addition rule for mutually exclusive events. More generally, if events E_1, E_2, \dots, E_k are all mutually exclusive, then

$$P(E_1 \text{ or } E_2 \text{ or } \dots \text{ or } E_k) = P(E_1 \cup E_2 \cup \dots \cup E_k) = P(E_1) + P(E_2) + \dots + P(E_k)$$

In words, the probability that any of these k mutually exclusive events occurs is the sum of the probabilities of the individual events.

Consider the chance experiment that consists of rolling a pair of fair dice. There are 36 possible outcomes for this chance experiment, such as $(1, 1)$, $(1, 2)$, and so on. Because these outcomes are equally likely, we can calculate the probability of observing a total of 5 on the two dice by counting the number of outcomes that result in a total of 5. There are four such outcomes— $(1, 4)$, $(2, 3)$, $(3, 2)$, and $(4, 1)$. Then

$$P(\text{total of } 5) = \frac{4}{36}$$

The probabilities for the other possible totals can be calculated in a similar way, and they are shown in the following table:

Total	Probability	Total	Probability
2	1/36	8	5/36
3	2/36	9	4/36
4	3/36	10	3/36
5	4/36	11	2/36
6	5/36	12	1/36
7	6/36		

What is the probability of getting a total of 3 or 5? Consider the two events

$$E = \text{total is } 3$$

and $F = \text{total is } 5$

Clearly E and F are mutually exclusive because the sum cannot simultaneously be both 3 and 5. Notice also that neither of these events is simple, because neither consists of only a single outcome from the sample space. We can apply the addition rule for mutually exclusive events as follows:

$$P(E \text{ or } F) = P(E \cup F) = P(E) + P(F) = \frac{2}{36} + \frac{4}{36} = \frac{6}{36}$$

Example 6.10 Car Choices

Understand the context ➤

A large auto center sells cars made by a number of different manufacturers. Three of these manufacturers are Japanese: Honda, Nissan, and Toyota. Consider a chance experiment that consists of observing the make and model of the next car purchased at this auto center.

The outcomes in the sample space would then be simple events such as Nissan Altima and Toyota Prius. Let's define events E_1 , E_2 , and E_3 by

$$E_1 = \text{Honda}$$

$$E_2 = \text{Nissan}$$

and $E_3 = \text{Toyota}$

Notice that E_1 is not a simple event because there is more than one model Honda (for example, Civic, Accord, and Fit).

Suppose that based on several years of sales data, the probabilities of these three events have been estimated empirically as $P(E_1) = 0.25$, $P(E_2) = 0.18$, and $P(E_3) = 0.14$. Because E_1 , E_2 , and E_3 are mutually exclusive, the addition rule gives

$$\begin{aligned} P(\text{Honda or Nissan or Toyota}) &= P(E_1 \cup E_2 \cup E_3) \\ &= P(E_1) + P(E_2) + P(E_3) \\ &= 0.25 + 0.18 + 0.14 = 0.57 \end{aligned}$$

The probability that the next car purchased is *not* made by one of these three manufacturers is

$$P(\text{not}(E_1 \text{ or } E_2 \text{ or } E_3)) = P((E_1 \cup E_2 \cup E_3)^C) = 1 - 0.57 = 0.43$$

In Section 6.6, we will see how $P(E \cup F)$ can be calculated when the two events E and F are not mutually exclusive.

EXERCISES 6.17 - 6.37

● Data set available online

- 6.17** A large department store offers online ordering. When a purchase is made online, the customer can select one of four different delivery options: expedited overnight delivery, expedited second-day delivery, standard delivery, or delivery to the nearest store for customer pick-up. Consider the chance experiment that consists of observing the selected delivery option for a randomly selected online purchase. What are the four simple events that make up the sample space for this experiment?

- 6.18** Consider the chance experiment described in the previous exercise. Suppose that the probability of an overnight delivery selection is 0.1, the probability of a second-day delivery selection is 0.3, and the probability of a standard-delivery selection is 0.4. Find the following probabilities. (Hint: See Example 6.9.)
- the probability that a randomly selected online purchase selects delivery to the nearest store for customer pick-up.
 - the probability that the customer selects a form of expedited delivery.
 - the probability that either standard delivery or delivery to the nearest store is selected.

- 6.19** The manager of an online music store has kept records of the number of songs downloaded in a single transaction by customers who make a purchase at the store. Consider the chance experiment of observing the number of songs downloaded by

a randomly selected customer. The accompanying table gives six possible outcomes and the estimated probability of each of these outcomes.

Number of songs downloaded	1	2	3	4	5	more
Estimated probability	0.45	0.25	0.10	0.10	0.07	0.03

- What is the estimated probability that a randomly selected customer downloads three or fewer songs?
- What is the estimated probability that a randomly selected customer downloads at most three songs? How does this compare to the probability calculated in Part (a)?

- 6.20** Consider the chance experiment described in the previous exercise.
- What is the estimated probability that a randomly selected customer downloads five or more songs?
 - What is the estimated probability that a randomly selected customer downloads one or two songs?
 - What is the estimated probability that a randomly selected customer downloads more than two songs? Show two different ways to compute this probability that use the probability rules of this section.

- 6.21** A bookstore sells two types of books (fiction and nonfiction) in several formats (hardcover, paperback, digital, and audio). For the chance experiment that consists of observing the type and format of a single-book purchase, two of the eight possible outcomes are a hardcover fiction book and an audio nonfiction book.

- There are eight outcomes in the sample space for this experiment. List these possible outcomes.
- Do you think it is reasonable to think that the outcomes for this experiment would be equally likely? Explain.

- 6.22** Consider the chance experiment described in the previous exercise.

- For customers who purchase a single book, the estimated probabilities for the different possible outcomes are given in the cells of the accompanying table. What is the probability that a randomly selected single-book purchase will be for a book in print format (hardcover or paperback)?

	Hardcover	Paperback	Digital	Audio
Fiction	0.15	0.45	0.10	0.10
Nonfiction	0.08	0.04	0.02	0.06

- Show two different ways to calculate the probability that a randomly selected single-book purchase will be for a book that is not in a print format. (Hint: See Example 6.9.)
- Calculate the probability that a randomly selected single-book purchase will be for a work of fiction.

- 6.23** Medical insurance status—covered (C) or not covered (N)—is determined for each individual arriving for treatment at a hospital's emergency room. Consider the chance experiment in which this determination is made for two randomly selected patients.

The simple events for this chance experiment are $O_1 = (C, C)$, meaning that the first patient selected was covered and the second patient selected was also covered, $O_2 = (C, N)$, $O_3 = (N, C)$, and $O_4 = (N, N)$. Suppose that probabilities are $P(O_1) = 0.81$, $P(O_2) = 0.09$, $P(O_3) = 0.09$, and $P(O_4) = 0.01$.

- Define A as the event that at most one patient is covered.
 - What simple events are in the event A ?
 - Calculate $P(A)$.
- Define B as the event that the two patients have the same coverage status.
 - What simple events are in the event B ?
 - Calculate $P(B)$.

- 6.24** Roulette is a game of chance that involves spinning a wheel that is divided into 38 equal segments, as shown in the accompanying picture.



Anna Baburkina/Shutterstock.com

A metal ball is tossed into the wheel as it is spinning, and the ball eventually lands in one of the 38 segments. Each segment has an associated color. Two segments are green. Half of the other 36 segments are red and the others are black. When a balanced roulette wheel is spun, the ball is equally likely to land in any one of the 38 segments.

- When a balanced roulette wheel is spun, what is the probability that the ball lands in a red segment?
- In the roulette wheel shown, black and red segments alternate. Suppose instead that the red segments were side-by-side and that the black segments were together. Does this increase the probability that the ball will land in a red segment? Explain.
- Suppose that you watch 1000 spins of a roulette wheel and note the color that results from each spin. What would be an indication that the wheel was not balanced?

- 6.25** Phoenix is a hub for a large airline. Suppose that on a particular day, 8000 passengers arrived in Phoenix on this airline. Phoenix was the final destination for 1800 of these passengers. The others were all connecting to flights to other cities. On this particular day, several inbound flights were late, and 480 connecting passengers missed their connecting flight and were delayed in Phoenix. Of the 480 who were delayed, 75 were delayed overnight and had to spend the night in Phoenix. Consider the chance experiment of choosing a passenger at random from these 8000 passengers. Calculate the following probabilities:

- the probability that the selected passenger had Phoenix as a final destination.
- the probability that the selected passenger did not have Phoenix as a final destination.
- the probability that the selected passenger was connecting and missed the connecting flight.

- d.** the probability that the selected passenger was a connecting passenger and did not miss the connecting flight.
- e.** the probability that the selected passenger either had Phoenix as a final destination or was delayed overnight in Phoenix.
- 6.26** A customer satisfaction survey is planned. The company carrying out the survey plans to contact 50 passengers selected at random from the 8000 passengers who arrived in Phoenix on the day described in the previous exercise. The airline knows that the survey results will not be favorable if too many people who were delayed overnight are included in the survey. Should the airline be worried? Write a few sentences explaining whether or not you think the airline should be worried, using relevant probabilities to support your answer.
- 6.27** A professor assigns five problems to be completed as homework. At the next class meeting, two of the five problems will be selected at random and collected for grading. You only completed the first three problems.
- What is the probability that you will be able to turn in both of the problems selected? (Hint: You can think of the problems as being labeled A, B, C, D, and E. Then one possible selection of two problems is A and B. If these are the two problems selected and you did problems A, B, and C, you will be able to turn in both problems. There are nine other possible selections to consider.)
 - Does the probability that you will be able to turn in both problems change if you had completed the last three problems instead of the first three problems? Explain.
 - What happens to the probability that you will be able to turn in both problems selected if you had completed four of the problems rather than just three?
- 6.28** Refer to the following information on full-term births in the United States over a given period of time:
- | Type of Birth | Number of Births |
|---------------|------------------|
| Single birth | 41,500,000 |
| Twins | 500,000 |
| Triplets | 5,000 |
| Quadruplets | 100 |
- Use this information to estimate the probability that a randomly selected pregnant woman who reaches full term
- Delivers twins
 - Delivers quadruplets
 - Gives birth to more than a single child
- 6.29** The report “*Teens, Social Media & Technology Overview 2015*” (Pew Research Center, April 9, 2015) summarized data from a large survey of teens age 13 to 17. Of those surveyed, 71% use Facebook and 52% use Instagram. Use these percentages to explain why the two events identified below cannot be mutually exclusive events.
- F = event that a randomly selected survey participant uses Facebook
and
 I = event that a randomly selected survey participant uses Instagram
- 6.30** According to *The Chronicle for Higher Education Almanac* (2016), there were 1,003,329 Associate degrees awarded by U.S. community colleges in the 2013–2014 academic year. A total of 613,034 of these degrees were awarded to women.
- If a person who received a degree in 2013–2014 was selected at random, what is the probability that the selected student is female?
 - What is the probability that the selected student is male?
- 6.31** The same issue of *The Chronicle for Higher Education Almanac* referenced in the previous exercise also reported the following information for Ph.D. degrees awarded by U.S. colleges in the 2013–2014 academic year:
- A total of 54,070 Ph.D. degrees were awarded.
 - 12,504 of these degrees were in the life sciences.
 - 9859 of these degrees were in the physical sciences.
 - The remaining degrees were in majors other than life or physical sciences.
- What is the probability that a randomly selected Ph.D. student who received a degree in 2013–2014
- received a degree in the life sciences?
 - received a degree that was not in a life or a physical science?
 - did not receive a degree in the physical sciences?
- 6.32** A deck of 52 playing cards is mixed well, and 5 cards are dealt.
- It can be shown that (disregarding the order in which the cards are dealt) there are 2,598,960 possible five-card hands, of which only 1287 are hands consisting entirely of spades. What is the probability that a hand will consist entirely of spades? What is the probability that a hand will consist entirely of a single suit?
 - It can be shown that exactly 63,206 hands contain only spades and clubs, with both suits represented. What is the probability that a hand consists entirely of spades and clubs with both suits represented?

- c. Using the result of Part (b), what is the probability that a hand contains cards from exactly two suits?
- 6.33** After all students have left the classroom, a statistics professor notices that four copies of the text were left under desks. At the beginning of the next class, the professor distributes the four books at random to the four students (1, 2, 3, and 4) who said they left books. One possible outcome is that 1 receives 2's book, 2 receives 4's book, 3 receives his or her own book, and 4 receives 1's book. This outcome can be abbreviated (2, 4, 3, 1).
- List the 23 other possible outcomes.
 - Which outcomes are contained in the event that exactly two of the books are returned to their correct owners? Assuming equally likely outcomes, what is the probability of this event?
- 6.34** Use the information given in the previous exercise to answer the following questions.
- What is the probability that exactly one of the four students receives his or her own book?
 - What is the probability that exactly three receive their own books?
 - What is the probability that at least two of the four students receive their own books?
- 6.35** The student council for a school of science and math has one representative from each of the five academic departments: biology (B), chemistry (C), mathematics (M), physics (P), and statistics (S). Two of these students are to be randomly selected for a university-wide student committee. The two students will be selected by placing five slips of paper (numbered 1, 2, 3, 4, and 5) into a bowl, mixing, and drawing out two of them.
- What are the 10 possible outcomes (simple events)?
 - From the description of the selection process, all simple events are equally likely. What is the probability of each simple event?
- c. What is the probability that one of the committee members is the statistics department representative?
- d. What is the probability that both committee members come from laboratory science departments?
- 6.36** A student placement center has requests from five students for interviews for employment with a consulting firm. Three of these students are math majors, and the other two students are statistics majors. Unfortunately, the interviewer only has time to talk with two of the students. These two will be randomly selected from among the five.
- What is the probability that both selected students are statistics majors?
 - What is the probability that both students selected are math majors?
 - What is the probability that at least one of the students selected is a statistics major?
 - What is the probability that the selected students have different majors?
- 6.37** Suppose that a six-sided die is weighted so that any even-numbered face is twice as likely to land face up as any odd-numbered face. Consider the chance experiment that consists of rolling this die.
- What are the probabilities of the six simple events?
(Hint: Denote these events by O_1, \dots, O_6 . Then $P(O_1) = p, P(O_2) = 2p, P(O_3) = p, \dots, P(O_6) = 2p$. Now use a condition on the sum of these probabilities to determine p .)
 - What is the probability that the number showing is an odd number? What is the probability that the number showing is at most three?
 - Now suppose that the die is weighted so that the probability of each simple event is proportional to the number showing on the corresponding upturned face; that is, $P(O_1) = c, P(O_2) = 2c, \dots, P(O_6) = 6c$. What are the probabilities of the six simple events? Calculate the probabilities of Part (b) for this die.

SECTION 6.4 Conditional Probability

Sometimes the knowledge that one event has occurred changes our assessment of the likelihood that another event occurs. For example, consider a population in which 0.1% of all individuals have a certain disease. It is difficult to tell if someone has the disease based on a physical exam, but there is a diagnostic test available. Unfortunately, the test is not always correct. Of those with positive test results, 80% actually have the disease and the other 20% who show positive test results are false-positives.

To put this in probability terms, consider the chance experiment of randomly selecting a person from the population. Define the following events:

E = event that the individual has the disease

F = event that the individual's diagnostic test is positive

We will use $P(E|F)$ to denote the probability of the event E given that the event F is known to have occurred. A new symbol has been used to indicate that a probability calculation has been made conditional on the occurrence of another event. The standard symbol for this is a vertical line, and it is read “given.” For example, we would say, “the probability that an individual has the disease given that the diagnostic test is positive.” This is represented symbolically as $P(\text{has disease}|\text{positive test})$ or $P(E|F)$. This probability is called a **conditional probability**.

The information provided about the disease and the diagnostic test implies that

$$\begin{aligned}P(E) &= 0.001 \\P(E|F) &= 0.8\end{aligned}$$

This means that before we have diagnostic test information, the occurrence of E is unlikely. However, once it is known that the test result is positive, the likelihood of the disease increases dramatically. (If this were not so, the diagnostic test would not be very useful!)

Example 6.11 College Housing Options

Understand the context ➤

At a small liberal arts college, students have three housing options. They can live in on-campus housing, off-campus housing, or at home with family. The following table gives the number of students in each housing option by year in school.

Consider the data ➤

	Freshman	Sophomore	Junior	Senior	Total
On-campus housing	150	160	140	150	600
Off-campus housing	100	90	120	125	435
Home with family	125	140	150	150	565
Total	375	390	410	425	1600

Consider each of the following statements, and make sure that you see how each follows from the information in the table:

1. There are 160 sophomores who live in on-campus housing.
2. The number of seniors who live in off-campus housing is 125.
3. There are 435 students who live in off-campus housing.
4. There are 410 juniors at the college.
5. The total number of students at the college is 1600.

Each week, the college president selects a student at random and invites him or her to have lunch with her to discuss various issues that might be of concern to them. She feels that random selection will give her the greatest chance of hearing from a diverse group of students. What is the probability that a randomly selected student is a senior who lives on campus? Assuming that each student is equally likely to be selected, we can calculate this probability as follows:

$$P(\text{senior who lives on campus}) = \frac{\text{number of seniors who live on campus}}{\text{total number of students}} = \frac{150}{1600} = 0.09375$$

Now suppose that the president’s assistant records not only the student’s name but also the student’s year in school. The assistant has indicated that the selected student is a senior. Does this information change our assessment of the likelihood that the selected student lives on campus? Because 150 of the 425 seniors live on campus, this suggests that

$$P(\text{live on campus|senior}) = \frac{\text{number of seniors who live on campus}}{\text{total number of seniors}} = \frac{150}{425} = 0.3529$$

The probability is calculated in this way because we know that the selected student is one of 425 seniors, each of whom is equally likely to have been the one selected.

Do the work ➤

- Interpret the results ➤ The interpretation of this conditional probability is that if we were to repeat the chance experiment of selecting a student at random, about 35.29% of the selections that resulted in a senior being selected would also result in the selection of someone who lives on campus.

Example 6.12 GFI Switches

- Understand the context ➤ A GFI (ground fault interrupt) switch turns off power to a system in the event of an electrical malfunction. A spa manufacturer currently has 25 spas in stock, each equipped with a single GFI switch. Suppose that two different companies supplied the switches, and that some of the switches are defective, as summarized in the following table:

Consider the data ➤

	Nondefective	Defective	Total
Company 1	10	5	15
Company 2	8	2	10
Total	18	7	25

A spa is randomly selected for testing. Consider the following events:

$$\begin{aligned} E &= \text{event that GFI switch in the selected spa is from Company 1} \\ F &= \text{event that GFI switch in the selected spa is defective} \end{aligned}$$

- Do the work ➤ Using the information in the table, we can calculate the following probabilities:

$$P(E) = \frac{15}{25} = 0.60 \quad P(F) = \frac{7}{25} = 0.28 \quad P(E \text{ and } F) = P(E \cap F) = \frac{5}{25} = 0.20$$

Now suppose that testing reveals a defective switch. This means that the chosen spa is one of the seven in the “defective” column. How likely is it that the switch came from the first company? Because five of the seven defective switches are from Company 1,

$$P(E|F) = P(\text{company 1}|\text{defective}) = \frac{5}{7} = 0.714$$

Notice that this is larger than the unconditional probability $P(E)$. This is because Company 1 has a much higher defective rate than Company 2.

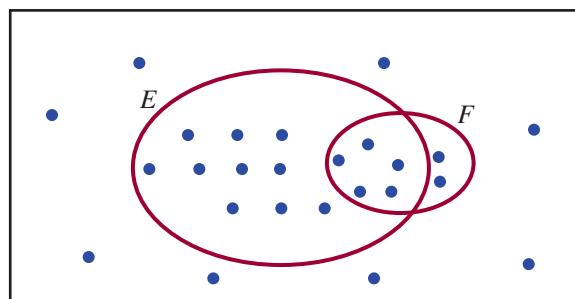
An alternative expression for the conditional probability is

$$P(E|F) = \frac{5}{7} = \frac{5/25}{7/25} = \frac{P(E \text{ and } F)}{P(F)} = \frac{P(E \cap F)}{P(F)}$$

Notice that $P(E|F)$ is a ratio of two previously specified probabilities. It is the probability that both events occur divided by the probability of the “conditioning event” F . Additional insight comes from the Venn diagram of Figure 6.8. Once it is known that the outcome lies in F , the chance that E also occurs is the “size” of $(E \text{ and } F)$ relative to the size of F .

FIGURE 6.8

Venn diagram for Example 6.12 (each dot represents one GFI switch).



The previous example leads to a general definition of conditional probability.

DEFINITION

Conditional probability: Suppose that E and F are two events with $P(F) > 0$.

The conditional probability of the event E given that the event F has occurred, denoted by $P(E|F)$, is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Notice the requirement that $P(F) > 0$. In addition to the standard warning about division by 0, there is another reason for requiring $P(F)$ to be positive. If the probability of F were 0, the event F would never occur. Therefore, it would not make sense to calculate the probability of another event conditional on F having occurred.

Example 6.13 Surviving a Heart Attack

Understand the context ➤

Medical guidelines recommend that a hospitalized patient who suffers a cardiac arrest (a serious condition where the heart stops beating) should receive defibrillation (an electric shock to the heart) within 2 minutes. The paper “Delayed Time to Defibrillation After In-Hospital Cardiac Arrest” (*The New England Journal of Medicine* [2008]: 9–17) describes a study of the time to defibrillation for hospitalized patients in hospitals of various sizes.

The authors examined medical records of 6716 patients who suffered cardiac arrest while hospitalized, recording the size of the hospital and whether or not defibrillation occurred in 2 minutes or less. Data from this study are summarized in the accompanying table.

Consider the data ➤

Hospital Size	Time to Defibrillation		
	2 minutes or less	More than 2 minutes	Total
Small (Less than 250 beds)	1,124	576	1,700
Medium (250–499 beds)	2,178	886	3,064
Large (500 or more beds)	1,387	565	1,952
Total	4,689	2,027	6,716

In this example, we assume that these data are representative of the larger group of all hospitalized patients who suffer a cardiac arrest. Suppose that a hospitalized patient who suffered a cardiac arrest is selected at random. The following events are of interest:

S = event that the selected patient is at a small hospital

M = event that the selected patient is at a medium-sized hospital

L = event that the selected patient is at a large hospital

D = event that the selected patient receives defibrillation in 2 minutes or less

Do the work ➤ We can use the information in the table to calculate

$$P(D) = \frac{4689}{6716} = 0.698$$

This probability is interpreted as the proportion of hospitalized patients who suffer cardiac arrest that receive defibrillation in 2 minutes or less. This means that 69.8% of these patients receive timely defibrillation.

Now suppose it is known that the selected patient was at a small hospital. How likely is it that this patient received defibrillation in 2 minutes or less? To answer this question, we need to compute $P(D|S)$, the probability of defibrillation in 2 minutes or less *given* that the patient is at a small hospital. Using the formula for conditional probability, we know

$$P(D|S) = \frac{P(D \cap S)}{P(S)}$$

From the information in the table, we can calculate

$$P(D \cap S) = \frac{1124}{6716} = 0.167$$

$$P(S) = \frac{1700}{6716} = 0.253$$

and so

$$P(D|S) = \frac{P(D \cap S)}{P(S)} = \frac{0.167}{0.253} = 0.660$$

Notice that this is smaller than the unconditional probability, $P(D) = 0.698$. This tells us that there is a smaller probability of timely defibrillation at a small hospital.

Two other conditional probabilities of interest are

$$P(D|M) = \frac{P(D \cap M)}{P(M)} = \frac{2178/6716}{3064/6716} = 0.711$$

and

$$P(D|L) = \frac{P(D \cap L)}{P(L)} = \frac{1387/6716}{1952/6716} = 0.711$$

From this, we see that the probability of timely defibrillation is the same for patients at medium-sized and large hospitals, and that this probability is higher than that for patients at small hospitals.

It is also possible to calculate $P(L|D)$, the probability that a patient is at a large hospital given that the patient received timely defibrillation:

$$P(L|D) = \frac{P(L \cap D)}{P(D)} = \frac{1387/6716}{4689/6716} = 0.296$$

Interpret the results ➤

Let's look carefully at the interpretation of some of these probabilities:

1. $P(D) = 0.698$ is interpreted as the proportion of *all hospitalized patients* suffering cardiac arrest who receive timely defibrillation. Approximately 69.8% of these patients receive timely defibrillation.
2. $P(D \cap L) = \frac{1387}{6716} = 0.207$ is the proportion of *all hospitalized patients* who suffer cardiac arrest who are at a large hospital *and* who receive timely defibrillation.
3. $P(D|L) = 0.711$ is the proportion of *patients at large hospitals* suffering cardiac arrest who receive timely defibrillation.
4. $P(L|D) = 0.296$ is the proportion of *patients who receive timely defibrillation* who were at large hospitals.

Notice the difference between the unconditional probabilities in Interpretations 1 and 2 and the conditional probabilities in Interpretations 3 and 4. The reference point for the unconditional probabilities is the entire group of interest (all hospitalized patients suffering cardiac arrest), whereas the conditional probabilities are interpreted in a more restricted context defined by the “given” event.

Example 6.14 demonstrates the calculation of conditional probabilities and also makes the point that we must be careful when translating real-world problems—especially probability problems—into mathematical form.

Example 6.14 Two-Kid Families

Understand the context ➤

Consider the population of all families with two children. Representing the sex of each child using G for girl and B for boy results in four possibilities: BB, BG, GB, GG. The sex

information is sequential, with the first letter indicating the sex of the older child. A family having a girl first and then a boy is denoted GB. If we assume that a child is equally likely to be a boy or a girl, each of the four possibilities in the sample space for the chance experiment that selects a family at random from families with two children is equally likely. Consider the following two questions:

1. What is the probability that the selected family has two girls, given that the family has at least one girl?
2. What is the probability that the selected family has two girls, given that the older child is a girl?

Do the work ➤ To many people, these questions *appear* to be identical. However, by calculating the appropriate probabilities, we can see that they are actually different.

For question 1 we calculate

$$\begin{aligned} P(\text{family has two girls} \mid \text{family has at least one girl}) \\ = \frac{P(\text{family has two girls and family has at least one girl})}{P(\text{family has at least one girl})} \\ = \frac{P(GG)}{P(GG \text{ or } BG \text{ or } GB)} \\ = \frac{1/4}{3/4} = \frac{0.25}{0.75} = 0.333 \end{aligned}$$

For question 2 we calculate

$$\begin{aligned} P(\text{family has two girls} \mid \text{family has a girl as the older child}) \\ = \frac{P(\text{family has two girls and family has a girl as the older child})}{P(\text{family has a girl as the older child})} \\ = \frac{P(GG)}{P(GB \text{ or } GG)} \\ = \frac{1/4}{1/2} = \frac{0.25}{0.50} = 0.50 \end{aligned}$$

Notice that these two probabilities are not the same. We would interpret them as follows:

1. One-third of families with two children that have at least one girl have two girls.
2. One-half of families with two children whose oldest child is a girl have two girls.

As we mentioned in the opening paragraphs of this section, one of the most important practical uses of conditional probability is in making diagnoses. Your mechanic diagnoses your car by hooking it up to a machine and reading the pressures and speeds of the various components. Doctors observe characteristics of their patients in an attempt to determine whether or not their patients have a certain disease. Many diseases are not actually observable—or at least not easily observable—and often the doctor must make a probabilistic judgment. As we will see in the next example, conditional probability plays a large role in evaluating diagnostic techniques.

Example 6.15 Diagnosing Tuberculosis

Understand the context ➤

To illustrate the calculations involved in evaluating a diagnostic test, we consider the case of tuberculosis (TB), an infectious disease that typically attacks lung tissue. Before 1998, culturing was the standard method for diagnosing TB. This method always resulted in a correct diagnosis, but it took 10 to 15 days to obtain the result. In 1998, investigators evaluated a DNA technique that turned out to be much faster (“[LCx: A Diagnostic Alternative for](#)

the Early Detection of *Mycobacterium tuberculosis* Complex," *Diagnostic Microbiology and Infectious Diseases* [1998]: 259–264).

The DNA technique for detecting tuberculosis was evaluated by comparing results from the test to the existing gold standard, with the following results for 207 patients exhibiting symptoms:

Consider the data ➤

	Gold Standard Test Shows Has Tuberculosis	Gold Standard Test Shows Does Not Have Tuberculosis
DNA Positive Indication	14	0
DNA Negative Indication	12	181

Converting these data to proportions and adding the column and row totals into the table, we get the following information:

Do the work ➤

	Gold Standard Test Shows Has Tuberculosis	Gold Standard Test Shows Does Not Have Tuberculosis	Total
DNA Positive Indication	0.068	0.000	0.068
DNA Negative Indication	0.058	0.874	0.932
Total	0.126	0.874	1.000

A quick look at the table indicates that the DNA technique seems to be effective as a diagnostic test. Patients who tested positive with the DNA test were also positive with the gold standard test in every case. Patients who tested negative with the DNA test generally tested negative with the gold standard test, but the table also indicates some false-negative results for the DNA test.

Now think about a randomly selected individual who is tested for TB. Consider the following events:

T = event that the individual has tuberculosis

N = event that the DNA test is negative

What is the probability that a person has tuberculosis, given that the DNA test was negative? To answer the question, we calculate $P(T|N)$, the probability of the event T given that the event N has occurred. We calculate this probability as follows:

$$\begin{aligned} P(T|N) &= P(\text{tuberculosis}|\text{negative DNA test}) \\ &= \frac{P(\text{tuberculosis} \cap \text{negative DNA test})}{P(\text{negative DNA test})} \\ &= \frac{0.058}{0.932} \\ &= 0.062 \end{aligned}$$

Interpret the results ➤

Notice that 0.126 is the proportion of those tested who had tuberculosis. The added information provided by the diagnostic test has altered probability—and provides some measure of relief for patients who test negative. Once it is known that the test result is negative, the estimated probability of having tuberculosis is only about half as large.

EXERCISES 6.38 - 6.52

• Data set available online

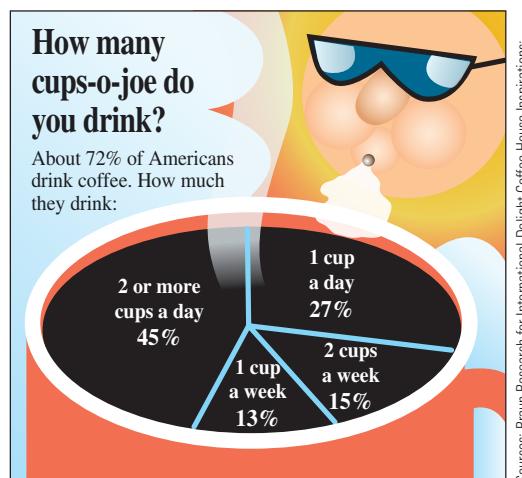
- 6.38** Two different airlines have a flight from Los Angeles to New York that departs each weekday morning at a certain time. Suppose that E denotes the event that

the first airline's flight is fully booked on a particular day, and F denotes the event that the second airline's flight is fully booked on that same day.

Suppose that $P(E) = 0.7$, $P(F) = 0.6$, and

$$P(E \cap F) = 0.54.$$

- a. Calculate $P(E|F)$, the probability that the first airline's flight is fully booked given that the second airline's flight is fully booked. (Hint: See Example 6.12.)
 - b. Calculate $P(F|E)$.
- 6.39** The article "Chances Are You Know Someone with a Tattoo, and He's Not a Sailor" (*Associated Press, June 11, 2006*) included results from a survey of adults aged 18 to 50. The accompanying data are consistent with summary values given in the article.
- | | At Least One Tattoo | No Tattoo |
|-----------|---------------------|-----------|
| Age 18–29 | 18 | 32 |
| Age 30–50 | 6 | 44 |
- Assuming these data are representative of adult Americans and that an adult American is selected at random, use the given information to estimate the following probabilities. (Hint: See Example 6.13.)
- a. $P(\text{tattoo})$
 - b. $P(\text{tattoo}|\text{age } 18\text{--}29)$
 - c. $P(\text{tattoo}|\text{age } 30\text{--}50)$
 - d. $P(\text{age } 18\text{--}29|\text{tattoo})$
- 6.40** The accompanying data are from the article "Characteristics of Buyers of Hybrid Honda Civic IMA: Preferences, Decision Process, Vehicle Ownership, and Willingness-to-Pay" (*Institute for Environmental Decisions, November 2006*). Each of 311 people who purchased a Honda Civic was classified according to sex and whether the car purchased had a hybrid engine or not.
- | | Hybrid | Not Hybrid |
|--------|--------|------------|
| Male | 77 | 117 |
| Female | 34 | 83 |
- Suppose one of these 311 individuals is to be selected at random.
- a. Find the following probabilities:
 - i. $P(\text{male})$
 - ii. $P(\text{hybrid})$
 - iii. $P(\text{hybrid}|\text{male})$
 - iv. $P(\text{hybrid}|\text{female})$
 - v. $P(\text{female}|\text{hybrid})$
 - b. For each of the probabilities calculated in Part (a), write a sentence interpreting the probability. (Hint: See Example 6.13.)
- 6.41** Using the probabilities calculated in the previous exercise, determine if the probabilities $P(\text{hybrid}|\text{male})$ and $P(\text{male}|\text{hybrid})$ equal. If not, write a sentence or two explaining the difference between these two probabilities.
- 6.42** The following graphical display is similar to one that appeared in *USA TODAY* (January 8, 2010).



Sources: Brain Research for International Delight Coffee House Inspirations; USA TODAY, January 8, 2010.

Use the information in this display to answer the following questions. Assume that the percentages in the graph are representative of adult Americans.

- a. What is the probability that a randomly selected adult American drinks coffee?
- b. The display associates 45% with the category "two or more cups a day." For the chance experiment that consists of selecting an adult American at random, is 0.45 the probability that the selected individual drinks two or more cups of coffee a day, or is it the conditional probability that the selected individual drinks two or more cups a day given that he or she drinks coffee? Explain.

- 6.43** The article "Americans Growing More Concerned About Head Injuries in Football" (*theharrispoll.com, December 21, 2015*) describes a survey of 2096 adult Americans. Survey participants were asked if they were football fans and also if they agreed or disagreed that the rules that the National Football League adopted in 2010 designed to limit head injuries have been effective. Data from the survey are summarized in the table below.

	Agree	Disagree	Total
Football Fan	693	523	1,216
Not a Football Fan	229	651	880
Total	922	1,174	2,096

Suppose that a survey participant is to be selected at random. Consider the following events:

A = event selected participant agrees that the rules have been effective

D = event selected participant disagrees that the rules have been effective

F = event selected participant is a football fan

Calculate the following probabilities

- a. $P(A)$
- b. $P(D)$
- c. $P(A|F)$
- d. $P(A|F^C)$

- 6.44** The events E and T_i are defined as $E =$ the event that someone who is out of work and actively looking for work will find a job within the next month and $T_i =$ the event that someone who is currently out of work has been out of work for i months. For example, T_2 is the event that someone who is out of work has been out of work for 2 months.

The following conditional probabilities are approximate and were read from a graph in the paper “**The Probability of Finding a Job**” (*American Economic Review: Papers & Proceedings* [2008]: 268–273):

$$\begin{array}{ll} P(E|T_1) = 0.30 & P(E|T_2) = 0.24 \\ P(E|T_3) = 0.22 & P(E|T_4) = 0.21 \\ P(E|T_5) = 0.20 & P(E|T_6) = 0.19 \\ P(E|T_7) = 0.19 & P(E|T_8) = 0.18 \\ P(E|T_9) = 0.18 & P(E|T_{10}) = 0.18 \\ P(E|T_{11}) = 0.18 & P(E|T_{12}) = 0.18 \end{array}$$

- a. Interpret the following two probabilities:
 - i. $P(E|T_1) = 0.30$
 - ii. $P(E|T_6) = 0.19$
- b. Construct a graph of $P(E|T_i)$ versus i . That is, plot $P(E|T_i)$ on the y -axis and $i = 1, 2, \dots, 12$ on the x -axis.
- c. Write a few sentences about how the probability of finding a job in the next month changes as a function of length of unemployment.

- 6.45** The newspaper article “**Folic Acid Might Reduce Risk of Down Syndrome**” (*USA TODAY*, September 29, 1999) makes the following statement: “Older women are at a greater risk of giving birth to a baby with Down Syndrome than are younger women. But younger women are more fertile, so most children with Down Syndrome are born to mothers under 30.”

Let D = event that a randomly selected baby is born with Down Syndrome and Y = event that a randomly selected baby is born to a young mother (under age 30). For each of the following probability statements, indicate whether the statement is consistent with the quote from the article, and if not, explain why not.

- a. $P(D|Y) = 0.001$, $P(D|Y^c) = 0.004$, $P(Y) = 0.7$
- b. $P(D|Y) = 0.001$, $P(D|Y^c) = 0.001$, $P(Y) = 0.7$
- c. $P(D|Y) = 0.004$, $P(D|Y^c) = 0.004$, $P(Y) = 0.7$
- d. $P(D|Y) = 0.001$, $P(D|Y^c) = 0.004$, $P(Y) = 0.4$
- e. $P(D|Y) = 0.001$, $P(D|Y^c) = 0.001$, $P(Y) = 0.4$
- f. $P(D|Y) = 0.004$, $P(D|Y^c) = 0.004$, $P(Y) = 0.4$

- 6.46** Suppose that an individual is randomly selected from the population of all adult males living in the United States. Let A be the event that the selected individual is over 6 feet in height, and let B be the event that the selected individual is a professional basketball player. Which do you think is larger, $P(A|B)$ or $P(B|A)$? Why?

- 6.47** Is ultrasound a reliable method for determining the gender of an unborn baby? The accompanying data on 1000 births are consistent with summary values that appeared in the *Journal of Statistics Education* (“**New Approaches to Learning Probability in the First Statistics Course**,” 2001).

	Ultrasound Predicted Female	Ultrasound Predicted Male
Actual Gender Is Female	432	48
Actual Gender Is Male	130	390

- a. Use the given information to estimate the probability that a newborn baby is female, given that the ultrasound predicted the baby would be female.
- b. Use the given information to estimate the probability that a newborn baby is male, given that the ultrasound predicted the baby would be male.
- c. Based on your answers to Parts (a) and (b), do you think that a prediction that a baby is male and a prediction that a baby is female are equally reliable? Explain.

- 6.48** The paper “**Accuracy and Reliability of Self-Reported Weight and Height in the Sister Study**” (*Public Health Nutrition* [2012]: 989–999) investigates whether women accurately report their weight. The table below is based on comparing actual weight to self-reported weight for women participating in a large-scale medical study. Each participant was classified into a category describing accuracy of reported weight and also by age.

Age	Accuracy Category			
	Under-reported by more than 7 lbs.	Under-reported by between 4 and 7 lbs.	Reported weight within 3 lbs.	Over-reported by 4 or more pounds
< 45 years	297	373	1,594	207
45–54 years	763	902	4,130	510
55–64 years	677	966	4,383	545
65+ years	263	444	2,285	300

Assume that it is reasonable to consider these data representative of adult women in the United States. Consider the following conclusion:

Most women reported their weight to within 3 pounds of their actual weight. Older women were less likely to under-report their weight and more likely to over-report their weight than younger women.

Provide a justification for this conclusion. Use the information in the table to calculate relevant probabilities.

- 6.49** The report “[2015 Utah Seat Belt Use Survey](#)” ([Utah Department of Public Safety—Highway Safety Office, September 14, 2015](#)) stated that based on observing a large number of vehicle occupants, the estimated percentage of Utah drivers and passengers who wear seatbelts is 87.2%. The report also gave information on seat belt use by sex and by whether the vehicle is traveling in an urban or rural area. The information in the following table is consistent with summary values given in the report.

	Urban Areas		Rural Areas	
	Male	Female	Male	Female
Seat belt	871	928	770	837
No Seat Belt	129	72	230	163

Assume that these data are representative of drivers and passengers in Utah. Consider the following conclusion:

Females are more likely to wear seat belts than males in both urban and rural areas. The difference in the percentage of females and the percentage of males who wear seat belts is greater for rural areas than for urban areas.

Provide a justification for this conclusion. Use the information in the table to calculate relevant probabilities.

- 6.50** The National Highway Traffic Safety Administration requires each U.S. state to carry out an observational study to assess the level of seat belt use in the state. The report “[2015 Utah Seat Belt Use Survey](#)” ([Utah Department of Public Safety, September 14, 2015](#)) summarized data from the study done in Utah. The proportions in the accompanying table are based on observations of over 25,000 drivers and passengers.

	Uses Seatbelt	Does Not Use Seat Belt
Male	0.423	0.077
Female	0.452	0.048

Assume that these proportions are representative of adults in Utah and that an adult from Utah is selected at random.

- a. What is the probability that the selected adult uses a seat belt?
 - b. What is the probability that the selected adult uses a seat belt given that the individual selected is male?
- 6.51** Use the information given in the previous exercise to answer the following questions.
- a. What is the probability that the selected adult does not use a seat belt given that the selected individual is female?

- b. What is the probability that the selected individual is female given that the selected individual does not use a seat belt?
- c. Are the probabilities from Parts (a) and (b) equal? Write a couple of sentences explaining why or why not.

- 6.52** The paper “[Good for Women, Good for Men, Bad for People: Simpson’s Paradox and the Importance of Sex-Specific Analysis in Observational Studies](#)” ([Journal of Women’s Health and Gender-Based Medicine \[2001\]: 867–872](#)) described the results of a medical study in which one treatment was shown to be better for men and better for women than a competing treatment. However, if the data for men and women are combined, it appears as though the competing treatment is better.

To see how this can happen, consider the accompanying data tables constructed from information in the paper. Subjects in the study were given either Treatment A or Treatment B, and survival was noted. Let S be the event that a patient selected at random survives, A be the event that a patient selected at random received Treatment A, and B be the event that a patient selected at random received Treatment B.

- a. The following table summarizes data for men and women combined:

	Survived	Died	Total
Treatment A	215	85	300
Treatment B	241	59	300
Total	456	144	

- i. Find $P(S)$.
 - ii. Find $P(S|A)$.
 - iii. Find $P(S|B)$.
 - iv. Which treatment appears to be better?
- b. Now consider the summary data for the men who participated in the study:

	Survived	Died	Total
Treatment A	120	80	200
Treatment B	20	20	40
Total	140	100	

- i. Find $P(S)$.
 - ii. Find $P(S|A)$.
 - iii. Find $P(S|B)$.
 - iv. Which treatment appears to be better?
- c. Now consider the summary data for the women who participated in the study:

	Survived	Died	Total
Treatment A	95	5	100
Treatment B	221	39	260
Total	316	144	

- i. Find $P(S)$.
 - ii. Find $P(S|A)$.
 - iii. Find $P(S|B)$.
 - iv. Which treatment appears to be better?
- d. You should have noticed from Parts (b) and (c) that for both men and women, Treatment A appears to be better. But in Part (a), when the data for men and women are combined, it

looks like Treatment B is better. This is an example of what is called Simpson's paradox. Write a brief explanation of why this apparent inconsistency occurs for this data set. (Hint: Do men and women respond similarly to the two treatments?)

SECTION 6.5 Independence

In Section 6.4, we saw that knowing one event has occurred can change our assessment of the probability that some other event has occurred. However, it is also possible that knowing that one event has occurred will not change our assessment of the probability of occurrence of a second event.

Example 6.16 Mortgage Choices

Understand the context ➤

A bank offers both adjustable-rate and fixed-rate mortgage loans on residential property, which it classifies into three categories: single-family houses, condominiums, and multi-family dwellings. The following table, called a *joint probability table*, displays probabilities based on the bank's long-run lending behavior:

Consider the data ➤

	Single-Family	Condo	Multifamily	Total
Adjustable	0.40	0.21	0.09	0.70
Fixed	0.10	0.09	0.11	0.30
Total	0.50	0.30	0.20	1.00

From the table we see that 70% of all mortgages are adjustable rate, 50% of all mortgages are for single-family houses, 40% of all mortgages are adjustable rate for single-family houses (adjustable-rate *and* single-family), and so on.

Define the events E and F as

E = event that a mortgage is adjustable rate

F = event that a mortgage is for a single-family house

Do the work ➤ Then $P(E) = 0.70$ and

$$P(E|F) = \frac{P(E \text{ and } F)}{P(F)} = \frac{0.40}{0.50} = 0.80$$

Interpret the results ➤ This means that 80% of loans made for single-family houses are adjustable-rate loans. Notice that $P(E|F)$ is larger than the original (unconditional) probability $P(E) = 0.70$. Also,

$$P(F|E) = \frac{P(E \text{ and } F)}{P(E)} = \frac{0.40}{0.70} = 0.571$$

which is larger than the unconditional probability $P(F) = 0.5$. Knowing that E (adjustable rate) has occurred has changed our assessment of how likely it is that F (single-family house) has also occurred.

Now consider another event C defined as

C = event that a mortgage is for a condominium

then

$$P(E|C) = \frac{P(E \text{ and } C)}{P(C)} = \frac{0.21}{0.30} = 0.70$$

Notice that $P(E|C) = P(E)$. In this case, knowing that a mortgage is for a condominium doesn't change our assessment of the probability that the mortgage has an adjustable interest rate.

When two events E and F are such that $P(E|F) = P(E)$, the probability that event E has occurred is the same after we learn that F has occurred as it was before we knew that F had occurred. If this is the case, we say that E and F are independent of one another.

DEFINITIONS

Independent events: Two events E and F are **independent** if

$$P(E|F) = P(E)$$

Dependent events: If two events E and F are not independent, they are **dependent** events.

If $P(E|F) = P(E)$, it is also true that $P(F|E) = P(F)$, and vice versa.

Independence of events E and F also implies the following additional three relationships:

$$P(\text{not } E|F) = P(\text{not } E)$$

$$P(E|\text{not } F) = P(E)$$

$$P(\text{not } E|\text{not } F) = P(\text{not } E)$$

This means that if E and F are independent, nothing we learn about F will change the probability of E or of $\text{not } E$.

Recall that the formula for conditional probability is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

which can be rearranged to give

$$P(E \cap F) = P(E|F)P(F)$$

When E and F are independent, $P(E|F) = P(E)$, so it follows that if E and F are independent,

$$P(E \cap F) = P(E|F)P(F) = P(E)P(F)$$

This result is called the multiplication rule for two independent events.

Multiplication Rule for Two Independent Events

If the events E and F are independent,

$$P(E \cap F) = P(E)P(F)$$

Example 6.17 Hitchhiker's Thumb

Understand the context ➤

In humans, there is a gene that controls a characteristic known as hitchhiker's thumb. Hitchhiker's thumb is the ability to bend the last joint of the thumb back at an angle of 60° or more. Whether a child has hitchhiker's thumb is determined by two random events: which of two alleles is contributed by the father and which of two alleles is contributed by the mother. You can think of these alleles as a parental vote of yes or no on the hitchhiker's thumb gene. If the votes by the two parents disagree, the dominant allele wins and the child does not have hitchhiker's thumb. These two random events, the results of cell division in two different biological parents, are independent of each other.

Do the work ➤

Suppose that there is a 0.10 probability that a parent contributes a positive hitchhiker's thumb allele. Because the events are independent, the probability that both parents contribute a positive hitchhiker's thumb allele, H^+ , to the offspring is

$$\begin{aligned} P(\text{mother contributes } H^+ \cap \text{father contributes } H^+) \\ = P(\text{mother contributes } H^+)P(\text{father contributes } H^+) \\ = (0.10)(0.10) \\ = 0.01 \end{aligned}$$

Interpret the results ➤

This means that only about 1% of all children in the population would have hitchhiker's thumb.



Bluehand/Dreamstime.com

Understand the context ➤

Example 6.18 Curious Guppies

Let's look at another example, this time from the field of animal behavior, to illustrate how an investigator could judge whether two events are independent. In a number of fish species, including guppies, a phenomenon known as predator inspection has been reported. It is thought that predator inspection allows a guppy to assess the risk posed by a potential predator. In a typical inspection a guppy moves toward a predator, presumably to acquire information and then (hopefully) depart to inspect again another day.

Investigators have observed that guppies sometimes approach and inspect a predator in pairs. Suppose that it is not known whether these predator inspections are independent or whether the guppies are operating as a team. We can use p to denote the probability that an individual guppy will inspect a predator. Suppose E_1 is the event that guppy 1 will approach and inspect a predator and E_2 is the event that guppy 2 will approach and inspect a predator. Then the probability of guppies 1 and 2 both approaching the predator at the same time by chance if they are acting *independently* is

$$P(E_1 \cap E_2) = P(E_1)P(E_2) = p \cdot p = p^2$$

Interpret the results ➤

Based on our analysis, if the inspections are in fact independent, we would expect the proportion of times that two guppies happen to simultaneously inspect a predator to be equal to the square of the proportion of times that a single fish does so. For example, if the probability a single guppy inspects a predator is 0.3, we would expect the proportion of the time that two guppies would inspect a predator simultaneously to be $(0.3)^2 = 0.09$. Based on observations of the inspection behavior of a large number of guppies, scientists found that inspections by two guppies occurred much more often than 9% of the time. As a result, they concluded that the inspection behavior of guppies does not appear to be independent.

The concept of independence extends to more than two events. Consider three events, E_1 , E_2 , and E_3 . Then independence means not only that

$$\begin{aligned} P(E_1|E_2) &= P(E_1) \\ P(E_3|E_2) &= P(E_3) \end{aligned}$$

and so on but also that

$$\begin{aligned} P(E_1|E_2 \text{ and } E_3) &= P(E_1) \\ P(E_1 \text{ and } E_3|E_2) &= P(E_1 \text{ and } E_3) \end{aligned}$$

and so on. There is also a multiplication rule for more than two independent events.

The independence of more than two events is an important concept in studying complex systems with many components. If these components are critical to the operation of a machine, assessing the probability of the machine's failure is undertaken by analyzing the failure probabilities of the individual components.

Multiplication Rule for k Independent Events

Events E_1, E_2, \dots, E_k are **independent** if knowledge that any of the events have occurred does not change the probabilities that any particular one or more of the other events has occurred.

Independence implies that

$$P(E_1 \cap E_2 \cap \cdots \cap E_k) = P(E_1)P(E_2) \cdots P(E_k)$$

This means that when events are independent, the probability that all occur together is the product of the individual probabilities. This relationship also holds if one or more of the events is replaced by its complement.

In Example 6.19 we take a rather simplified view of a desktop computer to illustrate the use of the multiplication rule.

Example 6.19 Computer Configurations

Understand the context ➤

Suppose that a desktop computer system consists of a monitor, a mouse, a keyboard, the computer processor itself, and storage devices such as a disk drive. Most computer system problems due to manufacturing defects occur soon in the system's lifetime. Purchasers of new computer systems are advised to turn their computers on as soon as they are purchased and then to let them run for a few hours to see if any problems occur.

Suppose

- E_1 = event that a newly purchased monitor is not defective
- E_2 = event that a newly purchased mouse is not defective
- E_3 = event that a newly purchased disk drive is not defective
- E_4 = event that a newly purchased computer processor is not defective

Suppose these four events are independent, with

$$P(E_1) = P(E_2) = 0.98 \quad P(E_3) = 0.95 \quad P(E_4) = 0.99$$

Do the work ➤ The probability that all these components are not defective and that the system will operate properly is then

$$\begin{aligned} P(E_1 \cap E_2 \cap E_3 \cap E_4) &= P(E_1)P(E_2)P(E_3)P(E_4) \\ &= (0.98)(0.98)(0.94)(0.99) \\ &= 0.89 \end{aligned}$$

Interpret the result ➤

We interpret this probability as follows: In the long run, 89% of such systems will run properly when tested shortly after purchase. (In reality, the reliability of these components is much higher than the numbers used in this example!) The probability that all components except the monitor will run properly is

$$\begin{aligned} P(E_1^C \cap E_2 \cap E_3 \cap E_4) &= P(E_1^C)P(E_2)P(E_3)P(E_4) \\ &= (1 - P(E_1))P(E_2)P(E_3)P(E_4) \\ &= (0.02)(0.98)(0.94)(0.99) \\ &= 0.018 \end{aligned}$$

Sampling With and Without Replacement

One area of statistics where the rules of probability are important is sampling. As we saw in Chapter 2, a well-designed sampling plan allows investigators to make inferences about a population based on information from a sample. Sampling methods can be classified into two categories: **sampling with replacement** and **sampling without replacement**.

Most inferential methods presented in an introductory statistics course are based on the assumption of sampling *with replacement*, but when sampling from real populations,

we almost always sample *without* replacement. This seemingly contradictory practice can be a source of confusion. Fortunately, under certain conditions, the distinction between sampling with and without replacement is not important.

DEFINITIONS

Sampling with replacement: Once selected, an individual or object is put back into the population before the next selection.

Sampling without replacement: Once selected, an individual or object is not returned to the population prior to subsequent selections.

Example 6.20 Sampling With and Without Replacement

Consider the process of selecting three cards from a standard deck of playing cards. This selection can be made in two ways. One method is to shuffle the cards and then deal three cards off the top of the deck. This would be sampling without replacement. A second method, rarely seen in real games, is to select a card at random, note which card is observed, replace it in the deck, and shuffle before selecting the next card. This method is sampling with replacement.

From the standpoint of probability, sampling with and without replacement are analyzed differently. To see this, consider these events:

$$\begin{aligned} H_1 &= \text{event that the first card is a heart} \\ H_2 &= \text{event that the second card is a heart} \\ H_3 &= \text{event that the third card is a heart} \end{aligned}$$

For sampling with replacement, the probability of H_3 is 0.25, regardless of whether either H_1 or H_2 occurs, because replacing selected cards gives the same deck for the third selection as for the first two selections. Whether either of the first two cards was a heart has no bearing on the third card selected, so the three events, H_1 , H_2 , and H_3 , are independent.

When sampling is without replacement, the chance of getting a heart on the third card selected does depend on the results of the first two selections. If both H_1 and H_2 occur, only 11 of the 50 remaining cards are hearts. Because any one of these 50 has the same chance of being selected, the probability of H_3 in this case is

$$P(H_3 | H_1 \text{ and } H_2) = \frac{11}{50} = 0.22$$

Alternatively, if neither of the first 2 cards is a heart, then all 13 hearts remain in the deck for the third draw, so

$$P(H_3 | \text{not } H_1 \text{ and not } H_2) = \frac{13}{50} = 0.26$$

Information about the occurrence of H_1 and H_2 affects the chance that H_3 has occurred. For sampling without replacement, the three events described here are not independent.

In opinion polls and other types of surveys, sampling is almost always done without replacement. For this method of sampling, the results of successive selections are not independent of one another. However, Example 6.21 suggests that, *under certain conditions*, the fact that selections in sampling without replacement are not independent is not a cause for concern.

Example 6.21 Independence and Sampling Without Replacement

Suppose that a shipment of 10,000 computer chips used in graphing calculators consists of 2500 manufactured by one firm and 7500 manufactured by a second firm, all mixed together. Three of the chips will be selected at random *without* replacement. Let

$$\begin{aligned} E_1 &= \text{event that first chip selected was manufactured by Firm 1} \\ E_2 &= \text{event that second chip selected was manufactured by Firm 1} \\ E_3 &= \text{event that third chip selected was manufactured by Firm 1} \end{aligned}$$

Following the reasoning we used in Example 6.20, we can calculate

$$P(E_3|E_1 \text{ and } E_2) = \frac{2498}{9998} = 0.24985$$

$$P(E_3|\text{not } E_1 \text{ and not } E_2) = \frac{2500}{9998} = 0.25005$$

Although these two probabilities differ slightly, when rounded to three decimal places they are both 0.250. We conclude that the occurrence or nonoccurrence of E_1 or E_2 has virtually no effect on the chance that E_3 will occur. *For practical purposes*, the three events can be considered independent.

The essential difference between the situations described in Example 6.20 and Example 6.21 is the size of the sample relative to the size of the population. In Example 6.20 a relatively large proportion of the population was sampled (3 out of 52), whereas in Example 6.21 the proportion of the population sampled was quite small (only 3 out of 10,000).

Usually, when a sample is selected, the sample size is small compared to the size of the population. The condition of sampling with replacement required for many inferential methods can coexist with the practice of sampling without replacement because of the following principle:

If a random sample of size n is selected from a population of size N , the probabilities of successive selections calculated on the basis of sampling with replacement and on the basis of sampling without replacement are nearly equal when n is small compared to N .

In practice, independence can be assumed for the purpose of calculating probabilities as long as n is not larger than 5% of N .

This principle justifies the assumption of independence in many statistical problems. The phrase *assumption of independence* does not signify that the investigators are in some sense fooling themselves. They are recognizing that, for practical purposes, the results will not differ from the “right” answers.

In some sampling situations, the sample size might be more than 5% of the population. For example, a newspaper editor at a small school might easily sample more than 5% of the students at the school to assess student opinion on a particular issue. In that case, the editors would be wise to consult a statistician before proceeding. Does this mean that an investigator should never sample more than 5% of a population? Certainly not! It is almost always the case that a larger sample results in better inferences about the population. The only disadvantage of sampling more than 5% of the population (other than an increase in the time and resources required) is that the analysis of the resulting data is slightly more complicated.

EXERCISES 6.53 - 6.68

● Data set available online

- 6.53** Many fire stations handle emergency calls for medical assistance as well as calls requesting firefighting equipment. A particular station says that the probability that an incoming call is for medical assistance is 0.85. This can be expressed as $P(\text{call is for medical assistance}) = 0.85$.

- Give a relative frequency interpretation of the given probability.
- What is the probability that a call is not for medical assistance?

- 6.54** Refer to the information given in the previous exercise.

- Assuming that successive calls are independent of one another, calculate the probability that two successive calls will both be for medical assistance. (Hint: See Example 6.18.)
- Still assuming independence, calculate the probability that for two successive calls, the first is for medical assistance and the second is not for medical assistance.

- c. Still assuming independence, calculate the probability that exactly one of the next two calls will be for medical assistance. (Hint: There are two different possibilities. The one call for medical assistance might be the first call, or it might be the second call.)
- d. Do you think that it is reasonable to assume that the requests made in successive calls are independent? Explain.

- 6.55** The paper “**Predictors of Complementary Therapy Use Among Asthma Patients: Results of a Primary Care Survey**” (*Health and Social Care in the Community* [2008]: 155–164) included the accompanying table. The table summarizes the responses given by 1077 asthma patients to two questions:

Question 1: Do conventional asthma medications usually help your asthma symptoms?

Question 2: Do you use complementary therapies (such as herbs, acupuncture, aroma therapy) in the treatment of your asthma?

	Doesn't Use Complementary Therapies	Does Use Complementary Therapies
Conventional Medications Usually Help	816	131
Conventional Medications Usually Do Not Help	103	27

Consider a chance experiment that consists of randomly selecting one of the 1077 survey participants.

- a. Construct a joint probability table by dividing the count in each cell of the table by the sample size $n = 1077$. (Hint: See Example 6.16.)
- b. The joint probability in the upper left cell of the table from Part (a) is $\frac{816}{1077} = 0.758$. This represents the probability of conventional medications usually help *and* does not use complementary therapies. Interpret the other three probabilities in the joint probability table of Part (a).
- c. Are the events

CH = event that the selected participant reports that conventional medications usually help and

CT = event that the selected participant reports using complementary therapies

independent events?

Use a probability argument to justify your choice. (Hint: See Example 6.18.)

- 6.56** The report “**TV Drama/Comedy Viewers and Health Information**” (cdc.gov/healthcommunication/pdf/healthstyles_2005.pdf, retrieved April 25, 2017) describes the results of a large survey that was conducted for the Centers for Disease Control and Prevention (CDC). The sample was selected in a way that the CDC believed would result in a sample that was representative of adult Americans.

One question on the survey asked respondents if they had learned something new about a health issue or disease from a TV show in the previous 6 months. Data from the survey was used to estimate the following probabilities, where

L = event that a randomly selected adult American reports learning something new about a health issue or disease from a TV show in the previous 6 months

and

F = event that a randomly selected adult American is female

$$P(L) = 0.58 \quad P(L \cap F) = 0.31$$

Assume that $P(F) = 0.5$. Are the events L and F independent events? Use probabilities to justify your answer.

- 6.57** The report “**Great Jobs, Great Lives. The Relationship Between Student Debt, Experiences and Perceptions of College Worth**” (Gallup-Purdue Index 2015 Report) gave information on the percentage of recent college graduates (those graduating between 2006 and 2015, inclusive) who strongly agree with the statement “My college education was worth the cost.” Suppose that a college graduate will be selected at random, and consider the following events:

A = event that the selected graduate strongly agrees that education was worth the cost

N = event that the selected graduate finished college with no student debt

H = event that the selected graduate finished college with high student debt (over \$50,000)

The following probability estimates were given in the report:

$$P(A) = 0.38 \quad P(A|N) = 0.49 \quad P(A|H) = 0.18$$

- a. Interpret the value of $P(A|N)$.

- b. Interpret the value of $P(A|H)$.

- c. Are the events A and H independent? Justify your answer.

- 6.58** In a small city, approximately 15% of those eligible are called for jury duty in any one calendar year. People are selected for jury duty at random from those eligible, and the same individual cannot be called more than once in the same year.

- a.** What is the probability that a particular eligible person in this city is selected in both of the next 2 years?
- b.** What is the probability that a particular eligible person in this city is selected in all three of the next 3 years?
- 6.59** Jeanie is a bit forgetful, and if she doesn't make a "to do" list, the probability that she forgets something she is supposed to do is 0.1. Tomorrow she intends to run three errands, and she fails to write them on her list.
- a.** What is the probability that Jeanie forgets all three errands? What assumptions did you make to calculate this probability?
- b.** What is the probability that Jeanie remembers at least one of the three errands?
- c.** What is the probability that Jeanie remembers the first errand but not the second or third?
- 6.60** Consider a system consisting of four components, as pictured in the following diagram:
-
- Components 1 and 2 form a series subsystem, as do Components 3 and 4. The two subsystems are connected in parallel. Suppose that
- $P(1 \text{ works}) = 0.9$
 $P(2 \text{ works}) = 0.9$
 $P(3 \text{ works}) = 0.9$
 $P(4 \text{ works}) = 0.9$
- and that the four components work independently of one another.
- a.** The 1–2 subsystem works only if both components work. What is the probability of this happening?
- b.** What is the probability that the 1–2 subsystem doesn't work? that the 3–4 subsystem doesn't work?
- c.** The system won't work if the 1–2 subsystem doesn't work and if the 3–4 subsystem also doesn't work. What is the probability that the system won't work? that it will work?
- 6.61** Consider the system described in the previous exercise.
- a.** How would the probability of the system working change if a 5–6 subsystem were added in parallel with the other two subsystems?
- b.** How would the probability that the system works change if there were three components in series in each of the two subsystems?
- 6.62** In a January 2016 Harris Poll, each of 2252 American adults was asked the following question: If you had to choose, which ONE of the following sports would you say is your favorite?" (["Pro Football is Still America's Favorite Sport," theharrispoll.com/sports/Americas_Fav_Sport_2016.html, retrieved April 15, 2017](http://www.profootballamerica.com/sports/Americas_Fav_Sport_2016.html)). Of the survey participants, 33% chose pro football as their favorite sport. The report also included the following statement "Adults with household incomes of \$75,000 – < \$100,000 (48%) are especially likely to name pro football as their favorite sport, while love of this particular game is especially low among those in \$100,000+ households (21%)."
- Suppose that the percentages from this poll are representative of American adults in general. Consider the following events:
- F = event that a randomly selected American adult names pro football as his or her favorite sport
- L = event that a randomly selected American has a household income of \$75,000 – < \$100,000
- H = event that a randomly selected American has a household income of \$100,000+
- a.** Use the given information to estimate the following probabilities:
- $P(F)$
 - $P(F|L)$
 - $P(F|H)$
- b.** Are the events F and L mutually exclusive? Justify your answer.
- c.** Are the events H and L mutually exclusive? Justify your answer.
- d.** Are the events F and H independent? Justify your answer.
- 6.63** Consider the following events:
- T = event that a randomly selected adult trusts credit card companies to safeguard his or her personal data
- M = event that a randomly selected adult is between the ages of 19 and 36
- O = event that a randomly selected adult is 37 or older
- Based on a June 9, 2016, Gallup survey (["Data Security: Not a Big Concern for Millennials," gallup.com, retrieved April 25, 2017](http://www.gallup.com/poll/1918/data-security-not-big-concern-millennials.aspx)), the following probability estimates are reasonable:
- $P(T|M) = 0.27 \quad P(T|O) = 0.22$
- Explain why $P(T)$ is not just the average of the two given probabilities.

- 6.64** The following case study was reported in the article “Parking Tickets and Missing Women,” which appeared in an early edition of the book *Statistics: A Guide to the Unknown*. In a Swedish trial on a charge of overtime parking, a police officer testified that he had noted the position of the two air valves on the tires of a parked car: To the closest hour, one was at the one o’clock position and the other was at the six o’clock position. After the allowable time for parking in that zone had passed, the policeman returned, noted that the valves were in the same position, and ticketed the car. The owner of the car claimed that he had left the parking place in time and had returned later. The valves just happened by chance to be in the same positions.

An “expert” witness computed the probability of this occurring as $(1/12)(1/12) = 1/144$.

- What reasoning did the expert use to arrive at the probability of $1/144$?
- Can you spot the error in the reasoning that leads to the stated probability of $1/144$?
- What effect does this error have on the probability of occurrence?
- Do you think that $1/144$ is larger or smaller than the correct probability of occurrence?

- 6.65** Three friends (A, B, and C) will participate in a round-robin tournament in which each one plays both of the others. Suppose that

$$P(\text{A beats B}) = 0.7$$

$$P(\text{A beats C}) = 0.8$$

$$P(\text{B beats C}) = 0.6$$

and that the outcomes of the three matches are independent of one another.

- What is the probability that A wins both her matches and that B beats C?
- What is the probability that A wins both her matches?
- What is the probability that A loses both her matches?
- What is the probability that each person wins one match? (Hint: There are two different ways for this to happen.)

- 6.66** A store sells two different brands of dishwasher soap, and each brand comes in three different sizes: small (S), medium (M), and large (L). The proportions of the two brands and of the three sizes purchased are displayed as marginal totals in the following table.

		Size			
		S	M	L	
Brand	B_1				.40
	B_2				.60
		.30	.50	.20	

Suppose that any event involving brand is independent of any event involving size. What is the probability of the event that a randomly selected purchaser buys the small size of Brand B_1 (the event $B_1 \cap S$)? What are the probabilities of the other brand-size combinations?

- 6.67** The National Public Radio show *Car Talk* used to have a feature called “The Puzzler.” Listeners were asked to send in answers to some puzzling questions—usually about cars but sometimes about probability (which, of course, must account for the incredible popularity of the program!).

Suppose that for a car question, 800 answers were submitted, of which 50 are correct.

- Suppose that the hosts randomly select two answers from those submitted *with replacement*. Calculate the probability that both selected answers are correct. (For purposes of this problem, keep at least five digits to the right of the decimal.)
- Suppose now that the hosts select the answers at random but *without replacement*. Use conditional probability to evaluate the probability that both answers selected are correct. How does this probability compare to the one computed in Part (a)?

- 6.68** Refer to the previous exercise. Suppose now that for a probability question, 100 answers are submitted, of which 50 are correct. Calculate the probabilities in Parts (a) and (b) of the previous exercise for the probability question.

SECTION 6.6 Some General Probability Rules

In previous sections, we saw how the probability of $P(E \cup F)$ could be easily calculated when E and F are mutually exclusive and how $P(E \cap F)$ could be calculated when E and F are independent. In this section, we develop more general rules: an addition rule that can be used even when events are not mutually exclusive and a multiplication rule that can be used even when events are not independent.

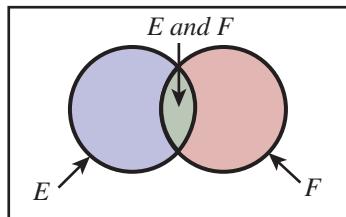
General Addition Rule

Calculating $P(E \cup F)$ when the two events are not mutually exclusive is a bit more complicated than in the case of mutually exclusive events. Consider Figure 6.9, in which E

and F overlap. The area of the colored region ($E \cup F$) is not the sum of the area of E and the area of F , because when the two individual areas are added, the area of the intersection ($E \cap F$) is counted twice. Similarly, $P(E) + P(F)$ includes $P(E \cap F)$ twice, so this intersection probability must be subtracted from the sum to obtain $P(E \cup F)$. This reasoning leads to the general addition rule.

FIGURE 6.9

The colored region is $P(E \cup F)$, and $P(E \cup F) \neq P(E) + P(F)$.



General Addition Rule for Two Events

For any two events E and F ,

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

When E and F are mutually exclusive, the general addition rule simplifies to the previous rule for mutually exclusive events. This is because when E and F are mutually exclusive, $E \cap F$ contains no outcomes and $P(E \cap F) = 0$. The general addition rule can be used to determine any one of the four probabilities $P(E)$, $P(F)$, $P(E \cap F)$, or $P(E \cup F)$ provided that the other three probabilities are known.

Example 6.22 Cable Services

Understand the context ➤

Suppose that 60% of all customers of a large cable company subscribe to Internet service, 40% subscribe to phone service, and 25% have both types of services. If a customer is selected at random, what is the probability that he or she has at least one of these two types of service? We can define the following events:

E = event that a selected customer has Internet service

F = event that a selected customer has phone service

Do the work ➤ The given information implies that

$$P(E) = 0.60 \quad P(F) = 0.40 \quad P(E \cap F) = 0.25$$

from which we can calculate

$$\begin{aligned} P(\text{customer has at least one of the two types of services}) \\ &= P(E \cup F) \\ &= P(E) + P(F) - P(E \cap F) \\ &= 0.60 + 0.40 - 0.25 \\ &= 0.75 \end{aligned}$$

The event that the customer has neither type of service is $(E \cup F)^c$, so

$$P(\text{customer has neither type of service}) = 1 - P(E \cup F) = 0.25$$

Now let's determine the probability that the selected customer has exactly one type of service. Referring to the Venn diagram in Figure 6.10, we see that the event *at least one* can be thought of as consisting of two mutually exclusive parts: *exactly one* and *both*. This means that

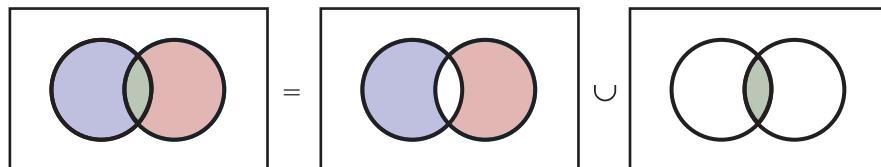
$$\begin{aligned} P(E \cup F) &= P(\text{at least one}) \\ &= P(\text{exactly one} \cup \text{both}) \\ &= P(\text{exactly one}) + P(\text{both}) \\ &= P(\text{exactly one}) + P(E \cap F) \end{aligned}$$

It follows that

$$\begin{aligned} P(\text{exactly one}) &= P(E \cup F) - P(E \cap F) \\ &= 0.75 - 0.25 \\ &= 0.50 \end{aligned}$$

FIGURE 6.10

Representing $P(E \cup F)$ as the union of two mutually exclusive events.



The general addition rule for more than two events is rather complicated. For example, in the case of three events,

$$\begin{aligned} P(E \cup F \cup G) &= P(E) + P(F) + P(G) - P(E \cap F) - P(E \cap G) \\ &\quad - P(F \cap G) + P(E \cap F \cap G) \end{aligned}$$

For more than three events, you should consult a book with more extensive coverage of probability.

General Multiplication Rule

In Section 6.4, we used the formula

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

to calculate conditional probabilities when $P(E \cap F)$ is known. Sometimes, however, conditional probabilities are known or can be estimated. When this is the case, they can be used to calculate the probability of the intersection of two events.

Multiplying both sides of the conditional probability formula by $P(F)$ gives a useful expression for the probability that both events E and F will occur.

General Multiplication Rule for Two Events

For any two events E and F ,

$$P(E \cap F) = P(E|F)P(F)$$

Example 6.23 Traffic School

Understand the context ➤

Suppose that 20% of all teenage drivers in a certain county received a citation for a moving violation in 2018. Also suppose that 80% of those receiving such a citation attended traffic school so that the citation would not appear on their driving record. If a teenage driver from this county is randomly selected, what is the probability that he or she received a citation and attended traffic school?

Formulate a plan ➤

Let's define two events E and F as follows:

E = selected driver attended traffic school

F = selected driver received a citation

The question posed can then be answered by calculating $P(E \cap F)$. The percentages given in the problem imply that

$$\begin{aligned} P(F) &= 0.20 \\ \text{and } P(E|F) &= 0.80. \end{aligned}$$

- Do the work ➤ Notice the difference between $P(E)$, which is the proportion in the entire population who attended traffic school (not given), and $P(E|F)$, which is the proportion of those receiving a citation who attended traffic school. Using the multiplication rule, we calculate

$$\begin{aligned} P(E \text{ and } F) &= P(E|F)P(F) \\ &= (0.80)(0.20) \\ &= 0.16 \end{aligned}$$

- Interpret the results ➤ This means that 16% of all teenage drivers in this county received a citation *and* attended traffic school.
-

Example 6.24 Smart Phone Warranties

- Understand the context ➤ The following table gives information on smart phones sold by a large electronics store:

	Percentage of Customers Purchasing	Of Those Who Purchase, Percentage Who Purchase Extended Warranty
Brand 1	70	20
Brand 2	30	40

Suppose a purchaser is randomly selected from among all those who bought a smart phone from this store. What is the probability that the selected customer purchased a Brand 1 model and an extended warranty?

- Formulate a plan ➤ To answer this question, we first define the following events:

$$\begin{aligned} B_1 &= \text{event that Brand 1 is purchased} \\ B_2 &= \text{event that Brand 2 is purchased} \\ E &= \text{event that an extended warranty is purchased} \end{aligned}$$

The information in the table implies that

$$\begin{aligned} P(\text{Brand 1 purchased}) &= P(B_1) = 0.70 \\ P(\text{extended warranty} | \text{Brand 1 purchased}) &= P(E|B_1) = 0.20 \end{aligned}$$

- Do the work ➤ Notice that the 20% is identified with a *conditional* probability. *Among purchasers of Brand 1*, this is the percentage opting for an extended warranty. Substituting these numbers into the general multiplication rule, we get

$$\begin{aligned} P(B_1 \text{ and } E) &= P(E|B_1)P(B_1) \\ &= (0.20)(0.70) \\ &= 0.14 \end{aligned}$$

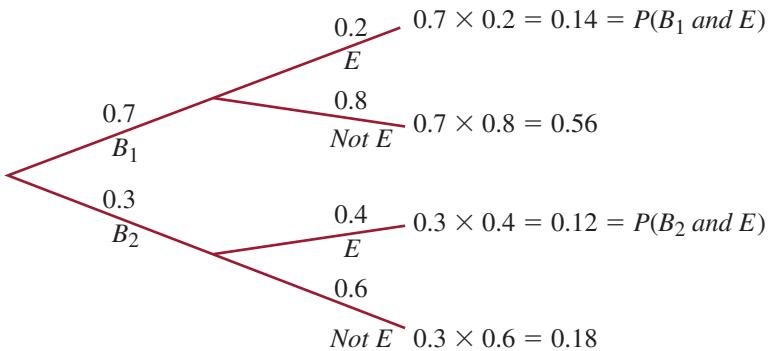
The tree diagram of Figure 6.11 gives a nice visual display of how the general multiplication rule is used here. The two first-generation branches are labeled with events B_1 and B_2 along with their probabilities. Two second-generation branches extend from each first-generation branch. These correspond to the two events E and *not E*. The *conditional* probabilities $P(E|B_1)$, $P(\text{not } E|B_1)$, $P(E|B_2)$, and $P(\text{not } E|B_2)$, appear on these branches. Application of the multiplication rule then consists of multiplying probabilities along the branches of the tree diagram. For example,

$$\begin{aligned} P(\text{Brand 2 and warranty purchased}) &= P(B_2 \text{ and } E) \\ &= P(B_2 \cap E) \\ &= P(E|B_2)P(B_2) \\ &= (0.4)(0.3) \\ &= 0.12 \end{aligned}$$

and this probability is displayed to the right of the E branch that comes from the B_2 branch.

FIGURE 6.11

A tree diagram for the probability calculations of Example 6.24.



We can now easily calculate $P(E)$, the probability that an extended warranty is purchased. The event E can occur in two different ways: Buy Brand 1 *and* warranty, or buy Brand 2 *and* warranty. These events are $B_1 \cap E$ and $B_2 \cap E$. Furthermore, if each customer purchased a single smart phone, he or she could not have simultaneously purchased both Brand 1 and Brand 2, so the two events $B_1 \cap E$ and $B_2 \cap E$ are mutually exclusive.

It follows that

$$\begin{aligned}
 P(E) &= P(B_1 \cap E) + P(B_2 \cap E) \\
 &= P(E|B_1)P(B_1) + P(E|B_2)P(B_2) \\
 &= (0.2)(0.7) + (0.4)(0.3) \\
 &= 0.14 + 0.12 \\
 &= 0.26
 \end{aligned}$$

- Interpret the results ➤ This probability is the sum of two of the probabilities shown on the right-hand side of the tree diagram. This means that 26% of all smart phone purchasers selected an extended warranty.

The general multiplication rule can be extended to give an expression for the probability that several events occur together. In the case of three events E , F , and G , we have

$$P(E \cap F \cap G) = P(E|F \cap G)P(F|G)P(G)$$

When the events are all independent, $P(E|F \cap G) = P(E)$ and $P(F|G) = P(F)$ so the right-hand side of the equation for $P(E \cap F \cap G)$ simplifies to the product of the three unconditional probabilities.

Example 6.25 Lost Luggage

- Understand the context ➤ Twenty percent of all passengers who fly from Los Angeles (LA) to New York (NY) do so on Airline G. This airline misplaces luggage for 10% of its passengers, and 90% of this lost luggage is found and returned to the owner. If a passenger who has flown from LA to NY is randomly selected, what is the probability that the selected individual flew on Airline G (event G), had luggage misplaced (event F), and the misplaced luggage was returned (event E)? The given information implies that

$$P(G) = 0.20 \quad P(F|G) = 0.10 \quad P(E|F \cap G) = 0.90$$

Then

$$P(E \cap F \cap G) = P(E|F \cap G)P(F|G)P(G) = (0.90)(0.10)(0.20) = 0.018$$

- Interpret the results ➤ This means that about 1.8% of passengers flying from LA to NY fly on Airline G, have their luggage misplaced, and the lost luggage is returned.

Law of Total Probability

Let's reconsider the information on smart phone sales from Example 6.24. In this example, the following events were defined:

- B_1 = event that Brand 1 is purchased
- B_2 = event that Brand 2 is purchased
- E = event that an extended warranty is purchased

Based on the information given in Example 6.24, the following probabilities are known:

$$P(B_1) = 0.7 \quad P(B_2) = 0.3 \quad P(E|B_1) = 0.2 \quad P(E|B_2) = 0.4$$

Notice that the conditional probabilities $P(E|B_1)$ and $P(E|B_2)$ are known but that the unconditional probability $P(E)$ is not known.

To find $P(E)$ we noted that the event E can occur in two ways:

- (1) A customer purchases an extended warranty *and* buys Brand 1 ($E \cap B_1$); or
- (2) a customer purchases an extended warranty *and* buys Brand 2 ($E \cap B_2$).

Because these are the only ways in which E can occur, we can write the event E as

$$E = (E \cap B_1) \cup (E \cap B_2)$$

The two events $(E \cap B_1)$ and $(E \cap B_2)$ are mutually exclusive (since B_1 and B_2 are mutually exclusive), so using the addition rule for mutually exclusive events gives

$$\begin{aligned} P(E) &= P((E \cap B_1) \cup (E \cap B_2)) \\ &= P(E \cap B_1) + P(E \cap B_2) \end{aligned}$$

Finally, using the general multiplication rule to evaluate $P(E \cap B_1)$ and $P(E \cap B_2)$ results in

$$\begin{aligned} P(E) &= P(E \cap B_1) + P(E \cap B_2) \\ &= P(E|B_1)P(B_1) + P(E|B_2)P(B_2) \end{aligned}$$

Substituting in the known probabilities results in

$$\begin{aligned} P(E) &= P(E|B_1)P(B_1) + P(E|B_2)P(B_2) \\ &= (0.2)(0.7) + (0.4)(0.3) \\ &= 0.26 \end{aligned}$$

We concluded that 26% of the smart phone customers purchased an extended warranty.

This illustrates that, when conditional probabilities are known, they can sometimes be used to calculate unconditional probabilities. The **law of total probability** formalizes this use of conditional probabilities.

The Law of Total Probability

If B_1 and B_2 are mutually exclusive events with $P(B_1) + P(B_2) = 1$, then for any event E

$$\begin{aligned} P(E) &= P(E \cap B_1) + P(E \cap B_2) \\ &= P(E|B_1)P(B_1) + P(E|B_2)P(B_2) \end{aligned}$$

More generally, if B_1, B_2, \dots, B_k are mutually exclusive events with $P(B_1) + P(B_2) + \dots + P(B_k) = 1$, then for any event E

$$\begin{aligned} P(E) &= P(E \cap B_1) + P(E \cap B_2) + \dots + P(E \cap B_k) \\ &= P(E|B_1)P(B_1) + P(E|B_2)P(B_2) + \dots + P(E|B_k)P(B_k) \end{aligned}$$

Example 6.26 Which Way to Jump?

Understand the context ➤

The paper “**Action Bias among Elite Soccer Goalkeepers: The Case of Penalty Kicks**” (*Journal of Economic Psychology* [2007]: 606–621) presents an interesting analysis of 286 penalty kicks in televised championship soccer games from around the world. In a penalty

kick, the only players involved are the kicker and the goalkeeper from the opposing team. The kicker tries to kick a ball into the goal from a point located 11 meters away. The goalkeeper tries to block the ball from reaching the goal.

For each penalty kick analyzed, the researchers recorded the direction that the goalkeeper moved (jumped to the left, stayed in the center, or jumped to the right) and whether or not the penalty kick was successfully blocked.

Consider the following events:

L = the event that the goalkeeper jumps to the left

C = the event that the goalkeeper stays in the center

R = the event that the goalkeeper jumps to the right

B = the event that the penalty kick is blocked

Consider the data ➤

Based on an analysis of the penalty kicks, the authors of the paper gave the following probability estimates:

$$\begin{array}{lll} P(B|L) = 0.142 & P(B|C) = 0.333 & P(B|R) = 0.126 \\ P(L) = 0.493 & P(C) = 0.063 & P(R) = 0.444 \end{array}$$

Formulate a plan ➤

What proportion of penalty kicks were blocked? We can use the law of total probability to answer this question. Here, the three events L , C , and R play the role of B_1 , B_2 , and B_3 and B plays the role of E in the formula for the law of total probability.

Substituting into the formula, we get

Do the work ➤

$$\begin{aligned} P(B) &= P(B \cap L) + P(B \cap C) + P(B \cap R) \\ &= P(B|L)P(L) + P(B|C)P(C) + P(B|R)P(R) \\ &= (0.142)(0.493) + (0.333)(0.063) + (0.126)(0.444) \\ &= 0.070 + 0.021 + 0.056 \\ &= 0.147 \end{aligned}$$

Interpret the results ➤

This means that only 14.7% of penalty kicks were successfully blocked. Two other interesting findings of this study were

1. The direction that the goalkeeper moves appears to be independent of whether the kicker kicked the ball to the left, center, or right of the goal. This was attributed to the fact that goalkeepers have to choose their action before they can clearly observe the direction of the kick.
2. Based on the three conditional probabilities— $P(B|L) = 0.142$, $P(B|C) = 0.333$, and $P(B|R) = 0.126$ —the optimal strategy for a goalkeeper appears to be to stay in the center of the goal. However, staying in the center was only chosen 6.3% of the time—much less often than jumping left or right. The authors believe that this is because a goalkeeper does not feel as bad about not successfully blocking a kick if some action (jumping left or right) is taken compared to if no action (staying in the center) is taken. This is the “action bias” referred to in the title of the paper.

Bayes' Rule

We conclude our discussion of probability rules by considering a formula discovered by the Reverend Thomas Bayes (1702–1761), an English Presbyterian minister. He discovered what is now known as Bayes' rule (or Bayes' theorem). Bayes' rule is a solution to what Bayes called the converse problem. To see what he meant by this, we return to the field of medical diagnosis.

Example 6.27 Lyme Disease

Understand the context ➤

Lyme disease is transmitted by infected ticks. Several tests are available for people with symptoms of Lyme disease. One of these tests is the EIA/IFA test. The paper “[Lyme Disease Testing by Large Commercial Laboratories in the United States](#) (*Clinical Infectious Disease* [2014]: 676–681) found that 11.4% of those tested had Lyme disease.

Consider the following events:

- + represents a positive result on the blood test
- represents a negative result on the blood test
- L represents the event that the patient actually has Lyme disease
- L^C represents the event that the patient actually does not have Lyme disease

The following probabilities were reported in the article:

Consider the data ➤

Probability	Interpretation
$P(L) = 0.114$	The prevalence of Lyme disease in the population. About 11.4% of the population actually has Lyme disease.
$P(L^C) = 0.886$	88.6% of the population does not have Lyme disease.
$P(+ L) = 0.933$	93.3% of those with Lyme disease test positive.
$P(- L) = 0.067$	6.7% of those with Lyme disease test negative.
$P(+ L^C) = 0.039$	3.9% of those who do not have Lyme disease test positive.
$P(- L^C) = 0.961$	96.1% of those who do not have Lyme disease test negative.

Notice the form of the known conditional probabilities. For example, $P(+|L)$ is the probability of a positive test *given* that a person selected at random from the population actually has Lyme disease. Bayes' converse problem poses a question of a different form: Given that a person tests positive for the disease, what is the probability that he or she actually has Lyme disease? This converse problem is the one that is of primary interest in medical diagnosis problems.

Bayes reasoned as follows to obtain the answer to the converse problem of finding $P(L|+)$. We know from the definition of conditional probability that

$$P(L|+) = \frac{P(L \cap +)}{P(+)}$$

Because $P(L \cap +) = P(+ \cap L)$, we can use the general multiplication rule to get

$$P(L \cap +) = P(+ \cap L) = P(+|L)P(L)$$

This helps, because both $P(+|L)$ and $P(L)$ are known. We now have

$$P(L|+) = \frac{P(+|L)P(L)}{P(+)}$$

The denominator $P(+)$ can be evaluated using the law of total probability, because L and L^C are mutually exclusive with $P(L) + P(L^C) = 1$. Applying the law of total probability to the denominator, we obtain

$$\begin{aligned} P(+) &= P(+ \cap L) + P(+ \cap L^C) \\ &= P(+|L)P(L) + P(+|L^C)P(L^C) \end{aligned}$$

We now have all we need to answer the converse problem:

Do the work ➤

$$\begin{aligned} P(L|+) &= \frac{P(+|L)P(L)}{P(+|L)P(L) + P(+|L^C)P(L^C)} \\ &= \frac{(0.933)(0.114)}{(0.933)(0.114) + (0.039)(0.886)} \\ &= \frac{0.106}{0.106 + 0.035} \\ &= 0.752 \end{aligned}$$

Interpret the results ➤

The probability $P(L|+)$ is a conditional probability. $P(L|+) = 0.752$ means that, in the long run, about 75.2% of those who test positive actually have the disease. Notice the difference between $P(L|+)$ and the previously reported conditional probability $P(+|L) = 0.933$, which means that 93.3% of those with Lyme disease test positive.

The accompanying box formalizes this reasoning in the statement of Bayes' rule.

Bayes' Rule

If B_1 and B_2 are mutually exclusive events with $P(B_1) + P(B_2) = 1$, then for any event E

$$P(B_1|E) = \frac{P(E|B_1)P(B_1)}{P(E|B_1)P(B_1) + P(E|B_2)P(B_2)}$$

More generally, if B_1, B_2, \dots, B_k are mutually exclusive events with $P(B_1) + P(B_2) + \dots + P(B_k) = 1$ then for any event E ,

$$P(B_i|E) = \frac{P(E|B_i)P(B_i)}{P(E|B_1)P(B_1) + P(E|B_2)P(B_2) + \dots + P(E|B_k)P(B_k)}$$

Example 6.28 Internet Addiction

Understand the context ➤

Internet addiction has been defined by researchers as a disorder characterized by excessive time spent on the Internet, impaired judgment and decision-making ability, social withdrawal, and depression. The paper “**The Association between Aggressive Behaviors and Internet Addiction and Online Activities in Adolescents**” (*Journal of Adolescent Health [2009]: 598–605*) describes a study of a large number of adolescents.

Consider the data ➤

Each participant in the study was assessed using the Chen Internet Addiction Scale to determine if he or she suffered from Internet addiction. The following statements are based on the study results:

1. 51.8% of the study participants were female and 48.2% were male.
2. 13.1% of the females suffered from Internet addiction.
3. 24.8% of the males suffered from Internet addiction.

Consider the chance experiment that consists of selecting a study participant at random, and define the following events:

F = the event that the selected participant is female

M = the event that the selected participant is male

I = the event that the selected participant suffers from Internet addiction

The three statements from the paper define the following probabilities:

$$\begin{aligned} P(F) &= 0.518 & P(M) &= 0.482 \\ P(I|F) &= 0.131 & P(I|M) &= 0.248 \end{aligned}$$

Formulate a plan ➤

Suppose that we want to know the proportion of those who suffer from Internet addiction who are female. This is equivalent to $P(F|I)$.

Notice that we know $P(I|F)$ but not $P(F|I)$. We can use Bayes' rule to evaluate $P(F|I)$ as follows (with F and M playing the role of B_1 and B_2 and I playing the role of E in the formula for Bayes' rule):

$$\begin{aligned} \text{Do the work ➤ } P(F|I) &= \frac{P(I|F)P(F)}{P(I|F)P(F) + P(I|M)P(M)} \\ &= \frac{(0.131)(0.518)}{(0.131)(0.518) + (0.482)(0.248)} \\ &= \frac{0.068}{0.068 + 0.120} \\ &= \frac{0.068}{0.188} \\ &= 0.362 \end{aligned}$$

Interpret the results ➤

This tells us that 36.2% of those who suffered from Internet addiction were female.

EXERCISES 6.69 - 6.87**● Data set available online**

- 6.69** A university has 10 vehicles available for use by faculty and staff. Six of these are vans and four are cars. On a particular day, only two requests for vehicles have been made. Suppose that the two vehicles to be assigned are chosen at random from the 10 vehicles available.
- Let E denote the event that the first vehicle assigned is a van. What is $P(E)$?
 - Let F denote the event that the second vehicle assigned is a van. What is $P(F|E)$?
 - Use the results of Parts (a) and (b) to calculate $P(E \text{ and } F)$ (Hint: See Example 6.23.)
- 6.70** A construction firm bids on two different contracts. Let E_1 be the event that the bid on the first contract is successful, and define E_2 analogously for the second contract. Suppose that $P(E_1) = 0.4$ and $P(E_2) = 0.3$ and that E_1 and E_2 are independent events.
- Calculate the probability that both bids are successful (the probability of the event $E_1 \text{ and } E_2$).
 - Calculate the probability that neither bid is successful (the probability of the event $(\text{not } E_1) \text{ and } (\text{not } E_2)$).
 - What is the probability that the firm is successful in at least one of the two bids? (Hint: See Example 6.22.)
- 6.71** There are two traffic lights on Darlene's route from home to work. Let E denote the event that Darlene must stop at the first light, and define the event F in a similar manner for the second light. Suppose that $P(E) = 0.4$, $P(F) = 0.3$, and $P(E \cap F) = 0.15$.
- What is the probability that Darlene must stop at at least one light; that is, what is the probability of the event $E \cup F$?
 - What is the probability that Darlene doesn't have to stop at either light?
 - What is the probability that Darlene must stop at exactly one of the two lights?
 - What is the probability that Darlene must stop just at the first light? (Hint: How is the probability of this event related to $P(E)$ and $P(E \cap F)$? A Venn diagram might help.)
- 6.72** Let F denote the event that a randomly selected registered voter in a certain city has signed a petition to recall the mayor. Also, let E denote the event that the randomly selected registered voter actually votes in the recall election. Describe the event $E \cap F$ in words. If $P(F) = 0.10$ and $P(E|F) = 0.80$, calculate $P(E \cap F)$.

- 6.73** According to a July 31, 2013, posting on [cnn.com](#), a 2010 study in the journal *Pediatrics* found that 8% of American children younger than age 18 have at least one food allergy. Among those with food allergies, about 39% had a history of severe reaction.
- If an American child younger than 18 is randomly selected, what is the probability that he or she has at least one food allergy and has a history of severe reaction?
 - It was also reported that 30% of those with an allergy are allergic to multiple foods. If an American child younger than 18 is randomly selected, what is the probability that he or she is allergic to multiple foods?
- 6.74** Suppose that Blue Cab operates 15% of the taxis in a certain city and Green Cab operates the other 85%. After a nighttime hit-and-run accident involving a taxi, an eyewitness said the vehicle was blue. Suppose, though, that under night vision conditions, only 80% of individuals can correctly distinguish between a blue and a green vehicle. What is the probability that the taxi at fault was blue? (Hint: A tree diagram might help.)
- 6.75** A large cable company reports the following:
- 80% of its customers subscribe to cable TV service
 - 42% of its customers subscribe to Internet service
 - 32% of its customers subscribe to telephone service
 - 25% of its customers subscribe to both cable TV and Internet service
 - 21% of its customers subscribe to both cable TV and phone service
 - 23% of its customers subscribe to both Internet and phone service
 - 15% of its customers subscribe to all three services
- Consider the chance experiment that consists of selecting one of the cable company customers at random. Calculate and interpret the following probabilities:
- $P(\text{cable TV only})$
 - $P(\text{Internet} | \text{cable TV})$
 - $P(\text{exactly two services})$
 - $P(\text{Internet and cable TV only})$
- 6.76** Refer to the information given in the previous exercise about customers of a large cable company.
- Suppose two customers are to be selected at random. Would it be reasonable to consider the events C_1 = event that the first customer selected subscribes to cable TV and C_2 = event that the second customer selected subscribes to cable TV as independent events? Explain.
 - With C_1 and C_2 as defined in Part (a), calculate $P(C_1 \cap C_2)$.

- 6.77** The authors of the paper “**Do Physicians Know When Their Diagnoses Are Correct?**” (*Journal of General Internal Medicine* [2005]: 334–339) presented detailed case studies to medical students and to faculty at medical schools. Each participant was asked to provide a diagnosis in the case and also to indicate whether his or her confidence in the correctness of the diagnosis was high or low. Define the events C , I , and H as follows:

C = event that diagnosis is correct

I = event that diagnosis is incorrect

H = event that confidence in the correctness of the diagnosis is high

- a. Data appearing in the paper were used to estimate the following probabilities for medical students:

$$P(C) = 0.261 \quad P(I) = 0.739$$

$$P(H|C) = 0.375 \quad P(H|I) = 0.073$$

Use Bayes’ rule to calculate the probability of a correct diagnosis given that the student’s confidence level in the correctness of the diagnosis is high.

- b. Data from the paper were also used to estimate the following probabilities for medical school faculty:

$$P(C) = 0.495 \quad P(I) = 0.505$$

$$P(H|C) = 0.537 \quad P(H|I) = 0.252$$

Calculate $P(C|H)$ for medical school faculty.

How does the value of this probability compare to the value of $P(C|H)$ for students computed in Part (a)?

- 6.78** A study of how people are using online services for medical consulting is described in the paper “**Internet Based Consultation to Transfer Knowledge for Patients Requiring Specialized Care**” (*British Medical Journal* [2003]: 696–699). Patients using a particular online site could request any combination of three services: specialist opinion, assessment of imaging studies (such as X-ray and MRI), and assessment of pathology results.

The accompanying table shows the combinations of services that were requested by 79 patients in their online consultations.

Combination of Services Provided	Number of Patients
Specialist opinion only	37
Assessment of pathology results only	1
Specialist opinion and assessment of pathology results only	11
Specialist opinion and assessment of imaging studies only	14
Specialist opinion, assessment of imaging studies, and assessment of pathology results	16

For a randomly selected patient from this study, define the events O , I , and A as follows:

O = event that the online consultation involves a specialist opinion

I = event that the online consultation involves the assessment of imaging studies

A = event that the online consultation involves the assessment of pathology results

Use the given information to calculate the following probabilities:

a. $P(O)$

b. $P(\text{not } A)$

c. $P(O \cap I)$

d. $P(I|O)$

e. $P(O|I)$

- 6.79** The report “**Twitter in Higher Education: Usage Habits and Trends of Today’s College Faculty**” (*Magna Publications*, September 2009) describes results of a survey of nearly 2000 college faculty.

The report indicates the following:

- 30.7% reported that they use Twitter and 69.3% said that they did not use Twitter.
- Of those who use Twitter, 39.9% said they sometimes use Twitter to communicate with students.
- Of those who use Twitter, 27.5% said that they sometimes use Twitter as a learning tool in the classroom.

Consider the chance experiment that selects one of the study participants at random and define the following events:

T = event that selected faculty member uses Twitter

C = event that selected faculty member sometimes uses Twitter to communicate with students

L = event that selected faculty member sometimes uses Twitter as a learning tool in the classroom

- a. Use the given information to calculate the following probabilities:

i. $P(T)$

ii. $P(T^c)$

iii. $P(C|T)$

iv. $P(L|T)$

v. $P(C \cap T)$

- b. Interpret each of the probabilities computed in Part (a).

- 6.80** Use the information given in the previous exercise to answer the following questions.

- a. What proportion of the faculty surveyed sometimes use Twitter to communicate with students? [Hint: Use the law of total probability to find $P(C)$.]

- b.** What proportion of faculty surveyed sometimes use Twitter as a learning tool in the classroom?
- 6.81** The accompanying table summarizes data from a medical expenditures survey carried out by the National Center for Health Statistics (“**Assessing the Effects of Race and Ethnicity on Use of Complementary and Alternative Therapies in the USA**,” *Ethnicity and Health* [2005]: 19–32).
- | Use of Alternative Therapies by Education Level | | |
|---|-------------------------------------|--|
| Education Level | Percent Using Alternative Therapies | |
| High school or less | 4.3% | |
| College—1 to 4 years | 8.2% | |
| College—5 years or more | 11.0% | |
- These percentages were based on data from 7320 people whose education level was high school or less, 4793 people with 1 to 4 years of college, and 1095 people with 5 or more years of college.
- a.** Use the information given to determine the *number* of respondents falling into each of the six cells of the table below.
- | | Does
Uses
Alternative
Therapies | Not Use
Alternative
Therapies | Total |
|-----------------------|--|-------------------------------------|-------|
| HS/less | | | 7,320 |
| College: 1–4 yrs | | | 4,793 |
| College: ≥ 5 yrs | | | 1,095 |
- b.** Construct a table of estimated probabilities by dividing the count in each of the six table cells by the total sample size, $n = 13,208$.
- c.** The authors of the study indicated that the sample was selected in a way that makes it reasonable to regard the estimated probabilities in the table from Part (b) as representative of the adult population in the United States. Use the information in that table to estimate the following probabilities for adults in the United States.
- i.** The probability that a randomly selected individual has 5 or more years of college.
 - ii.** The probability that a randomly selected individual uses alternative therapies.
- 6.82** Use the table of estimated probabilities from the previous exercise to answer the following questions.
- a.** Estimate the following probabilities for adults in the United States.
- i.** The probability that a randomly selected individual uses alternative therapies given that he or she has 5 or more years of college.
- ii.** The probability that a randomly selected individual uses alternative therapies given that he or she has an education level of high school or less.
 - iii.** The probability that a randomly selected individual who uses alternative therapies has an education level of high school or less.
 - iv.** The probability that a randomly selected individual with some college uses alternative therapies.
- b.** Are the events H = event that a randomly selected individual has an education level of high school or less and A = event that a randomly selected individual uses alternative therapies independent events? Explain.
- 6.83** Suppose that we define the following events:
- C = event that a randomly selected driver is observed to be using a cell phone
 - A = event that a randomly selected driver is observed driving a passenger automobile
 - V = event that a randomly selected driver is observed driving a van or SUV
 - T = event that a randomly selected driver is observed driving a pickup truck
- The probabilities given below are based on data from the **National Highway Traffic Safety Administration (“Traffic Safety Facts,” February 2014)**:
- $P(C) = 0.015 \quad P(C|A) = 0.017$
 $P(C|V) = 0.014 \quad P(C|T) = 0.010$
- Explain why $P(C)$ is not just the average of the three given conditional probabilities.
- 6.84** The article “**U.S. Investors Split Between Digital and Traditional Banking**” (gallup.com, August 5, 2016, retrieved April 25, 2017) summarized data from a Gallup survey of a random sample of 1019 U.S. adults with investments of \$10,000 or more. Based on the survey data, it was estimated that 31% of investors manage their investments by doing everything they possibly can online. But the authors of the article also noted that there was quite a difference between younger investors (age 18 to 49) and older investors (age 50 and older). For younger investors, 43% said they do everything they possibly can online, while the percentage for older investors was 23%.
- a.** Use the given information to estimate $P(O)$, $P(O|Y)$, and $P(O|F)$ where O = event that a randomly selected investor does everything possible online, Y = event that a randomly selected investor is age 18 to 49, and

F = event that a randomly selected investor is 50 years old or older.

- b. Suppose that 40% of investors are between the ages of 18 and 49. Use the probabilities from Part (a) and the estimate $P(Y) = 0.40$ to calculate $P(Y|O)$ and write a sentence interpreting this value.
- 6.85** Radiologists are often asked to predict the sex of a baby from ultrasound images made during pregnancy. The authors of the paper “**The Use of Three Dimensional Ultrasound for Fetal Gender Determination in the First Trimester**” (*The British Journal of Radiology*, [2003]: 448–451) followed up on 159 predictions made by a particular radiologist (Radiologist 1) to determine whether or not they were correct. Data from the paper is summarized in the accompanying table.

		Radiologist 1	
		Predicted Male	Predicted Female
Baby Is Male	74	12	
Baby Is Female	14	59	

- a. Assuming that these data are representative of sex predictions made by Radiologist 1, estimate the probability that a sex prediction is correct, given that the baby is male.
- b. Assuming that these data are representative of sex predictions made by Radiologist 1, estimate the probability that a sex prediction is correct, given that the baby is female.
- c. For Radiologist 1, is a sex prediction more likely to be correct if the baby is male? Explain.
- d. Estimate the probability that a sex prediction made by Radiologist 1 is correct.
- 6.86** The paper referenced in the previous exercise also included data for a second radiologist, Radiologist 2. Based on the data from the previous exercise for Radiologist 1 and the data in the accompanying table for Radiologist 2, write a

paragraph comparing the accuracy of sex predictions made by these two radiologists.

		Radiologist 2	
		Predicted Male	Predicted Female
Baby Is Male	81	8	
Baby Is Female	7	58	

- 6.87** In an article that appears on the web site of the American Statistical Association (amstat.org), Carlton Gunn, a public defender in Seattle, Washington, wrote about how he uses statistics in his work as an attorney. He states:

I personally have used statistics in trying to challenge the reliability of drug testing results. Suppose the chance of a mistake in the taking and processing of a urine sample for a drug test is just 1 in 100. And your client has a “dirty” (i.e., positive) test result. Only a 1 in 100 chance that it could be wrong? Not necessarily. If the vast majority of all tests given—say 99 in 100—are truly clean, then you get one false dirty and one true dirty in every 100 tests, so that half of the dirty tests are false.

Define the following events as:

TD = event that the test result is dirty

TC = event that the test result is clean

D = event that the person tested is actually dirty

C = event that the person tested is actually clean

- a. Using the information in the quote, what are the values of
- i. $P(TD|D)$
 - ii. $P(TD|C)$
 - iii. $P(C)$
 - iv. $P(D)$
- b. Use the law of total probability to find $P(TD)$.
- c. Use Bayes’ rule to evaluate $P(C|TD)$. Is this value consistent with the argument given in the quote? Explain.

SECTION 6.7 Estimating Probabilities Empirically and Using Simulation

In the examples presented so far, reaching conclusions required knowledge of the probabilities of various outcomes. In some cases, this is reasonable, and we know the actual long-run proportion of the time that each outcome will occur. In other situations, these probabilities are not known. Sometimes probabilities can be determined analytically, by using the probability properties and probability rules introduced in this chapter. However, when this is not possible, impractical, or just beyond the limited probability tools of the introductory course, we can *estimate* probabilities empirically through observation or by simulation.

Estimating Probabilities Empirically

It is fairly common practice to use observed long-run proportions to estimate probabilities. The process of estimating probabilities empirically is simple:

1. Observe a very large number of chance outcomes under controlled circumstances.
2. Interpreting probability as a long-run relative frequency, estimate the probability of an event by using the observed proportion of occurrence.

This process is illustrated in Examples 6.29 and 6.30.

Example 6.29 Fair Hiring Practices

Understand the context ➤

The Biology Department at a university plans to recruit a new faculty member and intends to advertise for someone with a Ph.D. in biology and at least 10 years of college-level teaching experience. A member of the department expresses the belief that the experience requirement will exclude many potential applicants and will exclude far more female applicants than male applicants. The Biology Department would like to determine the probability that an applicant with a Ph.D. in biology would be eliminated from consideration because of the experience requirement.

A similar university completed a search in which there was no requirement for prior teaching experience, but the information about prior teaching experience was recorded. The 410 applications yielded the following data:

Consider the data ➤

			Number of Applicants	
		Less Than 10 Years of Experience	10 Years of Experience or More	Total
	Male	178	112	290
	Female	99	21	120
	Total	277	133	410

Do the work ➤

Let's assume that the populations of applicants for the two positions can be regarded as the same. We can use the available information to approximate the probability that an applicant will fall into each of the four sex–experience combinations.

The estimated probabilities (obtained by dividing the number of applicants for each sex–experience combination by 410) are given in Table 6.1.

Table 6.1 Estimated Probabilities for Example 6.29

	Less Than 10 Years of Experience	10 Years of Experience or More
Male	0.434	0.273
Female	0.242	0.051

Interpret the results ➤

From Table 6.1, the estimate of $P(\text{candidate excluded because of the experience requirement}) = 0.434 + 0.242 = 0.676$.

We can also assess the impact of the experience requirement separately for male and for female applicants. From the given information, the proportion of male applicants who have less than 10 years of experience is $178/290 = 0.614$, whereas the corresponding proportion for females is $99/120 = 0.825$. Therefore, approximately 61% of the male applicants would be eliminated by the experience requirement, and about 83% of the female applicants would be eliminated.

These subgroup proportions—0.614 for males and 0.825 for females—are estimates of conditional probabilities, which show how the original probability changes in light of new information. In this example, the probability that a potential candidate has less than 10 years of experience is 0.676, but this probability changes to 0.825 if we know that a candidate is female. These probabilities can be expressed as

$P(\text{less than 10 years of experience}) = 0.676$ (an unconditional probability)

and

$P(\text{less than 10 years of experience} | \text{female}) = 0.825$ (a conditional probability)

Example 6.30 Who Has the Upper Hand?

Understand the context ➤

Men and women frequently express intimacy through the simple act of holding hands. Some researchers have suggested that hand-holding is not only an expression of intimacy but also communicates status differences. For two people to hold hands, one must assume an overhand grip and one an underhand grip. Research in this area has shown that it is predominantly the male who assumes the overhand grip.

In the view of some investigators, the overhand grip is seen to imply status or superiority. The authors of the paper “[Men and Women Holding Hands: Whose Hand Is Uppermost?](#)” (*Perceptual and Motor Skills* [1999]: 537–549) investigated an alternative explanation—perhaps the positioning of hands is a function of the heights of the individuals. Because men, on average, tend to be taller than women, maybe comfort, not status, dictates the positioning.

Investigators at two separate universities observed hand-holding male–female couples and recorded the data in the accompanying table.

Consider the data ➤

	Sex of Person with Uppermost Hand		
	Male	Female	Total
Man Taller	2,149	299	2,448
Equal Height	780	246	1,026
Woman Taller	241	205	446
Total	3,170	750	3,920

Do the work ➤

Assuming that these 3920 hand-holding couples are representative of hand-holding couples in general, we can use the available information to estimate various probabilities. For example, if a hand-holding couple is selected at random, then

$$P(\text{man's hand uppermost}) = \frac{3170}{3920} = 0.809$$

For a randomly selected hand-holding couple, if the man is taller, then the probability that the male has the uppermost hand is

$$2149/2448 = 0.878.$$

On the other hand—so to speak—if the woman is taller, the probability that the female has the uppermost hand is

$$205/446 = 0.460.$$

Interpret the results ➤

Notice that these are estimates of the conditional probabilities $P(\text{male uppermost} | \text{male taller})$ and $P(\text{female uppermost} | \text{female taller})$, respectively. Also, because $P(\text{male uppermost} | \text{male taller})$ is not equal to $P(\text{male uppermost})$, the events *male uppermost* and *male taller* are not independent events. Even when the female is taller, the male is still more likely to have the upper hand!

Estimating Probabilities Using Simulation

Simulation provides a way to estimate probabilities when we are unable to determine probabilities analytically and when it is impractical to estimate them empirically by observation. Simulation is a method that generates “observations” by performing a chance experiment that is as similar as possible in structure to the real situation of interest.

To illustrate the idea of simulation, consider the situation in which a professor wishes to estimate the probabilities of different possible scores on a 20-question true–false quiz when students are just guessing at the answers. Because each question is a true–false question, a person who is guessing should be equally likely to answer correctly or incorrectly on any given question.

Rather than asking a guessing student to select true or false and then comparing the choice to the correct answer, an equivalent process would be to pick a ball at random from a box that contains half red balls and half blue balls, with a blue ball representing a correct answer. Making 20 selections from the box (with replacement) and then counting the number of correct choices (the number of times a blue ball is selected) is a physical substitute for an observation from a student who has guessed at the answers to 20 true–false questions. Any particular number of blue balls in 20 selections should have the same probability as the same number of correct responses to the quiz when a student is guessing.

For example, 20 selections of balls might produce the following results:

Selection	1	2	3	4	5	6	7	8	9	10
	R	R	B	R	B	B	R	R	R	B
Selection	11	12	13	14	15	16	17	18	19	20
	R	R	B	R	R	B	B	R	R	B

Because 8 of the 20 selections resulted in a blue ball, this would correspond to a quiz with eight correct responses, and it would provide us with one observation for estimating the probabilities of interest. This process could then be repeated a large number of times to generate additional observations. For example, we might find the following:

Repetition	Number of “Correct” Responses
1	8
2	11
3	10
4	12
:	:
1,000	11

The 1000 simulated quiz scores could then be used to construct a table of estimated probabilities.

Taking this many balls out of a box and writing down the results would be cumbersome and tedious. The process can be simplified by using random digits to substitute for drawing balls from the box. For example, a single digit could be selected at random from the 10 digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. When using random digits, each of the 10 possibilities is equally likely to occur, so we can use the even digits (including 0) to indicate a correct response and the odd digits to indicate an incorrect response. This would maintain the important property that a correct response and an incorrect response are equally likely, because correct and incorrect are each represented by 5 of the 10 digits.

To aid in carrying out such a simulation, tables of random digits (such as Appendix A Table 1) or computer-generated random digits can be used. The numbers in Appendix A Table 1 were generated using a computer’s random number generator.

To see how a table of random numbers can be used to carry out a simulation, let’s reconsider the quiz example. We use a random digit to represent the guess on a single question, with an even digit representing a correct response. A sequence of 20 digits could represent the answers to the 20 quiz questions. We pick an arbitrary starting point in Appendix A Table 1. Suppose that we start at row 10 and take the 20 digits in a row to represent one quiz. The first five “quizzes” and the corresponding number correct (number of even digits) are:

Quiz	Random Digits	Number Correct
1	9 4 6 0 6 9 7 8 8 2 5 2 9 6 0 1 4 6 0 5	13
2	6 6 9 5 7 4 4 6 3 2 0 6 0 8 9 1 3 6 1 8	12
3	0 7 1 7 7 7 2 9 7 8 7 5 8 8 6 9 8 4 1 0	9
4	6 1 3 0 9 7 3 3 6 6 0 4 1 8 3 2 6 7 6 8	11
5	2 2 3 6 2 1 3 0 2 2 6 6 9 7 0 2 1 2 5 8	13

This process would be repeated to generate a large number of observations, which would then be used to construct a table of estimated probabilities.

The method for generating observations must preserve the important characteristics of the actual process being considered if simulation is to be successful. For example, it would be easy to adapt the simulation procedure for the true–false quiz to one for a multiple-choice quiz. Suppose that each of the 20 questions on the quiz has five possible responses, only one of which is correct. For any particular question, we would expect a student to be able to guess the correct answer only one-fifth of the time in the long run.

To simulate this situation, we could select at random from a box that contained four red balls and only one blue ball (or, more generally, four times as many red balls as blue balls). If we are using random digits for the simulation, we could use 0 and 1 to represent a correct response and 2, 3, . . . , 9 to represent an incorrect response.

Using Simulation to Approximate a Probability

1. Propose a method that uses a random mechanism (such as a random number generator or table, the selection of a ball from a box, the toss of a coin, etc.) to represent an observation. Be sure that the important characteristics of the actual process are preserved.
2. Generate an observation using the method from Step 1, and determine whether the outcome of interest has occurred.
3. Repeat Step 2 a large number of times.
4. Calculate the estimated probability by dividing the number of observations for which the outcome of interest occurred by the total number of observations generated.

The simulation process is illustrated in Examples 6.31–6.33.

Example 6.31 Building Permits

Understand the context ➤

Many California cities limit the number of building permits that are issued each year. Because of limited water resources, one such city plans to issue permits for only 10 dwelling units in the upcoming year. The city will decide who is to receive permits by holding a lottery.

Suppose that you are one of 39 individuals who apply for permits. Thirty of these individuals are requesting permits for a single-family home, eight are requesting permits for a duplex (which counts as two dwelling units), and one person is requesting a permit for a small apartment building with eight units (which counts as eight dwelling units). Each request will be entered into the lottery.

Requests will be selected at random one at a time, and if there are enough permits remaining, the request will be granted. This process will continue until all 10 permits have been issued. If your request is for a single-family home, what is the probability that you receive a permit? We can use simulation to estimate this probability. (It is not easy to determine analytically.)

To carry out the simulation, we can view the requests as being numbered from 1 to 39 as follows:

- 01–30 Requests for single-family homes
- 31–38 Requests for duplexes
- 39 Request for 8-unit apartment

For ease of discussion, let's assume that your request is number 01.

One method for simulating the permit lottery consists of these three steps:

Formulate a plan ➤

1. Choose a random number between 01 and 39 to indicate which permit request is selected first, and grant this request.
2. Select another random number between 01 and 39 to indicate which permit request is considered next. Determine the number of dwelling units for the selected request. Grant the request only if there are enough permits remaining to satisfy the request.
3. Repeat Step 2 until permits for 10 dwelling units have been granted.

Do the work ➤

We used Minitab to generate random numbers between 01 and 39 to imitate the lottery drawing. (The random number table in Appendix A Table 1 could also be used by selecting two digits and ignoring 00 and any value over 39.) The first sequence generated by Minitab was

Random Number	Type of Request	Total Number of Units So Far
25	Single-family home	1
07	Single-family home	2
38	Duplex	4
31	Duplex	6
26	Single-family home	7
12	Single-family home	8
33	Duplex	10

We would stop at this point, because permits for 10 units would have been issued. In this simulated lottery, Request 01 was not selected, so you would not have received a permit.

The next simulated lottery (using Minitab to generate the selections) was as follows:

Random Number	Type of Request	Total Number of Units So Far
38	Duplex	2
16	Single-family home	3
30	Single-family home	4
39	Apartment—not granted, since there are not 8 permits remaining	4
14	Single-family home	5
26	Single-family home	6
36	Duplex	8
13	Single-family home	9
15	Single-family home	10

Again, Request 01 was not selected, so you would not have received a permit in this simulated lottery.

Now that a strategy for simulating a lottery has been devised, the tedious part of the simulation begins. We would now simulate a large number of lottery drawings, determining for each one whether Request 01 was granted. We simulated 500 such drawings and found that Request 01 was selected in 85 of the lotteries. This results in

$$\text{estimated probability of receiving a building permit} = \frac{85}{500} = 0.17$$

Example 6.32 One-Boy Family Planning

Understand the context ➤

Suppose that couples who wanted children were to continue having children until a boy is born. Assuming that each newborn child is equally likely to be a boy or a girl, would this behavior change the proportion of boys in the population? This question was posed in an article that appeared in *The American Statistician (“What Some Puzzling Problems Teach About the Theory of Simulation and the Use of Resampling” [1994]: 290–293)*, and many people answered the question incorrectly.

We will use simulation to estimate the long-run proportion of boys in the population if families were to continue to have children until they have a boy. This proportion is an estimate of the probability that a randomly selected child from this population is a boy. Notice that in this population, every sibling group would have exactly one boy.

Formulate a plan ➤

We use a single-digit random number to represent a child. The odd digits (1, 3, 5, 7, 9) will represent a male birth, and the even digits will represent a female birth. An observation is constructed by selecting a sequence of random digits. If the first random number obtained is odd (a boy), the observation is complete. If the first selected number is even (a girl), another digit is chosen. We continue in this way until an odd digit is obtained. For example, reading across row 15 of the random number table (Appendix A Table 1), the first 10 digits are

0 7 1 7 4 2 0 0 0 1

Do the work ➤

Using these numbers to simulate sibling groups, we get

Sibling group 1	0 7	girl, boy
Sibling group 2	1	boy
Sibling group 3	7	boy
Sibling group 4	4 2 0 0 0 1	girl, girl, girl, girl, girl, boy

Continuing along row 15 of the random number table,

Sibling group 5	3	boy
Sibling group 6	1	boy
Sibling group 7	2 0 4 7	girl, girl, girl, boy
Sibling group 8	8 4 1	girl, girl, boy

Interpret the results ➤

After simulating eight sibling groups, we have 8 boys among 19 children. The proportion of boys is 8/19, which is close to 0.5. Continuing the simulation to obtain a large number of observations suggests that the long-run proportion of boys in the population would still be 0.5, which is indeed the case.

Example 6.33 ESP?

Understand the context ➤

Can a close friend read your mind? Try the following chance experiment. Write the word *blue* on one piece of paper and the word *red* on another, and place the two slips of paper in a box. Select one slip of paper from the box, look at the word written on it, and then try to convey the word by sending a mental message to a friend who is seated in the same room. Ask your friend to select either red or blue, and record whether the response is correct. Repeat this 9 more times and determine the number of correct responses. How did your friend do? Is your friend receiving your mental messages or just guessing?

Formulate a plan ➤

Let's investigate by using simulation to get the approximate probabilities of the various possible numbers of correct responses for someone who is guessing. Someone who is guessing should have an equal chance of responding correctly or incorrectly. We can use a random digit to represent a response, with an even digit representing a correct response (C) and an odd digit representing an incorrect response (X). A sequence of 10 digits can be used to simulate performing the chance experiment one time.

For example, using the last 10 digits in row 25 of the random number table (Appendix A Table 1) gives

5	2	8	3	4	3	0	7	3	5
X	C	C	X	C	X	C	X	X	X

which is a simulated chance experiment resulting in four correct responses. We used Minitab to generate 150 sequences of 10 random digits and obtained the following results:

Sequence Number	Digits	Number Correct
1	3996285890	5
2	1690555784	3
3	9133190550	2
:	:	:
149	3083994450	5
150	9202078546	7

Table 6.2 summarizes the results of our simulation.

Interpret the results ➤

The estimated probabilities in Table 6.2 are based on the assumption that a correct and an incorrect response are equally likely (guessing). Evaluate your friend's performance in light of the information in Table 6.2. Is it likely that someone who is guessing would have been able to get as many correct as your friend did? Do you think your friend was receiving your mental messages? How are the estimated probabilities in Table 6.2 used to support your answer?

TABLE 6.2
Estimated Probabilities for Example 6.33

Number Correct	Number of Sequences	Estimated Probability
0	0	0.000
1	1	0.007
2	8	0.053
3	16	0.107
4	30	0.200
5	36	0.240
6	35	0.233
7	17	0.113
8	7	0.047
9	0	0.000
10	0	0.000
Total	150	1.000

EXERCISES 6.88 - 6.99

● Data set available online

- 6.88** The report “Airline Quality Rating 2016” (airlinequalityrating.com/reports/2016_AQR_Final.pdf, retrieved April 25, 2017) provides an overview of the complaints about airlines received by the U.S. Department of Transportation. The table below gives the number of complaints received by type of complaint for the years 2014 and 2015.

Type of Complaint	2014 Number of Complaints Received	2015 Number of Complaints Received
Flight problems	4,304	5,506
Baggage handling	1,628	2,050
Reservations, ticketing, boarding	1,276	1,807
Customer service	1,201	1,728
Fares	699	1,300
Other	2,257	2,869

Suppose that one of these complaints is randomly selected for a follow-up interview. Use the given information to estimate the probabilities on the next page.

- a. The probability that a complaint made in 2014 was about baggage handling.
- b. The probability that a complaint made in 2015 was not about flight problems.
- c. The probability that two complaints made in 2015 were both about flight problems.
- d. The probability that a complaint made in 2014 was either about flight problems or customer service.
- 6.89** Five hundred first-year students at a state university were classified according to both high school GPA and whether they were on academic probation at the end of their first semester. The data are summarized in the accompanying table.

Probation	High School GPA			Total
	2.5 to <3.0	3.0 to <3.5	3.5 and Above	
Yes	50	55	30	135
No	45	135	185	365
Total	95	190	215	500

- a. Construct a table of the estimated probabilities for each GPA–probation combination. (Hint: See Example 6.25.)
- b. Use the table constructed in Part (a) to approximate the probability that a randomly selected first-year student at this university will be on academic probation at the end of the first semester.
- c. What is the estimated probability that a randomly selected first-year student at this university had a high school GPA of 3.5 or above?
- d. Are the two events *selected student has a high school GPA of 3.5 or above* and *selected student is on academic probation at the end of the first semester* independent events? How can you tell?

Table for Exercise 6.91

Gender	Education	College					
		Engineering	Liberal Arts	Science and Math	Agriculture	Business	Architecture
Male	200	3,200	2,500	1,500	2,100	1,500	200
Female	300	800	1,500	1,500	900	1,500	300

Table for Exercise 6.92

County	Race/Ethnicity				
	Caucasian	Hispanic	Black	Asian	American Indian
Monterey	163,000	139,000	24,000	39,000	4,000
San Luis Obispo	180,000	37,000	7,000	9,000	3,000
Santa Barbara	230,000	121,000	12,000	24,000	5,000
Ventura	430,000	231,000	18,000	50,000	7,000

- d. If one person is selected at random from this region, what is the estimated probability that the selected person is an Asian from San Luis Obispo County?

6.93 Refer to the information given in the previous exercise.

- a. If one person is selected at random from this region, what is the estimated probability that the person is either Asian or from San Luis Obispo County?
- b. If one person is selected at random from this region, what is the estimated probability that the person is Asian or from San Luis Obispo County but not both?
- c. If two people are selected at random from this region, what is the estimated probability that both are Caucasians?
- d. If two people are selected at random from this region, what is the estimated probability that neither is Caucasian?

6.94 Refer to the information given in Exercises 6.92 and 6.93.

- a. If two people are selected at random from this region, what is the estimated probability that exactly one is a Caucasian?
- b. If two people are selected at random from this region, what is the estimated probability that both are residents of the same county?
- c. If two people are selected at random from this region, what is the estimated probability that both are from different racial/ethnic groups?

6.95 A medical research team wishes to evaluate two different treatments for a disease. Subjects are selected two at a time, and then one of the pair is assigned to each of the two treatments. The treatments are applied, and each is either a success (S) or a failure (F).

The researchers keep track of the total number of successes for each treatment. They plan to continue the chance experiment until the number of successes for one treatment exceeds the number of successes or the other treatment by 2. For example, they might observe the results in the table for Exercise 6.95 given below. The chance experiment would stop after the sixth pair, because Treatment 1 has two more successes than Treatment 2. The researchers would conclude that Treatment 1 is preferable to Treatment 2.

Suppose that Treatment 1 has a success rate of 0.7 (that is, $P(\text{success}) = 0.7$ for Treatment 1) and that Treatment 2 has a success rate of 0.4. Use simulation to estimate the probabilities in Parts (a) and (b) by using the following procedure:

1. Use a pair of random digits to simulate one pair of subjects. Let the first digit represent Treatment 1 and use 1–7 as an indication of a success and 8, 9, and 0 to indicate a failure. Let the second digit represent Treatment 2, with 1–4 representing a success. For example, if the two digits selected to represent a pair were 8 and 3, you would record failure for Treatment 1 and success for Treatment 2.
2. Continue to select pairs, keeping track of the total number of successes for each treatment. Stop the trial as soon as the number of successes for one treatment exceeds that for the other by 2. This would complete one trial.
3. Repeat this whole process until you have results for at least 20 trials (more is better).
4. Use the simulation results to estimate the desired probabilities.
 - a. Estimate the probability that more than five pairs must be treated before a conclusion can be reached. (Hint: $P(\text{more than } 5) = 1 - P(5 \text{ or fewer})$.)
 - b. Estimate the probability that the researchers will incorrectly conclude that Treatment 2 is the better treatment.

Table for Exercise 6.95

Pair	Treatment 1	Treatment 2	Total Number of Successes for Treatment 1	Total Number of Successes for Treatment 2
1	S	F	1	0
2	S	S	2	1
3	F	F	2	1
4	S	S	3	2
5	F	F	3	2
6	S	F	4	2

- 6.96** Many cities regulate the number of taxi licenses, and there is a great deal of competition for both new and existing licenses. Suppose that a city has decided to sell 10 new licenses for \$25,000 each. A lottery will be held to determine who gets the licenses, and no one may request more than three licenses.

Twenty individuals and taxi companies have entered the lottery. Six of the 20 entries are requests for 3 licenses, 9 are requests for 2 licenses, and the rest are requests for a single license. The city will select requests at random, filling as many of the requests as possible.

For example, the city might fill requests for 2, 3, 1, and 3 licenses and then select a request for 3. Because there is only one license left, the last request selected would receive a license, but only one.

- An individual has put in a request for a single license. Use simulation to approximate the probability that the request will be granted. Perform *at least* 20 simulated lotteries (more is better!).
- Do you think that this is a fair way of distributing licenses? Can you propose an alternative procedure for distribution?

- 6.97** Four students must work together on a group project. They decide that each will take responsibility for a particular part of the project, as follows:

Person	Maria	Alex	Juan	Jacob
Task	Survey design	Data collection	Analysis	Report writing

Because of the way the tasks have been divided, one student must finish before the next student can begin work.

To ensure that the project is completed on time, a schedule is established, with a deadline for each team member. If any one of the team members is late, the timely completion of the project is jeopardized. Assume the following probabilities:

- The probability that Maria completes her part on time is 0.8.

- If Maria completes her part on time, the probability that Alex completes on time is 0.9, but if Maria is late, the probability that Alex completes on time is only 0.6.
- If Alex completes his part on time, the probability that Juan completes on time is 0.8, but if Alex is late, the probability that Juan completes on time is only 0.5.
- If Juan completes his part on time, the probability that Jacob completes on time is 0.9, but if Juan is late, the probability that Jacob completes on time is only 0.7.

Use simulation (with at least 20 trials) to estimate the probability that the project is completed on time. Think carefully about this one. For example, you might use a random digit to represent each part of the project (four in all). For the first digit (Maria's part), 1–8 could represent *on time* and 9 and 0 could represent *late*. Depending on what happened with Maria (late or on time), you would then look at the digit representing Alex's part. If Maria was on time, 1–9 would represent *on time* for Alex, but if Maria was late, only 1–6 would represent *on time*. The parts for Juan and Jacob could be handled similarly.

- 6.98** In Exercise 6.97, the probability that Maria completes her part on time was 0.8. Suppose that this probability is really only 0.6. Use simulation (with at least 20 trials) to estimate the probability that the project is completed on time.

- 6.99** Refer to Exercises 6.97 and 6.98. Suppose that the probabilities of timely completion are as in Exercise 6.97 for Maria, Alex, and Juan, but that Jacob has a probability of completing on time of 0.7 if Juan is on time and 0.5 if Juan is late.

- Use simulation (with at least 20 trials) to estimate the probability that the project is completed on time.
- Compare the probability from Part (a) to the one computed in Exercise 6.98. Which decrease in the probability of on-time completion (Maria's or Jacob's) made the bigger change in the probability that the project is completed on time?

CHAPTER ACTIVITIES

ACTIVITY 6.1 KISSES

Background: The paper “[What Is the Probability of a Kiss? \(It’s Not What You Think\)](#)” (*Journal of Statistics Education* (online) [2002]) posed the following question: What is the probability that a Hershey’s Kiss will land on its base (as opposed to its side) if it is flipped onto a table? Unlike flipping a coin, there is no reason to believe that this probability would be 0.5.

Working as a class, develop a plan that would enable you to estimate this probability empirically.

Once you have an acceptable plan, carry it out and use the resulting data to produce an estimate of the desired probability. Do you think that a kiss is equally likely to land on its base or on its side? Explain.

ACTIVITY 6.2 A CRISIS FOR EUROPEAN SPORTS FANS?

Background: The *New Scientist* (January 4, 2002) reported on a controversy surrounding the Euro coins that have been introduced as a common currency across Europe. Each country mints its own coins, but these coins are accepted in any of the countries that have adopted the Euro as their currency.

A group in Poland claims that the Belgium-minted Euro does not have an equal chance of landing heads or tails. This claim was based on 250 tosses of the Belgium-minted Euro, of which 140 (56%) came up heads. Should this be cause for alarm for European sports fans, who know that “important” decisions are made by the flip of a coin?

In this activity, we will investigate whether this should be cause for alarm by examining whether observing 140 heads out of 250 tosses is an unusual outcome if the coin is fair.

- For this first step, you can either (a) flip a U.S. penny 250 times, keeping a tally of the number of

heads and tails observed (this won’t take as long as you think), or (b) simulate 250 coin tosses by using your calculator or a statistics software package to generate random numbers (if you choose this option, give a brief description of how you carried out the simulation).

- For your sequence of 250 tosses, calculate the proportion of heads observed.
- Form a data set that consists of the values for proportion of heads observed in 250 tosses of a fair coin for the entire class. Summarize this data set by constructing a graphical display.
- Working with a partner, write a paragraph explaining why European sports fans should or should not be worried by the results of the Polish experiment. Your explanation should be based on the observed proportion of heads from the Polish experiment and the graphical display constructed in Step 3.

ACTIVITY 6.3 THE ‘HOT HAND’ IN BASKETBALL

Background: Consider a mediocre basketball player who has consistently made only 50% of his free throws over several seasons. If we were to examine his free throw record over the last 50 free throw attempts, is it likely that we would see a streak of 5 in a row where he is successful in making the free throw? In this activity, we will investigate this question. We will assume that the outcomes of successive free throw attempts are independent and that the probability that the player is successful on any particular attempt is 0.5.

- Begin by simulating a sequence of 50 free throws for this player. Because this player has probability of success of 0.5 for each attempt and the attempts are independent, we can model a free throw by tossing a coin. Using heads to represent a successful free throw and tails to represent a missed free throw, simulate 50 free throws by tossing a coin 50 times, recording the outcome of each toss.
- For your sequence of 50 tosses, identify the longest streak by looking for the longest string of heads in your sequence. Determine the length of this longest streak.
- Combine your longest streak value with those from the rest of the class and construct a histogram or dotplot of these longest streak values.
- Based on the graph from Step 3, does it appear likely that a player of this skill level would have a streak of

5 or more successes sometime during a sequence of 50 free throw attempts? Justify your answer based on the graph from Step 3.

- Use the combined class data to estimate the probability that a player of this skill level has a streak of at least 5 somewhere in a sequence of 50 free throw attempts.
- Using the multiplication rule for independent events, we can calculate the probability that a player of this skill level is successful on the *next 5* free throw attempts:

$$P(\text{SSSSS}) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^5 = 0.031$$

which is relatively small. At first this might seem inconsistent with your answer in Step 5, but the estimated probability from Step 5 and the computed probability of 0.031 are really considering different situations. Explain why it is plausible that both probabilities could be correct.

- Do you think that the assumption that the outcomes of successive free throws are independent is reasonable? Explain. (This is a hotly debated topic among both sports fans and statisticians!)

SUMMARY Key Concepts and Formulas

TERM OR FORMULA

Chance experiment

Sample space

Event

Simple event

Events

1. *not A, A^c*
2. *A or B, $A \cup B$*
3. *A and B, $A \cap B$*

Mutually exclusive events

Fundamental properties of probability

$$P(E) = \frac{\text{number of outcomes in } E}{N}$$

$$P(E \cup F) = P(E) + P(F)$$

$$P(E_1 \cup \dots \cup E_k) = P(E_1) + \dots + P(E_k)$$

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

Independence of events E and F

$$P(E|F) = P(E)$$

$$P(E \cap F) = P(E)P(F)$$

$$P(E_1 \cap \dots \cap E_k) = P(E_1)P(E_2) \cdots P(E_k)$$

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

$$P(E \cap F) = P(E|F)P(F)$$

$$P(E) = P(E|B_1)P(B_1) + P(E|B_2)P(B_2) + \dots + P(E|B_k)P(B_k)$$

$$P(B_i|E) = \frac{P(E|B_i)P(B_i)}{P(E|B_1)P(B_1) + P(E|B_2)P(B_2) + \dots + P(E|B_k)P(B_k)}$$

COMMENT

Any experiment for which there is uncertainty concerning the resulting outcome.

The collection of all possible outcomes from a chance experiment.

Any collection of possible outcomes from a chance experiment.

An event that consists of a single outcome.

1. The event consisting of all outcomes not in A .
2. The event consisting of all outcomes in at least one of the two events.
3. The event consisting of outcomes common to both events.

Events that have no outcomes in common.

Fundamental properties of probability

1. The probability of any event must be a number between 0 and 1.
2. If S is the sample space for a chance experiment, $P(S) = 1$
3. If E and F are mutually exclusive events, $P(E \cup F) = P(E) + P(F)$
4. $P(E) + P(E^c) = 1$

$P(E)$ when the outcomes are *equally likely* and where N is the number of outcomes in the sample space.

Addition rules when events are *mutually exclusive*.

The conditional probability of the event E given that the event F has occurred.

Events E and F are independent if the probability that E has occurred given F is the same as the probability that E will occur with no knowledge of F .

Multiplication rules for *independent* events.

The general addition rule for two events.

The general multiplication rule for two events.

The law of total probability, where B_1, B_2, \dots, B_k are mutually exclusive events with $P(B_1) + P(B_2) + \dots + P(B_k) = 1$

Bayes' rule, where B_1, B_2, \dots, B_k are mutually exclusive events with $P(B_1) + P(B_2) + \dots + P(B_k) = 1$

CHAPTER REVIEW Exercises 6.100 - 6.120

- 6.100** False positive results are not uncommon with mammograms, a test used to screen for breast cancer. For a woman who has a positive mammogram, the probability that she actually has breast cancer is less than 0.05 if she is under 40 years old, and ranges from 0.05 to 0.109 if she is over 40 years old (["Breast Cancer Screenings: Does the Evidence Support the Recommendations?," Significance \[August 2016\]: 24–37](#)). If a woman with a positive mammogram is selected at random, are the two events

B = event that selected woman has breast cancer and

A = event that selected woman is over 40 years old independent events? Justify your answer using the given information.

- 6.101** A company uses three different assembly lines— A_1, A_2 , and A_3 —to manufacture a particular component. Of those manufactured by A_1 , 5% need rework to remedy a defect, whereas 8% of A_2 's components and 10% of A_3 's components need rework. Suppose that 50% of all components are produced by A_1 , 30% are produced by A_2 and 20% are produced by A_3 .

- a. Construct a tree diagram with first-generation branches corresponding to the three lines. Leading from each branch, draw one branch for rework (R) and another for no rework (N). Then enter appropriate probabilities on the branches.
- b. What is the probability that a randomly selected component came from A_1 and needed rework?
- c. What is the probability that a randomly selected component needed rework?

6.102 Consider the following information about passengers on a cruise ship on vacation: 40% check work e-mail, 30% use a cell phone to stay connected to work, 25% bring a laptop with them on vacation, 23% both check work e-mail and use a cell phone to stay connected, and 51% neither check work e-mail nor use a cell phone to stay connected nor bring a laptop. In addition, 88% of those who bring a laptop also check work e-mail and 70% of those who use a cell phone to stay connected also bring a laptop. With

E = event that a traveler on vacation checks work e-mail

C = event that a traveler on vacation uses a cell phone to stay connected

L = event that a traveler on vacation brought a laptop
use the given information to determine the following probabilities. A Venn diagram may help.

- a. $P(E)$
- b. $P(C)$
- c. $P(L)$
- d. $P(E \text{ and } C)$

6.103 Use the information given in the previous exercise to determine the following probabilities.

- a. $P(E \text{ and } C \text{ or } L)$
- b. $P(E|L)$
- c. $P(L|C)$

6.104 Use the information given in exercise 6.102 to determine the following probabilities.

- a. $P(E \text{ and } C \text{ and } L)$
- b. $P(E \text{ and } L)$
- c. $P(C \text{ and } L)$
- d. $P(C|E \text{ and } L)$

6.105 The article “**Obesity, Smoking Damage U.S. Economy,**” which appeared in the *Gallup Online Business Journal* (gallup.com, September 7, 2016, retrieved April 25, 2017), reported that based on a large representative sample of adult Americans, 52.7% claimed that they exercised at least

30 minutes on 3 or more days per week during 2015. It also reported that the percentage for millennials (people age 19–35) was 57.1%, and for those over 35 it was 51.1%. If an adult American were to be selected at random, are the events *selected adult exercises at least 30 minutes 3 times per week* and *selected adult is a millennial* independent or dependent events? Justify your answer using the given information.

6.106 The following table summarizing data on smoking status and age group is consistent with summary quantities obtained in a Gallup Poll published in the online article “**In U.S., Young Adults’ Cigarette Use is Down Sharply**” (gallup.com, December 10, 2015, retrieved April 25, 2017).

Age Group	Smoking Status	
	Smoker	Nonsmoker
18 to 29	174	618
30 to 49	333	1,115
50 to 64	384	1,445
65 and older	211	1,707

Assume that it is reasonable to consider these data as representative of the American adult population. Consider the chance experiment or randomly selecting an adult American.

- a. What is the probability that the selected adult is a smoker?
- b. What is the probability that the selected adult is under 50 years of age?
- c. What is the probability that the selected adult is a smoker that is 65 or older?
- d. What is the probability that the selected adult is a smoker or is age 65 or older?

6.107 A study of the impact of seeking a second opinion about a medical condition is described in the paper “**Evaluation of Outcomes from a National Patient-Initiated Second-Opinion Program**” (*The American Journal of Medicine* [2015], 1138e25–1138e33). Based on a review of 6791 patient-initiated second opinions, the paper states the following: “Second opinions often resulted in changes in diagnosis (14.8%), treatment (37.4%), or changes in both (10.6%).”

Consider the following two events:

D = event that second opinion results in a change in diagnosis

T = event that second opinion results in a change in treatment

- a. What are the values of $P(D)$, $P(T)$, and $P(D \cap T)$?

- b. What is the probability that a second opinion results in neither a change in diagnosis nor a change in treatment?
- c. What is the probability that a second opinion results in a change in diagnosis or a change in treatment?
- 6.108** A company sends 40% of its overnight mail parcels by means of express mail service A_1 . Of these parcels, 2% arrive after the guaranteed delivery time. What is the probability that a randomly selected overnight parcel was shipped by mail service A_1 and was late?
- 6.109** Return to the context of the previous exercise and suppose that 50% of the overnight parcels are sent by means of express mail service A_2 and the remaining 10% are sent by means of A_3 . Of those sent by means of A_2 , only 1% arrived late, whereas 5% of the parcels handled by A_3 arrived late.
- What is the probability that a randomly selected parcel arrived late? (Hint: A tree diagram should help.)
 - Suppose that a randomly selected overnight parcel arrived late. What is the probability that the parcel was shipped using mail service A_1 ? That is, what is the probability of A_1 given L , denoted $P(A_1|L)$?
 - What is $P(A_2|L)$?
 - What is $P(A_3|L)$?
- 6.110** Two individuals, A and B, are finalists for a chess championship. They will play a sequence of games, each of which can result in a win for A, a win for B, or a draw. Suppose that the outcomes of successive games are independent, with $P(\text{A wins game}) = 0.3$, $P(\text{B wins game}) = 0.2$, and $P(\text{draw}) = 0.5$. Each time a player wins a game, he earns 1 point and his opponent earns no points. The first player to win 5 points wins the championship.
- What is the probability that A wins the championship in just five games?
 - What is the probability that it takes just five games to obtain a champion?
 - If a draw earns a half-point for each player, describe how you would perform a simulation to estimate $P(\text{A wins the championship})$. Assume that the championship will end in a draw if both players obtain 5 points at the same time.
- 6.111** A single-elimination tournament with four players is to be held. In Game 1, the players seeded (rated) first and fourth play. In Game 2, the players seeded second and third play. In Game 3, the winners of

Games 1 and 2 play, with the winner of Game 3 declared the tournament winner. Suppose that the following probabilities are given:

$$\begin{aligned}P(\text{seed 1 defeats seed 4}) &= 0.8 \\P(\text{seed 1 defeats seed 2}) &= 0.6 \\P(\text{seed 1 defeats seed 3}) &= 0.7 \\P(\text{seed 2 defeats seed 3}) &= 0.6 \\P(\text{seed 2 defeats seed 4}) &= 0.7 \\P(\text{seed 3 defeats seed 4}) &= 0.6\end{aligned}$$

- Describe how you would use random digits to simulate Game 1 of this tournament.
 - Describe how you would use random digits to simulate Game 2 of this tournament.
 - How would you use random digits to simulate Game 3 in the tournament? (This will depend on the outcomes of Games 1 and 2.)
 - Simulate one complete tournament, giving an explanation for each step in the process.
- 6.112** Refer to the simulation from the previous exercise.
- Simulate 10 tournaments, and use the resulting information to estimate the probability that the first seed wins the tournament.
 - Ask four classmates for their simulation results. Along with your own results, this should give you information for 50 simulated tournaments. Use this information to estimate the probability that the first seed wins the tournament.
 - Why do the estimated probabilities from Parts (a) and (b) differ? Which do you think is a better estimate of the true probability? Explain.
- 6.113** In a school machine shop, 60% of all machine breakdowns occur on lathes and 15% occur on drill presses. Let E denote the event that the next machine breakdown is on a lathe, and let F denote the event that a drill press is the next machine to break down. With $P(E) = 0.60$ and $P(F) = 0.15$, calculate:
- $P(E^c)$
 - $P(E \cup F)$
 - $P(E^c \cap F^c)$
- 6.114** There are five faculty members in a certain academic department. These individuals have 3, 6, 7, 10, and 14 years of teaching experience. Two of these individuals are randomly selected to serve on a personnel review committee. What is the probability that the chosen representatives have a total of at least 15 years of teaching experience? (Hint: Consider all possible committees.)

- 6.115** The general addition rule for three events states that

$$\begin{aligned} P(A \text{ or } B \text{ or } C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \text{ and } B) - P(A \text{ and } C) \\ &\quad - P(B \text{ and } C) + P(A \text{ and } B \text{ and } C) \end{aligned}$$

A new magazine publishes columns entitled “Art” (A), “Books” (B), and “Cinema” (C). Suppose that

14% of all subscribers read A
 23% read B
 37% read C
 8% read A and B
 9% read A and C
 13% read B and C
 5% read all three columns

What is the probability that a randomly selected subscriber reads at least one of these three columns?

- 6.116** A theater complex is currently showing four R-rated movies, three PG-13 movies, two PG movies, and one G movie. The following table gives the number of people at the first showing of each movie on a certain Saturday:

Theater	Rating	Number of Viewers
1	R	600
2	PG-13	420
3	PG-13	323
4	R	196
5	G	254
6	PG	179
7	PG-13	114
8	R	205
9	R	139
10	PG	87

Suppose that one of these people is randomly selected.

- a. What is the probability that the selected individual saw a PG movie?
- b. What is the probability that the selected individual saw a PG or a PG-13 movie?
- c. What is the probability that the selected individual did not see an R movie?

- 6.117** Refer to Exercise 6.116, but now suppose that two viewers are randomly selected (without replacement). Let R_1 and R_2 denote the events that the first and second individuals, respectively, watched an R-rated movie. Are R_1 and R_2 independent events? Explain. From a practical point of view, can these events be regarded as independent? Explain.

- 6.118** Suppose that a box contains 25 light bulbs, of which 20 are good and the other 5 are defective. Consider randomly selecting three bulbs without replacement. Let E denote the event that the first bulb selected is good, F be the event that the second bulb is good, and G represent the event that the third bulb selected is good.

- a. What is $P(E)$?
- b. What is $P(F|E)$?
- c. What is $P(G|E \cap F)$?
- d. What is the probability that all three selected bulbs are good?

- 6.119** Return to Exercise 6.118, and suppose that 4 bulbs are randomly selected from the 25.

- a. What is the probability that all 4 are good?
- b. What is the probability that at least 1 selected bulb is bad?

- 6.120** A transmitter is sending a message using a binary code (a sequence of 0's and 1's). Each transmitted bit (0 or 1) must pass through three relays to reach the receiver. At each relay, the probability is 0.20 that the bit sent on is different from the bit received (a reversal). Assume that the relays operate independently of one another:

transmitter \rightarrow relay 1 \rightarrow relay 2 \rightarrow relay 3 \rightarrow receiver

- a. If a 1 is sent from the transmitter, what is the probability that a 1 is sent on by all three relays?
- b. If a 1 is sent from the transmitter, what is the probability that a 1 is received by the receiver? (Hint: The eight experimental outcomes can be displayed on a tree diagram with three generations of branches, one generation for each relay.)

7

Random Variables and Probability Distributions



mikeledray/Shutterstock.com

T

his chapter is the first of two chapters that together link the basic ideas of probability introduced in Chapter 6 with the methods of statistical inference. Chapter 6 used probability to describe the long-run relative frequency of occurrence of various types of outcomes. In this chapter we introduce probability models that can be used to describe the distribution of values of a variable.

In Chapter 8, we will see how these same probability models can be used to describe the behavior of sample statistics. We will also see that this is what allows us to draw conclusions based on sample data.

In a chance experiment, we often focus on some numerical aspect of the outcome. For example, an environmental scientist who obtains an air sample from a specified location might be interested in the ozone concentration (ozone is a major constituent of smog). A quality control inspector who must decide whether to accept a large shipment of plastic cell phone cases may base the decision on the number of defective cases in a group of 20 cases randomly selected from the shipment.

Before selecting an air sample, the value of the ozone concentration is uncertain. Similarly, the number of defective cases among the 20 selected might be any whole number between 0 and 20. Because the value of a variable quantity such as ozone concentration or number of defective cases is subject to uncertainty, such variables are called *random variables*.

In this chapter we begin by distinguishing between discrete and continuous numerical random variables. We show how the behavior of both discrete and continuous numerical random variables can be described by a probability distribution. This distribution can then be used to make probability statements about values of the random variable. Three commonly encountered probability distributions (the binomial, geometric, and normal distributions) are also introduced.

LEARNING OBJECTIVES

Students will understand:

- That a probability distribution describes the long-run behavior of a random variable.
- That areas under a density curve for a continuous random variable are interpreted as probabilities.

Students will be able to:

- Distinguish between discrete and continuous random variables.
- Construct the probability distribution of a discrete random variable.

- Calculate and interpret the mean and standard deviation of a discrete random variable.
- Distinguish between binomial and geometric random variables.
- Calculate and interpret binomial probabilities.
- Calculate probabilities involving continuous random variables whose density curves have a simple form.
- Find an area under a normal curve and interpret this area as a probability.
- Construct and interpret a normal probability plot.

SECTION 7.1 Random Variables

In many chance experiments, we are interested in one or more variable quantities. For example, consider a management consultant who is studying the operation of a supermarket. A chance experiment might involve randomly selecting a customer leaving the store. One numerical variable that might be of interest is the number of items purchased by the customer. We can denote this variable using a letter, such as x . Possible values of this variable are 0, 1, 2, 3, and so on. The possible values of x are isolated points on the number line. Until a customer is selected and the number of items counted, there is uncertainty about what value of x will be observed.

Another variable of interest might be the time y (in minutes) spent in a checkout line. One possible value of y is 3.0 minutes and another is 4.0 minutes, but *any* other number between 3.0 and 4.0 is also a possibility. The possible y values form an entire interval (a continuum) on the number line.

DEFINITIONS

Random variable: A numerical variable whose value depends on the outcome of a chance experiment. A random variable associates a numerical value with each outcome of a chance experiment.

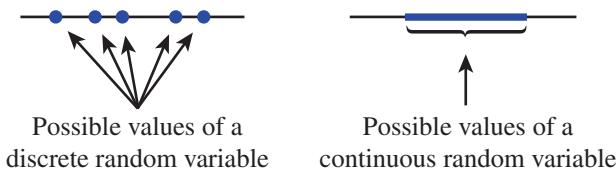
Discrete random variable: A random variable is discrete if its set of possible values is a collection of isolated points along the number line.

Continuous random variable: A random variable is continuous if its set of possible values includes an entire interval on the number line.

We use lowercase letters, such as x and y , to represent random variables.* Figure 7.1 shows a set of possible values for each type of random variable. In practice, a discrete random variable is usually the result of counting (for example, the number of items purchased, the number of gas pumps in use, or the number of broken eggs in a carton). A continuous random variable is one whose value is typically obtained by measurement (temperature in a freezer, weight of a pineapple, amount of time spent in a store, etc.).

FIGURE 7.1

Two different types of random variables.



Because there is a limit to the accuracy of any measuring instrument, such as a watch or a scale, it may seem that any variable should be regarded as discrete. For example, when weight is measured to the nearest pound, the observed values appear to be isolated points along the number line, such as 2 pounds or 3 pounds. But this is just a function of the

*In some books, uppercase letters are used to name random variables, with lowercase letters representing a particular value that the variable might assume. We have chosen to use a simpler and less formal notation.

accuracy with which weight is recorded and not because a weight between 2 and 3 pounds is impossible. In this case, the variable is continuous.

Example 7.1 Book Sales

Consider an experiment in which the type of book, print (P) or digital (D), chosen by each of three successive customers making a purchase of a single book from an online bookstore is noted. Define a random variable x by

$$x = \text{number of customers purchasing a digital book}$$

The outcome in which the first and third customers purchase a digital book and the second customer purchases a print book can be abbreviated DPD. The associated x value is 2, because two of the three customers selected a digital book. Similarly, the x value for the outcome DDD (all three purchase a digital book) is 3.

The eight possible outcomes and the corresponding values of x are displayed in the following table:

Outcome	PPP	DPP	PDP	PPD	DDP	DPD	PDD	DDD
x value	0	1	1	1	2	2	2	3

There are only four possible x values—0, 1, 2, and 3—and these are isolated points on the number line. This means that x is a discrete random variable.

In some situations, the random variable of interest is discrete, but the number of possible values is not finite. This is illustrated in Example 7.2.

Example 7.2 This Could Be a Long Game . . .

Two friends agree to play a game that consists of a sequence of trials. The game continues until one player wins two trials in a row. One random variable of interest might be

$$x = \text{number of trials required to complete the game}$$

Let A denote a win for Player 1 and B denote a win for Player 2. The simplest possible experimental outcomes are AA (the case in which Player 1 wins the first two trials and the game ends) and BB (the case in which Player 2 wins the first two trials). With either of these two outcomes, $x = 2$. There are also two outcomes for which $x = 3$: ABB and BAA. Some other possible outcomes and associated x values are shown in the following table.

Outcomes	x value
AA, BB	2
BAA, ABB	3
ABAA, BABB	4
ABABB, BABAA	5
.	.
ABABABABAA, BABABABABB	10

Any positive integer that is 2 or greater is a possible value. Because the values 2, 3, 4, ... are isolated points on the number line, x is a discrete random variable even though there is no upper limit to the number of possible values.

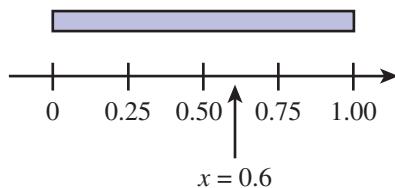
Example 7.3 Stress

In an engineering stress test, pressure is applied to a thin 1-foot-long bar until the bar snaps. The precise location where the bar will snap is uncertain. Let x be the distance from the left end of the bar to the break. Then $x = 0.25$ is one possibility, $x = 0.9$ is another, and in

fact any number between 0 and 1 is a possible value of x . (Figure 7.2 shows the case of the outcome $x = 0.6$.) The set of possible values is an entire interval on the number line, so x is a continuous random variable.

FIGURE 7.2

The bar for Example 7.3 and the outcome $x = 0.6$.



Even though in practice we may only be able to measure the distance to the nearest tenth of an inch or hundredth of an inch, the *actual* distance could be any number between 0 and 1. So, even though the recorded values might be rounded because of the accuracy of the measuring instrument, the variable is still continuous.

In data analysis, random variables often arise in the context of summarizing sample data when a sample is selected from some population. This is illustrated in Example 7.4.

Example 7.4 College Plans

Suppose that a high school counselor plans to select a random sample of 50 seniors and to ask each student in the sample whether he or she plans to attend college after graduation. The process of sampling is a chance experiment. The sample space for this experiment consists of all the different possible random samples of size 50 that might result (there are a very large number of these), and for simple random sampling, each of these outcomes is equally likely.

Suppose that the counselor is interested in the number of students in the sample who plan to attend college. We can use the letter x to represent this number:

$$x = \text{number of students in the sample who plan to attend college}$$

Then x is a random variable, because it associates a numerical value with each of the possible outcomes (random samples). Possible values of x are 0, 1, 2, . . . , 50, and x is a discrete random variable.

EXERCISES 7.1 - 7.7

- 7.1** State whether each of the following random variables is discrete or continuous:

- a. The number of defective tires on a car
- b. The body temperature of a hospital patient
- c. The number of pages in a book
- d. The number of draws (with replacement) from a deck of playing cards until a heart is selected
- e. The lifetime of a lightbulb

- 7.2** Classify each of the following random variables as either discrete or continuous:

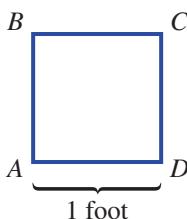
- a. The fuel efficiency (mpg) of an automobile
- b. The amount of rainfall at a particular location during the next year
- c. The distance that a person throws a baseball
- d. The number of questions asked during a 1-hour lecture

- e. The tension (in pounds per square inch) at which a tennis racket is strung
- f. The amount of water used by a household during a given month

- g. The number of traffic citations issued by the highway patrol in a particular county on a given day

- 7.3** Starting at a particular time, each car entering an intersection is observed to see whether it turns left (L) or right (R) or goes straight ahead (S). The experiment terminates as soon as a car is observed to go straight. Suppose that the random variable y denotes the number of cars observed.
- a. What are possible y values?
 - b. List five different outcomes and their associated y values. (Hint: See Example 7.2.)

- 7.4** A point is randomly selected from the interior of the square pictured.



Let x denote the distance from the lower left-hand corner A of the square to the selected point.

- What are possible values of x ?
 - Is x a discrete or a continuous variable? (Hint: See Example 7.3.)
- 7.5** A point is randomly selected on the surface of a lake that has a maximum depth of 100 feet. Let y be the depth of the lake at the randomly chosen point.
- What are possible values of y ?
 - Is y discrete or continuous?

- 7.6** A person stands at the corner marked A of the square pictured in Exercise 7.4 and tosses a coin. If it lands heads up, the person moves one corner clockwise, to B . If the coin lands tails up, the person moves one corner counterclockwise, to D . This process is then repeated until the person arrives back at A . Let y denote the number of coin tosses.

- What are possible values of y ?
- Is y discrete or continuous?

- 7.7** A box contains four slips of paper marked 1, 2, 3, and 4. Two slips are selected without replacement. List the possible values for each of the following random variables:
- x = sum of the two numbers
 - y = difference between the first and second numbers
 - z = number of slips selected that show an even number
 - w = number of slips selected that show a 4

SECTION 7.2 Probability Distributions for Discrete Random Variables

The probability distribution for a random variable describes the long-run behavior of the variable. For example, suppose that the county Department of Animal Regulation is interested in studying the variable

$$x = \text{number of licensed dogs or cats for a household}$$

Suppose that county regulations prohibit more than five dogs or cats per household. If we consider the chance experiment of randomly selecting a household in this county, then x is a discrete random variable because it associates a numerical value (0, 1, 2, 3, 4, or 5) with each of the possible outcomes (households) in the sample space.

Although we know what the possible values for x are, it would also be useful to know how this variable behaves in repeated observation. What would be the most common value? What proportion of the time would $x = 5$ be observed? $x = 3$? A probability distribution provides this type of information about the long-run behavior of a random variable.

The **probability distribution of a discrete random variable x** gives the probability associated with each possible x value. Each probability is the long-run relative frequency of occurrence of the corresponding x value when a chance experiment is performed a very large number of times.

Common ways to display a probability distribution for a discrete random variable are a table, a probability histogram, or a formula.

Notation: If one possible value of x is 2, we often write $p(2)$ in place of $P(x = 2)$. Similarly, $p(5)$ denotes the probability that $x = 5$, and so on.

Example 7.5 Energy Efficient Refrigerators

Understand the context ➤

Suppose that each of four randomly selected customers purchasing a refrigerator at a large appliance store chooses either an energy efficient model (E) or one from a less expensive group of models (G) that do not have an energy efficient rating. Suppose that these customers

make their choices independently of one another and that 40% of all customers select an energy efficient model. This implies that for any particular one of the four customers, $P(E) = 0.4$ and $P(G) = 0.6$.

One possible outcome is EGGE, where the first and fourth customers select energy efficient models and the other two choose less expensive models. Because the customers make their choices independently, the multiplication rule for independent events implies that

$$\begin{aligned} P(\text{EGGE}) &= P(\text{1st chooses E and 2nd chooses G and 3rd chooses G and 4th chooses E}) \\ &= P(E)P(G)P(G)P(E) \\ &= (0.4)(0.6)(0.6)(0.4) \\ &= 0.0576 \end{aligned}$$

Similarly,

$$\begin{aligned} P(\text{EGEG}) &= P(E)P(G)P(E)P(G) \\ &= (0.4)(0.6)(0.4)(0.6) \\ &= 0.0576 \quad (\text{which is equal to } P(\text{EGGE})) \end{aligned}$$

and

$$P(\text{GGGE}) = (0.6)(0.6)(0.6)(0.4) = 0.0864$$

The number among the four customers who purchase an energy efficient model is a random variable. We can denote this variable by x :

x = the number of energy efficient refrigerators purchased by the four customers

Table 7.1 displays the 16 possible outcomes, the probability of each outcome, and the value of the random variable x that is associated with each outcome.

TABLE 7.1 Outcomes and Probabilities for Example 7.5

Outcome	Probability	x Value	Outcome	Probability	x Value
GGGG	0.1296	0	GEEG	0.0576	2
EGGG	0.0864	1	GEGE	0.0576	2
GEGG	0.0864	1	GGEE	0.0576	2
GGEG	0.0864	1	GEEE	0.0384	3
GGGE	0.0864	1	EGEE	0.0384	3
EEGG	0.0576	2	EEGE	0.0384	3
ESEG	0.0576	2	EEEG	0.0384	3
EGGE	0.0576	2	EEEE	0.0256	4

Do the work ► The probability distribution of x is easily obtained from the information in Table 7.1. Consider the smallest possible x value, 0. The only outcome for which $x = 0$ is GGGG, so

$$p(0) = P(x = 0) = P(\text{GGGG}) = 0.1296$$

There are four different outcomes for which $x = 1$, so $p(1)$ is calculated by adding the four corresponding probabilities:

$$\begin{aligned} p(1) &= P(x = 1) = P(\text{EGGG or GEGG or GGEG or GGGE}) \\ &= P(\text{EGGG}) + P(\text{GEGG}) + P(\text{GGEG}) + P(\text{GGGE}) \\ &= 0.0864 + 0.0864 + 0.0864 + 0.0864 \\ &= 4(0.0864) \\ &= 0.3456 \end{aligned}$$

Similarly,

$$\begin{aligned} p(2) &= P(\text{EEGG}) + \dots + P(\text{GGEE}) = 6(0.0576) = 0.3456 \\ p(3) &= 4(0.0384) = 0.1536 \\ p(4) &= 0.0256 \end{aligned}$$

The probability distribution of x is summarized in the following table:

x Value	0	1	2	3	4
$p(x) = \text{Probability of Value}$	0.1296	0.3456	0.3456	0.1536	0.0256

Interpret the results ➤

To interpret $p(3) = 0.1536$, think of performing the chance experiment repeatedly, each time with a new group of four customers. In the long run, 15.36% of these groups will have exactly three customers purchasing an energy efficient refrigerator.

The probability distribution can be used to determine probabilities of various events involving x . For example, the probability that at least two of the four customers choose energy efficient models is

$$\begin{aligned} P(x \geq 2) &= P(x = 2 \text{ or } x = 3 \text{ or } x = 4) \\ &= p(2) + p(3) + p(4) \\ &= 0.5248 \end{aligned}$$

This means that, in the long run, a group of four refrigerator purchasers will include at least two who select energy efficient models 52.48% of the time.

A probability distribution table for a discrete variable shows the possible x values and also $p(x)$ for each possible x value. Because $p(x)$ is a probability, it must be a number between 0 and 1, and because the probability distribution lists all possible x values, the sum of all the $p(x)$ values must equal 1. These properties of discrete probability distributions are summarized in the following box.

Properties of Discrete Probability Distributions

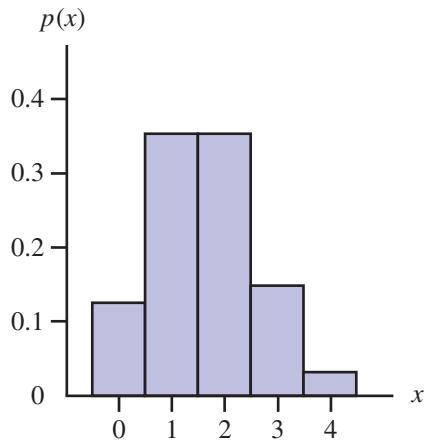
Properties of Discrete Probability Distributions

1. For every possible x value, $0 \leq p(x) \leq 1$.
2. $\sum_{\text{all } x \text{ values}} p(x) = 1$

A graphical representation of a discrete probability distribution is called a *probability histogram*. The probability histogram has a rectangle centered above each possible value of x , and the area of each rectangle is proportional to the probability of the corresponding value. Figure 7.3 displays the probability histogram for the probability distribution of Example 7.5.

FIGURE 7.3

Probability histogram for the distribution of Example 7.5.



In Example 7.5, the probability distribution was constructed by starting with a chance experiment and using probability rules. When it is not possible to construct a probability distribution in this way, an investigator can sometimes propose an approximate

probability distribution consistent with observed values of the random variable and prior knowledge. Probability distributions based on observed values must still be consistent with rules of probability:

1. $p(x) \geq 0$ for every x value.
2. $\sum_{\text{all } x \text{ values}} p(x) = 1$

This is illustrated in Examples 7.6 and 7.7.

Example 7.6 Paint Flaws

Understand the context ➤

In automobile manufacturing, one of the last steps in the process of assembling a new car is painting. Some minor blemishes in the paint surface are considered acceptable, but if there are too many, it becomes noticeable to a potential customer and the car must be repainted. Cars coming off the assembly line are carefully inspected and the number of minor blemishes in the paint surface is determined. We can define the random variable x to be the number of minor blemishes on a randomly selected car from a particular manufacturing plant. A large number of automobiles were evaluated, and a probability distribution consistent with these observations is

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)$	0.041	0.010	0.209	0.223	0.178	0.114	0.061	0.028	0.011	0.004	0.001

Interpret the results ➤

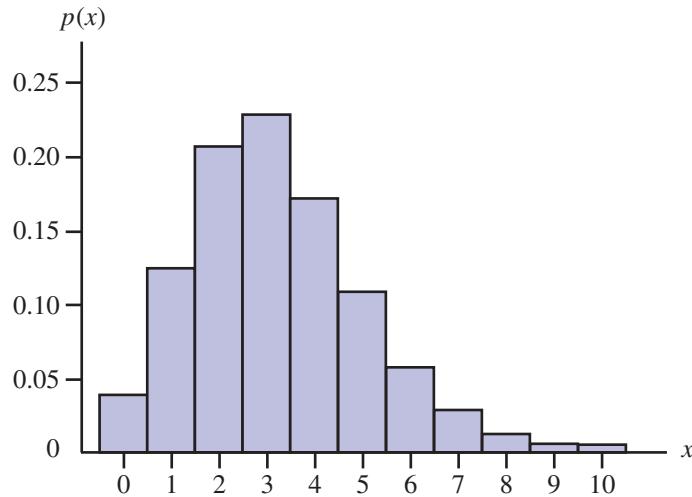
The corresponding probability histogram appears in Figure 7.4. The probabilities in this distribution reflect the car manufacturer's experience. For example, $p(3) = 0.223$ indicates that 22.3% of new automobiles have 3 minor paint blemishes. The probability that the number of minor paint blemishes is between 2 and 5 inclusive is

$$P(2 \leq x \leq 5) = p(2) + p(3) + p(4) + p(5) = 0.724$$

If car after car of this type were examined, in the long run, 72.4% would have 2, 3, 4, or 5 minor paint blemishes.

FIGURE 7.4

Probability histogram for the distribution of the number of minor paint blemishes on a randomly selected car.



Example 7.7 iPod Shuffles

Understand the context ➤

The paper “Does Your iPod Really Play Favorites?” (*American Statistician* [2009]: 263–268) investigated the shuffle feature of the iPod. The shuffle feature takes a group of songs, called a playlist, and plays them in a random order. Some have questioned the “randomness” of the shuffle, citing examples of situations where several songs from the same artist were played in close proximity to each other. One example appeared in a *Newsweek* article in 2005, where the author states that Steely Dan songs always seem to pop up in the first hour

of play. (For those young readers unfamiliar with Steely Dan, see steelydan.com.) Is this consistent with a “random” shuffle?

To investigate, suppose we create a playlist of 3000 songs that includes 50 songs by Steely Dan (this is the situation considered by the authors of the *American Statistician* paper). We could then carry out a simulation by creating a shuffle of 20 songs (about one hour of playing time) from the playlist and noting the number of songs by Steely Dan that were among the 20.

This could be done by thinking of the songs in the playlist as being numbered from 1 to 3000, with numbers 1 to 50 representing the Steely Dan songs. A random number generator could then be used to select 20 random numbers between 1 and 3000. We would then count how many times a number between 1 and 50 was included in this list. This would correspond to the number of Steely Dan songs in this particular shuffle of 20 songs. Repeating this process a large number of times would enable us to estimate the probabilities needed for the probability distribution of

$$x = \text{number of Steely Dan songs in a random shuffle consisting of 20 songs}$$

The probabilities in the accompanying probability distribution are from the *American Statistician* paper. Even though possible values of x are 0, 1, 2, ..., 20, the probability of x taking on a value of 4 or greater is very small, and so 4, 5, ..., 20 have been grouped into a single entry in the probability distribution table.

Probability Distribution of $x = \text{number of Steely Dan songs in a 20-song shuffle}$	
x	$p(x)$
0	0.7138
1	0.2435
2	0.0387
3	0.0038
4 or more	0.0002

Interpret the results ➤

Notice that $P(x \geq 1) = 1 - 0.7138 = 0.2862$. This means that about 28.6% of the time, at least one Steely Dan song would occur in a random shuffle of 20 songs. Given that there are only 50 Steely Dan songs in the playlist of 3000 songs, this result surprises many people!

We have seen examples in which the probability distribution of a discrete random variable has been given as a table or as a probability histogram. It is also possible to give a formula that allows calculation of the probability for each possible value of the random variable. Examples of this approach are given in Section 7.5.

EXERCISES 7.8 - 7.21

- 7.8** Define the random variable x to be the number of courses for which a randomly selected student at a certain university is registered. Suppose that the probability distribution of x is given in the following table:

x	1	2	3	4	5	6	7
$p(x)$	0.02	0.03	0.09	0.25	0.40	0.16	0.05

- a. What is $P(x = 4)$?
b. What is $P(x \leq 4)$?

- c. What is the probability that the selected student is taking at most five courses?
d. What is the probability that the selected student is taking at least five courses?
e. What is the probability that the selected student is taking more than five courses?

- 7.9** Using the probability distribution given in the previous exercise, calculate $P(3 \leq x \leq 6)$ and $P(3 < x < 6)$. Explain in words why these two probabilities are different.

- 7.10** Let y denote the number of broken eggs in a randomly selected carton of one dozen eggs. Suppose that the probability distribution of y is as follows:

y	0	1	2	3	4
$p(y)$	0.65	0.20	0.10	0.04	?

- a. Only y values of 0, 1, 2, 3, and 4 have positive probabilities. What is $p(4)$? (Hint: Consider the properties of a discrete probability distribution.)
- b. How would you interpret $p(1) = 0.20$?
- c. Calculate $P(y \leq 2)$, the probability that the carton contains at most two broken eggs, and interpret this probability.
- d. Calculate $P(y < 2)$, the probability that the carton contains *fewer than* two broken eggs. Why is this smaller than the probability in Part (c)?

- 7.11** Use the probability distribution given in the previous exercise to answer the following questions.

- a. What is the probability that the carton contains exactly 10 unbroken eggs?
- b. What is the probability that at least 10 eggs are unbroken?

- 7.12** Suppose that fund-raisers at a university call recent graduates to request donations for campus outreach programs. They report the following information for last year's graduates:

Size of donation	\$0	\$10	\$25	\$50
Proportion of calls	0.45	0.30	0.20	0.05

Three attempts were made to contact each graduate. A donation of \$0 was recorded both for those who were contacted but who declined to make a donation and for those who were not reached in three attempts. Consider the variable x = amount of donation for a person selected at random from the population of last year's graduates of this university.

- a. Write a few sentences describing what you think you might see if the value of x was observed for each of 1000 graduates.
- b. In the long run, what value of x would be observed most often?
- c. What is $P(x \geq 25)$?
- d. What is $P(x > 0)$?

- 7.13** Airlines sometimes overbook flights. Suppose that for a plane with 100 seats, an airline takes 110 reservations. Define the variable x as the number of people who actually show up for a sold-out flight. From past experience, the probability distribution of x is given in the following table:

x	95	96	97	98	99	100	101	102
$p(x)$	0.05	0.10	0.12	0.14	0.24	0.17	0.06	0.04
x	103	104	105	106	107	108	109	110
$p(x)$	0.03	0.02	0.01	0.005	0.005	0.005	0.0037	0.0013

- a. What is the probability that the airline can accommodate everyone who shows up for the flight?
- b. What is the probability that not all passengers can be accommodated?
- c. If you are trying to get a seat on such a flight and you are number 1 on the standby list, what is the probability that you will be able to take the flight? What if you are number 3?

- 7.14** Suppose that a computer manufacturer receives computer boards in lots of five. Two boards are selected from each lot for inspection. We can represent possible outcomes of the selection process by pairs. For example, the pair (1,2) represents the outcome where Boards 1 and 2 are selected for inspection.

- a. List the 10 different possible outcomes.
- b. Suppose that Boards 1 and 2 are the only defective boards in a lot of five. Two boards are to be chosen at random. Define x to be the number of defective boards observed among those inspected. Find the probability distribution of x .

- 7.15** Simulate the chance experiment described in the previous exercise using five slips of paper, with two marked *defective* and three marked *not defective*. Place the slips in a box, mix them well, and draw out two. Record the number of defective boards. Replace the slips and repeat until you have 50 observations of the value x . Construct a relative frequency distribution for the 50 observations, and compare this with the probability distribution obtained in the previous exercise.

- 7.16** Of all airline flight requests received by a certain discount ticket broker, 70% are for domestic travel (D) and 30% are for international flights (I). Define the random variable x to be the number of requests among the next three requests that are for domestic flights. Assuming independence of successive requests, determine the probability distribution of x . (Hint: See Example 7.5.)

- 7.17** Suppose that 20% of all homeowners in an earthquake-prone area of California are insured against earthquake damage. Four homeowners are selected at random. Suppose that x denotes the number among the four who have earthquake insurance.
- a. Find the probability distribution of x . (Hint: Let S represent a homeowner who has insurance and F one who does not. Then one possible outcome is SFSS, with probability $(0.2)(0.8)(0.2)(0.2)$ and associated x value of 3. There are 15 other outcomes.)
 - b. What is the most likely value of x ?
 - c. What is the probability that at least two of the four selected homeowners have earthquake insurance?

- 7.18** A box contains five slips of paper, marked \$1, \$1, \$1, \$10, and \$25. The winner of a contest selects two slips of paper at random and then gets the larger of the dollar amounts on the two slips. Define a random variable w by $w =$ amount awarded. Determine the probability distribution of w . (Hint: Think of the slips as numbered 1, 2, 3, 4, and 5, so that an outcome of the experiment consists of two of these numbers.)
- 7.19** Components coming off an assembly line are either free of defects (S, for success) or defective (F, for failure). Suppose that 70% of the components are defect-free. Components are independently selected and tested one by one. Let y denote the number of components that must be tested until a defect-free component is obtained.
- What is the smallest possible y value, and what outcome corresponds to this y value? What is the second smallest y value, and what is the corresponding outcome?
 - What is the set of all possible y values?
 - Calculate the probability of each of the five smallest y values. (You should see a pattern that leads to a simple formula for $p(y)$, the probability distribution of y .)
- 7.20** When applying for a building permit, a contractor is required by a county planning department to submit anywhere from one to five forms (depending on the nature of the project). Define the random variable y to be the number of forms required of the next

applicant. Suppose that the probability that y forms are required is known to be proportional to y with $p(y) = ky$ for $y = 1, \dots, 5$.

- What is the value of k ? (Hint: $\sum p(y) = 1$.)
- What is the probability that at most three forms are required?
- What is the probability that between two and four forms (inclusive) are required?
- Could $p(y) = y^2/50$ for $y = 1, 2, 3, 4, 5$ be the probability distribution of y ? Explain.

- 7.21** A library subscribes to two different weekly news magazines, each of which is supposed to arrive in Wednesday's mail. However, each one might actually arrive on Wednesday (W), Thursday (T), Friday (F), or Saturday (S). Suppose that the two magazines arrive independently of one another and that for each magazine

$$P(W) = 0.4, P(T) = 0.3, P(F) = 0.2, \text{ and } P(S) = 0.1$$

Define a random variable y by

$y =$ the number of days beyond Wednesday that it takes for both magazines to arrive.

For example, if the first magazine arrives on Friday and the second magazine arrives on Wednesday, then $y = 2$. If both magazines arrive on Thursday, $y = 1$. Determine the probability distribution of y . (Hint: Draw a tree diagram with two generations of branches, the first labeled with arrival days for Magazine 1 and the second for Magazine 2.)

SECTION 7.3 Probability Distributions for Continuous Random Variables

A continuous random variable is one that has possible values that form an entire interval on the number line. One example of a continuous random variable is the weight x (in pounds) of a full-term newborn baby. Suppose for the moment that weight is recorded only to the nearest pound. Then a reported weight of 7 pounds would be used for any weight greater than or equal to 6.5 pounds and less than 7.5 pounds.

The probability distribution of x can be pictured as a probability histogram with rectangles centered at 4, 5, and so on. The area of each rectangle represents the probability of the corresponding weight value, and the total area of all the rectangles is 1. The probability that a weight (to the nearest pound) is between two values, such as 6 and 8, is the sum of the areas of the corresponding rectangles. Figure 7.5(a) illustrates this.

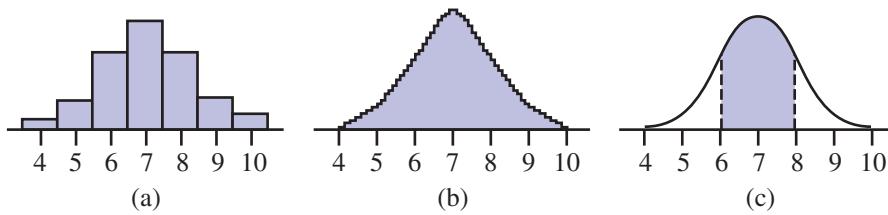
Now suppose that weight is measured to the nearest tenth of a pound. There are many more possible reported weight values than before, such as 5.0, 5.1, 5.7, 7.3, and 8.9. As shown in Figure 7.5(b), the rectangles in the probability histogram are much narrower, and this histogram has a much smoother appearance.

Now think about what would happen if weight were measured with greater accuracy. The probability histogram will look more like the smooth curve shown in Figure 7.5(c). This curve does not go below the horizontal measurement scale, and the total area under the curve is 1 (because this is true for all probability histograms). The probability that x falls in an interval such as $6 \leq x \leq 8$ is represented by the area under the curve and above that interval.

FIGURE 7.5

Probability distribution for birth weight:

- (a) weight measured to the nearest pound;
- (b) weight measured to the nearest tenth of a pound;
- (c) limiting curve as measurement accuracy increases; shaded area = $P(6 \leq \text{weight} \leq 8)$.



A **probability distribution for a continuous random variable x** is specified by a curve called a **density curve**. The function that defines this curve is denoted by $f(x)$ and it is called the **density function**.

The following are properties of all continuous probability distributions:

1. $f(x) \geq 0$ (the curve cannot dip below the horizontal axis).
2. The total area under the density curve is equal to 1.

The probability that x falls in any particular interval is equal to the area under the density curve and above the interval.

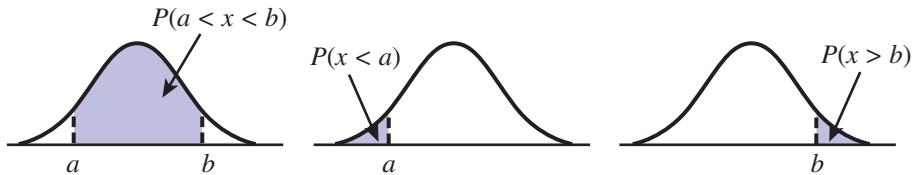
Many probability calculations for continuous random variables involve the following three types of events:

1. $a < x < b$, the event that the random variable x assumes a value between two given numbers, a and b
2. $x < a$, the event that the random variable x assumes a value less than a given number a
3. $x > b$, the event that the random variable x assumes a value greater than a given number b (this can also be written as $b < x$)

Figure 7.6 illustrates how the probabilities of these events are represented by areas under a density curve.

FIGURE 7.6

Probabilities as areas under a density curve.



Example 7.8 Application Processing Times

Understand the context ➤

Suppose x represents the continuous random variable

x = amount of time (in minutes) it takes to process a credit card application form

Suppose that x has a probability distribution with density function

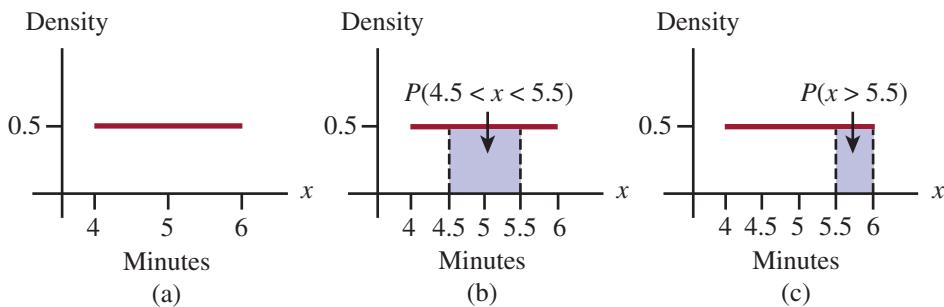
$$f(x) = \begin{cases} 0.5 & 4 < x < 6 \\ 0 & \text{otherwise} \end{cases}$$

The graph of the density curve $f(x)$ is shown in Figure 7.7(a). It is easy to use this density curve to calculate probabilities, because areas under this density curve are areas of rectangles. The area of a rectangle is calculated using the formula

$$\text{Area} = (\text{Base})(\text{Height})$$

FIGURE 7.7

The probability distribution for Example 7.8.



Do the work ➤

The curve has positive height, 0.5, only between $x = 4$ and $x = 6$. The total area under the curve is just the area of the rectangle with base extending from 4 to 6 and with height 0.5. We can verify that the total area under the density curve is equal to 1 by calculating

$$\text{Area} = (6 - 4)(0.5) = 1$$

When the density is constant over an interval (resulting in a horizontal density curve), the probability distribution is called a **uniform distribution**.

As illustrated in Figure 7.7(b), the probability that x is between 4.5 and 5.5 is

$$\begin{aligned} P(4.5 < x < 5.5) &= \text{Area of shaded rectangle} \\ &= (\text{Base})(\text{Height}) \\ &= (5.5 - 4.5)(0.5) \\ &= 0.5 \end{aligned}$$

Similarly (see Figure 7.7(c)), because in this context $x > 5.5$ is equivalent to $5.5 < x \leq 6$, we have

$$P(x > 5.5) = (6 - 5.5)(0.5) = 0.25$$

Interpret the results ➤

Based on this probability distribution, in the long run, 25% of all forms will have processing times that are greater than 5.5 minutes.

The probability that a *discrete* random variable x takes a value in an interval between two endpoints a and b can depend on whether either endpoint is included in the interval. For example, suppose that x is the number of major defects on a new automobile. Then

$$P(3 \leq x \leq 7) = p(3) + p(4) + p(5) + p(6) + p(7)$$

and

$$P(3 < x < 7) = p(4) + p(5) + p(6)$$

However, if x is a *continuous* random variable, such as task completion time, then

$$P(3 \leq x \leq 7) = P(3 < x < 7)$$

because the area under a density curve and above a single value such as 3 or 7 is 0. Geometrically, we can think of finding the area above a single point as finding the area of a rectangle with width = 0. This means that for a continuous random variable, the area above an interval of values does not depend on whether either endpoint is included.

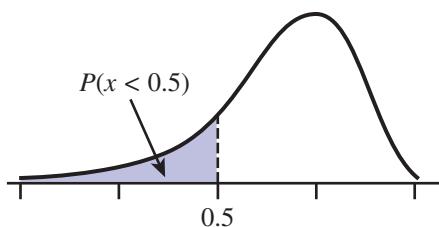
If x is a continuous random variable, then for any two numbers a and b with $a < b$,

$$P(a \leq x \leq b) = P(a < x \leq b) = P(a \leq x < b) = P(a < x < b)$$

Probabilities for continuous random variables are often calculated using cumulative areas. A cumulative area is all of the area under the density curve to the left of a particular value. Figure 7.8 illustrates the cumulative area to the left of 0.5, which represents $P(x < 0.5)$. The probability that x is in any particular interval, $P(a < x < b)$, can be expressed as the difference between two cumulative areas.

FIGURE 7.8

A cumulative area under a density curve.



The probability that a continuous random variable x lies between a lower limit a and an upper limit b is

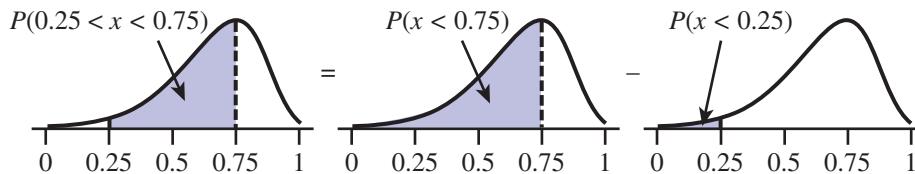
$$\begin{aligned} P(a < x < b) &= (\text{cumulative area to the left of } b) - (\text{cumulative area to the left of } a) \\ &= P(x < b) - P(x < a) \end{aligned}$$

For continuous random variables, the area above an interval of values does not depend on whether either endpoint is included, so this is also the probability of $P(a \leq x \leq b)$, $P(a < x \leq b)$, and $P(a \leq x < b)$.

This property is illustrated in Figure 7.9 for the case of $a = 0.25$ and $b = 0.75$. We will use this result often in Section 7.6 when we calculate probabilities for random variables that have a normal distribution.

FIGURE 7.9

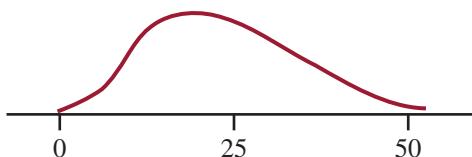
Calculation of $P(a < x < b)$ using cumulative areas.



For some continuous distributions, cumulative areas can be calculated using methods from the branch of mathematics called calculus. However, because we are not assuming knowledge of calculus, we will rely on technology or on tables that have been constructed for the commonly encountered continuous probability distributions. Many graphing calculators and statistical software packages will compute areas for widely used continuous probability distributions.

EXERCISES 7.22 - 7.32

- 7.22** Let x denote the lifetime (in thousands of hours) of a certain type of fan used in diesel engines. The density curve of x is as pictured:



Shade the area under the curve corresponding to each of the following probabilities (draw a new curve for each part). (Hint: See Example 7.8.)

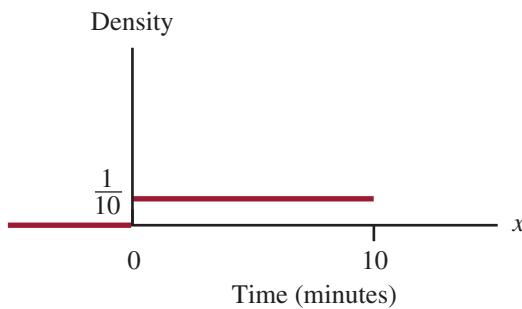
- $P(10 < x < 25)$
- $P(10 \leq x \leq 25)$
- $P(x < 30)$

- 7.23** Using the density curve for fan lifetime given in the previous exercise, shade the area under the curve

corresponding to each of the following probabilities (draw a new curve for each part).

- The probability that the lifetime is at least 25,000 hours
- The probability that the lifetime exceeds 25,000 hours

- 7.24** A particular professor never dismisses class early. Let x denote the amount of time past the class end time (in minutes) that passes before the professor dismisses class. Suppose that x has a uniform distribution on the interval from 0 to 10 minutes. The density curve is shown in the following figure:



- What is the probability that at most 5 minutes pass before dismissal? (Hint: See Example 7.8.)
- What is the probability that between 3 and 5 minutes pass before dismissal?

- 7.25** Refer to the probability distribution given in the previous exercise. Put the following probabilities in order, from smallest to largest:

$$P(2 < x < 3), P(2 \leq x \leq 3), P(x < 2), P(x > 7)$$

Explain your reasoning.

- 7.26** The article “[Probabilistic Risk Assessment of Infrastructure Networks Subjected to Hurricanes](#)” (*12th International Conference on Applications of Statistics and Probability in Civil Engineering*, 2015) suggests a uniform distribution as a model for the actual landfall position of the eye of a hurricane. Consider the random variable $x = \text{distance of actual landfall from predicted landfall}$. Suppose that a uniform distribution on the interval from 0 km to 400 km is a reasonable model for x .

- Draw the density curve for x .
- What is the height of the density curve?

- 7.27** Use the density curve of $x = \text{distance of actual landfall from predicted landfall}$ from the previous exercise to answer the following questions.

- What is the probability that x is at most 100?

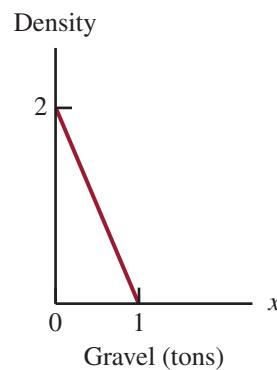
- What is the probability that x is between 200 and 300?

- What is the probability that x is between 50 and 150?
- Explain why the probabilities calculated in Parts (b) and (c) are equal.

- 7.28** Let x denote the amount of gravel sold (in tons) during a randomly selected week at a particular sales facility. Suppose that the density curve has height $f(x)$ above the value x , where

$$f(x) = \begin{cases} 2(1-x) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The density curve (the graph of $f(x)$) is shown in the following figure:



Use the fact that the

$$\text{Area of a triangle} = \frac{1}{2}(\text{Base})(\text{Height})$$

to calculate each of the following probabilities. (Hint: Drawing a picture and shading the appropriate area will help.)

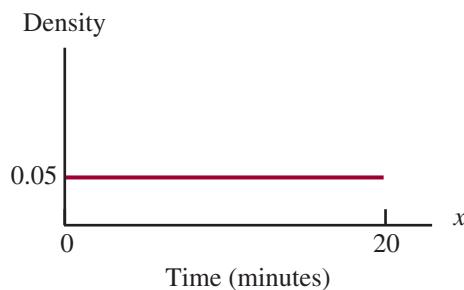
- $P\left(x < \frac{1}{2}\right)$
- $P\left(x \leq \frac{1}{2}\right)$
- $P\left(x < \frac{1}{4}\right)$
- $P\left(\frac{1}{4} < x < \frac{1}{2}\right)$ (Hint: Use the results of Parts (a)–(c).)

- 7.29** Use the density curve for $x = \text{amount of gravel sold}$ given in the previous exercise to determine each of the following probabilities.

- The probability that gravel sold exceeds $\frac{1}{2}$ ton
- The probability that gravel sold is at least $\frac{1}{4}$ ton

- 7.30** Let x be the amount of time (in minutes) that a particular San Francisco commuter must wait for

a train. Suppose that the density curve is as pictured (a uniform distribution):



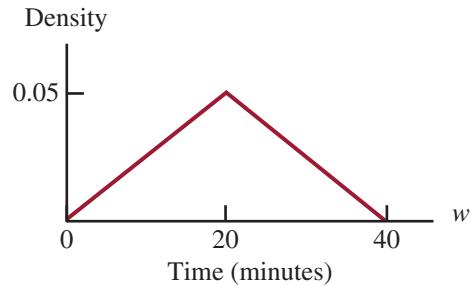
- What is the probability that x is less than 10 minutes? more than 15 minutes?
- What is the probability that x is between 7 and 12 minutes?
- Find the value c for which $P(x < c) = 0.9$.

7.31 Referring to the previous exercise, let x and y be waiting times on two independently selected days. Define a new random variable w by

$$w = x + y$$

the sum of the two waiting times. The set of possible values for w is the interval from 0 to 40 (because both x and y can range from 0 to 20). It can be shown that

the density curve of w is as pictured (this curve is called a triangular distribution, for obvious reasons!):



- Verify that the total area under the density curve is equal to 1.
(Hint: Area of a triangle = $\frac{1}{2}(\text{Base})(\text{Height})$.)
- What is the probability that w is less than 20?
- What is the probability that w is less than 10?
- What is the probability that w is greater than 30?

7.32 The density curve for the random variable w (the sum of two wait times) is given in the previous exercise. What is the probability that w is between 10 and 30? (Hint: It might be easier first to find the probability that w is not between 10 and 30.)

SECTION 7.4 Mean and Standard Deviation of a Random Variable

We study a random variable x , such as the number of insurance claims made by a homeowner (a discrete variable) or the birth weight of a baby (a continuous variable), to learn something about how its values are distributed along the measurement scale. The sample mean \bar{x} and sample standard deviation s summarize center and spread for the values in a sample. Similarly, the mean value and standard deviation of a random variable describe where the variable's probability distribution is centered and the extent to which it spreads out about the center.

The **mean value of a random variable x** , denoted by μ_x , describes where the probability distribution of x is centered.

The **standard deviation of a random variable x** , denoted by σ_x , describes variability in the probability distribution. When the value of σ_x is small, observed values of x will tend to be close to the mean value (little variability). When the value of σ_x is large, there will be more variability in observed x values.

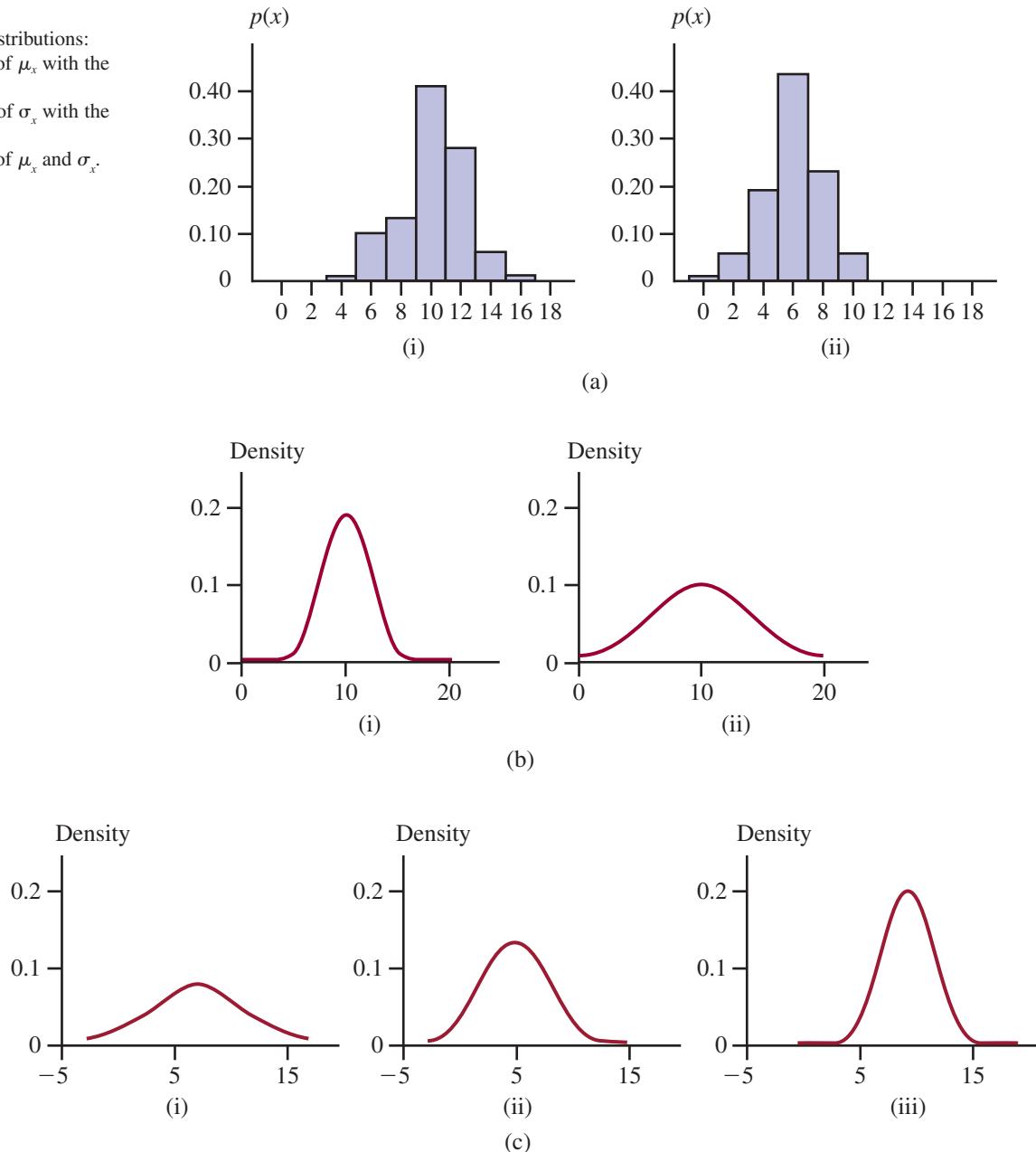
Consider the probability distributions in Figure 7.10.

- Figure 7.10(a) shows two discrete probability distributions with the same standard deviation (variability) but different means (center). One distribution has a mean of $\mu_x = 6$ and the other has $\mu_x = 10$. Which has the mean of 6 and which has the mean of 10?
- Figure 7.10(b) shows two continuous probability distributions that have the same mean but different standard deviations. Which distribution—(i) or (ii)—has the greater standard deviation?

3. Figure 7.10(c) shows three continuous distributions with different means and standard deviations. Which of the three distributions has the greatest mean? Which has a mean of about 5? Which distribution has the smallest standard deviation?

FIGURE 7.10

Some probability distributions:
 (a) different values of μ_x with the same value of σ_x ;
 (b) different values of σ_x with the same value of μ_x ;
 (c) different values of μ_x and σ_x .



The answers to these questions are the following: Figure 7.10(a)(ii) has a mean of 6, and Figure 7.10(a)(i) has a mean of 10; Figure 7.10(b)(ii) has the greater standard deviation; Figure 7.10(c)(iii) has the greatest mean, Figure 7.10(c)(ii) has a mean of about 5, and Figure 7.10(c)(iii) has the smallest standard deviation.

It is customary to use the terms *mean of the random variable x* and *mean of the probability distribution of x* interchangeably. Similarly, the *standard deviation of the random variable x* and the *standard deviation of the probability distribution of x* refer to the same thing. Although the mean and standard deviation are calculated differently for discrete and continuous random variables, the interpretation is the same in both cases.

Mean Value of a Discrete Random Variable

Consider a chance experiment consisting of randomly selecting an automobile licensed in Pennsylvania. Consider the discrete random variable x defined as

x = the number of low-beam headlights on the selected car that need adjustment

Possible x values are 0, 1, and 2, and the probability distribution of x might be as follows:

x value	0	1	2
Probability	0.5	0.3	0.2

The corresponding probability histogram is given in Figure 7.11. In a sample of 100 cars licensed in Pennsylvania, the sample relative frequencies might differ somewhat from the given probabilities (which are the limiting relative frequencies). For example, we might see:

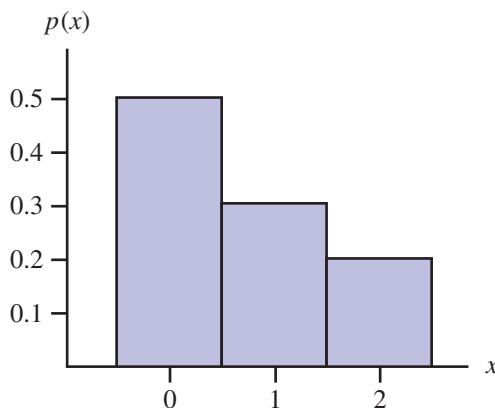
x value	0	1	2
Frequency	46	33	21

The sample mean value of x for these 100 observations is the sum of 46 zeros, 33 ones, and 21 twos, all divided by 100:

$$\begin{aligned}\bar{x} &= \frac{(46)(0) + (33)(1) + (21)(2)}{100} \\ &= \left(\frac{46}{100}\right)(0) + \left(\frac{33}{100}\right)(1) + \left(\frac{21}{100}\right)(2) \\ &= (\text{rel. freq. of } 0)(0) + (\text{rel. freq. of } 1)(1) + (\text{rel. freq. of } 2)(2) \\ &= 0.75\end{aligned}$$

FIGURE 7.11

Probability histogram for the distribution of the number of headlights needing adjustment.



As the sample size increases, each relative frequency will approach the corresponding probability. In a very long sequence of observations, the value of \bar{x} will approach

$$\begin{aligned}\text{mean value of } x &= P(x = 0)(0) + P(x = 1)(1) + P(x = 2)(2) \\ &= (0.5)(0) + (0.3)(1) + (0.2)(2) \\ &= 0.70\end{aligned}$$

Notice that the expression for \bar{x} is a weighted average of possible x values. The weight of each value is the observed relative frequency. Similarly, the mean value of the random variable x is a weighted average, but with weights that are the probabilities from the probability distribution.

DEFINITION

Mean value of a discrete random variable: The mean value of a discrete random variable x , denoted by μ_x , is calculated by first multiplying each possible x value by the probability of observing that value and then adding the resulting quantities:

$$\mu_x = \sum_{\text{all possible } x \text{ values}} x \cdot p(x)$$

The term **expected value** is sometimes used in place of mean value, and $E(x)$ is another way to denote μ_x .

Example 7.9 Exam Attempts

Understand the context ➤

Suppose that students taking an online course must take an exam and they are allowed to retake the exam up to three additional times in order to improve their scores. Consider the chance experiment of selecting a student at random. Suppose that the random variable x defined as

x = the number of times the selected student took the exam

Suppose the probability distribution of x is as follows:

x	1	2	3	4
$p(x)$	0.10	0.20	0.30	0.40

Do the work ➤

Then x has mean value

$$\begin{aligned}
 \mu_x &= \sum_{x=1,2,3,4} x \cdot p(x) \\
 &= (1)p(1) + (2)p(2) + (3)p(3) + (4)p(4) \\
 &= (1)(0.10) + (2)(0.20) + (3)(0.30) + (4)(0.40) \\
 &= 0.10 + 0.40 + 0.90 + 1.60 \\
 &= 3.00
 \end{aligned}$$

Interpret the results ➤

This means that for students in the online course, the mean number of times the exam is taken is 3.

Example 7.10 Apgar Scores

Understand the context ➤

At 1 minute after birth and again at 5 minutes, each newborn child is given a numerical rating called an Apgar score. Possible values of this score are 0, 1, 2, . . . , 9, 10. A child's score is determined by five factors: muscle tone, skin color, respiratory effort, strength of heartbeat, and reflex, with a high score indicating a healthy infant.

Consider the random variable x defined as

x = the Apgar score (at 1 minute) of a randomly selected newborn infant at a particular hospital

Suppose that x has the following probability distribution:

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)$	0.002	0.001	0.002	0.005	0.02	0.04	0.17	0.38	0.25	0.12	0.01

Do the work ➤

The mean value of x is

$$\begin{aligned}\mu_x &= (0)p(0) + (1)p(1) + \cdots + (9)p(9) + (10)p(10) \\ &= (0)(0.002) + (1)(0.001) + \cdots + (9)(0.12) + (10)(0.01) \\ &= 7.16\end{aligned}$$

Interpret the results ➤

The mean Apgar score for a *sample* of newborn children born at this hospital might be $\bar{x} = 7.05$, $\bar{x} = 8.30$, or any one of a number of other possible values between 0 and 10.

However, as child after child is born and rated, the mean score will approach the value 7.16. This value can be interpreted as the mean Apgar score for the population of all babies born at this hospital.

Standard Deviation of a Discrete Random Variable

The mean value μ_x provides only a partial summary of a probability distribution. Two different distributions can have the same value of μ_x , but there may be more variability in a long sequence of observations from one distribution than in a long sequence of observations from the other distribution.

Example 7.11 Glass Panels

Understand the context ➤

Flat screen TVs require high quality glass with very few flaws. Suppose a television manufacturer receives glass panels from two different suppliers. Let x denote the number of flaws in a randomly selected glass panel from the first supplier and y denote the number of flaws in a randomly selected glass panel from the second supplier. Suppose that the probability distributions for x and y are as follows:

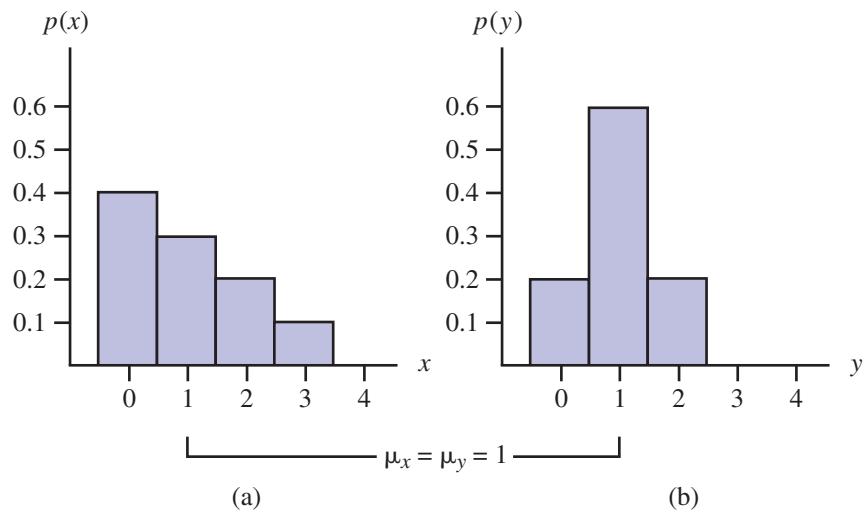
x	0	1	2	3	4	y	0	1	2	3	4
$p(x)$	0.4	0.3	0.2	0.1	0	$p(y)$	0.2	0.6	0.2	0	0

Probability histograms for x and y are given in Figure 7.12.

It is easy to verify that the mean values of both x and y are 1, so for either supplier the long-run average number of flaws per panel is 1. However, the two probability histograms show that the probability distribution for the second supplier is concentrated closer to the mean value than is the case for the first supplier.

FIGURE 7.12

Probability distribution for the number of flaws in a glass panel in Example 7.11:
 (a) Supplier 1;
 (b) Supplier 2.



Interpret the results ➤

The greater spread of the first distribution implies that there will be more variability in a long sequence of observed x values than in a long sequence of observed y values. For example, the y sequence will contain no 3's, but in the long run, 10% of the observed x values will be 3.

Calculation of the variance and standard deviation of a random variable x involves squared deviations from the mean. A value far from the mean results in a large squared deviation. However, such a value does not contribute substantially to variability in x if the probability associated with that value is small. For example, if $\mu_x = 1$ and $x = 25$ is a possible value, then the squared deviation is $(25 - 1)^2 = 576$. However, if $P(x = 25) = 0.000001$, the value

25 will hardly ever be observed, so it won't contribute much to variability in a long sequence of observations. This is why each squared deviation is multiplied by the probability associated with the value to obtain a measure of variability.

DEFINITIONS

Variance of a discrete random variable: The variance of a discrete random variable x , denoted by σ_x^2 , is calculated by

1. subtracting the mean from each possible x value to obtain the deviations
2. squaring each deviation
3. multiplying each squared deviation by the probability of the corresponding x value
4. adding these quantities

$$\sigma_x^2 = \sum_{\text{all possible } x \text{ values}} (x - \mu_x)^2 p(x)$$

Standard deviation of a discrete random variable: The standard deviation of a discrete random variable x , denoted by σ_x , is the square root of the variance.

Example 7.12 Glass Panels Revisited

For x = number of flaws in a glass panel from the first supplier in Example 7.11,

$$\begin{aligned}\sigma_x^2 &= (0 - 1)^2 p(0) + (1 - 1)^2 p(1) + (2 - 1)^2 p(2) + (3 - 1)^2 p(3) \\ &= (1)(0.4) + (0)(0.3) + (1)(0.2) + (4)(0.1) \\ &= 1.0\end{aligned}$$

The standard deviation of x is then $\sigma_x = \sqrt{\sigma_x^2} = \sqrt{1.0} = 1.0$.

For y = the number of flaws in a glass panel from the second supplier,

$$\sigma_y^2 = (0 - 1)^2(0.2) + (1 - 1)^2(0.6) + (2 - 1)^2(0.2) = 0.4$$

Then $\sigma_y = \sqrt{0.4} = 0.632$.

The fact that σ_x is greater than σ_y confirms the impression about the variability in x and y conveyed by the probability distributions shown in Figure 7.12.

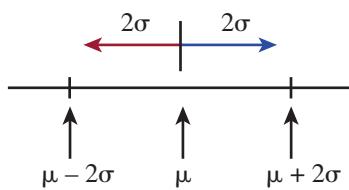


FIGURE 7.13

Values within 2 standard deviations of the mean.

Example 7.13 More on Apgar Scores

Reconsider the distribution of Apgar scores for children born at a certain hospital, introduced in Example 7.10. What is the probability that a randomly selected child's score will be within 2 standard deviations of the mean score? As Figure 7.13 shows, values of x within 2 standard deviations of the mean are those for which

$$\mu - 2\sigma < x < \mu + 2\sigma$$

From Example 7.10 we already know $\mu_x = 7.16$. The variance is

$$\begin{aligned}\sigma^2 &= \sum(x - \mu)^2 p(x) = \sum(x - 7.16)^2 p(x) \\ &= (0 - 7.16)^2(0.002) + (1 - 7.16)^2(0.001) + \cdots + (10 - 7.16)^2(0.01) \\ &= 1.5684\end{aligned}$$

and the standard deviation is

$$\sigma = \sqrt{1.5684} = 1.25$$

This gives (using the probabilities given in Example 7.10)

$$\begin{aligned}
 P(\mu - 2\sigma < x < \mu + 2\sigma) &= P(7.16 - 2.50 < x < 7.16 + 2.50) \\
 &= P(4.66 < x < 9.66) \\
 &= p(5) + \dots + p(9) \\
 &= 0.96
 \end{aligned}$$

Mean and Standard Deviation When x Is Continuous

For continuous probability distributions, μ_x and σ_x are defined and calculated using methods from calculus. In this text, we won't worry about calculating these values. What is important is knowing that μ_x and σ_x are interpreted in the same way as they are in the discrete case. The mean value μ_x locates the center of the continuous distribution. It gives the approximate long-run average of many observed x values. The standard deviation σ_x measures the extent that the continuous distribution (density curve) spreads out around μ_x . It provides information about the variability that can be expected in a long sequence of observed x values.

Example 7.14 Body Mass Index

Understand the context ➤

Body mass index (BMI) is a continuous variable calculated from height and weight that is measured in kilograms/meter². The authors of the paper “Concordance of Self-Report and Measured Height and Weight of College Students” (*Journal of Nutrition Education and Behavior* [2015]: 94–98) collected data on BMI from a large number of students at eight different colleges. For these students, the authors reported that the continuous random variables

x = BMI based on self-reported height and weight

and

y = BMI based on measured height and weight

have the following means and standard deviations:

$$\begin{array}{ll}
 \mu_x = 24.2 & \mu_y = 24.5 \\
 \sigma_x = 2.6 & \sigma_y = 3.9
 \end{array}$$

Interpret the results ➤

This suggests that if you were to observe self-reported and measured BMI for a large number of these college students, you would not see much difference in the means of the observations for these variables. But because the standard deviation of y = measured BMI is larger than the standard deviation of x = self-reported BMI, there would be more variability in the values of the measured BMI than the BMI that was based on self-reported height and weight.

Mean and Variance of Linear Functions and Linear Combinations

We have seen how the mean and standard deviation of one or more random variables provide useful information about the variables' long-run behavior. Sometimes we are also interested in the behavior of some function of random variables.

For example, consider the chance experiment in which a customer of a propane gas company is randomly selected. Suppose that the mean and standard deviation of the random variable

x = number of gallons required to fill a customer's propane tank

are known to be 318 gallons and 42 gallons, respectively. The company is considering two different pricing models:

Model 1: \$3 per gallon

Model 2: service charge of \$50 + \$2.80 per gallon

The company is interested in the variable

y = amount billed

For each of the two pricing models, y can be expressed as a function of the random variable x :

Model 1: $y_{model\ 1} = 3x$

Model 2: $y_{model\ 2} = 50 + 2.8x$

Both of these equations are examples of a linear function of the random variable x . The mean and standard deviation of a linear function of x can be calculated from the mean and standard deviation of x , as described in the following box.

The Mean, Variance, and Standard Deviation of a Linear Function

If x is a random variable with mean μ_x and variance σ_x^2 and a and b represent numbers, the random variable y defined by

$$y = a + bx$$

is called a **linear function of the random variable x** .

The mean of $y = a + bx$ is

$$\mu_y = \mu_{a+bx} = a + b\mu_x$$

The variance of y is

$$\sigma_y^2 = \sigma_{a+bx}^2 = b^2\sigma_x^2$$

It follows that the standard deviation of y is

$$\sigma_y = \sigma_{a+bx} = |b|\sigma_x$$

We can use the results in the preceding box to calculate the mean and standard deviation of the billing amount variable for the propane gas example, as follows:

For Model 1:

$$\mu_{model\ 1} = \mu_{3x} = 3\mu_x = 3(318) = 954$$

$$\sigma_{model\ 1}^2 = \sigma_{3x}^2 = 3^2\sigma_x^2 = 9(42)^2 = 15,876$$

$$\sigma_{model\ 1} = \sqrt{15,876} = 126, \text{ which is equal to } 3(42)$$

For Model 2:

$$\mu_{model\ 2} = \mu_{50+2.8x} = 50 + 2.8\mu_x = 50 + 2.8(318) = 940.40$$

$$\sigma_{model\ 2}^2 = \sigma_{50+2.8x}^2 = 2.8^2\sigma_x^2 = (2.8)^2(42)^2 = 13,829.76$$

$$\sigma_{model\ 2} = \sqrt{13,829.76} = 117.60, \text{ which is equal to } 2.8(42)$$

The mean billing amount for Model 1 is a bit higher than for Model 2, as is the variability in billing amounts. Billing amounts based on Model 2 will be slightly more consistent from bill to bill.

Linear Combinations

Now let's consider a different type of problem. Suppose that you have three tasks that you plan to complete on the way home from school: stop at the library to return an overdue book for which you must pay a fine, deposit your most recent paycheck at the

bank, and stop by the office supply store to buy paper for your printer. Define the following variables:

- x_1 = time required to return book and pay fine
- x_2 = time required to deposit paycheck
- x_3 = time required to buy printer paper

We can then define a new variable, y , to represent the total amount of time to complete these tasks:

$$y = x_1 + x_2 + x_3$$

Defined in this way, y is an example of a linear combination of random variables.

If x_1, x_2, \dots, x_n are random variables and a_1, a_2, \dots, a_n represent numbers, the random variable y defined as

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n$$

is a **linear combination of random variables**.

For example, $y = 10x_1 - 5x_2 + 8x_3$ is a linear combination of x_1, x_2 , and x_3 with $a_1 = 10$, $a_2 = -5$ and $a_3 = 8$.

It is easy to calculate the mean of a linear combination of random variables if the individual means are known. The variance and standard deviation of a linear combination of random variables are also easily calculated *if the random variables are independent*. Two random variables x_i and x_j are independent if any event defined solely by x_i is independent of any event defined solely by x_j . When the x_i 's are not independent, calculation of the variance and standard deviation of a linear combination of random variables is more complicated, and this case is not considered here.

Mean, Variance, and Standard Deviation for Linear Combinations

Suppose x_1, x_2, \dots, x_n are random variables with means $\mu_1, \mu_2, \dots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$, respectively. Define the random variable y as

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n$$

Then

$$1. \mu_y = \mu_{a_1x_1 + a_2x_2 + \dots + a_nx_n} = a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n$$

This result is true regardless of whether the x_i 's are independent.

$$2. \text{When } x_1, x_2, \dots, x_n \text{ are independent random variables,}$$

$$\sigma_y^2 = \sigma_{a_1x_1 + a_2x_2 + \dots + a_nx_n}^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2$$

$$\sigma_y = \sigma_{a_1x_1 + a_2x_2 + \dots + a_nx_n} = \sqrt{a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2}$$

These formulas are appropriate only when the x_i 's are independent.

Examples 7.15–7.17 illustrate the use of these rules.

Example 7.15 Freeway Traffic

Understand the context ▶

Three different roads feed into a particular freeway entrance. Suppose that during a fixed time period, the number of cars coming from each road onto the freeway is a random variable with mean values as follows:

Road	1	2	3
Mean	800	1,000	600

With x_i representing the number of cars entering from road i , we can define the total number of cars entering the freeway y as

$$y = x_1 + x_2 + x_3.$$

Do the work ➤ The mean value of y is

$$\begin{aligned}\mu_y &= \mu_{x_1+x_2+x_3} \\ &= \mu_{x_1} + \mu_{x_2} + \mu_{x_3} \\ &= 800 + 1000 + 600 \\ &= 2400\end{aligned}$$

Example 7.16 Combining Exam Subscores

Understand the context ➤

A nationwide standardized exam consists of a multiple-choice section and a free response section. Suppose that for each section, the mean and standard deviation are reported to be as shown in the following table:

	Mean	Standard Deviation
Multiple Choice	38	6
Free Response	30	7

Let's define x_1 as the multiple-choice score and x_2 as the free-response score for a student selected at random from those taking this exam. We are also interested in the variable y = overall exam score.

Suppose that the free-response score is given twice the weight of the multiple-choice score in determining the overall exam score. Then the overall score is calculated as

$$y = x_1 + 2x_2$$

Do the work ➤ What are the mean and standard deviation of y ?

Because $y = x_1 + 2x_2$ is a linear combination of x_1 and x_2 , the mean of y is

$$\begin{aligned}\mu_y &= \mu_{x_1+2x_2} \\ &= \mu_{x_1} + 2\mu_{x_2} \\ &= 38 + 2(30) \\ &= 98\end{aligned}$$

Interpret the results ➤

This means that the average overall score for students taking this exam is 98. What about the variance and standard deviation of y ? To use the formulas in the preceding box, x_1 and x_2 must be independent. It is unlikely that the value of x_1 (a student's multiple-choice score) would be unrelated to the value of x_2 (the same student's free-response score), because it seems likely that students who score well on one section of the exam will also tend to score well on the other section. Therefore, it would not be appropriate to use the formulas given previously to calculate the variance and standard deviation.

Example 7.17 Baggage Weights

Understand the context ➤

A commuter airline flies small planes between San Luis Obispo, California, and San Francisco. For small planes, baggage weight is a concern, especially on foggy mornings, because the weight of the plane has an effect on how quickly the plane can ascend. Suppose that it is known that the random variable x = weight (in pounds) of baggage checked by a randomly selected passenger has a mean of 42 and standard deviation of 16.

Consider a flight with 10 passengers, all traveling alone. If we use x_i to denote the baggage weight for passenger i (for i ranging from 1 to 10), the total weight of checked baggage, y , is then

$$y = x_1 + x_2 + \cdots + x_{10}$$

Do the work ➤ Notice that y is a linear combination of the x_i 's. The mean value of y is

$$\begin{aligned}\mu_y &= \mu_{x_1} + \mu_{x_2} + \cdots + \mu_{x_{10}} \\ &= 42 + 42 + \cdots + 42 \\ &= 420\end{aligned}$$

Since the 10 passengers are all traveling alone, it is reasonable to think that the 10 baggage weights are unrelated and that the x_i 's are independent. (This might not be a reasonable assumption if the 10 passengers were not traveling alone.) Then the variance of y is

$$\begin{aligned}\sigma_y^2 &= \sigma_{x_1}^2 + \sigma_{x_2}^2 + \cdots + \sigma_{x_{10}}^2 \\ &= 16^2 + 16^2 + \cdots + 16^2 \\ &= 2650\end{aligned}$$

and the standard deviation of y is

$$\sigma_y = \sqrt{2650} = 50.6$$

Interpret the results ➤ This means that the mean total weight of checked baggage is 420 pounds for flights with 10 passengers traveling alone. The large standard deviation of 50.6 pounds indicates that there will be a lot of variability in total checked baggage weight from flight to flight.

One Last Note on Linear Functions and Linear Combinations

In Example 7.17, the random variable of interest was x = weight of baggage checked by a randomly selected airline passenger. However, when we considered a flight with multiple passengers, we added subscripts to create variables like

$$x_1 = \text{baggage weight for passenger 1}$$

and

$$x_2 = \text{baggage weight for passenger 2}$$

It is important to note that the sum of the baggage weights for two different passengers does not result in the same value as doubling the baggage weight of a single passenger. *The linear combination $x_1 + x_2$ is different from the linear function $2x$.*

Although the mean of $x_1 + x_2$ and the mean of $2x$ are the same, the variances and standard deviations are different. For example, for the baggage weight distribution described in Example 7.17 (mean of 42 and standard deviation of 16), the mean of $2x$ is

$$\mu_{2x} = 2\mu_x = 2(42) = 84$$

and the mean of $x_1 + x_2$ is

$$\mu_{x_1 + x_2} = \mu_{x_1} + \mu_{x_2} = 42 + 42 = 84$$

However, the variance of $2x$ is

$$\sigma_{2x}^2 = (2)^2 \sigma_x^2 = 4(16^2) = 1024$$

and the variance of $x_1 + x_2$ is

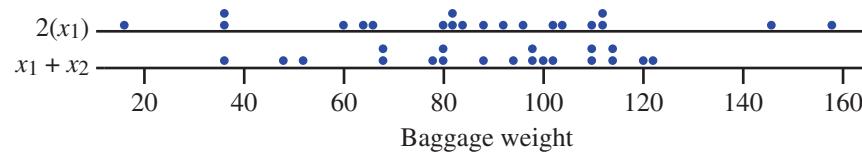
$$\sigma_{x_1 + x_2}^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 = 16^2 + 16^2 = 256 + 256 = 512$$

This means that observations of $2x$ for randomly selected passengers will show more variability than observations of $x_1 + x_2$ for two randomly selected passengers. This may seem counterintuitive, but it is true! For example, each row of the first two columns of the accompanying table show simulated baggage weights for two passengers (an x_1 and an x_2 value) that were selected at random from a probability distribution with mean 42 and standard deviation 16. The third column of the table contains the value of $2x_1$ (two times the passenger 1 baggage weight) and the last column contains the values of $x_1 + x_2$ (the sum of the passenger 1 baggage weight and the passenger 2 baggage weight).

x_1	x_2	$2x_1$	$x_1 + x_2$	x_1	x_2	$2x_1$	$x_1 + x_2$
56	41	112	97	44	23	88	67
30	5	60	35	48	54	96	102
32	15	64	47	52	69	104	121
55	45	110	100	56	63	112	119
40	54	80	94	79	35	158	114
41	68	82	109	18	34	36	52
42	35	84	77	8	71	16	79
33	47	66	80	51	58	102	109
46	52	92	98	18	49	36	67
73	40	146	113	41	46	82	87

Figure 7.14 shows dotplots of the 20 values of $2x_1$ and the 20 values of $x_1 + x_2$. You can see that there is less variability in the $x_1 + x_2$ values than in the values of $2x_1$.

FIGURE 7.14
Dotplots of $2x_1$ and $x_1 + x_2$.



EXERCISES 7.33 - 7.50

- 7.33** Consider selecting a household in rural Thailand at random. Define the random variable x to be

$$x = \text{number of individuals living in the selected household}$$

Based on information in an article that appeared in the *Journal of Applied Probability* (2011: 173–188), the probability distribution of x is as follows:

x	1	2	3	4	5
$p(x)$	0.140	0.175	0.220	0.260	0.155
x	6	7	8	9	10
$p(x)$	0.025	0.015	0.005	0.004	0.001

Calculate the mean value of the random variable x . (Hint: See Example 7.9.)

- 7.34** Suppose the probability distribution of x , the number of defective tires on a randomly selected car checked at an inspection station, is given in the following table:

x	0	1	2	3	4
$p(x)$	0.54	0.16	0.06	0.04	0.20

- Calculate the mean value of x .
- What is the probability that x exceeds its mean value?

- 7.35** Consider the following probability distribution for $y = \text{the number of broken eggs in a carton}$:

y	0	1	2	3	4
$p(y)$	0.65	0.20	0.10	0.04	0.01

- Calculate and interpret μ_y .
- In the long run, for what percentage of cartons is the number of broken eggs less than μ_y ? Does this surprise you?
- Why doesn't $\mu_y = (0 + 1 + 2 + 3 + 4)/5 = 2.0$? Explain.

- 7.36** Referring to the previous exercise, use the result of Part (a) along with the fact that a carton contains 12 eggs to determine the mean value of $z = \text{the number of unbroken eggs}$. (Hint: z can be written as a linear function of y ; see Example 7.15.)

- 7.37** Exercise 7.8 gave the following probability distribution for x = the number of courses for which a randomly selected student at a certain university is registered:

x	1	2	3	4	5	6	7
$p(x)$	0.02	0.03	0.09	0.25	0.40	0.16	0.05

For this probability distribution, $\mu = 4.66$ and $\sigma = 1.20$.

- a. Because $\mu - \sigma = 3.46$, the x values 1, 2, and 3 are more than 1 standard deviation below the mean. What is the probability that x is more than 1 standard deviation below its mean? (Hint: See Example 7.13.)
- b. What x values are more than 2 standard deviations away from the mean value (either less than $\mu - 2\sigma$ or greater than $\mu + 2\sigma$)?
- c. What is the probability that x is more than 2 standard deviations away from its mean value?

- 7.38** Example 7.11 gave the probability distributions of x = number of flaws in a randomly selected glass panel for two suppliers of glass used to make flat screen TVs. If the manufacturer wanted to select a single supplier for glass panels, which of these two suppliers would you recommend? Justify your choice based on consideration of both center and variability. (Hint: See Example 7.11.)

- 7.39** Consider a large ferry that can accommodate cars and buses. The toll for cars is \$3, and the toll for buses is \$10. Let x and y denote the number of cars and buses, respectively, carried on a single trip. Cars and buses are accommodated on different levels of the ferry, so the number of buses accommodated on any trip is independent of the number of cars on the trip. Suppose that x and y have the following probability distributions:

x	0	1	2	3	4	5
$p(x)$	0.05	0.10	0.25	0.30	0.20	0.10

y	0	1	2
$p(y)$	0.50	0.30	0.20

- a. Calculate the mean and standard deviation of x .
- b. Calculate the mean and standard deviation of y .

- 7.40** Refer to the information given in Exercise 7.39.

- a. Calculate the mean and variance of the total amount of money collected in tolls from cars.
- b. Calculate the mean and variance of the total amount of money collected in tolls from buses.

- 7.41** Refer to the information given in Exercise 7.39.

- a. Calculate the mean and variance of z = total number of vehicles (cars and buses) on the ferry.

- b. Calculate the mean and variance of w = total amount of money collected in tolls.

- 7.42** Suppose that for a particular computer salesperson, the probability distribution of x = the number of systems sold in 1 month is given by the following table:

x	1	2	3	4	5	6	7	8
$p(x)$	0.05	0.10	0.12	0.30	0.30	0.11	0.01	0.01

- a. Find the mean value of x (the mean number of systems sold).
- b. Find the variance and standard deviation of x . How would you interpret these values?
- c. What is the probability that the number of systems sold is within 1 standard deviation of its mean value?
- d. What is the probability that the number of systems sold is more than 2 standard deviations from the mean? (Hint: See Example 7.13.)

- 7.43** A local television station sells 15-second, 30-second, and 60-second advertising spots. Let x denote the length of a randomly selected commercial appearing on this station. Suppose that the probability distribution of x is given by the following table:

x	15	30	60
$p(x)$	0.1	0.3	0.6

- a. Find the mean length for commercials appearing on this station.
- b. If a 15-second spot sells for \$500, a 30-second spot for \$800, and a 60-second spot for \$1000, find the average amount paid for commercials appearing on this station. (Hint: Consider a new variable, y = cost, and then find the probability distribution and mean value of y .)

- 7.44** An author has written a book and submitted it to a publisher. The publisher offers to print the book and gives the author the choice between a flat payment of \$10,000 and a royalty plan. Under the royalty plan the author would receive \$1 for each copy of the book sold. The author thinks that the following table gives the probability distribution of the variable x = the number of books sold:

x	1000	5000	10,000	20,000
$p(x)$	0.05	0.30	0.40	0.25

Which payment plan should the author choose? Explain your reasoning.

- 7.45** A grocery store has an express line for customers purchasing five or fewer items. Let x be the number of items purchased by a randomly selected customer using this line. Give examples of two different assignments of probabilities such that the resulting distributions have the same mean but different standard deviations.

- 7.46** An appliance dealer sells three different models of upright freezers having 13.5, 15.9, and 19.1 cubic feet of storage space. Let x = the amount of storage space purchased by a customer who buys a freezer. Suppose that x has the following probability distribution:

x	13.5	15.9	19.1
$p(x)$	0.2	0.5	0.3

- a. Calculate the mean and standard deviation of x .
 - b. Suppose the price of the freezer depends on the size of the storage space, x . If $\text{Price} = 25x - 8.5$, what is the mean price paid by customers who buy freezers?
 - c. What is the standard deviation of the price paid?
- 7.47** To assemble a piece of furniture, a wood peg must be inserted into a predrilled hole. Suppose that the diameter of a randomly selected peg is a random variable with mean 0.25 inch and standard deviation 0.006 inch and that the diameter of a randomly selected hole is a random variable with mean 0.253 inch and standard deviation 0.002 inch. Let x_1 = peg diameter, and let x_2 = hole diameter.
- a. Why would the random variable y , defined as $y = x_2 - x_1$, be of interest to the furniture manufacturer?
 - b. What is the mean value of the random variable y ?
 - c. Assuming that x_1 and x_2 are independent, what is the standard deviation of y ?
 - d. Based on your answers to Parts (b) and (c), do you think that finding a peg that is too big to fit in the predrilled hole would be a relatively common or a relatively rare occurrence? Explain.

- 7.48** A multiple-choice exam consists of 50 questions. Each question has five choices, of which only one is correct. Suppose that the total score on the exam is calculated as

$$y = x_1 - \frac{1}{4}x_2$$

where x_1 = number of correct responses and x_2 = number of incorrect responses. (Calculating a total score by subtracting a term based on the number of incorrect responses is known as a correction for guessing and is designed to discourage test takers from choosing answers at random.)

- a. It can be shown that if a totally unprepared student answers all 50 questions by just selecting one of the five answers at random, then $\mu_{x_1} = 10$ and $\mu_{x_2} = 40$. What is the mean value of the total score, y ? Does this surprise you? Explain. (Hint: See Example 7.16.)

- b. Explain why it is unreasonable to use the formulas given in this section to compute the variance or standard deviation of y .

- 7.49** Consider a game in which a red die and a blue die are rolled. Let x_R denote the value showing on the uppermost face of the red die, and define x_B similarly for the blue die.

- a. The probability distribution of x_R is

x_R	1	2	3	4	5	6
$p(x_R)$	1/6	1/6	1/6	1/6	1/6	1/6

Find the mean, variance, and standard deviation of x_R .

- b. What are the values of the mean, variance, and standard deviation of x_B ? (You should be able to answer this question without doing any additional calculations.)

- 7.50** Consider the random variables x_R and x_B defined in the previous exercise.

- a. Suppose that you are offered a choice of the following two games:

Game 1: Pay \$7 to play, and you win y_1 dollars, where $y_1 = x_R + x_B$.

Game 2: Doesn't cost anything to play initially, but you "win" $3y_2$ dollars, where

$y_2 = x_R - x_B$. If y_2 is negative, you must pay that amount; if it is positive, you receive that amount.

For Game 1, the net amount won in a game is $w_1 = y_1 - 7 = x_R + x_B - 7$. What are the mean and standard deviation of w_1 ?

- b. For Game 2, the net amount won in a game is $w_2 = 3y_2 = 3(x_R - x_B)$. What are the mean and standard deviation of w_2 ?
- c. Based on your answers to Parts (a) and (b), if you had to play, which game would you choose? Explain your reasoning.

SECTION 7.5 Binomial and Geometric Distributions

In this section, we introduce two discrete probability distributions: the binomial distribution and the geometric distribution. These distributions arise when a chance experiment consists of making a sequence of observations when there are two possible values for each observation. The process of making a single observation is called a *trial*.

For example, one characteristic of blood type is Rh factor, which can be either positive or negative. We can think of a chance experiment that consists of noting the Rh factor for

each of 25 blood donors as a sequence of 25 trials. Each trial consists of observing the Rh factor of a single donor (there are two possible values—positive or negative).

We could also conduct a different chance experiment that consists of observing the Rh factor of blood donors until a donor who is Rh-negative is encountered. This second experiment can also be viewed as a sequence of trials, but the total number of trials in this experiment is not predetermined, as it was in the previous example, where we knew in advance that there would be 25 trials. Experiments of the two types just described are typical of those leading to the binomial distribution and to the geometric distribution.

Binomial Distributions

Suppose that we decide to record the sex of each of the next 25 babies born at a particular hospital. What is the chance that at least 15 are female? What is the chance that between 10 and 15 are female? How many among the 25 can we expect to be female? These and other similar questions can be answered by studying the **binomial probability distribution**.

The binomial distribution arises when the chance experiment of interest is a **binomial experiment**. Binomial experiments have the properties listed in the following box.

Properties of a Binomial Experiment

A **binomial experiment** consists of a sequence of trials with the following conditions:

1. There are a fixed number of trials.
2. Each trial can result in one of only two possible outcomes, labeled success (S) and failure (F).
3. Outcomes of different trials are independent.
4. The probability that a trial results in a success is the same for each trial.

The **binomial random variable x** is defined as

$x = \text{number of successes observed when a binomial experiment is performed}$

The probability distribution of x is called the **binomial probability distribution**.

The term *success* here does not necessarily have its usual meaning. Which of the two possible outcomes is labeled “success” is determined by the random variable of interest. For example, if the variable counts the number of female births among the next 25 births at a particular hospital, then a female birth would be labeled a success (because this is what the variable counts). If male births were counted instead, a male birth would be labeled a success and a female birth a failure.

One situation in which a binomial probability distribution arises was given in Example 7.5. In that example, we considered $x = \text{number among four customers who selected an energy efficient refrigerator (rather than a less expensive model)}$. This is a binomial experiment with four trials, where the purchase of an energy efficient refrigerator is considered a success and $P(\text{success}) = P(E) = 0.4$. The 16 possible outcomes, along with the associated probabilities, were displayed in Table 7.1.

Consider now the case of five customers, a binomial experiment with five trials. The possible values of

$x = \text{number who purchase an energy efficient refrigerator}$

are 0, 1, 2, 3, 4, and 5. There are 32 possible outcomes of the binomial experiment, each one a sequence of five successes and failures. Five of these outcomes result in $x = 1$: SFFFF, FSFFF, FFSSF, FFFSF, and FFFFS.

Because the trials are independent, the first of these outcomes has probability

$$\begin{aligned} P(\text{SFFFF}) &= P(S)P(F)P(F)P(F)P(F) \\ &= (0.4)(0.6)(0.6)(0.6)(0.6) \\ &= (0.4)(0.6)^4 \\ &= 0.05184 \end{aligned}$$

The probability calculation will be the same for any outcome with only one success ($x = 1$). It does not matter where in the sequence the single success occurs. It follows that

$$\begin{aligned} p(1) &= P(x = 1) \\ &= P(SFFFF \text{ or } FSFFF \text{ or } FFSFF \text{ or } FFFSF \text{ or } FFFFS) \\ &= 0.05184 + 0.05184 + 0.05184 + 0.05184 + 0.05184 \\ &= (5)(0.05184) \\ &= 0.25920 \end{aligned}$$

Similarly, there are 10 outcomes for which $x = 2$, because there are 10 ways to select two from among the five trials to be the S 's: $SSFFF$, $SFSFF$, . . . , and $FFFSS$. The probability of each results from multiplying together (0.4) two times and (0.6) three times. For example,

$$\begin{aligned} P(SSFFF) &= (0.4)(0.4)(0.6)(0.6)(0.6) \\ &= (0.4)^2(0.6)^3 \\ &= 0.03456 \end{aligned}$$

and so

$$\begin{aligned} p(2) &= P(x = 2) \\ &= P(SSFFF) + \dots + P(FFFSS) \\ &= (10)(0.4)^2(0.6)^3 \\ &= 0.34560 \end{aligned}$$

The general form of the formula for calculating the probabilities associated with the different possible values of x is

$$\begin{aligned} p(x) &= P(x S's \text{ among the five trials}) \\ &= (\text{number of outcomes with } x S's) \cdot (\text{probability of any given outcome with } x S's) \\ &= (\text{number of outcomes with } x S's) \cdot (0.4)^x(0.6)^{5-x} \end{aligned}$$

This form was seen previously where $p(2) = 10(0.4)^2(0.6)^3$.

The letter n is used to denote the number of trials in the binomial experiment. Then the number of outcomes with $x S$'s is the number of ways of selecting x from among the n trials to be the success trials. A simple expression for this quantity is

$$\text{number of outcomes with } x \text{ successes} = \frac{n!}{x!(n-x)!}$$

where, for any positive whole number m , the symbol $m!$ (read “ m factorial”) is defined by

$$m! = m(m-1)(m-2) \cdots (2)(1)$$

and $0! = 1$.

The Binomial Distribution

Notation:

- n = number of independent trials in a binomial experiment
- p = probability of success for each trial

Then

$$p(x) = P(x \text{ successes among } n \text{ trials})$$

$$= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad x = 0, 1, 2, \dots, n$$

The expressions $\binom{n}{x}$ or ${}_n C_x$ are sometimes used in place of $\frac{n!}{x!(n-x)!}$. Both are read as “ n choose x ” and represent the number of ways of choosing x items

(continued)

from a set of n . Using this notation, the binomial probability function can also be written as

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, 2, \dots, n$$

or

$$p(x) = {}_n C_x p^x (1-p)^{n-x} \quad x = 0, 1, 2, \dots, n$$

Notice that here the binomial probability distribution is specified using a formula that allows calculation of the various probabilities rather than by giving a table or a probability histogram.

Example 7.18 Recognizing Your Roommate's Scent

Understand the context ➤

An interesting experiment was described in the paper “**Sociochemosensory and Emotional Functions**” (*Psychological Science* [2009]: 1118–1123). The authors of this paper wondered if college students could recognize their roommate by scent. They carried out an experiment in which female college students used fragrance-free soap, deodorant, shampoo, and laundry detergent for a period of time. Their bedding was also laundered using a fragrance-free detergent. Each person was then given a new t-shirt that she slept in for one night. The shirt was then collected and sealed in an airtight bag. Later, the roommate was presented with three identical t-shirts (one worn by her roommate and two worn by other women) and asked to pick the one that smelled most like her roommate.

This process was repeated a second time, with the shirts refolded and rearranged before the second trial. The researchers recorded how many times (0, 1, or 2) that the shirt worn by the roommate was correctly identified.

This can be viewed as a binomial experiment consisting of $n = 2$ trials. Each trial results in either a correct identification or an incorrect identification. Because the researchers counted the number of correct identifications, a correct identification is considered a success. We can then define

$$x = \text{number of correct identifications}$$

Formulate a plan ➤

Suppose that a participant is not able to identify her roommate by smell. If this is the case, she is just picking one of the three shirts at random. In this case, the probability of success (picking the correct shirt) is $1/3$. And, if a participant can't identify her roommate by smell, it is also reasonable to regard the two trials as independent. Then this experiment satisfies the conditions of a binomial experiment, and x is a binomial random variable with $n = 2$ and $p = 1/3$.

Do the work ➤

We can use the binomial probability distribution formula to calculate the probability associated with each of the possible x values as follows:

$$p(0) = \binom{2}{0} \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^2 = \frac{2!}{0!2!} \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^2 = (1)(1) \left(\frac{2}{3}\right)^2 = 0.4444$$

$$p(1) = \binom{2}{1} \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^1 = \frac{2!}{1!1!} \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^1 = (2) \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^1 = 0.4444$$

$$p(2) = \binom{2}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^0 = \frac{2!}{2!0!} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^0 = (1) \left(\frac{1}{3}\right)^2 (1) = 0.1111$$

Summarizing in table form gives

<i>x</i>	<i>p(x)</i>
0	0.4444
1	0.4444
2	0.1111

- Interpret the results ➤ This means that about 44.4% of the time, a person who is just guessing would not pick the correct shirt on either trial, about 44.4% of the time the correct shirt would be identified on one of the two trials, and about 11.1% of the time the correct shirt would be identified on both trials.

The authors actually performed this experiment with 44 subjects. They reported that 47.7% of the subjects did not identify the correct shirt on either trial, 22.7% identified the correct shirt on one trial, and 31.7% identified the correct shirt on both trials. The fact that these observed percentages differed quite a bit from what would have been expected if participants were just guessing (as specified by the binomial probabilities in the table above) was interpreted by the authors as evidence that some women could in fact identify their roommates by smell.

Example 7.19 Computer Sales

- Understand the context ➤ Suppose that 60% of all computers sold at an electronics store are laptops and 40% are desktop models. The type of computer purchased by each of the next 12 customers will be noted. Define a random variable x as

x = number of computers among these 12 that are laptops

- Formulate a plan ➤ Because x counts the number of laptops, we use S to denote the sale of a laptop. Then x is a binomial random variable with $n = 12$ and $p = P(S) = 0.60$. The probability distribution of x is given by

$$p(x) = \frac{12!}{x!(12-x)!} (0.6)^x (0.4)^{12-x} \quad x = 0, 1, 2, \dots, 12$$

- Do the work ➤ The probability that exactly four computers are laptops is

$$\begin{aligned} p(4) &= P(x = 4) \\ &= \frac{12!}{4!8!} (0.6)^4 (0.4)^8 \\ &= (495)(0.6)^4 (0.4)^8 \\ &= 0.042 \end{aligned}$$

If group after group of 12 purchases is examined, the long-run percentage of those with exactly four laptops will be 4.2%.

The probability that between four and seven (inclusive) are laptops is

$$P(4 \leq x \leq 7) = P(x = 4 \text{ or } x = 5 \text{ or } x = 6 \text{ or } x = 7)$$

Since these outcomes are mutually exclusive, this is equal to

$$\begin{aligned} P(4 \leq x \leq 7) &= p(4) + p(5) + p(6) + p(7) \\ &= \frac{12!}{4!8!} (0.6)^4 (0.4)^8 + \dots + \frac{12!}{7!5!} (0.6)^7 (0.4)^5 \\ &= 0.042 + 0.101 + 0.177 + 0.227 \\ &= 0.547 \end{aligned}$$

Notice that

$$\begin{aligned} P(4 < x < 7) &= P(x = 5 \text{ or } x = 6) \\ &= p(5) + p(6) \\ &= 0.278 \end{aligned}$$

so the probability depends on whether $<$ or \leq appears in the inequality. (This is typical of *discrete* random variables.)

The binomial distribution formula can be tedious to use unless n is small. Statistical software and most graphing calculators can compute binomial probabilities. If you don't have access to technology, Appendix Table 9 can be used to find binomial probabilities for selected values of n and p .

Using Appendix Table 9

To find $p(x)$ for any particular value of x ,

1. Locate the part of the table corresponding to your value of n (5, 10, 15, 20, or 25).
2. Move down to the row labeled with your value of x .
3. Go across to the column headed by the specified value of p .

The desired probability is at the intersection of the designated x row and p column. For example, when $n = 20$ and $p = 0.8$,

$$p(15) = P(x = 15) = (\text{entry at intersection of } n = 15 \text{ row and } p = 0.8 \text{ column}) = 0.175$$

Although $p(x)$ is positive for every possible x value, many probabilities are zero when rounded to three decimal places, so they appear as 0.000 in the table. More extensive binomial tables are available.

Sampling Without Replacement

Usually, sampling is carried out without replacement. This means that once an element has been selected for the sample, it is not a candidate for future selection. However, if sampling is done by selecting an element from the population, observing whether it is a success or a failure, and then returning it to the population before the next selection is made, the variable

$$x = \text{number of successes observed in the sample}$$

would fit all the requirements of a binomial random variable.

When sampling is done without replacement, the trials (individual selections) are not independent. In this case, the number of successes observed in the sample does not have a binomial distribution but rather a different type of distribution called a *hypergeometric distribution*. The probability calculations for this distribution are even more tedious than for the binomial distribution. Fortunately, when the sample size n is much smaller than N , the population size, probabilities calculated using the binomial distribution and those calculated using the hypergeometric distribution are very close in value. They are so close, in fact, that we often ignore the difference and use the binomial probabilities in place of the hypergeometric probabilities. The following box provides a guideline for determining whether it is appropriate to use the binomial probability distribution when sampling without replacement.

Let x denote the number of S 's in a sample of size n selected without replacement from a population consisting of N individuals or objects. If $(n/N) \leq 0.05$ (that is, at most 5% of the population is sampled), then the binomial distribution provides a reasonable approximation to the actual probability distribution of x .*

*In Chapter 8, we will see a different situation where a similar condition is introduced, but where the requirement is that at most 10% of the population is included in the sample. Be careful not to confuse these rules.

Example 7.20 Saying I Do...

Understand the context ➤

A survey of 505 American women in 2016 found that only about 25% favor preserving the tradition of having the bride promise to obey her husband as part of wedding vows (yahoo.com/style/women-want-the-word-obey-dropped-from-wedding-162058081.html, retrieved May 2, 2017). Suppose that exactly 25% of American women favor preserving this tradition. Consider a random sample of $n = 20$ American women (much less than 5% of the population). Then

x = the number in the sample who favor preserving the promise to obey

Do the work ➤ has (approximately) a binomial distribution with $n = 20$ and $p = .25$. The probability that five of those sampled favor preserving the promise to obey is

$$\begin{aligned} P(5) &= P(x = 5) \\ &= \text{entry in } x = 5 \text{ row and } p = 0.25 \text{ column in Appendix Table 9 for } n = 20 \\ &= 0.202 \end{aligned}$$

The probability that at least half of those in the sample (that is, 10 or more) favor preserving the promise to obey is

$$\begin{aligned} P(x \geq 10) &= P(x = 10, 11, 12, \dots, 20) \\ &= p(10) + p(11) + \dots + p(20) \\ &= 0.010 + 0.003 + 0.001 + \dots + 0.000 \\ &= 0.014 \end{aligned}$$

Interpret the results ➤

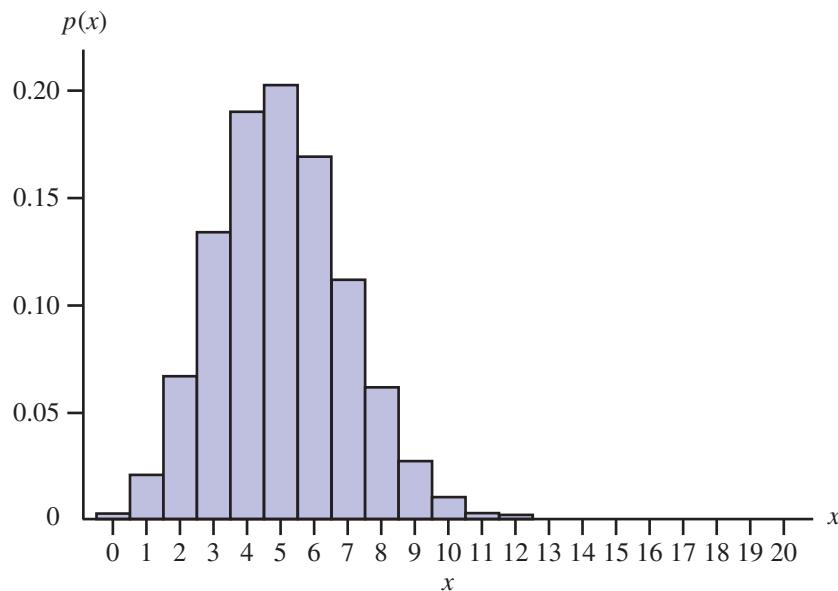
If $p = 0.25$, only about 1.4% of all samples of size 20 would result in at least 10 women who favor preserving the promise to obey. Because $P(x \geq 10)$ is so small when $p = 0.25$, if $x \geq 10$ were actually observed, we would have to wonder whether the reported value of $p = 0.25$ is correct.

Although it is possible that we would observe $x \geq 10$ when $p = 0.25$ (this would happen about 1.4% of the time in the long run), it might also be the case that p is actually greater than 0.25. In Chapter 10, we see how hypothesis-testing methods can be used to decide which of two contradictory claims about a population (such as $p = 0.25$ or $p > 0.25$) is more believable.

Technology, the binomial formula or tables can be used to find each of the 21 probabilities $p(0), p(1), \dots, p(20)$. Figure 7.15 shows the probability histogram for the binomial distribution with $n = 20$ and $p = 0.25$. Notice that the distribution is skewed to the right. (The binomial distribution is symmetric only when $p = 0.5$.)

FIGURE 7.15

The binomial probability histogram when $n = 20$ and $p = 0.25$.



Mean and Standard Deviation of a Binomial Random Variable

A binomial random variable x based on n trials has possible values 0, 1, 2, ..., n , so the mean value is

$$\mu_x = \sum x p(x) = (0)p(0) + (1)p(1) + \cdots + (n)p(n)$$

and the variance of x is

$$\begin{aligned}\sigma_x^2 &= \sum (x - \mu_x)^2 \cdot p(x) \\ &= (0 - \mu_x)^2 p(0) + (1 - \mu_x)^2 p(1) + \cdots + (n - \mu_x)^2 p(n)\end{aligned}$$

These expressions would be very tedious to evaluate for any particular values of n and p . Fortunately, there are simple formulas for the mean and standard deviation of a binomial random variable.

The mean value and the standard deviation of a binomial random variable x are

$$\mu_x = np \quad \text{and} \quad \sigma_x = \sqrt{np(1-p)}$$

Example 7.21 Budgets and Tracking Spending

Understand the context ➤

The report “[The 2016 Consumer Financial Literacy Survey](#)” ([The National Foundation for Credit Counseling, nfcc.org/wp-content/uploads/2016/04/NFCC_BECU_2016-FLS_datasheet-with-key-findings_041516.pdf](http://The-National-Foundation-for-Credit-Counseling-nfcc.org/wp-content/uploads/2016/04/NFCC_BECU_2016-FLS_datasheet-with-key-findings_041516.pdf), retrieved May 2, 2017) estimates that 40% of U.S. adults say that they have a budget and keep close track of their spending. This estimate was based on a representative sample of more than 1500 people. Suppose that 40% of all U.S. adults have a budget and keep close track of their spending. A random sample of $n = 25$ U.S. adults is to be selected. Consider the random variable

x = number in the sample who have a budget and keep track of spending.

Formulate a plan ➤

Even though sampling is done without replacement, the sample size $n = 25$ is very small compared to the total number of U.S. adults, so we can approximate the probability distribution of x using a binomial distribution with $n = 25$ and $p = 0.4$. Having a budget and keeping track of spending is identified as a success because this is the outcome counted by the random variable x .

Do the work ➤

The mean value of x is then

$$\mu_x = np = 25(0.40) = 10.0$$

and the standard deviation is

$$\sigma_x = \sqrt{np(1-p)} = \sqrt{25(0.40)(0.60)} = \sqrt{6} = 2.45$$

The probability that x is farther than 1 standard deviation from its mean value is

$$\begin{aligned}P(x < \mu_x - \sigma_x \text{ or } x > \mu_x + \sigma_x) &= P(x < 7.55 \text{ or } x > 12.45) \\ &= P(x \leq 7) + P(x \geq 13) \\ &= p(0) + \cdots + p(7) + p(13) + \cdots + p(25) \\ &= 0.307 \quad (\text{using technology or Appendix Table 9})\end{aligned}$$

The value of σ_x is 0 when $p = 0$ or $p = 1$. In these two cases, there is no uncertainty in x . We are sure to observe $x = 0$ when $p = 0$ and $x = n$ when $p = 1$. It is also easily verified that $p(1-p)$ is largest when $p = 0.5$. This means that the binomial distribution spreads out the most when sampling from a 50–50 population. The farther p is from 0.5, the less spread out and the more skewed is the binomial distribution.

Geometric Distributions

A binomial random variable is defined as the number of successes in n independent trials, where each trial can result in either a success or a failure and the probability of success is the same for each trial. Suppose, however, that we are not interested in the number of successes in a fixed number of trials but rather in the number of trials that must be carried out before a success occurs. Two examples are counting the number of boxes of cereal that must be purchased before finding one with a rare toy and counting the number of games that a professional bowler must play before achieving a score over 250.

The variable

x = number of trials to first success

is called a **geometric random variable**, and the probability distribution that describes its behavior is called a **geometric probability distribution**.

Suppose an experiment consists of a sequence of trials with the following conditions:

1. The trials are independent.
2. Each trial can result in one of two possible outcomes, success and failure.
3. The probability of success is the same for all trials.

A **geometric random variable** is defined as

x = number of trials until the first success is observed (including the success trial)

The probability distribution of x is called the **geometric probability distribution**.

For example, suppose that 40% of the students who drive to campus at your university carry jumper cables. Your car has a dead battery and you don't have jumper cables, so you decide to stop students who are headed to the parking lot and ask them whether they have a pair of jumper cables. You might be interested in the number of students you would have to stop before finding one who has jumper cables.

If we define success as a student with jumper cables, a trial would consist of asking an individual student for help. The random variable

x = number of students who must be stopped before finding one with jumper cables
is an example of a geometric random variable, because it can be viewed as the number of trials to the first success in a sequence of independent trials.

The probability distribution of a geometric random variable is easy to construct. We use p to denote the probability of success on any given trial. Possible outcomes can be denoted as follows:

Outcome	$x = \text{Number of Trials to First Success}$
S	1
FS	2
FFS	3
⋮	⋮
$FFFFFFS$	7
⋮	⋮

Each possible outcome consists of 0 or more failures followed by a single success. So,

$$\begin{aligned}
 p(x) &= P(x \text{ trials to first success}) \\
 &= P(FF\dots FS)
 \end{aligned}$$

\uparrow
 $x - 1$ failures followed by a success on trial x

Because the probability of success is p for each trial, the probability of failure for each trial is $1 - p$. Because the trials are independent,

$$\begin{aligned} p(x) &= P(x \text{ trials to first success}) = P(FF\dots FS) \\ &= P(F)P(F)\cdots P(F)P(S) \\ &= (1-p)(1-p)\cdots(1-p)p \\ &= (1-p)^{x-1}p \end{aligned}$$

This leads us to the formula for the geometric probability distribution.

Geometric Probability Distribution

If x is a geometric random variable with probability of success = p for each trial, then

$$p(x) = (1-p)^{x-1}p \quad x = 1, 2, 3, \dots$$



Example 7.22 Jumper Cables

Consider the jumper cable scenario described previously. For this problem, $p = 0.4$, because 40% of the students who drive to campus carry jumper cables. The probability distribution of

x = number of students who must be stopped before finding a student with jumper cables

is

$$p(x) = (0.6)^{x-1}(0.4) \quad x = 1, 2, 3, \dots$$

The probability distribution can now be used to calculate various probabilities. For example, the probability that the first student stopped has jumper cables (that is, $x = 1$) is

$$p(1) = (0.6)^{1-1}(0.4) = (0.6)^0(0.4) = 0.4$$

The probability that three or fewer students must be stopped is

$$\begin{aligned} P(x \leq 3) &= p(1) + p(2) + p(3) \\ &= (0.6)^0(0.4) + (0.6)^1(0.4) + (0.6)^2(0.4) \\ &= 0.4 + 0.24 + 0.144 \\ &= 0.784 \end{aligned}$$

EXERCISES 7.51 - 7.69

- 7.51** CBS News reported that 4% of adult Americans have a food allergy ([June 1, 2017, cbsnews.com /news/food-allergies-in-america-new-report -shellfish-peanut-dairy](#), retrieved March 25, 2018). Consider selecting 10 adult Americans at random. Define the random variable x as

x = number of people in the sample of 10 that have a food allergy.

Find the following probabilities. (Hint: See Examples 7.19 and 7.20.)

- a. $p(x < 3)$
- b. $p(x \leq 3)$

- c. $p(x \geq 4)$
- d. $p(1 \leq x \leq 3)$

- 7.52** The article “Should You Report That Fender-Bender?” ([Consumer Reports, 2013:15](#)) reported that 7 in 10 auto accidents involve a single vehicle. Suppose 15 accidents are randomly selected. (Hint: See Examples 7.19 and 7.20.)
- a. What is the probability that exactly four involve a single vehicle?
 - b. What is the probability that at most four involve a single vehicle?
 - c. What is the probability that exactly six involve multiple vehicles?

- 7.53** FlightView surveyed 2600 North American airline passengers and reported that approximately 80% said that they carry a smart phone when they travel (flightview.com/TravelersSurvey/downloads/survey_infographic_poster.pdf). Suppose that the actual percentage is 80%. Consider randomly selecting six passengers and define the random variable x to be the number of the six selected passengers who travel with a smart phone. The probability distribution of x is the binomial distribution with $n = 6$ and $p = 0.8$.
- Calculate $p(4)$, and interpret this probability.
 - Calculate $p(6)$, the probability that all six selected passengers travel with a smart phone.
 - Calculate $P(x \geq 4)$.
- 7.54** Refer to the previous exercise, and suppose that 10 rather than six passengers are selected ($n = 10$, $p = 0.8$). (Hint: Use technology or Appendix Table 9.)
- Calculate $p(8)$.
 - Calculate $P(x \leq 7)$.
 - Calculate the probability that more than half of the selected passengers travel with a smart phone.
- 7.55** Twenty-five percent of the customers of a grocery store use an express checkout. Consider five randomly selected customers, and let x denote the number among the five who use the express checkout.
- Calculate $p(2)$.
 - Calculate $P(x \leq 1)$.
 - Calculate $P(2 \leq x)$. (Hint: Make use of your answer from Part (b).)
 - Calculate $P(x \neq 2)$.
- 7.56** Example 7.18 described a study in which a person was asked to determine which of three t-shirts had been worn by her roommate by smelling the shirts ("Sociochemosensory and Emotional Functions," *Psychological Science* [2009]: 1118–1123). Suppose that instead of three shirts, each participant was asked to choose among four shirts and that the process was repeated five times. Then, assuming that the participant is choosing at random, x = number of correct identifications is a binomial random variable with $n = 5$ and $p = \frac{1}{4}$.
- What are the possible values of x ?
 - For each possible value of x , find the associated probability $p(x)$ and display the possible x values and $p(x)$ values in a table.
 - Construct a probability histogram for the probability distribution of x .
- 7.57** Information Security Buzz provides news for the information security community. In an article published on September 24, 2016, it reported that based on a large international survey of Internet users, 60% of Internet users have installed security solutions on all of the devices they use to access the Internet (informationsecuritybuzz.com/articles/21-29-60-kaspersky-lab-presents-first-cybersecurity-index/, retrieved May 2, 2017).
- Suppose that the true proportion of Internet users who have security solutions on all the devices they use to access the Internet is 0.60. If 20 Internet users are selected at random, what is the probability that more than 10 have security solutions installed on all devices used to access the Internet?
 - Suppose that a random sample of 20 Internet users is selected. Which is more likely—that more than 15 have security solutions on all devices used to access the Internet or that fewer than 5 have security solutions on all devices used to access the Internet? Justify your answer based on probability calculations.
- 7.58** A breeder of show dogs is interested in the number of female puppies in a litter. If a birth is equally likely to result in a male or a female puppy, give the probability distribution of the variable x = number of female puppies in a litter of size 5.
- 7.59** *Women's Health Magazine* surveyed 1187 readers to find out how often people wash their sheets (womenshealthmag.com/health/dirty-sheets, March 26, 2015, retrieved May 2, 2017). They found that even though microbiologists recommend that you wash your sheets at least once a week, only 44% said that they wash their sheets that often. Suppose this group is representative of adult Americans and define the random variable x to be the number of adult Americans you would have to ask before you found someone that washes his or her sheets at least once a week.
- Is the probability distribution of x binomial or geometric? Explain.
 - What is the probability that you would have to ask three people before finding one who washes sheets at least once a week?
 - What is the probability that fewer than four people would have to be asked before finding one who washes sheets at least once a week?
 - What is the probability that more than three people would have to be asked before finding one who washes sheets at least once a week?
- 7.60** Industrial quality control programs often include inspection of incoming parts from suppliers. If parts are purchased in large lots, a typical plan might be to select 20 parts at random from a lot and inspect them. Suppose that a lot is considered to be acceptable if one or fewer defective parts are found among those inspected. Otherwise, the lot is rejected and returned to the supplier. Use technology

or Appendix Table 9 to find the probability of accepting lots that have each of the following (Hint: Identify success with a defective part):

- 5% defective parts
- 10% defective parts
- 20% defective parts

7.61 Suppose that the probability is 0.1 that any given citrus tree will show measurable damage when the temperature falls to 30°F. (Hint: See Example 7.21.)

- If the temperature does drop to 30°F, what is the expected number of citrus trees showing damage in orchards of 2000 trees?
- What is the standard deviation of the number of trees that show damage?

7.62 Suppose that 30% of all automobiles undergoing an emissions inspection at an inspection station fail the inspection.

- Among 15 randomly selected cars, what is the probability that at most 5 fail the inspection?
- Among 15 randomly selected cars, what is the probability that between 5 and 10 (inclusive) fail to pass inspection?
- Among 25 randomly selected cars, what is the mean value of the number that pass inspection, and what is the standard deviation of the number that pass inspection?
- What is the probability that among 25 randomly selected cars, the number that pass is within 1 standard deviation of the mean value? (Hint: See Example 7.21.)

7.63 Suppose that you will take a multiple-choice exam consisting of 100 questions with five possible responses to each question. You have not studied and so must guess (select one of the five answers in a completely random fashion) on each question. Let x represent the number of correct responses on the test.

- What kind of probability distribution does x have?
- What is your expected score on the exam? (Hint: Your expected score is the mean value of the x distribution.)
- Calculate the variance and standard deviation of x .
- Based on your answers to Parts (b) and (c), is it likely that you would score over 50 on this exam? Explain the reasoning behind your answer.

7.64 Suppose that 20% of the 10,000 signatures on a recall petition are invalid. Would the number of invalid signatures in a sample of 2000 of these signatures have (approximately) a binomial distribution? Explain.

7.65 A city requires that smoke detectors be installed in all houses. There is concern that too many houses are still without detectors, so a costly inspection program is being considered. Let p be the proportion of all houses that have a detector. A random sample of 25 houses is selected. If the sample strongly suggests that $p < 0.80$ (less than 80% have detectors), as opposed to $p \geq 0.80$, the program will be implemented. Let x be the number of residences among the 25 that have a detector, and consider the following decision rule: Reject the claim that $p \geq 0.8$ and implement the program if $x \leq 15$.

- What is the probability that the program is implemented when $p = 0.80$?
- What is the probability that the program is not implemented if $p = 0.70$?
- What is the probability that the program is not implemented if $p = 0.60$?
- How do the “error probabilities” of Parts (b) and (c) change if the value 15 in the decision rule is changed to 14?

7.66 Suppose that 90% of all registered California voters favor banning the release of information from exit polls in presidential elections until after the polls in California close. A random sample of 25 registered California voters is to be selected.

- What is the probability that more than 20 favor the ban?
- What is the probability that at least 20 favor the ban?
- What are the mean value and standard deviation of the number of voters in the sample who favor the ban?
- If fewer than 20 voters in the sample favor the ban, is this inconsistent with the claim that (at least) 90% of California registered voters favors the ban? (Hint: Consider $P(x < 20)$ when $p = 0.9$.)

7.67 Suppose a playlist on a music player consists of 100 songs, of which eight are by a particular artist. Songs are played by selecting a song at random (with replacement) from the playlist. Let the random variable x represent the number of songs played until a song by this artist is played.

- Explain why the probability distribution of x is not binomial.
- Find the following probabilities. (Hint: See Example 7.22.)
 - $p(4)$
 - $P(x \leq 4)$
 - $P(x > 4)$
 - $P(x \geq 4)$
- Interpret each of the probabilities in Part (b).

7.68 Sophie is a dog that loves to play catch. Unfortunately, she isn't very good, and the probability that she catches a ball is only 0.1. Let x be the number of tosses required until Sophie catches a ball.

- Does x have a binomial or a geometric distribution?
- What is the probability that it will take exactly two tosses for Sophie to catch a ball?
- What is the probability that more than three tosses will be required?

7.69 Suppose that 5% of cereal boxes contain a prize and the other 95% contain the message, "Sorry, try again." Consider the random variable x , where $x =$ number of boxes purchased until a prize is found.

- What is the probability that at most two boxes must be purchased?
- What is the probability that exactly four boxes must be purchased?
- What is the probability that more than four boxes must be purchased?

SECTION 7.6 Normal Distributions

Normal distributions formalize the notion of mound-shaped histograms introduced in Chapter 4. Normal distributions are widely used for two reasons. First, they provide a reasonable approximation to the distribution of many different variables. They also play an important role in many of the inferential procedures that will be introduced in later chapters of this textbook.

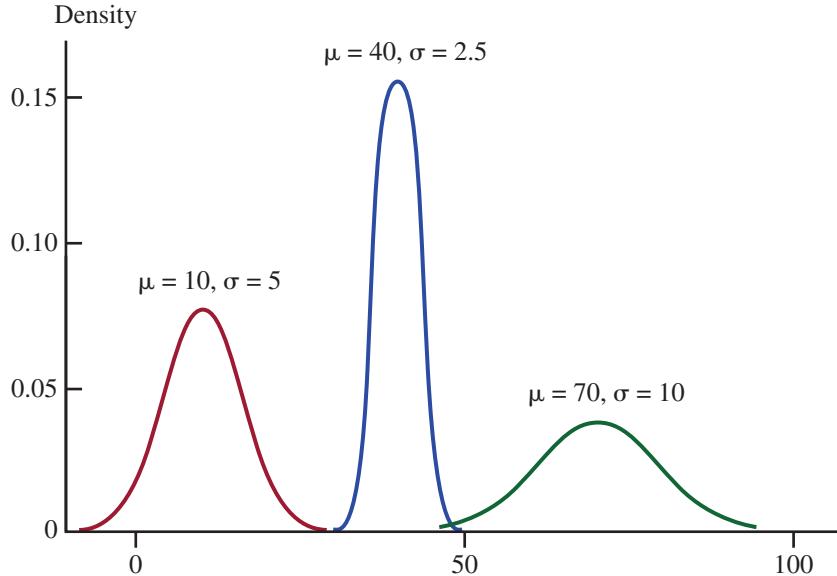
Normal distributions are continuous probability distributions that are bell-shaped and symmetric, as shown in Figure 7.16. Normal distributions are sometimes referred to as *normal curves*.

There are many different normal distributions, and they are distinguished from one another by their mean μ and standard deviation σ . The mean μ of a normal distribution describes where the corresponding curve is centered. The standard deviation σ describes how much the curve spreads out around that center. As with all continuous probability distributions, the total area under any normal curve is equal to 1.

Three normal distributions are shown in Figure 7.17. Notice that the smaller the standard deviation, the taller and narrower the corresponding curve. Remember that areas under a continuous probability distribution curve represent probabilities, so when the standard deviation is small, a larger area is concentrated near the center of the curve. This means that the chance of observing a value near the mean is much greater (because μ is at the center).

FIGURE 7.16
A normal distribution.

FIGURE 7.17
Three normal distributions.

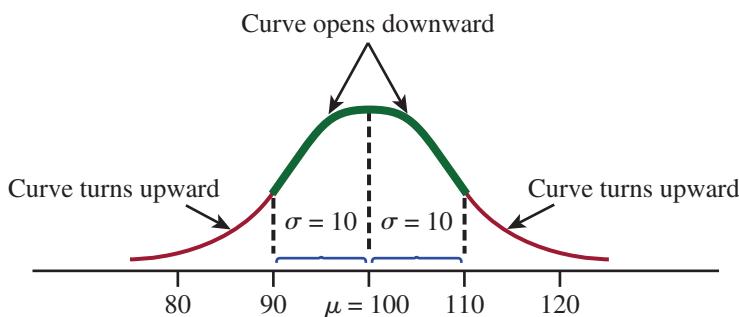


The value of μ is the number on the measurement axis lying directly below the top of the curve. The value of σ can be approximated from a picture of the curve. Consider the normal curve in Figure 7.18. Starting at the top (above $\mu = 100$) and moving to

the right, the curve turns downward until it is above the value 110. After that point, it continues to decrease in height but is turning upward rather than downward. Similarly, to the left of $\mu = 100$, the curve turns downward until it reaches 90 and then begins to turn upward. The curve changes from turning downward to turning upward at a distance of 10 on either side of μ . In general, σ is the distance to either side of μ at which a normal curve changes from turning downward to turning upward, so $\sigma = 10$ for the normal curve in Figure 7.18.

FIGURE 7.18

Mean μ and standard deviation σ for a normal curve.

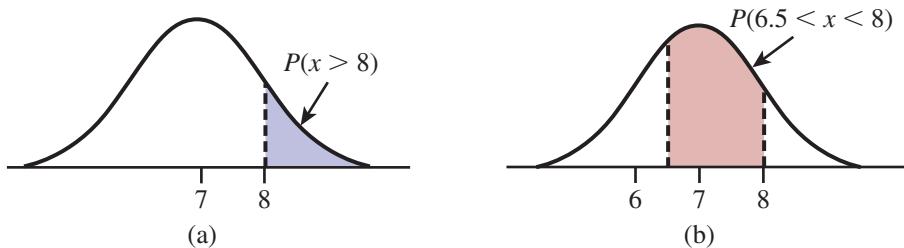


When a normal distribution is used to describe the behavior of a random variable, a mean and a standard deviation must be specified. For example, a normal distribution with mean 7 and standard deviation 1 might be used as a model for the distribution of x = birth weight (in pounds). If this model is a reasonable description of the probability distribution, we can use areas under the normal curve with $\mu = 7$ and $\sigma = 1$ to approximate various probabilities related to birth weight.

For example, the probability that a birth weight is over 8 pounds (expressed symbolically as $P(x > 8)$) corresponds to the shaded area in Figure 7.19(a). The shaded area in Figure 7.19(b) represents the probability of a birth weight between 6.5 and 8 pounds, $P(6.5 < x < 8)$.

FIGURE 7.19

Normal distribution for birth weight:
(a) shaded area = $P(x > 8)$;
(b) shaded area = $P(6.5 < x < 8)$.



Unfortunately, calculating these probabilities (areas under a normal curve) is not simple. To overcome this difficulty, we rely on technology or a table of areas for a reference normal distribution, called the **standard normal distribution**.

DEFINITIONS

Standard normal distribution: The normal distribution with

$$\mu = 0 \quad \text{and} \quad \sigma = 1$$

The corresponding density curve is called the **standard normal curve**.

The letter z is used to represent a variable whose distribution is described by the standard normal curve. The term **z curve** is often used in place of standard normal curve.

There aren't many naturally occurring variables with distributions that are well described by the standard normal distribution. However, this distribution is important because it is

also used in probability calculations for other normal distributions. When we are interested in finding a probability based on some other normal curve, we either rely on technology or we first translate our problem into an equivalent problem that involves finding an area under the standard normal curve. A table for the standard normal distribution is then used to find the desired area. To be able to do this, we must first learn to work with the standard normal distribution.

The Standard Normal Distribution

In working with normal distributions, two general skills are required:

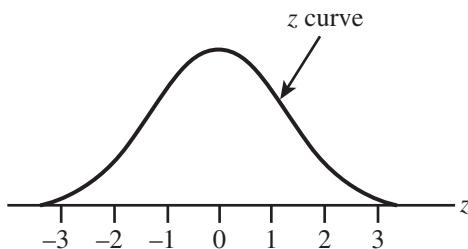
1. We must be able to use the normal distribution to calculate probabilities, which are areas under a normal curve and above given intervals.
2. We must be able to characterize extreme values in the distribution, such as the largest 5%, the smallest 1%, and the most extreme 5% (which would include both the largest 2.5% and the smallest 2.5%).

Let's begin by looking at how to accomplish these tasks when the distribution of interest is the standard normal distribution.

The standard normal or z curve is shown in Figure 7.20. It is centered at $\mu = 0$, and the standard deviation, $\sigma = 1$, is a measure of the extent to which it spreads out about its mean.

FIGURE 7.20

A standard normal (z) curve.



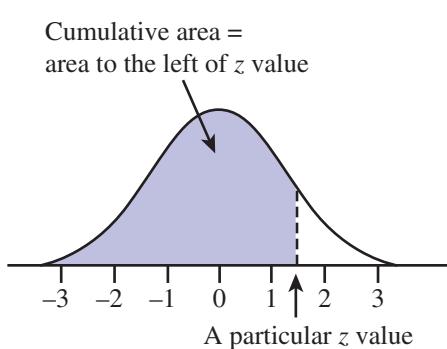
Notice that this picture is consistent with the Empirical Rule of Chapter 4. About 95% of the area (probability) is associated with values that are within 2 standard deviations of the mean (between -2 and 2), and almost all of the area is associated with values that are within 3 standard deviations of the mean (between -3 and 3).

In the examples that follow, Appendix Table 2 is used. If you have access to technology (a statistics software package or a graphing calculator), you can use technology to work these examples.

Appendix Table 2 tabulates cumulative z curve areas of the sort shown in Figure 7.21 for many different values of z . The smallest value for which the cumulative area is given is -3.89 , a value far out in the lower tail of the z curve. The next smallest value for which the area appears is -3.88 , then -3.87 , then -3.86 , and so on in increments of 0.01, ending with the cumulative area to the left of 3.89 .

FIGURE 7.21

A cumulative area under a standard normal (z) curve.



Using the Table of Standard Normal Curve Areas

For any number z^* between -3.89 and 3.89 and rounded to two decimal places, Appendix Table 2 gives

$$(\text{area under } z \text{ curve to the left of } z^*) = P(z < z^*) = P(z \leq z^*)$$

where the letter z is used to represent a random variable whose distribution is the standard normal distribution.

To find this probability using the table, locate the following:

1. The row labeled with the sign of z^* and the digit to either side of the decimal point (for example, -1.7 or 0.5)
2. The column identified with the second digit to the right of the decimal point in z^* (for example, $.06$ if $z^* = -1.76$)

The number at the intersection of this row and column is the probability, $P(z < z^*)$.

A portion of the table of standard normal curve areas appears in Figure 7.22. To find the area under the z curve to the left of 1.42 , look in the row labeled 1.4 and the column labeled $.02$ (the highlighted row and column in Figure 7.22). From the table, the corresponding cumulative area is 0.9222 . So

$$z \text{ curve area to the left of } 1.42 = 0.9222$$

We can also use the table to find the area to the right of a particular value. Because the total area under the z curve is 1 , it follows that

$$\begin{aligned} (z \text{ curve area to the right of } 1.42) &= 1 - (z \text{ curve area to the left of } 1.42) \\ &= 1 - 0.9222 \\ &= 0.0778 \end{aligned}$$

FIGURE 7.22

Portion of the table of standard normal curve areas.

z^*	.00	.01	.02	.03	.04	.05
0.0	.5000	.5040	.5080	.5120	.5160	.5199
0.1	.5398	.5438	.5478	.5517	.5557	.5596
0.2	.5793	.5832	.5871	.5910	.5948	.5987
0.3	.6179	.6217	.6255	.6293	.6331	.6368
0.4	.6554	.6591	.6628	.6664	.6700	.6736
0.5	.6915	.6950	.6985	.7019	.7054	.7088
0.6	.7257	.7291	.7324	.7357	.7389	.7422
0.7	.7580	.7611	.7642	.7673	.7704	.7734
0.8	.7881	.7910	.7939	.7967	.7995	.8023
0.9	.8159	.8186	.8212	.8238	.8264	.8289
1.0	.8413	.8438	.8461	.8485	.8508	.8531
1.1	.8643	.8665	.8686	.8708	.8729	.8749
1.2	.8849	.8869	.8888	.8907	.8925	.8944
1.3	.9032	.9049	.9066	.9082	.9099	.9115
1.4	.9192	.9207	.9222	.9236	.9251	.9265
1.5	.9332	.9345	.9357	.9370	.9382	.9394
1.6	.9452	.9463	.9474	.9484	.9495	.9505
1.7	.9554	.9564	.9573	.9582	.9591	.9599
1.8	.9641	.9649	.9656	.9664	.9671	.9678

$P(z < 1.42)$

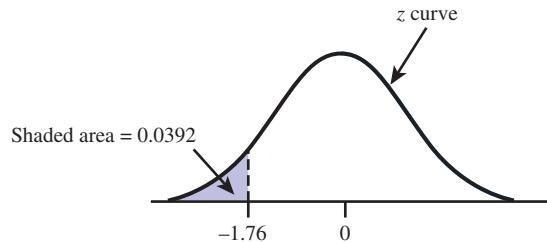
These probabilities can be interpreted to mean that in a long sequence of observations, approximately 92.22% of the observed z values will be less than 1.42 , and approximately 7.78% will be greater than 1.42 .

Example 7.23 Finding Standard Normal Curve Areas

The probability $P(z < -1.76)$ is found at the intersection of the -1.7 row and the $.06$ column of the z table. The result is

$$P(z < -1.76) = 0.0392$$

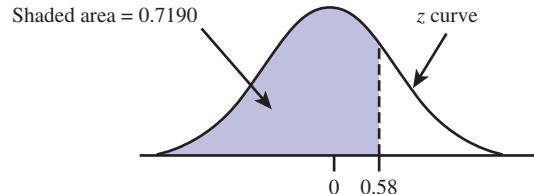
as shown in the following figure:



In other words, in a long sequence of observations, approximately 3.9% of the observed z values will be smaller than -1.76 . Similarly,

$$P(z \leq 0.58) = \text{entry in } 0.5 \text{ row and } .08 \text{ column of Appendix Table 2} = 0.7190$$

as shown in the following figure:



Now consider $P(z < -4.12)$. This probability does not appear in Appendix Table 2; there is no -4.1 row. However, it must be less than $P(z < -3.89)$, the smallest z value in the table, because -4.12 is farther out in the lower tail of the z curve. Since $P(z < -3.89) \approx 0$ (that is, zero when rounded to four decimal places), it follows that

$$P(z < -4.12) \approx 0$$

Similarly,

$$P(z < 4.18) > P(z < 3.89) \approx 1$$

from which we conclude that

$$P(z < 4.18) \approx 1$$

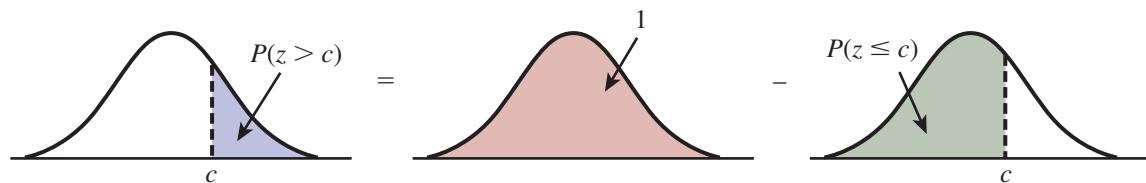
As illustrated in Example 7.23, we can use the cumulative areas tabulated in Appendix Table 2 to calculate other probabilities involving z . The probability that z is greater than a value c is

$$P(z > c) = \text{area under the } z \text{ curve to the right of } c = 1 - P(z \leq c)$$

FIGURE 7.23

The relationship between an upper-tail area and a cumulative area.

In other words, the area to the right of a value (a right-tail area) is 1 minus the corresponding cumulative area. This is illustrated in Figure 7.23.



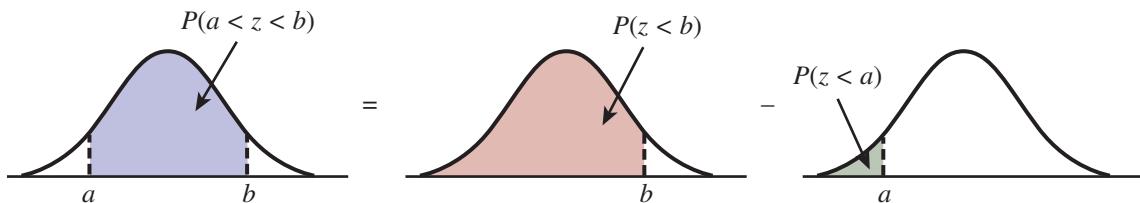
Similarly, the probability that z falls in the interval between a lower limit a and an upper limit b is

$$\begin{aligned} P(a < z < b) &= \text{area under the } z \text{ curve and above the interval from } a \text{ to } b \\ &= P(z < b) - P(z < a) \end{aligned}$$

That is, $P(a < z < b)$ is the difference between two cumulative areas, as illustrated in Figure 7.24.

FIGURE 7.24

$P(a < z < b)$ as the difference between the two cumulative areas.

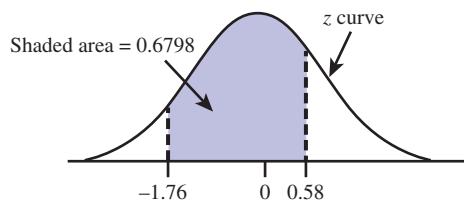


Example 7.24 More Standard Normal Curve Areas

The probability that z is between -1.76 and 0.58 is

$$\begin{aligned} P(-1.76 < z < 0.58) &= P(z < 0.58) - P(z < -1.76) \\ &= 0.7190 - 0.0392 \\ &= 0.6798 \end{aligned}$$

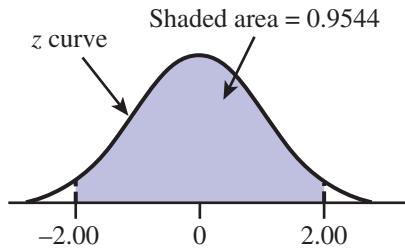
as shown in the following figure:



The probability that z is between -2 and $+2$ (within 2 standard deviations of its mean, since $\mu = 0$ and $\sigma = 1$) is

$$\begin{aligned} P(-2.00 < z < 2.00) &= P(z < 2.00) - P(z < -2.00) \\ &= 0.9772 - 0.0228 \\ &= 0.9544 \\ &\approx 0.95 \end{aligned}$$

as shown in the following figure:

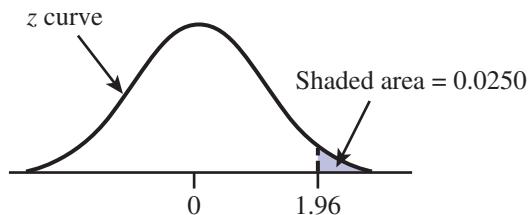


This last probability is the basis for one part of the Empirical Rule, which states that when a histogram is well approximated by a normal curve, approximately 95% of the values are within 2 standard deviations of the mean.

The probability that the value of z exceeds 1.96 is

$$\begin{aligned} P(z > 1.96) &= 1 - P(z < 1.96) \\ &= 1 - 0.9750 \\ &= 0.0250 \end{aligned}$$

as shown in the following figure:



That is, 2.5% of the area under the z curve lies to the right of 1.96 in the upper tail.

Similarly,

$$\begin{aligned} P(z > -1.28) &= \text{area to the right of } -1.28 \\ &= 1 - P(z < -1.28) \\ &= 1 - 0.1003 \\ &= 0.8997 \\ &\approx 0.90 \end{aligned}$$

Identifying Extreme Values

Suppose that we want to describe the values that make up the smallest 2% of a distribution or the values that make up the most extreme 5% (which includes the largest 2.5% and the smallest 2.5%). Examples 7.25 and 7.26 illustrate how this can be done.

Example 7.25 Identifying Extreme Values

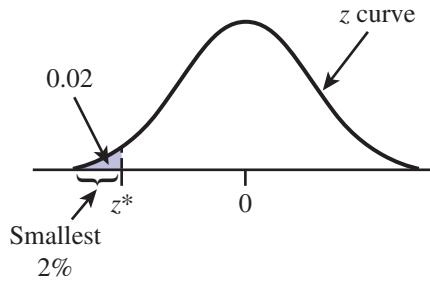
Suppose that we want to describe the values that make up the smallest 2% of the standard normal distribution. Symbolically, we are trying to find a value (call it z^*) such that

$$P(z < z^*) = 0.02$$

This is illustrated in Figure 7.25, which shows that the cumulative area for z^* is 0.02. Therefore, we look for a cumulative area of 0.0200 in the body of Appendix Table 2. The closest cumulative area in the table is 0.0202, in the -2.0 row and $.05$ column. We will use $z^* = -2.05$, the best approximation from the table. Variable values less than -2.05 make up the smallest 2% of the standard normal distribution.

FIGURE 7.25

The smallest 2% of the standard normal distribution.



Now suppose that we are interested in the largest 5% of all z values. We want to find a value of z^* for which

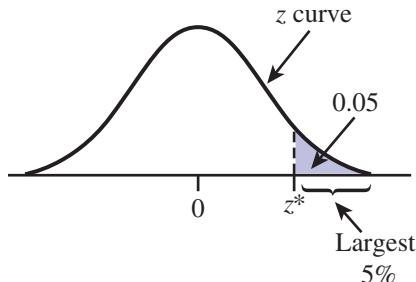
$$P(z > z^*) = 0.05$$

as illustrated in Figure 7.26. Because Appendix Table 2 always works with cumulative area (area to the left), the first step is to determine

$$\text{area to the left of } z^* = 1 - 0.05 = 0.95$$

FIGURE 7.26

The largest 5% of the standard normal distribution.



Looking for the cumulative area closest to 0.95 in Appendix Table 2, we find that 0.95 falls exactly halfway between 0.9495 (corresponding to a z value of 1.64) and 0.9505 (corresponding to a z value of 1.65). Because 0.9500 is exactly halfway between the two areas, we use a z value that is halfway between 1.64 and 1.65. (If one value had been closer to 0.9500 than the other, we would just use the z value corresponding to the closest area.) This gives

$$z^* = \frac{1.64 + 1.65}{2} = 1.645$$

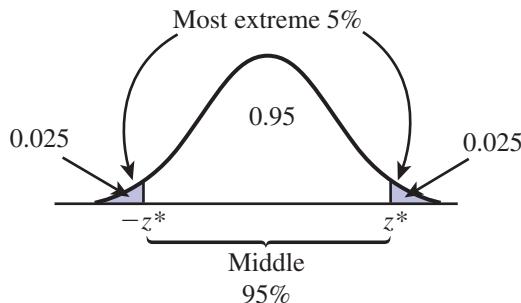
Values greater than 1.645 make up the largest 5% of the standard normal distribution. By symmetry, -1.645 separates the smallest 5% of all z values from the others.

Example 7.26 More Extremes

Sometimes we are interested in identifying the most extreme (unusually large *or* small) values in a distribution. Consider describing the values that make up the most extreme 5% of the standard normal distribution. That is, we want to separate the middle 95% from the extreme 5%. This is illustrated in Figure 7.27.

FIGURE 7.27

The most extreme 5% of the standard normal distribution.



Because the standard normal distribution is symmetric, the most extreme 5% is equally divided between the high side and the low side of the distribution, resulting in an area of 0.025 for each of the tails of the z curve. Symmetry about 0 implies that if z^* denotes the value that separates the largest 2.5%, the value that separates the smallest 2.5% is $-z^*$.

To find z^* , first determine the cumulative area for z^* , which is

$$\text{area to the left of } z^* = 0.95 + 0.025 = 0.975$$

The cumulative area 0.975 appears in the 1.9 row and .06 column of Appendix Table 2, so $z^* = 1.96$. For the standard normal distribution, 95% of the variable values fall between -1.96 and 1.96 . The most extreme 5% are those values that are either greater than 1.96 or less than -1.96 .

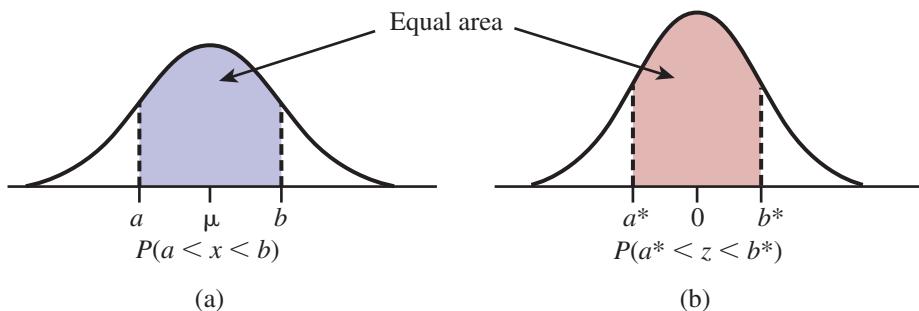
Other Normal Distributions

We now show how z curve areas can be used to calculate probabilities and to describe extreme values for any normal distribution. Remember that the letter z is reserved for those variables that have a standard normal distribution. The letter x is used more generally for any variable whose distribution is described by a normal curve with mean μ and standard deviation σ .

Suppose that we want to calculate $P(a < x < b)$, the probability that the variable x lies in a particular range. This probability corresponds to an area under a normal curve and above the interval from a to b , as shown in Figure 7.28(a).

FIGURE 7.28

Equality of nonstandard and standard normal curve areas.



The strategy for obtaining this probability is to find an equivalent problem involving the standard normal distribution. Finding an equivalent problem means determining an interval (a^*, b^*) that has the same probability for z (same area under the z curve) as the interval (a, b) in our original normal distribution (see Figure 7.28). The asterisk is used to distinguish a and b , the values for the original normal distribution with mean μ and standard deviation σ , from a^* and b^* , the values for the z curve.

To find a^* and b^* , we calculate z scores for the endpoints of the interval for which a probability is desired. This process is called *standardizing* the endpoints. For example, suppose that the variable x has a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 5$. To find

$$P(98 < x < 107)$$

we first translate this problem into an equivalent problem for the standard normal distribution.

Recall from Chapter 4 that a z score, or standardized score, tells how many standard deviations away from the mean a value lies. The z score is calculated by first subtracting the mean and then dividing by the standard deviation. Converting the lower endpoint $a = 98$ to a z score gives

$$a^* = \frac{98 - 100}{5} = \frac{-2}{5} = -0.40$$

and converting the upper endpoint yields

$$b^* = \frac{107 - 100}{5} = \frac{7}{5} = 1.40$$

Then

$$P(98 < x < 107) = P(-0.40 < z < 1.40)$$

The probability $P(-0.40 < z < 1.40)$ can now be evaluated using technology or Appendix Table 2.

Finding Normal Distribution Probabilities

To calculate probabilities for any normal distribution, standardize the appropriate interval endpoints and then use technology or the table of z curve areas. More specifically, if x is a variable whose behavior is described by a normal distribution with mean μ and standard deviation σ , then

$$\begin{aligned} P(x < b) &= P(z < b^*) \\ P(x > a) &= P(z > a^*) \\ P(a < x < b) &= P(a^* < z < b^*) \end{aligned}$$

where z is a variable whose distribution is standard normal and

$$a^* = \frac{a - \mu}{\sigma} \quad b^* = \frac{b - \mu}{\sigma}$$

Example 7.27 Newborn Birth Weights

Understand the context ➤

Data from the paper “Birth Weight Curves Tailored to Maternal World Region” (*Journal of Obstetrics and Gynaecology Canada* [2012]: 159–171) suggest that a normal distribution with mean $\mu = 3500$ grams and standard deviation $\sigma = 600$ grams is a reasonable model for the probability distribution of $x =$ birth weight of a randomly selected full-term baby born in Canada. What proportion of birth weights in Canada are between 2900 and 4700 grams?

To answer this question, we must find

$$P(2900 < x < 4700)$$

Do the work ➤

First, we translate the interval endpoints to equivalent endpoints for the standard normal distribution:

$$\begin{aligned} a^* &= \frac{a - \mu}{\sigma} = \frac{2900 - 3500}{600} = -1.00 \\ b^* &= \frac{b - \mu}{\sigma} = \frac{4700 - 3500}{600} = 2.00 \end{aligned}$$

Then

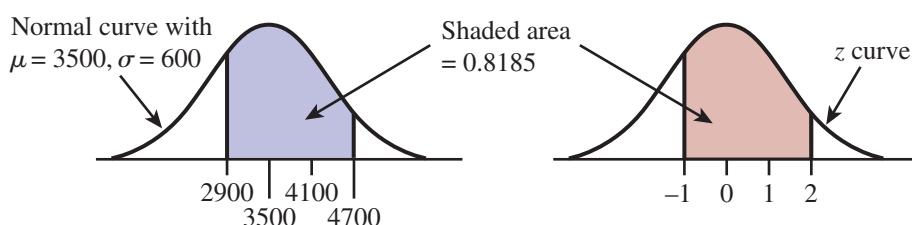
$$\begin{aligned} P(2900 < x < 4700) &= P(-1.00 < z < 2.00) \\ &= (z \text{ curve area to the left of } 2.00) \\ &\quad - (z \text{ curve area to the left of } -1.00) \\ &= 0.9772 - 0.1587 \\ &= 0.8185 \end{aligned}$$

Interpret the results ➤

The probabilities for x and z are shown in Figure 7.29. If birth weights were observed for many babies born in Canada, about 82% of them would fall between 2900 and 4700 grams.

FIGURE 7.29

$P(2900 < x < 4700)$ and corresponding z curve area for the birth weight distribution of Example 7.27.



What is the probability that a randomly selected baby born in Canada will have a birth weight greater than 4500? To evaluate $P(x > 4500)$, we first calculate

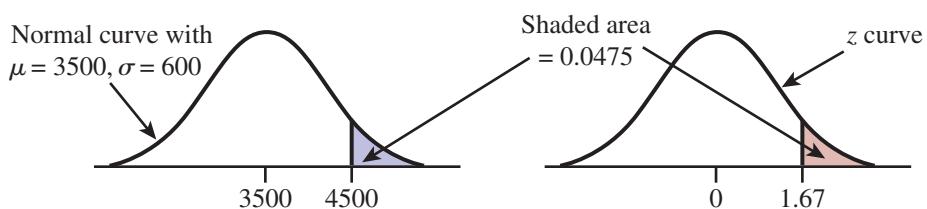
$$a^* = \frac{a - \mu}{\sigma} = \frac{4500 - 3500}{600} = 1.67$$

Then (see Figure 7.30)

$$\begin{aligned}
 P(x > 4500) &= P(z > 1.67) \\
 &= z \text{ curve area to the right of 1.67} \\
 &= 1 - (z \text{ curve area to the left of 1.67}) \\
 &= 1 - 0.9525 \\
 &= 0.0475
 \end{aligned}$$

FIGURE 7.30

$P(x > 4500)$ and corresponding z curve area for the birth weight distribution of Example 7.27.



Example 7.28 IQ Scores

Understand the context ➤

Although there is some controversy over the appropriateness of IQ scores as a measure of intelligence, IQ scores continue to be used for a variety of purposes. One commonly used IQ scale (the Stanford-Binet) has a mean of 100 and a standard deviation of 15, and IQ scores are approximately normally distributed. (IQ score is actually a discrete variable [because it is based on the number of correct responses on a test], but its population distribution closely resembles a normal curve.) If we define the random variable

x = IQ score of a randomly selected individual

then x has approximately a normal distribution with $\mu = 100$ and $\sigma = 15$.

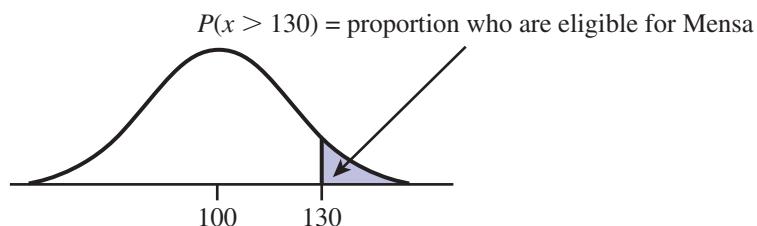
Do the work ➤

One way to become eligible for membership in Mensa, an organization advertised to be for those of high intelligence, is to have a Stanford-Binet IQ score above 130. What proportion of the population would qualify for Mensa membership? An answer to this question requires evaluating $P(x > 130)$. This probability is shown in Figure 7.31. With $a = 130$,

$$a^* = \frac{a - \mu}{\sigma} = \frac{130 - 100}{15} = 2.00$$

FIGURE 7.31

Normal distribution and desired proportion for Example 7.28.



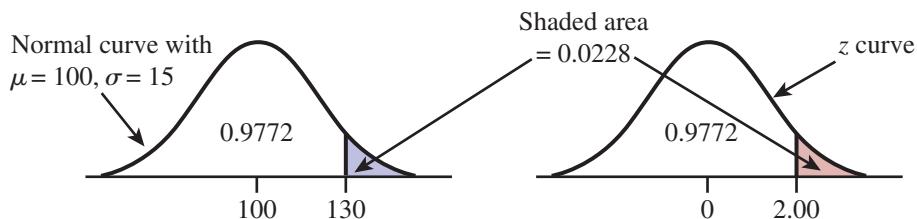
So (see Figure 7.32)

$$\begin{aligned}
 P(x > 130) &= P(z > 2.00) \\
 &= z \text{ curve area to the right of 2.00} \\
 &= 1 - (z \text{ curve area to the left of 2.00}) \\
 &= 1 - 0.9772 \\
 &= 0.0228
 \end{aligned}$$

Interpret the results ➤ Only 2.28% of the population would qualify for Mensa membership.

FIGURE 7.32

$P(x > 130)$ and corresponding z curve area for the IQ distribution of Example 7.28.



Suppose that we are interested in the proportion of the population with IQ scores below 80—that is, $P(x < 80)$. With $b = 80$,

$$b^* = \frac{b - \mu}{\sigma} = \frac{80 - 100}{15} = -1.33$$

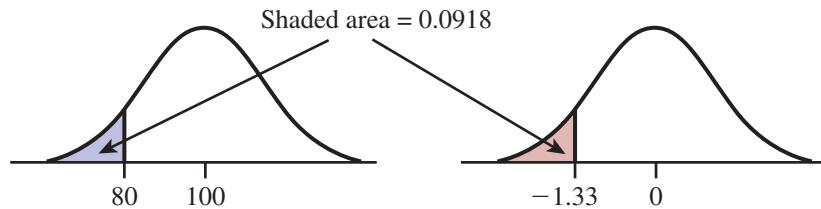
So

$$\begin{aligned} P(x < 80) &= P(z < -1.33) \\ &= z \text{ curve area to the left of } -1.33 \\ &= 0.0918 \end{aligned}$$

as shown in Figure 7.33. This probability (0.0918) tells us that just a little over 9% of the population has an IQ score below 80.

FIGURE 7.33

$P(x < 80)$ and corresponding z curve area for the IQ distribution of Example 7.28.



Now consider the proportion of the population with IQs between 75 and 125. Using $a = 75$ and $b = 125$, we calculate

$$a^* = \frac{75 - 100}{15} = -1.67 \quad b^* = \frac{125 - 100}{15} = 1.67$$

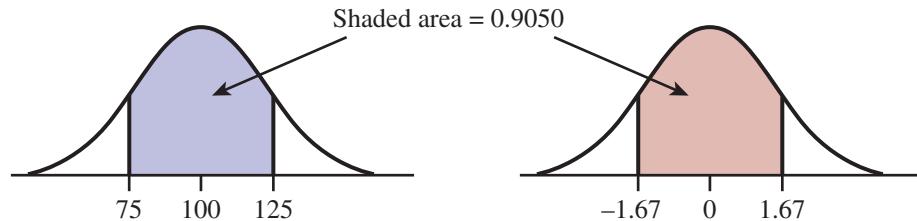
so

$$\begin{aligned} P(75 < x < 125) &= P(-1.67 < z < 1.67) \\ &= z \text{ curve area between } -1.67 \text{ and } 1.67 \\ &= (z \text{ curve area to the left of } 1.67) \\ &\quad - (z \text{ curve area to the left of } -1.67) \\ &= 0.9525 - 0.0475 \\ &= 0.9050 \end{aligned}$$

This is illustrated in Figure 7.34. The calculation tells us that 90.5% of the population has an IQ score between 75 and 125. Of the 9.5% whose IQ score is not between 75 and 125, half of them (4.75%) have scores over 125, and the other half have scores below 75.

FIGURE 7.34

$P(75 < x < 125)$ and corresponding z curve area for the IQ distribution of Example 7.28.



When we translate from a problem involving a normal distribution with mean μ and standard deviation σ to a problem involving the standard normal distribution, we convert to z scores:

$$z = \frac{x - \mu}{\sigma}$$

A z score can be interpreted as the distance of an x value from the mean in units of the standard deviation. For example, a z score of 1.4 corresponds to an x value that is 1.4 standard deviations above the mean, and a z score of -2.1 corresponds to an x value that is 2.1 standard deviations below the mean.

Suppose that we want to evaluate $P(x < 60)$ for a variable whose distribution is normal with $\mu = 50$ and $\sigma = 5$. Converting the endpoint 60 to a z score gives

$$z = \frac{60 - 50}{5} = 2$$

which tells us that the value 60 is 2 standard deviations above the mean. We then have

$$P(x < 60) = P(z < 2)$$

where z is a standard normal variable. Notice that for the standard normal distribution, the value 2 is also 2 standard deviations above the mean, because the mean is 0 and the standard deviation is 1. The value $z = 2$ is located the same distance (measured in standard deviations) from the mean of the standard normal distribution as is the value $x = 60$ from the mean in the normal distribution with $\mu = 50$ and $\sigma = 5$. This is why the translation using z scores results in an equivalent problem involving the standard normal distribution.

Describing Extreme Values in a Normal Distribution

To describe extreme values for a normal distribution with mean μ and standard deviation σ , we first solve a corresponding problem for the standard normal distribution and then translate our answer into one for the normal distribution of interest. This process is illustrated in Example 7.29.

Example 7.29 Registration Times

Understand the context ➤

Suppose data on the length of time required to complete registration for classes using an online registration system suggest that for students attending a particular university, the distribution of the variable

$$x = \text{time to register}$$

can be well approximated by a normal distribution with mean $\mu = 12$ minutes and standard deviation $\sigma = 2$ minutes. (The normal distribution might not be an appropriate model for $x = \text{time to register}$ at another university. There are many factors that might influence the shape, center, and variability of the registration time distribution.)

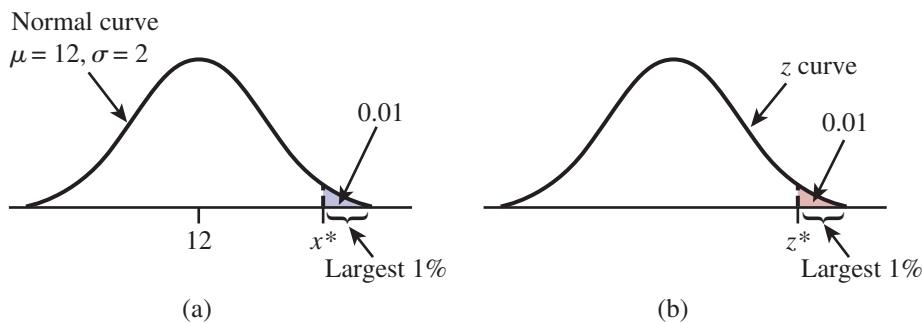
Because some students do not log off properly, the university would like to disconnect students automatically after some amount of time has elapsed. This time will be chosen so that only 1% of the students are disconnected while they are still attempting to register. To determine the amount of time that should be allowed before disconnecting a student, we need to describe the largest 1% of the registration times. These are the individuals who will be mistakenly disconnected. This is illustrated in Figure 7.35(a). To determine the value of x^* , we first solve the analogous problem for the standard normal distribution, as shown in Figure 7.35(b).

Do the work ➤

By looking at Appendix Table 2 for a cumulative area of 0.99, we find the closest entry (0.9901) in the 2.3 row and the .03 column, from which $z^* = 2.33$. For the standard normal distribution, the largest 1% of the distribution is made up of those values greater than 2.33.

FIGURE 7.35

Capturing the largest 1% of the registration times.



An equivalent statement is that the largest 1% are those with z scores greater than 2.33. This implies that in the distribution of x = time to register (or any other normal distribution), the largest 1% are those values with z scores greater than 2.33 or, equivalently, those x values more than 2.33 standard deviations above the mean. Here, the standard deviation is 2, so 2.33 standard deviations is 2.33(2), and it follows that

$$x^* = 12 + 2.33(2) = 12 + 4.66 = 16.66$$

- Interpret the results ➤ The largest 1% of the distribution for time to register is made up of values that are greater than 16.66 minutes. If the university system was to disconnect students after 16.66 minutes, only 1% of the students registering would be disconnected before completing their registration.

A general formula for converting a z score back to an x value results from solving

$$z^* = \frac{x^* - \mu}{\sigma} \text{ for } x^*, \text{ as shown in the accompanying box.}$$

To convert a z score z^* back to an x value x^* , use

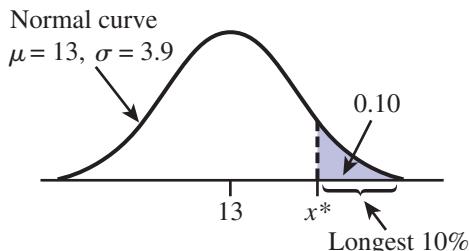
$$x^* = \mu + z^*\sigma$$

Example 7.30 Garbage Truck Processing Times

Understand the context ➤

Garbage trucks entering a waste management facility are weighed and then they offload garbage into a landfill. Data from the paper “[Estimating Waste Transfer Station Delays Using GPS](#)” ([Waste Management \[2008\]: 1742–1750](#)) suggest that a normal distribution with mean $\mu = 13$ minutes and $\sigma = 3.9$ minutes is a reasonable model for the probability distribution of the random variable x = total processing time for a garbage truck at the waste management facility (total processing time includes waiting time as well as the time required to weigh the truck and offload the garbage).

Suppose that we want to describe the total processing times of the trucks making up the 10% with the longest processing times. These trucks would be the 10% with times corresponding to the shaded region in the accompanying illustration.



Do the work ➤

For the standard normal distribution, the largest 10% are those with z values greater than $z^* = 1.28$ (from Appendix Table 2, based on a cumulative area of 0.90).

Then

$$\begin{aligned}x^* &= \mu + z^* \sigma \\&= 13 + 1.28(3.9) \\&= 13 + 4.992 \\&= 17.992\end{aligned}$$

Interpret the results ➤ About 10% of the garbage trucks using this facility would have a total processing time of more than 17.992 minutes.

The 5% with the fastest processing times would be those with z values less than $z^* = -1.645$ (from Appendix Table 2, based on a cumulative area of 0.05). Then

$$\begin{aligned}x^* &= \mu + z^* \sigma \\&= 13 + (-1.645)(3.9) \\&= 13 - 6.416 \\&= 6.584\end{aligned}$$

About 5% of the garbage trucks processed at this facility will have total processing times of less than 6.584 minutes.

EXERCISES 7.70 - 7.96

- 7.70** Determine the following standard normal (z) curve areas. (Hint: See Examples 7.23 and 7.24.)
- The area under the z curve to the left of 1.75
 - The area under the z curve to the left of -0.68
 - The area under the z curve to the right of 1.20
 - The area under the z curve to the right of -2.82
- 7.71** Determine the following standard normal (z) curve areas.
- The area under the z curve between -2.22 and 0.53
 - The area under the z curve between -1 and 1
 - The area under the z curve between -4 and 4
- 7.72** Determine each of the following areas under the standard normal (z) curve:
- To the left of -1.28
 - To the right of 1.28
 - Between -1 and 2
- 7.73** Determine each of the following areas under the standard normal (z) curve:
- To the right of 0
 - To the right of -5
 - Between -1.6 and 2.5
 - To the left of 0.23
- 7.74** Let z denote a random variable that has a standard normal distribution. Determine each of the following probabilities:
- $P(z < 2.36)$
 - $P(z \leq 2.36)$
 - $P(z < -1.23)$
 - $P(1.14 < z < 3.35)$
- 7.75** Let z denote a random variable that has a standard normal distribution. Determine each of the following probabilities:
- $P(-0.77 \leq z \leq -0.55)$
 - $P(z > 2)$
 - $P(z \geq -3.38)$
 - $P(z < 4.98)$
- 7.76** Let z denote a random variable having a normal distribution with $\mu = 0$ and $\sigma = 1$. Determine each of the following probabilities. (Hint: See Examples 7.27 and 7.28.)
- $P(z < 0.10)$
 - $P(z < -0.10)$
 - $P(0.40 < z < 0.85)$
- 7.77** Let z denote a random variable having a normal distribution with $\mu = 0$ and $\sigma = 1$. Determine each of the following probabilities.
- $P(-0.85 < z < -0.40)$
 - $P(-0.40 < z < 0.85)$
 - $P(z > -1.25)$
 - $P(z < -1.50 \text{ or } z > 2.50)$
- 7.78** Let z denote a variable that has a standard normal distribution. Determine the value z^* to satisfy the following conditions. (Hint: See Example 7.25.)
- $P(z < z^*) = 0.025$
 - $P(z < z^*) = 0.01$
 - $P(z < z^*) = 0.05$
 - $P(z > z^*) = 0.02$
 - $P(z > z^*) = 0.01$
 - $P(z > z^* \text{ or } z < -z^*) = 0.20$

7.79 Determine the value z^* that

- Separates the largest 3% of all z values from the others (Hint: See Example 7.26.)
- Separates the largest 1% of all z values from the others
- Separates the smallest 4% of all z values from the others
- Separates the smallest 10% of all z values from the others

7.80 Determine the value of z^* such that

- $-z^*$ and z^* separate the middle 95% of all z values from the most extreme 5% (Hint: See Example 7.26.)
- $-z^*$ and z^* separate the middle 90% of all z values from the most extreme 10%
- $-z^*$ and z^* separate the middle 98% of all z values from the most extreme 2%
- $-z^*$ and z^* separate the middle 92% of all z values from the most extreme 8%

7.81 Because $P(z < 0.44) = 0.67$, 67% of all z values are less than 0.44, and 0.44 is the 67th percentile of the standard normal distribution. Determine the value of each of the following percentiles for the standard normal distribution. (Hint: If the cumulative area that you must look for does not appear in the z table, use the closest entry, see Example 7.27.)

- The 91st percentile (Hint: Look for area 0.9100.)
- The 77th percentile
- The 50th percentile
- The 9th percentile
- What is the relationship between the 70th z percentile and the 30th z percentile?

7.82 Consider the population of all 1-gallon cans of dusty rose paint manufactured by a particular paint company. Suppose that a normal distribution with mean $\mu = 5$ ml and standard deviation $\sigma = 0.2$ ml is a reasonable model for the distribution of the variable x = amount of red dye in the paint mixture. Use the normal distribution model to calculate the following probabilities. (Hint: See Examples 7.27 and 7.28.)

- | | |
|--------------------|-----------------------|
| a. $P(x < 5.0)$ | b. $P(x < 5.4)$ |
| c. $P(x \leq 5.4)$ | d. $P(4.6 < x < 5.2)$ |
| e. $P(x > 4.5)$ | f. $P(x > 4.0)$ |

7.83 Consider babies born in the “normal” range of 37–43 weeks gestational age. The paper referenced in Example 7.27 “Birth Weight Curves Tailored to Maternal World Region” (*Journal of Obstetrics and Gynaecology Canada* [2012]: 159–171) suggests that a normal distribution with mean $\mu = 3500$ grams and standard deviation $\sigma = 600$ grams is a reasonable model for the probability distribution of the variable x = birth weight of a randomly selected full-term baby born in Canada.

a. What is the probability that the birth weight of a randomly selected full-term baby born in Canada exceeds 4000 g?

b. What is the probability that the birth weight of a randomly selected full-term baby born in Canada is between 3000 and 4000 g?

c. What is the probability that the birth weight of a randomly selected full-term baby born in Canada is either less than 2000 g or greater than 5000 g?

7.84 Use the information on birth weights for babies born in Canada given in the previous exercise to answer the following questions.

- What is the probability that the birth weight of a randomly selected full-term baby born in Canada exceeds 7 pounds? (Hint: 1 lb = 453.59 g.)
- How would you characterize the most extreme 0.1% of all full-term baby birth weights for babies born in Canada?
- If x is a random variable with a normal distribution and a is a numerical constant ($a \neq 0$), then $y = ax$ also has a normal distribution. Use this formula to determine the distribution of full-term baby birth weight expressed in pounds (shape, mean, and standard deviation), and then recalculate the probability from Part (a). How does this compare to your previous answer?

7.85 Emissions of nitrogen oxides, which are major constituents of smog, can be modeled using a normal distribution. Let x denote the amount of this pollutant emitted (in parts per billion) by a randomly selected vehicle. Suppose the distribution of x can be described by a normal distribution with $\mu = 1.6$ and $\sigma = 0.4$. A city wants to offer some sort of incentive to get the worst polluters off the road. What emission levels constitute the worst 10% of the vehicles?

7.86 The paper referenced in Example 7.30 (“Estimating Waste Transfer Station Delays Using GPS,” *Waste Management* [2008]: 1742–1750) describing processing times for garbage trucks also provided information on processing times at a second facility. At this second facility, the mean total processing time was 9.9 minutes and the standard deviation of the processing times was 6.2 minutes. Explain why a normal distribution with mean 9.9 and standard deviation 6.2 would not be an appropriate model for the probability distribution of the variable x = total processing time of a randomly selected truck entering this facility.

7.87 The size of the left upper chamber of the heart is one measure of cardiovascular health. When the upper

left chamber is enlarged, the risk of heart problems is increased. The paper “**Left Atrial Size Increases with Body Mass Index in Children**” (*International Journal of Cardiology* [2009]: 1–7) described a study in which the left atrial size was measured for a large number of children age 5 to 15 years. Based on these data, the authors concluded that for healthy children, left atrial diameter was approximately normally distributed with a mean of 26.4 mm and a standard deviation of 4.2 mm.

- a. Approximately what proportion of healthy children have left atrial diameters less than 24 mm?
- b. Approximately what proportion of healthy children have left atrial diameters greater than 32 mm?
- c. Approximately what proportion of healthy children have left atrial diameters between 25 and 30 mm?
- d. For healthy children, what is the value for which only about 20% have a larger left atrial diameter?

7.88 The paper referenced in the previous exercise also included data on left atrial diameter for children who were considered overweight. For these children, left atrial diameter was approximately normally distributed with a mean of 28 mm and a standard deviation of 4.7 mm.

- a. Approximately what proportion of overweight children have left atrial diameters less than 25 mm?
- b. Approximately what proportion of overweight children have left atrial diameters greater than 32 mm?
- c. Approximately what proportion of overweight children have left atrial diameters between 25 and 30 mm?
- d. What proportion of overweight children has left atrial diameters greater than the mean for healthy children?

7.89 The article “**New York City’s Graffiti-Removal Response Time Rises**” (*The Wall Street Journal*, September 16, 2016, wsj.com/articles/new-york-citys-graffiti-removal-response-time-rises-1473287392, retrieved May 1, 2017) states that the city took an average of 114 days to handle graffiti complaints in 2015. Suppose that the response time is approximately normally distributed with a mean of 114 days and a standard deviation of 20 days.

- a. Approximately what proportion of graffiti removal requests are handled within 60 days?
- b. Approximately what proportion of graffiti removal requests take more than 120 days?

7.90 A machine that cuts corks for wine bottles operates in such a way that the distribution of the diameter

for the corks produced is well approximated by a normal distribution with mean 3 cm and standard deviation 0.1 cm. The specifications call for corks with diameters between 2.9 and 3.1 cm. A cork not meeting the specifications is considered defective. (A cork that is too small leaks and causes the wine to deteriorate. A cork that is too large doesn’t fit in the bottle.) What proportion of corks produced by this machine are defective?

7.91 Refer to the previous exercise. Suppose that there are two machines available for cutting corks. The machine described in the preceding problem produces corks with diameters that are approximately normally distributed with mean 3 cm and standard deviation 0.1 cm. The second machine produces corks with diameters that are approximately normally distributed with mean 3.05 cm and standard deviation 0.01 cm. Which machine would you recommend? (Hint: Which machine would produce fewer defective corks?)

7.92 Purchases made at small “corner stores” were studied by the authors of the paper “**Changes in Quantity, Spending, and Nutritional Characteristics of Adult, Adolescent and Child Urban Corner Store Purchases After an Environmental Intervention**” (*Preventive Medicine* [2015]: 81–85). Corner stores were defined as stores that are less than 200 square feet in size, have only one cash register, and primarily sell food. After observing a large number of corner store purchases in Philadelphia, the authors reported that the average number of grams of fat in a corner store purchase was 21.1. Suppose that the variable x = number of grams of fat in a corner store purchase has a distribution that is approximately normal with a mean of 21.1 grams and a standard deviation of 7 grams.

- a. What is the probability that a randomly selected corner store purchase has more than 30 grams of fat?
- b. What is the probability that a randomly selected corner store purchase has between 15 and 25 grams of fat?
- c. If two corner store purchases are randomly selected, what is the probability that both of these purchases will have more than 25 grams of fat?

7.93 The time that it takes a randomly selected job applicant to perform a certain task has a distribution that can be approximated by a normal distribution with a mean value of 120 seconds and a standard deviation of 20 seconds. The fastest 10% are to be given advanced training. What task times qualify individuals for such training?

- 7.94** Suppose that the distribution of typing speed in words per minute (wpm) for experienced typists using a new type of split keyboard can be approximated by a normal curve with mean 60 wpm and standard deviation 15 wpm ([“The Effects of Split Keyboard Geometry on Upper body Postures,” Ergonomics \[2009\]: 104–111](#)).
- What is the probability that a randomly selected typist’s speed is at most 60 wpm?
 - What is the probability that a randomly selected typist’s speed is less than 60 wpm?
 - What is the probability that a randomly selected typist’s speed is between 45 and 90 wpm?

- 7.95** Consider the typing speed distribution described in the previous exercise. Would you be surprised to find a typist in this population whose speed exceeded 105 wpm?

- 7.96** Consider the typing speed distribution described in Exercise 7.94.
- Suppose that two typists are independently selected. What is the probability that both their typing speeds exceed 75 wpm?
 - Suppose that special training is to be made available to the slowest 20% of the typists. What typing speeds would qualify individuals for this training?

SECTION 7.7 Checking for Normality and Normalizing Transformations

Some of the most frequently used statistical methods are valid only when the sample has been selected from a population distribution that is at least approximately normal. One way to determine if it is reasonable to assume that the population distribution is approximately normal is to construct a **normal probability plot** of the sample data.

A normal probability plot uses quantities called **normal scores**. The values of the normal scores depend on the sample size n . For example, the normal scores when $n = 10$ are as follows:

−1.54	−1.00	−0.66	−0.38	−0.12
0.12	0.38	0.66	1.00	1.54

To interpret these numbers, think of selecting sample after sample from a standard normal distribution, each one consisting of $n = 10$ observations. Then −1.54 is the long-run average of the smallest observation from each sample, −1.00 is the long-run average of the second smallest observation from each sample, and so on.

Tables of normal scores for many different sample sizes are available. Alternatively, many statistical software packages (such as Minitab and JMP) and some graphing calculators can compute these scores on request and then use them to construct a normal probability plot. Not all calculators and software packages use the same algorithm to compute normal scores. However, these differences in normal scores do not change the overall character of a normal probability plot, so either values from a table or those provided by software or a calculator can be used.

After the sample observations are ordered from smallest to largest, the smallest normal score is paired with the smallest observation in the sample, the second smallest normal score with the second smallest observation in the sample, and so on. The first number in a pair is the normal score, and the second number in the pair is the observed data value. A normal probability plot is a scatterplot of these (normal score, observed value) pairs.

If the sample has been selected from a population with a *standard* normal distribution, the second number in each pair should be reasonably close to the first number (ordered observation \approx corresponding normal score). Then the n plotted points will fall near a line with slope equal to 1 (a 45° line) passing through $(0, 0)$. When the sample has been selected from *some* normal population distribution (but not necessarily the standard normal distribution), the plotted points will be close to *some* straight line (but not necessarily one with slope 1 and intercept 0).

DEFINITION

Normal probability plot: A scatterplot of the (normal score, observed value) pairs.

A strong linear pattern in a normal probability plot suggests it is reasonable to think that the population distribution is normal. On the other hand, systematic departure from a straight-line pattern (such as curvature in the plot) indicates that it is not reasonable to assume that the population distribution is normal.

Example 7.31 Egg Weights

Understand the context ➤

The following data represent egg weights (in grams) for a sample of 10 eggs. These data are consistent with summary quantities in the paper “Evaluation of Egg Quality Traits of Chickens Reared under Backyard System in Western Uttar Pradesh” (*Indian Journal of Poultry Science*, 2009).

53.04 53.50 52.53 53.00 53.07 52.86 52.66 53.23 53.26 53.16

Do the work ➤

Arranging the sample observations in order from smallest to largest results in

52.53 52.66 52.86 53.00 53.04 53.07 53.16 53.23 53.26 53.50

Pairing these ordered observations with the normal scores for a sample of size 10 (previously given) results in the following 10 pairs that can be used to construct the normal probability plot:

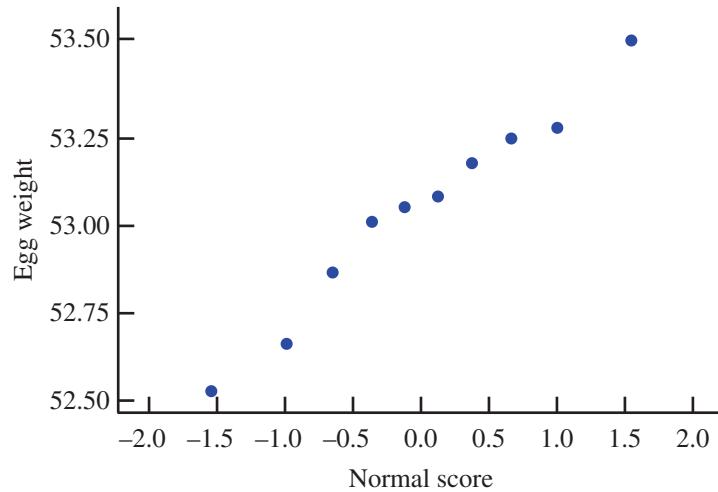
(−1.54, 52.53)	(−1.00, 52.66)
(−0.66, 52.86)	(−0.38, 53.00)
(−0.12, 53.04)	(0.12, 53.07)
(0.38, 53.16)	(0.66, 53.23)
(1.00, 53.26)	(1.54, 53.50)

Interpret the results ➤

The normal probability plot is shown in Figure 7.36. The linear pattern in the plot suggests that it is reasonable to think that the egg-weight distribution from which these observations were selected is approximately normal.

FIGURE 7.36

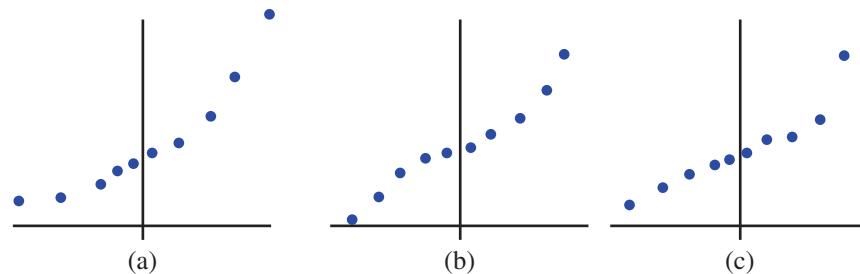
A normal probability plot for data of Example 7.31.



Deciding whether a normal probability plot shows a strong linear pattern is sometimes a matter of opinion. Particularly when n is small, normality should not be ruled out unless the pattern is clearly nonlinear. Figure 7.37 displays several plots that suggest that the population distribution may not be normal.

FIGURE 7.37

Plots suggesting nonnormality:
 (a) indication that the population distribution is skewed;
 (b) indication that the population distribution has heavier tails than a normal curve;
 (c) presence of an outlier.



Transforming Data to Obtain a Distribution That Is Approximately Normal (Optional)

Many statistical methods are valid only when the sample is selected at random from a population whose distribution is at least approximately normal. When a sample histogram shows a distinctly nonnormal shape, it is common to use a transformation or reexpression of the data.

By *transforming* data, we mean applying some specified mathematical function (such as the square root, logarithm, or reciprocal) to each data value to produce a set of transformed data. We then study and summarize the distribution of these transformed values.

We saw in Chapter 5 that, with bivariate data, one or both of the variables can be transformed in an attempt to find two variables that are linearly related. With univariate data, a transformation is usually chosen to produce a distribution of transformed values that is more nearly symmetric and more closely approximated by a normal distribution than the original distribution.

Example 7.32 Rainfall Data

Understand the context ➤

- Data that have been used by several investigators to introduce the concept of transformation consist of values of March precipitation for Minneapolis–St. Paul over a period of 30 years. These values are given in Table 7.2, along with the square root of each value. Histograms of both the original and the transformed data appear in Figure 7.38.

The distribution of the original data is clearly skewed, with a long upper tail. The square-root transformation results in a distribution that is more symmetric, with a typical value of around 1.1.

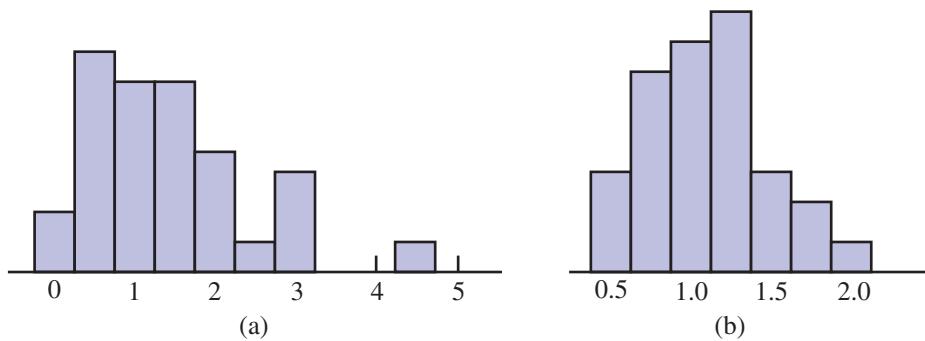
TABLE 7.2 Original and Square-Root-Transformed Values of March Precipitation in Minneapolis–St. Paul over a 30-year Period

Year	Precipitation	$\sqrt{\text{Precipitation}}$	Year	Precipitation	$\sqrt{\text{Precipitation}}$
1	0.77	0.88	16	1.62	1.27
2	1.74	1.32	17	1.31	1.14
3	0.81	0.90	18	0.32	0.57
4	1.20	1.10	19	0.59	0.77
5	1.95	1.40	20	0.81	0.90
6	1.20	1.10	21	2.81	1.68
7	0.47	0.69	22	1.87	1.37
8	1.43	1.20	23	1.18	1.09
9	3.37	1.84	24	1.35	1.16
10	2.20	1.48	25	4.75	2.18
11	3.00	1.73	26	2.48	1.57
12	3.09	1.76	27	0.96	0.98
13	1.51	1.23	28	1.89	1.37
14	2.10	1.45	29	0.90	0.95
15	0.52	0.72	30	2.05	1.43

● Data set available online

FIGURE 7.38

Histograms of the precipitation data used in Example 7.32:
 (a) untransformed data;
 (b) square-root transformed data.



Logarithmic transformations are also common and, as with bivariate data, either the natural logarithm or the base 10 logarithm can be used. A logarithmic transformation is often used with data that are positively skewed (a long upper tail). This affects values in the upper tail substantially more than values in the lower tail, resulting in a more symmetric—and often more nearly normal—distribution.

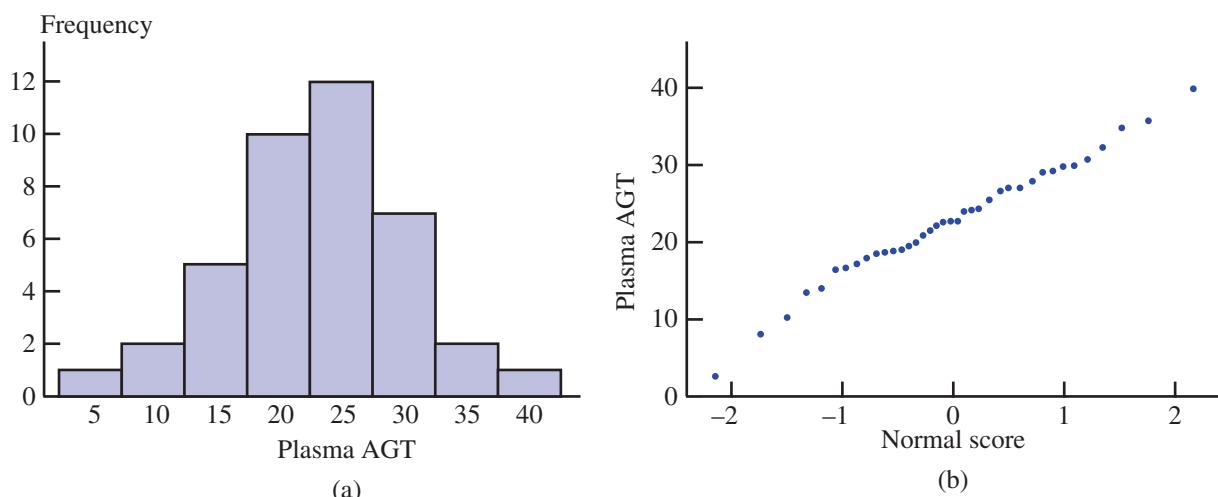
Example 7.33 Markers for Kidney Disease

Two measures of kidney function are the levels of a substance called AGT found in blood and in urine. The paper “[Urinary Angiotensinogen as a Potential Biomarker of Severity of Chronic Kidney Diseases](#)” (*Journal of the American Society of Hypertension* [2008]: 349–354) describes a study in which blood plasma AGT levels and urinary AGT levels were measured for a sample of adults with chronic kidney disease. Representative data (consistent with summary quantities and descriptions given in the paper) for 40 patients are given in Table 7.3.

TABLE 7.3 Plasma and Urinary AGT Levels

Plasma AGT	Plasma AGT	Urinary AGT	Urinary AGT
21.0	16.7	56.2	41.7
36.0	20.2	288.4	29.5
22.9	24.5	45.7	208.9
8.0	18.5	426.6	229.1
27.3	40.2	190.6	186.2
32.4	18.8	616.6	29.5
17.2	28.1	97.7	229.1
30.9	26.8	66.1	13.5
27.2	24.1	2.6	407.4
30.0	14.1	74.1	1122.0
35.1	18.9	14.5	66.1
21.6	25.6	56.2	7.4
22.7	10.2	812.8	177.8
2.5	29.2	11.5	6.2
30.2	29.5	346.7	67.6
27.3	24.3	9.6	20.0
19.6	22.3	288.4	28.8
19.0	16.5	147.9	186.2
13.4	25.6	17.0	141.3
18.0	23.0	575.4	724.4

The authors of the paper stated that the distribution of plasma AGT levels was approximately normal. Minitab was used to construct the histogram and normal probability plot for the plasma AGT levels shown in Figure 7.39. The histogram is

**FIGURE 7.39**

Graphical displays for the plasma AGT data of Example 7.33:

- (a) histogram;
- (b) normal probability plot.

reasonably symmetric and the normal probability plot shows a strong linear pattern. This is consistent with the authors' statement about the approximate normality of the plasma AGT levels.

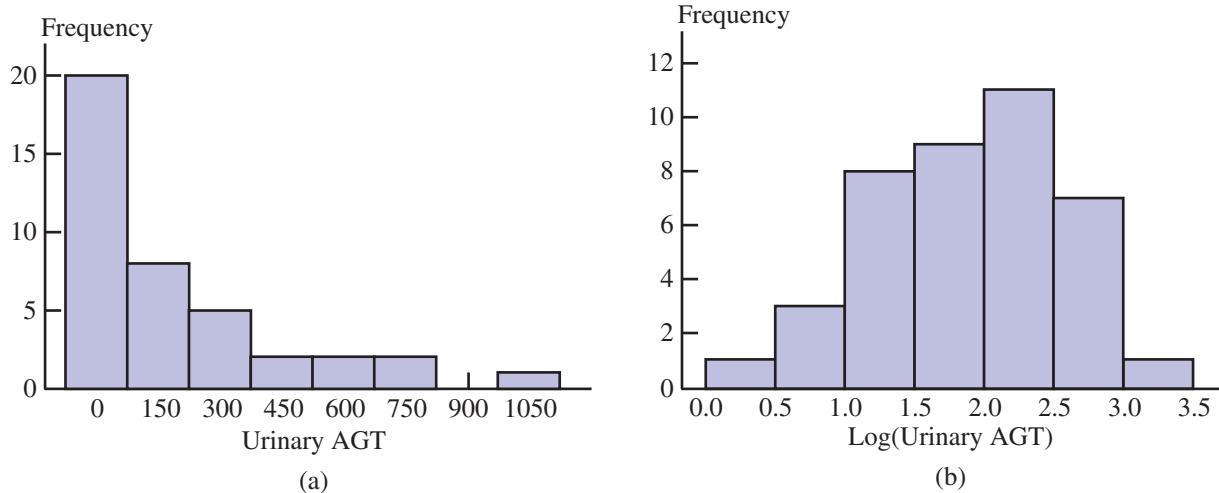
When the authors considered urinary AGT levels, they found that the distribution of the sample data was skewed, and they used a log transformation in order to obtain a distribution that was more approximately normal. Table 7.4 gives the urinary AGT levels along with the log-transformed data. Figure 7.40 shows histograms of the original urinary AGT data and the transformed urinary AGT data. Notice that the histogram for the transformed data is more nearly symmetric and more mound shaped than the histogram of the untransformed data.

TABLE 7.4 Urinary AGT Levels and Log-Transformed Levels

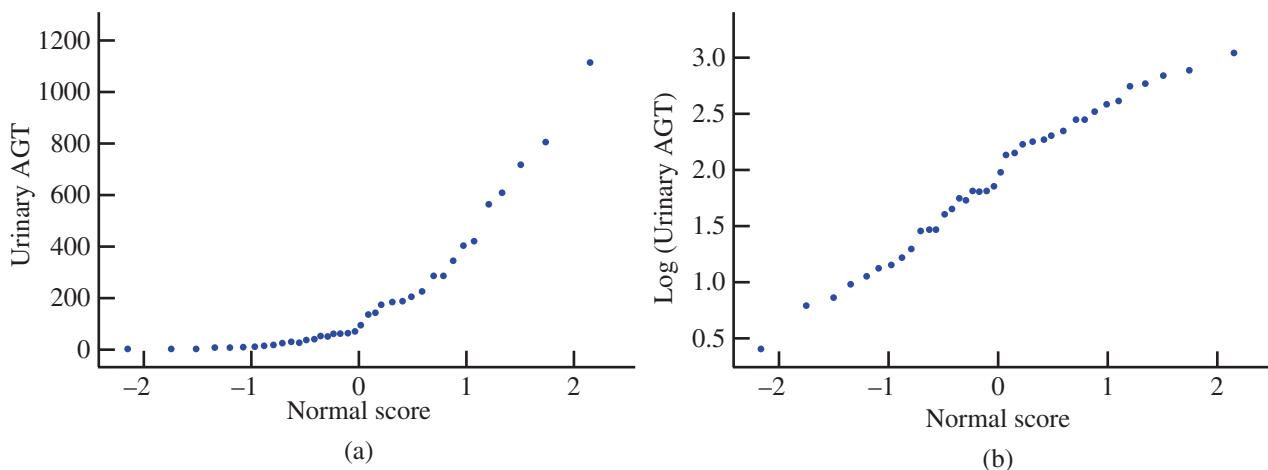
Urinary AGT	Log(Urinary AGT)	Urinary AGT	Log(Urinary AGT)
56.2	1.75	41.7	1.62
288.4	2.46	29.5	1.47
45.7	1.66	208.9	2.32
426.6	2.63	229.1	2.36
190.6	2.28	186.2	2.27
616.6	2.79	29.5	1.47
97.7	1.99	229.1	2.36
66.1	1.82	13.5	1.13
2.6	0.41	407.4	2.61
74.1	1.87	1122.0	3.05
14.5	1.16	66.1	1.82
56.2	1.75	7.4	0.87
812.8	2.91	177.8	2.25
11.5	1.06	6.2	0.79
346.7	2.54	67.6	1.83
9.6	0.98	20.0	1.30
288.4	2.46	28.8	1.46
147.9	2.17	186.2	2.27
17.0	1.23	141.3	2.15
575.4	2.76	724.4	2.86

FIGURE 7.40

Histograms of urinary AGT data from Example 7.33:
 (a) untransformed data;
 (b) transformed data.

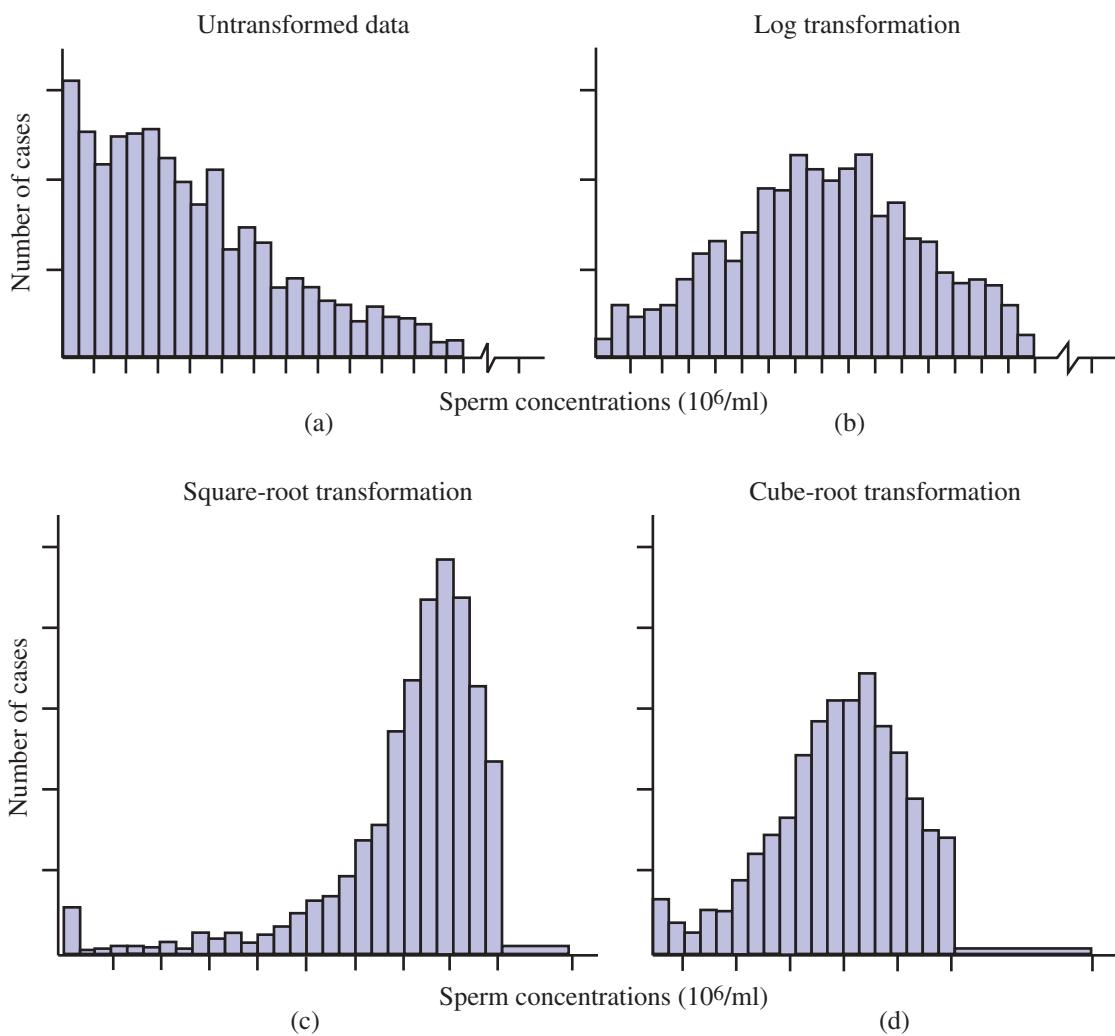
**FIGURE 7.41**

Minitab normal probability plots for the urinary AGT data of
 (a) original data;
 (b) transformed data.



Selecting a Transformation

When selecting a transformation, it is common to try several different transformations to find one that is satisfactory. Figure 7.42, from the article “**Distribution of Sperm Counts in Suspected Infertile Men**” (*Journal of Reproduction and Fertility* [1983]: 91–96), shows what can result from such a search. Investigators in this field have used all three of the transformations illustrated, but the log transformation shown in Figure 7.42(b) appears to be the most successful in creating a distribution that is approximately normal.

**FIGURE 7.42**

Histograms of sperm concentrations for 1711 suspected infertile men:

- (a) untransformed data (highly skewed);
- (b) log-transformed data (reasonably symmetric);
- (c) square-root-transformed data;
- (d) cube-root-transformed data.

EXERCISES 7.97 - 7.108

● Data set available online

7.97 ● The authors of the paper “[Development of Nutritionally At-Risk Young Children is Predicted by Malaria, Anemia, and Stunting in Pemba, Zanzibar](#)” (*The Journal of Nutrition* [2009]:763–772) studied factors that might be related to dietary deficiencies in children. Children were observed for a length of time and the time spent in various activities was recorded. One variable of interest was the length of time (in minutes) a child spent fussing.

The authors comment that the distribution of fussing times was skewed and that they used a square root transformation to create a distribution that was more approximately normal. Data consistent with summary quantities in the paper for 15 children are given in the accompanying table.

Normal scores for a samples size of 15 are also given.

Fussing Time	Normal Score
0.05	-1.739
0.10	-1.245
0.15	-0.946
0.40	-0.714
0.70	-0.515
1.05	-0.333
1.95	-0.165
2.15	0.000
3.70	0.165
3.90	0.335

(continued)

Fussing Time	Normal Score
4.50	0.515
6.00	0.714
8.00	0.946
11.00	1.245
14.00	1.739

- a. Construct a normal probability plot for the fussing time data. (Hint: See Example 7.31.)
- b. Does the plot from Part (a) look linear? Do you agree with the authors of the paper that the fussing time distribution is not normal?
- c. Transform the data by taking the square root of each data value. Construct a normal probability plot for the square root transformed data.
- d. How do the normal probability plots from Parts (a) and (c) compare?

7.98 ● The paper “Risk Behavior, Decision Making, and Music Genre in Adolescent Males” (Marshall University, May 2009) examined the effect of type of music playing and performance on a risky, decision-making task.

- a. Participants in the study responded to a questionnaire that was used to assign a risk behavior score. Risk behavior scores (read from a graph that appeared in the paper) for 15 participants follow. Use these data to construct a normal probability plot (the normal scores for a sample of size 15 are given in the previous exercise).

102 105 113 120 125 127 134 135
139 141 144 145 149 150 160

- b. Participants also completed a positive and negative affect scale (PANAS) designed to measure emotional response to music. PANAS values (read from a graph that appeared in the paper) for 15 participants follow. Use these data to construct a normal probability plot (the normal scores for a sample of size 15 are given in the previous exercise).

36 40 45 47 48 49 50 52
53 54 56 59 61 62 70

- c. The author of the paper states that he believes that it is reasonable to consider both risk behavior scores and PANAS scores to be approximately normally distributed. Do the normal probability plots from Parts (a) and (b) support this conclusion? Explain.

7.99 ● Measures of nerve conductivity are used in the diagnosis of certain medical conditions. The paper “Effects of Age, Gender, Height, and Weight on Late Responses and Nerve Conduction Study Parameters” (Acta Neurologica Taiwanica [2009]: 242–249) describes a study in which the ulnar nerve

was stimulated in healthy patients and the amplitude and velocity of the response was measured.

Representative data (consistent with summary quantities and descriptions given in the paper) for 30 patients for the variable x = response velocity (m/s) are given in the accompanying table. Also given are values of the log of x and the square root of x .

x	$\log(x)$	\sqrt{x}
60.1	1.78	7.75
48.7	1.69	6.98
51.7	1.71	7.19
52.9	1.72	7.27
50.5	1.70	7.11
58.5	1.77	7.65
53.6	1.73	7.32
60.3	1.78	7.77
64.5	1.81	8.03
50.4	1.70	7.10
56.5	1.75	7.52
55.5	1.74	7.45
53.0	1.72	7.28
50.5	1.70	7.11
54.0	1.73	7.35
53.6	1.73	7.32
55.2	1.74	7.43
57.9	1.76	7.61
61.5	1.79	7.84
58.0	1.76	7.62
57.6	1.76	7.59
67.1	1.83	8.19
56.2	1.75	7.50
53.8	1.73	7.33
55.7	1.75	7.46
52.9	1.72	7.27
54.0	1.73	7.35
52.6	1.72	7.25
61.8	1.79	7.86
62.8	1.80	7.92

- a. Construct a histogram of the untransformed data.
- b. Does the distribution of x appear to be approximately normal? Explain.
- c. Construct a histogram of the log-transformed data.
- d. Is the histogram of the log-transformed data more nearly symmetric than the histogram of the untransformed data?

7.100 Use the information given in the previous question to answer the following questions.

- a. Construct a histogram of the square root transformed data.
- b. Do either of the two transformations (square root or log) result in a histogram that is more nearly normal in shape? (Hint: See Example 7.33.)

- 7.101** • Macular degeneration is the most common cause of blindness in people older than 60 years. One variable thought to be related to a type of inflammation associated with this disease is level of a substance called soluble Fas ligand (sFasL) in the blood.

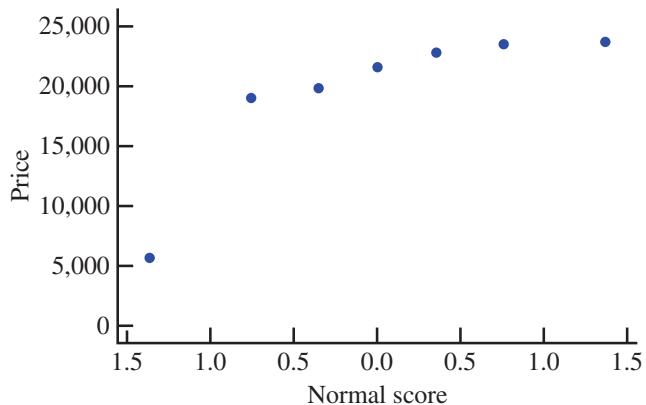
The accompanying table contains representative data on $x = \text{sFasL}$ level for 10 patients with age-related macular degeneration. These data are consistent with summary quantities and descriptions of the data given in the paper

"Associations of Plasma-Soluble Fas Ligand with Aging and Age-Related Macular Degeneration"
(Investigative Ophthalmology & Visual Science [2008]: 1345–1349).

The authors of the paper noted that the distribution of sFasL level was skewed and recommended a cube-root transformation. The cube-root values and the normal scores for a sample size of 10 are also given in the accompanying table.

x	Cube Root of x	Normal Score
0.069	0.41	-1.539
0.074	0.42	-1.001
0.176	0.56	-0.656
0.185	0.57	-0.376
0.216	0.60	-0.123
0.287	0.66	0.123
0.343	0.70	0.376
0.343	0.70	0.656
0.512	0.80	1.001
0.729	0.90	1.539

- a.** Construct a normal probability plot using the untransformed data.
b. Does the normal probability plot for the untransformed data appear linear or curved?
c. Construct a normal probability plot using the cube-root transformed data. Does the normal probability plot appear more nearly linear than the plot for the untransformed data?
- 7.102** The following normal probability plot was constructed using data on the price of seven 2015 Honda Accords with automatic transmissions that were listed for sale within 25 miles of the zip code 19383 ([from autotrader.com, search conducted on September 24, 2016](#)). For purposes of this exercise, you may assume that this sample is representative of 2015 Honda Accord prices in this area. Based on the normal probability plot, is it reasonable to think that the distribution of 2015 Honda Accord prices in this area is approximately normal? Explain.



- 7.103** • Consider the following 10 observations on the lifetime (in hours) for a certain type of power supply: 152.7, 172.0, 172.5, 173.3, 193.0, 204.7, 216.5, 234.9, 262.6, and 422.6. Construct a normal probability plot, and comment on the plausibility of a normal distribution as a model for power supply lifetime. (The normal scores for a sample of size 10 are $-1.539, -1.001, -0.656, -0.376, -0.123, 0.123, 0.376, 0.656, 1.001$, and 1.539 .)

- 7.104** • Consider the following sample of 25 observations on $x = \text{diameter}$ (in centimeters) of DVD disks produced by a particular manufacturer:

15.66 15.78 15.82 15.84 15.89 15.92 15.94 15.95 15.99
16.01 16.04 16.05 16.06 16.07 16.08 16.10 16.11 16.13
16.13 16.15 16.15 16.19 16.22 16.27 16.29

The 13 largest normal scores for a sample of size 25 are $1.965, 1.524, 1.263, 1.067, 0.905, 0.764, 0.637, 0.519, 0.409, 0.303, 0.200, 0.100$, and 0. The 12 smallest scores result from placing a negative sign in front of each of the given nonzero scores. Construct a normal probability plot. Does it appear plausible that disk diameter is normally distributed? Explain.

- 7.105** • Example 7.32 examined rainfall data for Minneapolis–St. Paul. The square-root transformation was used to obtain a distribution of values that was more nearly symmetric than the distribution of the original data. Another transformation that has been suggested by meteorologists is the cube root: transformed value = (original value) $^{1/3}$. The original values and their cube roots (the transformed values) are given in the following table:

Original	Transformed	Original	Transformed
0.32	0.68	0.59	0.84
0.47	0.78	0.77	0.92
0.52	0.80	0.81	0.93

(continued)

Original	Transformed	Original	Transformed
0.81	0.93	1.87	1.23
0.90	0.97	1.89	1.24
0.96	0.99	1.95	1.25
1.18	1.06	2.05	1.27
1.20	1.06	2.10	1.28
1.20	1.06	2.20	1.30
1.31	1.09	2.48	1.35
1.35	1.11	2.81	1.41
1.43	1.13	3.00	1.44
1.51	1.15	3.09	1.46
1.62	1.17	3.37	1.50
1.74	1.20	4.75	1.68

Construct a histogram of the transformed data. Compare your histogram to those given in Figure 7.38. Which of the cube-root and square-root transformations appear to result in a histogram that is more nearly symmetric?

7.106 The article “[The Distribution of Buying Frequency Rates](#)” (*Journal of Marketing Research* [1980]: 210–216) reported the results of a $3\frac{1}{2}$ -year study of toothpaste purchases. The investigators conducted their research using a national sample of 2071 households and recorded the number of toothpaste purchases for each household participating in the study. The results are given in the following frequency distribution:

Number of Purchases	Number of Households (Frequency)
10 to <20	904
20 to <30	500
30 to <40	258
40 to <50	167
50 to <60	94
60 to <70	56
70 to <80	26
80 to <90	20
90 to <100	13
100 to <110	9
110 to <120	7
120 to <130	6
130 to <140	6
140 to <150	3
150 to <160	0
160 to <170	2

- a. Draw a histogram for this frequency distribution. Would you describe the histogram as positively or negatively skewed?

- b. Does the square-root transformation result in a histogram that is more nearly symmetric than that of the original data? (Be careful! This one is a bit tricky because you don't have the raw data; transforming the endpoints of the class intervals will result in class intervals that are not necessarily of equal widths, so the histogram of the transformed values will have to be drawn with this in mind.)

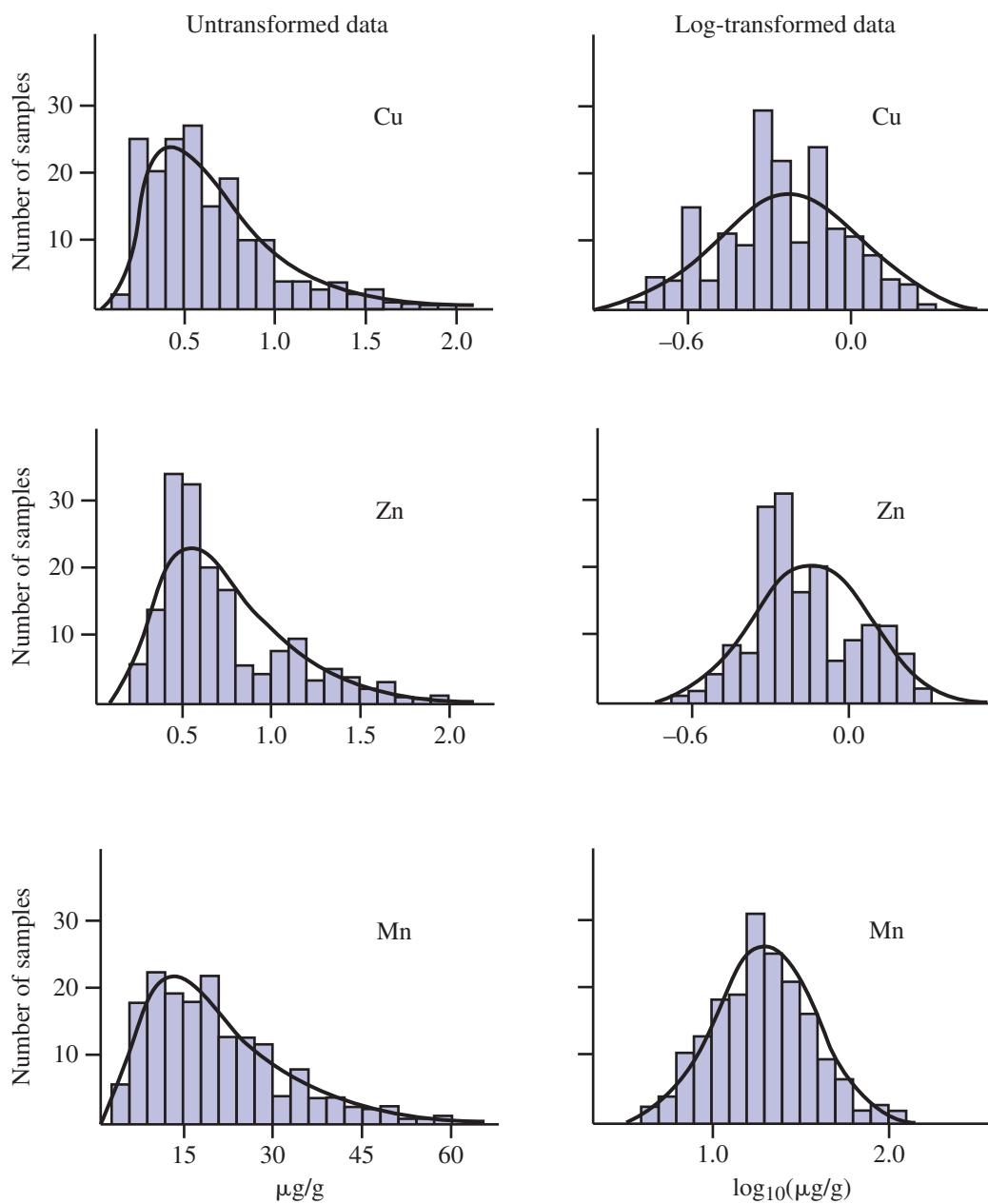
7.107 The paper “[Temperature and the Northern Distributions of Wintering Birds](#)” (*Ecology* [1991]: 2274–2285) gave the following body masses (in grams) for 50 different bird species:

7.7	10.1	21.6	8.6	12.0	11.4	16.6	9.4
11.5	9.0	8.2	20.2	48.5	21.6	26.1	6.2
19.1	21.0	28.1	10.6	31.6	6.7	5.0	68.8
23.9	19.8	20.1	6.0	99.6	19.8	16.5	9.0
448.0	21.3	17.4	36.9	34.0	41.0	15.9	12.5
10.2	31.0	21.5	11.9	32.5	9.8	93.9	10.9
19.6	14.5						

- a. Draw a histogram based on class intervals 5 to <10, 10 to <15, 15 to <20, 20 to <25, 25 to <30, 30 to <40, 40 to <50, 50 to <100, and 100 to <500. Is a transformation of the data desirable? Explain.
- b. Use a calculator or statistical software package to calculate logarithms of these observations, and construct a histogram. Is the log transformation successful in producing a more symmetric distribution?
- c. Consider transformed value = $\sqrt[1]{\text{original value}}$ and construct a histogram of the transformed data. Does the histogram appear to resemble a normal distribution?

7.108 The figure on the next page appeared in the paper “[EDTA-Extractable Copper, Zinc, and Manganese in Soils of the Canterbury Plains](#)” (*New Zealand Journal of Agricultural Research* [1984]: 207–217). A large number of topsoil samples were analyzed for manganese (Mn), zinc (Zn), and copper (Cu), and the resulting data were summarized using histograms.

The investigators transformed each data set using logarithms in an effort to obtain more nearly symmetric distributions of values. Do you think the transformations were successful? Explain.



SECTION 7.8 Using the Normal Distribution to Approximate a Discrete Distribution (Optional)

The distributions of many random variables can be approximated by a carefully chosen normal distribution. In this section, we show how probabilities for some discrete random variables can be approximated using a normal curve. The most important case of this is the approximation of binomial probabilities.

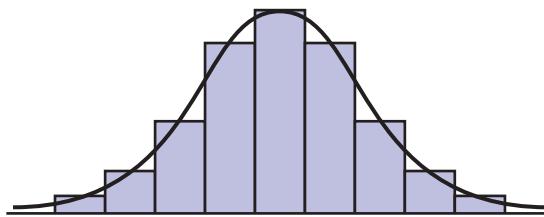
The Normal Curve and Discrete Variables

The probability distribution of a discrete random variable x is represented graphically by a probability histogram. Possible values of x are isolated points on the number line, usually whole numbers. The probability of a particular value is equal to the area of the rectangle centered at that value.

Sometimes a probability histogram can be well approximated by a normal curve, as illustrated in Figure 7.43. In such cases, it is customary to say that the distribution of x is approximately normal. The normal distribution can then be used to calculate approximate probabilities of events involving x .

FIGURE 7.43

A normal curve approximation to a probability histogram.



Example 7.34 Express Mail Packages

Understand the context ➤

Suppose the number of express mail packages mailed at a certain post office on a randomly selected day is approximately normally distributed with mean 18 and standard deviation 6. Let's look at how to approximate the probability that $x = 20$.

Figure 7.44(a) shows a portion of the probability histogram for x with the approximating normal curve superimposed. The area of the shaded rectangle is equal to $P(x = 20)$. The left edge of this rectangle is at 19.5 on the horizontal scale, and the right edge is at 20.5. Therefore, the desired probability is approximately the area under the normal curve between 19.5 and 20.5.

Standardizing these limits gives

$$\frac{20.5 - 18}{6} = 0.42 \quad \frac{19.5 - 18}{6} = 0.25$$

from which we get

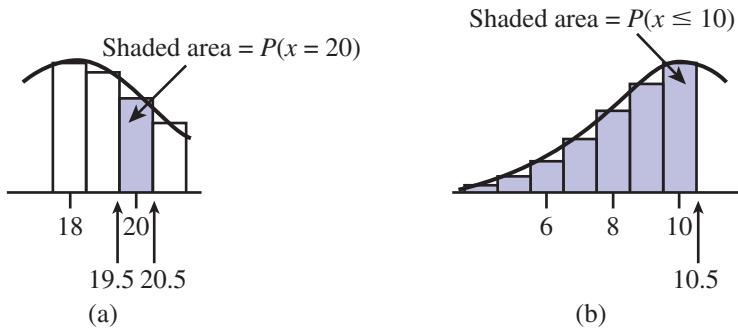
$$P(x = 20) \approx P(0.25 < z < 0.42) = 0.6628 - 0.5987 = 0.0641$$

In a similar fashion, Figure 7.44(b) shows that $P(x \leq 10)$ is approximately the area under the normal curve to the left of 10.5. Then

$$\begin{aligned} P(x \leq 10) &\approx P\left(z \leq \frac{10.5 - 18}{6}\right) = P(z \leq -1.25) \\ &= 0.1056 \end{aligned}$$

FIGURE 7.44

The normal approximation for Example 7.34.



The calculation of probabilities in Example 7.34 illustrates the use of what is known as a **continuity correction**. Because the rectangle for $x = 10$ extends to 10.5 on the right, we use the normal curve area to the left of 10.5 rather than 10. In general, if possible x values are consecutive whole numbers, then $P(a \leq x \leq b)$ will be approximately the normal curve area between limits $a - \frac{1}{2}$ and $b + \frac{1}{2}$.

Normal Approximation to a Binomial Distribution

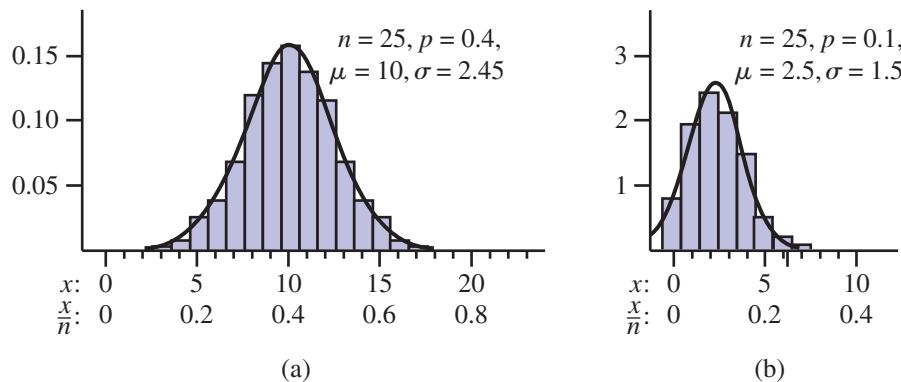
Figure 7.45 shows the probability histograms for two binomial distributions, one with $n = 25$, $p = 0.4$, and the other with $n = 25$, $p = 0.1$. For each distribution, we calculated $\mu = np$ and $\sigma = \sqrt{np(1-p)}$ and then we superimposed a normal curve with the corresponding μ and σ on the probability histogram.

A normal curve fits the probability histogram well in the first case (Figure 7.45(a)). When this happens, it is reasonable to approximate binomial probabilities using areas under the normal curve.

In the second case (Figure 7.45(b)), the normal curve does not give a good approximation because the probability histogram is skewed, whereas the normal curve is symmetric.

FIGURE 7.45

Normal approximations to binomial distributions.



Suppose x is a binomial random variable based on n trials and success probability p , so that

$$\mu = np \text{ and } \sigma = \sqrt{np(1-p)}$$

If n and p are such that

$$np \geq 10 \text{ and } n(1-p) \geq 10$$

then the distribution of x is approximately normal.

Combining this result with the continuity correction implies that

$$P(a \leq x \leq b) \approx P\left(\frac{a - \frac{1}{2} - \mu}{\sigma} \leq z \leq \frac{b + \frac{1}{2} - \mu}{\sigma}\right)$$

That is, the probability that x is between a and b inclusive is approximately the area under the approximating normal curve between $a - \frac{1}{2}$ and $b + \frac{1}{2}$.

Similarly,

$$P(x \leq b) \approx P\left(z \leq \frac{b + \frac{1}{2} - \mu}{\sigma}\right) \quad P(a \leq x) \approx P\left(\frac{a - \frac{1}{2} - \mu}{\sigma} \leq z\right)$$

When either $np < 10$ or $n(1-p) < 10$, the binomial distribution is too skewed for the normal approximation to provide reasonable approximations to binomial probabilities.

Example 7.35 Premature Babies

Understand the context ▶

Premature babies are those born before 37 weeks, and those born before 34 weeks are most at risk. The paper “Some Thoughts on the True Value of Ultrasound” (*Ultrasound in Obstetrics and Gynecology* [2007]: 671–674) reported that 2% of births in the United States occur before 34 weeks.

Suppose that 1000 births are randomly selected and that the number of these births that occurred prior to 34 weeks, x , is to be determined. Because

$$\begin{aligned}np &= 1000(0.02) = 20 \geq 10 \\n(1 - p) &= 1000(0.98) = 980 \geq 10\end{aligned}$$

the distribution of x is approximately normal with

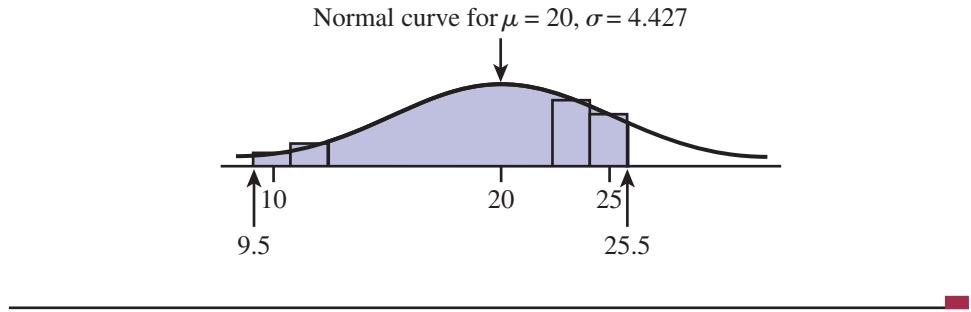
$$\begin{aligned}\mu &= np = 1000(0.02) = 20 \\\sigma &= \sqrt{np(1 - p)} = \sqrt{1000(0.02)(0.98)} = \sqrt{19.60} = 4.427\end{aligned}$$

The probability that the number of babies in a sample of 1000 born prior to 34 weeks will be between 10 and 25 (inclusive) is

Do the work ➤

$$\begin{aligned}P(10 \leq x \leq 25) &= P\left(\frac{9.5 - 20}{4.427} \leq z \leq \frac{25.5 - 20}{4.427}\right) \\&= P(-2.37 \leq z \leq 1.29) \\&= 0.9015 - 0.0089 \\&= 0.8926\end{aligned}$$

as shown in the following figure:



EXERCISES 7.109 - 7.120

- 7.109** Let x denote the IQ of an individual selected at random from a population. The value of x must be a whole number. Suppose that the distribution of x can be approximated by a normal distribution with mean value 100 and standard deviation 15. Approximate the following probabilities. (Hint: See Example 7.34.)

- a. $P(x = 100)$
- b. $P(x \leq 110)$
- c. $P(x < 110)$ (Hint: $x < 110$ is the same as $x \leq 109$.)
- d. $P(75 \leq x \leq 125)$

- 7.110** Suppose that the distribution of

x = the number of items produced by an assembly line during an 8-hour shift
can be approximated by a normal distribution with mean value 150 and standard deviation 10.

- a. What is the approximate probability that the number of items produced is at most 120?
- b. What is the approximate probability that at least 125 items are produced?
- c. What is the approximate probability that between 135 and 160 (inclusive) items are produced?

- 7.111** The number of vehicles leaving a turnpike at a certain exit during a particular time period has approximately a normal distribution with mean value 500 and standard deviation 75. What is the approximate probability that the number of cars exiting during this period is
- a. at least 650?
 - b. strictly between 400 and 550? (*Strictly* means that the values 400 and 550 are not included.)
 - c. between 400 and 550 (inclusive)?

- 7.112** Suppose that x has a binomial distribution with $n = 50$ and $p = 0.6$, so that $\mu = np = 30$ and $\sigma = \sqrt{np(1-p)} = 3.464$. Approximate the following probabilities using the normal approximation with the continuity correction. (Hint: See Example 7.35.)
- $P(x = 30)$
 - $P(x = 25)$
 - $P(x \leq 25)$

- 7.113** For the binomial distribution described in the previous exercise, approximate the following probabilities using the normal approximation with the continuity correction.
- $P(25 \leq x \leq 40)$
 - $P(25 < x < 40)$ (Hint: $25 < x < 40$ is the same as $26 \leq x \leq 39$.)

- 7.114** Symptom validity tests (SVTs) are sometimes used to confirm diagnosis of psychiatric disorders. The paper “Developing a Symptom Validity Test for Post-Traumatic Stress Disorder: Application of the Binomial Distribution” (*Journal of Anxiety Disorders* [2008]: 1297–1302) investigated the use of SVTs in the diagnosis of post-traumatic stress disorder.

One SVT proposed is a 60-item test (called the MENT test), where each item has only a correct or incorrect response. The MENT test is designed so that responses to the individual questions can be considered independent of one another. For this reason, the authors of the paper believe that the score on the MENT test can be viewed as a binomial random variable with $n = 60$.

The MENT test is designed to help in distinguishing fake claims of post-traumatic stress disorder. The items on the MENT test are written so that the correct response to an item should be relatively obvious, even to people suffering from stress disorders. Researchers have found that a patient with a fake claim of stress disorder will try to fake the test, and that the probability of a correct response to an item for these patients is 0.7 (compared to 0.96 for other patients).

The authors used a normal approximation to the binomial distribution with $n = 60$ and $p = 0.7$ to approximate various probabilities of interest, where

x = number of correct responses on the MENT test for a patient who is trying to fake the test.

- Verify that it is appropriate to use a normal approximation to the binomial distribution in this situation.
- Approximate the following probabilities:
 - $P(x = 42)$
 - $P(x < 42)$
 - $P(x \leq 42)$
- Explain why the probabilities computed in Part (b) are not all equal.

- 7.115** Consider the information on the MENT test given in the previous exercise.

- The exact binomial probability of a score of 42 or less for someone who is not faking the test ($p = 0.96$) is

$$P(x \leq 42) = 0.00000000001$$

Explain why this probability was calculated using the binomial formula rather than using a normal approximation.

- The authors of the study described in the previous exercise propose that someone who scores 42 or less on the MENT exam is faking the test. Explain why this is reasonable, using the probability from Part (a) and the probabilities from the previous exercise as justification.

- 7.116** Suppose that 70% of the bicycles sold by a certain store are mountain bikes. Among 100 randomly selected bike purchases, what is the approximate probability that

- At most 75 are mountain bikes?
- Between 60 and 75 (inclusive) are mountain bikes?
- More than 80 are mountain bikes?
- At most 30 are not mountain bikes?

- 7.117** Suppose that 25% of the fire alarms in a large city are false alarms. Let x denote the number of false alarms in a random sample of 100 alarms. Approximate the following probabilities:

- $P(20 \leq x \leq 30)$
- $P(20 < x < 30)$
- $P(x \geq 35)$
- The probability that x is farther than 2 standard deviations from its mean value

- 7.118** Suppose that 65% of all registered voters in a certain area favor a 7-day waiting period before purchase of a handgun. Among 225 randomly selected registered voters, what is the approximate probability that

- At least 150 favor such a waiting period?
- More than 150 favor such a waiting period?
- Fewer than 125 favor such a waiting period?

- 7.119** Flashlight bulbs manufactured by a certain company are sometimes defective.

- If 5% of all such bulbs are defective, could the techniques of this section be used to approximate the probability that at least five of the bulbs in a random sample of size 50 are defective? If so, calculate this probability; if not, explain why not.
- Reconsider the question posed in Part (a) for the probability that at least 20 bulbs in a random sample of size 500 are defective.

- 7.120** A company that manufactures mufflers for cars offers a lifetime warranty on its products, provided that ownership of the car does not change. Suppose that only 20% of its mufflers are replaced under this warranty.
- In a random sample of 400 purchases, what is the approximate probability that between 75 and 100 (inclusive) mufflers are replaced under warranty?

- Among 400 randomly selected purchases, what is the approximate probability that at most 70 mufflers are ultimately replaced under warranty?
- If you were told that fewer than 50 among 400 randomly selected purchases were ever replaced under warranty, would you question the 20% figure? Explain.

CHAPTER ACTIVITIES

ACTIVITY 7.1 IS IT REAL?

Background: Three students were asked to complete an assignment that requested they do the following:

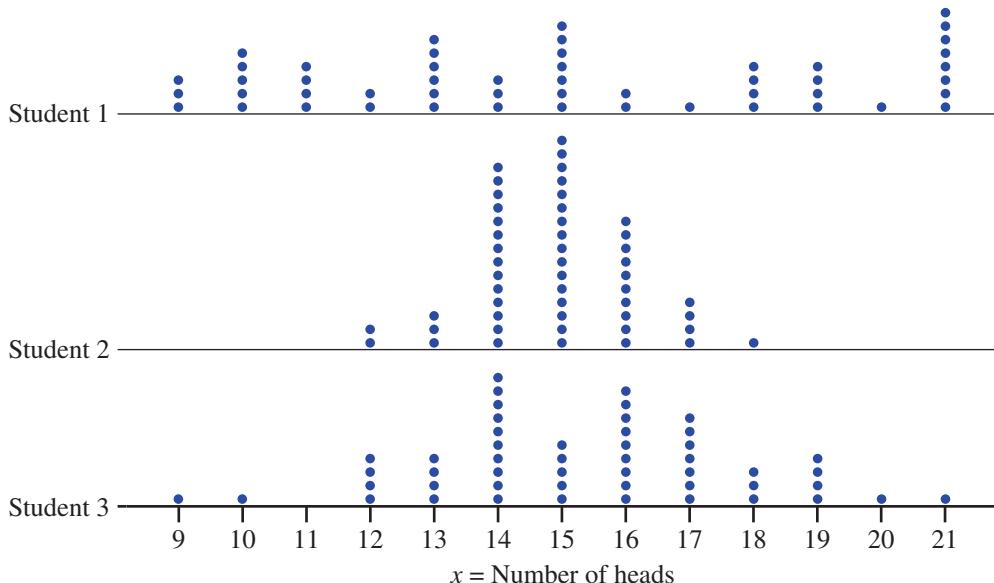
- Flip a coin 30 times, and note the number of heads observed in the 30 flips.
- Repeat Step (a) 100 times, to obtain 100 observations of the random variable x = number of heads in 30 flips.
- Construct a dotplot of the 100 x values.

Because this was a tedious assignment, one or more of the students did not really carry out the coin flipping and just made up 100 values of x that they thought would look “real.” The dotplots produced by these three students are shown below.

- Do you think that any of the three students made up the x values shown in their dotplot? If so, which

ones, and what about the dotplot makes you think the student did not actually do the coin flipping?

- Working as a group, each student in your class should flip a coin 30 times and note the number of heads in the 30 tosses. If there are fewer than 50 students in the class, each student should repeat this process until there are a total of at least 50 observations of x = number of heads in 30 flips. Using the data from the entire class, construct a dotplot of the x values.
- After looking at the dotplot in Step 2 that resulted from actually flipping a coin 30 times and observing number of heads, reconsider your answers in Question 1. For each of the three students, explain why you now think that he or she did or did not actually do the coin flipping.



ACTIVITY 7.2 ROTTEN EGGS?

Background: *The Salt Lake Tribune* (October 11, 2002) printed the following account of an exchange between a restaurant manager and a health inspector:

The recipe calls for four fresh eggs for each quiche. A Salt Lake County Health Department inspector paid a visit recently and pointed out that research by the Food and Drug Administration indicates that one in four eggs carries salmonella bacterium, so restaurants should never use more than three eggs when preparing quiche. The manager on duty wondered aloud if simply throwing out three eggs from each dozen and using the remaining nine in four-egg quiches would serve the same purpose.

1. Working in a group or as a class, discuss the folly of the above statement!
2. Suppose the following argument is made for three-egg quiches rather than four-egg quiches: Let x = number of eggs that carry salmonella. Then

$$p(0) = p(x = 0) = (0.75)^3 = 0.422$$

for three-egg quiches and

$$p(0) = p(x = 0) = (0.75)^4 = 0.316$$

for four-egg quiches. What assumption must be made to justify these probability calculations? Do you think this is reasonable or not? Explain.

3. Suppose that a carton of one dozen eggs does happen to have exactly three eggs that carry salmonella and

that the manager does as he proposes: selects three eggs at random and throws them out, then uses the remaining nine eggs in four-egg quiches. Let x = number of eggs that carry salmonella among four eggs selected at random from the remaining nine.

Working with a partner, conduct a simulation to approximate the distribution of x by carrying out the following sequence of steps:

- a. Take 12 identical slips of paper and write “Good” on nine of them and “Bad” on the remaining three. Place the slips of paper in a paper bag or some other container.
 - b. Mix the slips and then select three at random and remove them from the bag.
 - c. Mix the remaining slips and select four “eggs” from the bags.
 - d. Note the number of bad eggs among the four selected. (This is an observed x value.)
 - e. Replace all slips, so that the bag now contains all 12 “eggs.”
 - f. Repeat Steps (b)–(d) at least 10 times, each time recording the observed x value.
4. Combine the observations from your group with those from the other groups. Use the resulting data to approximate the distribution of x . Comment on the resulting distribution in the context of the risk of salmonella exposure if the manager’s proposed procedure is used.

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
Random variable: discrete or continuous	A numerical variable with a value determined by the outcome of a chance experiment. A random variable is discrete if its possible values are isolated points along the number line and continuous if its possible values form an entire interval on the number line.	μ_x and σ_x	The mean and standard deviation, respectively, of a random variable x . These quantities describe the center and extent of spread about the center of the variable’s probability distribution.
Probability distribution of a discrete random variable x	A formula, table, or graph that gives the probability associated with each possible x value. Conditions on $p(x)$ are (1) $p(x) \geq 0$, and (2) $\sum p(x) = 1$, where the sum is over all possible x values.	$\mu_x = \Sigma x p(x)$	The mean value of a discrete random variable x . It locates the center of the variable’s probability distribution.
Probability distribution of a continuous random variable x	Specified by a smooth (density) curve for which the total area under the curve is 1. The probability $P(a < x < b)$ is the area under the curve and above the interval from a to b ; this is also $P(a \leq x \leq b)$.	$\sigma_x^2 = \Sigma (x - \mu_x)^2 p(x)$ $\sigma_x = \sqrt{\sigma_x^2}$	The variance and standard deviation, respectively, of a discrete random variable. These are measures of the extent to which the variable’s distribution spreads out about the mean μ_x .
		Binomial probability distribution	This formula gives the probability of observing x successes ($x = 0, 1, \dots, n$) among n trials of a binomial experiment.
		$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$	

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
$\mu_x = np$ $\sigma_x = \sqrt{np(1-p)}$	The mean and standard deviation of a binomial random variable.	$z = \frac{x - \mu}{\sigma}$	z is obtained by “standardizing”: subtracting the mean and then dividing by the standard deviation. When x has a normal distribution, z has a standard normal distribution.
Normal distribution	A continuous probability distribution that has a bell-shaped density curve. A particular normal distribution is determined by specifying values of μ and σ .	Normal probability plot	A graph used to judge the plausibility of the assumption that a sample has been selected from a normal population distribution. If the plot is reasonably straight, this assumption is reasonable.
Standard normal distribution	This is the normal distribution with $\mu = 0$ and $\sigma = 1$. The density curve is called the z curve, and z is the letter commonly used to denote a variable having this distribution. Areas under the z curve to the left of various values are given in Appendix Table 2.	Normal approximation to the binomial distribution	When both $np \geq 10$ and $n(1-p) \geq 10$, binomial probabilities are well approximated by corresponding areas under a <u>normal curve</u> with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

CHAPTER REVIEW Exercises 7.121 - 7.141

- 7.121** Let x denote the duration of a randomly selected pregnancy (the time elapsed between conception and birth). Accepted values for the mean value and standard deviation of x are 266 days and 16 days, respectively. Suppose that the probability distribution of x is (approximately) normal.
- What is the probability that the duration of pregnancy is between 250 and 300 days?
 - What is the probability that the duration of pregnancy is at most 240 days?
 - What is the probability that the duration of pregnancy is within 16 days of the mean duration?
 - A “Dear Abby” column dated January 20, 1973, contained a letter from a woman who stated that the duration of her pregnancy was exactly 310 days. (She wrote that the last visit with her husband, who was in the Navy, occurred 310 days before birth.) What is the probability that duration of pregnancy is 310 days or more? Does this probability make you a bit skeptical of the claim?
 - Some insurance companies will pay the medical expenses associated with childbirth only if the insurance has been in effect for more than 9 months (275 days). This restriction is designed to ensure that the insurance company pays benefits for only those pregnancies for which conception occurred during coverage. Suppose that conception occurred 2 weeks after coverage began. What is the probability that the insurance company will refuse to pay benefits because of the 275-day insurance requirement?
- 7.122** A soft-drink machine dispenses only regular Coke and Diet Coke. Sixty percent of all purchases from

this machine are diet drinks. The machine currently has 10 cans of each type. If 15 customers want to purchase drinks before the machine is restocked, what is the probability that each of the 15 is able to purchase the type of drink desired? (Hint: Let x denote the number among the 15 who want a diet drink. For which possible values of x is everyone satisfied?)

- 7.123** A business has six customer service telephone lines. Let x denote the number of lines in use at a specified time. Suppose that the probability distribution of x is as follows:

x	0	1	2	3	4	5	6
$p(x)$	0.10	0.15	0.20	0.25	0.20	0.06	0.04

Write each of the following events in terms of x , and then calculate the probability of each one:

- At most three lines are in use
- Fewer than three lines are in use
- At least three lines are in use

- 7.124** Use the probability distribution for $x = \text{number of lines in use}$ given in the previous exercise to calculate the probability of each of the following events.

- Between two and five lines (inclusive) are in use
- Between two and four lines (inclusive) are not in use
- At least four lines are not in use

- 7.125** Refer to the probability distribution given in Exercise 7.123.

- Calculate the mean value and standard deviation of x .

- b.** What is the probability that the number of lines in use is more than 3 standard deviations from the mean value?
- 7.126** A new battery's voltage may be acceptable (A) or unacceptable (U). A certain flashlight requires two batteries, so batteries will be independently selected and tested until two acceptable ones have been found. Suppose that 80% of all batteries have acceptable voltages, and let y denote the number of batteries that must be tested.
- What is $p(2)$, that is, $P(y = 2)$?
 - What is $p(3)$? (Hint: There are two different outcomes that result in $y = 3$.)
 - In order to have $y = 5$, what must be true of the fifth battery selected? List the four outcomes for which $y = 5$, and then determine $p(5)$.
 - Use the pattern in your answers for Parts (a)–(c) to obtain a general formula for $p(y)$.
- 7.127** A pizza company advertises that it puts 0.5 pounds of real mozzarella cheese on its medium pizzas. Suppose that the amount of cheese on a randomly selected medium pizza is normally distributed with a mean value of 0.5 pounds and a standard deviation of 0.025 pounds.
- What is the probability that the amount of cheese on a medium pizza is between 0.525 and 0.550 pounds?
 - What is the probability that the amount of cheese on a medium pizza exceeds the mean value by more than 2 standard deviations?
 - What is the probability that three randomly selected medium pizzas all have at least 0.475 pounds of cheese?
- 7.128** Suppose that fuel efficiency for a particular model car under specified conditions is normally distributed with a mean value of 30.0 mpg and a standard deviation of 1.2 mpg.
- What is the probability that the fuel efficiency for a randomly selected car of this type is between 29 and 31 mpg?
 - Would it surprise you to find that the efficiency of a randomly selected car of this model is less than 25 mpg?
 - If three cars of this model are randomly selected, what is the probability that all three have efficiencies exceeding 32 mpg?
 - Find a number c^* such that 95% of all cars of this model have efficiencies exceeding c^* (that is, $P(x > c^*) = 0.95$).
- 7.129** A coin is flipped 25 times. Let x be the number of flips that result in heads (H). Consider the following rule for deciding whether or not the coin is fair:
- Judge the coin fair if $8 \leq x \leq 17$.
 Judge the coin biased if either $x \leq 7$ or $x \geq 18$.
- What is the probability of judging the coin biased when it is actually fair?
 - What is the probability of judging the coin fair when $P(H) = 0.9$, so that there is a substantial bias? Repeat for $P(H) = 0.1$.
 - What is the probability of judging the coin fair when $P(H) = 0.6$? When $P(H) = 0.4$? Why are these probabilities so large compared to the probabilities in Part (b)?
 - What happens to the “error probabilities” of Parts (a) and (b) if the decision rule is changed so that the coin is judged fair if $7 \leq x \leq 18$ and unfair otherwise? Is this a better rule than the one first proposed? Explain.
- 7.130** The probability distribution of x , the number of defective tires on a randomly selected automobile checked at a certain inspection station, is given in the following table:
- | x | 0 | 1 | 2 | 3 | 4 |
|--------|------|------|------|------|------|
| $p(x)$ | 0.54 | 0.16 | 0.06 | 0.04 | 0.20 |
- The mean value of x is $\mu_x = 1.2$. Calculate the values of σ_x^2 and σ_x .
- 7.131** The amount of time spent by a statistical consultant with a client at their first meeting is a random variable having a normal distribution with a mean value of 60 minutes and a standard deviation of 10 minutes.
- What is the probability that more than 45 minutes is spent at the first meeting?
 - What amount of time is exceeded by only 10% of all clients at a first meeting?
 - If the consultant assesses a fixed charge of \$10 (for overhead) and then charges \$50 per hour, what is the mean revenue from a client's first meeting?
- 7.132** The lifetime of a certain brand of battery is normally distributed with a mean value of 6 hours and a standard deviation of 0.8 hours when it is used in a particular DVD player. Suppose that two new batteries are independently selected and put into the player. The player stops functioning as soon as one of the batteries fails.
- What is the probability that the player functions for at least 4 hours?

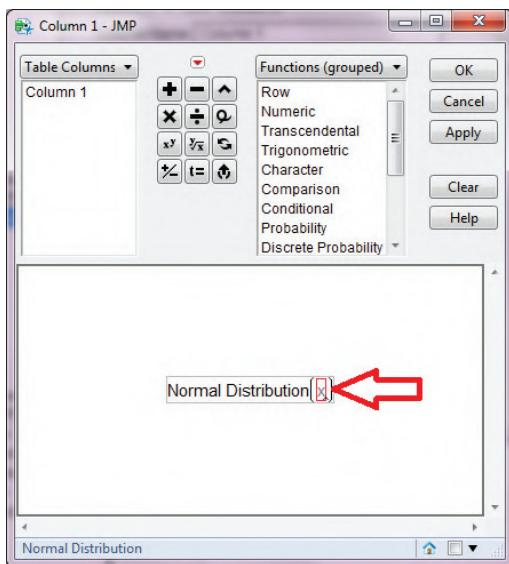
- b.** What is the probability that the DVD player works for at most 7 hours?
- c.** Find a number c^* such that only 5% of all DVD players will function without battery replacement for more than c^* hours.
- 7.133** A machine producing vitamin E capsules operates so that the actual amount of vitamin E in each capsule is normally distributed with a mean of 5 mg and a standard deviation of 0.05 mg. What is the probability that a randomly selected capsule contains less than 4.9 mg of vitamin E? at least 5.2 mg?
- 7.134** The *Wall Street Journal* (February 15, 1972) reported that General Electric was sued in Texas for sex discrimination over a minimum height requirement of 5 ft. 7 in. The suit claimed that this restriction eliminated more than 94% of adult females from consideration. Let x represent the height of a randomly selected adult woman. Suppose that x is approximately normally distributed with mean 66 inches (5 ft. 6 in.) and standard deviation 2 inches.
- a.** Is the claim that 94% of all women are shorter than 5 ft. 7 in. correct?
- b.** What proportion of adult women would be excluded from employment as a result of the height restriction?
- 7.135** The longest “run” of S ’s in the sequence SSFSSSSFFS has length 4, corresponding to the S ’s on the fourth, fifth, sixth, and seventh positions. Consider a binomial experiment with $n = 4$, and let y be the length in the longest run of S ’s.
- a.** When $p = 0.5$, the 16 possible outcomes are equally likely. Determine the probability distribution of y in this case (first list all outcomes and the y value for each one). Then calculate μ_y .
- b.** Repeat Part (a) for the case $p = 0.6$.
- c.** Let z denote the longest run of either S ’s or F ’s. Determine the probability distribution of z when $p = 0.5$.
- 7.136** Four people—a, b, c, and d—are waiting to give blood. Of these four, a and b have type AB blood, whereas c and d do not. An emergency call has just come in for some type AB blood. If blood donations are taken one by one from the four people in random order and x is the number of donations needed to obtain an AB individual (so possible x values are 1, 2, and 3), what is the probability distribution of x ?
- 7.137** Kyle and Lygia are going to play a series of Trivial Pursuit games. The first person to win four games will be declared the winner. Suppose that outcomes of successive games are independent and that the probability of Lygia winning any particular game is 0.6. Define a random variable x as the number of games played in the series.
- a.** What is $P(4)$? (Hint: Either Kyle or Lygia could win four straight games.)
- b.** What is $P(5)$? (Hint: For Lygia to win in exactly five games, what has to happen in the first four games and in Game 5?)
- c.** Determine the probability distribution of x .
- d.** On average, how many games will the series last?
- 7.138** Suppose that your statistics professor tells you that the scores on a midterm exam were approximately normally distributed with a mean of 78 and a standard deviation of 7. The top 15% of all scores have been designated A’s. Your score is 89. Did you receive an A? Explain.
- 7.139** Suppose that the pH of soil samples taken from a certain geographic region is normally distributed with a mean pH of 6.00 and a standard deviation of 0.10. If the pH of a randomly selected soil sample from this region is determined, answer the following questions about it:
- a.** What is the probability that the resulting pH is between 5.90 and 6.15?
- b.** What is the probability that the resulting pH exceeds 6.10?
- c.** What is the probability that the resulting pH is at most 5.95?
- d.** What value will be exceeded by only 5% of all such pH values?
- 7.140** The lightbulbs used to provide exterior lighting for a large office building have an average lifetime of 700 hours. If length of life is approximately normally distributed with a standard deviation of 50 hours, how often should all the bulbs be replaced so that no more than 20% of the bulbs will have already burned out?
- 7.141** Suppose there are approximately 40,000 travel agencies in the United States, of which 11,000 are members of the American Society of Travel Agents (ASTA).
- a.** If x is the number of ASTA members among 5000 randomly selected agencies, could you use the methods of Section 7.8 to approximate $P(1200 < x < 1400)$? Why or why not?
- b.** In a random sample of 100 agencies, what are the mean value and standard deviation of the number of ASTA members?
- c.** If the sample size in Part (b) is doubled, does the standard deviation double? Explain.

TECHNOLOGY NOTES

Finding Normal Probabilities

JMP

- After opening a new data table, click **Rows** then select **Add rows**
- Type 1 in the box next to **How many rows to add:**
- Click **OK**
- Double-click on the **Column 1** heading
- Click **Column Properties** and select **Formula**
- Click **Edit Formula**
- In the box under **Functions (grouped)** click **Probability** then select **Normal Distribution**
- In the white box at the bottom half of the screen, double-click in the red box around x



- Type the value that you would like to find a probability for and click **OK**
- Click **OK**

Note: This procedure outputs the value for $P(X \leq x)$. If you want to find the value for $P(X \geq x)$ you will need to subtract the output from one.

Minitab

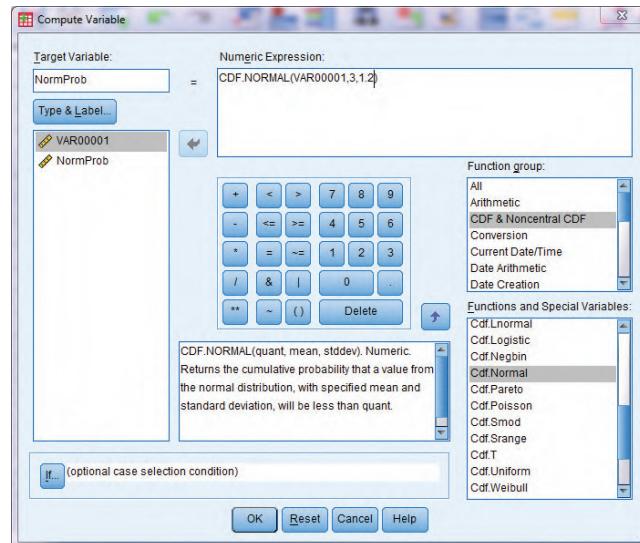
- Click **Calc** then click **Probability Distributions** then click **Normal...**
- In the box next to **Mean**: type the mean of the normal distribution that you are working with
- In the box next to **Standard deviation**: type the standard deviation of the normal distribution that you are working with
- Click the radio button next to **Input constant**:
- Click in the box next to **Input constant**: and type the value that you are finding the probability for
- Click **OK**

Note: This procedure outputs the value for $P(X \leq x)$. If you want to find the value for $P(X \geq x)$ you will need to subtract the output from one.

Note: You may also type a column of values for which you would like to find probabilities (these must ALL be from the SAME distribution) and use the **Input Column** option to find probabilities for each value in the column selected.

SPSS

- Type in the values that you would like to find the probabilities for in one column (this column will be automatically titled VAR00001)
- Click **Transform** and select **Compute Variable...**
- In the box under **Target Variable**: type NormProb (this will be the title of a new column that lists the cumulative probabilities for each value in VAR00001)
- Under **Function group**: select **CDF & Noncentral CDF**
- Under **Functions and Special Variables**: double-click **Cdf.Normal**
- In the box under **Numeric Expression**: you should see **CDF.NORMAL(?, ?, ?)**
- Highlight the first ? and double-click VAR00001
- Highlight the second ? and type the mean of the normal distribution that you are working with
- Highlight the third ? and type the standard deviation of the normal distribution that you are working with



- Click **OK**

Note: This procedure outputs the value for $P(X \leq x)$. If you want to find the value for $P(X \geq x)$ you will need to subtract the output from one.

Excel

- Click in an empty cell
- Click the **Formulas** ribbon
- Select **Insert Function**
- Select **Statistical** from the drop-down menu for category
- Select **NORMDIST** from the **Select a function:** box
- Click **OK**

7. Click in the box next to **X** and type the data value that you would like to find the probability for
8. Click in the box next to **Mean** and type the mean of the Normal distribution that you are working with
9. Click in the box next to **Standard_dev** and type the standard deviation of the Normal distribution that you are working with
10. Click in the box next to **Cumulative** and type **TRUE**
11. Click **OK**

Note: This procedure outputs the value for $P(X \leq x)$. If you want to find the value for $P(X \geq x)$ you will need to subtract the output from one.

TI-83/84

$P(x_1 < X < x_2)$

1. Press the **2nd** key then press the **VARS** key
2. Highlight **normalcdf(** and press **ENTER**
3. Type in the lower bound x_1
4. Type a comma
5. Type in the upper bound x_2
6. Type a comma
7. Type in the value of the mean
8. Type a comma
9. Type in the value of the standard deviation
10. Type a right parenthesis
11. Press **ENTER**

Note: If you are looking for $P(X < x)$, use a lower bound of $-10,000,000$. If you are looking for $P(X > x)$, use an upper bound of $10,000,000$.

TI-Nspire

$P(x_1 < X < x_2)$

1. Enter the Scratchpad
2. Press the **menu** key and select **5:Probability** then **5:Distributions** then **2:Normal Cdf...** then press **enter**
3. In the box next to **Lower Bound** type the value for x_1
4. In the box next to **Upper Bound** type the value for x_2
5. In the box next to μ type the value of the mean
6. In the box next to σ type the value of the standard deviation
7. Press **OK**

Note: If you are looking for $P(X < x)$, use a lower bound of $-10,000,000$. If you are looking for $P(X > x)$, use an upper bound of $10,000,000$.

Finding Binomial Probabilities

JMP

$P(X \leq x)$

1. After opening a new data table, click **Rows** then select **Add rows**
2. Type **1** in the box next to **How many rows to add:**
3. Click **OK**
4. Double-click on the **Column 1** heading
5. Click **Column Properties** and select **Formula**
6. Click **Edit Formula**
7. In the box under **Functions (grouped)** click **Discrete Probability** then select **Binomial Distribution**

8. In the white box at the bottom half of the screen, double-click in the red box around **p** and type the value for the success probability, **p**, and press **enter** on your keyboard
9. Double-click in the red box around **n** and type the value for the number of trials, **n**, and press **enter** on your keyboard
10. Double-click in the red box around **k** and type the value for the number of successes for which you would like to find the probability
11. Click **OK**
12. Click **OK**

$P(X = x)$

1. After opening a new data table, click **Rows** then select **Add rows**
2. Type **1** in the box next to **How many rows to add:**
3. Click **OK**
4. Double-click on the **Column 1** heading
5. Click **Column Properties** and select **Formula**
6. Click **Edit Formula**
7. In the box under **Functions (grouped)** click **Discrete Probability** then select **Binomial Probability**
8. In the white box at the bottom half of the screen, double-click in the red box around **p** and type the value for the success probability, **p**, and press **enter** on your keyboard
9. Double-click in the red box around **n** and type the value for the number of trials, **n**, and press **enter** on your keyboard
10. Double-click in the red box around **k** and type the value for the number of successes for which you would like to find the probability
11. Click **OK**
12. Click **OK**

Minitab

$P(X \leq x)$

1. Click **Calc** then click **Probability Distributions** then click **Binomial...**
2. In the box next to **Number of Trials:** input the value for **n**, the total number of trials
3. In the box next to **Probability of Success:** input the value for **p**, the success probability
4. Click the radio button next to **Input Constant**
5. In the box next to **Input Constant:** type the value that you want to find the probability for
6. Click **OK**

Note: You may also type a column of values for which you would like to find probabilities (these must ALL be from the SAME distribution) and use the **Input Column** option to find probabilities for each value in the column selected.

$P(X = x)$

1. Click **Calc** then click **Probability Distributions** then click **Binomial...**
2. Click the radio button next to **Probability**
3. In the box next to **Number of Trials:** input the value for **n**, the total number of trials
4. In the box next to **Probability of Success:** input the value for **p**, the success probability

5. Click the radio button next to **Input Constant**
6. In the box next to **Input Constant:** type the value that you want to find the probability for
7. Click **OK**

Note: You may also type a column of values for which you would like to find probabilities (these must ALL be from the SAME distribution) and use the **Input Column** option to find probabilities for each value in the column selected.

SPSS

$P(X = x)$

1. Type in the values that you would like to find the probabilities for in one column (this column will be automatically titled VAR00001)
2. Click **Transform** and select **Compute Variable...**
3. In the box under **Target Variable:** type **BinomProb** (this will be the title of a new column that lists the cumulative probabilities for each value in VAR00001)
4. Under **Function group:** select **CDF & Noncentral CDF**
5. Under **Functions and Special Variables:** double-click **Cdf. Binom**
6. In the box under **Numeric Expression:** you should see **CDF. BINOM(?, ?, ?)**
7. Highlight the first **?** and double-click **VAR00001**
8. Highlight the second **?** and type the value for n , the total number of trials
9. Highlight the third **?** and type success probability, p
10. Click **OK**

$P(X = x)$

1. Type in the values that you would like to find the probabilities for in one column (this column will be automatically titled VAR00001)
2. Click **Transform** and select **Compute Variable...**
3. In the box under **Target Variable:** type **BinProb** (this will be the title of a new column that lists the cumulative probabilities for each value in VAR00001)
4. Under **Function group:** select **PDF & Noncentral PDF**
5. Under **Functions and Special Variables:** double-click **Pdf. Binom**
6. In the box under **Numeric Expression:** you should see **PDF. BINOM(?, ?, ?)**
7. Highlight the first **?** and double-click **VAR00001**
8. Highlight the second **?** and type the value for n , the total number of trials
9. Highlight the third **?** and type success probability, p
10. Click **OK**

Excel

1. Click in an empty cell
2. Click the **Formulas** ribbon
3. Select **Insert Function**
4. Select **Statistical** from the drop-down menu for category
5. Select **BINOMDIST** from the **Select a function:** box
6. Click **OK**
7. Click in the box next to **Number_s** and type the number of successes that you are finding a probability for

8. Click in the box next to **Trials** and type the value for n , the total number of trials
9. Click in the box next to **Probability_s** and type success probability, p
10. Click in the box next to **Cumulative** and type **TRUE** if you are finding $P(X \leq x)$ or type **FALSE** if you are finding $P(X = x)$
11. Click **OK**

Note: This procedure outputs the value for $P(X \leq x)$ when **TRUE** is used as input for **Cumulative**. If you want to find the value for $P(X \geq x)$ you will need to subtract this output from one.

TI-83/84

$P(X \leq x)$

1. Press the **2nd** key then press the **VARS** key
2. Highlight the **binomcdf(** option and press the **ENTER** key
3. Type in the number of trials, n
4. Type a comma
5. Type in the success probability, p
6. Type a comma
7. Type the value for x
8. Type a right parenthesis
9. Press **ENTER**

$P(X = x)$

1. Press the **2nd** key then press the **VARS** key
2. Highlight the **binompdf(** option and press the **ENTER** key
3. Type in the number of trials, n
4. Type a comma
5. Type in the success probability, p
6. Type a comma
7. Type the value for x
8. Type a right parenthesis
9. Press **ENTER**

TI-Nspire

$P(x_1 \leq X \leq x_2)$

1. Enter the Scratchpad
2. Press the **menu** key and select **5:Probability** then select **5:Distributions** then select **E:Binomial Cdf...** and press the **enter** key
3. In the box next to **Num Trials, n** type in the number of trials, n
4. In the box next to **Prob Success, p** type in the success probability, p
5. In the box next to **Lower Bound** type in the value for x_1
6. In the box next to **Upper Bound** type in the value for x_2
7. Press **OK**

Note: In order to find $P(X \leq x)$ input a 0 for the lower bound.

$P(X = x)$

1. Enter the Calculate Scratchpad
2. Press the **menu** key and select **5:Probability** then select **5:Distributions** then select **D:Binomial Pdf...** and press the **enter** key
3. In the box next to **Num Trials, n** type in the number of trials, n

4. In the box next to **Prob Success, p** type in the success probability, p
5. In the box next to **X Value** type the value for x
6. Press **OK**

Normal Probability Plots

JMP

1. Enter the raw data into a column
2. Click **Analyze** then select **Distribution**
3. Click and drag the column name containing the data from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to the column name
6. Select **Normal Quantile Plot**

Minitab

1. Input the data for which you would like to check Normality into a column
2. Click **Graph** then click **Probability Plot...**
3. Highlight the **Single** plot
4. Click **OK**
5. Double click the column name for the column that contains your data to move it into the **Graph Variables:** box
6. Click **OK**

SPSS

1. Input the data for which you would like to check Normality into a column
2. Click **Analyze** then select **Descriptive Statistics** then select **Q-Q Plots...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Variables:** box
5. Click **OK**

Note: The normal probability plot is output with several other plots and statistics. This plot can be found under the title **Normal Q-Q Plot**.

Excel

Excel does not have the functionality to automatically produce Normal Probability Plots.

However, Excel can produce Normal Probability Plots in the course of running a regression using the Analysis ToolPak.

TI-83/84

1. Input the raw data into **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press the **2nd** key then press the **Y =** key
3. Highlight **Plot1** and press **ENTER**
4. Highlight **On** and press **ENTER**
5. Highlight the plot type on the second row, third column and press **ENTER**
6. Press **GRAPH**

TI-Nspire

1. Enter the data into a data list (In order to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press **menu** and select **3:Data** then **6:QuickGraph** then press **enter**
3. Press **menu** and select **1:Plot Type** then select **4:Normal Probability Plot** and press **enter**

CUMULATIVE REVIEW EXERCISES

CR7.1 - CR7.21

● Data set available online

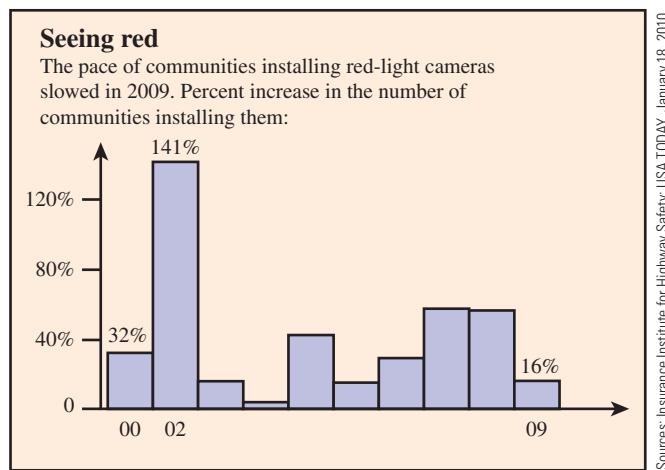
49	52	53	59	59	61	69	70	70	73	79	80
85	87	89	90	92	93	93	95	96	97	114	149
208	318	334									

- a. Use the given water consumption values to construct a boxplot. Describe any interesting features of the boxplot. Are there any outliers in the data set?
- b. Calculate the mean and standard deviation for this data set.
- c. If the two largest values were deleted from the data set, would the standard deviation of this new data set be greater than or less than the standard deviation you calculated for the entire data set?
- d. How do the values of the mean and median of the entire data set compare? Is this consistent with the shape of the boxplot from Part (a)? Explain.

CR7.3 Red-light cameras are used in many places to deter drivers from running red lights. The following graphical display is similar to one that appeared in the

CR7.2 ● The article “Water Consumption in the Bay Area” (*San Jose Mercury News*, January 16, 2010) reported the accompanying values for water consumption (gallons per person per day) for residential customers of 27 San Francisco Bay area water agencies:

article “[Communities Put a Halt to Red-Light Cameras](#)” (*USA TODAY*, January 18, 2010). Based on this graph, would it be correct to conclude that there were fewer red-light cameras in 2009 than in 2002? Explain.



CR7.4 The following quote is from [USA TODAY](#) (January 21, 2010):

Most Americans think the Census is important, and the majority say they will participate in the count this spring—although Hispanics, younger people and the less educated are not as enthusiastic, a Pew Research Center survey found. The poll of 1504 adults found that perceptions differ: 74% of blacks and 72% of Hispanics rate the Census “very important” vs. 57% of whites. Believing it’s important doesn’t necessarily mean participating: Fewer than half of Hispanics, 57% of blacks and 61% of whites say they will fill out and mail the forms. Age is the greatest factor in participation: Only 36% of respondents who are 30 or younger say they will definitely participate.

Suppose that it is reasonable to regard the survey participants as representative of adult Americans.

- a. Suppose that an adult American is to be selected at random and consider the following events:

I = the event that the selected individual thinks that the Census is very important

W = the event that the selected individual is white

Are the events I and W independent or dependent events? Explain.

- b. Give an example of two census-related events different than the ones defined in Part (a) and indicate whether they are independent or dependent events. Explain your reasoning.

CR7.5 Beverly is building a fireplace. She is using handmade decorative 8-inch bricks for the bottom row of the fireplace. Although the bricks are advertised as 8-inch bricks, the actual length of a brick is variable, and the brick lengths are approximately normally distributed with a mean of 8 inches and a standard deviation of 0.2 inches.

Unfortunately, Beverly didn’t buy enough of the 8-inch bricks, and needs to get two more bricks to try to fill the remaining 15.9 inches. The two additional bricks cannot have a combined length of more than 15.9 inches and must fit in the 15.9-inch space with no more than 0.1 inch of space left over.

What is the probability that two randomly selected bricks will allow Beverly to complete the bottom row of the fireplace? (Hint: Define b_1 = length of the first brick and b_2 = length of the second brick. A variable that is a linear combination of independent variables that have normal distributions also has a normal distribution.)

CR7.6 ● Manatees are large marine mammals found along the coast of Florida. Data on the total number of manatee deaths per year reported by the [Florida Marine Research Center](#) is given in the accompanying table. Construct a time series plot of these data and comment on any trends over time.

Year	Number of Manatee Deaths	Year	Number of Manatee Deaths
1974	7	1985	112
1975	28	1986	117
1976	58	1987	109
1977	109	1988	127
1978	77	1989	164
1979	72	1990	199
1980	56	1991	166
1981	112	1992	156
1982	111	1993	137
1983	71	1994	177
1984	119		

CR7.7 Two shipping services offer overnight delivery of parcels, and both promise delivery before 10 A.M. A mail-order catalog company ships 30% of its overnight packages using service 1 and 70% using service 2. Service 1 fails to meet the 10 A.M. delivery promise 10% of the time, whereas Service 2 fails to deliver by 10 A.M. 8% of the time.

Suppose that you made a purchase from this company and were expecting your package by 10 A.M., but it is late. Which shipping service was more likely to have been used?

CR7.8 The paper “[Examining Communication- and Media-Based Recreational Sedentary Behaviors Among Canadian Youth: Results from the COMPASS Study](#)” (*Preventive Medicine* [2015]: 74–80) estimated that the time spent playing video or computer games by high school boys had a mean of 123.4 minutes per day and a standard deviation of 117.1 minutes per day. Based on this mean and standard deviation, explain why it is not reasonable to think that the distribution of the random variable x = time spent playing video or computer games is approximately normal.

CR7.9 The article “[Doctors Misdiagnose More Women, Blacks](#)” (*San Luis Obispo Tribune*, April 20, 2000) gave the following information, which is based on a large study of more than 10,000 patients treated in emergency rooms in the eastern and midwestern United States:

1. Doctors misdiagnosed heart attacks in 2.1% of all patients.
2. Doctors misdiagnosed heart attacks in 4.3% of black patients.
3. Doctors misdiagnosed heart attacks in 7% of women under 55 years old.

Use the following event definitions:

M = event that a heart attack is misdiagnosed,

B = event that a patient is black, and

W = event that a patient is a woman under 55 years old.

Translate each of the three given statements into probability notation.

CR7.10 The *Cedar Rapids Gazette* (November 20, 1999) reported the following information on compliance with child restraint laws for cities in Iowa:

City	Number of Children Observed	Number Properly Restrained
Cedar Falls	210	173
Cedar Rapids	231	206
Dubuque	182	135
Iowa City (city)	175	140
Iowa City (interstate)	63	47

- a. Use the information provided to estimate the following probabilities:
 - i. The probability that a randomly selected child is properly restrained given that the child is observed in Dubuque.
 - ii. The probability that a randomly selected child is properly restrained given that the child is observed in a city that has “Cedar” in its name.
- b. Suppose that you are observing children in the Iowa City area. Use a tree diagram to illustrate the possible outcomes of an observation that considers both the location of the observation (city or interstate) and whether the child observed was properly restrained.

CR7.11 According to a study conducted by a risk assessment firm (*Associated Press*, December 8, 2005), drivers residing within 1 mile of a restaurant are 30% more likely to be in an accident in a given policy year. Consider the following two events:

A = event that a driver has an accident during a policy year

R = event that a driver lives within 1 mile of a restaurant

Which of the following four probability statements is consistent with the findings of this survey? Justify your choice.

- i. $P(A|R) = 0.3$
- ii. $P(A|R^C) = 0.3$
- iii. $\frac{P(A|R)}{P(A|R^C)} = 0.3$
- iv. $\frac{P(A|R) - P(A|R^C)}{P(A|R^C)} = 0.3$

CR7.12 The article “Men, Women at Odds on Gun Control” (*Cedar Rapids Gazette*, September 8, 1999) included the following statement: “The survey found that 56% of American adults favored stricter gun control laws. Sixty-six percent of women favored the tougher laws, compared with 45% of men.” These figures are based on a large telephone survey conducted by Associated Press Polls. If an adult is selected at random, are the events *selected adult is female* and *selected adult favors stricter gun control* independent events? Explain.

CR7.13 Suppose that a new Internet company Mumble.com requires all employees to take a drug test. Mumble.com can afford only the inexpensive drug test—the one with a 5% false-positive rate and a 10% false-negative rate. (That means that 5% of those who are not using drugs will incorrectly test positive and that 10% of those who are actually using drugs will test negative.) Suppose that 10% of those who work for Mumble.com are using the drugs for which Mumble is checking. (Hint: It may be helpful to draw a tree diagram to answer the questions that follow.)

- a. If one employee is chosen at random, what is the probability that the employee both uses drugs and tests positive?
- b. If one employee is chosen at random, what is the probability that the employee does not use drugs but tests positive anyway?
- c. If one employee is chosen at random, what is the probability that the employee tests positive?
- d. If we know that a randomly chosen employee has tested positive, what is the probability that he or she uses drugs?

CR7.14 Refer to the previous exercise. Suppose that because of the high rate of false-positives for the drug test, Mumble.com has instituted a mandatory independent second test for those who test positive on the first test.

- a. If one employee is selected at random, what is the probability that the selected employee uses drugs and tests positive twice?
- b. If one employee is selected at random, what is the probability that the employee tests positive twice?
- c. If we know that the randomly chosen employee has tested positive twice, what is the probability that he or she uses drugs?
- d. What is the chance that an individual who does use drugs doesn’t test positive twice (either this employee tests negative on the first round and doesn’t need a retest, or this employee tests positive the first time and that result is followed by a negative result on the retest)?
- e. Discuss the benefits and drawbacks of using a retest scheme such as the one proposed in this question.

CR7.15 A chemical supply company currently has in stock 100 pounds of a certain chemical, which it sells to customers in 5-pound lots. Let x = the number of lots ordered

by a randomly chosen customer. The probability distribution of x is as follows:

x	1	2	3	4
$p(x)$	0.2	0.4	0.3	0.1

- a. Calculate the mean value of x .
- b. Calculate the variance and standard deviation of x .

CR7.16 Return to the previous exercise, and let y denote the amount of material (in pounds) left after the next customer's order is shipped. Find the mean and variance of y . (Hint: y is a linear function of x .)

CR7.17 An experiment was conducted to investigate whether a graphologist (a handwriting analyst) could distinguish a normal person's handwriting from that of a psychotic. A well-known expert was given 10 files, each containing handwriting samples from a normal person and from a person diagnosed as psychotic, and asked to identify the psychotic's handwriting.

The graphologist made correct identifications in 6 of the 10 trials (data taken from *Statistics in the Real World, by R. J. Larsen and D. F. Stroup* [New York: Macmillan, 1976]). Does this evidence indicate that the graphologist has an ability to distinguish the handwriting of psychotics? (Hint: What is the probability of correctly guessing six or more times out of 10? Your answer should depend on whether this probability is relatively small or relatively large.)

CR7.18 A machine that produces ball bearings has initially been set so that the true average diameter of the bearings it produces is 0.500 inches. A bearing is acceptable if its diameter is within 0.004 inches of this target value. Suppose, however, that the setting has changed during the course of production, so that the distribution of the diameters produced is well approximated by a normal distribution with mean 0.499 inches and standard deviation 0.002 inches. What percentage of the bearings produced will not be acceptable?

CR7.19 Consider the variable

x = time required for a college student to complete a standardized exam

Suppose that for the population of students at a particular university, the distribution of x is well approximated by a normal distribution with mean 45 minutes and standard deviation 5 minutes.

- a. If 50 minutes is allowed for the exam, what proportion of students at this university would be unable to finish in the allotted time?
- b. How much time should be allowed for the exam if we wanted 90% of the students taking the test to be able to finish in the allotted time?
- c. How much time is required for the fastest 25% of all students to complete the exam?

CR7.20 The accompanying data on x = student-teacher ratio for a random sample of 25 high schools in Maine selected from a population of 85 high schools are consistent with summary values for the state of Maine that appeared in an article in the *Bangor Daily News* (September 22, 2016, bangordailynews.com/2016/09/22/maine-focus/we-discovered-a-surprise-when-we-looked-deeper-into-our-survey-of-maine-principals/?ref=moreInmidcoast). The corresponding normal scores are also shown.

Student-Teacher Ratio (x)	Normal Score
9.0	-1.868
10.0	-1.403
11.0	-1.128
11.2	-0.919
11.6	-0.744
11.7	-0.589
11.8	-0.448
11.9	-0.315
12.0	-0.187
12.1	-0.062
12.5	0.062
12.6	0.187
13.0	0.315
13.2	0.448
13.6	0.589
13.7	0.744
14.0	0.919
14.5	1.128
14.9	1.403
15.0	1.868

Construct a normal probability plot. Is it reasonable to assume that the distribution of student-teacher ratio is approximately normal?

CR7.21 • The following data are a sample of survival times (days from diagnosis) for patients suffering from chronic leukemia of a certain type:

7	47	58	74	177	232	273	285
317	429	440	445	455	468	495	497
532	571	579	581	650	702	715	779
881	900	930	968	1077	1109	1314	1334
1367	1534	1712	1784	1877	1886	2045	2056
2260	2429	2509					

- a. Construct a relative frequency distribution for this data set, and draw the corresponding histogram.
- b. Would you describe this histogram as having a positive or a negative skew?
- c. Would you recommend transforming the data? Explain.

8

Sampling Variability and Sampling Distributions



Kostyantyn Manzhura/EyeEm/Getty Images

T

he inferential methods presented in Chapters 9–15 use sample data to learn about population characteristics. For example, let μ denote the true mean fat content of quarter-pound hamburgers marketed by a national fast-food chain. To learn something about μ , we might select a sample of $n = 50$ hamburgers and determine the fat content for each one. The sample data might produce a mean of $\bar{x} = 28.4$ grams. How close is this sample mean to the population mean, μ ? If we select another sample of 50 quarter-pound burgers and then determine the sample mean fat content, would this second value be near 28.4, or might it be quite different?

These questions can be addressed by studying what is called the *sampling distribution* of \bar{x} . Just as the probability distribution of a numerical variable describes its long-run behavior, the sampling distribution of \bar{x} provides information about the long-run behavior of \bar{x} when samples are selected from a population.

In this chapter, we also consider the sampling distribution of a sample proportion (the fraction of individuals or objects in a sample that have some characteristic of interest). The sampling distribution of a sample proportion, \hat{p} , provides information about the long-run behavior of the sample proportion. This knowledge allows us to make inferences about a population proportion.

LEARNING OBJECTIVES

Students will understand:

- That the value of a sample statistic varies from sample to sample.
- That a sampling distribution describes sample-to-sample variability in the values of a statistic.
- How the standard deviations of the sampling distributions of \bar{x} and \hat{p} are related to sample size.

Students will be able to:

- Describe general properties of the sampling distribution of \bar{x} .
- Describe general properties of the sampling distribution of \hat{p} .
- Determine when the sampling distribution of \bar{x} or \hat{p} is approximately normal.
- Use properties of the sampling distribution to reason informally about the value of a population mean or a population proportion.

SECTION 8.1 Statistics and Sampling Variability

A number calculated from the values in a sample is called a **statistic**. Values of statistics such as the sample mean \bar{x} , the sample median, the sample standard deviation s , or the proportion of individuals in a sample that possess a particular property \hat{p} , are our primary sources of information about various population characteristics.

It is common to learn about the value of a population characteristic by selecting a sample from the population. For example, to learn about the mean credit card balance for students at a university, we might select a sample of 50 students. Each student could be asked about his or her credit card balance, resulting in a value of $x = \text{current balance}$.

We could construct a histogram of the 50 sample x values, and we could view this histogram as a rough approximation of the population distribution of x . In a similar way, we can view the sample mean \bar{x} (the mean of a sample of n values) as an estimate of μ , the mean of the population distribution.

It would be nice if the value of \bar{x} was equal to the value of the population mean μ , but this is not usually the case. Not only will the value of \bar{x} for a particular sample from a population usually differ from μ , but the \bar{x} values from different samples also usually differ from one another. For example, two different samples of 50 student credit card balances will usually result in different \bar{x} values. This sample-to-sample variability makes it challenging to generalize from a sample to the population from which it was selected.

DEFINITIONS

Statistic: A number calculated from values in a sample.

Sampling variability: The observed value of a statistic depends on the particular sample selected from the population and it will vary from sample to sample. This variability is called **sampling variability**.

EXAMPLE 8.1 Exploring Sampling Variability

Consider a small population consisting of the 20 students enrolled in a class. The amount of money (in dollars) each of the 20 students spent on textbooks for the current semester is shown in the following table:

Student	Amount Spent on Books	Student	Amount Spent on Books	Student	Amount Spent on Books
1	367	8	370	15	433
2	358	9	378	16	284
3	442	10	268	17	331
4	361	11	419	18	259
5	375	12	363	19	330
6	395	13	365	20	423
7	322	14	362		

For this population,

$$\mu = \frac{367 + 358 + \dots + 423}{20} = 360.25$$

Suppose we don't know the value of the population mean, so we decide to estimate μ by taking a random sample of five students and calculating \bar{x} , the sample mean amount spent on textbooks. Is this a reasonable thing to do? Is the estimate that results likely to be close to the value of μ , the population mean?

To answer these questions, consider a simple experiment that allows us to study the behavior of the statistic \bar{x} when random samples of size 5 are repeatedly selected. Begin by selecting a random sample of size 5 from this population. This can be done by writing the numbers

from 1 to 20 on otherwise identical slips of paper, mixing them well, and then selecting 5 slips without replacement. The numbers on the slips selected identify which of the 20 students will be included in the sample. Alternatively, either a table of random digits or a random number generator could be used to determine which 5 students should be selected.

We used Minitab to obtain five random numbers between 1 and 20, resulting in 17, 20, 7, 11, and 9. This results in the following sample of amounts spent on books:

331 423 322 419 378

For this sample,

$$\bar{x} = \frac{1873}{5} = 374.60$$

The sample mean is larger than the population mean of \$360.25 by about \$15. Is this difference typical, or is this particular sample mean unusually far away from μ ? Taking more samples will provide additional insight.

Four more random samples (Samples 2–5) from this same population are shown here.

Sample 2		Sample 3		Sample 4		Sample 5	
Student	x	Student	x	Student	x	Student	x
4	361	15	433	20	423	18	259
15	433	12	363	16	284	8	370
12	363	3	442	19	330	9	378
1	367	7	322	1	367	7	322
18	259	18	259	8	370	14	362
\bar{x}	356.60	\bar{x}	363.80	\bar{x}	354.80	\bar{x}	338.20

Because $\mu = 360.25$ was given, we can see the following:

1. The value of \bar{x} varies from one random sample to another (sampling variability).
2. Some samples produced \bar{x} values greater than μ (Samples 1 and 3), whereas others produced values less than μ (Samples 2, 4, and 5).
3. Samples 2, 3, and 4 produced \bar{x} values that were fairly close to the population mean, but Sample 5 resulted in a value that was \$22 less than the population mean.

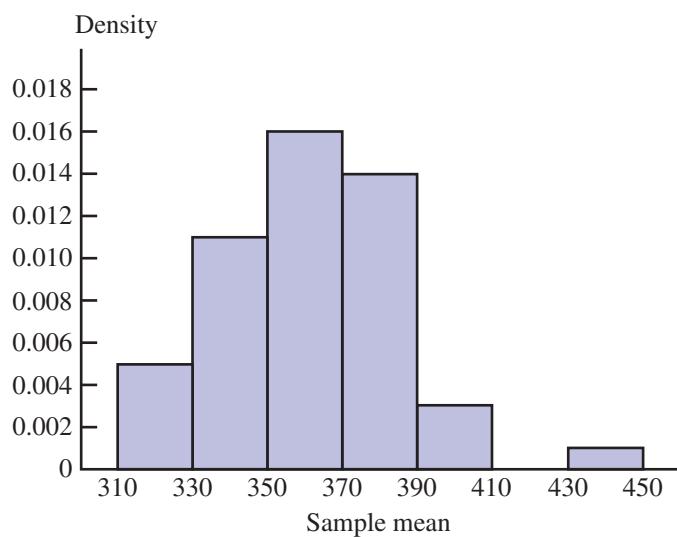
Continuing with the experiment, we selected 45 additional random samples of size $n = 5$. The resulting sample means are as follows:

Sample	\bar{x}	Sample	\bar{x}	Sample	\bar{x}
6	374.6	21	355.0	36	353.4
7	356.6	22	407.2	37	379.6
8	363.8	23	380.0	38	352.6
9	354.8	24	377.4	39	342.2
10	338.2	25	341.2	40	362.6
11	375.6	26	316.0	41	315.4
12	379.2	27	370.0	42	366.2
13	341.6	28	401.0	43	361.4
14	355.4	29	347.0	44	375.0
15	363.8	30	373.8	45	401.4
16	339.6	31	382.8	46	337.0
17	348.2	32	320.4	47	387.4
18	430.8	33	313.6	48	349.2
19	388.8	34	387.6	49	336.8
20	352.8	35	314.8	50	364.6

Figure 8.1 is a density histogram of the 50 sample means. It provides insight about the behavior of \bar{x} . Most samples resulted in \bar{x} values that are reasonably near $\mu = 360.25$, falling between 335 and 395. A few samples, however, produced values that were far from μ . If we were to take a sample of size 5 from this population and use \bar{x} as an estimate of the population mean μ , we should *not* necessarily expect \bar{x} to be close to μ .

FIGURE 8.1

Density histogram of \bar{x} values from the 50 random samples of Example 8.1.



The density histogram in Figure 8.1 visually conveys information about the sampling variability in the statistic \bar{x} . It provides an approximation to the distribution of \bar{x} values that would have been observed if we had considered every different possible sample of size 5 from this population.

In the example just considered, we obtained the approximate sampling distribution of the statistic \bar{x} by considering just 50 different samples. The actual **sampling distribution** comes from considering *all* possible samples of size n .

DEFINITION

Sampling distribution: The distribution that would be formed by considering the value of a sample statistic for every possible sample of a given size from a population.

The sampling distribution of a statistic, such as \bar{x} , provides important information about variability in the values of the statistic. The density histogram of Figure 8.1 is an *approximation* of the sampling distribution of the statistic \bar{x} for samples of size 5 from the population described in Example 8.1.

We could have determined the actual sampling distribution of \bar{x} by considering every possible sample of size 5 from the population of 20 students, calculating the mean for each sample, and then constructing a density histogram of the \bar{x} values. But this would have been a lot of work—there are 15,504 different possible samples of size 5. And, for more realistic cases with larger population and larger sample sizes, the situation becomes even worse because there are so many possible samples that must be considered.

Fortunately, as we look at more examples in the sections that follow, patterns emerge that enable us to describe some important aspects of the sampling distributions for some statistics without actually having to look at all possible samples.

EXERCISES 8.1 - 8.9

- 8.1** Explain the difference between a population characteristic and a statistic.
- 8.2** What is the difference between \bar{x} and μ ? between s and σ ?
- 8.3** For each of the following statements, identify the number that appears in boldface type as the value of either a population characteristic or a statistic:
- A department store reports that **84%** of all customers who use the store's credit plan pay their bills on time.
 - A sample of 100 students at a large university has a mean age of **24.1** years.
 - The Department of Motor Vehicles reports that **22%** of all vehicles registered in a particular state are imports.
 - A hospital reports that based on the 10 most recent cases, the mean length of stay for surgical patients is **6.4** days.
 - A consumer group, after testing 100 batteries of a certain brand, reported an average life of **63** hours of use.
- 8.4** Consider a population consisting of the following five values, which represent the number of video downloads during the academic year for each of five housemates:
- 8 14 16 10 11
- Compute the mean of this population.
 - Select a random sample of size 2 by writing the five numbers in this population on slips of paper, mixing them, and then selecting two. Calculate the mean for your sample.
 - Repeatedly select random samples of size 2, and calculate the \bar{x} value for each sample until you have the \bar{x} values for 25 samples.
 - Construct a density histogram using the 25 \bar{x} values. Are most of the \bar{x} values near the population mean? Do the \bar{x} values differ a lot from sample to sample, or do they tend to be similar? (Hint: See Example 8.1.)
- 8.5** Select 10 additional random samples of size 5 from the population of 20 students given in Example 8.1, and calculate the mean amount spent on books for each of the 10 samples. Are the \bar{x} values consistent with the results of the sampling experiment summarized in Figure 8.1?
- 8.6** Suppose that the sampling experiment described in Example 8.1 had used samples of size 10 rather than size 5.
- If 50 samples of size 10 were selected, the \bar{x} value for each sample calculated, and a density

histogram constructed, how do you think this histogram would differ from the density histogram constructed for samples of size 5 (Figure 8.1)?

- b.** In what way do you think the density histograms would be similar?

- 8.7** Consider the following population: {1, 2, 3, 4}. For this population the mean is

$$\mu = \frac{1 + 2 + 3 + 4}{4} = 2.5$$

Suppose that a random sample of size 2 is to be selected without replacement from this population. There are 12 possible samples (assuming that the order in which observations are selected is taken into account):

1, 2	1, 3	1, 4	2, 1	2, 3	2, 4
3, 1	3, 2	3, 4	4, 1	4, 2	4, 3

- Calculate the sample mean for each of the 12 possible samples.
- Use the sample means to construct the sampling distribution of \bar{x} . Display the sampling distribution as a density histogram.
- Suppose that a random sample of size 2 is to be selected, but this time sampling will be done with replacement. Using a method similar to that of Part (a), construct the sampling distribution of \bar{x} . (Hint: There are 16 different possible samples in this case.)
- In what ways are the two sampling distributions of Parts (b) and (c) similar? In what ways are they different?

- 8.8** Simulate sampling from the population of Exercise 8.7 by using four slips of paper individually marked 1, 2, 3, and 4. Select a sample of size 2 without replacement, and calculate \bar{x} . Repeat this process 50 times, and construct a density histogram of the 50 \bar{x} values. How does this sampling distribution compare to the actual sampling distribution of \bar{x} from Exercise 8.7, Part (b)?

- 8.9** Consider the following population: {2, 3, 3, 4, 4}. The value of μ is 3.2, but suppose that this is not known to an investigator. Three possible statistics for estimating μ are

Statistic 1: the sample mean, \bar{x}

Statistic 2: the sample median

Statistic 3: the average of the largest and the smallest values in the sample

A random sample of size 3 will be selected without replacement. If we disregard the order in which

the observations are selected, there are 10 possible samples that might result (writing 3 and 3*, 4 and 4* to distinguish the two 3's and the two 4's in the population):

$$\begin{array}{cccccc} 2, 3, 3^* & 2, 3, 4 & 2, 3, 4^* & 2, 3^*, 4 & 2, 3^*, 4^* \\ 2, 4, 4^* & 3, 3^*, 4 & 3, 3^*, 4^* & 3, 4, 4^* & 3^*, 4, 4^* \end{array}$$

- a. For each of these 10 samples, calculate the values of Statistics 1, 2, and 3.
- b. Construct the sampling distribution of each of these statistics.
- c. Which statistic would you recommend for estimating μ ? Explain your reasoning.

SECTION 8.2 The Sampling Distribution of a Sample Mean

When we want to learn about the value of a population mean μ , it is natural to consider the sample mean \bar{x} . To understand how inferential procedures based on \bar{x} work, we must first study how \bar{x} varies in value from one sample to another.

The behavior of \bar{x} is described by its sampling distribution. The sample size n and characteristics of the population—its shape, mean value μ , and standard deviation σ —are important in determining properties of the sampling distribution of \bar{x} .

It is helpful first to consider the results of some sampling experiments. In Examples 8.2 and 8.3, we start with a known population distribution, fix a sample size n , and select 500 different random samples of this size. We then calculate \bar{x} for each sample and construct a density histogram of these 500 \bar{x} values. Because 500 is reasonably large (a reasonably long sequence of samples), the density histogram of the \bar{x} values should closely resemble the actual sampling distribution of \bar{x} (which would be obtained by considering all possible samples).

We then repeat the experiment for several different values of n to see how the choice of sample size affects the sampling distribution. This enables us to identify some patterns that will be helpful in understanding important properties of the sampling distribution of \bar{x} .

EXAMPLE 8.2 Blood Platelet Volume

The paper “[Mean Platelet Volume Could Be Possible Biomarker in Early Diagnosis and Monitoring of Gastric Cancer](#)” (*Platelets* [2014]: 592–594) includes data that suggest that the distribution of platelet volume for patients who have gastric cancer is approximately normal with mean $\mu = 8.3$ and standard deviation $\sigma = 0.8$.

Figure 8.2 shows a normal curve centered at 8.3, the mean value of platelet volume. The value of the population standard deviation, 0.8, determines the extent to which the x distribution spreads out about its mean value.

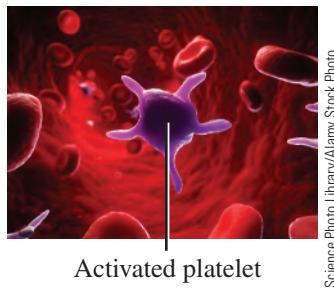
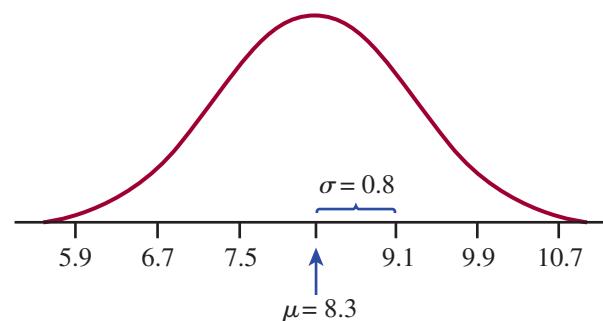
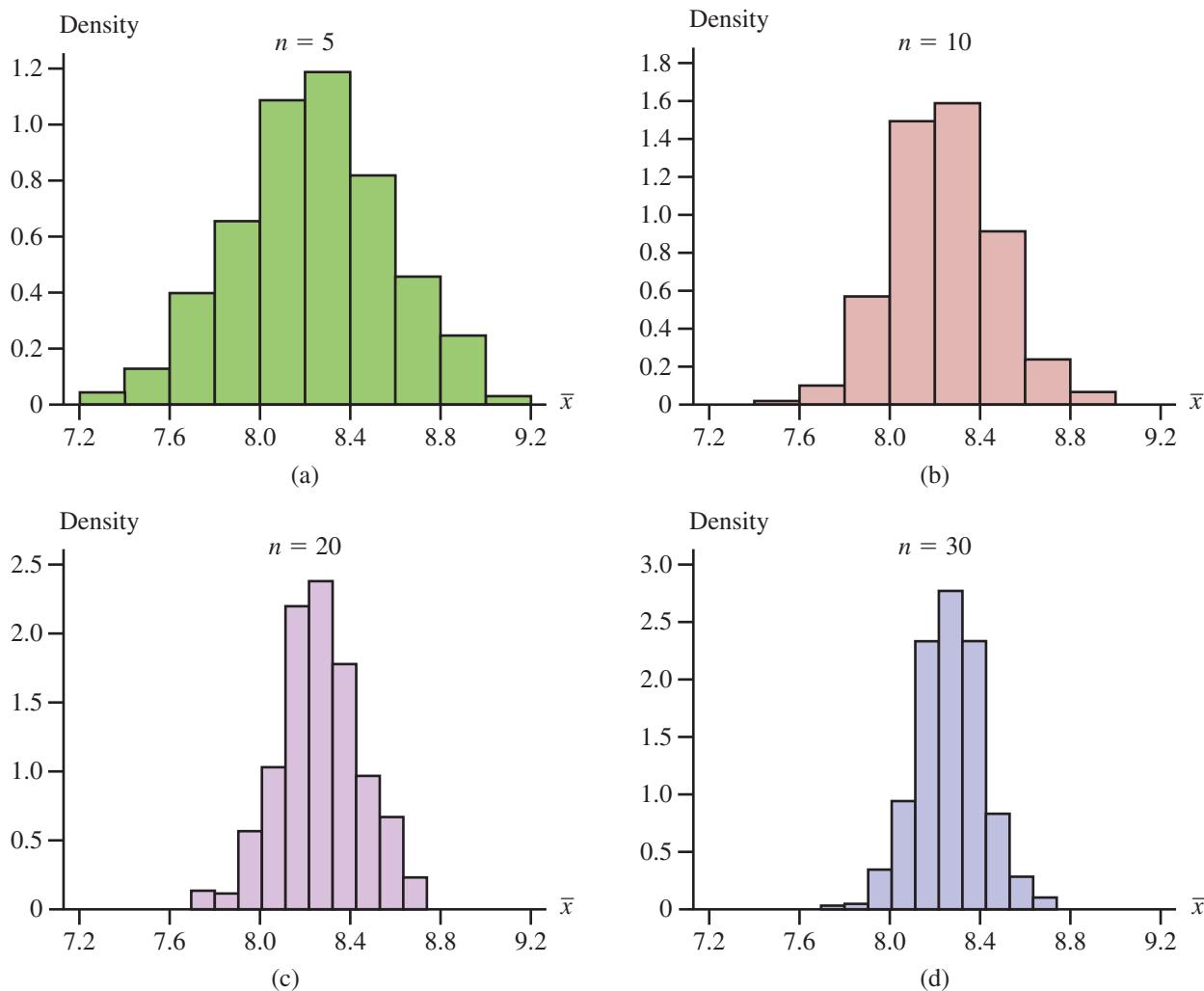


FIGURE 8.2
Normal distribution of platelet size
($\mu = 8.3$ and $\sigma = 0.8$).



We used Minitab to select 500 random samples from this population distribution, with each sample consisting of $n = 5$ observations. A density histogram of the resulting 500 \bar{x} values appears in Figure 8.3(a). This procedure was repeated for samples of size $n = 10$, $n = 20$, and $n = 30$. The resulting density histograms of the \bar{x} values are displayed in Figure 8.3(b)–(d).

**FIGURE 8.3**

Density histograms for \bar{x} based on 500 samples, each consisting of n observations, for Example 8.2: (a) $n = 5$; (b) $n = 10$; (c) $n = 20$; (d) $n = 30$.

Notice the shapes of the histograms. Each of the four histograms is approximately normal in shape. The resemblance would be even more striking if each histogram had been based on many more than 500 \bar{x} values.

Second, notice that each histogram is centered approximately at 8.3, the mean of the population being sampled. Had the histograms been constructed using \bar{x} values from every possible sample, their centers would have been exactly equal to the population mean, 8.3.

The final aspect of the histograms to note is their variability relative to one another. The smaller the value of n , the greater the extent to which the sampling distribution spreads out around the population mean value. This is why the histograms for $n = 20$ and $n = 30$ are based on narrower class intervals than those for the two smaller sample sizes. For the larger sample sizes, most of the \bar{x} values are quite close to 8.3. This is the effect of averaging. When n is small, a single unusual x value can result in an \bar{x} value far from the center. With a larger sample size, any unusual x values, when averaged with the other sample values, still tend to yield an \bar{x} value close to μ . Combining these insights yields a result that should appeal to your intuition: \bar{x} values based on a large sample will tend to be closer to μ than \bar{x} values based on a small sample.

EXAMPLE 8.3 Time to First Goal in Hockey

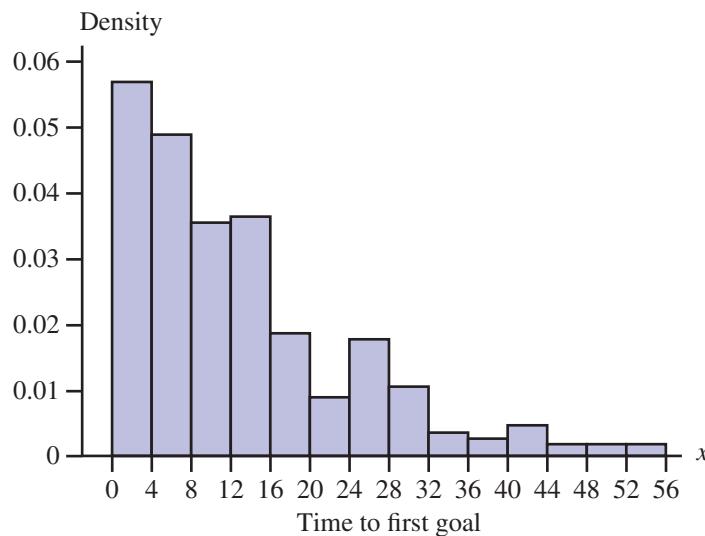
Now consider properties of the \bar{x} distribution when the population is quite skewed (and thus very unlike a normal distribution). The paper “[Is the Overtime Period in an NHL Game Long Enough?](#)” (*American Statistician* [2008]: 151–154) gave data on the time (in minutes)

from the start of the game to the first goal scored for the 281 regular season games from the 2005–2006 season that went into overtime.

Figure 8.4 displays a density histogram of the data (based on a graph that appeared in the paper). The histogram has a long upper tail, indicating that the first goal is scored in the first 20 minutes of most games, but that for some games, the first goal is not scored until much later in the game.

FIGURE 8.4

The population distribution for Example 8.3 ($\mu = 13$).



If we think of the 281 values as a population, the histogram in Figure 8.4 shows the distribution of values in that population. The skewed shape makes identification of the mean value from the picture more difficult than for a normal distribution. The mean of the 281 values was calculated to be $\mu = 13$ minutes.

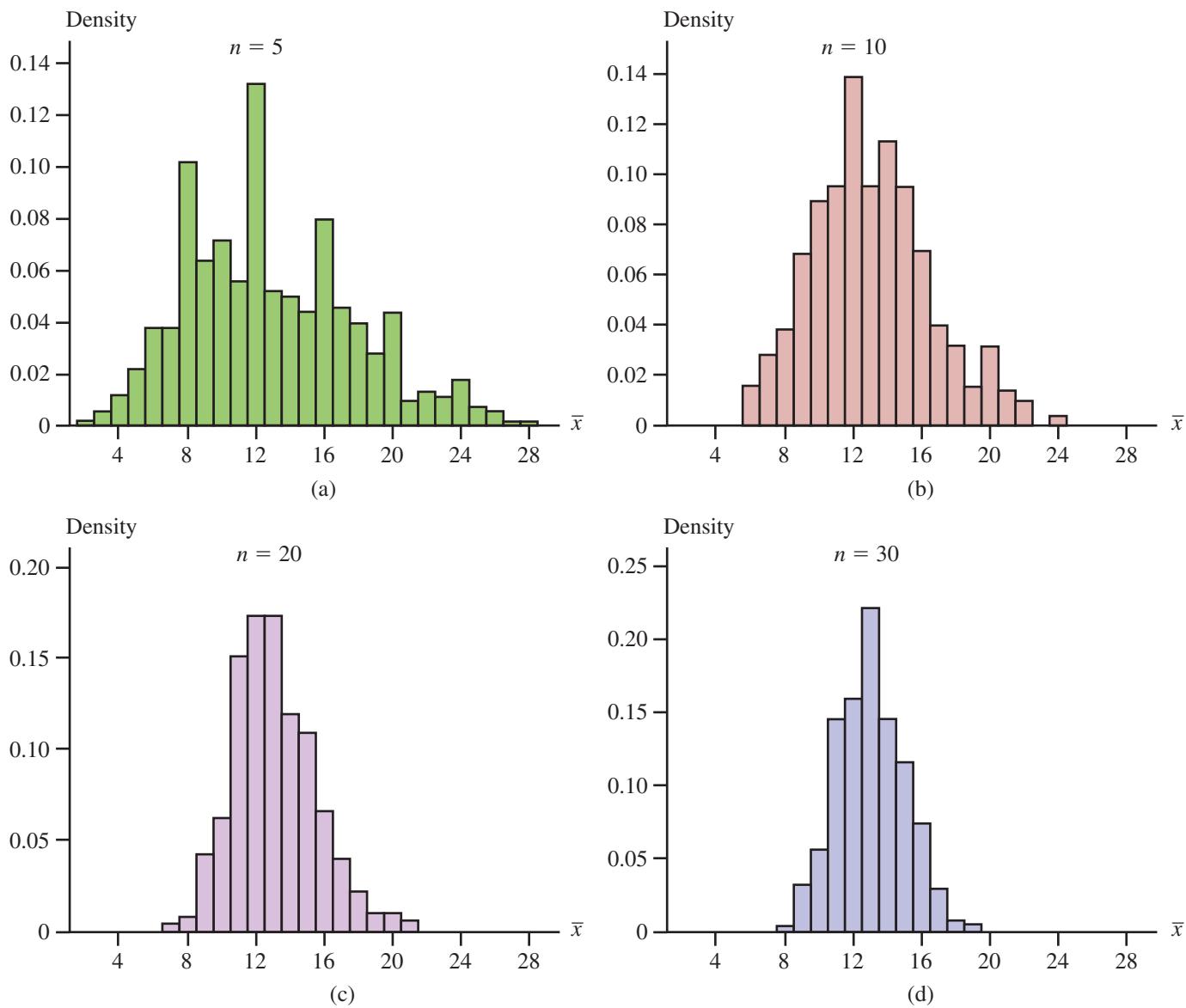
For each of the sample sizes $n = 5, 10, 20$, and 30 , we selected 500 random samples of size n . This was done with replacement to approximate more nearly the usual situation, in which the sample size n is only a small fraction of the population size. We then constructed a density histogram of the 500 \bar{x} values for each of the four sample sizes. These histograms are displayed in Figure 8.5.

As with samples from a normal population, the averages of the 500 \bar{x} values for the four different sample sizes are all close to the population mean $\mu = 13$. If each histogram had been using the sample means from all possible samples rather than just 500 of them, each histogram would have been centered at exactly 13.

Comparison of the four \bar{x} histograms in Figure 8.5 also shows that as n increases, the histogram's spread around its center decreases. This was also true of increasing sample sizes for the normal population in the previous example: \bar{x} is less variable (varies less from sample to sample) for a large sample size than it is for a small sample size.

One aspect of these histograms distinguishes them from the distribution of \bar{x} based on a sample from a normal population. They are skewed and differ in shape more, but they become progressively more symmetric as the sample size increases. We can also see that for $n = 30$, the histogram has a shape much like a normal curve. Again this is the effect of averaging. Even when n is large, one of the few large x values in the population doesn't appear in the sample very often. When one does appear, its contribution to \bar{x} is outweighed by the contributions of more typical sample values.

The normal shape of the histogram for $n = 30$ is what is predicted by the Central Limit Theorem, which will be introduced shortly. According to this theorem, even if the population distribution does not look like a normal distribution, the sampling distribution of \bar{x} is approximately normal in shape when the sample size n is reasonably large.

**FIGURE 8.5**

Four density histograms of 500 \bar{x} values for Example 8.3: (a) $n = 5$; (b) $n = 10$; (c) $n = 20$; (d) $n = 30$.

General Properties of the Sampling Distribution of \bar{x}

Examples 8.2 and 8.3 suggest that for any n , the center of the \bar{x} distribution (the mean value of \bar{x}) coincides with the mean of the population being sampled. We also see that the variability of the \bar{x} distribution decreases as n increases. This indicates that the standard deviation of \bar{x} is smaller for large n than for small n . The density histograms in Figures 8.3 and 8.5 also suggest that in some cases, the \bar{x} distribution is approximately normal in shape. These observations are stated more formally in the following general rules.

General Properties of the Sampling Distribution of \bar{x}

Notation

\bar{x}	Sample mean
n	Sample size
μ	Population mean
σ	Population standard deviation

(continued)

- $\mu_{\bar{x}}$ Mean of the sampling distribution of \bar{x}
 $\sigma_{\bar{x}}$ Standard deviation of the sampling distribution of \bar{x}

When random samples are selected from a population, the following are properties of the sampling distribution of \bar{x} :

Property 1. $\mu_{\bar{x}} = \mu$

Property 2. $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

This formula is exact if the population is infinite, and is approximately correct if the population is finite and no more than 10% of the population is included in the sample.

Property 3. When the population distribution is normal, the sampling distribution of \bar{x} is also normal for any sample size n .

Property 4. (Central Limit Theorem) When n is sufficiently large, the sampling distribution of \bar{x} is well approximated by a normal curve, even when the population distribution is not normal.

Property 1, $\mu_{\bar{x}} = \mu$, states that the sampling distribution of \bar{x} is always centered at the mean of the population sampled.

Property 2, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, not only states that the variability of the sampling distribution of \bar{x} decreases as n increases, but also gives a precise relationship between the standard deviation of the \bar{x} distribution and the population standard deviation and sample size. For example, when $n = 4$,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{4}} = \frac{\sigma}{2}$$

so the \bar{x} distribution has a standard deviation only half as large as the population standard deviation.

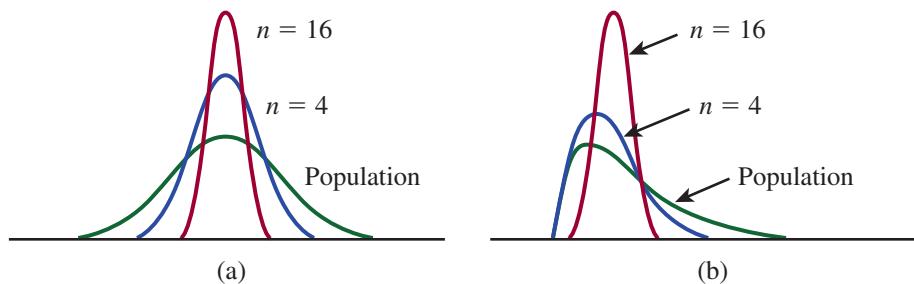
Properties 3 and 4 specify when the \bar{x} distribution is normal (when the population is normal) or approximately normal (when the sample size is large).

Figure 8.6 illustrates these rules by showing several \bar{x} distributions superimposed over a graph of the population distribution.

FIGURE 8.6

Population distribution and sampling distributions of \bar{x} :

- (a) symmetric population;
(b) skewed population.



The Central Limit Theorem, Property 4, states that when n is sufficiently large, the sampling distribution of \bar{x} is approximately normal for any population distribution. This result has enabled statisticians to develop procedures for drawing conclusions about a population mean μ using large samples, even when the shape of the population distribution is unknown.

Recall that a variable is standardized by subtracting the mean value and then dividing by its standard deviation. Using Properties 1 and 2 to standardize \bar{x} gives an important consequence of the last two properties.

If n is large or the population distribution is normal, the standardized variable

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has (at least approximately) a standard normal (z) distribution.

Applying the Central Limit Theorem requires a rule of thumb for deciding when n is large enough. Look back at Figure 8.5, which shows the approximate sampling distribution of \bar{x} for $n = 5, 10, 20$, and 30 when the population distribution is quite skewed. The histogram for $n = 5$ is not well described by a normal curve, and this is still true of the histogram for $n = 10$, particularly in the tails of the histogram. Among the four histograms, only the histogram for $n = 30$ has a shape that is reasonably well described by a normal curve.

On the other hand, when the population distribution is normal, the sampling distribution of \bar{x} is normal for any n . If the population distribution is somewhat skewed but not to the extent of Figure 8.4, we might expect the sampling distribution of \bar{x} to be a bit skewed for $n = 5$ but quite well described by a normal curve for n as small as 10 or 15.

How large n must be for the sampling distribution of \bar{x} to be approximately normal depends on how much the population distribution differs from a normal distribution. The closer the population distribution is to a normal distribution, the smaller the value of n necessary for the Central Limit Theorem approximation to be accurate.

Many statisticians recommend the following conservative rule:

The Central Limit Theorem can safely be applied if n is greater than or equal to 30.

EXAMPLE 8.4 Courting Scorpion Flies

Understand the context ➤

The authors of the paper “Should I Stay or Should I Go? Condition- and Status-Dependent Courtship Decisions in the Scorpion Fly *Panorpa Cognata*” (*Animal Behaviour* [2009]: 491–497) studied the courtship behavior of mating scorpion flies. One variable of interest was x = courtship time, which was defined as the time from the beginning of a female-male interaction until mating.

Data from the paper suggest that it is reasonable to think that the mean and standard deviation of x are $\mu = 117.1$ minutes and $\sigma = 109.1$ minutes. Notice that the distribution of courtship times cannot be normal because for a normal distribution centered at 117.1 and with such a large standard deviation, it would not be uncommon to observe negative values, but courtship time can't have a negative value.

The sampling distribution of

\bar{x} = mean courtship time for a random sample of 20 scorpion fly mating pairs

would have mean

$$\mu_{\bar{x}} = \mu = 117.1 \text{ minutes}$$

This tells you that the sampling distribution is centered at 117.1. The standard deviation of \bar{x} is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{109.1}{\sqrt{20}} = 24.40$$

which is smaller than the population standard deviation σ . Because the population distribution is not normal and because the sample size is smaller than 30, it is not reasonable to assume that the sampling distribution of \bar{x} is approximately normal in shape.

EXAMPLE 8.5 | Soft-Drink Volumes

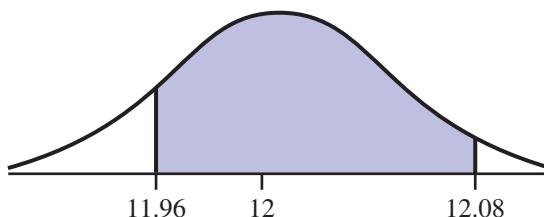
Understand the context ➤

A soft-drink bottler claims that, on average, cans contain 12 ounces of soda. Let x denote the actual volume of soda in a randomly selected can. Suppose that x is normally distributed with $\sigma = 0.16$ ounces. Sixteen cans are to be selected, and the soda volume will be determined for each one. Let \bar{x} denote the resulting sample mean soda volume. Because the x distribution is normal, the sampling distribution of \bar{x} is also normal. *If the bottler's claim is correct, the sampling distribution of \bar{x} has a mean value of $\mu_{\bar{x}} = \mu = 12$ and a standard deviation of*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.16}{\sqrt{16}} = 0.04$$

Formulate a plan ➤

To calculate a probability involving \bar{x} , we standardize by subtracting the mean value, 12, and dividing by the standard deviation (of \bar{x}), which is 0.04. For example, the probability that the sample mean soda volume is between 11.96 ounces and 12.08 ounces is the area between 11.96 and 12.08 under the normal curve with mean 12 and standard deviation 0.04, as shown in the following figure.



This area is calculated by first standardizing the interval limits:

Do the work ➤

$$\text{Lower limit } a^* = \frac{11.96 - 12}{0.04} = -1.0$$

$$\text{Upper limit } b^* = \frac{12.08 - 12}{0.04} = 2.0$$

Then (using Appendix Table 2)

$$\begin{aligned} P(11.96 \leq \bar{x} \leq 12.08) &= \text{area under the } z \text{ curve between } -1.0 \text{ and } 2.0 \\ &= (\text{area to the left of } 2.0) - (\text{area to the left of } -1.0) \\ &= 0.9772 - 0.1587 \\ &= 0.8185 \end{aligned}$$

The probability that the sample mean soda volume is at most 11.9 ounces is

$$\begin{aligned} P(\bar{x} \leq 11.9) &= P\left(z \leq \frac{11.9 - 12}{0.04} = -2.5\right) \\ &= (\text{area under the } z \text{ curve to the left of } -2.5) = 0.0062 \end{aligned}$$

Interpret the results ➤

If the x distribution is as described and the bottler's claim is correct, a sample mean soda volume based on a random sample of 16 observations is less than 11.9 ounces for fewer than 1% of all such samples. If we observed an \bar{x} value that is smaller than 11.9 ounces, this would cast doubt on the bottler's claim that the average soda volume is 12 ounces.

EXAMPLE 8.6 | Fat Content of Hot Dogs

Understand the context ➤

A hot dog manufacturer claims that one of its brands of hot dogs has an average fat content of $\mu = 18$ grams per hot dog. Consumers of this brand would probably not be disturbed if the mean is less than 18 but would be unhappy if it is greater than 18. Let x denote the fat content of a randomly selected hot dog, and suppose that σ , the standard deviation of the x distribution, is 1.

An independent testing organization is asked to analyze a random sample of 36 hot dogs. Let \bar{x} be the average fat content for this sample. The sample size, $n = 36$, is large enough to rely on the Central Limit Theorem and so it is reasonable to regard the \bar{x} distribution as approximately normal. The standard deviation of the \bar{x} distribution is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{36}} = 0.167$$

If the manufacturer's claim is correct, we know that $\mu_{\bar{x}} = \mu = 18$ grams.

Suppose that the sample resulted in a mean of $\bar{x} = 18.4$ grams. Does this result suggest that the manufacturer's claim is incorrect?

Formulate a plan ➤

We can answer this question by looking at the sampling distribution of \bar{x} . Because of sampling variability, even if $\mu = 18$, we know that \bar{x} will not usually be exactly 18. But, is it likely that we would see a sample mean at least as great as 18.4 when the population mean is really 18?

If the company's claim is correct,

Do the work ➤

$$\begin{aligned} P(\bar{x} \geq 18.4) &\approx P\left(z \geq \frac{18.4 - 18}{0.167}\right) \\ &= P(z \geq 2.40) \\ &= \text{area under the } z \text{ curve to the right of 2.40} \\ &= 1 - 0.9918 = 0.0082 \end{aligned}$$

Interpret the results ➤

Values of \bar{x} at least as great as 18.4 will be observed only about 0.82% of the time when a random sample of size 36 is taken from a population with mean 18 and standard deviation 1. The value $\bar{x} = 18.4$ is enough greater than 18 that we would question the manufacturer's claim.

Other Cases

We now know a great deal about the sampling distribution of \bar{x} in two cases: for a normal population distribution and for a large sample size. What happens when the population distribution is not normal and n is small? Although it is still true that $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, unfortunately there is no general result about the shape of the distribution.

When the objective is to draw a conclusion about the mean of such a population, one way to proceed is to make an assumption about the shape of the population distribution. Statisticians have proposed and studied a number of such models. Theoretical methods or simulation can be used to describe the sampling distribution of \bar{x} corresponding to the assumed model.

An alternative strategy is to use one of the transformations presented in Chapter 7 to create a data set that more closely resembles a sample from a normal population and then to base inferences on the transformed data. Yet another strategy is to use procedures based on a statistic other than \bar{x} . Consult a statistician or a more advanced text for more information.

EXERCISES 8.10 - 8.22

- 8.10** A random sample is selected from a population with mean $\mu = 100$ and standard deviation $\sigma = 10$. Determine the mean and standard deviation of the sampling distribution of \bar{x} for each of the following sample sizes:

- a. $n = 9$
- b. $n = 15$
- c. $n = 36$
- d. $n = 50$
- e. $n = 100$
- f. $n = 400$

- 8.11** For which of the sample sizes given in the previous exercise would it be reasonable to think that the sampling distribution of \bar{x} is approximately normal in shape?

- 8.12** Explain the difference between σ and $\sigma_{\bar{x}}$ and between μ and $\mu_{\bar{x}}$.

- 8.13** Suppose that a random sample of size 64 is to be selected from a population with mean 40 and standard deviation 5.
- What are the mean and standard deviation of the sampling distribution of \bar{x} ? Describe the shape of the sampling distribution of \bar{x} .
 - What is the approximate probability that \bar{x} will be within 0.5 of the population mean μ ? (Hint: See Examples 8.5 and 8.6.)
 - What is the approximate probability that \bar{x} will differ from μ by more than 0.7?
- 8.14** The time that a randomly selected individual waits for an elevator in an office building has a uniform distribution over the interval from 0 to 1 minute. For this distribution $\mu = 0.5$ and $\sigma = 0.289$.
- Let \bar{x} be the sample mean waiting time for a random sample of 16 individuals. What are the mean and standard deviation of the sampling distribution of \bar{x} ?
 - Answer Part (a) but for a random sample of 50 individuals.
 - Draw a picture of the approximate sampling distribution of \bar{x} when $n = 50$.
- 8.15** Let x denote the time (in minutes) that it takes a fifth-grade student to read a certain passage. Suppose that the mean value and standard deviation of x are $\mu = 2$ minutes and $\sigma = 0.8$ minutes, respectively.
- If \bar{x} is the sample mean time for a random sample of $n = 9$ students, where is the sampling distribution of \bar{x} centered, and how much does it spread out around the center (as described by its standard deviation)?
 - Repeat Part (a) for a sample of size of $n = 20$ and again for a sample of size $n = 100$. How do the centers and variability of the three \bar{x} distributions compare to one another?
 - Which of the sample sizes in Part (b) would be most likely to result in an \bar{x} value close to μ , and why?
- 8.16** In the library on a university campus, there is a sign in the elevator that indicates a limit of 16 persons. In addition, there is a weight limit of 2500 pounds. Assume that the average weight of students, faculty, and staff on campus is 150 pounds, that the standard deviation is 27 pounds, and that the distribution of weights of individuals on campus is approximately normal. Suppose a random sample of 16 persons from the campus will be selected.
- What is the mean of the sampling distribution of \bar{x} ?
 - What is the standard deviation of the sampling distribution of \bar{x} ?
 - What mean weights for a sample of 16 people will result in the total weight exceeding the weight limit of 2500 pounds?
- 8.17** Suppose that the mean value of interpupillary distance (the distance between the pupils of the left and right eyes) for adult males is 65 mm and that the population standard deviation is 5 mm.
- If the distribution of interpupillary distance is normal and a random sample of $n = 25$ adult males is to be selected, what is the probability that the sample mean distance \bar{x} for these 25 will be between 64 and 67 mm? at least 68 mm? (Hint: See Examples 8.5 and 8.6.)
 - Suppose that a random sample of 100 adult males is to be obtained. Without assuming that interpupillary distance is normally distributed, what is the approximate probability that the sample mean distance will be between 64 and 67 mm? at least 68 mm?
- 8.18** Suppose that a sample of size 100 is to be drawn from a population with standard deviation 10.
- What is the probability that the sample mean will be within 1 of the value of μ ?
 - For this example ($n = 100$, $\sigma = 10$), complete each of the following statements by computing the appropriate value:
 - Approximately 95% of the time, \bar{x} will be within _____ of μ .
 - Approximately 0.3% of the time, \bar{x} will be farther than _____ from μ .
- 8.19** A manufacturing process is designed to produce bolts with a 0.5-inch diameter. Once each day, a random sample of 36 bolts is selected and the bolt diameters are recorded. If the resulting sample mean is less than 0.49 inches or greater than 0.51 inches, the process is shut down for adjustment. The standard deviation for diameter is 0.02 inches. What is the probability that the manufacturing line will be shut down unnecessarily? (Hint: Find the probability of observing an \bar{x} in the shutdown range when the actual process mean really is 0.5 inches.)
- 8.20** College students with checking accounts typically write relatively few checks in any given month, whereas adults who are not students typically write many more checks during a month. Suppose that 50% of a bank's accounts are held by students and that 50% are held by adults who are not students. Let x denote the number of checks written in a given month by a randomly selected bank customer.
- Make a sketch of what the probability distribution of x might look like.
 - Suppose that the mean value of x is 22.0 and that the standard deviation is 16.5. A random sample of $n = 100$ customers is to be selected. Where is the sampling distribution of \bar{x} centered, and

- what is the standard deviation of the sampling distribution of \bar{x} ? Sketch a rough picture of the sampling distribution.
- c. Referring to Part (b), what is the approximate probability that \bar{x} is at most 20? at least 25?
- 8.21** An airplane with room for 100 passengers has a total baggage limit of 6000 pounds. Suppose that the total weight of the baggage checked by an individual passenger is a random variable x with a mean value of 50 pounds and a standard deviation of 20 pounds. If 100 passengers will board a flight, what is the approximate probability that the total weight of their baggage will exceed the limit? (Hint: With $n = 100$, the total weight exceeds the limit when the mean weight \bar{x} exceeds 6000/100.)
- 8.22** The thickness (in millimeters) of the coating applied to hard drives is one characteristic that determines the usefulness of the product. When no unusual circumstances are present, the thickness (x) has a normal distribution with a mean of 2 mm and a standard deviation of 0.05 mm. Suppose that the process will be monitored by selecting a random sample of 16 drives from each shift's production and determining \bar{x} , the mean coating thickness for the sample.
- a. Describe the sampling distribution of \bar{x} for a random sample of size 16.
- b. When no unusual circumstances are present, we expect \bar{x} to be within $3\sigma_{\bar{x}}$ of 2 mm, the desired value. An \bar{x} value farther from 2 mm than $3\sigma_{\bar{x}}$ is interpreted as an indication of a problem that needs attention. Calculate $2 \pm 3\sigma_{\bar{x}}$.
- c. Referring to Part (b), what is the probability that a sample mean will be outside $2 \pm 3\sigma_{\bar{x}}$ just by chance (that is, when there are no unusual circumstances)?
- d. Suppose that a machine used to apply the coating is out of adjustment, resulting in a mean coating thickness of 2.05 mm. What is the probability that a problem will be detected when the next sample is taken? (Hint: This will occur if $\bar{x} > 2 + 3\sigma_{\bar{x}}$ or $\bar{x} < 2 - 3\sigma_{\bar{x}}$ when $\mu = 2.05$.)

SECTION 8.3 The Sampling Distribution of a Sample Proportion

The objective of many statistical investigations is to draw a conclusion about the proportion of individuals or objects in a population that possess a specified property. For example, cell phones that don't require repair during the warranty period or coffee drinkers who regularly drink decaffeinated coffee. Because proportions are of interest in a variety of situations, it is helpful to introduce some general terminology. An individual or object that possesses the property of interest is labeled a success (S), and one that does not possess the property is termed a failure (F).

The letter p is used to denote the proportion of successes in the population. The value of p is a number between 0 and 1, and $100p$ is the percentage of successes in the population. For example, if $p = 0.75$, 75% of the population members are successes, and if $p = 0.01$, the population contains only 1% successes and 99% failures.

The value of p is usually unknown. When a random sample of size n is selected from a population that consists of successes and failures, some of the individuals in the sample are successes, and the rest are failures. The statistic that is used to draw conclusions about p is \hat{p} , the **sample proportion of successes**:

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

For example, if there are three successes in a sample of size 5, then $\hat{p} = 3/5 = 0.6$.

Drawing conclusions about p requires first learning about properties of the sampling distribution of the statistic \hat{p} . For example, when $n = 5$, the six possible values of \hat{p} are 0, 0.2 (from 1/5), 0.4, 0.6, 0.8, and 1. The sampling distribution of \hat{p} gives the probability of each of these six possible values. These probabilities are the long-run proportions of the time that these values would occur if random samples with $n = 5$ were selected over and over again.

As we did for the sample mean, we will look at some simulation experiments to develop an intuitive understanding of the sampling distribution of the sample proportion before stating general properties. In each example, 500 random samples of size n are selected from a population having a specified value of p . We calculate \hat{p} for each sample and then construct a density histogram of the 500 \hat{p} values.

EXAMPLE 8.7 STEM College Students

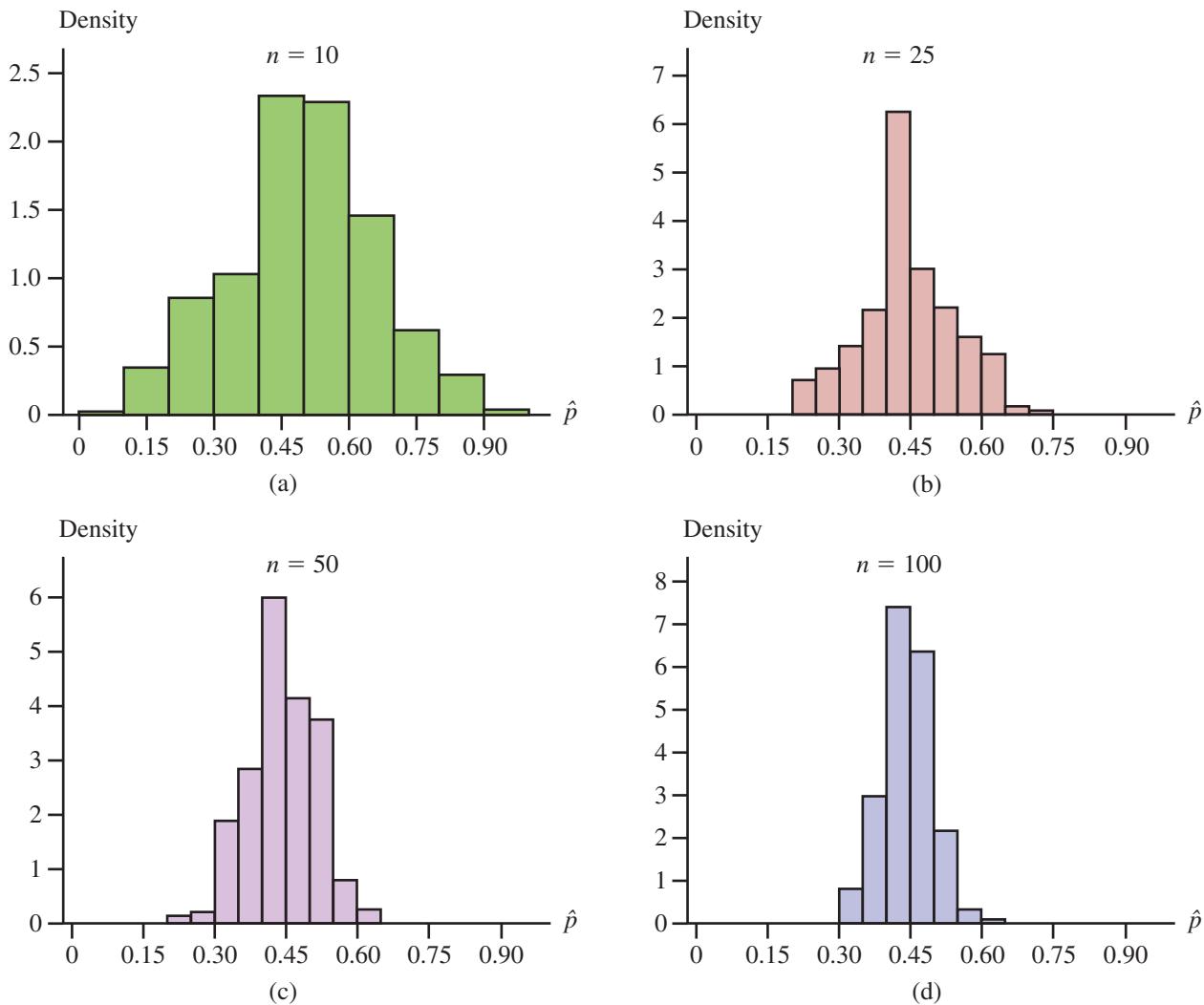
In the fall of 2015, there were 20,944 students enrolled at California Polytechnic State University, San Luis Obispo. Of these students, 9082 (43.4%) were enrolled in a science, technology, engineering or mathematics (STEM) major. To illustrate properties of the sampling distribution of a sample proportion, we will simulate sampling from this Cal Poly student population.

With success denoting a STEM major, the proportion of successes in the population is $p = 0.434$. A statistical software package was used to select 500 samples of size $n = 10$, then 500 samples of size $n = 25$, then 500 samples with $n = 50$, and finally 500 samples with $n = 100$. Density histograms of the 500 values of \hat{p} for each of the four sample sizes are displayed in Figure 8.7.

The most noticeable feature of the histogram shapes is that all are approximately symmetric. All four histograms appear to be centered at roughly 0.434, the value of p for the population sampled. Had the histograms been based on all possible samples, each histogram would have been centered at exactly 0.434. Finally, as was the case with the sampling distribution of \bar{x} , the histograms spread out more for small sample sizes than for large sample sizes. Not surprisingly, the value of \hat{p} based on a large sample size tends to be closer to p , the population proportion of successes, than does \hat{p} from a small sample.

FIGURE 8.7

Density histograms for 500 values of \hat{p} ($p = 0.434$) for Example 8.7:
 (a) $n = 10$; (b) $n = 25$; (c) $n = 50$;
 (d) $n = 100$.



EXAMPLE 8.8 Contracting Hepatitis from Blood Transfusion

The development of viral hepatitis after a blood transfusion can cause serious complications for a patient. The article “[An Assessment of Hepatitis E Virus in U.S. Blood Donors and Recipients](#)” (*Transfusion* [2013]: 2505–2511) reported that in a sample of 342 blood donors age 18 to 45, about 7% tested positive for the Hepatitis E virus. Suppose that the actual proportion that test positive for Hepatitis E in the population of all blood donors in this age group is 0.07. You can simulate sampling from this population of blood donors. For this example, a blood donor who tests positive for Hepatitis E will be considered a success, so $p = 0.07$. Figure 8.8 displays relative frequency histograms of 500 values of \hat{p} for the four sample sizes $n = 10, 25, 50$, and 100 .

As was the case in Example 8.7, all four histograms are centered at approximately the value of p for the population being sampled. (The average of the \hat{p} values for these simulations are 0.0690, 0.0677, 0.0707, and 0.0694.) If the histograms had been based on all possible samples, they would all have been centered at exactly $p = 0.07$.

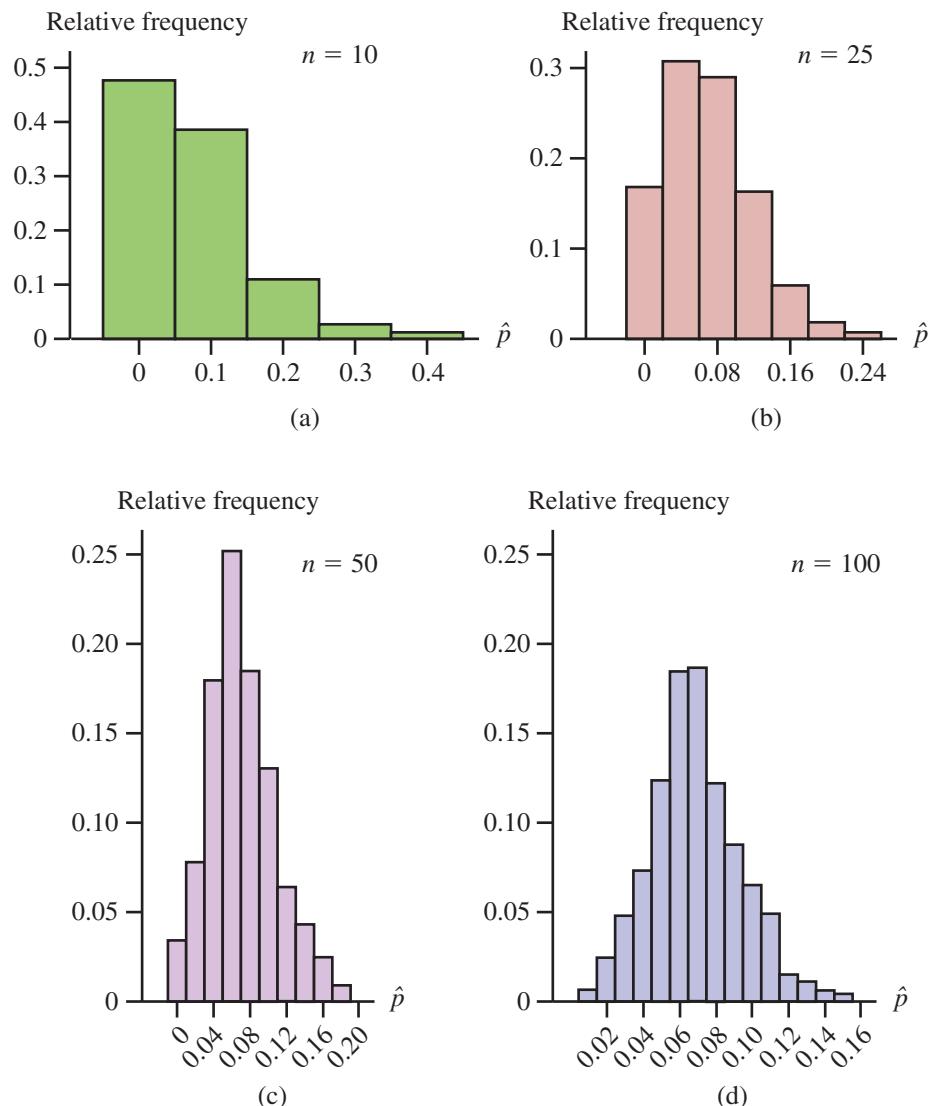
Notice that the scales on the axes are not the same for the four histograms in Figure 8.8. This was done so that it would be easier to see the behavior of the sample proportion for each sample size. Taking the differences in scales into account, you can see that the sample-to-sample variability decreases as the sample size increases. For example, the distribution is much less spread out in the histogram for $n = 100$ than for $n = 25$. The larger

FIGURE 8.8

Histograms of 500 values of \hat{p}

($p = 0.07$) for Example 8.8:

- (a) $n = 10$; (b) $n = 25$; (c) $n = 50$;
- (d) $n = 100$.



the value of n , the closer the sample proportion \hat{p} tends to be to the value of the population proportion p .

Another thing to notice about the histograms in Figure 8.8 is the progression toward the shape of a normal curve as n increases. The histograms for $n = 10$ and $n = 25$ are quite skewed, and the skew of the histogram for $n = 50$ is still moderate (compare Figure 8.8(c) with Figure 8.7(c)). Only the histogram for $n = 100$ looks approximately normal in shape. It appears that whether a normal curve provides a good approximation to the sampling distribution of \hat{p} depends on the values of both n and p . Knowing only that $n = 50$ is not enough to guarantee that the shape of the sampling distribution of \hat{p} will be approximately normal.

General Properties of the Sampling Distribution of \hat{p}

Examples 8.7 and 8.8 suggest that the sampling distribution of \hat{p} depends on both n , the sample size, and p , the proportion of successes in the population. Key results are stated more formally in the following general properties.

General Properties of the Sampling Distribution of \hat{p}

Notation

\hat{p} Sample proportion

n Sample size

p Population proportion

$\mu_{\hat{p}}$ Mean of the sampling distribution of \hat{p}

$\sigma_{\hat{p}}$ Standard deviation of the sampling distribution of \hat{p}

When random samples are selected from a population, the following are properties of the sampling distribution of \hat{p} :

Property 1. $\mu_{\hat{p}} = p$.

Property 2. $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

This formula is exact if the population is infinite, and is approximately correct if the population is finite and no more than 10% of the population is included in the sample.

Property 3. When n is large and p is not too near 0 or 1, the sampling distribution of \hat{p} is approximately normal.

The sampling distribution of \hat{p} is always centered at the value of the population success proportion p , and the extent to which the distribution spreads out about p decreases as the sample size n increases.

Examples 8.7 and 8.8 indicate that both p and n must be considered in judging whether the sampling distribution of \hat{p} is approximately normal.

The farther the value of p is from 0.5, the larger n must be in order for the sampling distribution of \hat{p} to be approximately normal.

A conservative rule of thumb is that if both

$$np \geq 10 \text{ and } n(1-p) \geq 10,$$

then a normal distribution provides a reasonable approximation to the sampling distribution of \hat{p} .

A sample size of $n = 100$ is not by itself sufficient to justify the use of a normal approximation. If $p = 0.01$, the distribution of \hat{p} is positively skewed even when $n = 100$, so a bell-shaped curve does not give a good approximation. Similarly, if $n = 100$ and $p = 0.99$ (so that $n(1 - p) = 1 < 10$), the distribution of \hat{p} has a substantial negative skew. The conditions $np \geq 10$ and $n(1 - p) \geq 10$ ensure that the sampling distribution of \hat{p} is not too skewed. If $p = 0.5$, the normal approximation can be used for n as small as 20, whereas for $p = 0.05$ or 0.95 , n should be at least 200.

EXAMPLE 8.9 Blood Transfusions Continued

Understand the context ➤

In the article referenced in Example 8.8, the proportion of blood donors testing positive for Hepatitis E was given as 0.07. Suppose that a new screening procedure is implemented and it is hoped that this will reduce the number of donors who test positive. Blood screened using this procedure is given to $n = 200$ blood recipients. Suppose that only 6 of the 200 patients contract hepatitis.

This appears to be a favorable result, because $\hat{p} = 6/200 = 0.03$. The question of interest to medical researchers is, Does this result indicate that the actual proportion of donors who test positive for Hepatitis E when the new screening procedure is used is less than 0.07, or could this result be plausibly attributed to sampling variability (that is, to the fact that \hat{p} typically differs from the population proportion, p)?

If the screening procedure is not effective then $p = 0.07$, and

$$\mu_{\hat{p}} = 0.07$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{200}} = \sqrt{\frac{(0.07)(0.93)}{200}} = 0.018$$

Because

$$np = 200(0.07) = 14 \geq 10$$

and

$$n(1 - p) = 200(0.93) = 186 \geq 10$$

the sampling distribution of \hat{p} is approximately normal.

Do the work ➤

Then, if the screening procedure is not effective,

$$\begin{aligned} P(\hat{p} \leq 0.03) &= P\left(z \leq \frac{0.03 - 0.07}{0.018}\right) \\ &= P(z \leq -2.22) \\ &= 0.0132 \end{aligned}$$

Interpret the results ➤

This small probability tells us that it is unlikely that a sample proportion 0.03 or smaller would be observed if the screening procedure was not effective. The new screening procedure appears to result in a smaller proportion of blood donors testing positive for Hepatitis E.

EXERCISES 8.23 - 8.31

- 8.23** A random sample is to be selected from a population that has a proportion of successes $p = 0.65$. Determine the mean and standard deviation of the sampling distribution of \hat{p} for each of the following sample sizes:
- a. $n = 10$
 - b. $n = 20$
 - c. $n = 30$
 - d. $n = 50$
 - e. $n = 100$
 - f. $n = 200$
- 8.24** a. For which of the sample sizes given in the previous exercise would the sampling distribution of \hat{p} be approximately normal if $p = 0.65$?
b. For which of the sample sizes given in the previous exercise would the sampling distribution be approximately normal if $p = 0.2$?
- 8.25** The article “Younger Adults More Likely than Their Elders to Prefer Reading the News” (October 6,

[2016, *pewresearch.org/fact-tank/2016/10/06/younger-adults-more-likely-than-their-elders-to-prefer-reading-news/*, retrieved March 28, 2018](https://www.pewresearch.org/fact-tank/2016/10/06/younger-adults-more-likely-than-their-elders-to-prefer-reading-news/)) estimated that only 3% of those age 65 and older who prefer to watch the news, rather than to read or listen to news, watch the news online. This estimate was based on a large sample of adult Americans conducted by the Pew Research Center. Consider the population consisting of all adult Americans age 65 and older who prefer to watch the news and suppose that for this population the actual proportion who prefer to watch online is 0.03. (Hint: See Example 8.9.)

- a. A random sample of $n = 100$ people will be selected from this population and \hat{p} , the proportion of people who prefer to watch online will be calculated. What are the mean and standard deviation of the sampling distribution of \hat{p} ?
- b. Is it reasonable to assume that the sampling distribution of \hat{p} is approximately normal for random samples of size $n = 100$? Explain.
- c. Suppose that the sample size is $n = 400$ rather than $n = 100$, as in Part (b). Does the change in sample size change the mean and standard deviation of the sampling distribution of \hat{p} ? If so, what are the new values for the mean and standard deviation? If not, explain why not.
- d. Is it reasonable to assume that the sampling distribution of \hat{p} is approximately normal for random samples of size $n = 400$? Explain.

8.26 The article referenced in the previous exercise reported that for people age 18 to 29 who prefer to watch the news, the proportion who prefer to watch online is 0.37. Answer the questions posed in Parts (a)–(d) of the previous exercise for the population of people age 18 to 29 who prefer to watch the news.

8.27 A certain chromosome defect occurs in only 1 in 200 adult Caucasian males. A random sample of $n = 100$ adult Caucasian males is to be obtained.

- a. What is the mean value of the sample proportion \hat{p} , and what is the standard deviation of the sample proportion?
- b. Does \hat{p} have approximately a normal distribution in this case? Explain.
- c. What is the smallest value of n for which the sampling distribution of \hat{p} is approximately normal?

8.28 The U.S. Census Bureau reported that in 2015, the proportion of adult Americans age 25 and older who have a bachelor's degree or higher is 0.325 ([“Educational Attainment in the United States: 2015,” *census.gov, retrieved January 22, 2017*](https://www.census.gov/popest/censuses/2015/index.html)).

Consider the population of all adult Americans age 25 and over in 2015 and define \hat{p} to the proportion of people in a random sample from this population who have a bachelor's degree or higher.

- a. Would \hat{p} based on a random sample of only 10 people from this population have approximately a normal distribution? Explain why or why not.
- b. What are the mean value and standard deviation of \hat{p} based on a random sample of size 400?
- c. Suppose that the sample size is $n = 200$ rather than $n = 400$. Does the change in sample size affect the mean and standard deviation of the sampling distribution of \hat{p} ? If so, what are the new values for the mean and standard deviation? If not, explain why not.

8.29 The article “Fewer Americans Are Reading, But Don’t Blame the Millennials” (*Los Angeles Times*, October 9, 2016) indicates that 80% of millennials (those age 18 to 29) have read a book in the last year. Suppose that this is the actual percentage for the population of all millennials. Consider a sample proportion \hat{p} that is based on a random sample of 225 millennials.

- a. If $p = 0.80$, what are the mean and standard deviation of the sampling distribution of \hat{p} ?
- b. If $p = 0.70$, what are the mean and standard deviation of the sampling distribution of \hat{p} ?
- c. Is the sampling distribution of \hat{p} approximately normal in both cases ($p = 0.80$ and $p = 0.70$)?

8.30 Suppose that a candidate for public office is favored by 48% of all registered voters in his district. A polling organization will take a random sample of 500 voters and will use \hat{p} , the sample proportion, to estimate p . What is the approximate probability that \hat{p} will be greater than 0.5, causing the polling organization to incorrectly predict the result of the upcoming election? (Hint: See Example 8.9.)

8.31 A manufacturer of computer printers purchases plastic ink cartridges from a vendor. When a large shipment is received, a random sample of 200 cartridges is selected, and each cartridge is inspected. If the sample proportion of defective cartridges is more than 0.02, the entire shipment is returned to the vendor.

- a. What is the approximate probability that a shipment will be returned if the true proportion of defective cartridges in the shipment is 0.05?
- b. What is the approximate probability that a shipment will not be returned if the true proportion of defective cartridges in the shipment is 0.10?

CHAPTER ACTIVITIES

ACTIVITY 8.1 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

Technology Activity

This activity uses the Shiny app titled Sampling Distribution Mean that is part of the collection of Shiny web apps that accompany this text. This app can be found at statistics.cengage.com/PSO6e/Apps.html.

In this activity, we will explore properties of the sampling distribution of the sample mean. Open the app. You should see a screen like the one shown here.

Sampling Distribution of the Sample Mean

Population Shape: Normal

Enter Population Mean: 1

Enter Population Standard Deviation: 1

Select Sample Size: 1

Number of Samples: 1 10 100 1,000 10,000

Choose type of plot for sample: Dotplot Histogram
 Adjust dot size or number of breaks

Buttons: Sample, Reset

This app can be used to carry out sampling experiments like the ones described in Examples 8.2 and 8.3. In those examples, random samples were selected from a specified population to learn about the behavior of the sample mean \bar{x} . This app allows you to specify a population distribution shape, as well as some characteristics of the population distribution, such as the mean. You can also specify a sample size and the number of samples you would like to generate.

Part 1: Normal Population

Step 1. Start by looking at the distribution of sample means when random samples are selected from a population that has a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 10$. Make sure that “Normal” is selected as the population shape and enter 100

for the population mean and 10 for the population standard deviation. Begin by entering 5 for the sample size and select 1000 for the number of samples to generate. Choose dotplot for the type of display for the sample.

Sampling Distribution of the Sample Mean

Population Shape

Normal

Enter Population Mean:

100

Enter Population Standard Deviation:

10

Select Sample Size:

5

Number of Samples:

1 10 100 1,000 10,000

Choose type of plot for sample

Dotplot Histogram

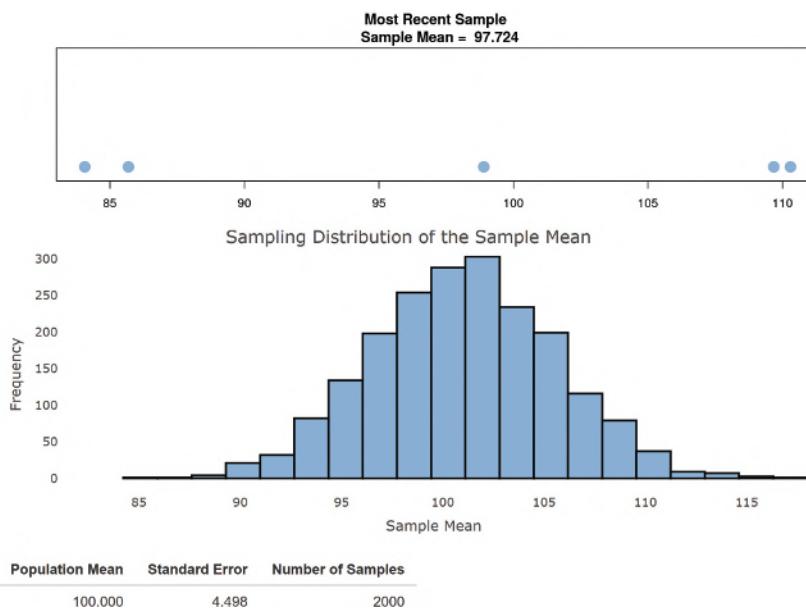
Adjust dot size or number of breaks

Buttons: Sample, Reset

When you click on the sample button, the app will generate 1000 random samples of size 5 from this normal population, calculate the sample mean for each sample, and display a histogram of the sample means. You should see a screen similar to the one that follows, although yours will be different since you will not have the same set of 1000 random samples.

Write a brief description of the histogram of the sample means, being sure to discuss center, variability, and shape.

Step 2. Next, click the reset button to clear the display. Generate a new histogram of 1000 sample means for random samples of size 10. Write a brief description of the histogram of the sample means, being sure to discuss center, variability, and shape. How does this distribution compare to the distribution of sample means from Step 1?



Step 3. Repeat Step 2 for sample sizes of 20 and 50.

Step 4. Consider the four distributions of sample means that you created (corresponding to sample sizes of 5, 10, 20, and 50). Explain how these histograms are consistent with the general properties of the sampling distribution of the sample mean given in Section 8.2.

Part 2: A Skewed Population

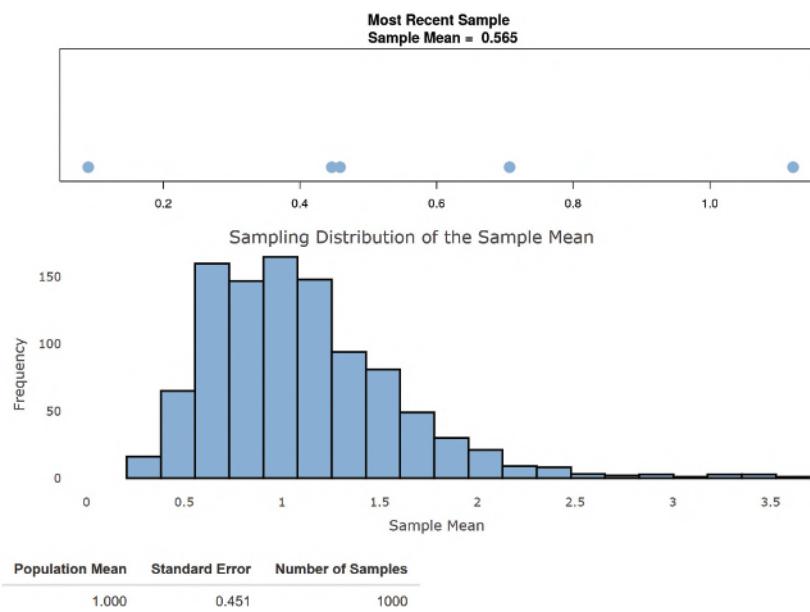
Step 1. In Part 2, you will consider the distribution of sample means when random samples are selected from a population that is skewed. Begin by selecting “Exponential” as the population shape and enter 1 for the population mean. Enter 5 for the sample size and select 1000 for the number of samples to generate. Choose dotplot for the type of display for the sample. When you click on the sample button, the app will generate 1000 random samples of size 5 from this population, calculate

the sample mean for each sample, and display a histogram of the sample means. You should see a screen similar to the one shown here, although yours will be different since you will not have the same set of 1000 random samples. Notice that the population distribution shown in the top graph is skewed.

Write a brief description of the histogram of the sample means, being sure to discuss center, variability, and shape.

Step 2. Next click the reset button to clear the display. Generate a new histogram of 1000 sample means for random samples of size 10. Write a brief description of the histogram of the sample means, being sure to discuss center, variability, and shape. How does this distribution compare to the distribution of sample means from Step 1?

Step 3. Repeat Step 2 for sample sizes of 20 and 50.



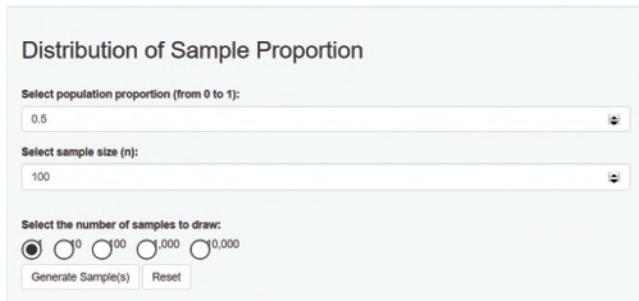
Step 4. Consider the four distributions of sample means that you created (corresponding to sample sizes of 5, 10, 20, and 50). Explain how these histograms are consistent

with the general properties of the sampling distribution of the sample mean given in Section 8.2.

ACTIVITY 8.2 SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION TECHNOLOGY ACTIVITY

This activity uses the Shiny app titled Sampling Distribution Proportion that is part of the collection of Shiny web apps that accompany this text. This app can be found at statistics.cengage.com/PSO6e/Apps.html.

In this activity, we will explore properties of the sampling distribution of the sample proportion. Open the app. You should see a screen like the one shown here.



This app can be used to carry out sampling experiments like the ones described in Examples 8.7 and 8.8. In those examples, random samples were selected from a specified population to learn about the behavior of the sample proportion \hat{p} . This app allows you to specify a population proportion. You can also specify a sample size and the number of samples you would like to generate.

Part 1: Exploring the effect of sample size

Step 1. Start by looking at the distribution of sample proportions when random samples are selected from a

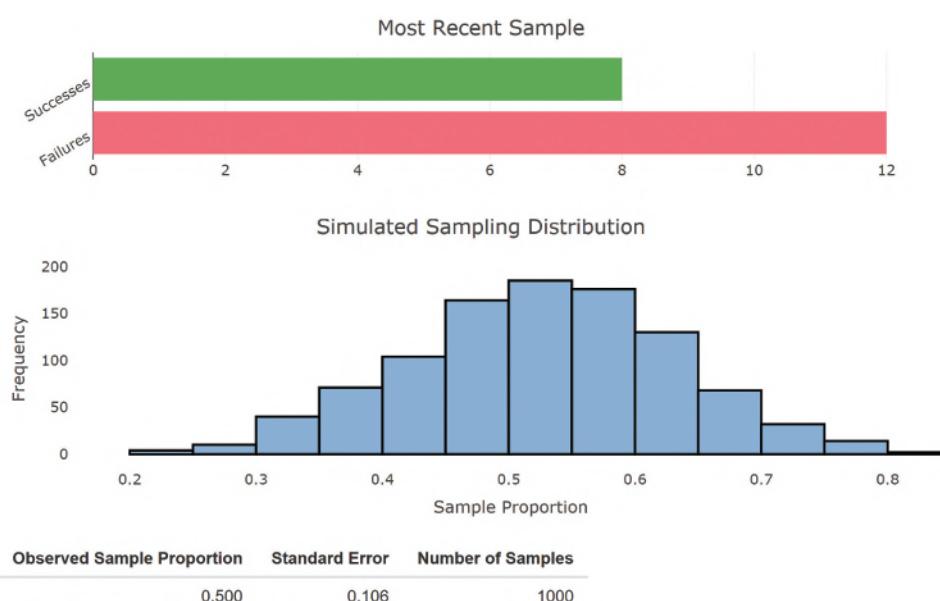
population that has a proportion of successes of 0.5. Enter 0.5 for the population proportion. Begin by entering 20 for the sample size and select 1000 for the number of samples to generate. When you click on the Generate Samples button, the app will generate 1000 random samples of size 20 from this population, calculate the sample proportion for each sample, and display a histogram of the sample proportions. You should see a screen similar to the one shown here, although yours will be different since you will not have the same set of 1000 random samples.

Write a brief description of the histogram of the sample proportions, being sure to discuss center, variability, and shape.

Step 2. Next, click the reset button to clear the display. Generate a new histogram of 1000 sample proportions for random samples of size 30. Write a brief description of the histogram of the sample proportions, being sure to discuss center, variability, and shape. How does this distribution compare to the distribution of sample proportions from Step 1?

Step 3. Repeat Step 2 for sample sizes of 50 and 100.

Step 4. Consider the four distributions of sample proportions that you created (corresponding to sample sizes of 20, 30, 50, and 100). Explain how these histograms are consistent with the general properties of the sampling distribution of the sample proportion given in Section 8.3.



Part 2: Exploring the effect of value of the population proportion

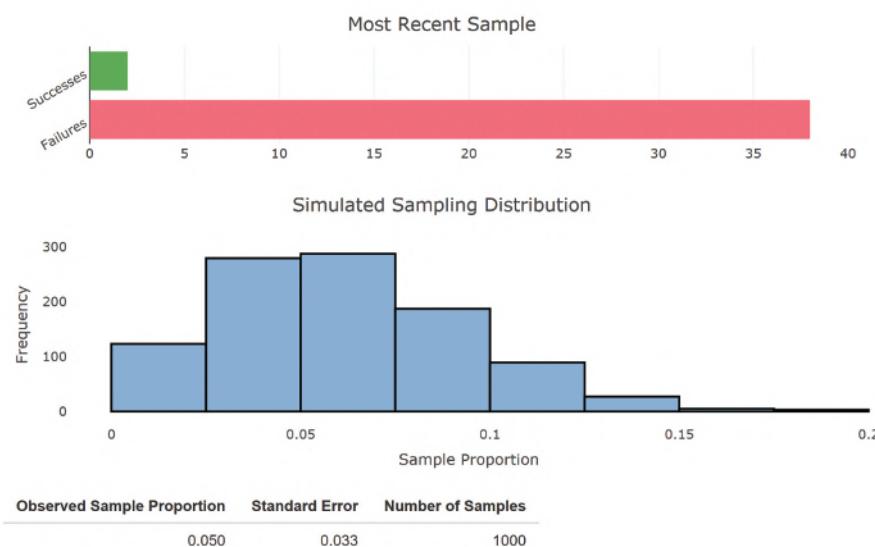
Step 1. Start by looking at the distribution of sample proportions when random samples of size 40 are selected from a population that has a proportion of success of 0.05. Enter 0.05 for the population proportion and 40 for the sample size. Select 1000 for the number of samples to generate. When you click on the Generate Samples button, the app will generate 1000 random samples of size 40 from this population, calculate the sample proportion for each sample, and display a histogram of the sample proportions. You should see a screen similar to the one shown here, although yours will be different since you will not have the same set of 1000 random samples.

Write a brief description of the histogram of the sample proportions, being sure to discuss center, variability, and shape.

Step 2. Next, click the reset button to clear the display. Generate a new histogram of 1000 sample proportions for random samples of size 40 from a population with a proportion of successes of 0.3. Write a brief description of the histogram of the sample proportions, being sure to discuss center, variability, and shape. How does this distribution compare to the distribution of sample proportions from Step 1?

Step 3. Repeat Step 2 for population proportions of 0.5 and 0.8.

Step 4. Consider the four distributions of sample proportions that you created (corresponding to population proportions of 0.05, 0.3, 0.5, and 0.8). Explain how these histograms are consistent with the general properties of the sampling distribution of the sample proportion given in Section 8.3.



ACTIVITY 8.3 DO STUDENTS WHO TAKE THE SATs MULTIPLE TIMES HAVE AN ADVANTAGE IN COLLEGE ADMISSIONS?

Technology activity: Requires use of a computer or a graphing calculator.

Background: *The Chronicle of Higher Education* (January 29, 2003) summarized an article that appeared on the *American Prospect* web site titled “College Try: Why Universities Should Stop Encouraging Applicants to Take the SATs Over and Over Again.” This paper argued that current college admission policies that permit applicants to take the SAT exam multiple times and then use the highest score for consideration of admission favor students from families with higher incomes (who can afford to take the exam many times). The author proposed two alternatives that he believes would be fairer than using the highest score: (1) Use the average of all test scores, or (2) use only the most recent score.

In this activity, you will investigate the differences between the three possibilities by looking at the sampling

distributions of three statistics for a test taker who takes the exam twice and for a test taker who takes the exam five times. The three statistics are

Max = maximum score

Mean = average score

Recent = most recent score

An individual’s score on the SAT exam fluctuates between test administrations. Suppose that a particular student’s “true ability” is reflected by an SAT score of 1200 but, because of chance fluctuations, the test score on any particular administration of the exam can be considered a random variable that has a distribution that is approximately normal with mean 1200 and standard deviation 30. If we select a sample from this normal distribution, the resulting set of observations can be viewed as a collection of test scores that might have been obtained by this student.

Part 1: Begin by considering what happens if this student takes the exam twice. You will use simulation to generate samples of two test scores, Score1 and Score2, for this student. Then you will compute the values of Max, Mean, and Recent for each pair of scores. The resulting values of Max, Mean, and Recent will be used to construct approximations to the sampling distributions of the three statistics.

The instructions that follow assume the use of Minitab. If you are using a different software package or a graphing calculator, your instructor will provide alternative instructions.

- Obtain 500 sets of two test scores by generating observations from a normal distribution with mean 1200 and standard deviation 30.

Minitab: Calc → Random Data → Normal

Enter 500 in the Generate box (to get 500 sets of scores)

Enter C1-C2 in the Store in Columns box (to get two test scores in each set)

Enter 1200 in the Mean box (because we want scores from a normal distribution with mean 1200)

Enter 30 in the Standard Deviation box (because we want scores from a normal distribution with standard deviation 30)

Click on OK

- Looking at the Minitab worksheet, you should now see 500 rows of values in each of the first two columns. The two values in any particular row can be regarded as the test scores that might be observed when the student takes the test twice. For each pair of test scores, we now calculate the values of Max, Mean, and Recent.
 - Recent is just the last test score, so the values in C2 are the values of Recent. Name this column recent2 by typing the name into the gray box at the top of C2.
 - Compute the maximum test score (Max) for each pair of scores, and store the values in C3, as follows:

Minitab: Calc → Row statistics

Click the button for maximum

Enter C1-C2 in the Input variables box

Enter C3 in the Store Result In box.

Click on OK

You should now see the maximum value for each pair in C3. Name this column max2.

- Compute the average test score (Mean) for each pair of scores, and store the values in C4, as follows:

Minitab: Calc → Row statistics

Click the button for mean

Enter C1-C2 in the Input Variables box

Enter C4 in the Store Result In box.

Click on OK

You should now see the average for each pair in C4. Name this column mean2.

- Construct density histograms for each of the three statistics (these density histograms approximate the sampling distributions of the three statistics), as follows:

Minitab: Graph → Histogram

Enter max2, mean2, and recent2 into the first three rows of the Graph Variables box

Click on the Options button. Select Density. Click on OK.

(This will produce histograms that use the density scale rather than the frequency scale.)

Click on the Frame drop-down menu, and select Multiple Graphs. Select Same X and Same Y. (This will cause Minitab to use the same scales for all three histograms, so that they can be easily compared.)

Click on OK.

Part 2: Now you will produce approximate sampling distributions for these same three statistics, but for the case of a student who takes the exam five times. Follow the same steps as in Part 1, with the following modifications:

- Obtain 500 sets of five test scores, and store these values in columns C11–C15.
- Recent will just be the values in C15; name this column recent5. Compute the Max and Mean values, and store them in columns C16 and C17. Name these columns max5 and mean5.
- Construct density histograms for max5, mean5, and recent5.

Part 3: Now use the approximate sampling distributions constructed in Parts 1 and 2 to answer the following questions.

- The statistic that is the average of the test scores is just a sample mean (for a sample of size 2 in Part 1 and for a sample of size 5 in Part 2). How do the sampling distributions of mean2 and mean5 compare to what is expected based on the general properties of the \bar{x} distribution given in Section 8.2? Explain.
- Based on the three distributions from Part 1, for a two-time test taker, describe the advantage of using the maximum score compared to using either the average score or the most recent score.
- Now consider the approximate sampling distributions of the maximum score for two-time and for five-time test takers. How do these two distributions compare?
- Does a student who takes the exam five times have a big advantage over a student of equal ability who takes the exam only twice if the maximum score is used for college admission decisions? Explain.
- If you were writing admission procedures for a selective university, would you recommend using the maximum test score, the average test score, or the most recent test score in making admission decisions? Write a paragraph explaining your choice.

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
Statistic	Any quantity whose value is calculated from sample data.	Central Limit Theorem	This important theorem states that when n is sufficiently large, the \bar{x} distribution will be approximately normal. The standard rule of thumb is that the theorem can safely be applied when n is greater than or equal to 30.
Sampling distribution	The probability distribution of a statistic. The sampling distribution describes the long-run behavior of the statistic.	Sampling distribution of \hat{p}	The probability distribution of the sample proportion \hat{p} , based on a random sample of size n . When the sample size is sufficiently large, the sampling distribution of \hat{p} is approximately normal, with
Sampling distribution of \bar{x}	The probability distribution of the sample mean \bar{x} based on a random sample of size n . Properties of the \bar{x} sampling distribution: $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ (where μ and σ are the population mean and standard deviation, respectively). In addition, when the population distribution is normal or the sample size is large, the sampling distribution of \bar{x} is (approximately) normal.	$\mu_{\hat{p}} = p \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ where p is the value of the population proportion.	

CHAPTER REVIEW Exercises 8.32 - 8.37

- 8.32** The nicotine content in a single cigarette of a particular brand has a distribution with mean 0.8 mg and standard deviation 0.1 mg. If 100 randomly selected cigarettes of this brand are analyzed, what is the probability that the resulting sample mean nicotine content will be less than 0.79? less than 0.77?
- 8.33** Let x_1, x_2, \dots, x_{100} denote the actual weights (in pounds) of 100 randomly selected bags of fertilizer. Suppose that the weight of a randomly selected bag has a distribution with mean 50 pounds and variance 1 pound². Let \bar{x} be the sample mean weight ($n = 100$).
- Describe the sampling distribution of \bar{x} .
 - What is the probability that the sample mean is between 49.75 pounds and 50.25 pounds?
 - What is the probability that the sample mean is less than 50 pounds?
- 8.34** Suppose that 20% of the subscribers of a cable television company watch the shopping channel at least once a week. The cable company is trying to decide whether to replace this channel with a new local station. A survey of 100 subscribers will be undertaken. The cable company has decided to keep the shopping channel if the sample proportion is greater than 0.25.
- What is the approximate probability that the cable company will keep the shopping channel, even though the proportion of all subscribers who watch it is only 0.20?
- 8.35** Water permeability of concrete can be measured by letting water flow across the surface and determining the amount lost (in inches per hour). Suppose that the permeability index x for a randomly selected concrete specimen is normally distributed with mean value 1000 and standard deviation 150.
- How likely is it that a single randomly selected specimen will have a permeability index between 850 and 1300?
 - If the permeability index is to be determined for each specimen in a random sample of size 10, how likely is it that the sample mean permeability index will be
 - between 950 and 1100?
 - between 850 and 1300?
- 8.36** Suppose that 40% of all U.S. employees contribute to a retirement plan ($p = 0.40$).
- In a random sample of 100 employees, what is the approximate probability that at least half of those in the sample contribute to a retirement plan?
 - Suppose you were told that at least 60 of the 100 employees in a sample from your state contribute to a retirement plan. Would you think $p = 0.40$ for your state? Explain.
- 8.37** The amount of money spent by a customer at a discount store has a mean of \$100 and a standard deviation of \$30. What is the probability that a randomly selected group of 50 shoppers will spend a total of more than \$5300? (Hint: The total will be more than \$5300 when the sample mean exceeds what value?)

9

Estimation Using a Single Sample



Anatoliy Cherkas/Shutterstock.com

College students use cell phones in multiple ways. The paper “[Sleeping with Technology: Cognitive, Affective and Technology Usage Predictors of Sleep Problems Among College Students](#)” (*Sleep Health* [2016]: 49–56) describes the results of a survey of a representative sample of 734 college students. Of the 734 students surveyed, 125 reported that they sleep with their cell phones near their bed and check their phones for something other than the time at least twice during the night. The authors of this paper were interested in using data from the sample to learn about the proportion of all college students who check their cell phones for something other than the time at least twice during the night.

The methods introduced in this chapter can be used to estimate this proportion. Because the estimate will be based only on a sample rather than on a census of all college students, it is important that this estimate be constructed in a way that also conveys information about the anticipated accuracy.

LEARNING OBJECTIVES

Students will understand:

- What it means for a statistic to be an unbiased estimator of a population characteristic.
- The relationships between sample size, margin of error, and the width of a confidence interval.
- The factors that affect the width of a confidence interval.
- The meaning of the confidence level associated with a confidence interval.

Students will be able to:

- Construct and interpret a confidence interval for a population proportion.
- Construct and interpret a confidence interval for a population mean.
- Determine the sample size necessary to achieve a desired margin of error when estimating a population proportion or a population mean.
- Calculate and interpret a bootstrap confidence interval for a population proportion. (Optional)
- Calculate and interpret a bootstrap confidence interval for a population mean. (Optional)

SECTION 9.1 Point Estimation

A simple way to estimate a population characteristic involves using sample data to calculate a single number that represents a plausible value of the population characteristic. For example, sample data might suggest that 1000 hours is a plausible value for μ , the mean lifetime for all lightbulbs of a particular brand. A survey of a sample of students at a particular university might lead to the statement that 0.41 is a plausible value for p , the proportion of all students at the university who favor a fee for recreational facilities.

DEFINITION

Point estimate: A single number calculated using sample data that represents a plausible value of a population characteristic.

A point estimate is obtained by first selecting an appropriate statistic. The value of the statistic for a given sample is then used as the point estimate. For example, the calculated value of the sample mean is one point estimate of a population mean μ , and the sample proportion is a point estimate of a population proportion, p .

Example 9.1 Internet Use by College Students

Understand the context ➤

One of the purposes of the survey described in the chapter introduction was to estimate the proportion of college students who check their phone for something other than the time more than twice during the night. Based on information given in the paper, 125 of the 734 students surveyed reported checking their phone more than twice during the night. We can use this information to estimate p , where p is the proportion of all college students who check their phone more than twice during the night.

With a success identified as a student who checks his or her phone more than twice during the night, p is the population proportion of successes. The statistic

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

Do the work ➤ which is the sample proportion of successes, is an obvious choice for obtaining a point estimate of p . Using the reported information, the point estimate of p is

$$\hat{p} = \frac{125}{734} = 0.170$$

Interpret the results ➤

Based on this random sample, we estimate that 17.0% of college students check their phone more than twice during the night.

For purposes of estimating a population proportion p , there is no obvious alternative to the statistic \hat{p} . In other situations, such as the one illustrated in Example 9.2, there may be several statistics that might be used to obtain an estimate.

Example 9.2 Academic Reading

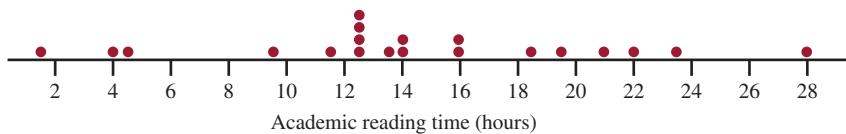
Understand the context ➤

The paper “[The Impact of Internet and Television Use on the Reading Habits and Practices of College Students](#)” (*Journal of Adolescent and Adult Literacy* [2009]: 609–619) investigates the reading habits of college students. The authors distinguished between recreational reading and academic reading and asked students to keep track of time spent reading. The following observations represent the number of hours spent on academic reading in 1 week by 20 college students (these data are compatible with summary values given in the paper and have been arranged in order from smallest to largest):

Consider the data ➤

1.7	3.8	4.7	9.6	11.7	12.3	12.3	12.4	12.6	13.4
14.1	14.2	15.8	15.9	18.7	19.4	21.2	21.9	23.3	28.2

A dotplot of the data is shown here:



From the dotplot, we can see that the distribution of academic reading time is approximately symmetric.

Suppose a point estimate of μ , the mean academic reading time per week for all college students, is desired. An obvious choice of a statistic for estimating μ is the sample mean, \bar{x} . However, there are other possibilities. If the population distribution is symmetric, the population mean and the population median are equal. Since the dotplot of the sample is approximately symmetric, we might consider using the sample median as an estimate of the population mean.

The estimates of μ based on the mean and the median for this sample are

Do the work ► sample mean = $\bar{x} = \frac{\sum x}{n} = \frac{287.2}{20} = 14.36$

$$\text{sample median} = \frac{13.4 + 14.1}{2} = 13.75$$

These two estimates of the mean academic reading time per week for college students are different. The choice between them should depend on which statistic tends, on average, to produce an estimate closest to the actual value of μ . The following subsection provides some criteria for choosing among competing statistics.

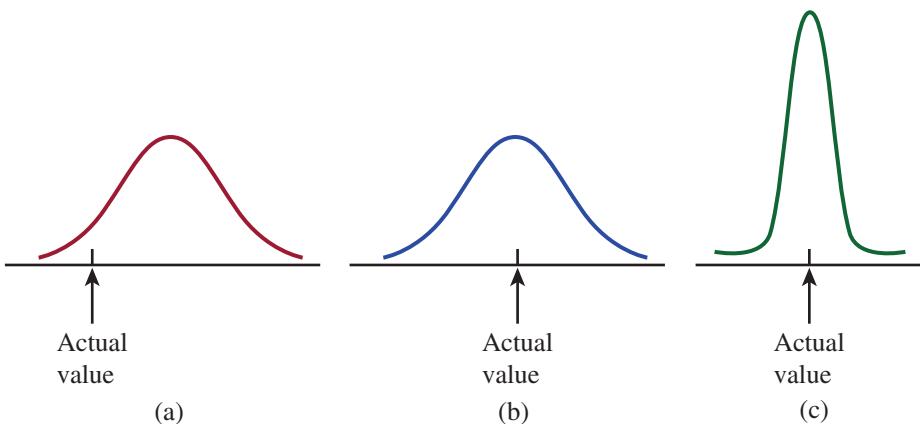
Choosing a Statistic for Calculating an Estimate

As illustrated in Example 9.2, there may be more than one statistic that is a reasonable choice as a point estimate of a population characteristic. We would like to use a statistic that tends to produce an accurate estimate—that is, an estimate close to the actual value of the population characteristic. Information about the accuracy of estimation for a particular statistic is provided by the statistic's sampling distribution.

Figure 9.1 displays the sampling distributions of three different statistics. The value of the population characteristic, which is denoted by *actual value* in the figure, is marked on the measurement axis. The distribution in Figure 9.1(a) is that of a statistic unlikely to result in an estimate close to the actual value. The distribution is centered to the right of the actual value, making it very likely that an estimate (a value of the statistic for a particular sample) will be larger than the actual value. If this statistic is used to calculate an estimate based

FIGURE 9.1

Sampling distributions of three different statistics for estimating a population characteristic.



on a first sample, then another estimate based on a second sample, and another estimate based on a third sample, and so on, the long-run average value of these estimates will be greater than the actual value.

The sampling distribution of Figure 9.1(b) is centered at the actual value. This means that although one estimate may be smaller than the actual value and another may be larger, when this statistic is used many times over with different random samples, there will be no long-run tendency to overestimate or underestimate the actual value. Notice that even though the sampling distribution is correctly centered, it spreads out quite a bit around the actual value. Because of this, some estimates resulting from the use of this statistic will be far above or far below the actual value, even though there is no systematic tendency to underestimate or overestimate the actual value.

In contrast, the mean value of the statistic with the distribution shown in Figure 9.1(c) is equal to the actual value of the population characteristic (implying no systematic tendency to overestimate or underestimate the actual value). The statistic's standard deviation is relatively small. This means that estimates based on this third statistic will almost always be quite close to the actual value—certainly more often than estimates resulting from the statistic with the sampling distribution shown in Figure 9.1(b).

DEFINITIONS

Unbiased statistic: A statistic whose mean value is equal to the value of the population characteristic being estimated.

Biased statistic: A statistic that is not unbiased.

As an example of a statistic that is biased, consider using the sample range as an estimate of the population range. Because the range of a population is defined as the difference between the largest value in the population and the smallest value, the range for a sample tends to underestimate the population range. This is because the largest value in a sample must be less than or equal to the largest value in the population and the smallest sample value must be greater than or equal to the smallest value in the population. The sample range equals the population range *only* if the sample happens to include both the largest and the smallest values in the population. In all other cases, the value of the sample range is smaller than the population range. This means that $\mu_{\text{sample range}}$ is less than the value of the population range, and the sample range is a biased statistic for estimating the population range.

One of the general results concerning the sampling distribution of \bar{x} , the sample mean, is that $\mu_{\bar{x}} = \mu$. This result says that the \bar{x} values from all possible random samples of size n center around μ , the population mean. For example, if $\mu = 100$, the \bar{x} distribution is centered at 100, whereas if $\mu = 5200$, then the \bar{x} distribution is centered at 5200. This means that \bar{x} is an unbiased statistic for estimating μ . Similarly, because the sampling distribution of \hat{p} is centered at p , it follows that \hat{p} is an unbiased statistic for estimating a population proportion.

Using an unbiased statistic that also has a small standard deviation ensures that there will be no systematic tendency to underestimate or overestimate the value of the population characteristic *and* that estimates will almost always be relatively close to the actual value of the population characteristic.

Given several unbiased statistics that might be used for estimating a population characteristic, the best choice is the statistic with the smallest standard deviation.

Consider the problem of estimating a population mean, μ . The obvious choice of statistic for obtaining a point estimate of μ is the sample mean, \bar{x} , an unbiased statistic for estimating the population mean. However, when the population distribution is symmetric, \bar{x} is not the only unbiased statistic for estimating the population mean. Other unbiased statistics for estimating μ in this case include the sample median. Which statistic should be

used? If the population distribution is normal, then \bar{x} has a smaller standard deviation than any other unbiased statistic for estimating μ . This means that if the population distribution is normal, the sample mean would be the best choice.

Now consider estimating another population characteristic, the population variance, σ^2 . The sample variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

is a good choice for obtaining a point estimate of the population variance, σ^2 . It can be shown that s^2 is an unbiased statistic for estimating σ^2 . This means that whatever the value of σ^2 , the sampling distribution of s^2 is centered at that value. This is the reason that the divisor $(n - 1)$ is used. An alternative statistic is the average squared deviation

$$\frac{\sum(x - \bar{x})^2}{n}$$

which has a more natural denominator than s^2 . However, the average squared deviation is biased, with its values tending to be smaller, on average, than the actual value of σ^2 .

Example 9.3 Airborne Times for Flights from San Francisco to Washington, D.C.

Understand the context ➤

The [Bureau of Transportation Statistics](#) provides data on U.S. airline flights. Suppose that the airborne times (in minutes) for nonstop flights from San Francisco to Washington Dulles airport for 10 randomly selected flights in June 2018 are:

270 256 267 285 274 275 266 258 271 281

For these data $\sum x = 2703$, $\sum x^2 = 731,373$, $n = 10$, and

$$\begin{aligned} \sum(x - \bar{x})^2 &= \sum x^2 - \frac{(\sum x)^2}{n} \\ &= 731,373 - \frac{(2703)^2}{10} \\ &= 752.1 \end{aligned}$$

Do the work ➤

We use σ^2 to denote the actual variance in airborne time for June 2018 nonstop flights from San Francisco to Washington Dulles airport. Using the sample variance s^2 to provide a point estimate of σ^2 yields

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{752.1}{9} = 83.57$$

Using the average squared deviation (with divisor $n = 10$), the resulting point estimate is

$$\frac{\sum(x - \bar{x})^2}{n} = \frac{752.1}{10} = 75.21$$

Interpret the results ➤

Because s^2 is an unbiased statistic for estimating σ^2 , many statisticians would recommend using the point estimate 83.57.

An obvious choice of a statistic for estimating the population standard deviation σ is the sample standard deviation s . For the data given in Example 9.3,

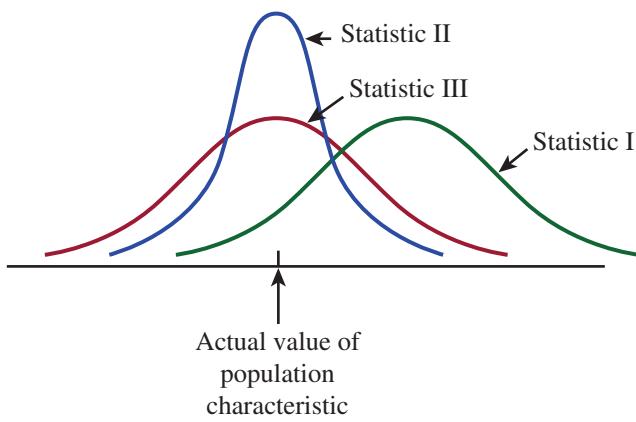
$$s = \sqrt{83.57} = 9.14$$

Unfortunately, the fact that s^2 is an unbiased statistic for estimating σ^2 does not imply that s is an unbiased statistic for estimating σ . The sample standard deviation tends to underestimate slightly the actual value of σ . However, unbiasedness is not the only criterion by which a statistic can be judged, and there are other good reasons for using s to estimate σ . Whenever we need to estimate σ based on a single random sample, we will use the statistic s to obtain a point estimate.

EXERCISES 9.1 - 9.9

• Data set available online

- 9.1** Three different statistics are being considered for estimating a population characteristic. The sampling distributions of the three statistics are shown in the following illustration:



Which statistic would you recommend? Explain your choice. (Hint: See the section on choosing a statistic.)

- 9.2** a. Why is an unbiased statistic generally preferred over a biased statistic for estimating a population characteristic?
 b. Does unbiasedness alone guarantee that the estimate will be close to the true value? Explain.
 c. Under what circumstances might you choose a biased statistic over an unbiased statistic if two statistics are available for estimating a population characteristic?
- 9.3** The report “[The 2016 Consumer Financial Literacy Survey](#)” ([The National Foundation for Credit Counseling](#) ([nfcc.org](#), retrieved October 28, 2016)) summarized data from a representative sample of 1668 adult Americans. When asked if they typically carry a credit card debt from month to month, 584 of these people responded “yes.” Estimate p , the proportion of adult Americans who carry credit card debt from month to month.
- 9.4** • The authors of the paper “[Influence of Biofeedback Weight Bearing Training in Sit to Stand to Sit and the Limits of Stability on Stroke Patients](#)” ([The Journal of Physical Therapy Science](#)

[2016]: 3011–2014) randomly selected two samples of patients admitted to the hospital after suffering a stroke. One sample was selected from patients who received biofeedback weight training for 8 weeks and the other sample was selected from patients who did not receive this training. At the end of 8 weeks, the time it took (in seconds) to stand from a sitting position and then to sit down again (called sit-stand-sit time) was measured for the people in each sample. Data consistent with summary quantities given in the paper are given below. For purposes of this exercise, you can assume that the samples are representative of the population of stroke patients who receive the biofeedback training and the population of stroke patients who do not receive this training.

Biofeedback Group

1.9	2.6	4.3	2.1	2.7	4.1	3.2	4.0
3.2	3.5	2.8	3.5	3.5	2.3	3.1	

No Biofeedback Group

5.1	4.7	3.9	4.2	4.7	4.3	4.2	5.1
3.4	4.2	5.1	4.4	4.0	3.4	3.9	

- a. Use the given data to estimate the mean sit-stand-sit time for stroke patients who receive biofeedback training.
 b. Use the given data to estimate the mean sit-stand-sit time for stroke patients who do not receive biofeedback training.
 c. Use the given information to estimate the standard deviation of sit-stand-sit times for stroke patients who receive biofeedback.

- 9.5** Each person in a random sample of 20 students at a particular university was asked whether he or she is registered to vote. The responses (R = registered, N = not registered) are given here:

R R N R N N R R R R N R R R R N R R N

Use these data to estimate p , the proportion of all students at the university who are registered to vote.

- 9.6** Suppose that each of 935 smokers received a nicotine patch, which delivers nicotine to the bloodstream but at a much slower rate than cigarettes.

Dosage was decreased to 0 over a 12-week period. Suppose that 245 of the subjects were still not smoking 6 months after treatment. Assuming it is reasonable to regard this sample as representative of all smokers, estimate the proportion of all smokers who, when given this treatment, would refrain from smoking for at least 6 months.

- 9.7** ● Given below are the sodium contents (in mg) for seven brands of hot dogs rated as “very good” by *Consumer Reports* (consumerreports.org):

420 470 350 360 270 550 530

- Use the given data to produce a point estimate of μ , the true mean sodium content for hot dogs.
- Use the given data to produce a point estimate of σ^2 , the variance of sodium content for hot dogs.
- Use the given data to produce an estimate of σ , the standard deviation of sodium content. Is the statistic you used to produce your estimate unbiased? (Hint: See the discussion following Example 9.3.)

- 9.8** ● A random sample of $n = 12$ four-year-old red pine trees was selected, and the diameter (in inches) of each tree’s main stem was measured. The resulting observations are given here:

11.3 10.7 12.4 15.2 10.1 12.1 16.2 10.5
11.4 11.0 10.7 12.0

- Calculate a point estimate of σ , the population standard deviation of main stem diameter. What statistic did you use to obtain your estimate?

- Suppose that the diameter distribution is normal. Then the 90th percentile of the diameter distribution is $\mu + 1.28\sigma$ (so 90% of all trees have diameters less than this value). Calculate a point estimate for this percentile. (Hint: First calculate an estimate of μ and then use it along with your estimate of σ from Part (a).)

- 9.9** ● A random sample of 10 houses heated with natural gas in a particular area is selected, and the amount of gas (in therms) used during the month of January is determined for each house. The resulting observations are as follows:

103 156 118 89 125 147 122 109 138 99

- Let μ_J denote the average gas usage during January by all houses in this area. Calculate a point estimate of μ_J .
- Suppose that 10,000 houses in this area use natural gas for heating. Let τ denote the total amount of gas used by all of these houses during January. Estimate τ using the given data. What statistic did you use to calculate your estimate?
- Use the given data to estimate p , the proportion of all houses that used at least 100 therms.
- Give a point estimate of the population median usage based on the given data. Which statistic did you use?

SECTION 9.2 Large-Sample Confidence Interval for a Population Proportion

In Section 9.1, we saw how to use a statistic to produce a point estimate of a population characteristic. The value of a point estimate depends on which sample, out of all the possible samples, happens to be selected. Different samples usually produce different estimates as a result of chance differences from one sample to another. Because of sampling variability, rarely is the point estimate from a sample exactly equal to the actual value of the population characteristic. We hope that the chosen statistic produces an estimate that is close, on average, to the true value.

Although a point estimate may represent our best single-number estimate for the value of the population characteristic, it is not the only plausible value. As an alternative to a point estimate, we can use the sample data to report an interval of plausible values for the population characteristic. For example, we might be confident that for all text messages sent from cell phones, the proportion p of messages that are longer than 50 characters is in the interval from 0.53 to 0.57. The narrowness of this interval implies that we have rather precise information about the value of p . If, with the same confidence, we could only state that p was between 0.32 and 0.74, we don’t really know much about the value of p .

DEFINITION

Confidence interval: An interval of plausible values for a population characteristic.

A confidence interval is constructed so that we have a chosen level of confidence that the actual value of the population characteristic will be between the lower endpoint and the upper endpoint of the interval.

Associated with each confidence interval is a **confidence level**. The confidence level provides information on how much “confidence” we can have in the *method* used to construct the interval estimate (*not* our confidence in any one particular interval).

Usual choices for confidence levels are 90%, 95%, and 99%, although other confidence levels are also possible. If we were to construct a 95% confidence interval using the method to be described shortly, we would be using a method that is “successful” 95% of the time. This means that if this method was used to generate an interval estimate over and over again with different random samples, in the long run 95% of the resulting intervals would include the actual value of the characteristic being estimated. Similarly, a 99% confidence interval is one that is constructed using a method that is, in the long run, successful in capturing the actual value of the population characteristic 99% of the time.

DEFINITION

Confidence level: The success rate of the *method* used to construct a confidence interval.

Many statistical studies are carried out to estimate the proportion of individuals or objects in a population that possess a particular property of interest (whatever we have defined as a “success”). For example, a university administrator might be interested in the proportion of students who prefer a new registration system to the previous registration method. In a different setting, a quality control engineer might be interested in the proportion of defective parts produced by a particular manufacturing process.

Recall that p denotes the proportion of individuals in the population that possess the property of interest (successes). Previously, we used the sample proportion

$$\hat{p} = \frac{\text{number in the sample that possess the property of interest}}{n}$$

to calculate a point estimate of p . We can also use \hat{p} to construct a confidence interval for p .

Although a small-sample confidence interval for p can be obtained, our focus is on the large-sample case. The construction of the large-sample interval is based on properties of the sampling distribution of the statistic \hat{p} :

1. The sampling distribution of \hat{p} is centered at p . This means that $\mu_{\hat{p}} = p$, so \hat{p} is an unbiased statistic for estimating p .

2. As long as the sample size is less than 10% of the population size, the standard

$$\text{deviation of } \hat{p} \text{ is well approximated by } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

3. As long as n is large ($np \geq 10$ and $n(1-p) \geq 10$), the sampling distribution of \hat{p} is well approximated by a normal curve.

The accompanying box summarizes these properties.

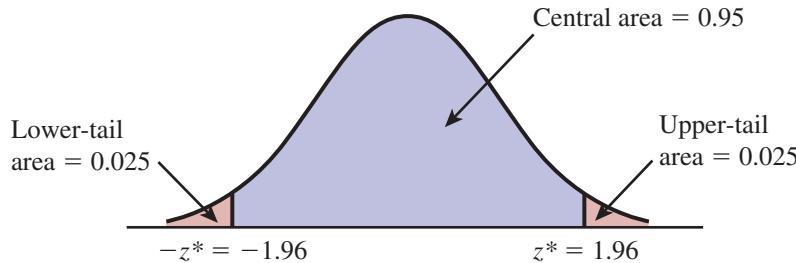
When n is large and the sample size is less than 10% of the population size, the statistic \hat{p} has a sampling distribution that is approximately normal with mean

$$p \text{ and standard deviation } \sqrt{\frac{p(1-p)}{n}}.$$

The development of a confidence interval for p is easier to follow if we begin by selecting a particular confidence level. For a confidence level of 95%, the table of standard normal (z) curve areas (Appendix Table 2) can be used to determine a value z^* such that a central area of 0.95 falls between $-z^*$ and z^* . In this case, the remaining area of 0.05 is divided equally between the two tails, as shown in Figure 9.2. The total area to the left of the desired z^* is 0.975 (0.95 central area + 0.025 area below $-z^*$). By locating 0.9750 in the body of Appendix Table 2, we find that the corresponding z critical value is $z^* = 1.96$. A statistics software package or a graphing calculator could also have been used to determine the value of z^* .

FIGURE 9.2

Capturing a central area of 0.95 under the z curve.



Generalizing this result to normal distributions other than the standard normal distribution tells us that for *any* normal distribution, about 95% of the values are within 1.96 standard deviations of the mean. For large random samples, the sampling distribution of

\hat{p} is approximately normal with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, which leads to the following result.

When n is large, approximately 95% of all random samples of size n will result in a value of \hat{p} that is within $1.96 \sigma_{\hat{p}} = 1.96 \sqrt{\frac{p(1-p)}{n}}$ of the actual value of the population proportion p .

If \hat{p} is within $1.96 \sqrt{\frac{p(1-p)}{n}}$ of p , this means the interval

$$\hat{p} - 1.96 \sqrt{\frac{p(1-p)}{n}} \quad \text{to} \quad \hat{p} + 1.96 \sqrt{\frac{p(1-p)}{n}}$$

will capture p . This will happen for 95% of all possible samples. However, if \hat{p} is farther away from p than $1.96 \sqrt{\frac{p(1-p)}{n}}$ (which will happen for about 5% of all possible samples), the interval will not include the value of p . This is shown in Figure 9.3.

Because \hat{p} is within $1.96\sigma_{\hat{p}}$ of p for 95% of all possible random samples, this means that in repeated random sampling, 95% of the intervals

$$\hat{p} - 1.96 \sqrt{\frac{p(1-p)}{n}} \quad \text{to} \quad \hat{p} + 1.96 \sqrt{\frac{p(1-p)}{n}}$$

will contain the value of p .

Since the value of p is unknown, $\sqrt{\frac{p(1-p)}{n}}$ must be estimated. As long as the sample size is large, the value of $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ can be used in place of $\sqrt{\frac{p(1-p)}{n}}$.

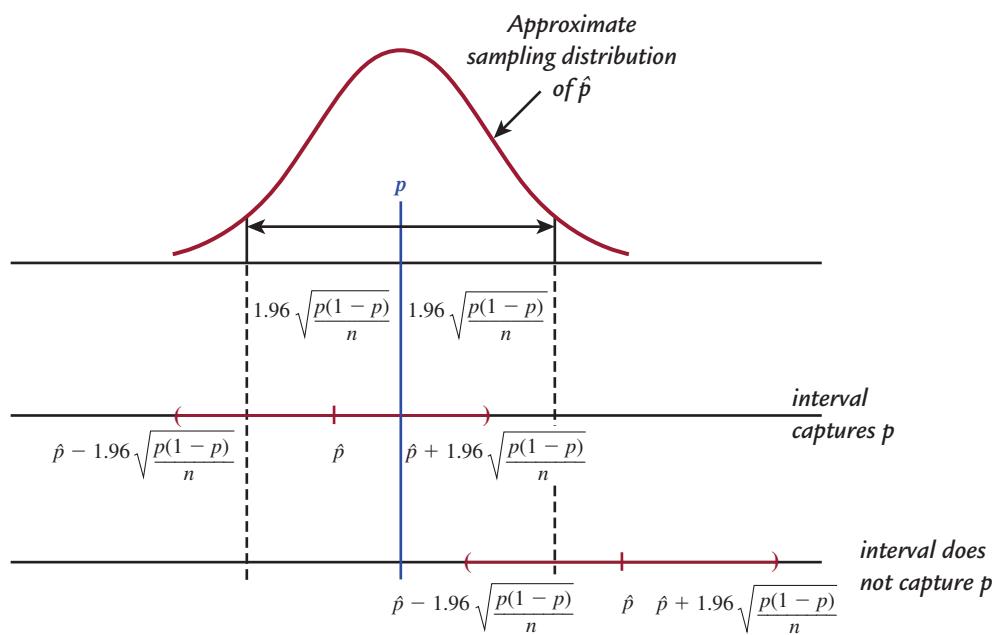
FIGURE 9.3

The population proportion p is captured in the interval from

$$\hat{p} - 1.96 \sqrt{\frac{p(1-p)}{n}} \text{ to}$$

$$\hat{p} + 1.96 \sqrt{\frac{p(1-p)}{n}} \text{ when}$$

$$\hat{p} \text{ is within } 1.96 \sqrt{\frac{p(1-p)}{n}} \text{ of } p.$$



When n is large, a **95% confidence interval for p** is

$$\left(\hat{p} - 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

This can also be written as

$$\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p} + 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ gives the upper endpoint of the interval and

$\hat{p} - 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ gives the lower endpoint of the interval.

This interval can be used as long as

1. $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$,
2. the sample size is less than 10% of the population size if sampling is without replacement, and
3. the sample can be regarded as a random sample from the population of interest.

Example 9.4 College Education Essential for Success?

Understand the context ▶

The article “How Well Are U.S. Colleges Run?” (*USA TODAY*, February 17, 2010) describes a survey of 1031 adult Americans. The survey was carried out by the National Center for Public Policy and the sample was selected in a way that makes it reasonable to regard the sample as representative of adult Americans. Of those surveyed, 567 indicated that they believed a college education is essential for success.

With p denoting the proportion of all adult Americans who believe that a college education is essential for success, a point estimate of p is

$$\hat{p} = \frac{567}{1031} = 0.55$$

Before calculating a 95% confidence interval to estimate p , we should check to make sure that the three necessary conditions are met:

1. $n\hat{p} = 1031(0.55) = 567$ and $n(1 - \hat{p}) = 1031(1 - 0.55) = 1031(0.45) = 364$ are both greater than or equal to 10, so the sample size is large enough to proceed.
2. The sample size of $n = 1031$ is much smaller than 10% of the population size (the number of adult Americans).
3. The sample was selected in a way designed to produce a representative sample. So, it is reasonable to regard the sample as a random sample from the population.

Formulate a plan ➤ Because all three conditions are met, it is appropriate to use the sample data to construct a 95% confidence interval for p .

Do the work ➤ A 95% confidence interval for p is

$$\begin{aligned}\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} &= 0.55 \pm 1.96 \sqrt{\frac{(0.55)(1 - 0.55)}{1031}} \\ &= 0.55 \pm (1.96)(0.015) \\ &= 0.55 \pm 0.029 \\ &= (0.521, 0.579)\end{aligned}$$

Interpret the results ➤ Based on this sample, we can be 95% confident that p , the proportion of adult Americans who believe a college education is essential for success, is between 0.521 and 0.579. We used a *method* to construct this estimate that will successfully capture the actual value of p 95% of the time in the long run.

The 95% confidence interval for p calculated in Example 9.4 is $(0.521, 0.579)$. It is tempting to say that there is a “probability” of 0.95 that p is between 0.521 and 0.579. *Do not yield to this temptation!* The 95% refers to the percentage of *all* possible samples resulting in an interval that includes the value of p . In other words, if we take random sample after random sample from the population and use each one separately to calculate a 95% confidence interval, in the long run roughly 95% of these intervals will capture the value of p .

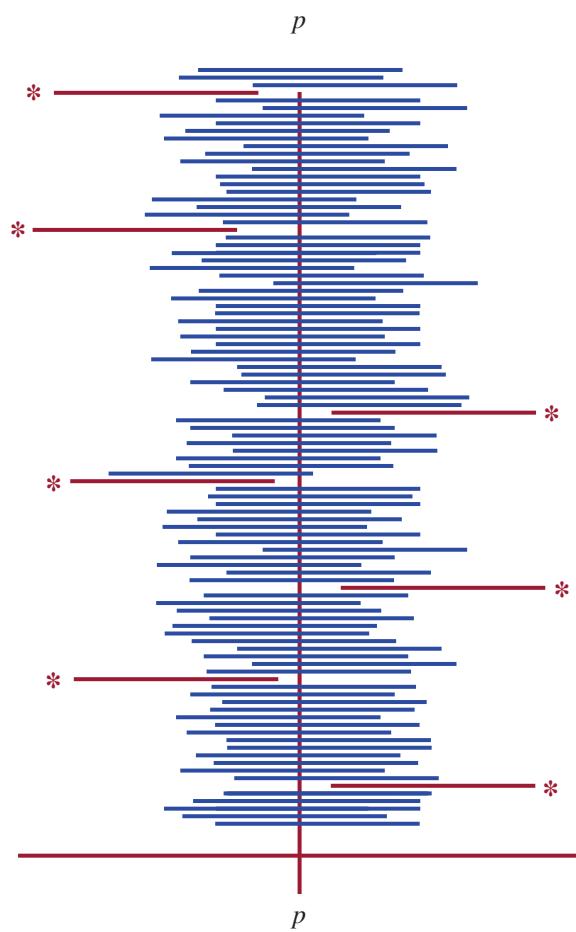
Figure 9.4 illustrates this concept for intervals generated from 100 different random samples. In this particular set of 100 intervals, 93 include p , whereas 7 do not. Any specific interval, and our interval $(0.521, 0.579)$ in particular, either includes p or it does not (remember, the value of p is a fixed number but not known to us). We cannot make a chance (probability) statement concerning this particular interval. *The confidence level 95% refers to the method used to construct the interval rather than to any particular interval, such as the one we obtained.*

The formula given for a 95% confidence interval can easily be adapted for other confidence levels. The choice of a 95% confidence level led to the use of the z value 1.96 (chosen to capture a central area of 0.95 under the standard normal curve) in the confidence interval formula. Any other confidence level can be obtained by using an appropriate z critical value in place of 1.96.

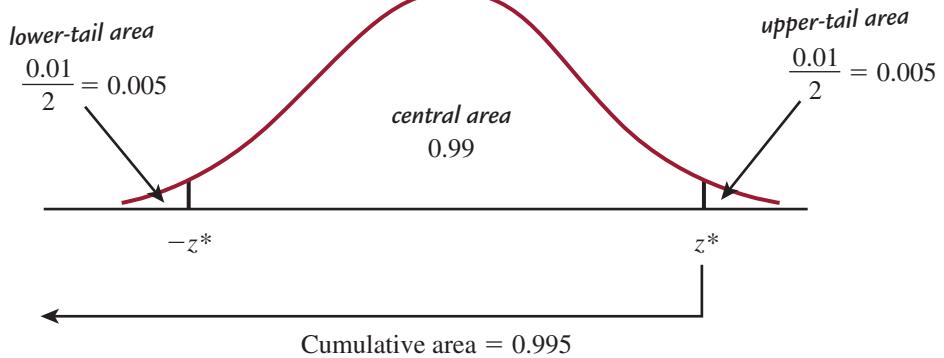
For example, suppose that we wanted to achieve a confidence level of 99%. To obtain a central area of 0.99, the appropriate z critical value would have a cumulative area (area to the left) of 0.995, as illustrated in Figure 9.5. From Appendix Table 2, we find that the corresponding z critical value is $z = 2.58$. A 99% confidence interval for p is then obtained by using 2.58 in place of 1.96 in the formula for the 95% confidence interval.

FIGURE 9.4

One hundred 95% confidence intervals for p calculated from 100 different random samples (asterisks identify intervals that do not include p).

**FIGURE 9.5**

Finding the z critical value for a 99% confidence level.



Why settle for 95% confidence when 99% confidence is possible? Because the higher confidence level comes at a price. The resulting interval is wider than the 95% interval.

The width of the 95% interval is $2\left(1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right)$, whereas the 99% interval has width

$2\left(2.58 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}\right)$. The higher *reliability* of the 99% interval (where “reliability” is specified by the confidence level) entails a loss in precision (as indicated by the wider interval). In the opinion of many investigators, a 95% confidence interval produces a reasonable compromise between reliability and precision.

The Large-Sample Confidence Interval for p

The general formula for a confidence interval for a population proportion p when

1. \hat{p} is the sample proportion from a **simple random sample**,
 2. the sample size **n is large** ($n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$), and
 3. if the sample is selected without replacement, **the sample size is small relative to the population size** (n is at most 10% of the population size)*
- is

$$\hat{p} \pm (z \text{ critical value}) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

The desired confidence level determines which z critical value is used. The three most commonly used confidence levels, 90%, 95%, and 99%, use z critical values 1.645, 1.96, and 2.58, respectively.

Note: This interval is not appropriate for small samples. It is possible to construct a confidence interval in the small-sample case, but this is beyond the scope of this textbook.

*In Chapter 7, we saw a different situation where a similar condition is introduced, but where the requirement was that at most 5% of the population is included in the sample. Be careful not to confuse these two rules.

Example 9.5 Babies on Social Media

Understand the context ➤



Westend61/Getty Images

The article “[Have a Social Media Account for Your Baby? 40% of Millennial Moms Do](#)” ([today.com/parents/have-social-media-account-your-baby-40-percent-millennial-moms-1D80224937](#), October 18, 2014) reported on a survey conducted by Gerber of 1000 women age 18 to 34 who had children under the age of 2. Of those surveyed, 400 had created a social media account for their babies before their first birthday. If it is reasonable to regard this sample of 1000 as representative of the population of millennial moms, we can use this information to construct an estimate of all millennial moms who have created a social media account for their babies before their first birthday.

For this sample

$$\hat{p} = \frac{400}{1000} = 0.400$$

Formulate a plan ➤

Because the sample size is less than 10% of the population size and $n\hat{p} = 400$ and $n(1 - \hat{p}) = 600$ are both greater than or equal to 10, the conditions necessary for appropriate use of the formula for a large-sample confidence interval are met.

A 90% confidence interval for p is then

$$\begin{aligned} \text{Do the work ➤ } \hat{p} \pm (z \text{ critical value}) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} &= 0.400 \pm 1.645 \sqrt{\frac{(0.400)(1 - 0.400)}{1000}} \\ &= 0.400 \pm (1.645)(0.015) \\ &= 0.400 \pm 0.025 \\ &= (0.375, 0.425) \end{aligned}$$

Interpret the results ➤

Based on these sample data, we can be 90% confident that the proportion of all millennial moms who have created a social media account for their baby before the baby’s first birthday is between 0.375 and 0.425. We have used a *method* to construct this interval estimate that has a 10% error rate.

The confidence level for the z confidence interval for a population proportion is only approximate. When we report a 95% confidence interval for a population proportion, the 95% confidence level implies that we have used a method that produces an interval that includes the actual value of the population proportion 95% of the time in repeated random sampling. In fact, because the normal distribution is only an approximation to the sampling distribution of \hat{p} , the actual confidence level may differ somewhat from the reported value. If the conditions (1) $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$ and (2) n is at most 10% of the population size if sampling without replacement are met, the normal approximation is reasonable and the actual confidence level is usually not too different from the reported level. This is why it is important to check these conditions before calculating and reporting a large-sample z confidence interval for a population proportion.

What should you do if these conditions are not met? If the sample size is too small to satisfy the $n\hat{p}$ and $n(1 - \hat{p})$ greater than or equal to 10 condition, an alternative procedure can be used. One alternative method is the bootstrap confidence interval introduced in Section 9.5. If the condition that the sample size is less than 10% of the population size when sampling without replacement is not satisfied, the large-sample z confidence interval tends to be conservative (that is, it tends to be wider than is necessary to achieve the desired confidence level). In this case, a finite population correction factor can be used to obtain a more precise interval. Consult a statistician or a more advanced textbook for more information.

An Alternative to the Large-Sample z Interval

Investigators have shown that in some instances, even when the sample size conditions of the large-sample z confidence interval for a population proportion are met, the actual confidence level associated with the method may be noticeably different from the reported confidence level.

A modified interval that has an actual confidence level that is closer to the reported confidence level is based on a modified sample proportion, \hat{p}_{mod} , the proportion of successes after adding two successes and two failures to the sample. Then \hat{p}_{mod} is

$$\hat{p}_{\text{mod}} = \frac{\text{number of successes} + 2}{n + 4}$$

\hat{p}_{mod} is used in place of \hat{p} in the usual confidence interval formula. Properties of this modified confidence interval are investigated in Activity 9.2 at the end of the chapter.

General Form of a Confidence Interval

Many confidence intervals have the same general form as the large-sample z confidence interval for p just considered. We started with a point estimate based on the statistic \hat{p} . The standard deviation of this statistic is $\sqrt{p(1 - p)/n}$. This resulted in a confidence interval of the form

$$\left(\begin{array}{l} \text{point estimate using} \\ \text{a specified statistic} \end{array} \right) \pm (\text{critical value}) \left(\begin{array}{l} \text{standard deviation} \\ \text{of the statistic} \end{array} \right)$$

Because p was unknown, we estimated the standard deviation of the statistic by $\sqrt{\hat{p}(1 - \hat{p})/n}$, which yielded an interval of the form

$$\left(\begin{array}{l} \text{point estimate using} \\ \text{a specified statistic} \end{array} \right) \pm (\text{critical value}) \left(\begin{array}{l} \text{estimated} \\ \text{standard deviation} \\ \text{of the statistic} \end{array} \right)$$

For a population characteristic other than p , a statistic for estimating the characteristic is selected. Then (drawing on statistical theory) a formula for the standard deviation of the

statistic is given. In practice, it is almost always necessary to estimate this standard deviation, so that the confidence interval has the following form:

$$\left(\begin{array}{l} \text{point estimate using} \\ \text{a specified statistic} \end{array} \right) \pm (\text{critical value}) \left(\begin{array}{l} \text{estimated} \\ \text{standard deviation} \\ \text{of the statistic} \end{array} \right)$$

It is common practice to refer to both the standard deviation of a statistic and the *estimated* standard deviation of a statistic as the **standard error**. In this textbook, when we use the term standard error, we mean the estimated standard deviation of a statistic.

DEFINITION

Standard error: The estimated standard deviation of a statistic.

The 95% confidence interval for p is based on the fact that, for approximately 95% of all random samples, \hat{p} is within $1.96 \sqrt{\frac{p(1-p)}{n}}$ of p . The quantity $1.96 \sqrt{\frac{p(1-p)}{n}}$ is sometimes called the **margin of error** associated with a 95% confidence level. This means that we have 95% confidence that the point estimate \hat{p} is no farther than this quantity from p .

DEFINITION

Margin of error: If the sampling distribution of a statistic is (at least approximately) normal, the **margin of error**, M , associated with a 95% confidence interval is $(1.96) \cdot (\text{standard error of the statistic})$.

Choosing the Sample Size

Before collecting any data, an investigator may wish to determine a sample size for which a particular value of the margin of error is achieved. For example, with p representing the actual proportion of students at a university who purchase textbooks online, the objective of an investigation may be to estimate p to within 0.05 with 95% confidence.

The value of n necessary to achieve this is obtained by setting $1.96 \sqrt{\frac{p(1-p)}{n}}$ equal to 0.05 and solving for n .

In general, suppose that we wish to estimate p to within an amount M (the specified margin of error) with 95% confidence. To find the necessary sample size, consider the equation

$$M = 1.96 \sqrt{\frac{p(1-p)}{n}}$$

Solving this equation for n results in

$$n = p(1-p) \left(\frac{1.96}{M} \right)^2$$

Unfortunately, the use of this formula requires the value of p , which is unknown. One possible way to proceed is to carry out a preliminary study and use the resulting data to get a rough estimate of p . In other cases, prior knowledge may suggest a reasonable estimate of p .

If there is no reasonable basis for estimating p and a preliminary study is not feasible, a conservative solution follows from the observation that $p(1-p)$ is never larger than 0.25 (its value when $p = 0.5$). Replacing $p(1-p)$ with 0.25, the maximum value, yields

$$n = 0.25 \left(\frac{1.96}{M} \right)^2$$

Using this formula to obtain n gives us a sample size for which we can be 95% confident that \hat{p} will be within M of p , no matter what the value of p .

The sample size required to estimate a population proportion p to within an amount M with 95% confidence is

$$n = p(1 - p) \left(\frac{1.96}{M} \right)^2$$

The value of p may be estimated using prior information. In the absence of any such information, using $p = 0.5$ in this formula gives a conservatively large value for the required sample size (this value of p gives a larger n than would any other value).

Example 9.6 Sniffing Out Cancer

Understand the context ➤

Dogs have a sense of smell that is much more powerful than humans. Because of this, dogs can be trained to identify the presence of odors unique to specific types of cancer. The article [“Meet the Dogs Who Can Sniff Out Cancer Better Than Some Lab Tests” \(cnn.com/2015/11/20/health/cancer-smelling-dogs, February 3, 2016, retrieved April 4, 2018\)](https://www.cnn.com/2015/11/20/health/cancer-smelling-dogs) describes a large-scale study that is being planned by researchers to discern if dogs really can identify the presence of cancer by sniffing urine samples.

Suppose we want to collect data that would allow us to estimate the long-run proportion of accurate identifications for a particular dog that has completed training. The dog has been trained to lie down when presented with a urine specimen from a patient with cancer and to remain standing when presented with a specimen from a person who does not have cancer. How many different urine specimens should be used if we want to estimate the long run proportion of correct identifications for this dog to within 0.05 with 95% confidence?

Do the work ➤

Using a conservative value of $p = 0.5$ in the formula for required sample size gives

$$n = p(1 - p) \left(\frac{1.96}{M} \right)^2 = (0.5)(0.5) \left(\frac{1.96}{0.05} \right)^2 = 384.16$$

Interpret the results ➤

This means that a sample of at least 385 breath specimens should be used. Notice that in sample size calculations, we always round up.

EXERCISES 9.10 - 9.35

- 9.10** Explain which would result in a wider large-sample confidence interval for p : a 90% confidence level or a 95% confidence level. (Hint: Consider the confidence interval formula.)
- 9.11** Explain which would result in a wider large-sample confidence interval for p : a sample size of $n = 100$ or $n = 400$?
- 9.12** The formula used to calculate a large-sample confidence interval for p is

$$\hat{p} \pm (z \text{ critical value}) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

What is the appropriate z critical value for each of the following confidence levels?

- | | |
|---------------|---------------|
| a. 95% | d. 80% |
| b. 90% | e. 85% |
| c. 99% | |

- 9.13** The use of the interval

$$\hat{p} \pm (z \text{ critical value}) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

requires a large sample. For each of the following combinations of n and \hat{p} , indicate whether the sample size is large enough for use of this interval to be appropriate.

- a. $n = 50$ and $\hat{p} = 0.30$
- b. $n = 50$ and $\hat{p} = 0.05$
- c. $n = 15$ and $\hat{p} = 0.45$
- d. $n = 100$ and $\hat{p} = 0.01$

- 9.14** For each of the following combinations of sample size and sample proportion, indicate whether the sample size is large enough for appropriate use of the large-sample confidence interval for a population proportion.
- a. $n = 100$ and $\hat{p} = 0.70$
 - b. $n = 40$ and $\hat{p} = 0.25$
 - c. $n = 60$ and $\hat{p} = 0.25$
 - d. $n = 80$ and $\hat{p} = 0.10$

- 9.15** Discuss how each of the following factors affects the width of the confidence interval for p . (Hint: Consider the confidence interval formula.)
- a. The confidence level
 - b. The sample size
 - c. The value of \hat{p}

- 9.16** The *USA TODAY Snapshot* titled “[Social Media Jeopardizing Your Job?](#)” (*USA TODAY*, November 12, 2014) summarized data from a survey of 1855 recruiters and human resource professionals. The Snapshot indicated that 53% of the people surveyed had reconsidered a job candidate based on his or her social media profile. Assume that the sample is representative of the population of recruiters and human resource professionals in the United States.
- a. Use the given information to estimate the proportion of recruiters and human resource professionals who have reconsidered a job candidate based on his or her social media profile using a 95% confidence interval. Give an interpretation of the interval in context and an interpretation of the confidence level of 95%.
 - b. Would a 90% confidence interval be wider or narrower than the 95% confidence interval from Part (a)? Explain.

- 9.17** Based on data from a survey of 1200 randomly selected Facebook users (*USA TODAY*, March 24, 2010), the following is a 90% confidence interval for the proportion of all Facebook users who say it is not OK to “friend” someone who reports to you at work: (0.60, 0.64). What is the meaning of the 90% confidence level associated with this interval?

- 9.18** The report “[Parents, Teens and Digital Monitoring](#)” (Pew Research Center, January 7, 2016, pewinternet.org/2016/01/07/parents-teens-and-digital-monitoring, retrieved April 4, 2018) reported that 61% of parents of teens age 13 to 17 said that they had checked which web sites their teens had visited. The 61% figure was based on a representative sample of 1060 parents of teens in this age group.
- a. What assumption must be made in order for it to be appropriate to use the formula of this section

to construct a confidence interval to estimate the proportion of all parents of teens age 13 to 17 who have checked which web sites their teens visit?

- b. Construct and interpret a 90% confidence interval for the proportion of all parents of teens age 13 to 17 who have checked which web sites their teens visit. (Hint: See Example 9.5.)
- c. Would a 99% confidence interval be narrower or wider than the interval calculated in Part (b)? Justify your answer.

- 9.19** If a hurricane was headed your way, would you evacuate? The headline of a press release issued [January 21, 2009](#) by the survey research company [International Communications Research \(icrsurvey.com\)](#) states, “Thirty-one Percent of People on High-Risk Coast Will Refuse Evacuation Order, Survey of Hurricane Preparedness Finds.” This headline was based on a survey of 5046 adults who live within 20 miles of the coast in high hurricane risk counties of eight southern states. In selecting the sample, care was taken to ensure that the sample would be representative of the population of coastal residents in these states.
- a. Use this information to estimate the proportion of coastal residents who would evacuate using a 98% confidence interval.
 - b. Write a few sentences interpreting the interval and the confidence level associated with the interval.

- 9.20** The *USA TODAY Snapshot* titled “[Big Bang Theory](#)” (*USA TODAY*, October 14, 2016) summarized data from a sample of 1003 American parents of children age 6 to 11. It reported that 53% of these parents view science-oriented TV shows as a good way to expose their children to science outside of school. Assume that this sample is representative of American parents of children age 6 to 11. Construct and interpret a 90% confidence interval for the proportion of American parents of children age 6 to 11 who view science-oriented TV shows as a good way to expose their children to science.

- 9.21** The article “[Most Dog Owners Take More Pictures of Their Pet Than Their Spouse](#)” (August 22, 2016, news.fastcompany.com/most-dog-owners-take-more-pictures-of-their-pet-than-their-spouse-4017458, retrieved May 6, 2017) indicates that in a sample of 1000 dog owners, 650 said that they take more pictures of their dog than their significant others or friends. In addition, 460 said that they are more likely to complain to their dog than to a friend. Suppose that it is reasonable to consider this sample as representative of the population of dog owners.

- a. Construct and interpret a 90% confidence interval for the proportion of dog owners who take more pictures of their dog than of their significant others or friends.
- b. Construct and interpret a 95% confidence interval for the proportion of dog owners who are more likely to complain to their dog than to a friend.
- c. Give two reasons why the confidence interval in Part (b) is wider than the interval in Part (a).

- 9.22** The *Princeton Review 2016 College Hopes and Worries Survey Report* (princetonreview.com/cms-content/final_cohowo2016survrpt.pdf, retrieved May 6, 2017) reported that 31% of students applying to college wanted to attend a college that was within 250 miles of their home and that 51% of parents of students applying to college wanted their child to attend a college that was within 250 miles from home. Suppose that the reported percentages were based on random samples of 8347 students applying to college and of 2087 parents of students applying to college.
- a. Construct and interpret a 90% confidence interval for the proportion of students applying to college who want to attend a college within 250 miles of their home.
 - b. Construct and interpret a 90% confidence interval for the proportion of parents of students applying to college who want their child to attend a college within 250 miles from home.
 - c. Explain why the two 90% confidence intervals are not the same width.

- 9.23** The *USA TODAY Snapshot* titled “**Baby’s First Photo Reveal**” (*USA TODAY*, October 17, 2014) summarized data from a survey of 1001 mothers with children under the age of 2. The Snapshot includes the following statement: “83% of moms post new baby photos from the delivery room.” This information could be used to provide an estimate of the proportion of new mothers who post pictures on social media from the delivery room. Construct and interpret a 99% confidence interval for the proportion of mothers of children under the age of 2 who posted pictures of their new baby on social media from the delivery room.

- 9.24** The report “**Job Seeker Nation Study 2016**” (jobvite.com/wp-content/uploads/2016/03/Jobvite_Jobseeker_Nation_2016.pdf, retrieved May 6, 2017) summarized a survey of 2305 working adults. The report indicates that 484 of the working adults surveyed said they were very concerned that their job will be automated, outsourced, or otherwise made obsolete in the next 5 years. The sample was selected in a way designed to produce a

representative sample of working adults. Construct and interpret a 95% confidence interval for the proportion of working adults who are very concerned that their job will be automated, outsourced, or otherwise made obsolete in the next 5 years.

- 9.25** *USA TODAY* reported that the proportion of Americans who prefer cheese on their burgers is 0.84 (*USA TODAY*, September 7, 2016). This estimate was based on a survey of a representative sample of 1000 adult Americans. Calculate and interpret a 90% confidence interval for the proportion of Americans who prefer cheese on their burgers.
- 9.26** The *USA TODAY Snapshot* titled “**Have a Nice Trip**” (*USA TODAY*, November 17, 2015) summarized data from a survey of 1000 U.S. adults who had traveled by air at least once in the previous year. The *Snapshot* includes the following statement: “38% admit to yelling at a complete stranger while traveling.” Suppose that the sample was selected to be representative of the population of adults who have traveled by air at least once in the previous year. Calculate and interpret a 95% confidence interval for the population proportion who have yelled at a complete stranger while traveling.
- 9.27** *Business Insider* reported that a study commissioned by eBay motors found that nearly 40% of millennials who drive a car that is more than 5 years old have named their cars (“**Millennials Have an Odd Habit When It Comes to Their Cars**,” April 14, 2016). Suppose that this statement was based on a sample of size 800 and that 312 reported that they had named their car. Assuming that the sample was selected to be representative of the population of millennials who drive a car that is more than 5 years old, calculate and interpret a 90% confidence interval for the proportion of all millennials who drive a car that is more than 5 years old who have named their cars.
- 9.28** In 2010, the National Football League adopted new rules designed to limit head injuries. In a survey conducted in 2015 by the Harris Poll, 1216 of 2096 adults indicated that they were football fans and followed professional football. Of those who were football fans, 692 said they thought that the new rules were effective in limiting head injuries.
- a. If the sample is representative of adults in the United States, construct and interpret a 95% confidence interval for the proportion of U.S. adults who consider themselves to be football fans.
 - b. Construct and interpret a 95% confidence interval for the proportion of football fans who think

- that the new rules have been effective in limiting head injuries.
- c. Explain why the confidence intervals in Parts (a) and (b) are not the same width even though they both have a confidence level of 95%.
- 9.29** The article “[Most Americans Don’t Understand the Cloud, But They Should](#)” ([foxbusiness.com, October 17, 2016, retrieved November 12, 2016](#)) reported that in a sample of 1000 people, 22% said they have pretended to know what the cloud is or how it works. If it is reasonable to regard the sample as representative of adult Americans, an estimate of the proportion who have pretended to know what the cloud is or how it works is 0.22. The margin of error associated with this estimate is 0.026. Interpret the value of this margin of error.
- 9.30** The Gallup Organization conducts an annual survey on crime. It was reported that 25% of all households experienced some sort of crime during the past year. This estimate was based on a sample of 1002 randomly selected households. The report states, “One can say with 95% confidence that the margin of sampling error is ± 3 percentage points.” Explain how this statement can be justified.
- 9.31** The article “[Hospitals Dispute Medtronic Data on Wires](#)” ([The Wall Street Journal, February 4, 2010](#)) describes several studies of the failure rate of defibrillators used in the treatment of heart problems. In one study conducted by the Mayo Clinic, it was reported that failures were experienced within the first 2 years by 18 of 89 patients under 50 years old and 13 of 362 patients age 50 and older who received a particular type of defibrillator. Assume it is reasonable to regard these two samples as representative of patients in the two age groups who receive this type of defibrillator.
- a. Construct and interpret a 95% confidence interval for the proportion of patients under 50 years old who experience a failure within the first 2 years after receiving this type of defibrillator.
- b. Construct and interpret a 99% confidence interval for the proportion of patients age 50 and older who experience a failure within the first 2 years after receiving this type of defibrillator.
- c. Suppose that the researchers wanted to estimate the proportion of patients under 50 years old who experience a failure within the first 2 years after receiving this type of defibrillator to within 0.03 with 95% confidence. How large a sample should be used? Use the results of the study as a preliminary estimate of the population proportion. (Hint: See Example 9.6.)
- 9.32** Based on survey of a representative sample of 1000 adult Americans, YouGov estimated that the proportion of adult Americans who have less than \$1000 in savings is 0.430 (“[People More Likely to Save with an Opt-Out System](#),” [today.yougov.com/news/2016/04/25/savings/](#)). The margin of error for a 95% confidence level associated with this estimate is 0.03. Show how this margin of error was calculated.
- 9.33** A discussion of digital ethics appears in the article “[Academic Cheating, Aided by Cell Phones or Web, Shown to be Common](#)” ([Los Angeles Times, June 17, 2009](#)). One question posed in the article is: What proportion of college students have used cell phones to cheat on an exam? Suppose you have been asked to estimate this proportion for students enrolled at a large university. How many students should you include in your sample if you want to estimate this proportion to within 0.02 with 95% confidence?
- 9.34** In spite of the potential safety hazards, some people would like to have an Internet connection in their car. A preliminary survey of adult Americans has estimated this proportion to be somewhere around 0.30 ([USA TODAY, May 1, 2009](#)).
- a. Use the given preliminary estimate to determine the sample size required to estimate the proportion of adult Americans who would like an Internet connection in their car to within 0.02 with 95% confidence.
- b. The formula for determining sample size given in this section corresponds to a confidence level of 95%. How would you modify this formula if a 99% confidence level was desired?
- c. Use the given preliminary estimate to determine the sample size required to estimate the proportion of adult Americans who would like an Internet connection in their car to within 0.02 with 99% confidence.
- 9.35** In 2010, the online security firm Symantec estimated that 63% of computer users don’t change their passwords very often ([cnet.com/news/survey-63-don-t-change-passwords-very-often, retrieved November 19, 2016](#)). Suppose that you want to carry out a new survey to estimate the proportion of students at your school who do not change their password.
- a. What is the required sample size if you want to estimate this proportion with a margin of error of 0.05? Calculate the required sample size first using 0.63 as a preliminary estimate of p and then using the conservative value of 0.5. How do the two sample sizes compare?
- b. What sample size would you recommend? Justify your answer.

SECTION 9.3 Confidence Interval for a Population Mean

In this section, we consider how to use information from a random sample to construct a confidence interval estimate of a population mean, μ . We begin by considering the case in which σ , the population standard deviation, is known and the sample size n is large enough for the Central Limit Theorem to apply. In this case, the following three properties about the sampling distribution of \bar{x} hold:

1. The sampling distribution of \bar{x} is centered at μ , so \bar{x} is an unbiased statistic for estimating μ ($\mu_{\bar{x}} = \mu$).
2. The standard deviation of \bar{x} is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
3. As long as n is large (generally $n \geq 30$), the sampling distribution of \bar{x} is approximately normal, even when the population distribution itself is not normal.

The same reasoning that was used to develop the large-sample confidence interval for a population proportion p can be used to obtain a confidence interval estimate for μ .

The One-Sample z Confidence Interval for μ

The general formula for a confidence interval for a population mean μ when

1. \bar{x} is the sample mean from a **simple random sample**,
2. the **sample size n is large** (generally $n \geq 30$), and
3. σ , the **population standard deviation**, is **known**

is

$$\bar{x} \pm (z \text{ critical value}) \left(\frac{\sigma}{\sqrt{n}} \right)$$

Example 9.7 Cosmic Radiation

Understand the context ▶

Cosmic radiation levels rise with increasing altitude, prompting researchers to consider how pilots and flight crews are affected by increased exposure to cosmic radiation. The Centers for Disease Control and Prevention (CDC) reports that the National Council on Radiation Protection and Measurements estimates that flight crew members have a mean annual radiation exposure of 3.07 millisievert (mSv) per year (cdc.gov/niosh/topics/aircrew/cosmicionizingradiation.html, retrieved December 18, 2016). Suppose that the estimated mean exposure for flight crew members is based on a random sample of 100 flight crew members.

Here, μ will represent the mean annual cosmic radiation exposure for all flight crew members. Although σ , the actual value of the population standard deviation, is not usually known, for purposes of this example, suppose that σ is known to be 0.35 mSv. Because the sample size is large and σ is known, a 95% confidence interval for μ is

z critical value for 95% confidence level

Do the work ▶

$$\begin{aligned} \bar{x} \pm (z \text{ critical value}) \frac{\sigma}{\sqrt{n}} &= 3.07 \pm (1.96) \frac{0.35}{\sqrt{100}} \\ &= 3.07 \pm 0.069 \\ &= (3.001, 3.139) \end{aligned}$$

Interpret the results ▶

Based on this sample, *plausible* values of μ , the mean annual cosmic radiation exposure for all flight crew members, are those between 3.001 and 3.139 mSv. One mSv is the average

annual exposure in the United States due to normal exposure to background radiation. This means that it is estimated that the mean annual exposure of flight crew members is somewhere between about 3 and 3.14 times greater than for the general population. A confidence level of 95% is associated with the method used to produce this interval estimate.

The confidence interval just introduced is appropriate when σ is known and n is large, and it can be used regardless of the shape of the population distribution. This is because this confidence interval is based on the Central Limit Theorem, which says that when n is sufficiently large, the sampling distribution of \bar{x} is approximately normal for any population distribution.

When n is small, the Central Limit Theorem cannot be used to justify the normality of the \bar{x} sampling distribution, so the z confidence interval cannot be used. One way to proceed in the small-sample case is to make a specific assumption about the shape of the population distribution and then to use a method that is valid under this assumption.

One instance where this is easy to do is when the population distribution is normal in shape. Recall that for a normal population distribution, the sampling distribution of \bar{x} is normal even for small sample sizes. So, if n is small but the population distribution is normal, the same confidence interval formula just introduced can still be used.

If it is reasonable to believe that the distribution of values in the population is normal, a confidence interval for μ (when σ is known) is

$$\bar{x} \pm (z \text{ critical value}) \left(\frac{\sigma}{\sqrt{n}} \right)$$

This interval is appropriate even when n is small, as long as it is reasonable to think that the population distribution is normal in shape.

There are several ways that sample data can be used to assess the plausibility of normality. Two common ways are to look at a normal probability plot of the sample data (looking for a plot that is reasonably straight) or to construct a dotplot or a boxplot of the data (looking for approximate symmetry and no outliers).

Confidence Interval for μ When σ Is Unknown

The confidence interval just developed has an obvious drawback: To calculate the interval endpoints, σ must be known. Unfortunately, this is rarely the case in practice. We now turn our attention to the situation when σ is unknown. The development of the confidence interval in this instance depends on the assumption that the population distribution is normal. This assumption is not critical if the sample size is large, but it is important when the sample size is small.

To understand the derivation of this confidence interval, begin by taking another look at the 95% confidence interval when σ is known. We know that $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Also, when the population distribution is normal, the \bar{x} distribution is normal. These facts imply that the standardized variable

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has approximately a standard normal distribution. Because the interval from -1.96 to 1.96 captures an area of 0.95 under the z curve, approximately 95% of all samples result in an \bar{x} value that satisfies

$$-1.96 < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96$$

Rewriting these inequalities to isolate μ in the middle results in the equivalent inequalities:

$$\bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

The term $\bar{x} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$ is the lower endpoint of the 95% large-sample confidence interval for μ , and $\bar{x} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$ is the upper endpoint.

If σ is unknown, we must use the sample data to estimate σ . If we use the sample standard deviation as our estimate, the result is a different standardized variable denoted by t :

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The value of s may not be all that close to σ , especially when n is small. As a consequence, the use of s in place of σ introduces extra variability. The value of z varies from sample to sample, because different samples generally result in different \bar{x} values. There is even more variability in t , because different samples may result in different values of both \bar{x} and s . Because of this, the distribution of t is more spread out than the standard normal (z) distribution.

To develop an appropriate confidence interval, we must investigate the probability distribution of the standardized variable t for a sample from a normal population. This requires that we learn about probability distributions called *t distributions*.

t Distributions

Just as there are many different normal distributions, there are also many different *t* distributions. While normal distributions are distinguished from one another by their mean μ and standard deviation σ , *t* distributions are distinguished by a positive whole number called the number of *degrees of freedom* (df). There is a *t* distribution with 1 df, another with 2 df, and so on.

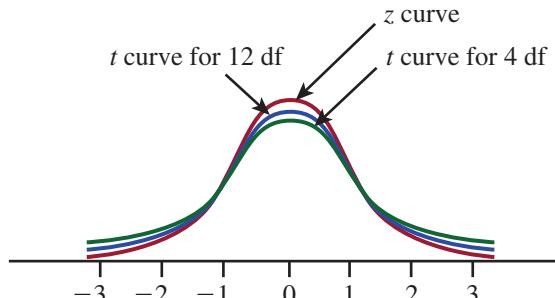
Important Properties of *t* Distributions

1. The *t* distribution corresponding to any particular number of degrees of freedom is bell shaped and centered at zero (just like the standard normal (z) distribution).
2. Each *t* distribution is more spread out than the standard normal (z) distribution.
3. As the number of degrees of freedom increases, the variability of the corresponding *t* distribution decreases.
4. As the number of degrees of freedom increases, the corresponding sequence of *t* distributions approaches the standard normal (z) distribution.

The properties discussed in the preceding box are illustrated in Figure 9.6, which shows two *t* curves along with the z curve.

FIGURE 9.6

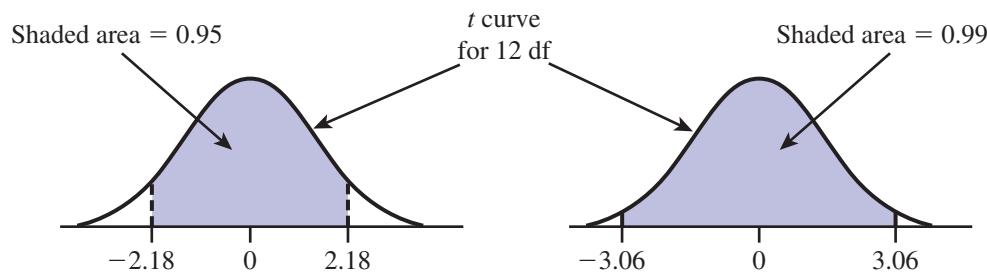
Comparison of the z curve and *t* curves for 12 df and 4 df.



Appendix Table 3 gives selected critical values for various t distributions. The central areas for which values are tabulated are 0.80, 0.90, 0.95, 0.98, 0.99, 0.998, and 0.999. To find a particular critical value, go down the left margin of the table to the row labeled with the desired number of degrees of freedom. Then move over in that row to the column headed by the desired central area.

For example, the value in the 12-df row under the column corresponding to central area 0.95 is 2.18, so 95% of the area under the t curve with 12 df lies between -2.18 and 2.18 . Moving over two columns, we find the critical value for central area 0.99 (still with 12 df) to be 3.06 (see Figure 9.7). Moving down the 0.99 column to the 20-df row, we see the critical value is 2.85, so the area between -2.85 and 2.85 under the t curve with 20 df is 0.99.

FIGURE 9.7
 t critical values illustrated.



Notice that the critical values increase from left to right in each row of Appendix Table 3. This makes sense because as we move to the right, we capture larger central areas. In each column, the critical values decrease as we move downward, reflecting decreasing variability for t distributions with larger degrees of freedom.

The larger the number of degrees of freedom, the more closely the t curve resembles the z curve. To emphasize this, we have included the z critical values as the last row of the t table. Also notice that, once the number of degrees of freedom is greater than 30, the critical values change very little as the number of degrees of freedom increases. For this reason, Appendix Table 3 jumps from 30 df to 40 df, then to 60 df, then to 120 df, and finally to the row of z critical values.

If we need a critical value for a number of degrees of freedom between those tabulated, we just use the critical value for the closest df. For $df > 120$, we use the z critical values. Many graphing calculators and statistical software packages calculate t critical values for any number of degrees of freedom, so if you are using such a calculator, it is not necessary to approximate the t critical values as described.

One-Sample t Confidence Interval

The fact that the sampling distribution of $\frac{\bar{x} - \mu}{(\sigma/\sqrt{n})}$ is approximately the z (standard normal) distribution when n is large led to the z confidence interval when σ is known. In the same way, the following proposition allows us to obtain a confidence interval when the population distribution is normal but σ is unknown.

If \bar{x} and s are the mean and standard deviation of a random sample from a normal population distribution, then the probability distribution of the standardized variable

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

is the t distribution with $df = n - 1$.

To see how this result leads to the desired confidence interval, consider the case $n = 25$. We use the t distribution with $df = n - 1 = 24$. From Appendix Table 3, the interval between -2.06 and 2.06 captures a central area of 0.95 under the t curve with 24 df. This means that 95% of all samples of size $n = 25$ from a normal population result in values of \bar{x} and s for which

$$-2.06 < \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} < 2.06$$

Rewriting these inequalities to isolate μ yields

$$\bar{x} - 2.06\left(\frac{s}{\sqrt{25}}\right) < \mu < \bar{x} + 2.06\left(\frac{s}{\sqrt{25}}\right)$$

The 95% confidence interval for μ in this situation extends from the lower endpoint $\bar{x} - 2.06\left(\frac{s}{\sqrt{25}}\right)$ to the upper endpoint $\bar{x} + 2.06\left(\frac{s}{\sqrt{25}}\right)$. This interval can also be written

$$\bar{x} \pm 2.06\left(\frac{s}{\sqrt{25}}\right)$$

The differences between this interval and the interval when σ is known are the use of the t critical value 2.06 rather than the z critical value 1.96 and the use of the sample standard deviation as an estimate of σ . The extra uncertainty that results from estimating σ is why the t interval is wider than the z interval.

If the sample size is something other than 25 or if the desired confidence level is something other than 95% , a different t critical value (obtained from Appendix Table 3 or by using technology) is used in place of 2.06 .

The One-Sample t Confidence Interval for μ

The general formula for a confidence interval for a population mean μ based on a sample of size n when

1. \bar{x} is the sample mean from a **simple random sample**,
2. the **population distribution is normal, or the sample size n is large** (generally $n \geq 30$), and
3. **σ , the population standard deviation, is unknown**

is

$$\bar{x} \pm (t \text{ critical value})\left(\frac{s}{\sqrt{n}}\right)$$

where the t critical value is based on $df = n - 1$. Appendix Table 3 gives critical values appropriate for each of the confidence levels 90% , 95% , and 99% , as well as several other less frequently used confidence levels.

If n is large (generally $n \geq 30$), the normality of the population distribution is not critical. *However, this confidence interval is appropriate for small n only when the population distribution is (at least approximately) normal.* If this is not the case, as might be suggested by a normal probability plot or boxplot, a different method (for example, the method introduced in Section 9.6) should be used.

Example 9.8 Drive-Through Medicine

Understand the context ➤

During a flu outbreak, many people visit emergency rooms, where they often must wait in crowded waiting rooms where other patients may be exposed. The paper “Drive-Through

“Medicine: A Novel Proposal for Rapid Evaluation of Patients during an Influenza Pandemic” (*Annals of Emergency Medicine* [2010]: 268–273) describes an interesting study of the feasibility of a drive-through model where flu patients are evaluated while they remain in their cars. One of the interesting observations from this study was that not only were patients kept relatively isolated and away from other patients, but the time to process a patient was shorter because delays related to turning over examination rooms were eliminated.

In the study, 38 volunteers were each given a scenario from a randomly selected set of flu cases seen in the emergency room. The scenarios provided the volunteer with a medical history and a description of symptoms that would allow the volunteer to respond to questions from the examining physician. These volunteer patients were then processed using a drive-through procedure that was implemented in the parking structure of Stanford University Hospital. The time to process each case from admission to discharge was recorded.

Consider the data ➤

Data read from a graph that appears in the paper were used to calculate the following summary statistics for admission-to-discharge processing times (in minutes):

$$n = 38 \quad \bar{x} = 26 \quad s = 1.57$$

Formulate a plan ➤

A boxplot of the 38 processing times did show a couple of outliers on the high end, corresponding to unusually long processing times, suggesting that it is probably not reasonable to think of the population distribution of drive-through processing times as being approximately normal. However, because the sample size is greater than 30 and the distribution of sample processing times was not extremely skewed, it is appropriate to consider using the *t* confidence interval to estimate the mean admission-to-discharge processing time for flu patients using the drive-through procedure. Because the 38 flu scenarios were thought to be representative of the population of flu patients seen in emergency rooms and the sample size is large, we can use the formula for the *t* confidence interval to calculate a 95% confidence interval.

For this example, $n = 38$, $df = 37$, and the appropriate *t* critical value is 2.02 (from the 40-df row of Appendix Table 3). The 95% confidence interval is then

Do the work ➤

$$\begin{aligned} \bar{x} \pm (t \text{ critical value}) \left(\frac{s}{\sqrt{n}} \right) &= 26 \pm (2.02) \left(\frac{1.57}{\sqrt{38}} \right) \\ &= 26 \pm 0.514 \\ &= (25.486, 26.514) \end{aligned}$$

Interpret the results ➤

Based on the sample data, we believe that the actual mean admission-to-discharge processing time for flu patients processed using the drive-through procedure is between 25.486 minutes and 26.514 minutes. We used a method that has a 5% error rate to construct this interval. The authors of the paper indicated that the average processing time for flu patients seen in the emergency room was about 90 minutes, so it appears that the drive-through procedure has promise both in terms of keeping flu patients isolated and also in reducing processing time, on average.

Example 9.9 Waiting for Surgery

Understand the context ➤

The authors of the paper “Length of Stay, Wait Time to Surgery and 30-Day Mortality for Patients with Hip Fractures After Opening of a Dedicated Orthopedic Weekend Trauma Room” (*Canadian Journal of Surgery* [2016]: 337–341) were interested in estimating the mean time that patients who broke a hip had to wait for surgery after the opening of a new hospital facility. They reported that for a representative sample of 204 people with a fractured hip, the sample mean time between arriving at the hospital and surgery to repair the hip was 28.5 hours and that the sample standard deviation of the wait times was 16.8 hours.

Formulate a plan ➤

If we had access to the raw data (the 204 individual wait-time observations), we might begin by looking at a boxplot. The authors of the paper commented that there were several outliers

in the data set, which might cause us to question the normality of the wait-time distribution, but because the sample size is large, it is still appropriate to use the t confidence interval.

We can use the confidence interval of this section to estimate the actual mean wait time for surgery. We know the following:

$$\begin{aligned}\text{sample size} &= n = 204 \\ \text{sample mean wait time} &= \bar{x} = 28.5 \text{ hours} \\ \text{sample standard deviation} &= s = 16.8 \text{ hours}\end{aligned}$$

The sample was thought to be representative of the population of patients with a fractured hip. So, with μ denoting the mean wait time for surgery for patients with a fractured hip, we can estimate μ using a 90% confidence interval.

- Do the work ➤** From Appendix Table 3, we use t critical value = 1.645 (from the z critical value row because $df = n - 1 = 203 > 120$, the largest number of degrees of freedom in the table). The 90% confidence interval for μ is

$$\begin{aligned}\bar{x} \pm (t \text{ critical value}) \left(\frac{s}{\sqrt{n}} \right) &= 28.5 \pm (1.645) \left(\frac{16.8}{\sqrt{204}} \right) \\ &= 28.5 \pm 1.94 \\ &= (26.56, 30.44)\end{aligned}$$

- Interpret the results ➤** Based on this sample, we are 90% confident that μ is between 26.56 hours and 30.44 hours.

The paper referenced in Example 9.9 also gave data on surgery wait times for a representative sample of 405 patients with fractured hips who were seen at this hospital before the new facility was opened. The mean wait time for the patients in this sample was 31.5 hours and the standard deviation of wait times was 27.0 hours. Because the sample was a representative sample and the sample size was large, it is appropriate to use the one-sample t confidence interval to estimate the mean wait time for surgery before the new facility was opened.

A graphing calculator or statistical software can be used to produce a one-sample t confidence interval. Using a 90% confidence level, output from Minitab for the sample of patients who had surgery before the new facility was opened is shown in Figure 9.8. The 90% confidence interval for the mean wait time before the new facility extends from 29.29 hours to 33.71 hours.

FIGURE 9.8

Minitab output for wait time for patients who had surgery before the new facility was opened.

One-Sample T

N	Mean	StDev	SE Mean	90% CI
405	31.50	27.00	1.34	(29.29, 33.71)



Alan and Sandy Carey/Getty Images

Example 9.10 Selfish Chimps?

- The article “[Chimps Aren’t Charitable](#)” (*Newsday*, November 2, 2005) summarized the results of a research study published in the journal *Nature*. In this study, chimpanzees learned to use an apparatus that dispensed food when either of two ropes was pulled. When one of the ropes was pulled, only the chimp controlling the apparatus received food. When the other rope was pulled, food was dispensed both to the chimp controlling the apparatus and also to a chimp in the adjoining cage.

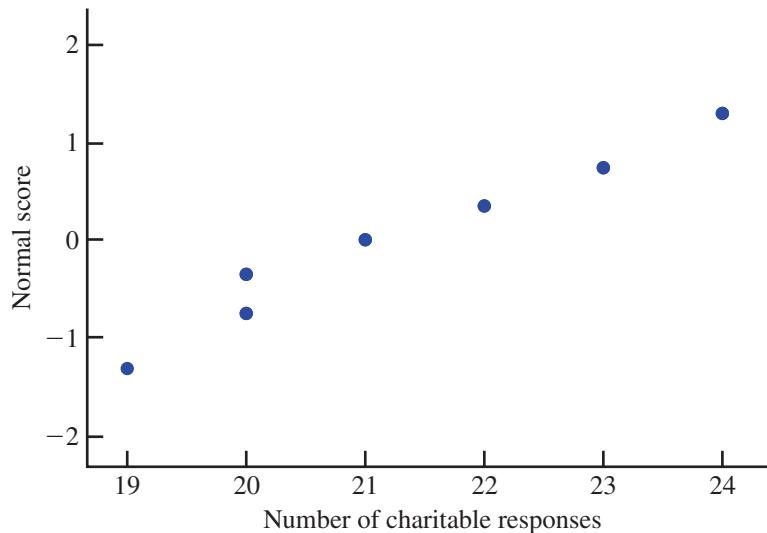
The accompanying data (approximated from a graph in the paper) represent the number of times out of 36 trials that each of seven chimps chose the option that would provide food to both chimps (the “charitable” response).

Formulate a plan ➤

Figure 9.9 is a normal probability plot of these data. The plot is reasonably straight, so it seems plausible that the population distribution of number of charitable responses is approximately normal.

FIGURE 9.9

Normal probability plot for data of Example 9.10



For purposes of this example, let's suppose it is reasonable to regard this sample of seven chimps as representative of the population of all chimpanzees. Calculation of a confidence interval for the mean number of charitable responses for the population of all chimps requires \bar{x} and s . From the given data, we calculate

$$\bar{x} = 21.29 \quad s = 1.80$$

The t critical value for a 99% confidence interval based on 6 df is 3.71. The interval is

Do the work ➤

$$\begin{aligned}\bar{x} \pm (t \text{ critical value})\left(\frac{s}{\sqrt{n}}\right) &= 21.29 \pm (3.71)\left(\frac{1.80}{\sqrt{7}}\right) \\ &= 21.29 \pm 2.52 \\ &= (18.77, 23.81)\end{aligned}$$

A statistical software package could also have been used to calculate the 99% confidence interval. The following is output from SPSS. The slight discrepancy between the hand-calculated interval and the one reported by SPSS occurs because SPSS uses more decimal accuracy in \bar{x} , s , and t critical values.

One-Sample Statistics					
	N	Mean	Std. Deviation	Std. Error Mean	
CharitableResponses	7	21.2857	1.79947	.68014	
One-Sample					
99% Confidence Interval					
	Lower	Upper			
CharitableResponses	18.7642	23.8073			

Interpret the results ➤

With 99% confidence, we estimate the population mean number of charitable responses (out of 36 trials) to be between 18.77 and 23.81. Remember that the 99% confidence level implies that if the same formula is used to calculate intervals for sample after sample randomly selected from the population of chimps, in the long run 99% of these intervals will capture the value of μ between the lower and upper confidence limits.

Notice that based on this interval, we would conclude that, on average, chimps choose the charitable option more than half the time (more than 18 out of 36 trials). The *Newsday*

headline “**Chimps Aren’t Charitable**” was based on additional data from the study indicating that chimps’ charitable behavior was no different when there was another chimp in the adjacent cage than when the adjacent cage was empty. We will revisit this study in Chapter 11 to investigate further.

Choosing the Sample Size

When estimating μ using a large sample or using a small sample from a normal population, the margin of error, M , associated with a 95% confidence interval is

$$M = 1.96 \left(\frac{\sigma}{\sqrt{n}} \right)$$

Before collecting any data, an investigator may wish to determine a sample size for which a particular value of the margin of error is achieved. For example, with μ representing the average fuel efficiency (in miles per gallon, mpg) for all cars of a certain type, the objective of an investigation may be to estimate μ to within 1 mpg with 95% confidence. The value of n necessary to achieve this is obtained by setting $M = 1$ and then solving

$$1 = 1.96 \left(\frac{\sigma}{\sqrt{n}} \right)$$

for n .

In general, suppose that we wish to estimate μ to within an amount M (the specified margin of error) with 95% confidence. Finding the necessary sample size requires solving the equation

$$M = 1.96 \left(\frac{\sigma}{\sqrt{n}} \right)$$

for n . The result is

$$n = \left(\frac{1.96\sigma}{M} \right)^2$$

Notice that the greater the variability in the population (larger σ), the greater the required sample size will be. And, of course, the smaller the desired margin of error is, the larger the required sample size will be.

Use of the sample-size formula requires that σ be known, but this is rarely the case in practice. One possible strategy for estimating σ is to carry out a preliminary study and use the resulting sample standard deviation (or a somewhat larger value, to be conservative) to determine n for the main part of the study. Another possibility is simply to make an educated guess about the value of σ and to use that value to calculate n . For a population distribution that is not too skewed, the anticipated range (the difference between the largest and the smallest values) divided by 4 is sometimes used as a rough estimate of the value of the standard deviation.

The sample size required to estimate a population mean μ to within an amount M with 95% confidence is

$$n = \left(\frac{1.96\sigma}{M} \right)^2$$

If σ is unknown, it may be estimated based on previous information or, for a population that is not too skewed, by using (range)/4.

If the desired confidence level is something other than 95%, 1.96 is replaced by the appropriate z critical value (for example, 2.58 for 99% confidence).

Example 9.11 Cost of Textbooks

A college financial aid advisor wants to estimate the mean cost of textbooks per quarter for students at a particular university. For the estimate to be useful, it should be within $M = \$20$ of the actual population mean. How large a sample should be used to be 95% confident while achieving this level of accuracy?

To determine the required sample size, we need a value for σ . Suppose the advisor thinks that the amount spent on books varies widely, with most values between \$150 and \$550. A reasonable estimate of σ is then

$$\frac{\text{range}}{4} = \frac{550 - 150}{4} = \frac{400}{4} = 100$$

The required sample size is

$$n = \left(\frac{1.96\sigma}{M} \right)^2 = \left(\frac{(1.96)(100)}{20} \right)^2 = (9.8)^2 = 96.04$$

Rounding up, a sample size of 97 or larger is recommended.

EXERCISES 9.36 - 9.54

● Data set available online

- 9.36** Given a variable that has a t distribution with the specified degrees of freedom, what percentage of the time will its value fall in the indicated region?

- a. 10 df, between -1.81 and 1.81
- b. 10 df, between -2.23 and 2.23
- c. 24 df, between -2.06 and 2.06
- d. 24 df, between -2.80 and 2.80
- e. 24 df, outside the interval from -2.80 to 2.80
- f. 24 df, to the right of 2.80
- g. 10 df, to the left of -1.81

- 9.37** The formula used to calculate a confidence interval for the mean of a normal population when n is small is

$$\bar{x} \pm (t \text{ critical value}) \frac{s}{\sqrt{n}}$$

What is the appropriate t critical value for each of the following confidence levels and sample sizes?

- a. 95% confidence, $n = 17$
- b. 90% confidence, $n = 12$
- c. 99% confidence, $n = 24$
- d. 90% confidence, $n = 25$
- e. 90% confidence, $n = 13$
- f. 95% confidence, $n = 10$

- 9.38** The two intervals $(114.4, 115.6)$ and $(114.1, 115.9)$ are confidence intervals (calculated using the same sample data) for μ = mean resonance frequency (in hertz) for all tennis rackets of a certain type.

- a. What is the value of the sample mean resonance frequency? (Hint: Where is the confidence interval centered?)

- b.** The confidence level for one of these intervals is 90% and for the other it is 99%. Which is which, and how can you tell?

- 9.39** Samples of two different models of cars were selected, and the actual speed for each car was determined when the speedometer registered 50 mph. The resulting 95% confidence intervals for mean actual speed were $(51.3, 52.7)$ and $(49.4, 50.6)$. Assuming that the two sample standard deviations are equal, which confidence interval is based on the larger sample size? Explain your reasoning.

- 9.40** *USA TODAY* reported that the average amount of money spent on coffee drinks each month is \$78.00 (*USA TODAY Snapshot, November 4, 2016*).

- a. Suppose that this estimate was based on a representative sample of 20 adult Americans. Would you recommend using the one-sample t confidence interval to estimate the population mean amount spent on coffee for the population of all adult Americans? Explain why or why not.
- b. If the sample size had been 200, would you recommend using the one-sample t confidence interval to estimate the population mean amount spent on coffee for the population of all adult Americans? Explain why or why not.

- 9.41** The paper “**The Effects of Adolescent Volunteer Activities on the Perception of Local Society and Community Spirit Mediated by Self-Conception**” (*Advanced Science and Technology Letters [2016]: 19–23*) describes a survey of a large representative sample of middle school children in South Korea.

One question in the survey asked how much time per year the children spent in volunteer activities. The sample mean was 14.76 hours and the sample standard deviation was 16.54 hours.

- Based on the reported sample mean and sample standard deviation, explain why it is not reasonable to think that the distribution of volunteer times for the population of South Korean middle school students is approximately normal.
- The sample size was not given in the paper, but the sample size was described as large. Suppose that the sample size was 500. Explain why it is reasonable to use a one-sample t confidence interval to estimate the population mean even though the population distribution is not approximately normal.
- Calculate and interpret a confidence interval for the mean number of hours spent in volunteer activities per year for South Korean middle school children.

- 9.42** Medical research has shown that repeated wrist extensions beyond 20 degrees increase the risk of wrist and hand injuries. Each of 24 students at Cornell University used a proposed new computer mouse design, and while using the mouse, each student's wrist extension was recorded. Data consistent with summary values given in the paper **"Comparative Study of Two Computer Mouse Designs"** (Cornell Human Factors Laboratory Technical Report RP7992) are given here.

27	28	24	26	27	25	25	24	24	24	25	28
22	25	24	28	27	26	31	25	28	27	27	25

- Use these data to estimate the mean wrist extension for people using this new mouse design using a 90% confidence interval.
- What assumptions are required in order for it to be appropriate to generalize your estimate to the population of Cornell students? To the population of all university students?
- Based on your interval from Part (a), do you think there is reason to believe that the mean wrist extension for people using the new mouse design is greater than 20 degrees? Explain why or why not.

- 9.43** Students in a representative sample of 65 first-year students selected from a large university in England participated in a study of academic procrastination (**"Study Goals and Procrastination Tendencies a Different Stages of the Undergraduate Degree,"** *Studies in Higher Education* [2016]: 2028–2043). Each student in the sample completed the Tuckman Procrastination Scale, which measures procrastination tendencies. Scores on this scale can range from 16 to 64, with

scores over 40 indicating higher levels of procrastination. For the 65 first-year students in this study, the mean score on the procrastination scale was 37.0 and the standard deviation was 6.44.

- Construct a 95% confidence interval estimate of μ , the mean procrastination scale for first-year students at this college.
- Based on your interval, is 40 a plausible value for the population mean score? What does this imply about the population of first-year students?

- 9.44** The paper referenced in the previous exercise also reported that for a representative sample of 68 second-year students at the university, the sample mean procrastination score was 41.0 and the sample standard deviation was 6.82.

- Construct a 95% confidence interval estimate of μ , the population mean procrastination scale for second-year students at this college.
- How does the confidence interval for the population mean score for second-year students compare to the confidence interval for first-year students calculated in the previous exercise? What does this tell you about the difference between first-year and second-year students in terms of mean procrastination score?

- 9.45** Suppose that a random sample of 50 bottles of a particular brand of cough medicine is selected and the alcohol content of each bottle is determined. Let μ denote the mean alcohol content (in percent) for the population of all bottles of the brand under study. Suppose that the sample of 50 results in a 95% confidence interval for μ of (7.8, 9.4).

- Would a 90% confidence interval have been narrower or wider than the given interval? Explain your answer.
- Consider the following statement: There is a 95% chance that μ is between 7.8 and 9.4. Is this statement correct? Why or why not?
- Consider the following statement: If the process of selecting a random sample of size 50 and then computing the corresponding 95% confidence interval is repeated 100 times, 95 of the resulting intervals will include μ . Is this statement correct? Why or why not?

- 9.46** The authors of the paper **"Driving Performance While Using a Mobile Phone: A Simulation Study of Greek Professional Drivers"** (*Transportation Research Part F* [2016]: 164–170) describe a study to evaluate the effect of mobile phone use by taxi drivers in Greece. Fifty taxi drivers drove in a driving simulator where they were following a lead car. The drivers were asked to carry out a conversation while driving, and the following distance (the distance between the taxi and the lead car) was recorded. The sample mean following distance

was 3.2 meters and the sample standard deviation was 1.11 meters.

- Construct and interpret a 95% confidence interval for μ , the population mean following distance while talking on a mobile phone for the population of taxi drivers.
- What assumption must be made to generalize this confidence interval to the population of all taxi drivers in Greece?

- 9.47** The article “**The Association Between Television Viewing and Irregular Sleep Schedules Among Children Less Than 3 Years of Age**” (*Pediatrics* [2005]: 851–856) reported the accompanying 95% confidence intervals for average TV viewing time (in hours per day) for three different age groups.

Age Group	95% Confidence Interval
Less than 12 months	(0.8, 1.0)
12 to 23 months	(1.4, 1.8)
24 to 35 months	(2.1, 2.5)

- Suppose that the sample sizes for each of the three age group samples were equal. Based on the given confidence intervals, which of the age group samples had the greatest variability in TV viewing time? Explain your choice. (Hint: Consider the formula for the confidence interval.)
- Now suppose that the sample standard deviations for the three age group samples were equal, but that the three sample sizes might have been different. Which of the three age-group samples had the largest sample size? Explain your choice.
- The interval (0.768, 1.032) is either a 90% confidence interval or a 99% confidence interval for the mean TV viewing time calculated using the sample data for children less than 12 months old. Is the confidence level for this interval 90% or 99%? Explain your choice.

- 9.48** The paper “**Patterns and Composition of Weight Change in College Freshmen**” (*College Student Journal* [2015]: 553–564) reported that the freshman year weight gain for the students in a representative sample of 103 freshmen at a midwestern university was 5.7 pounds and that the standard deviation of the weight gains was 6.8 pounds. The authors also reported that 75.7% of these students gained more than 1.1 pounds, 17.4% maintained their weight within 1.1 pounds, and 6.8% lost more than 1.1 pounds.
- Based on the reported sample mean and sample standard deviation, explain why it is not reasonable to think that the distribution of weight gains for the population of freshmen students at this university is approximately normal.
 - Explain why it is reasonable to use a one-sample t confidence interval to estimate the population

mean even though the population distribution is not approximately normal.

- Calculate and interpret a confidence interval for the population mean weight gain of freshmen students at this university.

- 9.49** Because of safety considerations, in May 2003 the Federal Aviation Administration (FAA) changed its guidelines for how small commuter airlines must estimate passenger weights. Under the old rule, airlines used 180 pounds as a typical passenger weight (including carry-on luggage) in warm months and 185 pounds as a typical weight in cold months.

The *Alaska Journal of Commerce* (May 25, 2003) reported that Frontier Airlines conducted a study to estimate average passenger plus carry-on weights. They found an average summer weight of 183 pounds and a winter average of 190 pounds. Suppose that each of these estimates was based on a random sample of 100 passengers and that the sample standard deviations were 20 pounds for the summer weights and 23 pounds for the winter weights.

- Construct and interpret a 95% confidence interval for the mean summer weight (including carry-on luggage) of Frontier Airlines passengers.
- Construct and interpret a 95% confidence interval for the mean winter weight (including carry-on luggage) of Frontier Airlines passengers.
- The new FAA recommendations are 190 pounds for summer and 195 pounds for winter. Comment on these recommendations in light of the confidence interval estimates from Parts (a) and (b).

- 9.50** ● Example 9.3 gave the following airborne times (in minutes) for 10 randomly selected flights from San Francisco to Washington Dulles airport:

270 256 267 285 274 275 266 258 271 281

- Calculate and interpret a 90% confidence interval for the mean airborne time for flights from San Francisco to Washington Dulles. (Hint: See Example 9.10.)
- Give an interpretation of the 90% confidence level associated with the interval estimate in Part (a).
- If a flight from San Francisco to Washington Dulles is scheduled to depart at 10 A.M., what would you recommend for the published arrival time? Explain.

- 9.51** ● *Consumer Reports* gave the following mileage ratings (in miles per gallon) for seven midsize hybrid 2016 model cars (consumerreports.org/cro/cars/new-cars/hybrids-evs/ratings-reliability/ratings-overview.htm, retrieved December 21, 2016). Is it reasonable to use these data and the t confidence interval of this section to construct a confidence interval for the mean mileage rating of 2016 midsize hybrid cars? Explain why or why not.

- 9.52** • Five students visiting the student health center for a free dental examination during National Dental Hygiene Month were asked how many months had passed since their last visit to a dentist. Their responses were as follows:

6 17 11 22 29

Assuming that these five students can be considered a random sample of all students participating in the free checkup program, construct a 95% confidence interval for the mean number of months elapsed since the last visit to a dentist for the population of students participating in the program.

- 9.53** The Bureau of Alcohol, Tobacco, and Firearms (BATF) has been concerned about lead levels in California wines. In a previous testing of wine specimens, lead levels ranging from 50 to 700 parts per billion were recorded. How many wine specimens should be tested if the BATF wishes to estimate the mean lead level for California wines to within 10 parts per billion with 95% confidence? (Hint: See Example 9.11.)

- 9.54** The formula described in this section for determining sample size corresponds to a confidence level of 95%. What would be the appropriate formula for determining sample size when the desired confidence level is 90%? 98%?

SECTION 9.4 Interpreting and Communicating the Results of Statistical Analyses

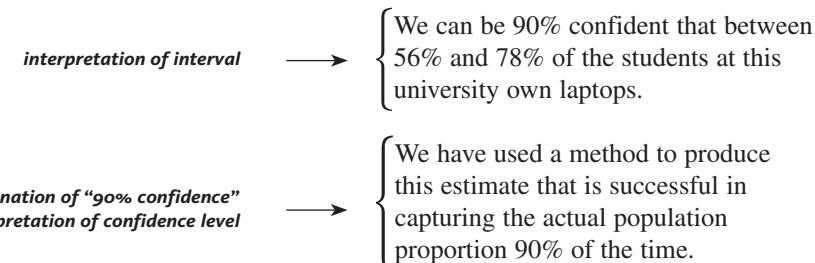
The purpose of most surveys and many research studies is to produce estimates of population characteristics. One way of providing an estimate is to construct and report a confidence interval for the population characteristic of interest.

Communicating the Results of Statistical Analyses

When using sample data to estimate a population characteristic, a point estimate or a confidence interval estimate might be used. Confidence intervals are generally preferred because a point estimate by itself does not convey any information about the precision of the estimate. For this reason, whenever you report the value of a point estimate, it is a good idea to also include an estimate of the margin of error.

Reporting and interpreting a confidence interval estimate requires a bit of care. First, always report both the confidence interval and the confidence level associated with the method used to produce the interval. Then, remember that both the confidence interval and the confidence level should be interpreted.

A good strategy is to begin with an interpretation of the confidence interval in the context of the problem and then to follow that with an interpretation of the confidence level. For example, if a 90% confidence interval for p , the proportion of students at a particular university who own a laptop computer, is (0.56, 0.78), we might say



When providing an interpretation of a confidence interval, remember that the interval is an estimate of a population characteristic and be careful *not* to say that the interval applies to individual values in the population or to the values of sample statistics. For example, if a 99% confidence interval for μ , the mean amount of ketchup in bottles labeled as 12 ounces, is (11.94, 11.98), this does *not* tell us that 99% of 12-ounce ketchup bottles contain between 11.94 and 11.98 ounces of ketchup. Nor does it tell us that 99% of samples of the same size would have sample means in this particular range. The confidence interval is an estimate of the *mean* for *all* bottles in the *population* of interest.

Interpreting the Results of Statistical Analyses

Unfortunately, there is no customary way of reporting the estimates of population characteristics in published sources. Possibilities include

- confidence interval
- estimate \pm margin of error
- estimate \pm standard error

If the population characteristic being estimated is a population mean, then you may also see

$$\text{sample mean} \pm \text{sample standard deviation}$$

If the interval reported is described as a confidence interval, a confidence level should accompany it. These intervals can be interpreted just as we have interpreted the confidence intervals in this chapter. The confidence level specifies the long-run error rate associated with the method used to construct the interval. For example, a 95% confidence level specifies a 5% long-run error rate.

A form particularly common in news articles is estimate \pm margin of error. The margin of error reported is usually two times the standard deviation of the estimate. This method of reporting is a little less formal than a confidence interval and, if the sample size is reasonably large, is roughly equivalent to reporting a 95% confidence interval. You can interpret these intervals as you would a confidence interval with approximate confidence level of 95%.

You must use care in interpreting intervals reported in the form of an estimate \pm standard error. Recall from Section 9.2 that the general form of a confidence interval is

$$\text{estimate} \pm (\text{critical value})(\text{standard deviation of the estimate})$$

In journal articles, the estimated standard deviation of the estimate is usually referred to as the *standard error*. The critical value in the confidence interval formula was determined by the form of the sampling distribution of the estimate and by the confidence level. Notice that the reported form, estimate \pm standard error, is equivalent to a confidence interval with the critical value set equal to 1. For a statistic whose sampling distribution is approximately normal (such as the sample mean for a large sample or the sample proportion for a large sample), a critical value of 1 corresponds to an approximate confidence level of about 68%. Because a confidence level of 68% is rather low, you may want to use the given information and the confidence interval formula to convert to an interval with a higher confidence level.

When researchers are trying to estimate a population mean, they sometimes report sample mean \pm sample standard deviation. Be particularly careful here. To convert this information into a useful interval estimate of the population mean, you must first convert the sample standard deviation to the standard error of the sample mean (by dividing by \sqrt{n}) and then use the standard error and an appropriate critical value to construct a confidence interval.

For example, suppose that a random sample of size 100 is used to estimate the population mean. If the sample resulted in a sample mean of 500 and a sample standard deviation of 20, you might find the published results summarized in any of the following ways:

- 95% confidence interval for the population mean: (496.08, 503.92)
- mean \pm margin of error: 500 \pm 4
- mean \pm standard error: 500 \pm 2
- mean \pm standard deviation: 500 \pm 20

What to Look for in Published Data

Here are some questions to ask when you encounter interval estimates in published reports.

- Is the reported interval a confidence interval, mean \pm margin of error, mean \pm standard error, or mean \pm standard deviation? If the reported interval is not a confidence interval, you may want to construct a confidence interval from the given information.

- What confidence level is associated with the given interval? Is the choice of confidence level reasonable? What does the confidence level say about the long-run error rate of the method used to construct the interval?
- Is the reported interval relatively narrow or relatively wide? Has the population characteristic been estimated precisely?

For example, the article “**Use of a Cast Compared with a Functional Ankle Brace After Operative Treatment of an Ankle Fracture**” (*Journal of Bone and Joint Surgery* [2003]: 205–211) compared two different methods of immobilizing an ankle after surgery to repair damage from a fracture. The article includes the following statement:

The mean duration (and standard deviation) between the operation and return to work was 63 ± 13 days (median, sixty-three days; range, thirty three to ninety-eight days) for the cast group and 65 ± 19 days (median, sixty-two days; range, eight to 131 days) for the brace group; the difference was not significant.

This is an example of a case where we must be careful—the reported intervals are of the form estimate \pm standard deviation. We can use this information to construct a confidence interval for the mean time between surgery and return to work for each method of immobilization. One hundred patients participated in the study, with 50 wearing a cast after surgery and 50 wearing an ankle brace (random assignment was used to assign patients to treatment groups). Because the sample sizes are both large, we can use the t confidence interval formula

$$\text{mean} \pm (t \text{ critical value}) \left(\frac{s}{\sqrt{n}} \right)$$

Each sample has $df = 50 - 1 = 49$. The closest df value in Appendix Table 3 is for $df = 40$, and the corresponding t critical value for a 95% confidence level is 2.02. The corresponding intervals are

$$\text{Cast: } 63 \pm 2.02 \left(\frac{13}{\sqrt{50}} \right) = 63 \pm 3.71 = (59.29, 66.71)$$

$$\text{Brace: } 65 \pm 2.02 \left(\frac{19}{\sqrt{50}} \right) = 65 \pm 5.43 = (59.57, 70.43)$$

The chosen confidence level of 95% implies that the method used to construct each of the intervals has a 5% long-run error rate. Assuming that it is reasonable to view these samples as representative of the patient population, we can interpret these intervals as follows: We can be 95% confident that the mean return-to-work time for those treated with a cast is between 59.29 and 66.71 days. We can be 95% confident that the mean return-to-work time for those treated with an ankle brace is between 59.57 and 70.43 days.

These intervals are relatively wide, indicating that the values of the treatment means have not been estimated as precisely as we might like. This is not surprising, given the sample sizes and the variability in each sample. Note that the two intervals overlap. This supports the statement that the difference between the two immobilization methods was not significant. Formal methods for directly comparing two groups, covered in Chapter 11, could be used to further investigate this issue.

A Word to the Wise: Cautions and Limitations

When working with point and confidence interval estimates, here are a few things you need to keep in mind:

1. In order for an estimate to be useful, we must know something about accuracy. You should beware of point estimates that are not accompanied by a margin of error or some other measure of accuracy.
2. A confidence interval estimate that is wide indicates that we don’t have very precise information about the population characteristics being estimated. Don’t be

fooled by a high confidence level if the resulting interval is wide. High confidence, while desirable, is not the same thing as saying we have precise information about the value of a population characteristic.

The width of a confidence interval is affected by the confidence level, the sample size, and the standard deviation of the statistic used to construct the interval. The best strategy for decreasing the width of a confidence interval is to take a larger sample. It is far better to think about this before collecting data and to use the required sample size formulas to determine a sample size that will result in a confidence interval estimate that is narrow enough to provide useful information.

3. The accuracy of estimates depends on the sample size, not the population size. This may be counter to intuition, but as long as the sample size is small relative to the population size (n less than 10% of the population size), the margin of error for estimating a population proportion with 95% confidence is approximately

$$2\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

and for estimating a population mean with 95% confidence is approximately $2\frac{s}{\sqrt{n}}$.

Notice that each of these involves the sample size n , and both margins of error decrease as the sample size increases. Neither approximate margin of error depends on the population size.

The size of the population does need to be considered if sampling is without replacement and the sample size is more than 10% of the population size. In this case, a **finite population correction factor** $\sqrt{\frac{N-n}{N-1}}$ is used to adjust the margin of error (the given margin is multiplied by the correction factor). Since this correction factor is always less than 1, the adjusted margin of error is smaller.

4. Assumptions and “plausibility” conditions are important. The confidence interval procedures of this chapter require certain assumptions. If these assumptions are met, the confidence intervals provide a method for using sample data to estimate population characteristics with confidence. When the assumptions associated with a confidence interval procedure are in fact true, the confidence level specifies a correct success rate for the method. However, assumptions (such as the assumption of a normal population distribution) are rarely exactly met in practice. Fortunately, in most cases, as long as the assumptions are approximately met, the confidence interval procedures still work well.

In general, we can only determine if assumptions are “plausible” or approximately met, and that we are in the situation where we expect the procedure to work reasonably well. This is usually confirmed by knowledge of the data collection process and by using the sample data to check certain “plausibility conditions.”

The formal assumptions for the large-sample z confidence interval for a population proportion are

- a. The sample is a random sample from the population of interest.
- b. The sample size is large enough for the sampling distribution of \hat{p} to be approximately normal.
- c. Sampling is without replacement.

Whether the random sample assumption is plausible will depend on how the sample was selected and the intended population. Plausibility conditions for the other two assumptions are the following:

$n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$ (so the sampling distribution of \hat{p} is approximately normal), and

n is less than 10% of the population size (so that the formula for the standard deviation of \hat{p} provides a good approximation to the actual standard deviation).

The formal assumptions for the t confidence interval for a population mean are

- The sample is a random sample from the population of interest.
- The population distribution is normal, so that the distribution of $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ has a t distribution.

As was the case for proportions, the plausibility of the random sample assumption will depend on how the sample was selected and the population of interest. The plausibility conditions for the normal population distribution assumption are the following:

A normal probability plot of the data is reasonably straight (indicating that the population distribution is approximately normal)

or

The data distribution is approximately symmetric and there are no outliers. This may be confirmed by looking at a dotplot, boxplot, stem-and-leaf display, or histogram of the data.

Alternatively, if n is large ($n \geq 30$), the sampling distribution of \bar{x} will be approximately normal even for nonnormal population distributions. This implies that use of the t interval is appropriate even if population normality is not plausible.

In the end, you must decide that the assumptions are met or that they are “plausible” and that the inferential method used will provide reasonable results. This is also true for the inferential methods introduced in the chapters that follow.

- Watch out for the “ \pm ” when reading published reports. Don’t fall into the trap of thinking confidence interval every time you see a \pm in an expression. As was discussed earlier in this section, published reports are not consistent, and in addition to confidence intervals, it is common to see estimate \pm standard error and estimate \pm sample standard deviation reported.

EXERCISES 9.55 - 9.57

- 9.55** The following quote is from the article [“Credit Card Debt Rises Faster for Seniors” \(USA TODAY, July 28, 2009\)](#):

The study, which will be released today by Demos, a liberal public policy group, shows that low- and middle-income consumers 65 and older carried \$10,235 in average credit card debt last year.

What additional information would you want in order to evaluate the accuracy of this estimate? Explain.

- 9.56** Authors of the news release titled [“Major Gaps Still Exist Between the Perception and the Reality of Americans’ Internet Security Protections, Study Finds”](#) (The National Cyber Security Alliance) estimated the proportion of Americans who claim to have a firewall installed on their computer to protect them from computer hackers to be 0.80 based on a survey conducted by the Zogby market research firm. They also estimated the proportion of those who actually have a firewall installed to be 0.42, based on

checkups performed by Norton’s PC Help software. The following quote is from the news release:

For the study, NCSA commissioned a Zogby survey of more than 3000 Americans and Symantec conducted checkups of 400 Americans’ personal computers performed by PC Help by Norton (www.norton.com/tuneup). The Zogby poll has a margin of error of $+/- 1.6\%$ and the checkup has a margin of error of $+/- 5\%$.

Explain why the margins of error for the two estimated proportions are different.

- 9.57** The paper [“The Curious Promiscuity of Queen Honey Bees \(*Apis mellifera*\): Evolutionary and Behavioral Mechanisms”](#) (*Annals of Zoology Fennici* [2001]: 255–265) describes a study of the mating behavior of queen honeybees. The following quote is from the paper:

Queens flew for an average of 24.2 ± 9.21 minutes on their mating flights, which is consistent with previous findings. On those flights, queens effectively mated with 4.6 ± 3.47 males (mean \pm SD).

- a. The intervals reported in the quote from the paper were based on data from the mating flights of $n = 30$ queen honeybees. One of the two intervals reported is stated to be a confidence interval for a population mean. Which interval is this? Justify your choice.
- b. Use the given information to construct a 95% confidence interval for the mean number of partners on a mating flight for queen honeybees. For purposes of this exercise, assume that it is reasonable to consider these 30 queen honeybees as representative of the population of queen honeybees.

SECTION 9.5 Bootstrap Confidence Intervals for a Population Proportion (Optional)

In Section 9.2, data from a large random sample were used to construct a confidence interval for a population proportion, p . The large-sample confidence interval has the form

$$\hat{p} \pm (z \text{ critical value}) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

To use this confidence interval formula, we need to know that the sample is a random sample from a population (or is selected in a way that makes it reasonable to think that the sample is representative of a population). In addition, we need to know that the sampling distribution of \hat{p} is approximately normal. This is reasonable when the sample size is large ($n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$) but isn't necessarily the case when the sample size is small. This section introduces an alternative method that can be used to obtain a confidence interval for a population proportion but that doesn't require a large sample size.

A Bootstrap Confidence Interval for a Population Proportion

When a sample proportion \hat{p} is used to estimate a population proportion, we know that the value of \hat{p} is not likely to be exactly equal to the value of the population proportion. But if the sample is selected in a reasonable way, we expect that the value of \hat{p} will be somewhere around the value of the population proportion. A confidence interval quantifies what is meant by “around the value of the population proportion.” When the assumptions of the large-sample confidence interval are reasonable, the confidence interval based on a sampling distribution of \hat{p} that is approximately normal can be used to determine the largest value that the difference between the observed sample proportion and the actual value of the population proportion is likely to be for a specific confidence level.

However, sometimes the distribution of \hat{p} is not approximately normal. In this case, constructing a confidence interval still requires knowing something about how far away the sample proportion is likely to be from the value of the population proportion. For example, suppose that the sample proportion is $\hat{p} = 0.32$ and that it is unlikely to be smaller than the actual value of the population proportion by more than 0.04. Then we think that the population proportion is less than 0.36 (from $0.32 + 0.04$). If it is also unlikely that the sample proportion is greater than the population proportion by more than 0.05, then we could say that we think that the population proportion is greater than 0.27 (from $0.32 - 0.05$). A reasonable interval estimate of the population proportion based on the sample would then be $(0.27, 0.36)$. **Bootstrapping** is a way to determine what number to add and what number to subtract from the sample proportion to form a confidence interval.

Bootstrapping works by considering how far sample proportions from random samples of size n tend to be from the value of the population proportion if the value of the population proportion were known. Taking many different random samples of size n from such a population and calculating the sample proportion for each one creates a distribution of these sample proportions. Looking at how the sample proportions cluster around the value of the population proportion provides information about variability. But of course, in practice we can't do this because the value of the population proportion is unknown.

Instead, we consider a hypothetical population that we expect to be very similar to the population that the sample is actually from. Examining sample proportions from this hypothetical population provides information about variability in \hat{p} values. To do this, bootstrapping uses the observed sample proportion as the proportion for the hypothetical population.

To create a bootstrap confidence interval, many random samples of size n are taken from this hypothetical population to form a **bootstrap distribution**. The variability in this distribution indicates how far \hat{p} values for samples from this hypothetical population might be from the original observed value of \hat{p} . Knowing how far these simulated \hat{p} values tend to fall from the observed value of \hat{p} indicates how far the observed value of \hat{p} is likely to be from the value of p in the actual population.

For a 95% confidence level, using the boundaries that capture the middle 95% of the simulated bootstrap distribution is equivalent to determining the endpoints of a confidence interval, which is represented as \hat{p} minus a number and \hat{p} plus a number. For bootstrap confidence intervals, the number subtracted from \hat{p} and the number added to \hat{p} won't always be equal since the bootstrap distribution may not be symmetric. This interval is called a **bootstrap confidence interval** estimate for p , and it is based on simulation rather than on knowing that the sampling distribution of \hat{p} is approximately normal.

Taking a random sample from a hypothetical population that has a proportion of successes equal to the original sample proportion is equivalent to selecting a random sample with replacement from the original sample. The process of drawing new samples from an original representative sample is called **resampling**, because new samples are taken directly from the original sample that was collected. This method is also called "bootstrapping" because it is like "pulling yourself up by the bootstraps," in the sense that nothing more than the original data collected in one sample is used to generate information about sample-to-sample variability in the sample proportion, \hat{p} .

Example 9.12 Generating a Bootstrap Distribution

Suppose that a random sample of 10 students at a school is selected, and each student is asked if they spend more than 4 hours a day online. The sample is used to estimate the proportion of students at the school who would respond "Yes" to this question. If the responses for the people in the sample were those shown below, the value of the sample proportion of "Yes" responses would be $\hat{p} = \frac{5}{10} = 0.5$.

Student Number	1	2	3	4	5	6	7	8	9	10
Response	No	Yes	Yes	No	No	Yes	No	No	Yes	Yes

To create a bootstrap distribution, consider a hypothetical population in which the proportion of successes is 0.5, and then take samples of size 10 from this hypothetical population. One way to do this is to select a random sample of size 10 with replacement from the original sample using random numbers. Ten random numbers, selected with replacement from the numbers from 1 to 10, are shown below. Also shown are the responses from the original sample associated with each of these random numbers.

Random Number	10	9	3	8	3	2	8	3	6	9
Response	Yes	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes

The proportion of "Yes" responses in this simulated sample is $\hat{p} = \frac{8}{10} = 0.8$, even though the proportion in the original sample was 0.5.

To form a bootstrap distribution, we resample from the original sample many times. These simulated samples provide information about the sample-to-sample variability in the sample proportions, \hat{p} , and the bootstrap distribution can then be used to determine an interval of plausible values for p associated with a specified confidence level.

Bootstrap Confidence Intervals for One Proportion

This section explains how to generate a bootstrap distribution and construct a bootstrap confidence interval for a population proportion using one of the Shiny web apps that accompany this text. These web apps are located at statistics.cengage.com/PSO6e/Apps.html. There are also a number of other resources for constructing bootstrap confidence intervals (for example, see the StatKey apps at lock5stat.com/StatKey/).

Example 9.13 Generating a Bootstrap Confidence Interval for a Population Proportion (Example 9.12 continued)

Recall that in the previous example, a random sample of 10 students produced a sample proportion who spend more than 4 hours a day online of $\hat{p} = 0.50$. To use the Shiny app to construct a bootstrap confidence interval for the proportion of students at the school who spend more than four hours a day online, follow the instructions below.

Enter the number of observations and the number of successes into the Shiny app titled “Bootstrap Confidence Interval for One Proportion.” In this example, the number of observations, also known as the sample size, is $n = 10$, and five successes were observed in the original sample. The confidence level in this example will be 95%, so this value should be entered as well, or recognize that 95% is the default.

Bootstrap Confidence Interval for One Proportion

Select number of observations: 10

Select the number of successes: 5

Select confidence level (in %): 95

Select the number of simulated samples to generate:

1 10 100 1,000 10,000

Generate Simulated Sample(s) Reset

Select the number of simulated bootstrap samples to be generated. In this example, choose 1000 simulated samples. Then, click “Generate Simulated Sample(s).”

Bootstrap Confidence Interval for One Proportion

Select number of observations: 10

Select the number of successes: 5

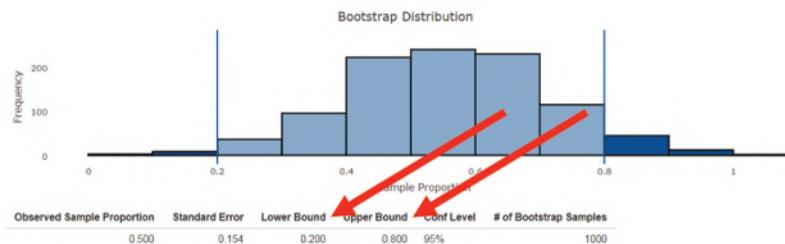
Select confidence level (in %): 95

Select the number of simulated samples to generate:

1 10 100 1,000 10,000

Generate Simulated Sample(s) Reset

The Shiny app identifies the “Lower Bound” to be the value of the sample proportion that has 2.5% of the simulated proportions below and the “Upper Bound” to be the value that has 2.5% of the simulated proportions above:



Different sets of simulated samples may produce slightly different results. Based on this output from the Shiny app, 2.5% of the simulated sample proportions fall at or below 0.2, and 2.5% of the simulated sample proportions fall at or above 0.8. A 95% bootstrap confidence interval for the proportion of students at the school who would respond that they spend more than 4 hours a day online is (0.2, 0.8). We can be 95% confident that the actual proportion of “Yes” responses falls between 0.2 and 0.8. Notice that this interval is very wide, with plausible values for the population proportion ranging from 0.2 all the way to 0.8. This is a function of the very small sample size. It is difficult to estimate population proportions accurately with a small sample, no matter what method is used!

Because a bootstrap confidence interval is based on a distribution of simulated proportions, the interval that is generated may vary from one simulation to another. But if the confidence interval is based on many simulated proportions, different bootstrap confidence intervals based on the same sample won’t differ substantially from one another.

Now let’s look at a more realistic example.

Example 9.14 College Education Revisited

Recall that in Example 9.4 we used the large-sample confidence interval formula to find a 95% confidence interval for p , the population proportion of adult Americans who believe that a college education is essential for success. The observed value of the sample proportion was $\hat{p} = \frac{567}{1031} = 0.55$, and the resulting confidence interval was (0.521, 0.579).

We can use the bootstrap method as an alternative way to find a 95% confidence interval for p .

In the Shiny app “Bootstrap Confidence Interval for One Proportion,” enter the sample size and the number of successes. In Example 9.4, the sample size is $n = 1031$, and 567 successes were observed in the original sample. Select a 95% confidence level and then choose 1000 simulated samples. Click on “Generate Simulated Sample(s).”

Bootstrap Confidence Interval for One Proportion

Select number of observations: 1031

Select the number of successes: 567

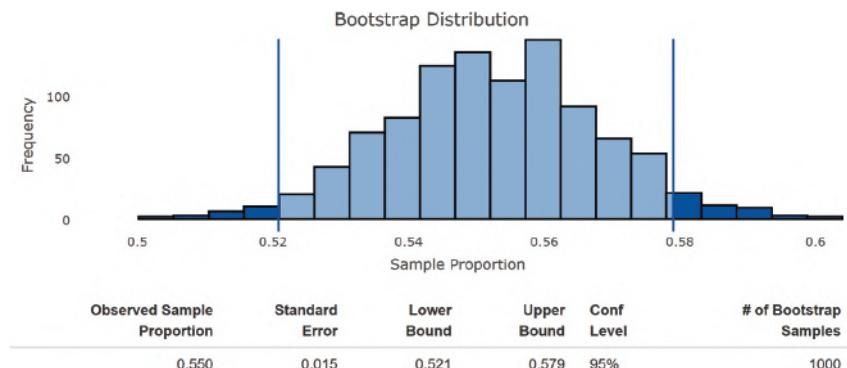
Select confidence level (in %): 95

Select the number of simulated samples to generate:

- 1
- 10
- 100
- 1,000
- 10,000

Generate Simulated Sample(s) Reset

For the simulation shown here, the Shiny app identifies the value that has 2.5% of the simulated proportions below and the value that has 2.5% of the simulated proportions above:



The Shiny app reports that 2.5% of the simulated sample proportions fall at or below 0.521, and that 2.5% of the simulated sample proportions fall at or above 0.579. A bootstrap confidence interval for the actual proportion of adult Americans who believe a college education is essential for success is (0.521, 0.579). Assuming that the sample is representative of the population, we can be 95% confident that the actual proportion of adult Americans who believe a college education is essential for success falls between 0.521 and 0.579 (between 52.1% and 57.9%).

Notice that this bootstrap confidence interval is the same as the interval provided by the large-sample method in Section 9.2, which is not surprising since the conditions for the large-sample confidence interval are satisfied.

What Happens When the Conditions for the Large-Sample Interval Are Not Met?

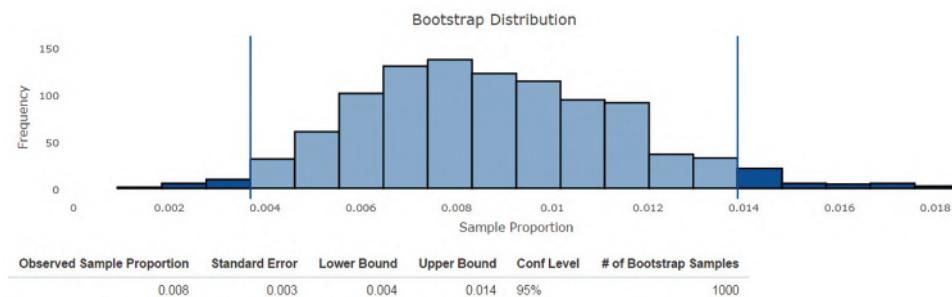
In the previous example, the bootstrap confidence interval turned out to be the same as the large-sample confidence interval. It is not surprising to find that the bootstrap confidence interval and the large-sample confidence interval are similar when the sample size is large. In this case, both methods provide appropriate ways to obtain a confidence interval. But the bootstrap method can still be used even if the sample size is not large enough to satisfy the conditions for the large-sample confidence interval. This is illustrated in the following example.

Example 9.15 Liver Injuries in Newborns

The article “[Severe Liver Injury While Using Umbilical Venous Catheter: Case Series and Literature Review](#)” (*American Journal of Perinatology* [2014]: 965–974) describes a study of newborns who were placed in intensive care and required insertion of an umbilical vein catheter so that fluids could be administered. Researchers found that 9 out of the 1081 newborns studied developed catheter-associated liver injury. The authors were interested in estimating the proportion of newborns who suffer liver injury as a result of the use of umbilical vein catheters.

The researchers considered this sample of 1081 infants to be representative of the population of newborns who required use of the catheter. The sample proportion with liver injury is $\hat{p} = \frac{9}{1081} = 0.00833$. Notice that although the sample size is 1081, because $n\hat{p} = (1081)(0.00833) = 9$, the sample size is not large enough to justify the use of the large-sample confidence interval for a population proportion.

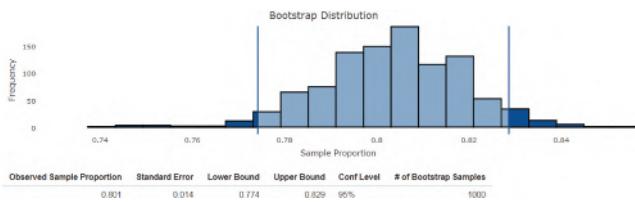
The Shiny app “Bootstrap Confidence Interval for One Proportion” was used to generate a bootstrap distribution based on 1000 simulated sample proportions.



A 95% bootstrap confidence interval for the population proportion, p , is $(0.004, 0.014)$. Based on this sample, we can be 95% confident that the actual proportion of newborns with umbilical catheters who suffer liver injury is somewhere between 0.004 and 0.014. Because the conditions for the large-sample interval are not met, the bootstrap interval is a more appropriate choice.

EXERCISES 9.58 - 9.65

- 9.58** A survey on SodaHead (sodahead.com/survey/featured/anonymous-advice/?results=1, retrieved May 13, 2016) reported that 603 out of 753 respondents replied “No” to the question “Should You Be Friends with Your Boss on Facebook?”
- Use the accompanying output from the “Bootstrap Confidence Interval for One Proportion” Shiny app to report a 95% bootstrap confidence interval for the population proportion who would reply “No” to the question. Interpret the confidence interval in context.



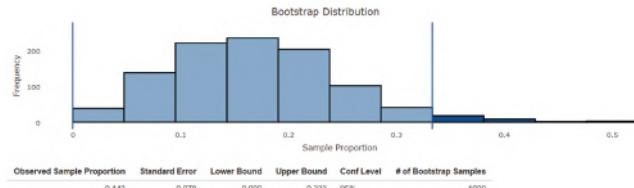
- SodaHead provides summaries for anonymous and voluntary responses to survey questions. Does the proportion of respondents who reply “No” to the question in an anonymous and voluntary situation tend to underestimate or overestimate the actual population proportion of interest? Explain.

- 9.59** The article “Report: More Than Half of DC-Area Millennials Are Using Ride-Hailing Apps” (washingtonian.com/2016/06/23/report-half-dc-area-millennials-using-ride-hailing-apps/, June 23, 2016, retrieved May 4, 2017) refers to a study summarized at the following site: wbaresearch.com/wp-content/uploads/2016/06/Transportation-MarkeTrak-Spring-20161.pdf (retrieved May 4, 2017). The study indicates that 21% of Washington-area

adults who are 55 years old and older have used transportation apps such as Uber or Lyft at least once.

Suppose that a small local transportation service for older residents is monitoring usage of app-based transportation. The service conducted a survey of a random sample of 21 of its regular customers who are 55 or older and found that 3 of them had tried Uber or Lyft at least once.

- Would it be appropriate to use the large-sample confidence interval for a population proportion to estimate the proportion of the transportation service customers who have tried Uber or Lyft at least once? Explain.
- Would it be appropriate to use a bootstrap confidence interval for a population proportion to estimate the proportion of the transportation service customers who have tried Uber or Lyft at least once? Explain.
- Use the accompanying output from the Shiny app to report a 95% bootstrap confidence interval for the population proportion of customers 55 or older who have used Uber or Lyft at least once. Interpret the confidence interval in context.



- Is the value obtained in the study for Washington-area adults who are over 55 years old who have used Uber or Lyft at least once

(21%) in the bootstrap confidence interval? What does this indicate?

- 9.60** An article titled “[The Latest on Workplace Monitoring and Surveillance](#)” ([American Management Association, November 17, 2014](#)) referred to the “2007 Electronic Monitoring & Surveillance Survey.” In a summary of survey results submitted by 304 U.S. businesses, 85 of these businesses had fired workers for e-mail misuse.

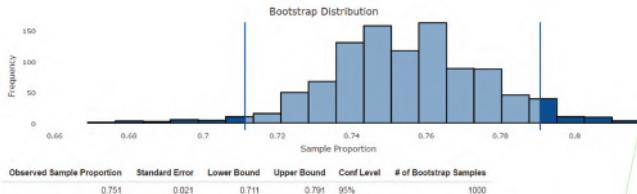
Suppose that it is reasonable to regard these 304 businesses as a representative sample of businesses in the United States. Use the “Bootstrap Confidence Interval for One Proportion” Shiny app to generate a bootstrap confidence interval for the proportion of U.S. businesses who have fired workers for e-mail misuse. Interpret the interval in context.

- 9.61** During the 2016 NBA Finals, Kevin Love of the Cleveland Cavaliers successfully made 5 three-point shots out of 19 attempts. Assume that these attempts comprise a sample that is representative of his ability during the entire 2016 season.

- Explain why it would not be appropriate to use a large-sample confidence interval for one proportion to estimate Kevin Love’s success rate for three-point shots during the 2016 season.
- Use the “Bootstrap Confidence Interval for One Proportion” Shiny app to generate a 90% bootstrap confidence interval for Kevin Love’s three-point shot success rate during the 2016 NBA season. Interpret the interval in context.

- 9.62** A survey of a representative sample of 478 U.S. employers determined that 359 ranked stress as their top health and productivity concern ([globenewswire.com/news-release/2016/06/29/852338/0/en/Seventy-five-percent-of-U-S-employers-say-stress-is-their-number-one-workplace-health-concern.html?print=1](#), June 29, 2016, retrieved May 4, 2017).

- Use the accompanying output from the “Bootstrap Confidence Interval for One Proportion” Shiny app to report a 95% bootstrap confidence interval for the proportion of all U.S. employers who would rank stress as their top health and productivity concern. Interpret the confidence interval in context.



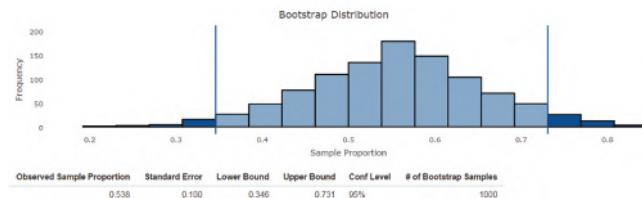
- A number of international employers were also surveyed. If the international employers had a similar rate of identifying stress as their

top health and productivity concern, and if the results from international employers were included in the sample, would the width of the resulting confidence interval remain the same, decrease, or increase? Explain.

- 9.63** In mid-2016 the United Kingdom (UK) withdrew from the European Union (an event known as “Brexit”), causing economic concerns throughout the world. One indicator that economists use to monitor the health of the economy is the proportion of residential properties offered for sale at auction that are successfully sold.

An article titled “[Going, Going, Gone through the Roof—Sky’s the Limit at Auction](#)” ([estateagenttoday.co.uk/features/2016/10/going-going-gone-through-the-roof--the-skys-the-limit-at-auction](#), October 22, 2016, retrieved May 4, 2017) reported the success rate of a sample of 26 residential properties offered for sale at auctions in the UK in the summer of 2016. For this sample of properties, 14 of the 26 residential properties were successfully sold. Suppose it is reasonable to consider these 26 properties as representative of residential properties offered at auction in the post-Brexit UK.

- Would it be appropriate to use the large-sample confidence interval for a population proportion to estimate the proportion of residential properties successfully sold at auction in the post-Brexit UK? Explain.
- Would it be appropriate to use a bootstrap confidence interval for a population proportion to estimate the proportion of residential properties successfully sold at auction in the post-Brexit UK? Explain.
- Use the accompanying output from the “Bootstrap Confidence Interval for One Proportion” Shiny app to report a 95% bootstrap confidence interval for the population proportion of residential properties successfully sold at auction in the post-Brexit UK. Interpret the confidence interval in context.



- The success rate for properties sold at auction throughout the UK during one stretch a year earlier—in 2015—was 72%. Does this value fall within the bootstrap confidence interval reported in Part (c)? What does this indicate?

- 9.64** The report “[One in Three American Households Are Stuck in a Relationship with a Financial Services](#)

Provider They Don't Trust" (businesswire.com/news/home/20160629005198/en/American-Households-Stuck-Relationship-Financial-Services-Provider, June 29, 2016, retrieved May 4, 2017) estimated that 31% of American households feel obliged to do business with one or more financial services companies they distrust. This estimate is based on a representative sample of 1056 consumers age 18 and older.

Use the "Bootstrap Confidence Interval for One Proportion" Shiny app to generate a 95% bootstrap confidence interval for the proportion of all U.S. households that feel obliged to do business with one or more financial services companies they distrust. Interpret the interval in context.

9.65 A 2016 study of 120 U.S. brand-name products found that 70% were active on Snapchat (businessinsider.com/what-exactly-are-brands-posting-on-snapchat-2016-6, June 15, 2016, retrieved May 4, 2017). The researchers conducting the study used bootstrap methods to determine a confidence interval.

Suppose that it is reasonable to consider this sample of brand-name products as representative of all brand-name products. Use the "Bootstrap Confidence Interval for One Proportion" Shiny app to find a 95% confidence interval for the proportion of all brand-name products that are active on Snapchat and interpret the interval in context.

SECTION 9.6 Bootstrap Confidence Intervals for a Population Mean (Optional)

The sample mean \bar{x} is an estimate of the population mean μ , but we don't expect that the value of \bar{x} will be exactly equal to the population mean. However, for representative samples, we expect that the value of \bar{x} will be near the value of μ . The notion of being near the value of a population mean can be represented by a confidence interval.

In Section 9.3, data from a sample were used to construct a confidence interval for a population mean, μ . The confidence interval introduced there has the form

$$\bar{x} \pm (t \text{ critical value}) \frac{s}{\sqrt{n}}$$

To use this confidence interval, we need to know that the population distribution is approximately normal or the sample size is large. If you are not certain about whether the assumptions needed for a large-sample confidence interval for μ are reasonable, a simulation-based approach called **bootstrapping** can be used.

Consider a hypothetical population that is similar to the population that the observed sample is actually from. By examining the distribution of sample means for samples taken from this hypothetical population, it is possible to use bootstrapping to find a confidence interval for the population mean. Bootstrapping uses random samples (called bootstrap samples) from a hypothetical population represented by the data in the observed sample, which is thought to be representative of the population. The distribution of sample means from the bootstrap samples is called a **bootstrap distribution**. The variability in this distribution indicates how far \bar{x} values for samples from this hypothetical population might be from the original observed value of \bar{x} . Knowing how far simulated \bar{x} values tend to fall from the observed value of \bar{x} provides information about how far the observed value of \bar{x} is likely to be from the value of μ in the actual population.

For a 95% confidence level, using the boundaries that capture the middle 95% of the bootstrap distribution is equivalent to getting a value for \bar{x} minus a number, and for \bar{x} plus a number (although the numbers subtracted and added won't always be the same since the bootstrap distribution may not be symmetric). This interval is a **bootstrap confidence interval** estimate for μ , and it is based on simulation rather than on knowing that the sampling distribution \bar{x} is at least approximately normal.

Example 9.16 Selfish Chimps? (Revisited)

Recall the data presented in Example 9.10 of Section 9.3, regarding the number of times out of 36 trials that each of $n = 7$ chimps chose the option to provide food to both chimps (the "charitable" response).

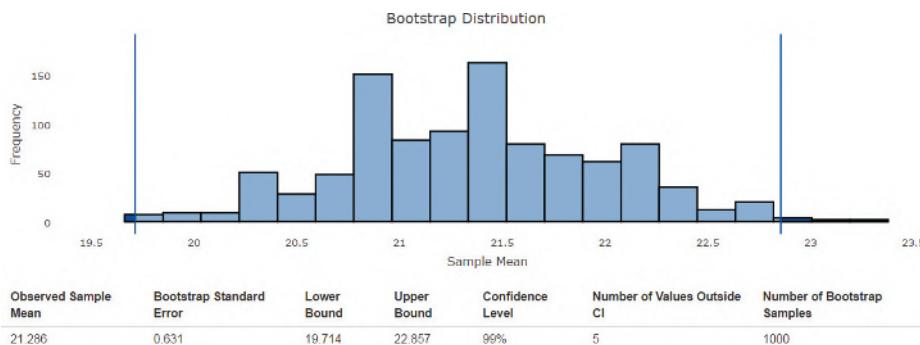
The original data values are displayed below. The sample mean of this original sample is $\bar{x} = 21.29$.

Chimp ID	1	2	3	4	5	6	7
Trials	23	22	21	24	19	20	20

We begin by resampling from this original sample. That is, we select bootstrap samples by selecting at random with replacement from the original sample. Here is one bootstrap sample from the original sample:

Resampled Chimp ID	3	2	1	4	3	5	1
Trials	21	22	23	24	21	19	23

The sample mean for this bootstrap sample is $\bar{x} = 21.86$. We can repeat this process many times, and the resulting bootstrap distribution of \bar{x} values provides information about sampling variability that can be used to find confidence intervals. Here is the output from the Shiny app titled “Bootstrap Confidence Interval for One Mean,” used to obtain a 99% confidence interval for μ .



The bootstrap approach produces a simulation-based bootstrap confidence interval of $(19.71, 22.86)$ for the population mean, μ . We can be 99% confident that this interval contains the population mean. This interval represents a set of plausible values for the population mean, μ .

Notice that the 99% one-sample t confidence interval reported in Example 9.10 is $(18.76, 23.81)$ trials. The bootstrap confidence interval is slightly narrower than the one-sample t confidence interval.

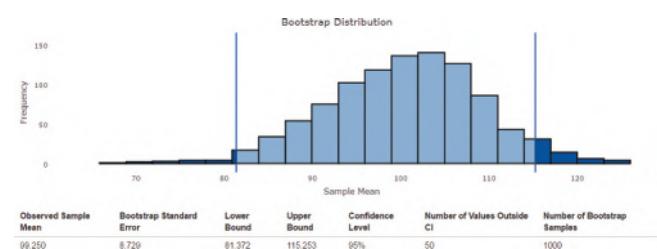
EXERCISES 9.66 - 9.71

● Data set available online

- 9.66** ● *Consumer Reports* published the following gas mileage values (“Overall MPG”) for a sample of electric or plug-in hybrid car models (consumerreports.org/cro/cars/new-cars/hybrids-evs/ratings-reliability/ratings-overview.htm, retrieved December 23, 2016).

Make and model	Overall MPG
Tesla Model S	87
BMW i3	139
Ford C-MAX	47
Nissan Leaf	106
Tesla Model X	92
Chevrolet Volt	105
Ford Focus	107
Mitsubishi i-MiEV	111

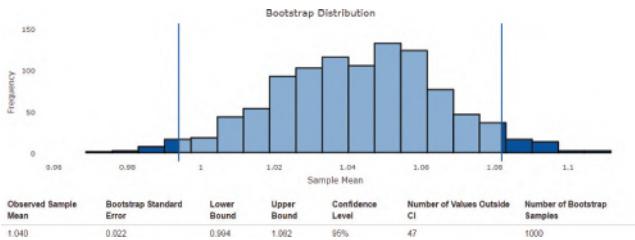
Use the accompanying output from the Shiny app to construct and interpret a 95% confidence interval for the population mean gas mileage for electric or plug-in hybrid cars. You may assume that this sample is representative of the population of electric and plug-in hybrid cars.



- 9.67** ● The authors of the paper “**Sex Differences in Time Perception During Smoking Abstinence**” (*Nicotine and Tobacco Research* [2015]: 449–454) investigated how nicotine withdrawal affects time perception and decision making. In this study, 21 male smokers were asked to abstain from smoking for 24 hours. After 24 hours, they were shown a demo screen with a green cross that changed to a red cross after a period of time. They were then shown a green cross and asked to indicate when they thought the same amount of time had passed as in the demo. This process was repeated 15 more times with varying times. A time discrimination score was calculated for each man by dividing the total of the estimated times by the total of the actual times. Suppose that the 21 time discrimination scores were the ones shown below. These values are consistent with summary values given in the paper.

1.12 1.03 1.09 1.03 1.09 0.97 0.98 1.20 1.16 1.03 1.10
1.11 0.98 1.02 1.20 0.96 0.78 1.05 0.90 1.08 0.95

- a. What characteristic of the sample size indicates that the methods based on the t distribution may not be appropriate?
- b. Use the given Shiny app output to estimate the mean time discrimination score for the population of male smokers who abstain from smoking for 24 hours using a 95% confidence interval.



- 9.68** ● Teams in the National Football League (NFL) are given a “bye” during one week of the season, when they can rest and not play a game. This may provide an advantage for the team in the next game they play after a bye.

In 2016, each of the 32 NFL teams was granted a bye during one of the weeks of the season. The following table contains the team name and the number of points they won by or lost by in the game after the bye (espn.com/nfl/, retrieved December 22, 2016). A positive value indicates that the team coming off the bye won the game, and a negative value means that they lost. You may consider these results to be a representative sample from a population of possible NFL matchups between teams where one of the teams is coming off a bye week.

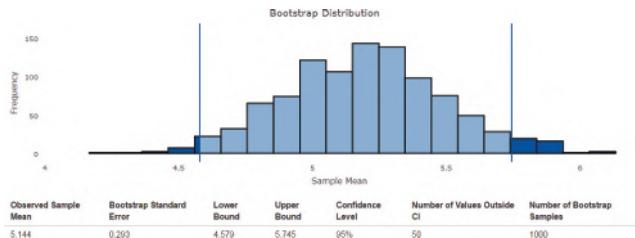
Team	Points	Team	Points
ARI	3	LA	-3
ATL	19	MIA	4
BAL	7	MIN	-11
BUF	4	NE	-7
CAR	10	NO	3
CHI	-26	NYG	5
CIN	-1	NYJ	-5
CLE	-13	OAK	7
DAL	6	PHI	-1
DEN	-3	PIT	-7
DET	7	SD	8
GB	7	SEA	2
HOU	3	SF	-18
IND	7	TB	17
JAX	1	TEN	3
KC	16	WSH	6

- a. Construct a graphical display for the data. Although the sample size is at least on the borderline of being adequate for t distribution methods, what characteristic of the distribution indicates that the methods based on the t distribution may not be appropriate?
- b. Use a bootstrap confidence interval to estimate the population mean point difference for NFL teams coming off a bye week using a 95% confidence interval.
- c. Use the results from Part (b) to explain whether or not you believe that teams coming off a bye week have a significant advantage in points scored over their opponents.

- 9.69** ● *Consumer Reports* provides ratings for televisions, including energy cost per year (consumerreports.org/products/lcd-led-oled-tvs/ratings-overview/, retrieved December 23, 2016). Energy cost data for a sample of 13 small televisions (29-inch and smaller) are displayed in the following table:

Make and Model	Energy Cost (dollars)
Samsung UN28H4000	7.17
LG 28LF4520	5.17
LG 24LF4520	4.47
LG 28LH4530	5.00
Vizio D28h-C1	5.06
Vizio D28h-D1	7.00
LG 22LH4530	4.00
Element ELEFW248	4.00
Vizio D24-D1	6.00
LG 24LH4530	4.00
Insignia NS-24ER310NA17	6.00
Seiki SE23HEB2	4.00
Insignia NS-24D510NA17	5.00

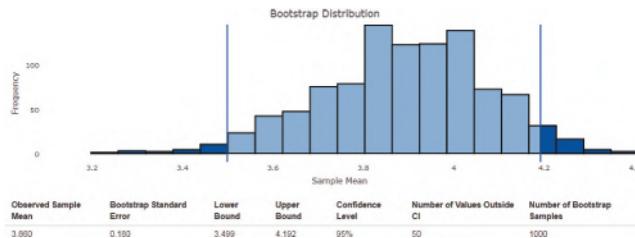
Suppose that this sample is representative of the population of all small televisions. Use the accompanying output from the Shiny app to construct and interpret a 95% bootstrap confidence interval for the mean annual energy cost for the population of all small televisions.



- 9.70** The *Economist* collects data each year on the price of a Big Mac in various countries around the world. A sample of McDonald's restaurants in Europe in 2016 resulted in the following Big Mac prices (after conversion to U.S. dollars):

4.44 3.15 2.42 3.96 4.35 4.51
4.17 3.69 4.62 3.80 3.36 3.85

- What characteristic of the sample indicates that the methods based on the t distribution may not be appropriate?
- Use the Shiny app given output to estimate the mean price of a Big Mac in Europe using a 95% bootstrap confidence interval.



- 9.71** Major League Baseball (MLB) includes two groups of teams, in “leagues.” There are 15 teams in each of the American League (AL) and the National League (NL). Since 1997, teams in each of the

leagues play teams from the other league in “interleague” regular-season games.

One way to determine whether one league is stronger than the other is to consider the interleague winning percentages for all the teams in one of the two leagues, say, the National League, for one season. For purposes of this exercise, consider the interleague games played in the 2016 season to be a representative sample of the performance of the teams in a population of potential future seasons.

Here are the 2016 interleague winning percentages for the 15 NL teams:

Team	W	L	Winning Percentage
ARI	5	15	25%
ATL	8	12	40%
CHC	15	5	75%
CIN	5	15	25%
COL	9	11	45%
LAD	10	10	50%
MIA	6	14	30%
MIL	11	9	55%
NYM	12	8	60%
PHI	11	9	55%
PIT	9	11	45%
SD	6	14	30%
SF	8	12	40%
STL	8	12	40%
WSH	12	8	60%

- What characteristic of this sample of NL team interleague winning percentages indicates that the methods based on the t distribution may not be appropriate?
- Estimate the mean interleague winning percentage for the population of NL teams using a 95% bootstrap confidence interval.
- Use the results from Part (b) to explain whether it is reasonable to say that the National League or the American League performs significantly better than the other in interleague play.

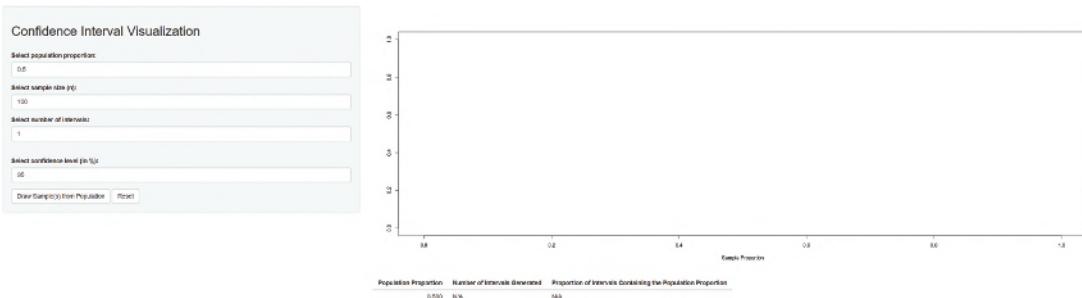
CHAPTER ACTIVITIES

ACTIVITY 9.1 GETTING A FEEL FOR CONFIDENCE LEVEL

Technology Activity

This activity uses the Shiny app titled CI Proportion Visualization that is part of the collection of Shiny web apps that accompany this text. This app can be found at statistics.cengage.com/PSO6e/Apps.html.

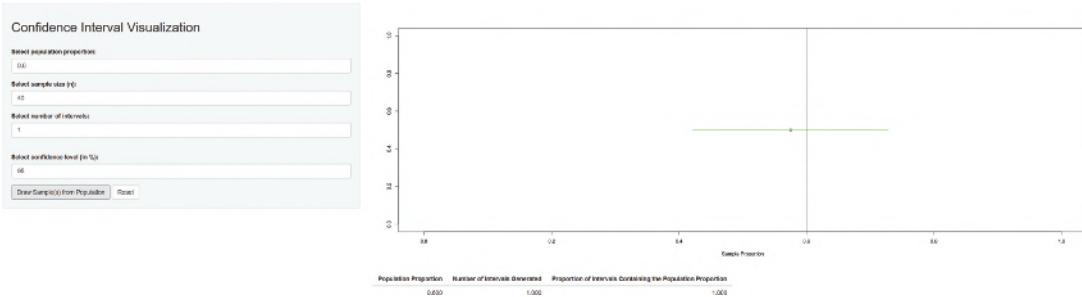
Open the app. You should see a screen like the one shown here.



This app will select a random sample from a population consisting of Successes and Failures. It uses the sample data to construct a confidence interval for the population proportion. This interval is then plotted on the display, and you can see if the resulting confidence interval contains the actual value of the population proportion.

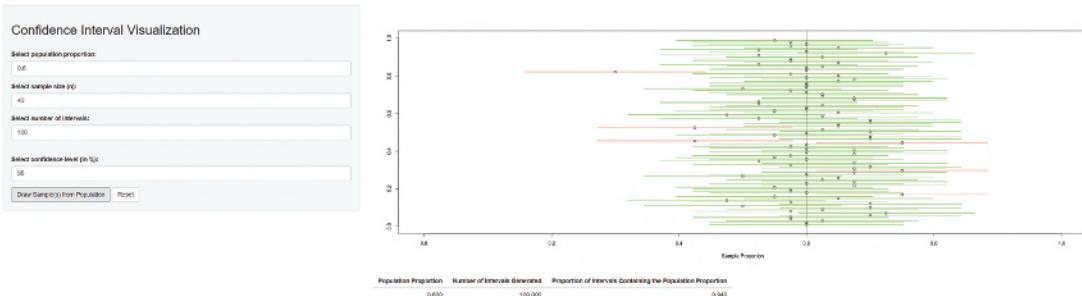
For purposes of this activity, we will take a random sample of size $n = 40$ from a population with a proportion of successes of $p = 0.60$. Change the value in the “Select population proportion” box to 0.60 and change the number in the “Select sample size” box to 40. Leave the number of intervals at 1 and the confidence level at 95. Now click on the “Draw sample(s) from population” button.

You should see a confidence interval appear on the display on the right-hand side. If the interval contains the actual value of the population proportion, the interval is drawn in green. If the value of the population proportion is not in the interval, the interval is shown in red. Your screen should look something like the following:



Click on the “Draw sample(s) from population” button several more times. Notice how the confidence interval estimate changes from sample to sample. If we were to construct a large number of intervals, the percentage of intervals that contain the value of the population proportion should approximate the capture rate for the confidence interval method.

To look at more than one interval at a time, click on the reset button. Then change the number of intervals to 100. When you click on the “Draw sample(s) from population” button, you should see a screen similar to the one below.



Notice that the table below the display includes the proportion of the intervals generated that include the value of the population proportion.

Next, click on the reset button and change the number of intervals to 1000. Click on the “Draw sample(s) from population” button to see the result of generating 1000 different 95% confidence intervals.

- How does the proportion of intervals constructed that include $p = 0.60$, the value of the population proportion, compare to 0.95?
- Experiment with several other confidence levels of your choice, and then answer the following question: In general, is the proportion of confidence intervals that contain $p = 0.60$ close to the stated confidence level?

ACTIVITY 9.2 AN ALTERNATIVE CONFIDENCE INTERVAL FOR A POPULATION PROPORTION

Technology Activity

This activity presumes that you have already completed Activity 9.1. This activity uses the Shiny apps titled CIProportionVisualization and AlternativeConfidenceIntervalVisualization that are part of the collection of Shiny web apps that accompany this text. These apps can be found at statistics.cengage.com/PSO6e/Apps.html.

Background: In Section 9.2, it was suggested that a confidence interval of the form

$$\hat{p}_{\text{mod}} \pm (z \text{ critical value}) \sqrt{\frac{\hat{p}_{\text{mod}}(1 - \hat{p}_{\text{mod}})}{n}}$$

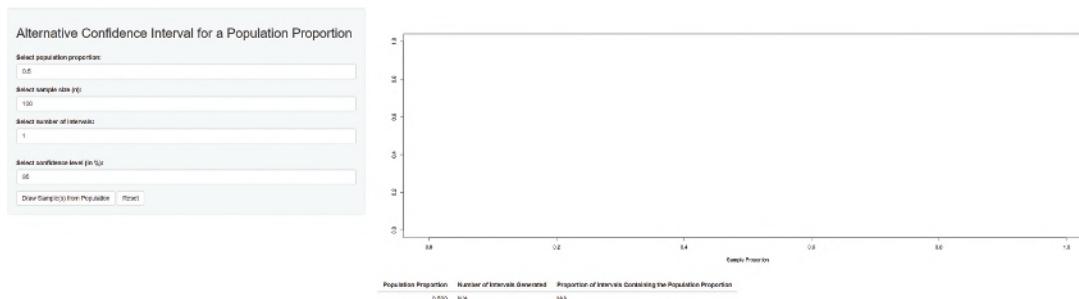
where $\hat{p}_{\text{mod}} = \frac{\text{successes} + 2}{n + 4}$ is an alternative to the usual large-sample z confidence interval. This alternative interval is

preferred by many statisticians because, in repeated sampling, the proportion of intervals constructed that include the actual value of the population proportion, p , tends to be closer to the stated confidence level. In this activity, we will explore how the “capture rates” for the two different interval estimation methods compare.

Open the app CIProportionVisualization that was also used in Activity 9.1. Use this app to generate 1000 95% confidence intervals for samples of size $n = 40$ from a population with $p = 0.3$. Notice that $n = 40$ is large enough to satisfy the sample size conditions for the large-sample confidence interval for a population proportion ($np \geq 10$ and $n(1 - p) \geq 10$).

- How does the proportion of intervals constructed that include $p = 0.6$, the value of the population proportion, compare to 0.95? Does this surprise you? Explain.

Now open the app Alternative Confidence Interval Visualization. This app is similar to the one used to generate large-sample confidence intervals, except this app selects a random sample and then uses the alternative method based on \hat{p}_{mod} to generate the confidence intervals. You should see a screen similar to the one below.



- Use the app to generate 1000 95% confidence intervals for samples of size $n = 40$ from a population with $p = 0.3$. How does the proportion of intervals constructed that include $p = 0.3$, the population proportion, compare to 0.95? Is this proportion closer to 0.95 than was the case for the large-sample z interval?
- Experiment with different combinations of values of sample size and population proportion p . Can you find a combination for which the large sample z interval has a capture rate that is close to 95%? Can you find a combination for which it has a capture rate that is even farther from 95% than it was for $n = 40$ and $p = 0.3$? How does the modified interval perform in each of these cases?

ACTIVITY 9.3 VERIFYING SIGNATURES ON A RECALL PETITION

Background: In 2003, petitions were submitted to the California Secretary of State calling for the recall of Governor Gray Davis. Each of California's 58 counties then had to report the number of valid signatures on the petitions from that county so that the State could determine whether there were enough valid signatures to certify the recall and set a date for the recall election. The following paragraph appeared in the *San Luis Obispo Tribune* (July 23, 2003):

In the campaign to recall Gov. Gray Davis, the secretary of state is reporting 16,000 verified signatures from San Luis Obispo County. In all, the County Clerk's Office received 18,866 signatures on recall petitions and was instructed by the state to check a random sample of 567. Out of those, 84.48% were good. The verification process includes checking whether

the signer is a registered voter and whether the address and signature on the recall petition match the voter registration.

1. Use the data from the random sample of 567 San Luis Obispo County signatures to construct a 95% confidence interval for the proportion of petition signatures that are valid.
2. How do you think that the reported figure of 16,000 verified signatures for San Luis Obispo County was obtained?
3. Based on your confidence interval from Step 1, explain why you think that the reported figure of 16,000 verified signatures is or is not reasonable.

ACTIVITY 9.4 A MEANINGFUL PARAGRAPH

Write a meaningful paragraph that includes the following six terms: **sample**, **population**, **confidence level**, **estimate**, **mean**, **margin of error**.

A “meaningful paragraph” is a coherent piece writing in an appropriate context that uses all of the listed words. The paragraph should show that you understand

the meanings of the terms and their relationship to one another. A sequence of sentences that just define the terms is *not* a meaningful paragraph. When choosing a context, think carefully about the terms you need to use. Choosing a good context will make writing a meaningful paragraph easier.

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
Point estimate	A single number, based on sample data, that represents a plausible value of a population characteristic.	$n = p(1 - p) \left(\frac{1.96}{M}\right)^2$	A formula used to calculate the sample size necessary for estimating p to within an amount M with 95% confidence. (For other confidence levels, replace 1.96 with an appropriate z critical value.)
Unbiased statistic	A statistic that has a sampling distribution with a mean equal to the value of the population characteristic to be estimated.	$\bar{x} \pm (z \text{ critical value}) \frac{\sigma}{\sqrt{n}}$	A formula used to construct a confidence interval for μ when σ is known and either the sample size is large or the population distribution is normal.
Confidence interval	An interval calculated from sample data that provides a range of plausible values for a population characteristic.	$\bar{x} \pm (t \text{ critical value}) \frac{s}{\sqrt{n}}$	A formula used to construct a confidence interval for μ when σ is unknown and either the sample size is large or the population distribution is normal.
Confidence level	A number that provides information on how much “confidence” we can have in the method used to construct a confidence interval estimate. The confidence level specifies the percentage of all possible samples that will produce an interval containing the value of the population characteristic.	$n = \left(\frac{1.96\sigma}{M}\right)^2$	A formula used to calculate the sample size necessary for estimating μ to within an amount M with 95% confidence. (For other confidence levels, replace 1.96 with an appropriate z critical value.)
$\hat{p} \pm (z \text{ critical value}) \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$	A formula used to construct a confidence interval for p when the sample size is large.		

CHAPTER REVIEW Exercises 9.72 - 9.82

- 9.72** The article “[Write It by Hand to Make It Stick](#)” (*Advertising Age*, July 27, 2016) summarizes data from a survey of 1001 students age 13 to 19. Of the students surveyed, 851 reported that they learn best using a mix of digital and nondigital tools. Construct and interpret a 95% confidence interval for the proportion of students age 13 to 19 who would say that they learn best using a mix of digital and non-digital tools. For the method used to construct the interval to be valid, what assumption about the sample must be reasonable?
- 9.73** The article referenced in the previous exercise also indicated that 811 of the 1001 students surveyed said that they would feel restricted if they could only work on digital devices. Would a 95% confidence interval for the proportion of students age 13 to 19 who would say that they would feel restricted if they could only use digital devices be narrower or wider than the interval constructed in the previous exercise for the proportion who would say that they learn best using a mix of digital and non-digital tools? Explain your reasoning—you should be able to answer this question without computing the second confidence interval.
- 9.74** The report “[The 2016 Consumer Financial Literacy Survey](#)” ([The National Foundation for Credit Counseling](#), [nfcc.org](#), retrieved October 28, 2016) summarized data from a representative sample of 1668 adult Americans. Based on data from this sample, it was reported that over half of U.S. adults would give themselves a grade of A or B on their knowledge of personal finance. This statement was based on observing that 934 people in the sample would have given themselves a grade of A or B.
- Construct and interpret a 95% confidence interval for the proportion of all adult Americans who would give themselves a grade of A or B on their financial knowledge of personal finance.
 - Is the confidence interval from Part (a) consistent with the statement that a majority of adult Americans would give themselves a grade of A or B? Explain why or why not.
- 9.75** The report “[The Politics of Climate](#)” ([Pew Research Center](#), October 4, 2016, [pewinternet.org/2016/10/04/the-politics-of-climate](#)) summarized data from a survey on public opinion of renewable and other energy sources. It was reported that 52% of the people in a sample from western states said that they have considered installing solar panels on their homes. This percentage was based on a representative sample of 369 homeowners in the western United States. Use the given information to construct and interpret a 90% confidence interval for the proportion of all homeowners in western states who have considered installing solar panels.
- 9.76** The report referenced in the previous exercise also indicated that 33% of those in a representative sample of 533 homeowners in southern states said that they had considered installing solar panels.
- Use the given information to construct and interpret a 90% confidence interval for the proportion of all homeowners in the southern states who have considered installing solar panels.
 - Give two reasons why the confidence interval in Part (a) is narrower than the confidence interval calculated in the previous exercise.
- 9.77** Data from a survey of a representative sample was used to estimate that 32% of computer users in 2011 had tried to get on a Wi-Fi network that was not their own in order to save money (*USA TODAY*, May 16, 2011). Suppose you decide to conduct a survey to estimate this proportion for the current year. What is the required sample size if you want to estimate this proportion with a margin of error of 0.05? Calculate the required sample size first using 0.32 as a preliminary estimate of p and then using the conservative value of 0.5. How do the two sample sizes compare? What sample size would you recommend for this study?
- 9.78** A manufacturer of small appliances purchases plastic handles for coffeepots from an outside vendor. If a handle is cracked, it is considered defective and must be discarded. A large shipment of plastic handles is received. The proportion of defective handles p is of interest. How many handles from the shipment should be inspected to estimate p to within 0.1 with 95% confidence?
- 9.79** A manufacturer of college textbooks is interested in estimating the strength of the bindings produced by a particular binding machine. Strength can be measured by recording the force required to pull the pages from the binding. If this force is measured in pounds, how many books should be tested to estimate the mean force required to break the binding to within 0.1 pounds with 95% confidence? Assume that σ is known to be 0.8 pound.
- 9.80** The confidence intervals presented in this chapter give both lower and upper bounds on plausible values for the population characteristic being estimated. In some instances, only an upper bound or only a lower bound is appropriate. Using the same

reasoning that gave the large sample interval in Section 9.3, we can say that when n is large, 99% of all samples have

$$\mu < \bar{x} + 2.33 \frac{s}{\sqrt{n}}$$

(because the area under the z curve to the left of 2.33 is 0.99). Thus, $\bar{x} + 2.33 \frac{s}{\sqrt{n}}$ is a 99% upper confidence bound for μ . For a sample of 539 patients waiting for bypass surgery, the mean wait time was 19 days and the standard deviation was 10 days. Calculate the 99% upper confidence bound for the population mean wait time for bypass patients.

- 9.81** When n is large, the statistic s is approximately unbiased for estimating σ and has approximately a normal distribution. The standard deviation of this statistic when the population distribution is normal is $\sigma_s \approx \frac{\sigma}{\sqrt{2n}}$ which can be estimated by $\frac{s}{\sqrt{2n}}$. An

approximate large-sample confidence interval for the population standard deviation σ is then

$$s \pm (z \text{ critical value}) \frac{s}{\sqrt{2n}}$$

For a sample of 847 patients waiting for angiography, the standard deviation of wait time was 9 days. Calculate a 95% confidence interval for the population standard deviation of waiting time for angiography.

- 9.82** The interval from -2.33 to 1.75 captures an area of 0.95 under the z curve. This implies that another large-sample 95% confidence interval for μ has lower limit $\bar{x} - 2.33 \frac{\sigma}{\sqrt{n}}$ and upper limit $\bar{x} + 1.75 \frac{\sigma}{\sqrt{n}}$. Would you recommend using this 95% interval over the 95% interval $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ discussed in the text? Explain. (Hint: Look at the width of each interval.)

TECHNOLOGY NOTES

Confidence Intervals for Proportions

JMP

Summarized data

- Enter the data table into the JMP data table with categories in the first column and counts in the second column

	Column 1	Column 2
1	Yes	130
2	No	45

- Click **Analyze** and select **Distribution**
- Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Columns**
- Click and drag the second column name from the box under **Select Columns** to the box next to **Freq**
- Click **OK**
- Click the red arrow next to the column name and click **Confidence Interval** then select the appropriate level or select **Other** to input a level that is not listed

Raw data

- Enter the raw data into a column
- Click **Analyze** and select **Distribution**
- Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Columns**
- Click **OK**

- Click the red arrow next to the column name and click **Confidence Interval** then select the appropriate level or select **Other** to input a level that is not listed

Minitab

Summarized data

- Click **Stat** then click **Basic Statistics** then click **1 Proportion...**
- Click the radio button next to **Summarized data**
- In the box next to **Number of Trials:** type the value for n , the total number of trials
- In the box next to **Number of events:** type the value for the number of successes
- Click **Options...**
- Input the appropriate confidence level in the box next to **Confidence Level**
- Check the box next to **Use test and interval based on normal distribution**
- Click **OK**
- Click **OK**

Raw data

- Input the raw data into a column
- Click **Stat** then click **Basic Statistics** then click **1 Proportion...**
- Click in the box under **Samples in columns:**
- Double click the column name where the raw data is stored
- Click **Options...**
- Input the appropriate confidence level in the box next to **Confidence Level**
- Check the box next to **Use test and interval based on normal distribution**

8. Click **OK**
9. Click **OK**

SPSS

SPSS does not have the functionality to automatically produce confidence intervals for a population proportion.

Excel

Excel does not have the functionality to automatically produce confidence intervals for a population proportion. However, you can manually type in the formula for the lower and upper limits separately into two different cells to have Excel calculate the result for you.

TI-83/84

1. Press the **STAT** key
2. Highlight **TESTS**
3. Highlight **1-PropZInterval** and press **ENTER**
4. Next to **x** type the number of successes
5. Next to **n** type the number of trials, **n**
6. Next to **C-Level** type the value for the confidence level
7. Highlight **Calculate** and press **ENTER**

TI-Nspire

1. Enter the Calculate Scratchpad
2. Press the **menu** key then select **6:Statistics** then select **6:Confidence Intervals** then **5:1-Prop z Interval...** then press **enter**
3. In the box next to **Successes, x** type the number of successes
4. In the box next to **n** type the number of trials, **n**
5. In the box next to **C Level** type the confidence level
6. Press **OK**

Confidence Interval for μ Based on *t*-distribution

JMP

1. Input the data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name from the box under **Select Columns** to the box next to **Y, Response**
4. Click **OK**
5. Click the red arrow next to the column name and select **Confidence Interval** then select the appropriate confidence level or click **Other** to specify a level

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **1-sample t...**
2. Click the radio button next to **Summarized data**
3. In the box next to **Sample size:** type the value for **n**, the sample size
4. In the box next to **Mean:** type the value for the sample mean
5. In the box next to **Standard deviation:** type the value for the sample standard deviation
6. Click **Options...**

7. Input the appropriate confidence level in the box next to **Confidence Level**
8. Click **OK**
9. Click **OK**

Raw data

1. Input the raw data into a column
2. Click **Stat** then click **Basic Statistics** then click **1-sample t...**
3. Click in the box under **Samples in columns:**
4. Double click the column name where the raw data are stored
5. Click **Options...**
6. Input the appropriate confidence level in the box next to **Confidence Level**
7. Click **OK**
8. Click **OK**

SPSS

1. Input the data into a column
2. Click **Analyze** then select **Compare Means** then select **One-Sample T Test...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Test Variable(s):** box
5. Click **Options...**
6. Input the appropriate confidence level in the box next to **Confidence Interval Percentage:**
7. Click **Continue**
8. Click **OK**

Note: The confidence interval for the population means is in the **One-Sample Test** table.

Excel

Note: Excel does not have the functionality to produce a confidence interval for a single population mean automatically. However, you can use Excel to produce the estimate for the sample mean and the margin of error for the confidence interval using the steps below.

1. Input the raw data into a column
2. Click the **Data** ribbon
3. Click **Data Analysis** in the **Analysis** group

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.

4. Select **Descriptive Statistics** from the dialog box and click **OK**
5. Click in the box next to **Input Range:** and select the data (if you used a column title, check the box next to **Labels in first row**)
6. Click the box next to **Confidence Level for Mean:**
7. In the box to the right of **Confidence Level for Mean:** type in the confidence level you are using
8. Click **OK**

Note: The margin of error can be found in the row titled Confidence Level. In order to find the lower limit of the confidence interval, subtract this from the mean shown in the output. To find the upper limit of the confidence interval, add this to the mean shown in the output.

TI-83/84

Summarized data

1. Press **STAT**
2. Highlight **TESTS**
3. Highlight **TInterval...** and press **ENTER**
4. Highlight **Stats** and press **ENTER**
5. Next to \bar{x} input the value for the sample mean
6. Next to s_x input the value for the sample standard deviation
7. Next to n input the value for the sample size
8. Next to **C-Level** input the appropriate confidence level
9. Highlight **Calculate** and press **ENTER**

Raw data

1. Enter the data into **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**
4. Highlight **TInterval...** and press **ENTER**
5. Highlight **Data** and press **ENTER**
6. Next to **C-Level** input the appropriate confidence level
7. Highlight **Calculate** and press **ENTER**

TI-Nspire

Summarized data

1. Enter the Calculate Scratchpad
2. Press the **menu** key and select **6:Statistics** then **6:Confidence Intervals** then **2:t Interval...** and press **enter**
3. From the drop-down menu select **Stats**
4. Press **OK**
5. Next to \bar{x} input the value for the sample mean
6. Next to s_x input the value for the sample standard deviation
7. Next to n input the value for the sample size
8. Next to **C Level** input the appropriate confidence level
9. Press **OK**

Raw data

1. Enter the data into a data list (In order to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then **3:Confidence Intervals** then **2:t Interval...** and press **enter**
3. From the drop-down menu select **Data**
4. Press **OK**
5. Next to **List** select the list containing your data
6. Next to **C-Level** input the appropriate confidence level
7. Press **OK**

10 Hypothesis Testing Using a Single Sample



PictureNet/Spirit/Corbis

In Chapter 9, we considered situations in which the primary goal was to estimate the unknown value of some population characteristic. Sample data can also be used to decide whether a claim or *hypothesis* about a population characteristic is plausible.

For example, how do employers rate the career readiness of college students? Is there evidence that less than half of employers would rate college seniors as proficient in oral and written communication? The article “[Overconfident Students, Dubious Employers](https://insidehighered.com/news/2018/02/23/study-students-believe-they-are-prepared-workplace-employers-disagree)” (insidehighered.com/news/2018/02/23/study-students-believe-they-are-prepared-workplace-employers-disagree, February 23, 2018, retrieved April 6, 2018) summarizes results from a survey of a representative sample of 201 employers and 4213 graduating seniors. Of the 201 employers surveyed, only 84 rated college seniors as proficient in oral and written communication. With p representing the proportion of all employers who would rate college seniors as proficient, we can use the hypothesis testing methods of this chapter to decide whether the sample data provide convincing evidence that p is less than 0.50.

As another example, a report released by the National Association of Colleges and Employers stated that the average starting salary for students graduating with a bachelor’s degree in 2017 is \$51,022 (“[Early Salaries for Class of 2017, A Mix of Bumps, Dips,](https://naceweb.org)” naceweb.org, retrieved April 6, 2018). Suppose that you are interested in investigating whether the mean starting salary for students graduating from your university this year is greater than the 2017 average of \$51,022. You select a random sample of $n = 40$ graduates from the 2017 graduating class of your university and determine the starting salary of each one.

If this sample produced a mean starting salary of \$52,136 and a standard deviation of \$1214, is it reasonable to conclude that μ , the mean starting salary for all graduates in the 2017 graduating class at your university, is greater than \$51,022? We will see in this chapter how the sample data can be used to decide whether $\mu > 51,022$ is a reasonable conclusion.

LEARNING OBJECTIVES

Students will understand:

- That rejecting the null hypothesis implies strong support for the alternative hypothesis.
- Why failing to reject the null hypothesis does not imply strong support for the null hypothesis.
- The reasoning used to reach a decision in a hypothesis test.

Students will be able to:

- Translate a research question into null and alternative hypotheses.
- Describe Type I and Type II errors in context.
- Carry out a large-sample z test for a population proportion and interpret the results in context.
- Carry out a t test for a population mean and interpret the results in context.
- Describe the effect of the significance level and the sample size on the power of a test.
- Carry out a randomization test for a population proportion. (Optional)
- Carry out an exact binomial test for a population proportion. (Optional)
- Carry out a randomization test for a population mean. (Optional)

SECTION 10.1 Hypotheses and Test Procedures

A hypothesis is a claim or statement about the value of a single population characteristic or the values of several population characteristics. The following are examples of legitimate hypotheses:

$$\mu = 1000, \text{ where } \mu \text{ is the mean number of characters in an e-mail message}$$

$$p < 0.01, \text{ where } p \text{ is the proportion of e-mail messages that are undeliverable}$$

In contrast, the statements $\bar{x} = 1000$ and $\hat{p} = 0.01$ are *not* hypotheses, because \bar{x} and \hat{p} are *sample* characteristics.

A test of hypotheses is a method that uses sample data to decide between two competing claims (hypotheses) about a population characteristic. One hypothesis might be $\mu = 1000$ and the other $\mu \neq 1000$ or one hypothesis might be $p = 0.01$ and the other $p < 0.01$. If it were possible to carry out a census of the entire population, we would know which of the two hypotheses is correct, but usually we must decide between them using information from a sample.

A criminal trial is a familiar situation in which a choice between two contradictory claims must be made. The person accused of the crime must be judged either guilty or not guilty. Under the U.S. system of justice, the individual on trial is initially presumed not guilty. Only strong evidence to the contrary causes the not guilty claim to be rejected in favor of a guilty verdict. The burden is thus put on the prosecution to provide convincing evidence that supports the guilty claim. In some other countries, the perspective in criminal proceedings is the opposite. Once enough evidence has been presented to justify bringing an individual to trial, the initial assumption is that the accused is guilty. The burden then falls on the accused to provide convincing evidence to support a not guilty claim.

As in a judicial proceeding, we initially assume that a particular hypothesis, called the **null hypothesis**, is the correct one. We then consider the evidence (the sample data) and reject the null hypothesis in favor of the competing hypothesis, called the **alternative hypothesis**, only if there is *convincing* evidence against the null hypothesis.

DEFINITIONS

Null hypothesis: A claim about a population characteristic that is initially assumed to be true. The null hypothesis is denoted by H_0 .

Alternative hypothesis: A competing claim about a population characteristic. The alternative hypothesis is denoted by H_a .

In carrying out a test of H_0 versus H_a , the null hypothesis H_0 will be rejected in favor of H_a only if the sample provides convincing evidence that H_0 is false.

If the sample does not provide such evidence, H_0 will not be rejected.

The two possible conclusions in a test of hypotheses are *reject H_0* or *fail to reject H_0* .

Example 10.1 Tennis Ball Diameters

Understand the context ➤

The International Tennis Federation requires that tennis balls have a diameter between 2.57 and 2.70 inches. Because of variation in the manufacturing process, tennis balls produced by a particular machine do not have identical diameters. Let μ denote the mean diameter for all tennis balls currently being produced. Suppose that the machine was initially calibrated to achieve the design specification $\mu = 2.64$ inches. However, the manufacturer is now concerned that the diameters no longer conform to this specification. If sample evidence suggests that $\mu \neq 2.64$ inches, the production process will be halted while the machine is recalibrated. Because stopping production is costly, the manufacturer wants to be quite sure that $\mu \neq 2.64$ inches before undertaking recalibration. Under these circumstances, a sensible choice of hypotheses is

$$\begin{aligned} H_0: \mu &= 2.64 \text{ (the specification is being met, so recalibration is unnecessary)} \\ H_a: \mu &\neq 2.64 \text{ (the specification is not being met, so recalibration is necessary)} \end{aligned}$$

H_0 would be rejected in favor of H_a only if the sample provides convincing evidence against the null hypothesis.

Example 10.2 Compact Fluorescent Lightbulb Lifetimes

Understand the context ➤

Compact fluorescent (cfl) lightbulbs are more energy efficient than standard incandescent light bulbs. Ecobulb brand 60-watt cfl lightbulbs state on the package “Average life 8,000 hours.” Let μ denote the true mean life of Ecobulb 60-watt cfl lightbulbs. Then the advertised claim is $\mu = 8000$ hours. People who purchase this brand would be unhappy if μ is actually less than the advertised value.

Suppose that a sample of Ecobulb cfl lightbulbs is selected and tested. The lifetime for each bulb in the sample is recorded. The sample data can then be used to test the hypothesis $\mu = 8000$ hours against the hypothesis $\mu < 8000$ hours. The accusation that the company is overstating the mean lifetime is a serious one, and it is reasonable to require convincing evidence before concluding that $\mu < 8000$. This suggests that the claim $\mu = 8000$ should be selected as the null hypothesis and that $\mu < 8000$ should be selected as the alternative hypothesis. Then

$$H_0: \mu = 8000$$

would be rejected in favor of

$$H_a: \mu < 8000$$

only if sample provides convincing evidence that the initial assumption, $\mu = 8000$ hours, is not plausible.

Because the alternative hypothesis in Example 10.2 is $\mu < 8000$ (the actual mean lifetime is less than the advertised value), it might have seemed sensible to state H_0 as the inequality $\mu \geq 8000$. The claim $\mu \geq 8000$ is in fact the *implicit* null hypothesis, but we will state H_0 as a claim of equality. There are several reasons for this. First, the development of a decision rule is most easily understood if there is only a single hypothesized value of μ (or p or whatever other population characteristic is under consideration). Second, suppose that the sample data provided compelling evidence that $H_0: \mu = 8000$ should be rejected in favor of $H_a: \mu < 8000$. This means that we were convinced by the sample data that the population mean was smaller than 8000. It follows that we would have also been convinced that the population mean could not have been 8001 or 8010 or any other value that was greater than 8000. This means that the conclusion when testing $H_0: \mu = 8000$ versus $H_a: \mu < 8000$ is always the same as the conclusion for a test where the null hypothesis is $H_0: \mu \geq 8000$. For these reasons, the null hypothesis H_0 is usually stated as a claim of equality.

The form of a null hypothesis is

$$H_0: \text{population characteristic} = \text{hypothesized value}$$

where the hypothesized value is a specific number determined by the problem context.

The alternative hypothesis will have one of the following three forms:

$$H_a: \text{population characteristic} > \text{hypothesized value}$$

$$H_a: \text{population characteristic} < \text{hypothesized value}$$

$$H_a: \text{population characteristic} \neq \text{hypothesized value}$$

Notice that we can test $H_0: p = 0.1$ versus $H_a: p < 0.1$, but we can't test $H_0: \mu = 50$ versus $H_a: \mu > 100$. The number appearing in the alternative hypothesis must be the same as the hypothesized value in H_0 .

Example 10.3 illustrates how the selection of H_0 (the claim initially assumed to be true) and H_a depends on the objectives of a study.

Example 10.3 Evaluating a New Medical Treatment

Understand the context ➤

A medical research team has been given the task of evaluating a new laser treatment for certain types of tumors. Consider the following two scenarios:

Scenario 1: The current standard treatment is considered reasonable and safe by the medical community, has no major side effects, and has a known success rate of 0.85 (85%).

Scenario 2: The current standard treatment sometimes has serious side effects, is costly, and has a known success rate of 0.30 (30%).

In the first scenario, research efforts would probably be directed toward determining whether the new treatment has a greater success rate than the standard treatment. Unless there is convincing evidence of this, it is unlikely that current medical practice would change. With p representing the true proportion of successes for the new laser treatment, the following hypotheses would be tested:

$$H_0: p = 0.85 \quad \text{versus} \quad H_a: p > 0.85$$

In this case, rejecting the null hypothesis indicates convincing evidence that the success rate is greater for the new treatment.

In the second scenario, the current standard treatment does not have much to recommend it. The new laser treatment may be considered preferable because of cost or because it has fewer or less serious side effects, as long as the success rate for the new procedure is no worse than the success rate of the standard treatment. Here, researchers might decide to test the hypothesis

$$H_0: p = 0.30 \quad \text{versus} \quad H_a: p < 0.30$$

If the null hypothesis is rejected, the new treatment will not be recommended as an alternative to the standard treatment, because there is convincing evidence that the laser method has a lower success rate.

If the null hypothesis is not rejected, we are able to conclude only that there is not convincing evidence that the success rate for the laser treatment is less than the success rate for the standard treatment. This is *not* the same as saying that we have evidence that the laser treatment is as good as the standard treatment. If medical practice were to embrace the new procedure, it would not be because it has a greater success rate but rather because it costs less or has fewer side effects, and there is not convincing evidence that it has a lower success rate than the standard treatment.

A statistical hypothesis test is only capable of demonstrating strong support for the alternative hypothesis (by rejecting the null hypothesis). When the null hypothesis is not rejected, it does not mean strong support for the null hypothesis—only that there is not convincing evidence against it.

In the lightbulb scenario of Example 10.2, if $H_0: \mu = 8000$ is rejected in favor of $H_a: \mu < 8000$, it is because we have convincing evidence that actual mean lifetime is less than the advertised value. However, not rejecting H_0 does not necessarily provide strong support for the advertised claim. If the objective is to demonstrate that the mean lifetime is greater than 8000 hours, the hypotheses that would be tested are $H_0: \mu = 8000$ versus $H_a: \mu > 8000$. Then rejecting H_0 indicates convincing evidence that $\mu > 8000$. *When deciding which alternative hypothesis to use, keep the research objectives in mind.*

EXERCISES 10.1 - 10.11

- 10.1** Explain why the statement $\bar{x} = 50$ is not a legitimate hypothesis.
- 10.2** For the following pairs, indicate which do not comply with the rules for setting up hypotheses, and explain why:
- $H_0: \mu = 15, H_a: \mu = 15$
 - $H_0: p = 0.4, H_a: p > 0.6$
 - $H_0: \mu = 123, H_a: \mu < 123$
 - $H_0: \mu = 123, H_a: \mu = 125$
 - $H_0: \hat{p} = 0.1, H_a: \hat{p} \neq 0.1$
- 10.3** To determine whether the pipe welds in a nuclear power plant meet specifications, a random sample of welds is selected and tests are conducted on each weld in the sample. Weld strength is measured as the force required to break the weld. Suppose that the specifications state that the mean strength of welds should exceed 100 lb/in². The inspection team decides to test $H_0: \mu = 100$ versus $H_a: \mu > 100$. Explain why this alternative hypothesis was chosen rather than $\mu < 100$.
- 10.4** According to an article in *Science Daily* (“**Still No Strong Evidence That Adjunct Treatment with HGH in IVF Improves Results**,” sciencedaily.com, July 4, 2016, retrieved November 26, 2016), women who are having difficulty becoming pregnant sometimes use human growth hormone (HGH) in addition to in-vitro fertilization (IVF) to try to have a baby. A large study found that “there was no strong evidence” that the proportion of women who became pregnant while taking HGH along with IVF was greater than the success rate for IVF alone.
- Is this consistent with testing $H_0: \text{HGH in addition to IVF increases the chance of getting pregnant}$ versus $H_a: \text{HGH in addition to IVF does not increase the chance of getting pregnant}$ or with testing $H_0: \text{HGH in addition to IVF does not increase the chance of getting pregnant}$ versus $H_a: \text{HGH in addition to IVF increases the chance of getting pregnant}$ Explain.
 - Does the stated conclusion of “no strong evidence” indicate that the null hypothesis was rejected? Explain.
- 10.5** A press release about a paper that appeared in [The Journal of Youth and Adolescence \(springer.com /about+springer/media/springer+select?SGW_ID=0-11001-6-1433942-0, August 26, 2013, retrieved May 8, 2017\)](https://link.springer.com/article/10.1007/s10626-013-0143-0) was titled “Video Games Do Not Make Vulnerable Teens More Violent.” The press release includes the following statement about the study described in the paper: “Study finds no evidence that violent video games increase antisocial behavior in youths with pre-existing psychological conditions.” In the context of a hypothesis test with the null hypothesis being that video games do not increase antisocial behavior, explain why the title of the press release is misleading.
- 10.6** CareerBuilder.com conducted a survey to learn about the proportion of employers who perform background checks when evaluating a job candidate (“**Majority of Employers Background Check Employees...Here's Why**,” November 17, 2016, retrieved November 19, 2016). Suppose you are interested in determining if the resulting data provide convincing evidence in support of the claim that more than two-thirds of

employers perform background checks. What pair of hypotheses should you test?

- 10.7** A national survey of 1012 adult Americans conducted by Gallup ([“Americans Still Generally Upbeat About Personal Finances,” gallup.com, January 25, 2016, retrieved November 16, 2016](#)) asked survey participants if they thought they were in better financial shape than they were 1 year ago. Suppose that you want to determine if the survey data provide convincing evidence that a majority of adult Americans believe they are in better financial shape than 1 year ago. With p = the population proportion of adult Americans who believe they are better off, what hypotheses should you test?
- 10.8** A researcher speculates that because of differences in diet, Japanese children may have a lower mean blood cholesterol level than U.S. children do. Suppose that the mean level for U.S. children is known to be 170. Let μ represent the mean blood cholesterol level for all Japanese children. What hypotheses should the researcher test?
- 10.9** A county commissioner must vote on a resolution that would commit substantial resources to the construction of a sewer in an outlying residential area. Her fiscal decisions have been criticized in the past, so she decides to take a survey of people in her district to find out whether they favor spending money for a sewer system. She will vote to appropriate funds only if she is convinced that a majority (more than 50%) of the people in her district favor the measure. What hypotheses should she test?
- 10.10** A cruise ship charges passengers \$3 for a can of soda. Because of passenger complaints, the ship manager has decided to try out a plan with a lower price. He thinks that with a lower price, more cans will be sold, which would mean that the ship would still make a reasonable total profit. With the old pricing, the mean number of cans sold per passenger for a 10-day trip was 10.3 cans. Suppose μ represents the mean number of cans per passenger for the new pricing. What hypotheses should the ship manager test if he wants to determine if the mean number of cans sold is greater for the new pricing plan?
- 10.11** The article [“Facebook Use and Academic Performance Among College Students,” Computers in Human Behavior \[2015\]: 265–272](#)) estimated that 70 percent of students at a large public university in California who are Facebook users log into their Facebook profiles at least six times a day. Suppose that you plan to select a random sample of 400 students at your college. You will ask each student in the sample if he or she is a Facebook user and if they log into their Facebook profile at least six times a day. You plan to use the resulting data to decide if there is evidence that the proportion for your college is different from the proportion reported in the article for the college in California. What hypotheses should you test?

SECTION 10.2 Errors in Hypothesis Testing

Once hypotheses have been formulated, a **test procedure** uses sample data to determine whether the null hypothesis, H_0 , should be rejected. Just as a jury may reach the wrong verdict in a trial, there is some chance that using a test procedure with sample data may lead us to the wrong conclusion about a population characteristic. In this section, we discuss the types of errors that can occur and consider how the choice of a test procedure influences the chances of these errors.

One incorrect conclusion in a criminal trial is for a jury to convict an innocent person. Another is for a guilty person to be set free. Similarly, there are two different types of errors that might be made when making a decision in a hypothesis test. One type of error involves rejecting H_0 even though the null hypothesis is true. The second type of error results from failing to reject H_0 when it is false. These errors are known as Type I and Type II errors, respectively.

DEFINITIONS

Type I error: The error of rejecting H_0 when H_0 is true

Type II error: The error of failing to reject H_0 when H_0 is false

The only way to guarantee that neither type of error occurs is to base the decision on a census of the entire population. Risk of error is the price paid for basing the decision on sample data.

Example 10.4 On-Time Arrivals

Understand the context ➤

The **U.S. Bureau of Transportation Statistics** reports that for 2015, 79.9% of all domestic passenger flights arrived within 15 minutes of the scheduled arrival time (*Air Travel Consumer Reports, February 2016*). Flights that arrive within 15 minutes of the scheduled time are considered to be on-time. Suppose that an airline with a poor on-time record decides to offer its employees a bonus if the airline's proportion of on-time flights exceeds the overall industry rate of 0.799 in an upcoming month. We can use p to represent the actual proportion of the airline's flights that are on time during the month of interest. A random sample of flights might be selected and used as a basis for choosing between

$$H_0: p = 0.799 \quad \text{and} \quad H_a: p > 0.799$$

In this context, a Type I error (rejecting a true H_0) is concluding that the airline on-time rate exceeds the overall industry rate, when in fact the airline does not have a better record. This Type I error would result in the airline rewarding its employees when the actual proportion of on-time flights was not actually greater than 0.799. A Type II error (not rejecting a false H_0) is not concluding that the airline's on-time proportion is greater than the industry proportion when the airline really did have a better on-time record. A type II error would result in the airline employees *not* receiving a reward that they deserved. Notice that the consequences associated with Type I and Type II errors are different.

Example 10.5 Slowing the Growth of Tumors

Understand the context ➤

In 2004, Vertex Pharmaceuticals, a biotechnology company, issued a press release announcing that it had filed an application with the Food and Drug Administration to begin clinical trials of an experimental drug VX-680 that had been found to reduce the growth rate of pancreatic and colon cancer tumors in animal studies (*New York Times, February 24, 2004*).

In this context, we can use μ to represent the actual mean growth rate of tumors for patients receiving the experimental drug. Data resulting from the planned clinical trials could be used to test

$$H_0: \mu = \text{mean growth rate of tumors without the experimental drug}$$

versus

$$H_a: \mu < \text{mean growth rate of tumors without the experimental drug}$$

The null hypothesis states that the experimental drug is not effective—that the mean growth rate of tumors for patients receiving the experimental drug is the same as for patients who do not take the experimental drug. The alternative hypothesis states that the experimental drug is effective in reducing the mean growth rate of tumors.

In this context, a Type I error is to incorrectly conclude that the experimental drug is effective in slowing the growth rate of tumors. A potential consequence of making a Type I error is that the company would continue to devote resources to the development of the drug when it really is not effective.

A Type II error is not concluding that the experimental drug is effective when the mean growth rate of tumors actually is reduced. A potential consequence of making a Type II error is that the company might abandon development of a drug that was effective.

The accompanying box introduces the terminology and notation used to describe error probabilities.

DEFINITIONS

The **probability of a Type I error** is denoted by α and is called the **significance level** of the test. For example, a test with $\alpha = 0.01$ is said to have a significance level of 0.01.

The **probability of a Type II error** is denoted by β .

Example 10.6 Early Detection of Lung Cancer

Understand the context ➤

Early detection has been shown to increase the chance of survival for patients with lung cancer. The paper “**Urinary Protein Biomarkers in the Early Detection of Lung Cancer**” (*Cancer Prevention Research* [2015]: 111–117) includes data from a study to determine if it is possible to accurately diagnose lung cancer using biomarkers found in a patient’s urine. The researchers developed a screening method and then tested the method with a group of people who were known to have lung cancer and a group of patients who were known to not have lung cancer. The paper included the following information:

- The test correctly identified lung cancer in 39 of 54 patients known to have lung cancer.
- The test correctly identified as cancer-free all of the 49 people tested who were known not to have lung cancer.

You can think of using this test to choose between two hypotheses:

$$\begin{aligned} H_0 &: \text{patient has lung cancer} \\ H_a &: \text{patient does not have lung cancer} \end{aligned}$$

Although these are not “statistical hypotheses” (statements about a population characteristic), the possible decision errors are analogous to Type I and Type II errors.

In this situation, believing that a patient with lung cancer is cancer-free would be a Type I error—rejecting the hypothesis of lung cancer when it is actually true. Believing that a cancer-free patient does have lung cancer is a Type II error—not rejecting the null hypothesis when it is actually false.

Based on the study results, we can estimate the error probabilities. The probability of a Type I error, α , is approximately $(54 - 39)/54 = 0.278$. The probability of a Type II error, β , is approximately $0/49 = 0$.

An ideal test procedure would result in both $\alpha = 0$ and $\beta = 0$. However, if we must base our decision on incomplete information—a sample rather than a census—it is impossible to achieve this ideal. The standard test procedures allow us to control α , but they provide no direct control over β . Because α represents the probability of rejecting a true null hypothesis, selecting a significance level $\alpha = 0.05$ results in a test procedure that, used over and over with different random samples, rejects a *true* H_0 about 5 times in 100. Selecting $\alpha = 0.01$ results in a test procedure with a Type I error rate of 1% in long-term repeated use. Choosing a small value for α implies that we want a procedure for which the risk of a Type I error is quite small.

One question arises naturally at this point: If we can select α , the probability of making a Type I error, why would we ever select $\alpha = 0.05$ rather than $\alpha = 0.01$? Why not always select a very small value for α ? To achieve a small probability of making a Type I error, we would need the corresponding test procedure to require the evidence against H_0 to be very strong before the null hypothesis can be rejected. Although this makes a Type I error unlikely, it increases the risk of a Type II error (*not* rejecting H_0 when it should have been rejected). The choice will depend on the consequences of Type I and Type II errors. If a Type II error has serious consequences, it may be a good idea to select a somewhat larger value for α .

In general, there is a compromise between small α and small β , leading to the following widely accepted principle for selecting the significance level to be used in a hypothesis test.

After assessing the consequences of Type I and Type II errors, identify the largest α that is acceptable. Then use a test procedure with this maximum acceptable value as the level of significance (because using a smaller α increases β).

Example 10.7 Lead in Tap Water

Understand the context ➤

The Environmental Protection Agency (EPA) has adopted what is known as the Lead and Copper Rule, which defines drinking water as unsafe if the concentration of lead is 15 parts per billion (ppb) or greater or if the concentration of copper is 1.3 parts per million (ppm) or greater.

With μ denoting the mean concentration of lead, the manager of a community water system might use lead level measurements from a sample of water specimens to test

$$H_0: \mu = 15 \quad \text{versus} \quad H_a: \mu < 15$$

The null hypothesis (which also implicitly includes the $\mu > 15$ case) states that the mean lead concentration is excessive by EPA standards. The alternative hypothesis states that the mean lead concentration is at an acceptable level and that the water system meets EPA standards for lead.

In this context, a Type I error leads to the conclusion that a water source meets EPA standards for lead when in fact it does not. Possible consequences of this type of error include health risks associated with excessive lead consumption (for example, increased blood pressure, hearing loss, and, in severe cases, anemia and kidney damage).

A Type II error is concluding that the water does not meet EPA standards for lead when it actually does. Possible consequences of a Type II error include elimination of a community water source. Because a Type I error might result in potentially serious public health risks, a small value of α (Type I error probability), such as $\alpha = 0.01$, could be selected. Of course, selecting a small value for α increases the risk of a Type II error. If the community has only one water source, a Type II error could also have very serious consequences for the community, and we might want to rethink the choice of α .

EXERCISES 10.12 - 10.22

- 10.12** Researchers at Boston's Children's Hospital and Harvard Medical School analyzed records of breast cancer screening and diagnostic evaluations ([National Expenditure for False-Positive Mammograms and Breast Cancer Overdiagnoses Estimated at \\$4 Billion a Year](#), *Health Affairs* [2015]: 576–583). Discussing the downsides of the screening process, the article states that the rate of false-positives is higher than previously thought, and that false-positives lead to unnecessary medical follow-up that can be costly.

Suppose that screening is used to decide between a null hypothesis of

$$H_0: \text{no cancer is present}$$

and an alternative hypothesis of

$$H_a: \text{cancer is present.}$$

(Although these are not hypotheses about a population characteristic, this exercise illustrates the definitions of Type I and Type II errors.) (Hint: See Example 10.6.)

- a. Would a false-positive (thinking that cancer is present when in fact it is not) be a Type I error or a Type II error?
- b. Describe a Type I error in the context of this problem, and discuss the possible consequences of making a Type I error.
- c. Describe a Type II error in the context of this problem, and discuss the possible consequences of making a Type II error.
- d. Which type of error are the researchers con-

cerned about when they say that false-positives lead to unnecessary medical follow-up? Explain why it would be reasonable to use a small significance level.

- 10.13** The paper “[Breast MRI as an Adjunct to Mammography for Breast Cancer Screening in High-Risk Patients](#)” (*American Journal of Roentgenology* [2015]: 889–897) describes a study that investigated the usefulness of MRI (magnetic resonance imaging) to diagnose breast cancer. MRI exams from 650 women were reviewed. Of the 650 women, 13 had breast cancer, and the MRI exam detected breast cancer in 12 of these women. Of the 637 women who did not have breast cancer, the MRI correctly identified that no cancer was present for 547 of them. The accompanying table summarizes this information.

	Breast Cancer Present	Breast Cancer Not Present	Total
MRI Indicated Breast Cancer	12	90	102
MRI Did Not Indicate Breast Cancer	1	547	548
Total	13	637	650

Suppose that an MRI exam is used to decide between the two hypotheses

$$H_0: \text{A woman does not have breast cancer}$$

$$H_a: \text{A woman has breast cancer}$$

(Although these are not hypotheses about a population characteristic, this exercise illustrates the definitions of Type I and Type II errors.)

- a. One possible error would be deciding that a woman who has breast cancer is cancer-free. Is this a Type I error or a Type II error? Use the information in the table to approximate the probability of this type of error.
 - b. There is a second type of error that is possible in this context. Describe this error and use the information in the table to approximate the probability of this type of error.
- 10.14** Medical personnel are required to report suspected cases of child abuse. Because some diseases have symptoms that mimic those of child abuse, doctors who see a child with these symptoms must decide between two competing hypotheses:

$$H_0: \text{symptoms are due to child abuse}$$

$$H_a: \text{symptoms are due to disease}$$

(Although these are not hypotheses about a population characteristic, this exercise illustrates the definitions of Type I and Type II errors.) The article “[Blurred Line Between Illness, Abuse Creates Problem for Authorities](#)” (*Macon Telegraph*, February 28, 2000) included the following quote from a doctor in Atlanta regarding the consequences of making an incorrect decision: “If it’s disease, the worst you have is an angry family. If it is abuse, the other kids (in the family) are in deadly danger.”

- a. For the given hypotheses, describe Type I and Type II errors.
- b. Based on the quote regarding consequences of the two kinds of error, which type of error does the doctor quoted consider more serious? Explain.

- 10.15** How accurate are DNA paternity tests? By comparing the DNA of the baby and the DNA of a man that is being tested, one maker of DNA paternity tests claims that their test is 100% accurate if the man is not the father and 99.99% accurate if the man is the father ([IDENTIGENE, dnatesting.com/paternity-test-questions/paternity-test-accuracy/](#)).

a. Consider using the results of DNA paternity testing to decide between the following two hypotheses:

$$H_0: \text{a particular man is not the father}$$

$$H_a: \text{a particular man is the father}$$

In the context of this problem, describe Type I and Type II errors. (Although these are not hypotheses about a population characteristic, this exercise illustrates the definitions of Type I and Type II errors.)

- b. Based on the information given, what are the values of α , the probability of a Type I error, and β , the probability of a Type II error?

- 10.16** A television manufacturer claims that (at least) 90% of its TV sets will not need service during the first 3 years of operation. A consumer agency wishes to check this claim, so it obtains a random sample of $n = 100$ purchasers and asks each whether the set purchased needed repair during the first 3 years after purchase. Let \hat{p} be the sample proportion of responses indicating no repair (so that no repair is identified with a success). Let p denote the actual proportion of successes for all sets made by this manufacturer.

The agency does not want to claim false advertising unless sample evidence strongly suggests that $p < 0.9$. The appropriate hypotheses are then

$$H_0: p = 0.9 \text{ versus } H_a: p < 0.9.$$

- a. In the context of this problem, describe Type I and Type II errors, and discuss the possible consequences of each.
- b. Would you recommend a test procedure that uses $\alpha = 0.10$ or one that uses $\alpha = 0.01$? Explain.

- 10.17** A manufacturer of hand-held calculators receives large shipments of printed circuits from a supplier. It is too costly and time-consuming to inspect all incoming circuits, so when each shipment arrives, a sample is selected for inspection. Information from the sample is then used to test $H_0: p = 0.01$ versus $H_a: p > 0.01$, where p is the actual proportion of defective circuits in the shipment.

If the null hypothesis is not rejected, the shipment is accepted, and the circuits are used in the production of calculators. If the null hypothesis is rejected, the entire shipment is returned to the supplier because of inferior quality. (A shipment is defined to be of inferior quality if it contains more than 1% defective circuits.)

- a. In this context, define Type I and Type II errors.
- b. From the calculator manufacturer’s point of view, which type of error is considered more serious?
- a. From the printed circuit supplier’s point of view, which type of error is considered more serious?

- 10.18** Water specimens are taken from water used for cooling as it is being discharged from a power plant into a river. It has been determined that as long as the mean temperature of the discharged water is at most 150°F, there will be no negative effects on the river’s ecosystem. To investigate whether the plant is in compliance with regulations that prohibit a mean discharge water temperature above 150°F, temperature will be determined for 50 water specimens at randomly selected times. The resulting data will be used to test the hypotheses $H_0: \mu = 150^\circ\text{F}$ versus $H_a: \mu > 150^\circ\text{F}$.

- a. In the context of this example, describe Type I and Type II errors.
- b. Which type of error would you consider more serious? Explain.

- 10.19** Suppose that for a particular hypothesis test, the consequences of a Type I error are very serious. Would you want to carry out the test using a small significance level α (such as 0.01) or a larger significance level (such as 0.10)? Explain the reason for your choice. (Hint: See discussion just before Example 10.7.)

- 10.20** Suppose that you are an inspector for the Fish and Game Department and that you are given the task of determining whether to prohibit fishing along part of the Oregon coast. You will close an area to fishing if it is determined that fish in that region have an unacceptably high mercury content.

- a. Assuming that a mercury concentration of 5 ppm is considered the maximum safe concentration, which of the following pairs of hypotheses would you test:

$$H_0: \mu = 5 \text{ versus } H_a: \mu > 5$$

or

$$H_0: \mu = 5 \text{ versus } H_a: \mu < 5$$

Give the reasons for your choice.

- b. Would you prefer a significance level of 0.10 or 0.01 for your test? Explain. (Hint: See discussion just before Example 10.7.)

- 10.21** The paper “[Living Near Nuclear Power Plants and Thyroid Cancer Risks](#)” (*Environmental International* [2016]: 42–48) investigated whether living near a nuclear power plant increases the risk of thyroid cancer. The authors of this paper concluded that there was no evidence of increased risk of thyroid cancer in areas that were near a nuclear power plant.
- a. Let p denote the proportion of the population in areas near nuclear power plants who are diagnosed with thyroid cancer during a given year. The researchers who wrote this paper might have considered the two rival hypotheses of the form

$$H_0: p \text{ is equal to the corresponding value for areas without nuclear power plants}$$

$$H_a: p \text{ is greater than the corresponding value for areas without nuclear power plants}$$

Did the researchers reject H_0 or fail to reject H_0 ?

- b. If the researchers are incorrect in their conclusion that there is no evidence of increased risk of thyroid cancer associated with living near a nuclear power plant, are they making a Type I or a Type II error? Explain.
- c. Can the result of this hypothesis test be interpreted as meaning that there is strong evidence that the risk of thyroid cancer is not higher for people living near nuclear power plants? Explain.

- 10.22** An automobile manufacturer is considering using robots for part of its assembly process. Converting to robots is an expensive process, so it will be undertaken only if there is strong evidence that the proportion of defective installations is lower for the robots than for human assemblers. Let p denote the proportion of defective installations for the robots. It is known that human assemblers have a defect proportion of 0.02.

- a. Which of the following pairs of hypotheses should the manufacturer test:

$$H_0: p = 0.02 \text{ versus } H_a: p < 0.02$$

or

$$H_0: p = 0.02 \text{ versus } H_a: p > 0.02$$

Explain your answer.

- b. In the context of this exercise, describe Type I and Type II errors.
- c. Would you prefer a test with $\alpha = 0.01$ or $\alpha = 0.10$? Explain your reasoning.

SECTION 10.3 Large-Sample Hypothesis Tests for a Population Proportion

Now that some basic concepts have been introduced, we are ready to consider how to use sample data to decide between a null and an alternative hypothesis. In a hypothesis test, there are two possible conclusions: We either reject H_0 or we fail to reject H_0 . The logic of hypothesis-testing procedures is this: *We reject the null hypothesis if the observed sample is very unlikely to have occurred when H_0 is true.*

In this section, we consider testing hypotheses about a population proportion when the sample size n is large. As before, p denotes the proportion of individuals or objects in a specified population that possess a certain property. A random sample of n individuals is selected from the population. The sample proportion

$$\hat{p} = \frac{\text{number in the sample that possess the property}}{n}$$

serves as the basis for testing hypotheses about p .

The large-sample test procedure is based on the same properties of the sampling distribution of \hat{p} that were used to obtain a confidence interval for p :

1. $\mu_{\hat{p}} = p$
2. $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

3. When n is large, the sampling distribution of \hat{p} is approximately normal.

These three results imply that the standardized variable

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

has approximately a standard normal distribution when n is large. Example 10.8 shows how this information allows us to make a decision in a hypothesis test.

Example 10.8 Impact of Food Labels

Understand the context ➤



An **Associated Press** survey was conducted to investigate how people use the nutritional information provided on food package labels (*Spartansburg Herald, March 19, 2016*).

Interviews were conducted with 1003 randomly selected adult Americans, and each participant was asked a series of questions, including the following two:

Question 1: When purchasing packaged food, how often do you check the nutrition labeling on the package?

Question 2: How often do you purchase foods that are bad for you, even after you've checked the nutrition labels?

It was reported that 582 responded “frequently” to the question about checking labels and 441 responded very often or somewhat often to the question about purchasing “bad” foods even after checking the label.

Let's start by looking at the responses to the first question. Based on these data, is it reasonable to conclude that a majority of adult Americans frequently check the nutritional labels when purchasing packaged foods? We can answer this question by considering

p = actual proportion of all adult Americans who frequently check nutritional labels

and testing the following hypotheses:

$$H_0: p = 0.5$$

$H_a: p > 0.5$ (The proportion of adult Americans who frequently check nutritional labels is greater than 0.5. That is, more than half (a majority) frequently check nutritional labels.)

Recall that in a hypothesis test, the null hypothesis is rejected only if there is convincing evidence against it—in this case, convincing evidence that $p > 0.5$. If H_0 is rejected, there is strong support for the claim that a majority of adult Americans frequently check nutritional labels when purchasing packaged foods.

For this sample,

$$\hat{p} = \frac{582}{1003} = 0.58$$

The observed sample proportion is certainly greater than 0.5, but this could just be due to sampling variability. That is, when $p = 0.5$ (meaning H_0 is true), the sample proportion \hat{p} usually differs somewhat from 0.5 simply because of chance variation from one sample to another. Is it plausible that a sample proportion of $\hat{p} = 0.58$ occurred because of this chance variation, or is it unusual to observe a sample proportion this large when $p = 0.5$?

To answer this question, we form a *test statistic*, the quantity used as a basis for making a decision between H_0 and H_a . Creating a test statistic involves replacing p with the hypothesized value in $z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ to obtain

$$z = \frac{\hat{p} - 0.5}{\sqrt{\frac{(0.5)(1-0.5)}{n}}}$$

If the null hypothesis is true and the sample size is large, this statistic should have approximately a standard normal distribution, because in this case

1. $\mu_{\hat{p}} = 0.5$

2. $\sigma_{\hat{p}} = \sqrt{\frac{(0.5)(1-0.5)}{n}}$

3. \hat{p} has approximately a normal distribution.

The calculated value of z is the difference between \hat{p} and the hypothesized value of p in terms of the standard deviation. For example, if $z = 3$, then the value of the sample proportion \hat{p} is 3 standard deviations (of \hat{p}) greater than what we would have expected if the null hypothesis were true. How likely is it that a z value at least this inconsistent with H_0 would be observed if H_0 is true? If H_0 is true, the test statistic has (approximately) a standard normal distribution. This means that

$$P(z \geq 3 \text{ when } H_0 \text{ is true}) = \text{area under the } z \text{ curve to the right of } 3.00 = 0.0013$$

It follows that if H_0 is true, fewer than 1% of all samples would result in a value of z at least as inconsistent with H_0 as $z = 3$. Because this z value is in the most extreme 1% of the z distribution, it is reasonable to reject H_0 .

For the given sample data,

Do the work ➤

$$z = \frac{\hat{p} - 0.5}{\sqrt{\frac{(0.5)(1-0.5)}{n}}} = \frac{0.58 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{1003}}} = \frac{0.08}{\sqrt{0.016}} = 5.00$$

This means that, $\hat{p} = 0.58$ is 5 standard deviations greater than what we would expect it to be if the null hypothesis $H_0: p = 0.5$ was true. The sample data appear to be much more consistent with the alternative hypothesis, $H_a: p > 0.5$. In particular,

$$\begin{aligned} P(\text{value of } z \text{ is at least as contradictory to } H_0 \text{ as } 5.00 \text{ when } H_0 \text{ is true}) \\ = P(z \geq 5.00 \text{ when } H_0 \text{ is true}) \\ = \text{area under the } z \text{ curve to the right of } 5.00 \\ \approx 0 \end{aligned}$$

Interpret the results ➤ When the null hypothesis is true, there is virtually no chance of seeing a sample proportion and corresponding z value this extreme as a result of chance variation alone. If \hat{p} is 5 standard deviations or more away from 0.5, it is hard to believe that $p = 0.5$? The evidence for rejecting H_0 in favor of H_a is very compelling.

In spite of the fact that there is strong evidence that a majority of adult Americans frequently check nutritional labels, the responses to the second question suggest that the percentage of people who then ignore the information on the label and purchase “bad” foods anyway is not small—the sample proportion who responded very often or somewhat often was 0.44.

The preceding example illustrates the reasoning behind large-sample procedures for testing hypotheses about p (and other test procedures as well). We begin by assuming that the null hypothesis is true. The sample is then examined in light of this assumption. If the

observed sample proportion would not be unusual when H_0 is true, then chance variability from one sample to another is a plausible explanation for what has been observed, and H_0 is not rejected. On the other hand, if the observed sample proportion would have been unlikely when H_0 is true, the sample provides evidence against the null hypothesis and we reject H_0 . We base a decision to reject or to fail to reject the null hypothesis on an assessment of how extreme or unlikely the observed sample is if H_0 is true.

The assessment of how inconsistent the observed sample proportion is with the null hypothesis, H_0 , is based on the value of the test statistic

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}}$$

The value of the test statistic is then used to calculate the **P-value**, the probability, assuming that H_0 is true, of obtaining a z value at least as inconsistent with H_0 as what was actually observed.

DEFINITIONS

Test statistic: A value calculated using sample data. It is the value used to make the decision to reject or fail to reject H_0 .

P-value: A measure of inconsistency between the hypothesized value for a population characteristic and the observed sample. It is the probability, assuming that H_0 is true, of obtaining a test statistic value at least as inconsistent with H_0 as what was observed. The *P*-value is also sometimes called the **observed significance level**.

Example 10.9 Detecting Plagiarism

Understand the context ➤

Plagiarism is a growing concern among college and university faculty members, and many universities are now using software tools to detect student work that is not original.

An Australian university introduced the use of plagiarism detection software in a number of courses. Researchers surveyed 171 students enrolled in those courses (“[Student and Staff Perceptions of the Effectiveness of Plagiarism Detection Software](#),” *Australian Journal of Educational Technology* [2008]: 222–240). In the survey, 58 of the 171 students indicated that they believed that the use of plagiarism-detection software unfairly targeted students.

Assuming it is reasonable to regard the sample as representative of students at this university, does the sample provide convincing evidence that more than one-third of the students at the university believe that the use of plagiarism-detection software unfairly targets students?

With

p = proportion of all students at the university who believe that the use of plagiarism-detection software unfairly targets students

the relevant hypotheses are

$$\begin{aligned} H_0: p &= \frac{1}{3} = 0.33 \\ H_a: p &> 0.33 \end{aligned}$$

The sample proportion is $\hat{p} = \frac{58}{171} = 0.34$.

Formulate a plan ➤

Is the value of \hat{p} enough greater than one-third that we should reject H_0 ?

Because the sample size is large, the statistic

$$z = \frac{\hat{p} - 0.33}{\sqrt{\frac{(0.33)(1 - 0.33)}{n}}}$$

has approximately a standard normal distribution when H_0 is true. The calculated value of the test statistic is

Do the work ➤

$$z = \frac{0.34 - 0.33}{\sqrt{\frac{(0.33)(1 - 0.33)}{171}}} = \frac{0.01}{0.036} = 0.28$$

The probability that a z value at least this inconsistent with H_0 would be observed if H_0 is true is

$$\begin{aligned} P\text{-value} &= P(z \geq 0.28 \text{ when } H_0 \text{ is true}) \\ &= \text{area under the } z \text{ curve to the right of 0.28} \\ &= 1 - 0.6103 \\ &= 0.3897 \end{aligned}$$

Interpret the results ➤

This probability indicates that when $p = 0.33$, it would not be unusual to observe a sample proportion as large as 0.34. When H_0 is true, about 40% of all samples would have a sample proportion as large as or larger than 0.34. This means that a sample proportion of 0.34 is reasonably consistent with the null hypothesis. Although 0.34 is larger than the hypothesized value of $p = 0.33$, chance variation from sample to sample is a plausible explanation for what was observed. There is not convincing evidence that the proportion of students who believe that the use of plagiarism-detection software unfairly targets students is greater than one-third.

As illustrated by Examples 10.8 and 10.9, small P -values indicate that sample results are inconsistent with H_0 , whereas larger P -values are interpreted as meaning that the data are consistent with H_0 and that sampling variability alone is a plausible explanation for what was observed in the sample. As you probably noticed, the two cases examined (P -value ≈ 0 and P -value = 0.3897) were such that a decision between rejecting or not rejecting H_0 was clear-cut. A decision in other cases might not be so obvious. For example, what if the sample had resulted in a P -value of 0.04? Is this unusual enough that H_0 should be rejected? How small does the P -value need to be before H_0 should be rejected?

The answers to these questions depend on the significance level, α (the probability of a Type I error), selected for the test. For example, suppose that we set $\alpha = 0.05$. This implies that the probability of rejecting a true null hypothesis is 0.05. To obtain a test procedure with this probability of Type I error, we would reject the null hypothesis if the sample result is among the most unusual 5% of all samples when H_0 is true. We would reject H_0 if the calculated P -value ≤ 0.05 . If we had selected $\alpha = 0.01$, H_0 would be rejected only if we observed a sample result so extreme that it would be among the most unusual 1% if H_0 is true (which occurs when P -value ≤ 0.01).

A decision to reject or to fail to reject H_0 results from comparing the P -value to the chosen significance level α :

Reject H_0 if P -value $\leq \alpha$.
Fail to reject H_0 if P -value $> \alpha$.

Suppose, for example, that the P -value = 0.0352 and that a significance level of 0.05 is chosen. Then, because

$$P\text{-value} = 0.0352 \leq 0.05 = \alpha$$

H_0 would be rejected. This would not be the case, though, for $\alpha = 0.01$, because then P -value $> \alpha$.

Calculating a *P*-Value for a Large-Sample Test About *p*

The calculation of the *P*-value depends on the form of the inequality in the alternative hypothesis, H_a . For example, suppose that we want to test

$$H_0: p = 0.6 \quad \text{versus} \quad H_a: p > 0.6$$

based on a large sample. The appropriate test statistic is

$$z = \frac{\hat{p} - 0.6}{\sqrt{\frac{(0.6)(1 - 0.6)}{n}}}$$

Values of \hat{p} inconsistent with H_0 and much more consistent with H_a are those much *greater* than 0.6 (because $p = 0.6$ when H_0 is true and $p > 0.6$ when H_0 is false and H_a is true). These values of \hat{p} correspond to *z* values considerably greater than 0.

If $n = 400$ and $\hat{p} = 0.679$, then

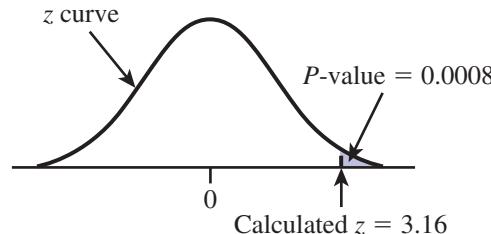
$$z = \frac{0.679 - 0.6}{\sqrt{\frac{(0.6)(1 - 0.6)}{400}}} = \frac{0.079}{0.025} = 3.16$$

The value $\hat{p} = 0.679$ is more than 3 standard deviations larger than what we would have expected if H_0 were true. Then,

$$\begin{aligned} P\text{-value} &= P(z \text{ at least as inconsistent with } H_0 \text{ as } 3.16 \text{ when } H_0 \text{ is true}) \\ &= P(z \geq 3.16) \\ &= \text{area under the } z \text{ curve to the right of } 3.16 \\ &= 1 - 0.9992 \\ &= 0.0008 \end{aligned}$$

This *P*-value is illustrated in Figure 10.1. If H_0 is true, in the long run, only 8 out of 10,000 samples would result in a *z* value as or more extreme than what actually resulted. Most people would consider this quite unusual. Using a significance level of 0.01, we would reject the null hypothesis because $P\text{-value} = 0.0008 \leq 0.01 = \alpha$.

FIGURE 10.1
Calculating a *P*-value.



Now consider testing $H_0: p = 0.3$ versus $H_a: p \neq 0.3$. A value of \hat{p} either much greater than 0.3 or much less than 0.3 is inconsistent with H_0 and provides support for H_a . These \hat{p} values correspond to a *z* value far out in either tail of the *z* curve. If

$$z = \frac{\hat{p} - 0.3}{\sqrt{\frac{(0.3)(1 - 0.3)}{n}}} = 1.75$$

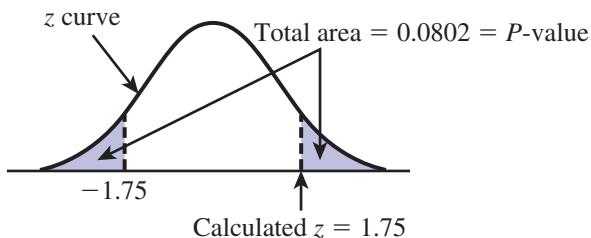
then (as shown in Figure 10.2)

$$\begin{aligned} P\text{-value} &= P(z \text{ value at least as inconsistent with } H_0 \text{ as } 1.75 \text{ when } H_0 \text{ is true}) \\ &= P(z \geq 1.75 \text{ or } z \leq -1.75) \\ &= (\text{z curve area to the right of } 1.75) + (\text{z curve area to the left of } -1.75) \\ &= (1 - 0.9599) + 0.0401 \\ &= 0.0802 \end{aligned}$$

If $z = -1.75$, the P -value in this situation is also 0.0802 because 1.75 and -1.75 are equally inconsistent with H_0 .

FIGURE 10.2

P -value as the sum of two tail areas.



The symmetry of the z curve implies that when the test is two-tailed (the “not equal” alternative), it is not necessary to add two curve areas. Instead,

If z is positive, P -value = 2(area to the right of z).

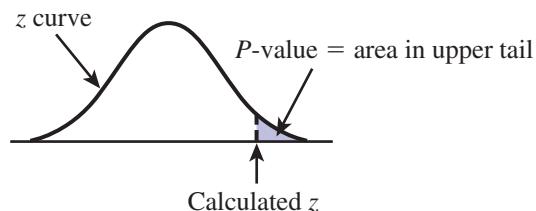
If z is negative, P -value = 2(area to the left of z).

Determination of the P -Value When the Test Statistic Is z

1. Upper-tailed test:

$H_a: p >$ hypothesized value

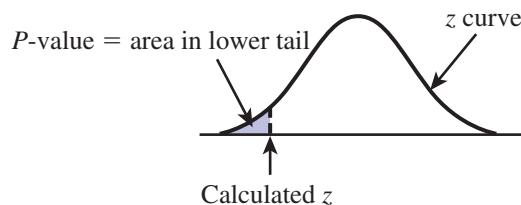
P -value calculated as illustrated:



2. Lower-tailed test:

$H_a: p <$ hypothesized value

P -value calculated as illustrated:

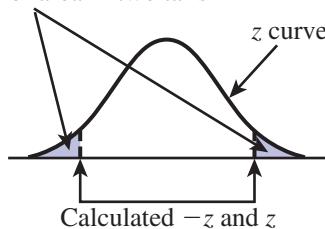


3. Two-tailed test:

$H_a: p \neq$ hypothesized value

P -value calculated as illustrated:

P -value = sum of area in two tails



Example 10.10 Water Conservation

Understand the context ➤

Suppose that in December 2017, a county-wide water conservation campaign was conducted in a particular county. In January 2018, a random sample of 500 homes is selected, and water usage is recorded for each home in the sample. The county supervisors wanted to know whether the data supported the claim that less than half of the households in the county reduced water consumption. The relevant hypotheses are

$$H_0: p = 0.5 \quad \text{versus} \quad H_a: p < 0.5$$

where p is the proportion of all households in the county with reduced water usage.

Formulate a plan ➤

Suppose that the sample results were $n = 500$ and $\hat{p} = 0.440$. Because the sample size is large and this is a lower-tailed test, we can calculate the P -value by first calculating the value of the z test statistic

$$z = \frac{\hat{p} - 0.5}{\sqrt{\frac{(0.5)(1 - 0.5)}{n}}}$$

and then finding the area under the z curve to the left of this z .

Based on the observed sample data,

$$\text{Do the work ➤} \quad z = \frac{0.440 - 0.5}{\sqrt{\frac{(0.5)(1 - 0.5)}{500}}} = \frac{-0.060}{0.0224} = -2.68$$

The P -value is then equal to the area under the z curve and to the left of -2.68 . From the entry in the -2.6 row and 0.08 column of Appendix Table 2, we find that

$$P\text{-value} = 0.0037$$

Interpret the results ➤

Using a 0.01 significance level, we reject H_0 (because $0.0037 \leq 0.01$). This leads us to conclude that there is convincing evidence that the proportion with reduced water usage was less than 0.5. Notice that H_0 would not have been rejected if a *very* small significance level, such as 0.001, had been selected.

Example 10.10 illustrates the calculation of a P -value for a lower-tailed test. The use of P -values in upper-tailed and two-tailed tests will be illustrated in Examples 10.11 and 10.12. But first we summarize the large-sample hypothesis test for a population proportion and introduce a step-by-step procedure for carrying out a hypothesis test.

Summary of Large-Sample z Test for a population proportion p

Null hypothesis: $H_0: p = \text{hypothesized value}$

Test statistic:
$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}}$$

Alternative Hypothesis:

$H_a: p > \text{hypothesized value}$

$H_a: p < \text{hypothesized value}$

$H_a: p \neq \text{hypothesized value}$

P -Value:

Area under z curve to right of calculated z

Area under z curve to left of calculated z

- (1) 2(area to right of z) if z is positive, or
- (2) 2(area to left of z) if z is negative

Assumptions: 1. \hat{p} is the sample proportion from a *random sample*.

2. The *sample size is large*. This test can be used if n satisfies both $n(\text{hypothesized value}) \geq 10$ and $n(1 - \text{hypothesized value}) \geq 10$.

3. If sampling is without replacement, the sample size is no more than 10% of the population size.

We recommend that the following sequence of steps be used when carrying out a hypothesis test.

Steps in a Hypothesis Test

1. Describe the population characteristic of interest.
2. State the null hypothesis H_0 .
3. State the alternative hypothesis H_a .
4. Select the significance level α for the test.
5. Display the test statistic to be used, with substitution of the hypothesized value identified in Step 2.
6. Check to make sure that any assumptions required for the test are reasonable.
7. Calculate all quantities appearing in the test statistic and the value of the test statistic itself.
8. Determine the P -value associated with the observed value of the test statistic.
9. State the conclusion (which is reject H_0 if P -value $\leq \alpha$ and fail to reject H_0 otherwise). The conclusion should then be stated in the context of the problem. The level of significance should be included when stating the conclusion.

Steps 1–4 constitute a statement of the problem, Steps 5–8 give the analysis that leads to a decision, and Step 9 provides the conclusion.

Example 10.11 Love Those Cell Phones . . .

Understand the context ➤

The article “**You Now Have a Shorter Attention Span Than a Goldfish**” (*Time*, May 14, 2015) describes a study of 2000 Canadians over the age of 18 that was carried out by Microsoft. Study participants were asked whether the following statement described them: “When nothing is occupying my attention, the first thing I do is reach for my phone.” Of the study participants in the age group 18 to 24 years old, 77% responded “yes” to this question. Suppose that this group of 18- to 24-year-olds can be considered as a representative sample of Canadians in this age group, and suppose that 800 of the study participants were in this age group.

Does this sample support the claim that more than 75% of all Canadians in this age group would respond “yes” to the given statement?

We answer this question by following the nine steps for carrying out a hypothesis test. We will use a 0.05 significance level for this example.

1. Population characteristic of interest:

p = proportion of all Canadians age 18 to 24 who would say that the given statement describes them

2. Null hypothesis: $H_0: p = 0.75$

3. Alternative hypothesis: $H_a: p > 0.75$ (the percentage of Canadians age 18 to 24 who would say that the given statement describes them is greater than 75%)

4. Significance level: $\alpha = 0.05$

5. Test statistic:

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}} = \frac{\hat{p} - 0.75}{\sqrt{\frac{(0.75)(1 - 0.75)}{n}}}$$

6. Assumptions: This test requires a random sample and a large sample size. The given sample was considered to be representative of Canadians age 18 to 24, and if this is the case, it is reasonable to regard the sample as if it were a random sample. The sample size was $n = 800$. Since $800(0.75) = 600 \geq 10$ and $800(1 - 0.75) = 200 \geq 10$, the large-sample test is appropriate. The population is all Canadians age 18 to 24, so the sample size is small compared to the population size.

Formulate a plan ➤

Do the work ➤ **7. Calculations:** $n = 800$ and $\hat{p} = 0.77$, so

$$z = \frac{\hat{p} - 0.75}{\sqrt{\frac{0.75(1 - 0.75)}{800}}} = \frac{0.77 - 0.75}{\sqrt{\frac{0.75(0.25)}{800}}} = \frac{0.02}{\sqrt{\frac{0.015}{800}}} = \frac{0.02}{0.015} = 1.33$$

8. P-value: This is an upper-tailed test (the inequality in the null hypothesis is $>$), so the *P*-value is the area under the *z* curve and to the right of the calculated *z* value.

$$\begin{aligned} P\text{-value} &= \text{area to the right of } 1.33 \\ &= P(z > 1.33) \\ &= 0.0918 \end{aligned}$$

Interpret the results ➤

9. Conclusion: Because the *P*-value is greater than the selected significance level, we fail to reject the null hypothesis. We conclude that the sample does not provide convincing evidence that more than 75% of Canadians age 18 to 24 think that the statement “When nothing is occupying my attention, the first thing I do is reach for my phone” describes them.

Example 10.12 College Attendance

Understand the context ➤

The report “Average Won’t Do: Performance Trends in Higher Education as a Foundation for Action” (January 2014, files.eric.ed.gov/fulltext/ED574485.pdf, retrieved April 11, 2018) indicates that 53% of students graduating from California high schools go on to attend a 2-year or 4-year college the year after graduation. This college-going rate for students graduating from high schools in the San Francisco Bay Area was estimated to be 50.7%. Suppose that the estimate of 50.7% was based on a representative sample of 1500 graduates from high schools in the San Francisco Bay Area.

Can we reasonably conclude that the proportion of San Francisco Bay Area high school graduates who attended college the year after graduation is different from the statewide proportion of 0.53? We will use the nine-step hypothesis testing procedure and a significance level of $\alpha = 0.01$ to answer this question.

1. p = proportion of all San Francisco Bay Area high school graduates who attended college the year after graduation
2. H_0 : $p = 0.53$
3. H_a : $p \neq 0.53$ (differs from the national proportion)
4. Significance level: $\alpha = 0.01$
5. Test statistic:

$$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{\frac{(\text{hypothesized value})(1 - \text{hypothesized value})}{n}}} = \frac{\hat{p} - 0.53}{\sqrt{\frac{(0.53)(1 - 0.53)}{n}}}$$

6. Assumptions: This test requires a random sample and a large sample size. The sample was said to be representative of San Francisco Bay Area high school graduates, so it is reasonable to regard the sample as if it were a random sample. The sample size was $n = 1500$, and the population size is much larger than the sample size. Because $1500(0.507) \geq 10$ and $1500(1 - 0.507) \geq 10$, the large-sample test is appropriate.

Do the work ➤

7. Calculations: $\hat{p} = 0.507$, so

$$z = \frac{\hat{p} - 0.530}{\sqrt{\frac{(0.530)(1 - 0.530)}{1500}}} = \frac{0.507 - 0.530}{\sqrt{\frac{(0.530)(0.470)}{1500}}} = \frac{-0.023}{\sqrt{\frac{0.013}{1500}}} = \frac{-0.023}{0.013} = -1.77$$

8. *P*-value: The area under the *z* curve to the left of -1.77 is 0.0384 , so P -value = $2(0.0384) = 0.0768$.

Interpret the results ➤

9. Conclusion: At significance level 0.01 , we fail to reject H_0 because the *P*-value is $2(0.0384) = 0.0768$, which is not less than or equal to the significance level of $\alpha = 0.01$. The data do not provide convincing evidence that the actual proportion of San Francisco Bay Area high school graduates who attended college during the year after graduation differs from the statewide proportion.

Most statistical software packages and graphing calculators can calculate and report *P*-values for a variety of hypothesis-testing situations, including the large-sample test for a population proportion. Minitab was used to carry out the test of Example 10.12, and the resulting computer output follows:

Test and CI for One Proportion

Test of $p = 0.53$ vs $p \neq 0.53$

Sample	X	N	Sample p	95% CI	Z-Value	P-Value
1	761	1500	0.507333	(0.482033, 0.532634)	-1.76	0.079

From the Minitab output, $z = -1.76$, and the associated *P*-value is 0.079 . The small difference between the *P*-value here and the one found in Example 10.12 (0.0768) is the result of rounding.

EXERCISES 10.23 - 10.42

- 10.23** Use the definition of the *P*-value to explain the following:

- a. Why H_0 would be rejected if *P*-value = 0.0003
- b. Why H_0 would not be rejected if *P*-value = 0.350

- 10.24** For which of the following *P*-values will the null hypothesis be rejected when performing a test with a significance level of 0.05 :

- | | |
|----------|----------|
| a. 0.001 | d. 0.047 |
| b. 0.021 | e. 0.148 |
| c. 0.078 | |

- 10.25** Pairs of *P*-values and significance levels, α , are given. For each pair, state whether you would reject H_0 at the given significance level.

- a. *P*-value = 0.084 , $\alpha = 0.05$
- b. *P*-value = 0.003 , $\alpha = 0.001$
- c. *P*-value = 0.498 , $\alpha = 0.05$

- 10.26** Pairs of *P*-values and significance levels are given. For each pair, state whether you would reject the null hypothesis at the given significance level.

- a. *P*-value = 0.084 , $\alpha = 0.10$
- b. *P*-value = 0.039 , $\alpha = 0.01$
- c. *P*-value = 0.218 , $\alpha = 0.10$

- 10.27** Let p denote the proportion of students at a particular university that use the fitness center on campus on a regular basis. For a large-sample *z* test of $H_0: p = 0.5$ versus $H_a: p > 0.5$, find the *P*-value associated with each of the given values of the test statistic:

- a. 1.40
- d. 2.45
- b. 0.93
- e. -0.17
- c. 1.96

- 10.28** Assuming a random sample from a large population, for which of the following null hypotheses and sample sizes n is the large-sample *z* test appropriate:

- a. $H_0: p = 0.2$, $n = 25$
- b. $H_0: p = 0.6$, $n = 210$
- c. $H_0: p = 0.9$, $n = 100$
- d. $H_0: p = 0.05$, $n = 75$

- 10.29** In a survey conducted by [CareerBuilder.com](#), employers were asked if they had ever fired an employee for shopping online while at work (“[Cyber Monday Shopping at Work? You’re Not Alone,” November 22, 2016](#), retrieved November 30, 2016). Of the 2379 employers responding to the survey, 262 said they had fired an employee for shopping online while at work. Suppose that it is reasonable to assume that the sample is representative of employers in the United States. Do the sample data provide convincing evidence that more than 10% of employers have fired an employee for shopping online at work? Test the relevant hypotheses using $\alpha = 0.01$.

- 10.30** In a survey of 1000 women age 22 to 35 who work full time, 540 indicated that they would be willing to give up some personal time in order to make more money ([USA TODAY](#), March 4, 2010). The

sample was selected in a way that was designed to produce a sample that was representative of women in the targeted age group.

- a. Do the sample data provide convincing evidence that the majority of women age 22 to 35 who work full-time would be willing to give up some personal time for more money? Test the relevant hypotheses using $\alpha = 0.01$.
 - b. Would it be reasonable to generalize the conclusion from Part (a) to all working women? Explain why or why not.
- 10.31** The paper “[Debt Literacy, Financial Experiences and Over-Indebtedness](#)” (*Social Science Research Network, Working paper W14808, 2008*) included analysis of data from a national sample of 1000 Americans. One question on the survey was:
- “You owe \$3000 on your credit card. You pay a minimum payment of \$30 each month. At an Annual Percentage Rate of 12% (or 1% per month), how many years would it take to eliminate your credit card debt if you made no additional charges?”
- Answer options for this question were: (a) less than 5 years; (b) between 5 and 10 years; (c) between 10 and 15 years; (d) never—you will continue to be in debt; (e) don’t know; and (f) prefer not to answer.
- a. Only 354 of the 1000 respondents chose the correct answer of never. For purposes of this exercise, assume that the sample is representative of adult Americans. Is there convincing evidence that the proportion of adult Americans who can answer this question correctly is less than 0.40 (40%)? Use $\alpha = 0.05$ to test the appropriate hypotheses. (Hint: See Example 10.10.)
 - b. The paper also reported that 37.8% of those in the sample chose one of the wrong answers (a, b, and c) as their response to this question. Is it reasonable to conclude that more than one-third of adult Americans would select a wrong answer to this question? Use $\alpha = 0.05$.
- 10.32** “[Most Like it Hot](#)” is the title of a press release issued by the [Pew Research Center \(March 18, 2009, pewsocialtrends.org/2009/03/18/most-like-it-hot, retrieved April 11, 2018\)](#). The press release states that “by an overwhelming margin, Americans want to live in a sunny place.” This statement is based on data from a nationally representative sample of 2260 adult Americans. Of those surveyed, 1288 indicated that they would prefer to live in a hot climate rather than a cold climate.
- Suppose that you want to determine if there is convincing evidence that a majority of all adult Americans prefer a hot climate over a cold climate.
- a. What hypotheses should be tested to answer this question?
- 10.33** In a survey of 1005 adult Americans, 46% indicated that they were somewhat interested or very interested in having web access in their cars ([USA TODAY, May 1, 2009](#)). Suppose that the marketing manager of a car manufacturer claims that the 46% is based only on a sample and that 46% is close to half, so there is no reason to believe that the proportion of all adult Americans who want car web access is less than 0.50. Is the marketing manager correct in his claim? Provide statistical evidence to support your answer. For purposes of this exercise, assume that the sample can be considered as representative of adult Americans.
- 10.34** The article “[Euthanasia Still Acceptable to Solid Majority in U.S.](#)” ([gallup.com, June 24, 2016, retrieved November 29, 2016](#)) summarized data from a survey of 1025 adult Americans. When asked if doctors should be able to end a terminally ill patient’s life by painless means if requested to do so by the patient, 707 of those surveyed responded yes. For purposes of this exercise, assume that it is reasonable to regard this sample as a random sample of adult Americans. Suppose that you want to use the data from this survey to decide if there is convincing evidence that more than two-thirds of adult Americans believe that doctors should be able to end a terminally ill patient’s life if requested to do so by the patient.
- a. What hypotheses should be tested to answer this question?
 - b. The P -value for this test is 0.058. What conclusions would you reach if $\alpha = 0.05$?
 - c. Would you have reached a different conclusion if $\alpha = 0.10$? Explain.
- 10.35** The report “[Digital Democracy Survey](#)” ([Deloitte Development LLC, 2016, deloitte.com/us/en.html, retrieved November 30, 2016](#)) says that 69% of U.S. teens access the social media from a mobile phone. Suppose you plan to select a random sample of students at the local high school and will ask each student in the sample if he or she accesses social media from a mobile phone. You want to determine if there is evidence that the proportion of students at the high school who access social media using a mobile phone differs from the national figure of 0.69 given in the Nielsen report. What hypotheses should you test?
- 10.36** The article “[How to Block Nuisance Calls](#)” ([The Guardian, November 7, 2015](#)) reported that in a survey of mobile phone users, 70% of those surveyed said they had received at least one nuisance call to their mobile phone in the last month. Suppose that

this estimate was based on a representative sample of 600 mobile phone users. These data can be used to determine if there is evidence that more than two-thirds all mobile phone users have received at least one nuisance call in the last month. The large-sample test for a population proportion was used to test $H_0: p = 0.667$ versus $H_a: p > 0.667$. The resulting P -value was 0.043. Using a significance level of 0.05, the null hypothesis was rejected.

- Based on the hypothesis test, what can you conclude about the proportion of mobile phone users who received at least one nuisance call on their mobile phones within the last month?
- Is it reasonable to say that the data provide strong support for the alternative hypothesis?
- Is it reasonable to say that the data provide strong support against the null hypothesis?

10.37 The article “Facebook Use and Academic Performance Among College Students” (*Computers in Human Behavior* [2015]: 265–272) estimated that 87% percent of students at a large public university in California who are Facebook users update their status at least two times a day. Suppose that you plan to select a random sample of 400 students at your college. You will ask each student in the sample if he or she is a Facebook user and if they update their status at least two times a day. You plan to use the resulting data to decide if there is evidence that the proportion for your college is different from the proportion reported in the article for the college in California. What hypotheses should you test?

10.38 The article “Public Acceptability in the UK and the USA of Nudging to Reduce Obesity: The Example of Reducing Sugar-Sweetened Beverages” (*PLOS One*, June 8, 2016) describes a survey in which each person in a representative sample of 1082 adult Americans was asked about whether they would find different types of interventions acceptable to reduce consumption of sugary beverages. When asked about a tax on sugary beverages, 459 of the people in the sample said they thought that this would be an acceptable intervention. These data were used to test $H_0: p = 0.5$ versus $H_a: p < 0.5$ and the null hypothesis was rejected. Based on the hypothesis test, what can you conclude about the proportion of adult Americans who think that taxing sugary beverages is an acceptable intervention to reduce consumption of sugary beverages?

10.39 The article “Cops Get Screened for Digital Dirt” (*USA TODAY*, November 12, 2010) summarizes a report of law enforcement agencies regarding the use of social media to screen applicants for employment. The report was based on a survey of 728 law enforcement agencies. One question on the survey asked if the agency routinely reviewed applicants’

social media activity during background checks. For purposes of this exercise, suppose that the 728 agencies were selected at random, and that you want to use the survey data to decide if there is convincing evidence that more than 25% of law enforcement agencies review applicants’ social media activity as part of routine background checks.

- The sampling distribution of \hat{p} describes the behavior of \hat{p} when random samples are selected from a particular population. Describe the shape, center, and spread of the sampling distribution of \hat{p} for samples of size 728 if the null hypothesis $H_0: p = 0.25$ is true.
- Would you be surprised to observe a sample proportion of $\hat{p} = 0.27$ for a sample of size 728 if the null hypothesis $H_0: p = 0.25$ is true? Explain why or why not.
- Would you be surprised to observe a sample proportion of $\hat{p} = 0.31$ for a sample of size 728 if the null hypothesis $H_0: p = 0.25$ is true? Explain why or why not.

10.40 Refer back to the previous exercise. The actual sample proportion observed in the study was $\hat{p} = 0.33$. Based on this sample proportion, is there convincing evidence that more than 25% of law enforcement agencies review social media activity as part of background checks, or is this sample proportion consistent with what you would expect to see when the null hypothesis is true?

10.41 The report “2007 Electronic Monitoring & Surveillance Survey: Many Companies Monitoring, Recording, Videotaping—and Firing—Employees” (*American Management Association*, 2007) summarized the results of a survey of 304 U.S. businesses. Of these companies, 201 indicated that they monitor employees’ web site visits. For purposes of this exercise, assume that it is reasonable to regard this sample as representative of businesses in the United States.

- Is there sufficient evidence to conclude that more than 60% of U.S. businesses monitor employees’ web site visits? Test the appropriate hypotheses using a significance level of 0.01.
- Is there sufficient evidence to conclude that a majority of U.S. businesses monitor employees’ web site visits? Test the appropriate hypotheses using a significance level of 0.01.

10.42 The United States Elections Project (electproject.org/2016g, retrieved January 22, 2017) reported that 57.8% of registered voters in California voted in the 2016 presidential election and that this was less than the national percentage of 60.0%. Explain why it is not necessary to carry out a hypothesis test to determine if the proportion of registered voters in California who voted in the election is less than the national proportion of 0.600.

SECTION 10.4 Hypothesis Tests for a Population Mean

We now turn our attention to developing a method for testing hypotheses about a population mean. The test procedures in this case are based on the same two results that led to the z and t confidence intervals in Chapter 9:

1. When n is large or the population distribution is approximately normal, then

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has approximately a standard normal distribution.

2. When n is large or the population distribution is approximately normal, then

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

has approximately a t distribution with $df = n - 1$.

This means that if we are interested in testing a null hypothesis of the form

$$H_0: \mu = \text{hypothesized value}$$

and n is large or the population distribution is approximately normal, one of the following z or t test statistics can be used:

Case 1: σ known

$$\text{Test statistic: } z = \frac{\bar{x} - \text{hypothesized value}}{\frac{\sigma}{\sqrt{n}}}$$

P -value: Calculated as an area under the z curve

Case 2: σ unknown

$$\text{Test statistic: } t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}}$$

P -value: Calculated as an area under the t curve with $df = n - 1$

Because it is rarely the case that σ , the population standard deviation, is known, we focus on the test procedure for the case in which σ is unknown.

When testing a hypothesis about a population mean, the null hypothesis specifies a particular hypothesized value for μ . We write the null hypothesis as $H_0: \mu = \text{hypothesized value}$. The alternative hypothesis has one of the following three forms, depending on the research question being addressed:

$$\begin{aligned} H_a: \mu &> \text{hypothesized value} \\ H_a: \mu &< \text{hypothesized value} \\ H_a: \mu &\neq \text{hypothesized value} \end{aligned}$$

If n is large or if the population distribution is approximately normal, the test statistic

$$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}}$$

can be used. For example, if the null hypothesis is $H_0: \mu = 100$, the test statistic becomes

$$t = \frac{\bar{x} - 100}{\frac{s}{\sqrt{n}}}$$

Consider the alternative hypothesis $H_a: \mu > 100$, and suppose that a sample of size $n = 24$ results in $\bar{x} = 104.20$ and $s = 8.23$. Then the test statistic value is

$$t = \frac{104.20 - 100}{\frac{8.23}{\sqrt{24}}} = \frac{4.20}{1.6799} = 2.50$$

Because this is an upper-tailed test, the P -value is the area under an appropriate t curve (here with $df = 24 - 1 = 23$) to the right of 2.50.

Appendix Table 4 gives t curve tail areas. Each column of the table is for a different number of degrees of freedom: 1, 2, 3, ..., 30, 35, 40, 60, 120, and a last column for $df = \infty$, which is the same as for the z curve. The table gives the area under each t curve to the right of values ranging from 0.0 to 4.0 in increments of 0.1. Part of this table appears in Figure 10.3. For example,

$$\begin{aligned} \text{area under the } 23\text{-df } t \text{ curve to the right of } 2.5 &= 0.010 \\ &= P\text{-value for an upper-tailed } t \text{ test} \end{aligned}$$

Suppose that $t = -2.7$ for a lower-tailed test based on 23 df. Then, because each t curve is symmetric about 0,

$$P\text{-value} = \text{area to the left of } -2.7 = \text{area to the right of } 2.7 = 0.006$$

As is the case for z tests, we double the tail area to obtain the P -value for two-tailed t tests. This means that if $t = 2.6$ or if $t = -2.6$ for a two-tailed t test with 23 df, then

$$P\text{-value} = 2(0.008) = 0.016$$

Once past 30 df, the tail areas change very little, so the last column (∞) in Appendix Table 4 provides a good approximation.

FIGURE 10.3

Part of Appendix Table 4: t curve tail areas.

df \ t	1	2	...	22	23	24	...	60	120
0.0									
0.1									
⋮				⋮	⋮	⋮			
2.5		010	.010	.010	...		
2.6		008	.008	.008	...		
2.7		007	.006	.006	...		
2.8		005	.005	.005	...		
⋮				⋮	⋮	⋮			
4.0									

Area under 23-df
 t curve to right of 2.7

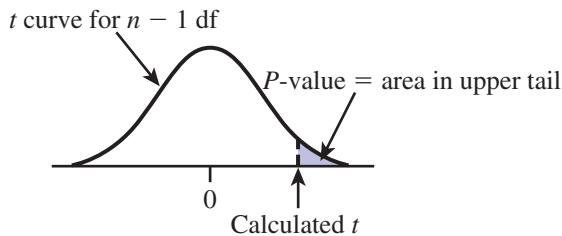
A graphing calculator or a statistics software package can also be used to compute P -values for t tests.

The following two boxes show how the P -value is obtained as a t curve area and give a general description of the test procedure.

Finding P -Values for a t Test

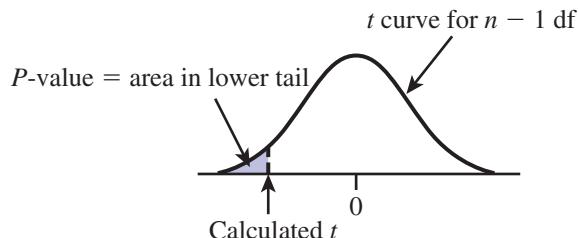
1. Upper-tailed test:

$$H_a: \mu > \text{hypothesized value}$$



2. Lower-tailed test:

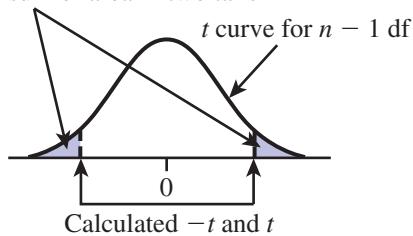
$$H_a: \mu < \text{hypothesized value}$$



3. Two-tailed test:

$$H_a: \mu \neq \text{hypothesized value}$$

$$P\text{-value} = \text{sum of area in two tails}$$



Appendix Table 4 gives upper-tail t curve areas to the right of values 0.0, 0.1, . . . , 4.0. These areas are P -values for upper-tailed tests and, by symmetry, also for lower-tailed tests. Doubling an area gives the P -value for a two-tailed test.

The One-Sample t Test for a Population Mean

Null hypothesis: $H_0: \mu = \text{hypothesized value}$

Test statistic: $t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}}$

Alternative Hypothesis:

$$H_a: \mu > \text{hypothesized value}$$

$H_a: \mu < \text{hypothesized value}$

$$H_a: \mu \neq \text{hypothesized value}$$

P-Value:

Area to the right of calculated t under t curve with $\text{df} = n - 1$

Area to the left of calculated t under t curve with $\text{df} = n - 1$

(1) 2(area to the right of t) if t is positive, or

(2) 2(area to the left of t) if t is negative

Assumptions: 1. \bar{x} and s are the sample mean and sample standard deviation from a *random sample*.

2. The *sample size is large* (generally $n \geq 30$) or the population distribution is at least approximately normal.

Example 10.13 Time Perception During Nicotine Withdrawal

- The authors of the paper “Sex Differences in Time Perception During Smoking Abstinence” (*Nicotine and Tobacco Research* [2015]: 449–454) carried out a study to investigate how nicotine withdrawal affects time perception and decision-making. In this study, 21 male smokers were asked to abstain from smoking for 24 hours. They were then shown a demo screen with a green cross that changed to a red cross after a period of time. They were then shown the green cross and then asked to indicate when they thought the same amount of time had passed as in the demo. This process was repeated for 15 more trials, with varying times, and then a time discrimination score was calculated as follows for each of the 21 men:

$$\text{time discrimination score} = \frac{\text{total estimated time}}{\text{total actual time of demos}}$$

A time discrimination score greater than 1 would result for someone who tended to overestimate the actual times, and a score of less than 1 would result for someone who tended to underestimate the actual times.

Suppose that the data were as follows (these data are artificial but are consistent with summary quantities given in the paper):

1.12	1.03	1.09	1.03	1.09	0.97	0.98	1.20	1.16	1.03	1.10
1.11	0.98	1.02	1.20	0.96	0.78	1.05	0.90	1.08	0.95	

These data were used to calculate the sample mean and standard deviation:

$$n = 21$$

$$\bar{x} = 1.04$$

$$s = 0.100$$

Suppose that it is reasonable to consider the people in this sample as representative of male smokers in general. These data can be used to determine if there is evidence that male smokers tend to overestimate time after having abstained from smoking for 24 hours.

With μ representing the mean time discrimination score for male smokers who have not smoked for 24 hours, we can answer this question by testing

$$H_0: \mu = 1 \text{ (no consistent tendency to overestimate time)}$$

versus

$$H_a: \mu > 1 \text{ (tendency for time to be overestimated)}$$

The null hypothesis is rejected only if there is convincing evidence that $\mu > 1$. The observed sample mean, 1.04, is greater than 1, but can a sample mean as large as this be explained by chance variation from one sample to another when $\mu = 1$?

To answer this question, we carry out a hypothesis test with a significance level of 0.05 using the nine-step procedure described in Section 10.3.

- Population characteristic of interest:

Understand the context ➤

μ = mean time discrimination score for male smokers who have not smoked for 24 hours

- Null hypothesis: $H_0: \mu = 1$

- Alternative hypothesis: $H_a: \mu > 1$

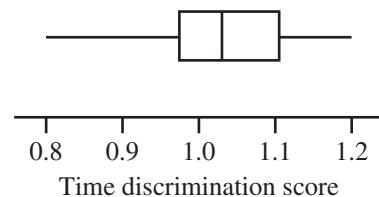
- Significance level: $\alpha = 0.05$

Formulate a plan ➤

- Test statistic: $t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - 1}{\frac{s}{\sqrt{n}}}$

- Assumptions: The one-sample t test requires a random sample and either a large sample size or a normal population distribution. The problem statement indicated that it was reasonable to consider this sample as representative of male smokers in general, so it is reasonable to regard the sample as if it were a random

sample. Because the sample size is only 21, for the t test to be appropriate, we must be willing to assume that the population distribution of time discrimination scores is at least approximately normal. Is this reasonable? The following graph is a boxplot of the data:



Although the boxplot is not perfectly symmetric, it does not appear to be too skewed and there are no outliers, so the use of the t test is reasonable.

7. Calculations: $n = 21$, $\bar{x} = 1.04$, and $s = 0.100$, so

Do the work ►

$$t = \frac{1.04 - 1}{\frac{0.100}{\sqrt{21}}} = \frac{0.04}{0.022} = 1.82$$

8. P -value: This is an upper-tailed test (the inequality in H_a is “greater than”), so the P -value is the area to the right of the calculated t value. Because $df = 21 - 1 = 20$, we can use the $df = 20$ column of Appendix Table 4 to find the P -value. With $t = 1.82$, we obtain P -value = area to the right of 1.82. The area to the right of 1.82 is approximately 0.043 (using 1.8, which is the closest value to 1.82 that appears in the table). The P -value ≈ 0.043 .

- Interpret the results ► 9. Conclusion: Because P -value $\leq \alpha$, we reject H_0 at the 0.05 level of significance. It would be unlikely to see a sample mean this extreme as a result of just chance variation when H_0 is true. There is convincing evidence that the mean time discrimination score is greater than 1. This is evidence that male smokers who have not smoked for 24 hours tend to overestimate time.

A statistical software package or a graphing calculator could also be used to carry out the calculate step of the hypothesis test process. For example, in the accompanying Minitab output, you can see that the value of the test statistic is given as $t = 1.83$ and the associated P -value is given as 0.041.

One-Sample T

Test of $\mu = 1$ vs > 1						
N	Mean	StDev	SE Mean	95% Lower Bound	T	P
21	1.0400	0.1000	0.0218	1.0024	1.83	0.041

The difference between what is given in the Minitab output and what was obtained by calculating the value of the test statistic and using Appendix Table 4 is due to differences in rounding in the calculations.

Example 10.14 Goofing Off at Work

- Many employers are concerned about employees wasting time by surfing the Internet and e-mailing friends during work hours. The article [“Who Goofs Off 2 Hours a Day? Most Workers, Survey Says”](#) (*San Luis Obispo Tribune*, August 3, 2006) summarized data from a large sample of workers. Suppose that the CEO of a large company wants to determine whether the mean wasted time during an 8-hour work day for employees of her company is less than the mean of 120 minutes reported in the article. Each person in a random sample of 10 employees was contacted and asked about daily wasted time

at work (in minutes). Participants would probably have to be guaranteed anonymity to obtain truthful responses. The resulting data are:

108 112 117 130 111 131 113 113 105 128

Summary quantities are $n = 10$, $\bar{x} = 116.80$, and $s = 9.45$.

Do these data provide evidence that the mean wasted time for this company is less than 120 minutes? To answer this question, carry out a hypothesis test with $\alpha = 0.05$.

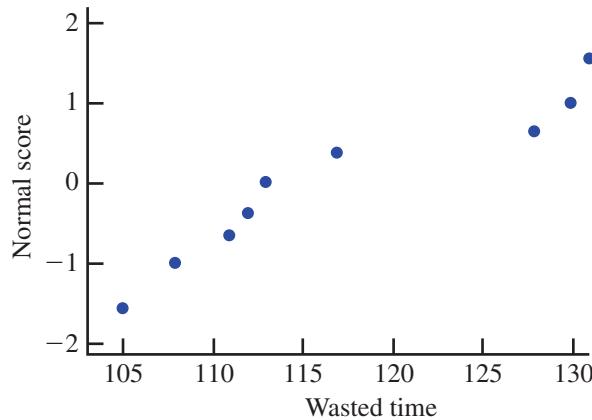
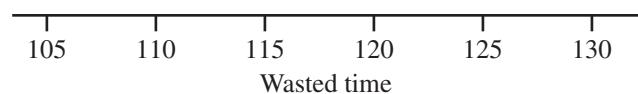
Understand the context ➤

1. μ = mean daily wasted time for employees of this company
2. $H_0: \mu = 120$
3. $H_a: \mu < 120$
4. $\alpha = 0.05$

Formulate a plan ➤

$$5. t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - 120}{\frac{s}{\sqrt{n}}}$$

6. This test requires a random sample and either a large sample or a normal population distribution. The given sample was a random sample of employees. Because the sample size is small, we must be willing to assume that the population distribution of times is at least approximately normal. The accompanying normal probability plot appears to be reasonably straight, and although the normal probability plot and the boxplot reveal some skewness in the sample, there are no outliers.



Based on these observations, it is plausible that the population distribution is approximately normal, so we proceed with the t test.

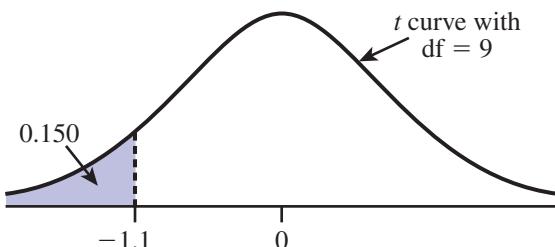
Do the work ➤

$$7. \text{ Test statistic: } t = \frac{116.80 - 120}{\frac{9.45}{\sqrt{10}}} = -1.07$$

8. From the $df = 9$ column of Appendix Table 4 and by rounding the test statistic value to -1.1 , we get

P -value = area to the left of -1.1 = area to the right of $1.1 = 0.150$

as shown:



Interpret the results ➤

9. Because the P -value $> \alpha$, we fail to reject H_0 . There is not sufficient evidence to conclude that the mean wasted time per 8-hour work day for employees at this company is less than 120 minutes.

Minitab could also have been used to carry out the test, as shown in the output below.

One-Sample T: Wasted Time

Test of $\mu = 120$ vs < 120

Variable	N	Mean	StDev	SE Mean	Upper Bound	T	P
Wasted Time	10	116.800	9.449	2.988	122.278	-1.07	0.156

Although we had to round the computed t value to -1.1 to use Appendix Table 4, Minitab was able to compute the P -value corresponding to the actual value of the test statistic, which was P -value = 0.156.

Example 10.15 Cricket Love



Dynamic Graphics Group/Creatas/Alamy Stock Photo

The article “Well-Fed Crickets Bowl Maidens Over” (*Nature Science Update*, February 11, 1999) reported that female field crickets are attracted to males that have high chirp rates and hypothesized that chirp rate is related to nutritional status. The usual chirp rate for male field crickets was reported to vary around a mean of 60 chirps per second. To investigate whether chirp rate was related to nutritional status, investigators fed male crickets a high protein diet for 8 days and the chirp rate was measured.

The mean chirp rate for the crickets on the high protein diet was reported to be 109 chirps per second. Is this convincing evidence that the mean chirp rate for crickets on a high protein diet is greater than 60 (which would then imply an advantage in attracting the ladies)?

Suppose that the sample size and sample standard deviation are $n = 32$ and $s = 40$. Let’s test the relevant hypotheses with $\alpha = 0.01$.

Understand the context ➤

1. μ = mean chirp rate for crickets on a high protein diet
2. $H_0: \mu = 60$
3. $H_a: \mu > 60$
4. $\alpha = 0.01$

Formulate a plan ➤

$$5. t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - 60}{\frac{s}{\sqrt{n}}}$$

6. This test requires a random sample and either a large sample or a normal population distribution. Because the sample size is large ($n = 32$), it is reasonable to proceed with the t test as long as we are willing to consider the 32 male field

crickets in this study as if they were a random sample from the population of male field crickets.

Do the work ➤

7. Test statistic: $t = \frac{109 - 60}{\sqrt{\frac{40}{32}}} = \frac{49}{7.07} = 6.93$

8. This is an upper-tailed test, so the P -value is the area under the t curve with $df = 31$ and to the right of 6.93. From Appendix Table 4, P -value ≈ 0 .

Interpret the results ➤

9. Because P -value ≈ 0 , which is less than the significance level, α , we reject H_0 . There is convincing evidence that the mean chirp rate is greater for male field crickets that eat a high protein diet.

Statistical Versus Practical Significance

Carrying out a hypothesis test amounts to deciding whether the value obtained for the test statistic is plausible when H_0 is true. When the value of the test statistic leads to rejection of H_0 , we say that the result is **statistically significant** at the chosen significance level α . The finding of statistical significance means that the observed deviation from what was expected when H_0 is true cannot reasonably be explained by only chance variation. However, statistical significance is not the same as concluding that the true situation differs from what the null hypothesis states in any practical sense. That is, even after H_0 has been rejected, the data may suggest that there is no *practical* difference between the actual value of the population characteristic and what the null hypothesis states that value to be. This is illustrated in Example 10.16.

Example 10.16 “Significant” but Unimpressive Test Score Improvement

Understand the context ➤

Let μ denote the average score on a standardized test for all children in a large school district. The average score for all children in the United States is 100. District administrators are interested in testing $H_0: \mu = 100$ versus $H_a: \mu > 100$ using a significance level of 0.001.

A sample of 2500 children resulted in the values $n = 2500$, $\bar{x} = 101.0$, and $s = 15.0$. Then

$$t = \frac{101.10 - 100}{\sqrt{\frac{15}{2500}}} = \frac{1.10}{0.3} = 3.33$$

This is an upper-tailed test, so (using the z column of Appendix Table 4 because $df = 2499$) P -value = area to the right of 3.33 ≈ 0.000 . Because P -value < 0.001 , we reject H_0 . There is convincing evidence that the mean score for this region is greater than 100.

Interpret the results ➤

However, with $n = 2500$, the point estimate $\bar{x} = 101.0$ is likely to be very close to the actual value of μ . This means that it looks as though H_0 was rejected because $\mu \approx 101$ rather than 100. From a practical point of view, a 1-point difference may not be important. A statistically significant result does not necessarily mean that there are any practical consequences.

EXERCISES 10.43 - 10.62

● Data set available online

- 10.43** Give as much information as you can about the P -value of a t test in each of the following situations:
- Upper-tailed test, $df = 8$, $t = 2.0$
 - Upper-tailed test, $n = 14$, $t = 3.2$
 - Lower-tailed test, $df = 10$, $t = -2.4$
 - Lower-tailed test, $n = 22$, $t = -4.2$

- 10.44** Give as much information as you can about the P -value of a t test in each of the following situations:
- Two-tailed test, $df = 15$, $t = -1.6$
 - Two-tailed test, $n = 16$, $t = 1.6$
 - Two-tailed test, $n = 16$, $t = 6.3$

- 10.45** Give as much information as you can about the P -value of a t test in each of the following situations:
- Two-tailed test, $df = 9, t = 0.73$
 - Upper-tailed test, $df = 10, t = -0.5$
 - Lower-tailed test, $n = 20, t = -2.1$
 - Lower-tailed test, $n = 20, t = -5.1$
 - Two-tailed test, $n = 40, t = 1.7$
- 10.46** Paint used to paint lines on roads must reflect enough light to be clearly visible at night. Let μ denote the mean reflectometer reading for a new type of paint. A test of $H_0: \mu = 20$ versus $H_a: \mu > 20$ based on a sample of 15 observations resulted in $t = 3.2$. What conclusion is appropriate at each of the following significance levels?
- $\alpha = 0.05$
 - $\alpha = 0.01$
 - $\alpha = 0.001$
- 10.47** A certain pen has been designed so that actual mean writing lifetime under controlled conditions (involving the use of a writing machine) is at least 10 hours. A random sample of 18 pens is selected, the writing lifetime of each is determined, and a normal probability plot of the resulting data supports the use of a one-sample t test. The relevant hypotheses are $H_0: \mu = 10$ versus $H_a: \mu < 10$.
- If $t = -2.3$ and $\alpha = 0.05$, what conclusion is appropriate?
 - If $t = -1.83$ and $\alpha = 0.01$, what conclusion is appropriate?
 - If $t = 0.47$, what conclusion is appropriate?
- 10.48** The average diameter of ball bearings of a certain type is supposed to be 0.5 inch. What conclusion is appropriate when testing $H_0: \mu = 0.5$ versus $H_a: \mu \neq 0.5$ inch each of the following situations:
- $n = 13, t = 1.6, \alpha = 0.05$
 - $n = 13, t = -1.6, \alpha = 0.05$
 - $n = 25, t = -2.6, \alpha = 0.01$
 - $n = 25, t = -3.6$
- 10.49** The paper “[Playing Active Video Games Increases Energy Expenditure in Children](#)” (*Pediatrics* [2009]: 534–539) describes an interesting investigation of the possible cardiovascular benefits of active video games. Mean heart rate for healthy boys age 10 to 13 after walking on a treadmill at 2.6 km/hour for 6 minutes is 98 beats per minute (bpm). For each of 14 boys, heart rate was measured after 15 minutes of playing Wii Bowling. The resulting sample mean and standard deviation were 101 bpm and 15 bpm, respectively. For purposes of this exercise, assume that it is reasonable to regard the sample of boys as representative of boys age 10 to 13 and that the distribution of heart rates after 15 minutes of Wii Bowling is approximately normal.
- Does the sample provide convincing evidence that the mean heart rate after 15 minutes of Wii Bowling is different from the known mean heart rate after 6 minutes walking on the treadmill?
- Carry out a hypothesis test using $\alpha = 0.01$.
- The known resting mean heart rate for boys in this age group is 66 bpm. Is there convincing evidence that the mean heart rate after Wii Bowling for 15 minutes is higher than the known mean resting heart rate for boys of this age? Use $\alpha = 0.01$. (Hint: See Example 10.15.)
 - Based on the outcomes of the tests in Parts (a) and (b), write a paragraph comparing the benefits of treadmill walking and Wii Bowling in terms of raising heart rate over the resting heart rate.
- 10.50** A study of fast-food intake is described in the paper “[What People Buy from Fast-Food Restaurants](#)” (*Obesity* [2009]: 1369–1374). Adult customers at three hamburger chains (McDonald’s, Burger King, and Wendy’s) at lunchtime in New York City were approached as they entered the restaurant and asked to provide their receipt when exiting. The receipts were then used to determine what was purchased and the number of calories consumed was determined. In all, 3857 people participated in the study. The sample mean number of calories consumed was 857 and the sample standard deviation was 677.
- The sample standard deviation is quite large. What does this tell you about number of calories consumed in a hamburger-chain lunchtime fast-food purchase in New York City?
 - Given the values of the sample mean and standard deviation and the fact that the number of calories consumed can’t be negative, explain why it is *not* reasonable to assume that the distribution of calories consumed is normal.
 - Based on a recommended daily intake of 2000 calories, the online [Healthy Dining Finder](#) (healthydiningfinder.com) recommends a target of 750 calories for lunch. Assuming that it is reasonable to regard the sample of 3857 fast-food purchases as representative of all hamburger-chain lunchtime purchases in New York City, carry out a hypothesis test to determine if the sample provides convincing evidence that the mean number of calories in a New York City hamburger-chain lunchtime purchase is greater than the lunch recommendation of 750 calories. Use $\alpha = 0.01$. (Hint: See Example 10.15.)
- 10.51** Refer to the study and hypothesis test of the previous exercise.
- Would it be reasonable to generalize the conclusion of the test in Part (c) of the previous exercise to the lunchtime fast-food purchases of all adult Americans? Explain why or why not.
 - Explain why it is better to use the customer receipt to determine what was ordered rather than just asking a customer leaving the restaurant what he or she purchased.

- c. Do you think that asking customers before they order to provide a receipt when they leave could have introduced a potential bias? Explain.
- 10.52** The report “[2016 Salary Survey Executive Summary](#)” ([National Association of Colleges and Employers, naceweb.org/uploadedfiles/files/2016/publications/executive-summary/2016-nace-salary-survey-fall-executive-summary.pdf](#), retrieved December 24, 2016) states that the mean yearly salary offer for students graduating with mathematics and statistics degrees in 2016 is \$62,985. Suppose that a random sample of 50 mathematics and statistics graduates at a large university who received job offers resulted in a mean offer of \$63,500 and a standard deviation of \$3300. Do the sample data provide strong support for the claim that the mean salary offer for mathematics and statistics graduates of this university is greater than the 2016 national average of \$62,985? Test the relevant hypotheses using $\alpha = 0.05$.
- 10.53** ● [The Economist](#) collects data each year on the price of a Big Mac in various countries around the world. The price of a Big Mac for a sample of McDonald’s restaurants in Europe in July 2016 resulted in the following Big Mac prices (after conversion to U.S. dollars):
- | | | | | | |
|------|------|------|------|------|------|
| 4.44 | 3.15 | 2.42 | 3.96 | 4.35 | 4.51 |
| 4.17 | 3.69 | 4.62 | 3.80 | 3.36 | 3.85 |
- The mean price of a Big Mac in the U.S. in July 2016 was \$5.04. For purposes of this exercise, assume it is reasonable to regard the sample as representative of European McDonald’s restaurants. Does the sample provide convincing evidence that the mean July 2016 price of a Big Mac in Europe is less than the reported U.S. price? Test the relevant hypotheses using $\alpha = 0.05$.
- 10.54** The report “[Majoring in Money: How American College Students Manage Their Finances](#)” ([Sallie Mae, 2016, news.salliemae.com](#), retrieved December 24, 2016) includes data from a survey of college students. Each person in a representative sample of 793 college students was asked if they had one or more credit cards and if so, whether they paid their balance in full each month. There were 500 who paid in full each month. For this sample of 500 students, the sample mean credit card balance was reported to be \$825. The sample standard deviation of the credit card balances for these 500 students was not reported, but for purposes of this exercises, suppose that it was \$200. Is there convincing evidence that college students who pay their credit card balance in full each month have mean balance that is lower than \$906, the value reported for all college students with credit cards? Carry out a hypothesis test using a significance level of 0.01.
- 10.55** The authors of the paper “[Changes in Quantity, Spending, and Nutritional Characteristics of Adult Adolescent and Child Urban Corner Store Purchases After an Environmental Intervention](#)” ([Preventative Medicine \[2015\]: 81–85](#)) wondered if increasing the availability of healthy food options would also increase the amount people spend at the corner store. They collected data from a representative sample of 5949 purchases at corner stores in Philadelphia after the stores increased their healthy food options. The sample mean amount spent for this sample of purchases was \$2.86 and the sample standard deviation was \$5.40.
- a. Notice that for this sample, the sample standard deviation is greater than the sample mean. What does this tell you about the distribution of purchase amounts?
- b. Before the stores increased the availability of healthy foods, the population mean total amount spent per purchase was thought to be about \$2.80. Do the data from this study provide convincing evidence that the population mean amount spent per purchase is greater after the change to increase healthy food options? Carry out a hypothesis test with a significance level of 0.05.
- 10.56** ● Medical research has shown that repeated wrist extension beyond 20 degrees increases the risk of wrist and hand injuries. Each of 24 students at Cornell University used a proposed new computer mouse design. While using the mouse, each student’s wrist extension was recorded. Data consistent with summary values given in the paper “[Comparative Study of Two Computer Mouse Designs](#)” ([Cornell Human Factors Laboratory Technical Report RP7992](#)) are given. Use these data to test the hypothesis that the mean wrist extension for people using this new mouse design is greater than 20 degrees. Are any assumptions required in order for it to be appropriate to generalize the results of your test to the population of Cornell students? To the population of all university students? (Hint: See Example 10.13.)
- | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|
| 27 | 28 | 24 | 26 | 27 | 25 | 24 | 24 | 24 | 25 | 28 |
| 22 | 25 | 24 | 28 | 27 | 26 | 31 | 25 | 28 | 27 | 27 |
- 10.57** A comprehensive study conducted by the National Institute of Child Health and Human Development tracked more than 1000 children from an early age through elementary school ([The New York Times, November 1, 2005](#)). The study concluded that children who spent more than 30 hours a week in child care before entering school tended to score higher in math and reading when they were in the third grade. The researchers cautioned that the findings should not be a cause for alarm because the effects of child care were found to be small.
- Explain how the difference between the sample mean math score for third graders who spent long hours in child care and the known overall mean for third graders could be small but the researchers

could still reach the conclusion that the mean for the child care group is significantly higher than the overall mean for third graders. (Hint: See discussion of statistical versus practical significance.)

- 10.58** In a study of media use, each person in a large representative sample of male Canadian high school students was asked how much time they spent playing video or computer games (in minutes per day). The sample mean was 123.4 minutes and the sample standard deviation was 117.1 minutes.

- Based on the given sample mean and standard deviation, is it reasonable to think that the distribution of time spent playing video or computer games for the population of male Canadian high school students is approximately normal? Explain why or why not.
- Suppose you wanted to use the sample data to decide if there is evidence that the average time spent playing video or computer games for male Canadian high school students is greater than 2 hours (120 minutes). What would you need to know to determine if the one-sample t test is an appropriate method?

- 10.59** Refer to the study description and sample statistics given in the previous exercise.

- Suppose that the sample size was 500. Carry out a hypothesis test to decide if there is evidence that the average time spent playing video or computer games for male Canadian high school students is greater than 2 hours. Use a significance level of 0.05.
- Now suppose that the sample standard deviation had been 37.1 rather than 117.1. Carry out a hypothesis test to decide if there is evidence that the average time spent playing video or computer games for male Canadian high school students is greater than 2 hours. Use a significance level of 0.05.
- Explain why the null hypothesis was rejected in the test of Part (d) but not in the test of Part (c).

- 10.60** The paper titled “Music for Pain Relief” (*The Cochrane Database of Systematic Reviews*, April 19, 2006) concluded, based on a review of 51 studies of the effect of music on pain intensity, that

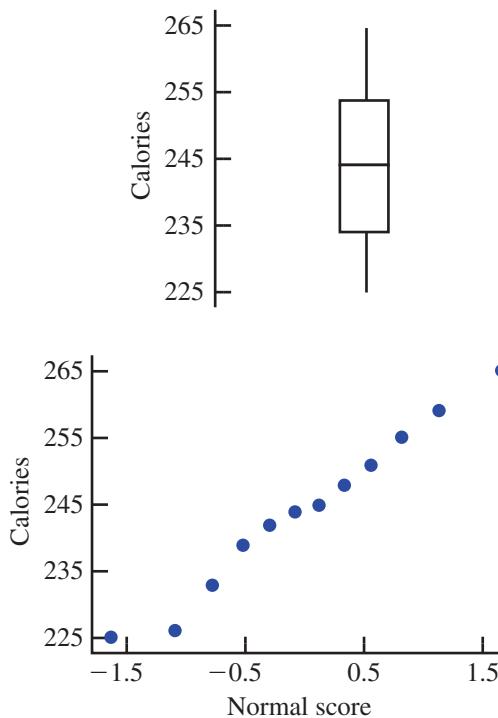
“Listening to music reduces pain intensity levels . . . However, the magnitude of these positive effects is small, the clinical relevance of music for pain relief in clinical practice is unclear.”

Are the authors of this paper claiming that the pain reduction attributable to listening to music is not statistically significant, not practically significant, or neither statistically nor practically significant? Explain.

- 10.61** Many consumers pay careful attention to stated nutritional contents on packaged foods when making purchases. It is therefore important that the information on packages be accurate. A random

sample of $n = 12$ frozen dinners of a certain type was selected from production during a particular period, and the calorie content of each one was determined. Here are the resulting observations, along with a boxplot and normal probability plot:

255 244 239 242 265 245 259 248
225 226 251 233



- Is it reasonable to test hypotheses about mean calorie content μ by using a t test? Explain why or why not.
- The stated calorie content is 240. Does the boxplot suggest that actual average content differs from the stated value? Explain your reasoning.
- Carry out a formal test of the hypotheses suggested in Part (b).

- 10.62** Much concern has been expressed regarding the practice of using nitrates as meat preservatives. In one study involving possible effects of these chemicals, bacteria cultures were grown in a medium containing nitrates. The rate of uptake of radio-labeled amino acid (in dpm, disintegrations per minute) was then determined for each culture, yielding the following observations:

7251 6871 9632 6866 9094 5849 8957 7978
7064 7494 7883 8178 7523 8724 7468

Suppose that it is known that the mean rate of uptake for cultures without nitrates is 8000.

Do the data suggest that the addition of nitrates results in a decrease in the mean rate of uptake? Test the appropriate hypotheses using a significance level of 0.10.

SECTION 10.5 Power and Probability of Type II Error

In this chapter, we have introduced test procedures for testing hypotheses about population characteristics, such as μ and p . What characterizes a “good” test procedure? It makes sense to think that a good test procedure is one that has both a small probability of rejecting the null hypothesis when it is true (a Type I error) and a high probability of rejecting the null hypothesis when it is false.

The test procedures presented in this chapter allow us to directly control the probability of rejecting a true null hypothesis by our choice of the significance level α . But what about the probability of rejecting the null hypothesis when it is false? As we will see, several factors influence this probability. Let’s begin by considering an example.

Suppose that the student body president at a university is interested in studying the amount of money that students spend on textbooks each semester. The director of the financial aid office believes that the average amount spent on books is \$500 per semester and uses this figure to determine the amount of financial aid for which a student is eligible. The student body president plans to ask each individual in a random sample of students how much he or she spent on books this semester. He will then use the resulting data to test

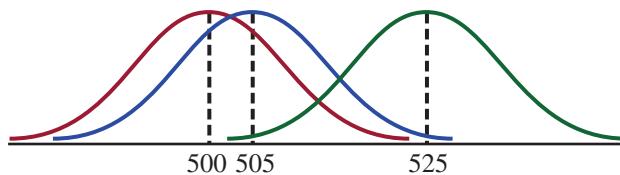
$$H_0: \mu = 500 \text{ versus } H_a: \mu > 500$$

using a significance level of 0.05. If the actual mean is 500 (or less than 500), the correct decision is to fail to reject the null hypothesis. Incorrectly rejecting the null hypothesis is a Type I error. On the other hand, if the actual mean is 525 or 505 or even 501, the correct decision is to reject the null hypothesis. Not rejecting the null hypothesis is a Type II error. How likely is it that the null hypothesis will in fact be rejected?

If the actual mean is 501, the probability that we reject $H_0: \mu = 500$ is not very great. This is because when we carry out the test, we consider the sample mean and ask if it looks like what we would expect to see if the population mean were 500. As illustrated in Figure 10.4, if the actual mean is greater than but very close to 500, chances are that the sample mean will look pretty much like what we would expect to see if the population mean were 500, and we will not be convinced that the null hypothesis should be rejected. If the true mean is 525, it is less likely that the sample will be mistaken for a sample from a population with mean 500 because sample means will tend to cluster around 525. In this case, it is more likely that we will correctly reject H_0 . If the actual mean is 550, rejection of H_0 is even more likely.

FIGURE 10.4

Sampling distribution of \bar{x} for $\mu = 500, 505$, and 525 .



When we consider the probability of rejecting the null hypothesis, we are looking at what statisticians refer to as the **power** of the test.

DEFINITION

Power of a test: The probability of rejecting the null hypothesis.

From the previous discussion, it should be apparent that when a hypothesis about a population mean is being tested, the power of the test depends on the actual value of the population mean, μ . Because the actual value of μ is unknown (if we knew the value of μ we wouldn’t be doing the hypothesis test!), we cannot know what the power is for the actual value of μ . However, it is possible to gain some insight into the power of a test by looking at a number of “what if” scenarios. For example, we might ask, What is the power if the actual mean is 525? or What is the power if the actual mean is 505? and so

on. We can determine the power at $\mu = 525$, the power at $\mu = 505$, and the power at any other value of interest. Although it is technically possible to consider power when the null hypothesis is true, an investigator is usually concerned about the power only at values for which the null hypothesis is false.

In general, when testing a hypothesis about a population characteristic, there are three factors that influence the power of the test:

1. The size of the difference between the actual value of the population characteristic and the hypothesized value (the value that appears in the null hypothesis).
2. The choice of significance level, α .
3. The sample size.

Effect of Various Factors on the Power of a Test

1. The larger the size of the difference between the hypothesized value and the actual value of the population characteristic, the greater the power.
2. The larger the significance level, α , the greater the power of the test.
3. The larger the sample size, the greater the power of the test.

Let's consider each of the statements in the box above. The first statement has already been discussed in the context of the textbook example. Because power is the probability of rejecting the null hypothesis, it makes sense that the power will be higher when the actual value of a population characteristic is quite different from the hypothesized value than when it is close to the hypothesized value.

The effect of significance level on power is not as obvious. To understand the relationship between power and significance level, it helps to consider the relationship between power and β , the probability of a Type II error.

When H_0 is false, power = $1 - \beta$.

This relationship follows from the definitions of power and Type II error. A Type II error results from *not* rejecting a false H_0 . Because power is the probability of rejecting H_0 , it follows that *when H_0 is false*

$$\begin{aligned}\text{power} &= \text{probability of rejecting a false } H_0 \\ &= 1 - \text{probability of not rejecting a false } H_0 \\ &= 1 - \beta\end{aligned}$$

Recall from Section 10.2 that the choice of α , the Type I error probability, affects the value of β , the Type II error probability. Choosing a larger value for α results in a smaller value for β (and therefore a larger value for $1 - \beta$). In terms of power, this means that choosing a larger value for α results in a larger value for the power of the test. The larger the Type I error probability we are willing to tolerate, the more likely it is that the test will be able to detect any particular departure from H_0 .

The third factor that affects the power of a test is the sample size. When H_0 is false, the power of a test is the probability that we will “detect” that H_0 is false and, based on the observed sample, reject H_0 . Intuition suggests that we will be more likely to detect a departure from H_0 with a large sample than with a small sample. This is in fact the case—the larger the sample size, the higher the power.

For example, consider testing the hypotheses:

$$H_0: \mu = 500 \quad \text{versus} \quad H_a: \mu > 500$$

The observations about power imply the following:

1. For any value of μ exceeding 500, the power of a test based on a sample of size 100 is greater than the power of a test based on a sample of size 75 (assuming the same significance level).

2. For any value of μ exceeding 500, the power of a test using a significance level of 0.05 is greater than the power of a test using a significance level of 0.01 (assuming the same sample size).
3. For any value of μ exceeding 500, the power of the test is greater if the actual mean is 550 than if the actual mean is 525 (assuming the same sample size and significance level).

As was mentioned previously in this section, it is impossible to calculate the *exact* power of a test because in practice we do not know the values of population characteristics. However, we can evaluate the power at a selected alternative value which would tell us whether the power would be high or low if this alternative value is the actual value.

The following optional subsection shows how Type II error probabilities and power can be evaluated for selected tests.

Calculating Power and Type II Error Probabilities for Selected Tests (Optional)

The test procedures presented in this chapter are designed to control the probability of a Type I error (rejecting H_0 when H_0 is true) at the desired significance level α . However, little has been said so far about calculating the value of β , the probability of a Type II error (not rejecting H_0 when H_0 is false). Here, we consider the calculation of β and power for the hypothesis tests introduced in this chapter.

When we carry out a hypothesis test, we specify the desired value of α , the probability of a Type I error. The probability of a Type II error, β , is the probability of not rejecting H_0 even though it is false. Suppose that we are testing

$$H_0: \mu = 1.5 \quad \text{versus} \quad H_a: \mu > 1.5$$

Because we do not know the actual value of μ , we cannot calculate the actual value of β . However, the probability of Type II error can be investigated by calculating β for several different potential values of μ , such as $\mu = 1.55$, $\mu = 1.6$, and $\mu = 1.7$. Once a value of β has been determined, the power of the test at the corresponding alternative value is just $1 - \beta$.

Example 10.17 Calculating Power

An airline claims that the mean time on hold for callers to its customer service phone line is 1.5 minutes. We might investigate this claim by testing

$$H_0: \mu = 1.5 \quad \text{versus} \quad H_a: \mu > 1.5$$

where μ is the actual mean customer hold time. A random sample of $n = 36$ calls is to be selected, and the resulting data will be used to reach a conclusion.

Suppose that the standard deviation of hold time (σ) is known to be 0.20 minutes and that a significance level of 0.01 is to be used. Our test statistic is

$$z = \frac{\bar{x} - 1.5}{\frac{0.20}{\sqrt{n}}} = \frac{\bar{x} - 1.5}{\frac{0.20}{\sqrt{36}}} = \frac{\bar{x} - 1.5}{0.0333}$$

The inequality in H_a implies that

P -value = area under z curve to the right of calculated z

From Appendix Table 2, it is easily verified that the z critical value 2.33 captures an upper-tail z curve area of 0.01. This means that P -value ≤ 0.01 only when $z \geq 2.33$. This is equivalent to the decision rule

reject H_0 if calculated $z \geq 2.33$

which becomes

$$\text{reject } H_0 \text{ if } \frac{\bar{x} - 1.5}{0.0333} \geq 2.33$$

Solving this inequality for \bar{x} we get

$$\bar{x} \geq 1.5 + 2.33(0.0333)$$

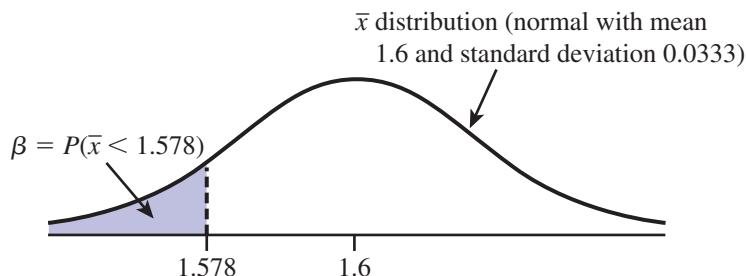
or

$$\bar{x} \geq 1.578$$

So if $\bar{x} \geq 1.578$, we will reject H_0 , and if $\bar{x} < 1.578$, we will fail to reject H_0 . This decision rule corresponds to $\alpha = 0.01$.

Suppose now that $\mu = 1.6$ (so that H_0 is false). A Type II error will then occur if $\bar{x} < 1.578$. What is the probability that this occurs? If $\mu = 1.6$, the sampling distribution of \bar{x} is approximately normal, centered at 1.6, and has a standard deviation of 0.0333. The probability of observing an \bar{x} value less than 1.578 can then be determined by finding an area under a normal curve with mean 1.6 and standard deviation 0.0333, as illustrated in Figure 10.5.

FIGURE 10.5
 β for $\mu = 1.6$ in Example 10.17.



Because the curve in Figure 10.5 is not the standard normal (z) curve, we must first convert to a z score before using Appendix Table 2 or technology to find the area. Here,

$$z \text{ score for } 1.578 = \frac{1.578 - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{1.578 - 1.6}{0.0333} = -0.66$$

and

$$\text{area under } z \text{ curve to left of } -0.66 = 0.2546$$

So, if $\mu = 1.6$, $\beta = 0.2546$. This means that if μ is 1.6, about 25% of all samples would still result in \bar{x} values less than 1.578, and we would not reject H_0 .

The power of the test at $\mu = 1.6$ is then

$$\begin{aligned} (\text{power at } \mu = 1.6) &= 1 - (\beta \text{ when } \mu \text{ is } 1.6) \\ &= 1 - 0.2546 \\ &= 0.7454 \end{aligned}$$

This means that if the actual mean is 1.6, the probability of rejecting H_0 : $\mu = 1.5$ in favor of H_a : $\mu > 1.5$ is 0.7454. If μ is 1.6 and the test is used repeatedly with random samples selected from the population, in the long run about 75% of the samples will result in the correct conclusion to reject H_0 .

Now consider β and power when $\mu = 1.65$. The normal curve in Figure 10.5 would then be centered at 1.65. Because β is the area to the left of 1.578 and the curve has shifted to the right, β decreases. Converting 1.578 to a z score and using Appendix Table 2 gives $\beta = 0.0154$. Also,

$$(\text{power at } \mu = 1.65) = 1 - 0.0154 = 0.9846$$

As expected, the power at $\mu = 1.65$ is greater than the power at $\mu = 1.6$ because 1.65 is farther from the hypothesized value of 1.5.

Statistical software packages and some graphing calculators can calculate the power for specified values of σ , α , n , and the difference between the actual and hypothesized values of μ . The following Minitab output shows power calculations corresponding to those in Example 10.17:

```
1-Sample Z Test
Testing mean = null (versus > null)
Alpha = 0.01    Sigma = 0.2      Sample Size = 36
Difference      Power
 0.10          0.7497
 0.15          0.9851
```

The slight differences between the power values computed by Minitab and those previously obtained are due to rounding in Example 10.17.

The probability of a Type II error and the power for z tests concerning a population proportion are calculated in a similar manner.

Example 10.18 Power for Testing Hypotheses About Proportions

A package delivery service advertises that at least 90% of all packages brought to its office by 9 A.M. for delivery in the same city are delivered by noon that day. Let p denote the actual proportion of all such packages that are delivered by noon. The hypotheses of interest are

$$H_0: p = 0.9 \text{ versus } H_a: p < 0.9$$

where the alternative hypothesis states that the company's claim is not true.

The value $p = 0.8$ represents a substantial departure from the company's claim. If the hypotheses are tested using significance level 0.01 and a sample of $n = 225$ packages, what is the probability that the departure from H_0 represented by this alternative value will go undetected?

At significance level 0.01, H_0 is rejected if $P\text{-value} \leq 0.01$. For the case of a lower-tailed test, this is the same as rejecting H_0 if

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - 0.9}{\sqrt{\frac{(0.9)(1 - 0.9)}{225}}} = \frac{\hat{p} - 0.9}{0.02} \leq -2.33$$

(Because -2.33 captures a lower-tail z curve area of 0.01, the smallest 1% of all z values satisfy $z \leq -2.33$.) This inequality is equivalent to $\hat{p} \leq 0.853$, so H_0 is *not* rejected if $\hat{p} > 0.853$. When $p = 0.8$, \hat{p} has approximately a normal distribution with

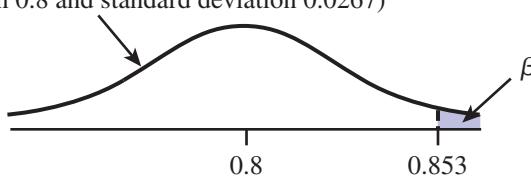
$$\mu_{\hat{p}} = 0.8$$

$$\sigma_{\hat{p}} = \sqrt{\frac{(0.8)(1 - 0.8)}{225}} = 0.0267$$

Then β is the probability of obtaining a sample proportion greater than 0.853, as illustrated in Figure 10.6.

FIGURE 10.6
 β for $p = 0.8$ in Example 10.18.

Sampling distribution of \hat{p} (normal with mean 0.8 and standard deviation 0.0267)



Converting to a z score results in

$$z = \frac{0.853 - 0.8}{0.0267} = 1.99$$

and Appendix Table 2 gives

$$\beta = 1 - 0.9767 = 0.0233$$

When $p = 0.8$ and a significance level of 0.01 is used, less than 3% of all samples of size $n = 225$ will result in a Type II error. The power of the test at $p = 0.8$ is $1 - 0.0233 = 0.9767$. This means that the probability of rejecting $H_0: p = 0.9$ in favor of $H_a: p < 0.9$ when p is really 0.8 is 0.9767, which is quite high.

β and Power for the t Test (Optional)

The power and β values for t tests can be approximated by using a set of curves constructed for this purpose or by using appropriate software. As with the z test, the value of β depends not only on the actual value of μ but also on the selected significance level α . β increases as α is made smaller. In addition, β depends on the number of degrees of freedom, $n - 1$. For any fixed significance level α , it should be easier for the test to detect a specific departure from H_0 when n is large than when n is small. For a fixed alternative value, β decreases as $n - 1$ increases.

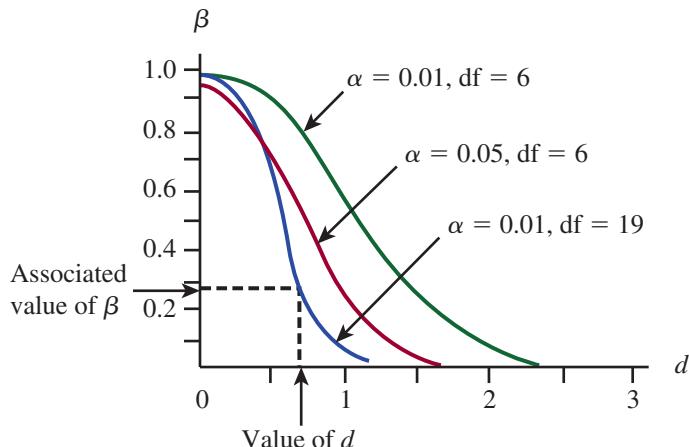
Unfortunately, there is one other quantity on which β depends: the population standard deviation σ . As σ increases, so does $\sigma_{\bar{x}}$. This in turn makes it more likely that an \bar{x} value far from μ will be observed just by chance, resulting in an incorrect conclusion. Once α is specified and n is fixed, the determination of β at a particular alternative value of μ requires that a value of σ be chosen, because each different value of σ yields a different value of β . (This did not present a problem with the z test because when using a z test, the value of σ is known.) If it is possible to specify a range of plausible values for σ , then using the largest such value will give a conservative estimate of β (one on the high side) and a conservative estimate of power (one on the low side).

Figure 10.7 shows three different β curves for a one-tailed t test (appropriate for $H_a: \mu >$ hypothesized value or for $H_a: \mu <$ hypothesized value). A more complete set of curves for both one-tailed and two-tailed tests when $\alpha = 0.05$ and when $\alpha = 0.01$ appears in Appendix Table 5. To determine β , first calculate the quantity

$$d = \frac{|\text{alternative value} - \text{hypothesized value}|}{\sigma}$$

Then locate d on the horizontal axis, move directly up to the curve for $n - 1$ df, and move over to the vertical axis to obtain an estimate of β .

FIGURE 10.7
 β curves for the one-tailed t test.



Example 10.19 β and Power for t Tests

Consider testing

$$H_0: \mu = 100 \text{ versus } H_a: \mu > 100$$

and focus on the alternative value $\mu = 110$. Suppose that $\sigma = 10$, the sample size is $n = 7$, and a significance level of 0.01 has been selected. For $\sigma = 10$,

$$d = \frac{|110 - 100|}{10} = \frac{10}{10} = 1$$

Figure 10.7 (using $df = 7 - 1 = 6$) gives $\beta \approx 0.6$. This means that if $\sigma = 10$, the significance level is 0.01 and $n = 7$, when $\mu = 110$, roughly 60% of all samples result in an incorrect decision to not reject H_0 ! Equivalently, the power of the test at $\mu = 110$ is only $1 - 0.6 = 0.4$. The probability of rejecting H_0 when $\mu = 110$ is not very large.

If a 0.05 significance level is used instead, then $\beta \approx 0.3$, which is still rather large. Using a 0.01 significance level with $n = 20$ ($df = 19$) yields, from Figure 10.7, $\beta \approx 0.05$. At the alternative value $\mu = 110$, for $\sigma = 10$, a significance level of 0.01, and $n = 20$, β is smaller than for a significance level of 0.05 with $n = 7$. Substantially increasing n counterbalances using the smaller α .

Now consider the alternative $\mu = 105$, again with $\sigma = 10$, so that

$$d = \frac{|105 - 100|}{10} = \frac{5}{10} = 0.5$$

Then, from Figure 10.7, $\beta = 0.95$ when $\alpha = 0.01$, $n = 7$; $\beta = 0.7$ when $\alpha = 0.05$, $n = 7$; and $\beta = 0.65$ when $\alpha = 0.01$, $n = 20$. These values of β are all quite large. With $\sigma = 10$, $\mu = 105$ is too close to the hypothesized value of 100 for any of these three tests to have a good chance of detecting this departure from H_0 . A substantial decrease in β would require using a much larger sample size. For example, from Appendix Table 5, $\beta = 0.08$ when $\alpha = 0.05$ and $n = 40$.

The curves in Figure 10.7 also give β when testing $H_0: \mu = 100$ versus $H_a: \mu < 100$. If the alternative value $\mu = 90$ is of interest and $\sigma = 10$,

$$d = \frac{|90 - 100|}{10} = \frac{10}{10} = 1$$

and values of β are the same as those given in the first paragraph of this example.

Because curves for only selected degrees of freedom appear in Appendix Table 5, other degrees of freedom require a visual approximation. For example, the 27-df curve (for $n = 28$) lies between the 19-df and 29-df curves, which do appear, and it is closer to the 29-df curve. This type of approximation is adequate because it is the general magnitude of β —large, small, or moderate—that is of primary concern.

Minitab can also evaluate power for the t test. For example, the following output shows Minitab calculations for power at $\mu = 110$ for samples of size 7 and 20 when $\alpha = 0.01$. The corresponding approximate values from Appendix Table 5 found in Example 10.19 are fairly close to the Minitab values.

1-Sample t Test

Testing mean = null (versus $>$ null)

Calculating power for mean = null + 10

Alpha = 0.01 Sigma = 10

Sample Size Power

7 0.3968

20 0.9653

The β curves in Appendix Table 5 are those for t tests. When the alternative value in H_a corresponds to a value of d relatively close to 0, β for a t test may be rather large. One might wonder whether there is another type of test that has the same level of significance α as does the t test and smaller values of β . The following result provides the answer to this question.

When the population distribution is normal, the t test for testing hypotheses about μ has smaller β than does any other test procedure that has the same level of significance α .

Stated another way, among all tests with level of significance α , the t test makes β as small as it can possibly be when the population distribution is normal. In this sense, the t test is a best test. Statisticians have also shown that when the population distribution is not too far from a normal distribution, no test procedure can improve on the t test by very much (i.e., no test procedure can have the same α and substantially smaller β). However, when the population distribution is believed to be strongly nonnormal (heavy-tailed, highly skewed, or multimodal), the t test should not be used. One alternative method that could be used in this situation is introduced in Section 10.8.

EXERCISES 10.63 - 10.71

- 10.63** The power of a test is influenced by the sample size and the choice of significance level.

- Explain how increasing the sample size affects the power (when significance level is held fixed).
- Explain how increasing the significance level affects the power (when sample size is held fixed).

- 10.64** Water specimens are taken from water used for cooling as it is being discharged from a power plant into a river. It has been determined that as long as the mean temperature of the discharged water is at most 150°F , there will be no negative effects on the river's ecosystem. To investigate whether the plant is in compliance with regulations that prohibit a mean discharge water temperature above 150°F , a scientist will take 50 water specimens at randomly selected times and will record the water temperature of each specimen. She will then use a z statistic

$$z = \frac{\bar{x} - 150}{\frac{\sigma}{\sqrt{n}}}$$

to decide between the hypotheses $H_0: \mu = 150$ and $H_a: \mu > 150$, where μ is the mean temperature of discharged water. Assume that σ is known to be 10.

- Explain why use of the z statistic is appropriate in this setting.
- Describe Type I and Type II errors in this context.
- The rejection of H_0 when $z \geq 1.8$ corresponds to what value of α ? (That is, what is the area under the z curve to the right of 1.8?)

- Suppose that the actual value for μ is 153 and that H_0 is to be rejected if $z \geq 1.8$. Draw a sketch (similar to that of Figure 10.5) of the sampling distribution of \bar{x} , and shade the region that would represent β , the probability of making a Type II error.

- 10.65** Consider the water temperature situation described in the previous exercise.

- For the hypotheses and test procedure described in the previous exercise, calculate the value of β when $\mu = 153$. (Hint: See Example 10.17.)
- For the hypotheses and test procedure described in the previous exercise, what is the value of β if $\mu = 160$?
- What would be the conclusion of the test if H_0 is rejected when $z \geq 1.8$ and $\bar{x} = 152.4$? What type of error might have been made in reaching this conclusion?

- 10.66** Let μ denote the mean lifetime (in hours) for a certain type of battery under controlled laboratory conditions. A test of $H_0: \mu = 10$ versus $H_a: \mu < 10$ will be based on a sample of size 36. Suppose that σ is known to be 0.6, so $\sigma_{\bar{x}} = 0.1$. The appropriate test statistic is then

$$z = \frac{\bar{x} - 10}{0.1}$$

- What is α for the test procedure that rejects H_0 if $z \leq -1.28$?

- b.** If the test procedure of Part (a) is used, calculate β when $\mu = 9.8$, and interpret this error probability.
- c.** Without doing any calculation, explain how β when $\mu = 9.5$ compares to β when $\mu = 9.8$. Then check your assertion by calculating β when $\mu = 9.5$.
- d.** What is the power of the test when $\mu = 9.8$? when $\mu = 9.5$?
- 10.67** The city council in a large city has become concerned about the trend toward exclusion of renters with children in apartments within the city. The housing coordinator has decided to select a random sample of 125 apartments and determine for each whether children are permitted. Let p be the proportion of all apartments that prohibit children. If the city council is convinced that p is greater than 0.75, it will consider appropriate legislation.
- a.** If 102 of the 125 sampled apartments exclude renters with children, would a level 0.05 test lead you to the conclusion that more than 75% of all apartments exclude children?
- b.** What is the power of the test when $p = 0.8$ and $\alpha = 0.05$? (Hint: See Example 10.18.)
- 10.68** The amount of shaft wear after a fixed mileage was determined for each of seven randomly selected internal combustion engines, resulting in a mean of 0.0372 inch and a standard deviation of 0.0125 inch.
- a.** Assuming that the distribution of shaft wear is normal, use $\alpha = 0.05$ to test the hypotheses $H_0: \mu = 0.035$ versus $H_a: \mu > 0.035$.
- b.** Using $\sigma = 0.0125$, $\alpha = 0.05$, and Appendix Table 5, what is the approximate value of β , the probability of a Type II error, when $\mu = 0.04$? (Hint: See Example 10.19.)
- c.** What is the approximate power of the test when $\mu = 0.04$ and $\alpha = 0.05$?
- 10.69** Optical fibers are used in telecommunications to transmit light. Suppose current technology allows production of fibers that transmit light about 50 km. Researchers are trying to develop a new type of glass fiber that will increase this distance. In evaluating a new fiber, it is of interest to test $H_0: \mu = 50$ versus $H_a: \mu > 50$, with μ denoting the mean transmission distance for the new optical fiber.
- a.** Assuming $\sigma = 10$ and $n = 10$, use Appendix Table 5 to approximate β , the probability of a Type II error, for each of the given alternative values of μ for a test with significance level 0.05:
- i. 52 ii. 55 iii. 60 iv. 70
- b.** What happens to β in each of the cases in Part (a) if σ is actually larger than 10? Explain your reasoning.
- 10.70** Let μ denote the mean diameter for bearings of a certain type. A test of $H_0: \mu = 0.5$ versus $H_a: \mu \neq 0.5$ will be based on a sample of n bearings. The diameter distribution is believed to be normal. Determine the approximate value of β in each of the following cases:
- a.** $n = 15, \alpha = 0.05, \sigma = 0.02, \mu = 0.52$
- b.** $n = 15, \alpha = 0.05, \sigma = 0.02, \mu = 0.48$
- c.** $n = 15, \alpha = 0.01, \sigma = 0.02, \mu = 0.52$
- 10.71** Use the information given in the previous exercise to determine the approximate value of β in each of the following cases.
- a.** $n = 15, \alpha = 0.05, \sigma = 0.02, \mu = 0.54$
- b.** $n = 15, \alpha = 0.05, \sigma = 0.04, \mu = 0.54$
- c.** $n = 20, \alpha = 0.05, \sigma = 0.04, \mu = 0.54$

SECTION 10.6 Interpreting and Communicating the Results of Statistical Analyses

The nine-step procedure that we have proposed for testing hypotheses provides a systematic approach for carrying out a complete test. However, you rarely see the results of a hypothesis test reported in publications in such a complete way.

Communicating the Results of Statistical Analyses

When summarizing the results of a hypothesis test, it is important that you include several things in the summary in order to provide all the relevant information. These are:

- 1. Hypotheses.** Whether specified in symbols or described in words, it is important that both the null and the alternative hypotheses be clearly stated. If you are using symbols to define the hypotheses, be sure to describe them in the context of the problem at hand (for example, μ = population mean calorie intake).
- 2. Test procedure.** You should be clear about what test procedure was used (for example, large-sample z test for proportions) and why you think it was reasonable

to use this procedure. The plausibility of any required assumptions should be satisfactorily addressed.

3. *Test statistic.* Be sure to include the value of the test statistic and the associated *P*-value. Including the *P*-value allows a reader who may have chosen a different significance level to see whether she would have reached the same or a different conclusion.
4. *Conclusion in context.* Never end the report of a hypothesis test with the statement “I rejected (or did not reject) H_0 .” Always provide a conclusion that is in the context of the problem and that answers the original research question that the hypothesis test was designed to answer. Be sure also to indicate the level of significance used as a basis for the decision.

Interpreting the Results of Statistical Analyses

When the results of a hypothesis test are reported in a journal article or other published source, it is common to find only the value of the test statistic and the associated *P*-value accompanying the discussion of conclusions drawn from the data. Often, especially in newspaper articles, only sample summary statistics are given, with the conclusion immediately following. You may have to fill in some of the intermediate steps for yourself to see whether or not the conclusion is justified.

For example, the article “[Physicians’ Knowledge of Herbal Toxicities and Adverse Herb-Drug Interactions](#)” (*European Journal of Emergency Medicine*, August 2004) summarizes the results of a study to assess doctors’ familiarity with adverse effects of herbal remedies as follows:

A total of 142 surveys and quizzes were completed by 59 attending physicians, 57 resident physicians, and 26 medical students. The mean subject score on the quiz was only slightly higher than would have occurred from random guessing.

The quiz consisted of 16 multiple-choice questions. If each question had four possible choices, the statement that the mean quiz score was only slightly higher than would have occurred from random guessing suggests that the researchers considered the hypotheses $H_0: \mu = 4$ and $H_a: \mu > 4$, where μ represents the mean score for the population of all physicians and medical students and the null hypothesis corresponds to the expected number of correct choices for someone who is guessing. Assuming that it is reasonable to regard this sample as representative of the population of interest, the data from the sample could be used to carry out a test of these hypotheses.

What to Look For in Published Data

Here are some questions to consider when you are reading a report that contains the results of a hypothesis test:

- What hypotheses are being tested? Are the hypotheses about a population mean, a population proportion, or some other population characteristic?
- Was the appropriate test used? Does the validity of the test depend on any assumptions about the sample or about the population from which the sample was selected? If so, are the assumptions reasonable?
- What is the *P*-value associated with the test? Was a significance level reported (as opposed to simply reporting the *P*-value)? Is the chosen significance level reasonable?
- Are the conclusions drawn consistent with the results of the hypothesis test?

For example, consider the following statement from the paper “[Didgeridoo Playing as Alternative Treatment for Obstructive Sleep Apnoea Syndrome](#)” (*British Medical Journal* [2006]: 266–270): “We found that four months of training of the upper airways by didgeridoo playing reduces daytime sleepiness in people with snoring and obstructive apnoea syndrome.” This statement was supported by data on a measure of daytime sleepiness called the Epworth scale. For the 14 participants in the study, the mean improvement in Epworth scale was 4.4 and the standard deviation was 3.7.

The paper does not indicate what test was performed or what the value of the test statistic was. It appears that the hypotheses of interest are $H_0: \mu = 0$ (no improvement) versus $H_a: \mu > 0$, where μ represents the mean improvement in Epworth score after four months of didgeridoo playing for all people with snoring and obstructive sleep apnoea. Because the sample size is not large, the one-sample t test would be appropriate if the sample can be considered a random sample and the distribution of Epworth scale improvement scores is approximately normal. If these assumptions are reasonable (something that was not addressed in the paper), the t test results in $t = 4.45$ and an associated P -value of 0.000. Because the reported P -value is so small H_0 would be rejected, supporting the conclusion in the paper that didgeridoo playing is an effective treatment. (In case you are wondering, a didgeridoo is an Australian Aboriginal woodwind instrument.)

A Word to the Wise: Cautions and Limitations

There are several things you should watch for when conducting a hypothesis test or when evaluating a written summary of a hypothesis test.

1. A hypothesis test can never show strong support for the null hypothesis. Make sure that you don't confuse "There is no reason to believe the null hypothesis is not true" with the statement "There is convincing evidence that the null hypothesis is true." These are very different statements!
2. If you have complete information for the population, don't carry out a hypothesis test! It should be obvious that no test is needed to answer questions about a population if you have complete information and don't need to generalize from a sample, but people sometimes forget this fact. For example, in an article on growth in the number of prisoners by state, the *San Luis Obispo Tribune (August 13, 2001)* reported "California's numbers showed a statistically insignificant change, with 66 fewer prisoners at the end of 2000." The use of the term "statistically insignificant" implies some sort of statistical inference, which is not appropriate when a complete accounting of the entire prison population is known. Perhaps the author confused statistical and practical significance. Which brings us to . . .
3. Don't confuse statistical significance with practical significance. When statistical significance has been declared, be sure to step back and evaluate the result in light of its practical importance. For example, we may be convinced that the proportion who respond favorably to a proposed medical treatment is greater than 0.4, the known proportion that responds favorably for the currently recommended treatments. But if our estimate of this proportion for the proposed treatment is 0.405, is this of any practical interest? It might be if the proposed treatment is less costly or has fewer side effects, but in other cases it may not be of any real interest. Conclusions should always be interpreted in context.

EXERCISES 10.72 - 10.73

- 10.72** In 2006, Boston Scientific sought approval for a new heart stent (a medical device used to open clogged arteries) called the Liberte. This stent was being proposed as an alternative to a stent called the Express that was already on the market. The following excerpt is from an article that appeared in *The Wall Street Journal (August 14, 2008)*:

Boston Scientific wasn't required to prove that the Liberte was 'superior' than a previous treatment, the agency decided—only that it wasn't 'inferior' to Express. Boston Scientific proposed—and the FDA

okayed—a benchmark in which Liberte could be up to three percentage points worse than Express—meaning that if 6% of Express patients' arteries reclog, Boston Scientific would have to prove that Liberte's rate of reclogging was less than 9%. Anything more would be considered 'inferior.' . . . In the end, after nine months, the Atlas study found that 85 of the patients suffered reclogging. In comparison, historical data on 991 patients implanted with the Express stent show a 7% rate. Boston Scientific then had to answer this question: Could the study have gotten such results if the Liberte were truly inferior to Express?"

Assume a 7% reclogging rate for the Express stent. Explain why it would be appropriate for Boston Scientific to carry out a hypothesis test using the following hypotheses:

$$\begin{aligned} H_0: p &= 0.10 \\ H_a: p &< 0.10 \end{aligned}$$

where p is the proportion of patients receiving Liberte stents that suffer reclogging. Be sure to address both the choice of the hypothesized value and the form of the alternative hypothesis in your explanation.

- 10.73** The article “Boy or Girl: Which Gender Baby Would You Pick?” (*LiveScience*, March 23, 2005, livescience.com) summarized the findings of a study that was published in *Fertility and Sterility*. The *LiveScience* article makes the following statements: “When given the opportunity to choose the sex of their baby, women are just as likely to choose pink socks as blue, a new study shows” and “Of the 561 women who participated in the study, 229 said they would like to choose the sex of a future child. Among these 229, there was no greater demand for boys or girls.” These statements are equivalent to the claim that for women who would like to choose the baby’s sex, the proportion who would choose a girl is 0.50 or 50%.

- a. The journal article on which the *LiveScience* summary was based (“[Preimplantation Sex-Selection Demand and Preferences in an Infertility Population](#),” *Fertility and Sterility* [2005]: 649–658) states that of the 229 women who wanted to select the baby’s sex, 89 wanted a boy and 140 wanted a girl. Does this provide convincing evidence against the statement of no preference in the *LiveScience* summary? Test the relevant hypotheses using $\alpha = 0.05$. Be sure to state any assumptions you must make about the way the sample was selected in order for your test to be appropriate.
- b. The journal article also provided the following information about the study:
- A survey with 19 questions was mailed to 1385 women who had visited the Center for Reproductive Medicine at Brigham and Women’s Hospital.
 - 561 women returned the survey.

Do you think it is reasonable to generalize the results from this survey to a larger population? Do you have any concerns about the way the sample was selected or about potential sources of bias? Explain.

SECTION 10.7 Randomization Test and Exact Binomial Test for a Population Proportion (Optional)

In Section 10.3, we considered how data from a random sample could be used to carry out a large-sample z test for one proportion. For that test to be appropriate, the sample must be a random sample from a population (or is selected in a way that makes it reasonable to think that the sample is representative of a population). In addition, the distribution of the sample proportion, \hat{p} , should be approximately normal. It is reasonable to think that the distribution of \hat{p} is approximately normal when the sample size is large ($np \geq 10$ and $n(1 - p) \geq 10$), but the distribution of \hat{p} is not always approximately normal when the sample size is small.

When the sample size is large, a z test makes use of the normal distribution to calculate a P -value. But even when the sampling distribution of \hat{p} is well approximated by a normal distribution, the resulting P -value for the large-sample z test is still just an approximation to the actual P -value.

When the sample size is not large enough to assume that the sampling distribution of \hat{p} is normal, there are other methods that can be used to approximate a P -value for the hypothesis test that do not require a large sample size. Two of these methods are a randomization test and an “exact” test that is based on the binomial distribution.

A Randomization Test for One Proportion

In a hypothesis test for one proportion, the null hypothesis is of the form $H_0: p =$ hypothesized value, where p is the population proportion and the hypothesized value is determined by the question of interest. For example, the null hypothesis might be $H_0: p = 0.50$.

Recall that in a hypothesis test, the P -value is a measure of how likely it would be to see something as extreme or more extreme as what was observed in the sample data

if the null hypothesis were true. One way to approximate a P -value is to assume that the population proportion is in fact equal to the value specified in the null hypothesis and then simulate random samples from such a population. The distribution of these simulated sample proportions would give a sense of what values for the sample proportion would be expected when the null hypothesis is true and what values would be unlikely to occur. The distribution of simulated proportions also provides a way to find an approximate P -value. This distribution of simulated sample proportions is called a **randomization distribution**.

Example 10.20 Using a Shiny App to Create a Randomization Distribution

Suppose that when deciding if there is evidence that a majority of the students at a school are registered to vote, a random sample of 10 students includes 8 who are registered to vote. A majority of the students in the sample are registered, but does this mean it is reasonable to conclude that there is convincing evidence that a majority of *all* students are registered to vote? To answer this question, we would test the null hypothesis $H_0: p = 0.50$ against the alternative hypothesis $H_a: p > 0.50$.

Assuming that $H_0: p = 0.50$ is true, then values of \hat{p} can be generated for many different simulated samples of size $n = 10$. Open the Shiny app called “Randomization Test for One Proportion,” one of the Shiny web apps that accompany this text. These web apps are located at statistics.cengage.com/PSO6e/Apps.html. Enter 10 for the number of observations and 8 for the number of successes. The default hypothesized value for p is already 0.5. Choose “Upper-Tailed ($>$)” for the form of the alternative hypothesis, and request 1000 simulated samples. Click “Generate Simulated Samples.”

Randomization Test for One Proportion

Select number of observations: 10

Select the number of successes: 8

Enter hypothesized value: 0.5

Select form of alternative hypothesis:

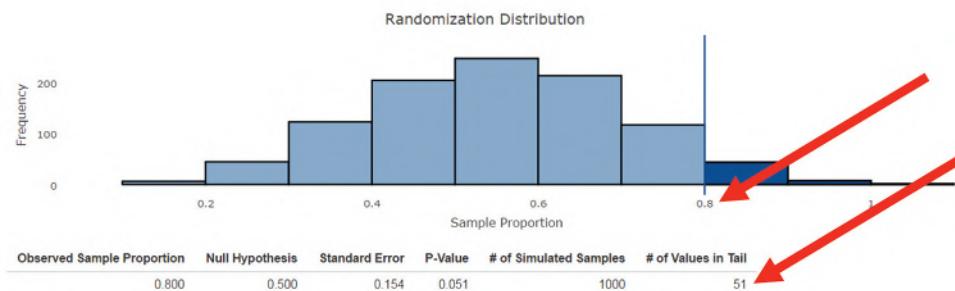
- Two-Tailed (Not Equal)
- Upper-Tailed ($>$)
- Lower-Tailed ($<$)

Select number of simulated samples to generate:

- 1
- 10
- 100
- 1,000
- 10,000

Generate Simulated Samples Reset

The Shiny app generates 1000 simulated random samples of size 10 from a population with $p = 0.50$ and provides a histogram of the values of the 1000 simulated sample proportions. The Shiny app also indicates where the observed value of the sample proportion, $\hat{p} = 0.8$, falls in the randomization distribution.



The app reports that for this particular simulation, 51 of the 1000 simulated samples, or about 5.1%, were as or more extreme than what was observed in the actual sample. This value, $\frac{51}{1000} = 0.051$, is an approximation for the *P*-value for the one-sided test. The approximate *P*-value is also shown in the output produced by the app.

Based on this randomization distribution of 1000 simulated values of \hat{p} when $n = 10$ and the null hypothesis $H_0: p = 0.5$ is true, the approximate *P*-value is $51/1000 = 0.051$. This *P*-value is compared to a selected significance level in order to make a decision about whether the null hypothesis should be rejected. For example, when using a significance level of 0.05, we will fail to reject H_0 because the *P*-value of 0.051 is close to, but greater than $\alpha = 0.05$. This means that even though 8 out of 10 students in the sample were registered to vote, it would not have been surprising to have seen this just by chance when the null hypothesis is true.

Example 10.21 Cell Phones Revisited

The paper referenced in Example 10.11 described a study of 2000 Canadians over the age of 18 that was carried out by Microsoft. Study participants were asked whether the following statement described them: “When nothing is occupying my attention, the first thing I do is reach for my phone.” Of the study participants in the age group 18 to 24 years old, 77% responded “yes” to this question. In Example 10.11, it was assumed that the 77% was based on a representative sample of 800 Canadians age 18 to 24 years. A large-sample *z* test was used to decide if the sample data support the claim that more than 75% of all Canadians in this age group would respond “yes” when asked if the given statement describes them. In this version of the example, a randomization test will be used.

Let p represent the proportion of Canadians ages 18 to 24 years who would respond “yes” for the population represented by the sample of 800. The appropriate hypotheses for the test are $H_0: p = 0.75$ and $H_a: p > 0.75$.

Open the Shiny app called “Randomization Test for One Proportion.” This is one of the Shiny web apps that accompany this text. These web apps are located at statistics.cengage.com/PSO6e/Apps.html. Enter 800 for the Number of observations, and 616 (77% of 800) for the Number of successes. The default hypothesized value for p is 0.5, so change it to 0.75. Choose “Upper-Tailed (>)” for the form of the alternative hypothesis, and request 1000 simulated samples.

Confirm that the value of $\hat{p} = 616/800 = 0.770$ appears below “Observed Sample Proportion” in the simulation output. Notice that the value of “# of Values in Tail” in the run of the simulation shown in the accompanying figure is 107, and that the *P*-value is $107/1000 = 0.107$. This represents the randomization test approximation for the *P*-value for the hypothesis test.

The approximate *P*-value is 0.107. Because the *P*-value is greater than 0.05, we fail to reject H_0 . The sample does not provide convincing evidence that more than 75% of Canadians age 18 to 24 would respond “yes” when asked if the statement “When nothing is occupying my attention the first thing I do is reach for my phone” describes them. This is consistent with the *P*-value obtained for the large-sample *z* test in Example 10.11 (which

Randomization Test for One Proportion

Select number of observations: (Red arrow)

Select the number of successes: (Red arrow)

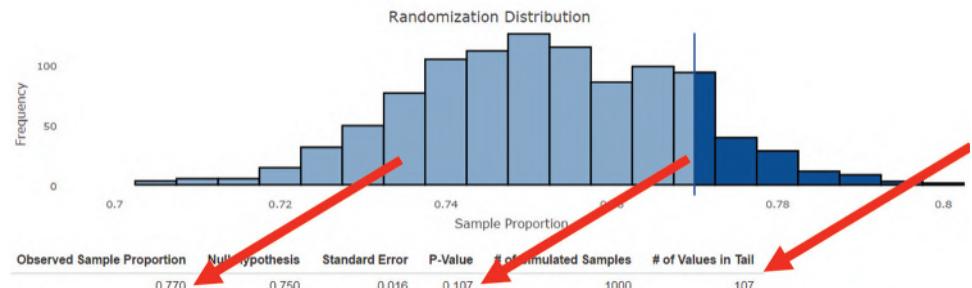
Enter hypothesized value: (Red arrow)

Select form of alternative hypothesis:

- Two-Tailed (Not Equal)
- Upper-Tailed (>)
- Lower-Tailed (<)

Select number of simulated samples to generate:

- 1
- 10
- 100
- 1,000
- 10,000



was 0.092), and so the same conclusion is reached using either method to approximate the P -value. This is not surprising because the sample size is large enough for the large-sample test to be appropriate. The advantage of the randomization test is that it can be used even when the sample size is not large.

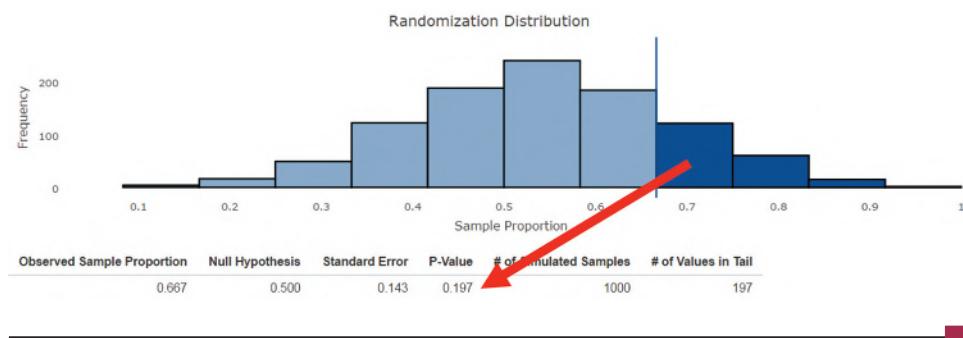
Example 10.22 Vaccination Coverage

Following a 2-year period when the Australian government offered women age 13 to 26 free vaccinations to protect against cervical cancer, researchers carried out surveys of random samples of women in each of the territories in Australia (“[Human Papillomavirus \(HPV\) Vaccination Coverage in Young Australian Women Is Higher Than Previously Estimated](#),” *Vaccine* [2014]: 592–597). While the sample sizes were large for most territories, for the Northern Territory the sample size was only 12. The researchers found that 8 of the 12 women in the sample from the Northern Territory had received at least one of the three recommended doses of HPV vaccine. The data in this sample can be used to test the hypothesis that more than half of women age 13 to 26 in the Northern Territory have received at least one dose of the HPV vaccine.

Let p represent the proportion of women age 13 to 26 in the Northern Territory who have received at least one dose of HPV vaccine. The hypotheses of interest are $H_0: p = 0.50$ and $H_a: p > 0.50$. The observed value of the sample proportion is $\hat{p} = 8/12 = 0.667$.

Notice that the large-sample z test of Section 10.3 is not an appropriate choice because the sample size condition of $np \geq 10$ is not met ($np = (12)(0.50) = 6 < 10$). However, because the sample was a random sample and the randomization test does not require a large sample, it is appropriate to proceed with a randomization test.

From the accompanying output from the Shiny app “Randomization Test for One Proportion,” the approximate P -value for this upper-tailed randomization test is 0.197. This means that 197 of the 1000 simulated sample proportions were at least as large as 0.667. Using a significance level of 0.05, the null hypothesis would not be rejected. The sample does not provide convincing evidence that more than half of women age 13 to 26 in the Northern Territory of Australia have received at least one dose of HPV vaccine.



Example 10.22 illustrates that randomization tests may be used even for small random samples when the conditions for large-sample inference may not be met. For this example, if a large-sample z test had been used, then the resulting P -value of 0.124 would be quite different than the P -value from the randomization test. This illustrates why a hypothesis test should not be used when its assumptions are not met.

An Exact Binomial Test for One Proportion

Another way to obtain a P -value when testing hypotheses about a population proportion is to use an exact probability approach that uses the binomial distribution. The binomial probability distribution was introduced in Section 7.5, Binomial and Geometric Distributions.

Example 10.23 An Exact Binomial Test

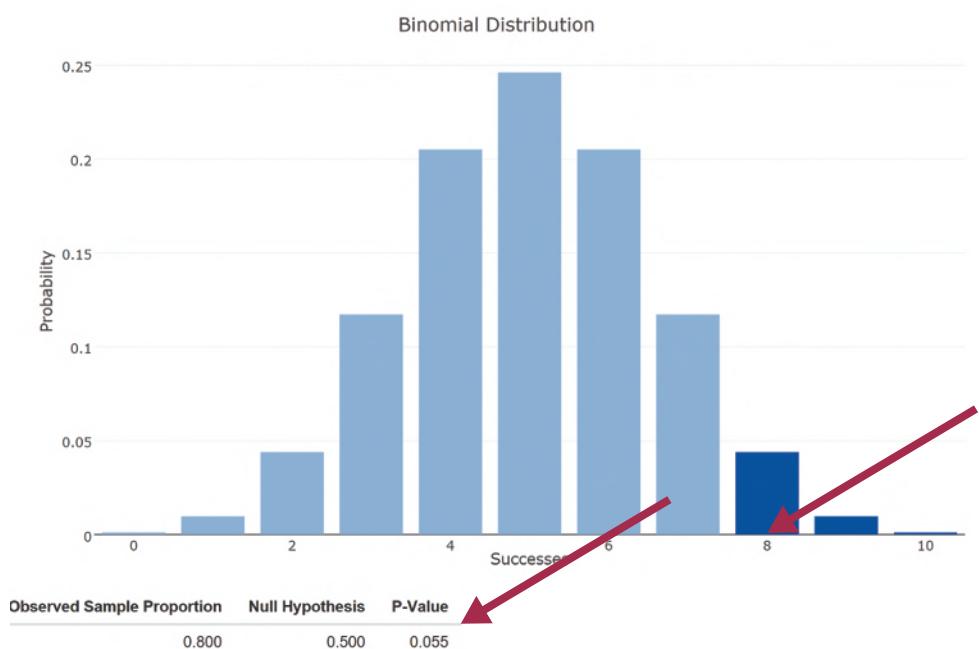
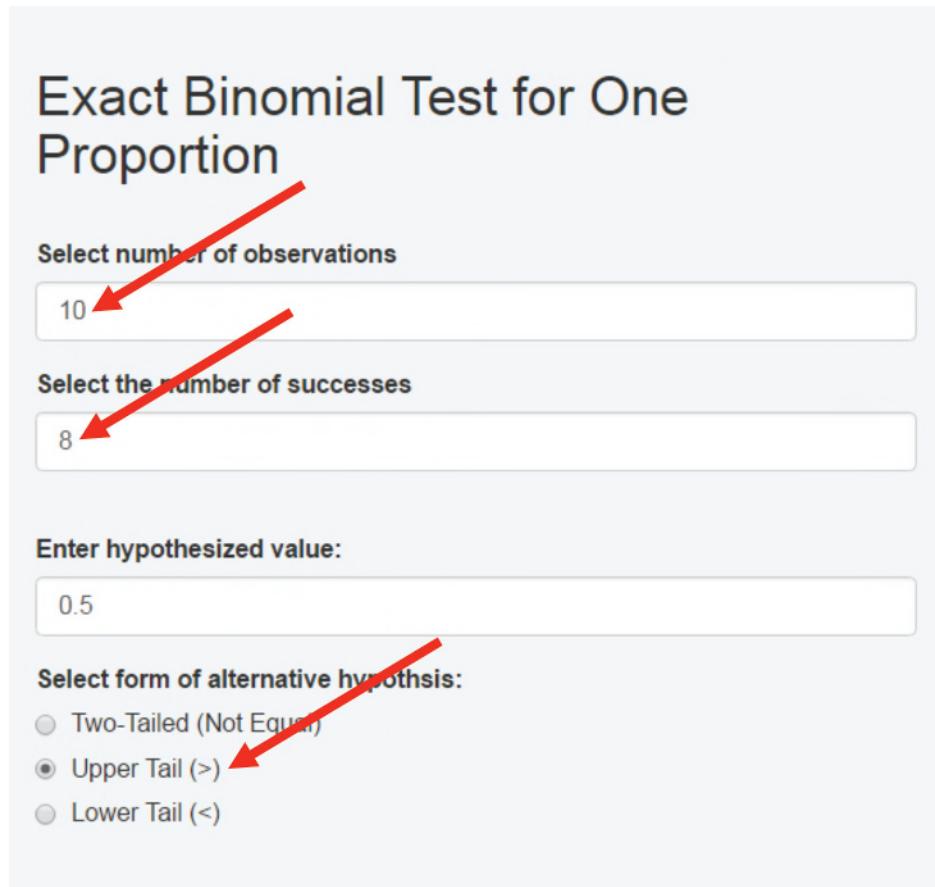
Consider a test of the null hypothesis $H_0: p = 0.50$. The value of p that is specified in the null hypothesis, when combined with an observed sample size, identifies a specific binomial distribution that can be used to calculate an “exact” P -value. Suppose that the alternative hypothesis of interest is $H_a: p > 0.50$ and that the test will be carried out using data from a random sample of $n = 10$ independent success (S) or failure (F) observations.

Because the sample size is small, the large-sample z test is not an appropriate way to test these hypotheses. The exact binomial test does not require a large sample, so it can be used when the sample size is small and the sampling distribution of \hat{p} may not be approximately normal.

Suppose that we observe $x = 8$ successes in the sample of size $n = 10$. This means that $\hat{p} = 8/10 = 0.8$. The binomial distribution with $n = 10$ and $p = 0.50$ (the hypothesized proportion) can be used to calculate the probability of observing a sample proportion as or more extreme than what was observed in the sample. This is the P -value for the hypothesis test.

To calculate the P -value for an exact binomial test, use the Shiny app “Exact Binomial Test for One Proportion.” This is one of the Shiny web apps that accompany this text. These web apps are located at statistics.cengage.com/PSO6e/Apps.html. Enter 10 for the Number of observations and 8 for the Number of successes. The default hypothesized value is already 0.5. Specify “Upper Tail (>)” for the form of the alternative hypothesis.

The Shiny app automatically updates the probability histogram to reflect your choices and displays the upper-tail P -value as a binomial probability. In this example, the P -value is equal to 0.055.



The probability that $x \geq 8$ represents the exact one-sided P -value for testing $H_0: p = 0.50$ versus $H_a: p > 0.50$. From the Shiny app, this probability is 0.055, which is the value of the P -value for the exact binomial test.

Since this P -value = 0.055 is greater than $\alpha = 0.05$, using a significance level of 0.05 we would fail to reject H_0 . Observing eight successes in $n = 10$ trials is not convincing evidence that $p > 0.50$.

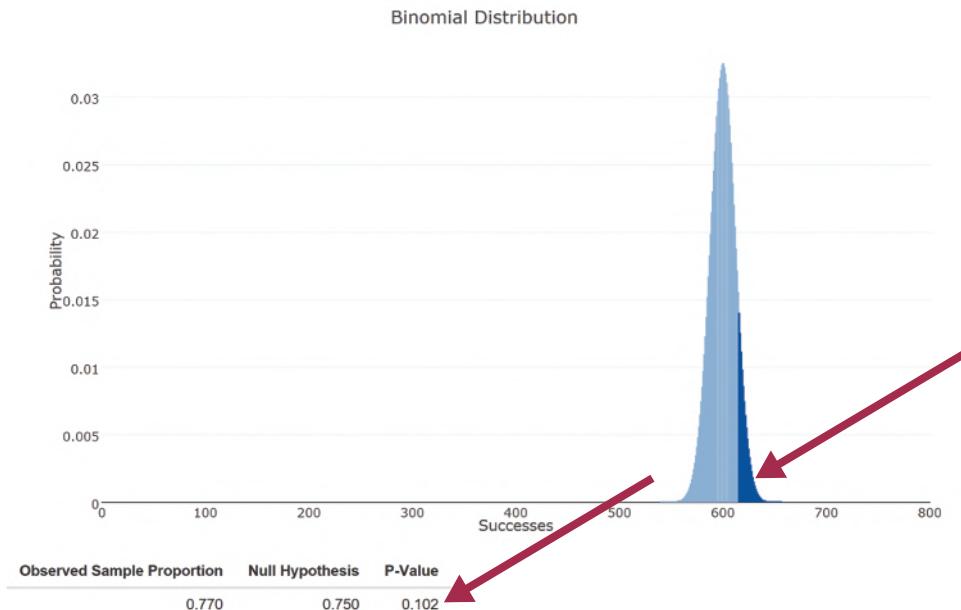
The following examples illustrate the use of the exact binomial test. By revisiting the examples previously used to illustrate the randomization test, comparisons can be made between the approximate P -value from the randomization test and the P -value from the exact binomial test.

Example 10.24 Cell Phones Revisited Again

In the study described in Examples 10.11 and 10.21, 77% of 800 Canadians age 18 to 24 years responded “yes” when asked if the statement “When nothing is occupying my attention the first thing I do is reach for my phone” describes them. As was done in the previous examples, this data will be used to decide whether the sample supports the claim that more than 75% of Canadians in this age group would respond “yes.” In this example, an exact binomial test will be used.

For the exact binomial test, the assumption is made that the Canadians sampled are representative of the population of interest, but there are no sample size conditions that must be satisfied.

The Shiny app “Exact Binomial Test for One Proportion” provides the binomial probability that at least 616 of the people in a random sample of size 800 would have responded “yes” if the null hypothesis of $p = 0.75$ were true.



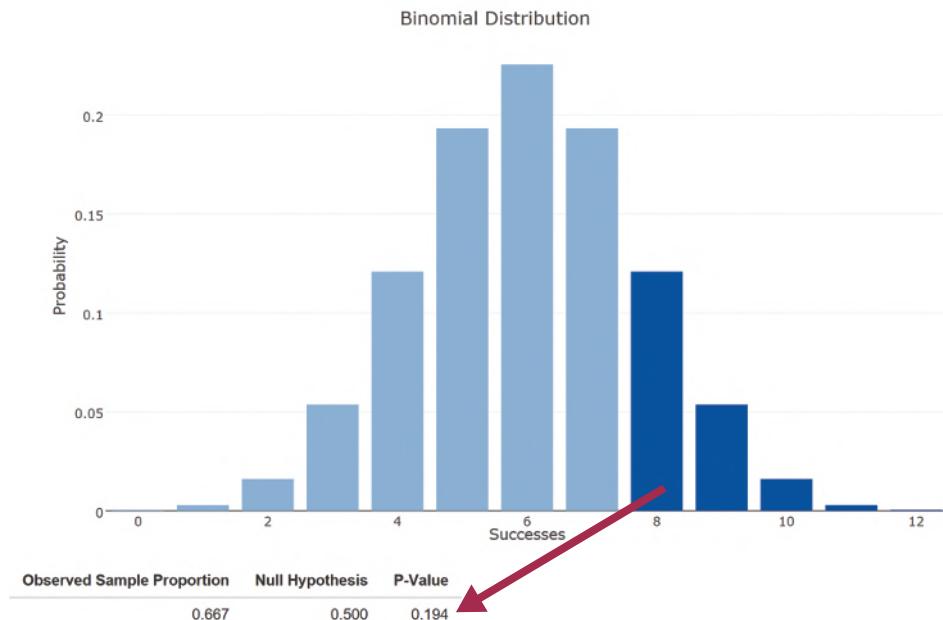
From the Shiny app, the P -value = $P(x \geq 616 \text{ when } n = 800 \text{ and } H_0: p = 0.75 \text{ is true})$ is 0.102. This is greater than 0.05, so we fail to reject H_0 . The sample does not provide convincing evidence that more than 75% of Canadians ages 18 to 24 would respond “yes” when asked if the statement, “When nothing is occupying my attention the first thing I do is reach for my phone” describes them. The P -value from the exact binomial test is similar to the approximate P -values for the large-sample z test (0.092) and the randomization test (0.107). This is not surprising because any of these three methods are appropriate for this sample size.

Example 10.25 Vaccination Coverage Revisited

Recall that in Example 10.22, data from a random sample of 12 women from the Northern Territory of Australia were used to test the claim that more than 50% of women age 13 to 26 had received at least one dose of the HPV vaccine. In this example, the exact binomial

test will be used to test the hypotheses $H_0: p = 0.5$ and $H_a: p > 0.5$, where p represents the proportion of women age 13 to 26 in the Northern Territory who have had at least one dose of the HPV vaccine. It is reasonable to use the exact binomial test because the sample was a random sample and the exact binomial test does not have any sample size requirements.

Recall that 8 of the 12 women in the sample reported that they had received at least one dose of the vaccine. The binomial distribution with $n = 12$ and $p = 0.5$ (from the null hypothesis) can be used to determine the probability of observing 8 or more successes (which is equivalent to more than 7 successes) in a sample of size 12 when the null hypothesis is true and the actual proportion of successes is 0.5.



From the accompanying Shiny app output, the probability of at least 8 (more than 7) successes in a sample of size 12 when $p = 0.50$ is 0.194. This is the P -value for the exact binomial test, and because the P -value is large, the null hypothesis would not be rejected.

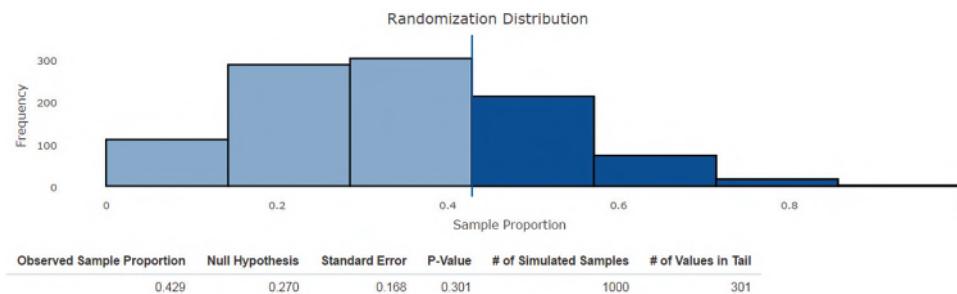
Notice that the P -value for the exact binomial test and the approximate P -value for the randomization test (0.197) are very close, and either method would be an appropriate choice for this sample. The P -value for the large-sample z test was quite a bit smaller (0.124), and we would not want to base a conclusion on this P -value because the assumptions required for the large-sample z test are not met.

EXERCISES 10.74 - 10.79

- 10.74** We are only beginning to learn about the long-term effects of space travel on human health. A study published in 2016 (*Nature Scientific Reports* 6, Article number: 29901, nature.com/articles/srep29901, July 28, 2016, retrieved May 6, 2017) found that by 2014, seven of the U.S. astronauts who traveled to the moon during Apollo lunar missions of the 1960s and 1970s had died, and that three of these ($3/7 = 43\%$) had died from cardiovascular disease (CVD). The overall U.S. death rate due to CVD for adults age 55 to 64 in 2013 was 27%. Do the data for lunar astronauts

indicate that, as a group, they are at increased risk of death caused by CVD? Assume that it is reasonable to regard this sample as representative of the population of past, present, and future U.S. lunar astronauts.

- Explain why the data in this example should not be analyzed using a large-sample hypothesis test for a population proportion.
- Use the output at the top of the next page from the Shiny app “Randomization Test for One Proportion” to complete an appropriate hypothesis test.

Output for Exercise 10.74

- 10.75** A study of hospitalized patients who develop pneumonia reported that 1 in 5 (20%) are readmitted to the hospital within 30 days after discharge (“Comparison of Therapist-Directed and Physician-Directed Respiratory Care in COPD Subjects with Acute Pneumonia,” *Respiratory Care* [2015]: 151–154).

The study reported that 15 out of $n = 162$ hospital patients who had been treated for pneumonia using a respiratory therapist protocol were readmitted to the hospital within 30 days after discharge. These data can be used to decide if the proportion readmitted is less than 0.20.

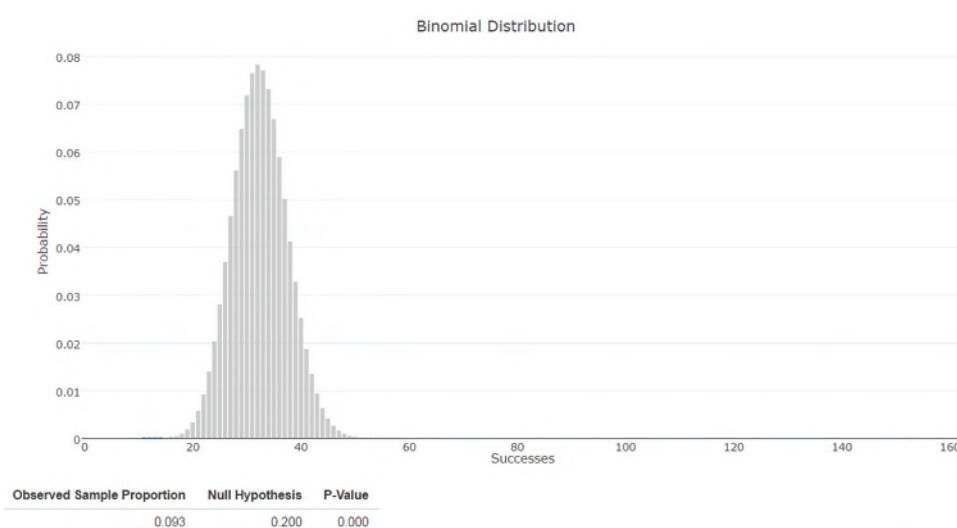
- What hypotheses should be tested?
- Discuss whether the conditions necessary for a large-sample hypothesis test for one proportion are satisfied.
- The exact binomial test can be used even in cases when the sample size condition for the large-sample test is met. Use the output at the bottom of the page from the Shiny app “Exact Binomial Test for One Proportion” to carry out an exact binomial test for one proportion, using the null hypothesis that the population proportion of subjects who will be readmitted to a hospital within 30 days after following a respiratory therapist protocol for treatment of pneumonia is equal to $1/5 = 0.2$.

- Calculate the P -value for the same hypotheses using the large-sample approach for testing one population proportion. Compare this P -value to the P -value obtained in Part (b). Would the same conclusion be reached in either case? Explain.

- 10.76** A sample of dogs were trained using a “Do as I do” method, in which the dog observes the trainer performing a simple task (such as climbing onto a chair or touching a chair) and is expected to perform the same task on the command “Do it!” In a separate training session, the same dogs were trained to lie down regardless of the trainer’s actions.

Later, the trainer demonstrated a new simple action and said “Do it!” The dog then either repeated the new action, or repeated a previous trained action (such as lying down). The dogs were retested on the new simple action after one minute had passed, and after one hour had passed. A “success” was recorded if a dog performed the new simple action on the command “Do it!” before performing a previously trained action.

The article “Your Dog Remembers More Than You Think” (*Science*, November 23, 2016, scienmag.org/news/2016/11/your-dog-remembers-more-you-think, retrieved May 6, 2017) reports that dogs trained using this method recalled the correct new

Output for Exercise 10.75

action in 33 out of 35 trials. Suppose the data from this study will be used to determine if more than half of all dogs trained using this method would recall the correct new action.

- Explain why the data in this example should not be analyzed using a large-sample hypothesis test for one population proportion.
- Perform an exact binomial test for the null hypothesis that the proportion of all dogs trained using this method who would perform the correct new action is 0.5, versus the alternative hypothesis that the proportion is greater than 0.5.

10.77 Data from a large study carried out in 2008 were used to estimate that 10% of all smokers who quit smoking are smoking again after 1 year (["Relapse to Smoking After 1 Year of Abstinence: A Meta-analysis," ncbi.nlm.nih.gov/pmc/articles/PMC2577779/, June 8, 2008, retrieved May 6, 2017](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2577779/)).

The outcomes of many surgical procedures are improved for patients who are not smoking. In a University of Kansas Medical Center study (["Recidivism Rates After Smoking Cessation Before Spinal Fusion," healio.com/orthopedics/journals/ortho/2016-3-39-2/%7B28bf5c50-c2b8-413a-a662-f10efce6c9ef%7D/recidivism-rates-after-smoking-cessation-before-spinal-fusion, March 31, 2016, retrieved May 6, 2017](https://healio.com/orthopedics/journals/ortho/2016-3-39-2/%7B28bf5c50-c2b8-413a-a662-f10efce6c9ef%7D/recidivism-rates-after-smoking-cessation-before-spinal-fusion, March 31, 2016, retrieved May 6, 2017)), patients needing spinal fusion surgery were required to quit smoking before their surgery was scheduled. After 1 year, $n = 25$ of the patients responded to a follow-up survey, and 17 were smoking again. Assume it is reasonable to regard this sample as representative of people who quit smoking before surgery. The data from this sample are used to decide if there is convincing evidence that the proportion of people who quit smoking prior to surgery who are smoking again after 1 year is greater than 0.10.

- Explain why the data in this example should not be analyzed using a large-sample hypothesis test for one population proportion.

- Use the output at the bottom of the page from the Shiny app "Randomization Test for One Proportion" to help to complete an appropriate hypothesis test.

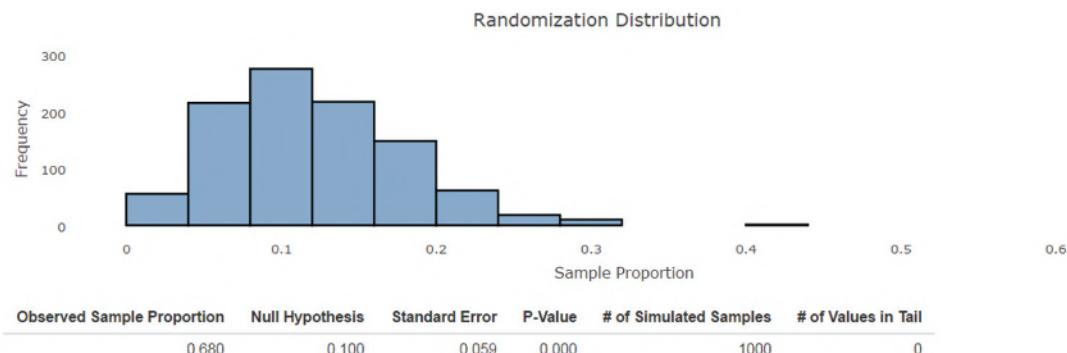
- Explain how the result of the hypothesis test performed may be related to the fact that the spinal fusion patients were required to stop smoking before their surgeries.

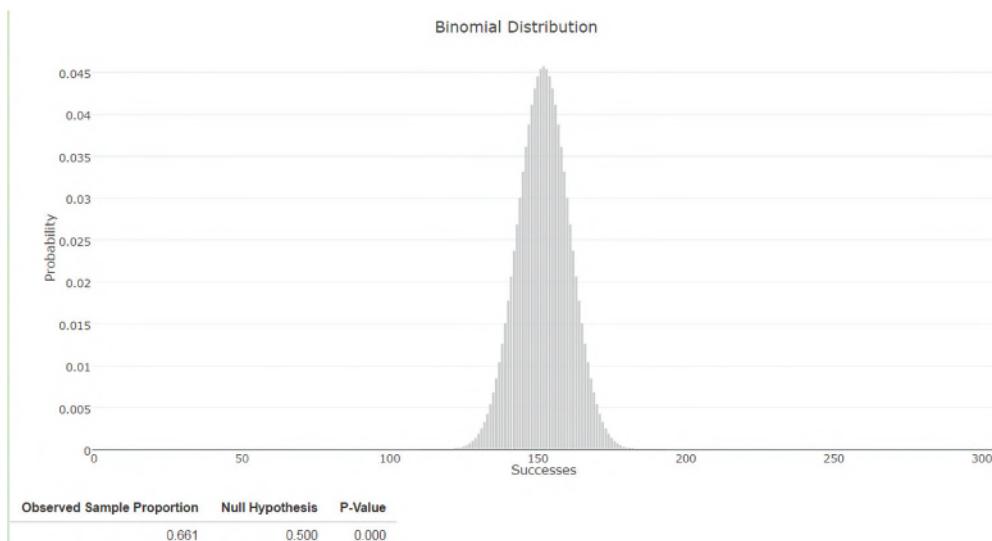
10.78 Recall that in Exercise 10.41, a survey of 304 U.S. businesses found that 201 indicated that they monitor employees' web site visits. In part (b) of that exercise, these data were used to determine if there is convincing evidence that a majority of businesses monitor employees' web site visits.

- What hypotheses were tested?
- Discuss whether the conditions necessary for a large-sample hypothesis test for one proportion are satisfied.
- The exact binomial test can be used even in cases when the sample size condition for the large-sample test is met. Use the output at the top of the following page from the Shiny app "Exact Binomial Test for One Proportion" to help to complete an exact binomial test for one proportion, testing whether there is sufficient evidence to conclude that a majority of U.S. businesses monitor employees' web site visits.
- Compare the P -value obtained in Part (c) with the P -value obtained in Exercise 10.41 Part (b) in Section 10.3. Would the same conclusion be reached in either case? Explain.

10.79 At one point during the 2015 NFL season, head coach Bill Belichick and the New England Patriots had won 19 of their past 25 called coin flips at the beginning of NFL games (["For Bill Belichick, Patriots' Strategy Is No Flip of the Coin," bostonglobe.com/sports/2015/11/04/pnotes/vFNt235bsK8x3JLZ6FJdtK/story.html, November 4, 2015, retrieved May 6, 2017](https://www.bostonglobe.com/sports/2015/11/04/pnotes/vFNt235bsK8x3JLZ6FJdtK/story.html, November 4, 2015, retrieved May 6, 2017)). Suppose that these 25 coin toss calls can be considered as

Output for Exercise 10.77



Output for Exercise 10.78

representative of all coin toss calls made by this team.

- a. Perform an exact binomial test to determine if there is convincing evidence that the proportion of all coin flip calls that the Patriots win is greater than 0.5.

- b. Discuss the conditions required for the exact binomial version of the hypothesis test. Write a brief explanation of why the results of the test performed in Part (a) do not necessarily mean that Coach Belichick is able to predict the results of coin flips better than other coaches.

SECTION 10.8 Randomization Test for a Population Mean (Optional)

The randomization approach introduced in Section 10.7 can also be used to approximate P -values for tests of hypotheses about a population mean.

Example 10.26 Goofing Off at Work (Revisited)

Recall the data presented in Example 10.14 of Section 10.4, regarding the amount of time a random sample of $n = 10$ employees spent wasting time at work, in minutes, on one day. The original sample data values are:

Employee ID	1	2	3	4	5	6	7	8	9	10
Minutes	108	112	117	130	111	131	113	113	105	128

The observed sample mean from the original sample is $\bar{x} = 116.8$ minutes.

Recall that the CEO wants to determine whether the mean wasted time per day at her company is less than 120 minutes. The distance between the claimed mean of 120 minutes and the sample mean from the original sample is $120 - 116.8 = 3.2$ minutes.

The data values in the original sample can be shifted to represent a sample from a hypothetical population with mean $\mu = 120$ minutes by adding 3.2 minutes to each, as follows:

Employee ID	1	2	3	4	5	6	7	8	9	10
Minutes	108	112	117	130	111	131	113	113	105	128
Minutes + 3.2	111.2	115.2	120.2	133.2	114.2	134.2	116.2	116.2	108.2	131.2

Let μ = Mean daily wasted time for employees of this company.

The relevant hypotheses are

$$\begin{aligned} H_0: \mu &= 120 \text{ minutes} \\ H_a: \mu &< 120 \text{ minutes} \end{aligned}$$

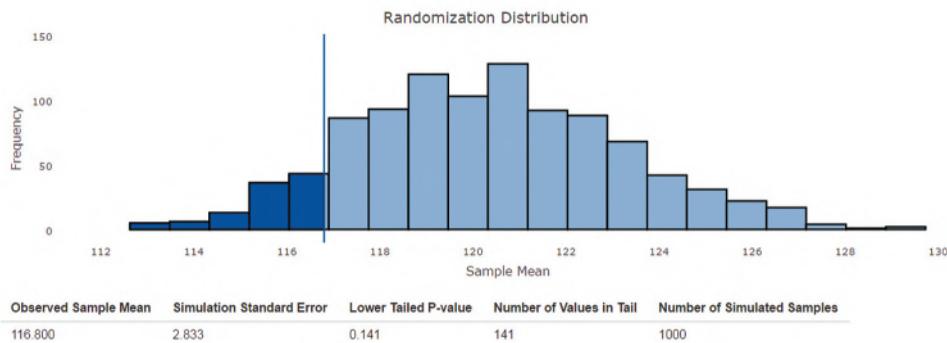
A randomization approach based on resampling can be used in this example, instead of a one-sample t -test. The sample was a random sample of employees of the company, and this is the only condition needed for the randomization test.

Begin by resampling from the sample that has been shifted to have a mean of 120 minutes. That is, draw new simulated samples, selected at random with replacement from the shifted sample. Here is one simulated sample from the shifted sample:

Employee ID	9	2	5	9	3	2	2	2	1	4
Minutes + 3.2	108.2	115.2	114.2	108.2	120.2	115.2	115.2	115.2	111.2	133.2

The sample mean for this simulated sample is $\bar{x} = 115.6$ minutes.

The randomization distribution is generated by selecting many random samples from the shifted sample. This distribution can be viewed as an approximate sampling distribution for \bar{x} under the assumption that the null hypothesis $\mu = 120$ is true. A P -value is the probability of obtaining a test statistic value at least as inconsistent with H_0 as what actually resulted. This means that the P -value for testing the null hypothesis $H_0: \mu = 120$ minutes against the one-sided alternative $H_a: \mu < 120$ minutes is the probability of observing $\bar{x} \leq 116.8$ when H_0 is true. This probability can be approximated using the proportion of simulated values of \bar{x} that are less than or equal to 116.8 in the randomization distribution. Below is output from the Shiny app “Randomization Test for One Mean.” This is one of the Shiny apps that accompany this text. These web apps are located at statistics.cengage.com/PSO6e/Apps.html.



The randomization test one-sided P -value is $P\text{-value} = 141/1000 = 0.141$. This P -value is greater than $\alpha = 0.05$, so we fail to reject H_0 . The P -value obtained from the one-sample t test in Example 10.14 was 0.150, so in this case, the conclusion in the hypothesis test does not change regardless of which of these two methods is used to carry out the test. Because the null hypothesis was not rejected, there is not convincing evidence that the mean time spent wasting time at the CEO’s company is less than the value from the larger study, 120 minutes.

EXERCISES 10.80 - 10.85

● Data set available online

- 10.80** ● The dodo was a species of flightless bird that lived on the island of Mauritius in the Indian Ocean. The first record of human interaction with the dodo occurred in 1598, and within 100 years the dodo was extinct due to hunting by humans and other newly introduced invasive species. After the extinction,

the word “dodo” became synonymous with stupidity, implying that the birds lacked the intelligence to avoid or escape extinction. The closest existing relatives of the dodo are pigeons and doves.

Researchers at the American Museum of Natural History used computed tomography scans to

measure the brain size (“endocranial capacity”) of one of the few existing preserved dodo birds, and to measure the brain sizes (reported in log mm³) in samples of eight birds that are close relatives of dodos (**“The First Endocast of the Extinct Dodo and an Anatomical Comparison Amongst Close Relatives,”** *Zoological Journal of the Linnean Society* [2016]: 950–953).

The brain size for the dodo was 4.17 log mm³. The following table contains the brain sizes for the sample of birds from related species (approximate values from a graph in the paper).

Brain Size	Brain Size
3.08	3.52
3.16	3.78
3.33	4.00
3.35	

- a. Use the Shiny app output at the bottom of the page to carry out a randomization test of the hypothesis that the population mean brain size for birds that are relatives of the dodo differs from the established dodo brain size of 4.17.
- b. What does the result of the test indicate about the brain size of the dodo?

- 10.81** Example 10.13 provided the following 21 time discrimination scores for male smokers who had abstained from smoking for 24 hours.

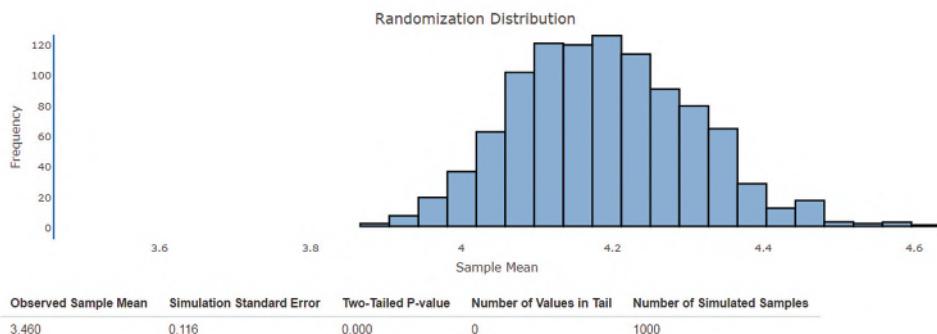
1.12 1.03 1.09 1.03 1.09 0.97 0.98 1.20 1.16
1.03 1.10 1.11 0.98 1.02 1.20 0.96 0.78 1.05
0.90 1.08 0.95

- a. What characteristic of the sample size indicates that the methods based on the *t* distribution may not be appropriate?
- b. Recall that a time discrimination score of 1 indicates that there is no tendency for time to be overestimated or underestimated. Use the following Shiny app output to help to test the hypothesis that the population mean time discrimination score for male smokers who abstain from smoking for 24 hours is significantly greater than 1.

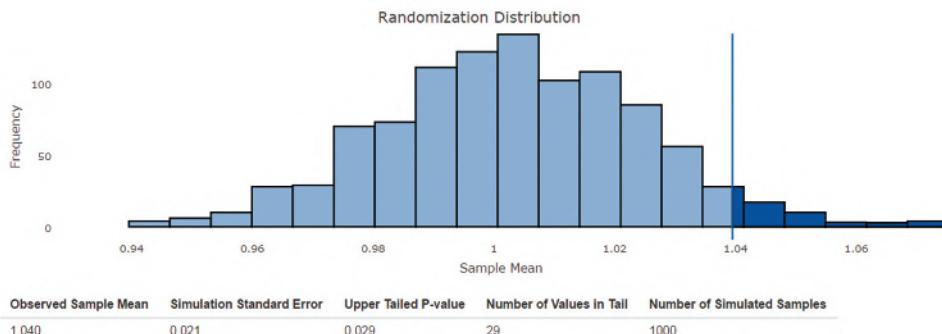
- 10.82** Teams in the National Football League (NFL) are given a “bye” during 1 week of the season, when they can rest and not play a game. This may provide an advantage for the team in the next game they play after a bye.

In 2016, each of the 32 NFL teams was granted a bye during one of the weeks of the season. The following table contains the team name and the number of points they won by or lost by in the game after the bye (espn.com/nfl/, retrieved December 22, 2016). A positive value indicates that the team coming off the bye won the game, and a negative value means that they lost. Consider these results to be a representative sample from a population of possible NFL match-ups between teams where one of the teams is coming off a bye week.

Output for Exercise 10.80



Output for Exercise 10.81



Team	Points	Team	Points
ARI	3	LA	-3
ATL	19	MIA	4
BAL	7	MIN	-11
BUF	4	NE	-7
CAR	10	NO	3
CHI	-26	NYG	5
CIN	-1	NYJ	-5
CLE	-13	OAK	7
DAL	6	PHI	-1
DEN	-3	PIT	-7
DET	7	SD	8
GB	7	SEA	2
HOU	3	SF	-18
IND	7	TB	17
JAX	1	TEN	3
KC	16	WSH	6

- a. Construct a graphical display for the data.

Although the sample size is at least on the borderline of being adequate for t distribution methods, what characteristic of the distribution indicates that the methods based on the t distribution may not be appropriate?

- b. Use a randomization test (Shiny app:

“Randomization Test for One Mean”) to perform a test of the hypothesis that the population mean points difference for NFL teams coming off a bye week differs from zero. Use a 0.05 significance level.

- c. Use the results from Part (b) to explain whether or not teams coming off a bye week have a significant advantage in points scored over their opponents.

- 10.83** Researchers studied ergometer (rowing machine) times for international male competitors in the 2007 World Junior Rowing Championships. They found that the mean time to row 2000 meters on an ergometer for the population of international sculls competitors was 387 seconds ([“Does 2000-m](#)

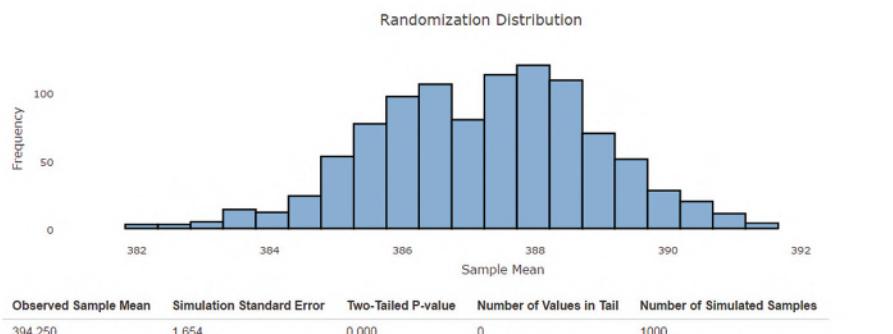
Rowing Ergometer Performance Time Correlate with Final Rankings at the World Junior Rowing Championship? A Case Study of 398 Elite Junior Rowers,” *Journal of Sports Sciences* [2009]: 361–366).

A related research team studied a sample of 24 junior male sculls rowers from only the United States in 2013–2014 and reported summary statistics for their 2000-meter ergometer times ([“Correlates of Performance at the U.S. Rowing Youth National Championships: A Case Study of 152 Junior Rowers,” *The Sport Journal*, March 3, 2014](#)). Data for a representative sample of 24 rowers that are consistent with summary statistics given in the paper are shown in the accompanying table.

Rower ID	Time (seconds)	Rower ID	Time (seconds)
1	391	13	404
2	389	14	399
3	386	15	390
4	396	16	395
5	395	17	392
6	409	18	401
7	406	19	381
8	402	20	392
9	393	21	376
10	392	22	400
11	382	23	388
12	401	24	402

- a. Use the Shiny app output at the bottom of the page to carry out a randomization test of the hypothesis that the population mean 2000-meter ergometer time for U.S. junior male sculls rowers differs from the 2007 international standard of 387 seconds. Use a significance level of 0.05.
- b. Based on the result of the hypothesis test, does it seem that the U.S. junior male sculls rowers have “caught up,” on average, with the international championship rowers from 2007? Explain.

Output for Exercise 10.83



- 10.84** • Exercise 10.53 asks whether a representative sample of Big Mac prices (after conversion to U.S. dollars) from countries in Europe provides evidence that the mean European price is less than the reported U.S. price of \$5.04. Here are the data:

4.44 3.15 2.42 3.96 4.35 4.51 4.17 3.69 4.62
3.80 3.36 3.85

- What characteristic of the sample indicates that the methods based on the t distribution may not be appropriate?
- Use the Shiny app output at the bottom of the page to test the hypothesis that the population mean price of a Big Mac in Europe is less than the reported U.S. price of \$5.04. Use a significance level of 0.05.

- 10.85** • Major League Baseball (MLB) includes two groups of teams, in two “leagues.” There are 15 teams in each of the American League (AL) and the National League (NL). Since 1997, teams in each of the leagues play teams from the other league in “interleague” regular-season games.

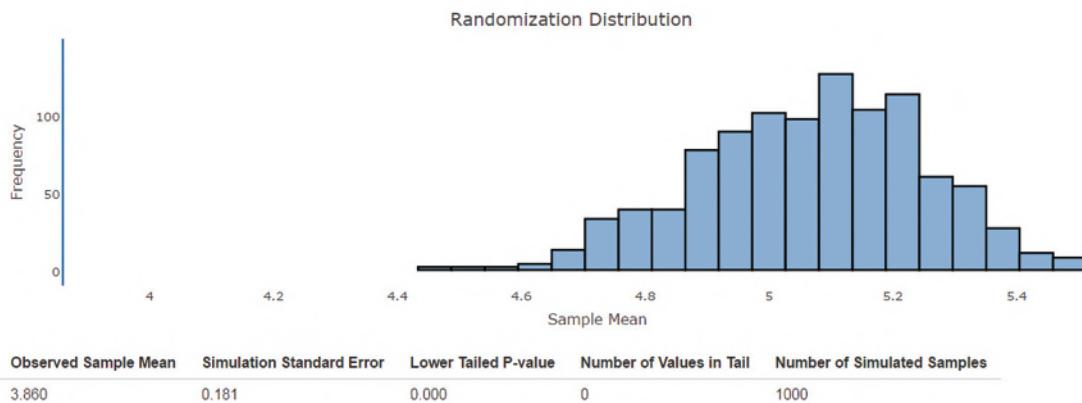
One way to determine whether one league is stronger than the other is to consider the interleague winning percentages for all the teams in one of the two leagues, say, the National League, for one season. For purposes of this exercise, consider the interleague games played in the 2016 season to be a representative sample of the performance of the teams in a population of potential future seasons.

Here are the 2016 interleague winning percentages for the 15 NL teams:

Team	W	L	Winning Percentage
ARI	5	15	25%
ATL	8	12	40%
CHC	15	5	75%
CIN	5	15	25%
COL	9	11	45%
LAD	10	10	50%
MIA	6	14	30%
MIL	11	9	55%
NYM	12	8	60%
PHI	11	9	55%
PIT	9	11	45%
SD	6	14	30%
SF	8	12	40%
STL	8	12	40%
WSH	12	8	60%

- What characteristic of this sample of NL team interleague winning percentages indicates that the methods based on the t distribution may not be appropriate?
- Use a randomization test (Shiny app: “Randomization Test for One Mean”) to perform a test of the hypothesis that the population mean interleague winning percentage for NL teams differs from 50%.
- Use the results from Part (b) to explain whether it is reasonable to say that the National League or the American League performs significantly better than the other in interleague play.

Output for Exercise 10.84



CHAPTER ACTIVITIES

ACTIVITY 10.1 COMPARING THE t AND z DISTRIBUTIONS

Technology Activity: Requires use of a computer or a graphing calculator.

The instructions that follow assume the use of Minitab. If you are using a different software package or a graphing calculator, your instructor will provide alternative instructions.

Background: Suppose a random sample will be selected from a population that is known to have a normal distribution. Then the statistic

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

has a standard normal (z) distribution. Since it is rarely the case that σ is known, inferences for population means are usually based on the statistic $t = \frac{\bar{x} - \mu}{(s/\sqrt{n})}$, which has a t distribution rather than a z distribution. The informal justification for this was that the use of s to estimate σ introduces additional variability, resulting in a statistic whose distribution is more spread out than is the z distribution.

In this activity, you will use simulation to sample from a known normal population and then investigate how the

behavior of $t = \frac{\bar{x} - \mu}{s/(\sqrt{n})}$ compares with the behavior of

$$z = \frac{\bar{x} - \mu}{\sigma/(\sqrt{n})}$$
.

- Generate 200 random samples of size 5 from a normal population with mean 100 and standard deviation 10.

Using Minitab, go to the Calc Menu. Then

Calc → Random Data → Normal

In the “Generate” box, enter 200

In the “Store in columns” box, enter c1-c5

In the mean box, enter 100

In the standard deviation box, enter 10

Click on OK

You should now see 200 rows of data in each of the first 5 columns of the Minitab worksheet.

- Each row contains five values that have been randomly selected from a normal population with mean 100 and standard deviation 10. Viewing each row as a sample of size 5 from this population, calculate the mean and standard deviation for each of the 200 samples (the 200 rows) by using Minitab’s row statistics functions, which can also be found under the Calc menu:

Calc → Row statistics

Choose the “Mean” button

In the “Input Variables” box, enter c1-c5

In the “Store result in” box, enter c7

Click on OK

You should now see the 200 sample means in column 7 of the Minitab worksheet. Name this column “ \bar{x} -bar” by typing the name in the gray box at the top of c7.

Now follow a similar process to compute the 200 sample standard deviations, and store them in c8. Name c8 “ s .”

- Next, calculate the value of the z statistic for each of the 200 samples. We can calculate z in this example because we know that the samples were selected from a population for which $\sigma = 10$.

Use the calculator function of Minitab to compute

$$z = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})} = \frac{\bar{x} - 100}{(10/\sqrt{5})} \text{ as follows:}$$

Calc → Calculator

In the “Store results in” box, enter c10

In the “Expression box” type in the following: (c7-100)/(10/sqrt(5))

Click on OK

You should now see the z values for the 200 samples in c11. Name c11 “ z .”

- Now calculate the value of the t statistic for each of the 200 samples. Use the calculator function of Minitab to compute $t = \frac{\bar{x} - \mu}{(s/\sqrt{n})} = \frac{\bar{x} - 100}{(s/\sqrt{5})}$ as follows:

Calc → Calculator

In the “Store results in” box, enter c11

In the “Expression box” type in the following: (c7-100)/(c8/sqrt(5))

Click on OK

You should now see the t values for the 200 samples in c10. Name c10 “ t .”

- Graphs, at last! Now construct histograms of the 200 z values and the 200 t values. These two graphical displays will provide insight about how each of these two statistics behaves in repeated sampling. Use the same scale for the two histograms so that it will be easier to compare the two distributions.

Graph → Histogram

In the “Graph variables” box, enter c10 for graph 1 and c11 for graph 2

Click the Frame dropdown menu and select multiple graphs.

Then under the scale choices, select “Same X and same Y.”

6. Now use the histograms from Step 5 to answer the following questions:
- Write a brief description of the shape, center, and spread for the histogram of the z values. Is what you see in the histogram consistent with what you would have expected to see? Explain. (Hint: In theory, what is the distribution of the z statistic?)
 - How does the histogram of the t values compare to the z histogram? Be sure to comment on center, shape, and spread.
 - Is your answer to Part (b) consistent with what would be expected for a statistic that has a t distribution? Explain.
 - The z and t histograms are based on only 200 samples, and they only approximate the corresponding sampling distributions. The 5th percentile for the standard normal distribution is -1.645 and the 95th percentile is $+1.645$. For a t distribution

with $df = 5 - 1 = 4$, the 5th and 95th percentiles are -2.13 and $+2.13$, respectively. How do these percentiles compare to those of the distributions displayed in the histograms? (Hint: Sort the 200 z values—in Minitab, choose “Sort” from the Manip menu. Once the values are sorted, percentiles from the histogram can be found by counting in 10 [which is 5% of 200] values from either end of the sorted list. Then repeat this with the t values.)

- Are the results of your simulation and analysis consistent with the statement that the statistic $z = \frac{\bar{x} - \mu}{(\sigma/\sqrt{n})}$ has a standard normal (z) distribution and the statistic $t = \frac{\bar{x} - \mu}{(s/\sqrt{n})}$ has a t distribution? Explain.

ACTIVITY 10.2 A MEANINGFUL PARAGRAPH

Write a meaningful paragraph that includes the following six terms: **hypotheses, P -value, reject H_0 , Type I error, statistical significance, practical significance**.

A “meaningful paragraph” is a coherent piece of writing in an appropriate context that uses all of the listed words. The paragraph should show that you understand the

meaning of the terms and their relationship to one another. A sequence of sentences that just define the terms is *not* a meaningful paragraph. When choosing a context, think carefully about the terms you need to use. Choosing a good context will make writing a meaningful paragraph easier.

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
Hypothesis	A claim about the value of a population characteristic.	$z = \frac{\hat{p} - \text{hypothesized value}}{\sqrt{(\text{hyp val})(1 - \text{hyp val})}}$	A test statistic for testing $H_0: p = \text{hypothesized value}$ when the sample size is large. The P -value is determined as an area under the z curve.
Null hypothesis, H_0	The hypothesis initially assumed to be true. It has the form $H_0: \text{population characteristic} = \text{hypothesized value}$.	$z = \frac{\bar{x} - \text{hypothesized value}}{\frac{\sigma}{\sqrt{n}}}$	A test statistic for testing $H_0: \mu = \text{hypothesized value}$ when σ is known and either the population distribution is normal or the sample size is large. The P -value is determined as an area under the z curve.
Alternative hypothesis, H_a	A hypothesis that specifies a claim that is contradictory to H_0 and is judged the more plausible claim when H_0 is rejected.	$t = \frac{\bar{x} - \text{hypothesized value}}{\frac{s}{\sqrt{n}}}$	A test statistic for testing $H_0: \mu = \text{hypothesized value}$ when σ is unknown and either the population distribution is normal or the sample size is large. The P -value is determined as an area under the t curve with $df = n - 1$.
Type I error	Rejecting H_0 when H_0 is true. The probability of a Type I error is denoted by α and is referred to as the significance level for the test.		The power of a test is the probability of rejecting the null hypothesis. Power is affected by the size of the difference between the hypothesized value and the actual value, the sample size, and the significance level.
Type II error	Not rejecting H_0 when H_0 is false. The probability of a Type II error is denoted by β .		
Test statistic	A value calculated from sample data that is then used as the basis for making a decision between H_0 and H_a .		
P-value	The probability, computed assuming H_0 is true, of obtaining a value of the test statistic at least as contradictory to H_0 as what actually resulted. H_0 is rejected if P -value $\leq \alpha$ and not rejected if P -value $> \alpha$, where α is the chosen significance level.		

CHAPTER REVIEW Exercises 10.86 - 10.100**● Data set available online**

- 10.86** The report “*A Crisis in Civic Education*” ([American Council of Trustees and Alumni, January 2016, goacta.org/images/download/A_Crisis_in_Civic_Education.pdf](http://americantrustees.org/images/download/A_Crisis_in_Civic_Education.pdf), retrieved November 30, 2016) summarizes data from a survey of a representative sample of 1000 adult Americans regarding their understanding of U.S. government. Only 459 of the adults in the sample were able to give a correct response to a question asking them to choose a correct definition of the Bill of Rights from a list of five possible answers. Using a significance level of 0.01, determine if there is convincing evidence that less than half of adult Americans could identify the correct definition of the Bill of Rights.
- 10.87** In a national survey of 2013 adults, 1590 responded that lack of respect and courtesy in American society is a serious problem, and 1283 indicated that they believe that rudeness is a more serious problem than in past years ([Associated Press, April 3, 2002](http://www.associatedpress.org/article/111313)). Is there convincing evidence that less than three-quarters of U.S. adults believe that rudeness is a worsening problem? Test the relevant hypotheses using a significance level of 0.05.
- 10.88** Students at the Akademia Podlaka conducted an experiment to determine whether the Belgium-minted Euro coin was equally likely to land heads up or tails up. Coins were spun on a smooth surface, and in 250 spins, 140 landed with the heads side up ([New Scientist, January 4, 2002](http://www.newscientist.com/article/mg17923690.100)).
- Should the students interpret this result as convincing evidence that the proportion of the time the coin would land heads up is not 0.5? Test the relevant hypotheses using $\alpha = 0.01$.
 - Would your conclusion be different if a significance level of 0.05 had been used? Explain.
- 10.89** The authors of the paper “*Mean Platelet Volume Could Be Possible Biomarker in Early Diagnosis and Monitoring of Gastric Cancer*” ([Platelets \[2014\]: 592–594](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3990523/)) wondered if mean platelet volume (MPV) might be a way to distinguish patients with gastric cancer from patients who did not have gastric cancer. MPV was recorded for 31 patients with gastric cancer. The sample mean was reported to be 8.31 femtoliters (fL) and the sample standard deviation was reported to be 0.78 fL. For a group of healthy people, the mean MPV was 7.85 fL. Is there convincing evidence that the mean MPV for patients with gastric cancer is greater than 7.85 fL? For purposes of this exercise, you can assume that the sample of 31 patients with gastric cancer is representative of the population of all patients with gastric cancer.
- 10.90** People in a random sample of 236 students enrolled at a liberal arts college were asked questions about how many hours of sleep they get each night (“*Alcohol Consumption, Sleep, and Academic Performance Among College Students*,” *Journal of Studies on Alcohol and Drugs* [2009]: 355–363). The sample mean sleep duration (average hours of daily sleep) was 7.71 hours and the sample standard deviation was 1.03 hours. The recommended number of hours of sleep for college-age students is 8.4 hours. Is there convincing evidence that the population mean sleep duration for students at this college is less than the recommended number of 8.4 hours? Test the relevant hypotheses using $\alpha = 0.01$.
- 10.91** *USA TODAY* reported that Americans spend 4.1 hours per weekday checking work email (October 14, 2016). This was an estimate based on a survey of 1004 white-collar workers in the United States.
- Suppose that you would like to know if there is evidence that the mean time spent checking work email for white-collar workers in the United States is more than half of the 8-hour workday. What would you need to assume about the sample to use the given sample data to answer this question?
 - Given that any concerns about the sample were satisfactorily addressed, carry out a test to decide if there is evidence that the mean time spent checking work email for white-collar workers in the United States is more than half of the 8-hour workday.
- 10.92** According to a large national survey conducted by the Pew Research Center (“*What Americans Think About NSA Surveillance, National Security and Privacy*,” May 2, 2015, [pewresearch.org, retrieved December 1, 2016](http://www.pewresearch.org/fact-tank/2015/05/02/what-americans-think-about-nsa-surveillance-national-security-and-privacy/)), 54% of adult Americans disapprove of the National Security Agency collecting records of phone and Internet data. Suppose that this estimate was based on a random sample of 1000 adult Americans.
- Is there convincing evidence that a majority of adult Americans feel this way? Test the relevant hypotheses using a 0.05 significance level.
 - The actual sample size was much larger than 1000. If you had used the actual sample size when doing the calculations for the test in Part (a), would the *P*-value have been larger than, the same as, or smaller than the *P*-value you obtained in Part (a)? Provide a justification for your answer.

- 10.93** In a representative sample of adult Americans age 26 to 32 years, 27% indicated that they owned a fitness band that kept track of the number of steps walked each day and their daily activity levels (*"Digital Democracy Survey," Deloitte Development LLC, 2016, deloitte.com/us/en.html, retrieved November 30, 2016.*) Suppose that the sample size was 500. Is there convincing evidence that more than one-quarter of all adult Americans in this age group own a fitness band? Test the relevant hypotheses using a significance level of 0.05.
- 10.94** The article *"Facebook Use and Academic Performance Among College Students"* (*Computers in Human Behavior [2015]: 265–272*) estimated that 87% percent of students at a large public university in California who are Facebook users update their status at least two times a day. This estimate was based on a random sample of size 261. Assume that this sample is representative of the students at this university.
- Does this sample provide convincing evidence that more than 80% of the students at this college who are Facebook users update their status at least two times a day? Test the relevant hypotheses using $\alpha = 0.05$.
 - Would it be reasonable to generalize the conclusion from the test in Part (a) to all college students in the United States? Explain why or why not.
- 10.95** A number of initiatives on the topic of legalized gambling have appeared on state ballots. A political candidate has decided to support legalization of casino gambling if he is convinced that more than two-thirds of U.S. adults approve of casino gambling. Suppose that 1523 adults selected at random from households with telephones were asked whether they approved of casino gambling. The number in the sample who approved was 1035. Does the sample provide convincing evidence that more than two-thirds of U.S. adults approve?
- 10.96** Duck hunting in populated areas faces opposition on the basis of safety and environmental issues. In a survey to assess public opinion regarding duck hunting on Morro Bay (located along the central coast of California), a random sample of 750 local residents included 560 who strongly opposed hunting on the bay. Does this sample provide sufficient evidence to conclude that the majority of local residents oppose hunting on Morro Bay? Test the relevant hypotheses using $\alpha = 0.01$.
- 10.97** Past experience has indicated that the response rate is 40% when individuals are approached with a request to fill out a questionnaire and return it in a stamped and addressed envelope. An investigator believes that if the person distributing the questionnaire is stigmatized in some obvious way, potential respondents would feel sorry for the distributor and tend to respond at a rate higher than 40%. To investigate this theory, a distributor is fitted with an eye patch. Of the 200 questionnaires distributed by this individual, 109 were returned. Does this provide convincing evidence that the response rate in this situation exceeds the rate in the past? State and test the appropriate hypotheses at significance level 0.05.
- 10.98** • An automobile manufacturer who wishes to advertise that one of its models achieves 30 mpg (miles per gallon) decides to carry out a fuel efficiency test. Six nonprofessional drivers were selected, and each one drove a car from Phoenix to Los Angeles. The resulting fuel efficiencies (in miles per gallon) are:
- | | | | | | |
|------|------|------|------|------|------|
| 27.2 | 29.3 | 31.2 | 28.4 | 30.3 | 29.6 |
|------|------|------|------|------|------|
- Assuming that fuel efficiency is normally distributed under these circumstances, do the data contradict the claim that true average fuel efficiency is (at least) 30 mpg?
- 10.99** A student organization uses the proceeds from a particular soft-drink dispensing machine to finance its activities. The price per can had been \$0.75 for a long time, and the average daily revenue during that period was \$75.00. The price was recently increased to \$1.00 per can. A random sample of $n = 20$ days after the price increase yielded a sample mean daily revenue and sample standard deviation of \$70.00 and \$4.20, respectively.
- Does this information suggest that the mean daily revenue has decreased from its value before the price increase? Test the appropriate hypotheses using $\alpha = 0.05$.
- 10.100** A hot tub manufacturer advertises that with its heating equipment, a temperature of 100°F can be achieved on average in 15 minutes or less. A random sample of 25 tubs is selected, and the time necessary to achieve a 100°F temperature is determined for each tub. The sample mean time and sample standard deviation are 17.5 minutes and 2.2 minutes, respectively. Does this information cast doubt on the company's claim? Carry out a test of hypotheses using significance level 0.05.

TECHNOLOGY NOTES

***z* Test for Proportions**

JMP

Summarized data

1. Enter the data into the JMP data table with categories in the first column and counts in the second column
2. Click **Analyze** and select **Distribution**
3. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Columns**
4. Click and drag the second column name from the box under **Select Columns** to the box next to **Freq**
5. Click **OK**
6. Click the red arrow next to the column name and click **Test Probabilities**
7. Under the **Test Probabilities** section that appears in the output, click in the box across from **Yes** under the **Hypothe Prob** and type the hypothesized value for p
8. Select the appropriate option for alternative
9. Click **Done**

Note: In the two-sided case, JMP uses the square of the z test statistic, called the Chi-Square test statistic. The two methods are mathematically identical.

Note: In the one-sided cases, JMP uses the exact binomial test rather than the z test.

Raw data

1. Enter the raw data into a column

	Column 1	Column 2
1	Yes	130
2	No	45

2. Click **Analyze** and select **Distribution**
3. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Columns**
4. Click **OK**
5. Click the red arrow next to the column name and click **Test Probabilities**
6. Under the **Test Probabilities** section that appears in the output, click in the box across from **Yes** under the **Hypothe Prob** and type the hypothesized value for p
7. Select the appropriate option for alternative
8. Click **Done**

Note: In the two-sided case, JMP uses the square of the z test statistic, called the Chi-Square test statistic. The two methods are mathematically identical.

Note: In the one-sided cases, JMP uses the exact binomial test rather than the z test.

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **1 Proportion...**
2. Click the radio button next to **Summarized data**
3. In the box next to **Number of Trials:** type the value for n , the sample size
4. In the box next to **Number of events:** type the value for the number of successes
5. Click **Options...**
6. Input the appropriate hypothesized value in the box next to **Test proportion:**
7. Select the appropriate alternative hypothesis from the drop-down menu next to **Alternative**
8. Check the box next to **Use test and interval based on normal distribution**
9. Click **OK**
10. Click **OK**

Raw data

1. Input the raw data into a column
2. Click **Stat** then click **Basic Statistics** then click **1 Proportion...**
3. Click in the box under **Samples in columns:**
4. Double click the column name where the raw data are stored
5. Click **Options...**
6. Select the appropriate alternative hypothesis from the drop-down menu next to **Alternative**
7. Check the box next to **Use test and interval based on normal distribution**
8. Click **OK**
9. Click **OK**

SPSS

SPSS does not have the functionality to automatically calculate a z test for a testing a single proportion.

Excel

Excel does not have the functionality to automatically calculate a z test for a testing a single population proportion. You may type in the formula by hand into a cell to have Excel calculate the value of the test statistic for you. Then use methods from Chapter 6 to find the P -value using the Normal Distribution.

TI-83/84

1. Press the **STAT** key
2. Highlight **TESTS**
3. Highlight **1-PropZTest...** and press **ENTER**
4. Next to **p0** type the hypothesized value for p
5. Next to **x** type the number of successes
6. Next to **n** type the sample size, n
7. Next to **prop**, highlight the appropriate alternative hypothesis
8. Highlight **Calculate** and press **ENTER**

TI-Nspire

1. Enter the Calculate Scratchpad
2. Press the **menu** key then select **6:Statistics** then select **7:Stat Tests** then **5:1-Prop z Test...** then press **enter**
3. In the box next to **p0** type the hypothesized value for p
4. In the box next to **Successes, x** type the number of successes
5. In the box next to **n** type the sample size, n
6. In the box next to **Alternate Hyp** choose the appropriate alternative hypothesis from the drop-down menu
7. Press **OK**

***t* Test for Population Mean, μ** **JMP**

1. Input the data into a column
2. Click **Analyze** and select **Distribution**
3. Click and drag the column name from the box under **Select Columns** to the box next to **Y, Response**
4. Click **OK**
5. Click the red arrow next to the column name and select **Test Mean**
6. In the box next to **Specify Hypothesized Mean**, type the hypothesized value of the mean, μ_0
7. Click **OK**

Note: The output provides results for all three possible alternative hypotheses.

Minitab**Summarized data**

1. Click **Stat** then click **Basic Statistics** then click **1-sample t...**
2. Click the radio button next to **Summarized data**
3. In the box next to **Sample size:** type the value for n , the sample size
4. In the box next to **Mean:** type the value for the sample mean
5. In the box next to **Standard deviation:** type the value for the sample standard deviation
6. In the box next to **Test mean:** type the hypothesized value of the population mean
7. Click **Options...**
8. Select the appropriate alternative hypothesis from the drop-down menu next to **Alternative:**
9. Click **OK**
10. Click **OK**

Raw data

1. Input the raw data into a column
2. Click **Stat** then click **Basic Statistics** then click **1-sample t...**
3. Click in the box under **Samples in columns:**
4. Double click the column name where the raw data are stored
5. In the box next to **Test mean:** type the hypothesized value of the population mean
6. Click **Options...**
7. Select the appropriate alternative hypothesis from the drop-down menu next to **Alternative:**
8. Click **OK**
9. Click **OK**

SPSS

1. Input the data into a column
2. Click **Analyze** then select **Compare Means** then select **One-Sample T Test...**
3. Highlight the column name for the variable
4. Click the arrow to move the variable to the **Test Variable(s):** box
5. In the box next to **Test Value:** input the hypothesized test value
6. Click **OK**

Note: This procedure produces a two-sided P -value.

Excel

Excel does not have the functionality to automatically produce a t test for a single population mean. However, you may type the formula into an empty cell manually to have Excel calculate the value of the test statistic for you. You can then use the steps below to find a P -value for the test statistic.

1. Select an empty cell
2. Click on **Formulas**
3. Click **Insert Function**
4. Select **Statistical** from the drop-down box for category
5. Select **TDIST** and click **OK**
6. Click in the box next to **X** and select the cell containing your test statistic or type it manually
7. Click in the box next to **Deg_freedom** and type the number of degrees of freedom ($n-1$)
8. Click in the box next to **Tails** and type 1 for a one-tailed P -value or 2 for a two-tailed P -value
9. Click **OK**

Note: Choosing a one-tailed distribution in Step 8 will result in returning $P(X \geq x)$.

TI-83/84**Summarized data**

1. Press **STAT**
2. Highlight **TESTS**
3. Highlight **T-Test...**
4. Highlight **Stats** and press **ENTER**
5. Next to μ_0 type the hypothesized value for the population mean
6. Next to \bar{x} input the value for the sample mean
7. Next to s_x input the value for the sample standard deviation
8. Next to n input the value for the sample size
9. Next to μ highlight the appropriate alternative hypothesis and press **ENTER**
10. Highlight **Calculate** and press **ENTER**

Raw data

1. Enter the data into **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**
4. Highlight **T-Test...**
5. Highlight **Data** and press **ENTER**
6. Next to μ_0 type the hypothesized value for the population mean
7. Next to μ highlight the appropriate alternative hypothesis and press **ENTER**
8. Highlight **Calculate** and press **ENTER**

TI-Nspire**Summarized data**

1. Enter the Calculate Scratchpad
2. Press the **menu** key and select **6:Statistics** then **7:Stat Tests** then **2:t test...** and press **enter**
3. From the drop-down menu select **Stats**
4. Press **OK**
5. Next to μ_0 type the hypothesized value for the population mean
6. Next to \bar{x} input the value for the sample mean
7. Next to s_x input the value for the sample standard deviation
8. Next to n input the value for the sample size
9. Next to **Alternate Hyp** select the appropriate alternative hypothesis from the drop-down menu
10. Press **OK**

Raw data

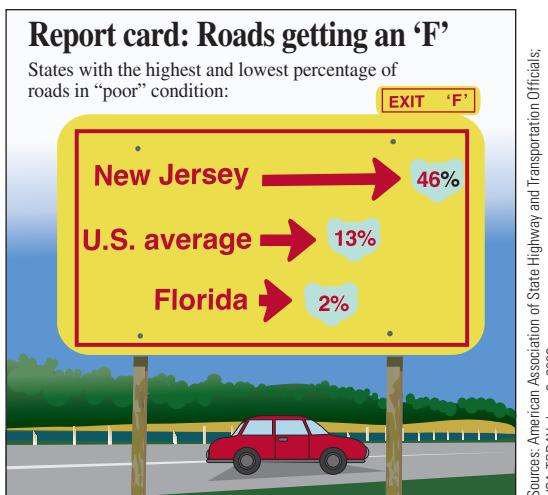
1. Enter the data into a data list (In order to access data lists select the spreadsheet option and press **enter**)
Note: Be sure to title the list by selecting the top row of the column and typing a title.
2. Press the **menu** key and select **4:Statistics** then **7:Stat Tests** then **2:t test...** and press **enter**
3. From the drop-down menu select **Data**
4. Press **OK**
5. Next to μ_0 input the hypothesized value for the population mean
6. Next to **List** select the list containing your data
7. Next to **Alternate Hyp** select the appropriate alternative hypothesis from the drop-down menu
8. Press **OK**

CUMULATIVE REVIEW EXERCISES**CR10.1 - CR10.16**

● Data set available online

CR10.1 The *AARP Bulletin* (March 2010) included the following short news brief: “Older adults who did 1 hour of tai chi twice weekly cut their pain from knee osteoarthritis considerably in a 12-week study conducted at Tufts University School of Medicine.” Suppose you were asked to design a study to investigate this claim. Describe an experiment that would allow comparison of the reduction in knee pain for those who did 1 hour of tai chi twice weekly to the reduction in knee pain for those who did not do tai chi. Include a discussion of how study participants would be selected, how pain reduction would be measured, and how participants would be assigned to experimental groups.

CR10.2 The following graphical display is similar to one that appeared in *USA TODAY*, (June 3, 2009). Write a few sentences critiquing this graphical display. Do you think it does a good job of creating a visual representation of the three percentages in the display?



CR10.3 ● The article “Flyers Trapped on Tarmac Push for Rules on Release” (*USA TODAY*, July 28, 2009) included the

accompanying data on the number of flights with a tarmac delay of more than 3 hours between October 2008 and May 2009 for U.S. airlines.

Airline	Number of Flights	Rate per 100,000 Flights
AirTran	7	0.4
Alaska	0	0.0
American	48	1.3
American Eagle	44	1.6
Atlantic Southeast	11	0.6
Comair	29	2.7
Continental	72	4.1
Delta	81	2.8
ExpressJet	93	4.9
Frontier	5	0.9
Hawaiian	0	0.0
JetBlue	18	1.4
Mesa	17	1.1
Northwest	24	1.2
Pinnacle	13	0.7
SkyWest	29	0.8
Southwest	11	0.1
United	29	1.1
US Airways	46	1.6

- a. Construct a dotplot of the data on number of flights delayed for more than 3 hours. Are there any unusual observations that stand out in the dot plot? What airlines appear to be the worst in terms of number of flights delayed on the tarmac for more than 3 hours?
- b. Construct a dotplot of the data on rate per 100,000 flights. Write a few sentences describing the interesting features of this plot.

- c. If you wanted to compare airlines on the basis of tarmac delays, would you recommend using the data on number of flights delayed or on rate per 100,000 flights? Explain the reason for your choice.

CR10.4 • The article “[Wait Times on Rise to See Doctor](#)” (*USA TODAY*, June 4, 2009) gave the accompanying data on average wait times in days to get an appointment with a medical specialist in 15 U.S. cities. Construct a boxplot of the average wait-time data. Are there any outliers in the data set?

City	Average Appointment Wait Time
Atlanta	11.2
Boston	49.6
Dallas	19.2
Denver	15.4
Detroit	12.0
Houston	23.4
Los Angeles	24.2
Miami	15.4
Minneapolis	19.8
New York	19.2
Philadelphia	27.0
Portland	14.4
San Diego	20.2
Seattle	14.2
Washington, D.C.	22.6

CR10.5 The report “[New Study Shows Need for Americans to Focus on Securing Online Accounts and Backing up Critical Data](#)” (*PRNewswire*, October 29, 2009) states that only 25% of Americans change computer passwords quarterly, in spite of a recommendation from the National Cyber Security Alliance that passwords be changed at least once every 90 days. For purposes of this exercise, assume that the 25% figure is correct for the population of adult Americans.

- a. If a random sample of 20 adult Americans is selected, what is the probability that exactly 3 of them change passwords quarterly?
- b. What is the probability that more than 8 people in a random sample of 20 adult Americans change passwords quarterly?
- c. What is the mean and standard deviation of the variable $x = \text{number of people in a random sample of 100 adult Americans who change passwords quarterly}$?
- d. Find the approximate probability that the number of people who change passwords quarterly in a random sample of 100 adult Americans is less than 20.

CR10.6 The article “[Should Canada Allow Direct-to-Consumer Advertising of Prescription Drugs?](#)” (*Canadian*

Family Physician [2009]: 130–131) calls for the legalization of advertising of prescription drugs in Canada. Suppose you wanted to conduct a survey to estimate the proportion of Canadians who would support allowing this type of advertising. How large a random sample would be required to estimate this proportion to within 0.02 with 95% confidence?

CR10.7 The [National Association of Colleges and Employers](#) carries out a student survey each year. A summary of data from the 2009 survey included the following information:

- 26% of students graduating in 2009 intended to go on to graduate or professional school.
- Only 40% of those who graduated in 2009 received at least one job offer prior to graduation.
- Of those who received a job offer, only 45% had accepted an offer by the time they graduated.

Consider the following events:

O = event that a randomly selected 2009 graduate received at least one job offer

A = event that a randomly selected 2009 graduate accepted a job offer prior to graduation

G = event that a randomly selected 2009 graduate plans to attend graduate or professional school

Calculate the following probabilities.

- a. $P(O)$
- b. $P(A)$
- c. $P(G)$
- d. $P(A|O)$
- e. $P(O|A)$
- f. $P(A \cap O)$

CR10.8 It probably wouldn’t surprise you to know that Valentine’s Day means big business for florists, jewelry stores, and restaurants. But would it surprise you to know that it is also a big day for pet stores? In January 2010, the National Retail Federation conducted a survey of consumers who they believed were selected in a way that would produce a sample representative of the population of adults in the United States (“[This Valentine’s Day, Couples Cut Back on Gifts to Each Other, According to NRF Survey](#),” *nrf.com*). One of the questions in the survey asked if the respondent planned to spend money on a Valentine’s Day gift for his or her pet this year.

- a. The proportion who responded that they did plan to purchase a gift for their pet was 0.173. Suppose that the sample size for this survey was $n = 200$. Construct and interpret a 95% confidence interval for the proportion of all U.S. adults who planned to purchase a Valentine’s Day gift for their pet in 2010.
- b. The actual sample size for the survey was much larger than 200. Would a 95% confidence interval computed using the actual sample size have been narrower or wider than the confidence interval computed in Part (a)?
- c. Still assuming a sample size of $n = 200$, carry out a hypothesis test to determine if the data provides

convincing evidence that the proportion who planned to buy a Valentine's Day gift for their pet in 2010 was greater than 0.15. Use a significance level of 0.05.

CR10.9 The article “[Doctors Cite Burnout in Mistakes](#)” (*San Luis Obispo Tribune*, March 5, 2002) reported that many doctors who are completing their residency have financial struggles that could interfere with training. In a sample of 115 residents, 38 reported that they worked moonlighting jobs and 22 reported a credit card debt of more than \$3000. Suppose that it is reasonable to consider this sample of 115 as a random sample of all medical residents in the United States.

- a. Construct and interpret a 95% confidence interval for the proportion of U.S. medical residents who work moonlighting jobs.
- b. Construct and interpret a 90% confidence interval for the proportion of U.S. medical residents who have a credit card debt of more than \$3000.
- c. Give two reasons why the confidence interval in Part (a) is wider than the confidence interval in Part (b).

CR10.10 The National Geographic Society conducted a study that included 3000 respondents, age 18 to 24, in nine different countries (*San Luis Obispo Tribune*, November 21, 2002). The society found that 10% of the participants could not identify their own country on a blank world map.

- a. Construct a 90% confidence interval for the proportion who can identify their own country on a blank world map.
- b. What assumptions are necessary for the confidence interval in Part (a) to be valid?
- c. To what population would it be reasonable to generalize the confidence interval estimate from Part (a)?

CR10.11 “[Heinz Plays Catch-up After Under-Filling Ketchup Containers](#)” is the headline of an article that appeared on CNN.com (November 30, 2000). The article stated that Heinz had agreed to put an extra 1% of ketchup into each ketchup container sold in California for a 1-year period. Suppose that you want to make sure that Heinz is in fact fulfilling its end of the agreement. You plan to take a sample of 20-oz bottles shipped to California, measure the amount of ketchup in each bottle, and then use the resulting data to estimate the mean amount of ketchup in 20-oz bottles. A small pilot study showed that the amount of ketchup in 20-oz bottles varied from 19.9 to 20.3 oz. How many bottles should be included in the sample if you want to estimate the true mean amount of ketchup to within 0.1 oz with 95% confidence?

CR10.12 In a survey conducted by Yahoo Small Business, 1432 of 1813 adults surveyed said that they would alter their shopping habits if gas prices remain high ([Associated Press, November 30, 2005](#)). The article did not say how the sample was selected, but for purposes of this exercise, assume that it is reasonable to regard this sample as representative of adult Americans. Based on these survey data, is it reasonable to conclude that more than three-quarters of adult Americans plan to alter their shopping habits if gas prices remain high?

CR10.13 In an AP-AOL sports poll ([Associated Press, December 18, 2005](#)), 272 of 394 randomly selected baseball fans stated that they thought the designated hitter rule should either be expanded to both baseball leagues or eliminated. Based on the given information, is there sufficient evidence to conclude that a majority of baseball fans feel this way?

CR10.14 The article titled “[13% of Americans Don't Use the Internet. Who Are They?](#)” describes a study conducted by the Pew Research Center ([pewresearch.org, September 7, 2016, retrieved December 1, 2016](#)). Suppose that the title of this article is based on a representative sample of 600 adult Americans. Does this support the claim that the proportion of adult Americans who do not use the Internet is greater than 0.10 (10%)?

CR10.15 A survey of teenagers and parents in Canada conducted by the polling organization Ipsos (“[Untangling the Web: The Facts About Kids and the Internet](#),” January 25, 2006) included questions about Internet use. It was reported that for a sample of 534 randomly selected teens, the mean number of hours per week spent online was 14.6 and the standard deviation was 11.6.

- a. What does the large standard deviation, 11.6 hours, tell you about the distribution of online times for this sample of teens?
- b. Do the sample data provide convincing evidence that the mean number of hours that teens spend online is greater than 10 hours per week?

CR10.16 The same survey referenced in the previous exercise reported that for a random sample of 676 parents of Canadian teens, the mean number of hours parents thought their teens spent online was 6.5 and the sample standard deviation was 8.6.

- a. Do the sample data provide convincing evidence that the mean number of hours that parents think their teens spend online is less than 10 hours per week?
- b. Write a few sentences commenting on the results of the test in Part (a) and of the test in Part (b) of the previous exercise.



Jacob Lund/Shutterstock.com

Many investigations are carried out for the purpose of comparing two populations or treatments. For example, the article “[What Do Happy People Do?](#)” (*Social Indicators Research [2008]: 565–571*) investigates differences in the way happy people and unhappy people spend their time. By comparing data from a large national sample of people who described themselves as very happy to data from a large national sample of people who described themselves as not happy, the authors were able to investigate whether the mean amount of time spent in various activities was greater for one group than for the other.

Using hypothesis tests to be introduced in this chapter, the authors were able to conclude that there was no significant difference in the mean number of hours per day spent on the Internet for happy and unhappy people but that the mean number of hours per day spent watching TV was significantly higher for unhappy people.

In this chapter, we will see hypothesis tests and confidence intervals that can be used to compare two populations or treatments.

LEARNING OBJECTIVES

Students will understand:

- The difference between paired and independent samples.

Students will be able to:

- Distinguish between paired and independent samples.
- Carry out a test for a difference in population means using independent samples or paired samples and interpret the results in context.
- Carry out a test for a difference in population proportions and interpret the results in context.
- Construct a confidence interval estimate of a difference in population means using independent samples or paired samples and interpret the interval in context.
- Construct a confidence interval estimate of a difference in population proportions using independent samples and interpret the interval in context.
- Calculate and interpret a bootstrap confidence interval for a difference in proportions. (Optional)
- Carry out a randomization test for a difference in proportions. (Optional)
- Calculate and interpret a bootstrap confidence interval for a difference in means. (Optional)
- Carry out a randomization test for a difference in means. (Optional)

SECTION 11.1

Inferences Concerning the Difference Between Two Population or Treatment Means Using Independent Samples

In this section, we consider using sample data to compare two population means or two treatment means. For example, many researchers have studied the ways in which college students use Facebook. As part of a study described in the paper “**Facebook Use and Academic Performance Among College Students: A Mixed-Methods Study with a Multi-Ethnic Sample**” (*Computers in Human Behavior* [2015]: 265–272), each person in a sample of 195 female Facebook users and an independent sample of 66 male Facebook users was asked to report the amount of time per day he or she spent on Facebook. The samples were chosen to be representative of female and male college students in Southern California. The authors of the paper were interested in learning whether the mean time spent by female Facebook users was greater than the mean time spent by male Facebook users. In this chapter, you will see how sample data can be used to answer questions like this that involve comparing two population means.

In other situations, an experiment might be carried out to compare two different treatments or to compare the effect of a treatment with the effect of no treatment. For example, an agricultural researcher might want to compare weight gains for animals placed on two different diets (each diet is a treatment). An educational researcher might want to compare online instruction to traditional classroom instruction by studying the difference in mean scores on a common final exam (each type of instruction is a treatment).

In previous chapters, the symbol μ was used to denote the mean of a single population under study. When comparing two populations or treatments, we use notation that distinguishes between the characteristics of the first and those of the second. This is accomplished by using subscripts, as shown in the accompanying box.

Notation

	Mean	Variance	Standard Deviation
Population or Treatment 1	μ_1	σ_1^2	σ_1
Population or Treatment 2	μ_2	σ_2^2	σ_2

	Sample Size	Sample Mean	Sample Variance	Standard Deviation
Sample from Population or Treatment 1	n_1	\bar{x}_1	s_1^2	s_1
Sample from Population or Treatment 2	n_2	\bar{x}_2	s_2^2	s_2

A comparison of means focuses on the difference, $\mu_1 - \mu_2$. When $\mu_1 - \mu_2 = 0$, the two population or treatment means are equal. That is,

$$\mu_1 - \mu_2 = 0 \text{ is equivalent to } \mu_1 = \mu_2$$

Similarly,

$$\mu_1 - \mu_2 > 0 \text{ is equivalent to } \mu_1 > \mu_2$$

and

$$\mu_1 - \mu_2 < 0 \text{ is equivalent to } \mu_1 < \mu_2$$

Before developing methods for drawing conclusions about $\mu_1 - \mu_2$, we must consider how the two samples, one from each population or treatment, are selected. Two samples

are said to be **independent** samples if the selection of the individuals or objects that make up one sample does not influence the selection of individuals or objects in the other sample.

When observations from the first sample are linked in some meaningful way with observations in the second sample, the samples are said to be **paired**. For example, to study the effectiveness of a speed-reading course, the reading speed of subjects could be measured before they take the course and again after they complete the course. This results in two related samples—one from the population of individuals who have not taken this particular course (the “before” measurements) and one from the population of individuals who have taken the course (the “after” measurements). These samples are paired. The two samples are not independently chosen, because the selection of individuals from the first (before) population completely determines which individuals make up the sample from the second (after) population. In Section 11.1, we consider procedures based on independent samples. Methods for analyzing data resulting from paired samples are presented in Section 11.2.

Because \bar{x}_1 provides an estimate of μ_1 and \bar{x}_2 gives an estimate of μ_2 , it is natural to use $\bar{x}_1 - \bar{x}_2$ as a point estimate of $\mu_1 - \mu_2$. The value of \bar{x}_1 varies from sample to sample (it is a *statistic*), as does the value of \bar{x}_2 . Since the difference $\bar{x}_1 - \bar{x}_2$ is calculated from sample values, it is also a statistic and has a sampling distribution.

Properties of the Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

For independent random samples:

$$1. \mu_{\bar{x}_1 - \bar{x}_2} = \left(\begin{array}{l} \text{mean value} \\ \text{of } \bar{x}_1 - \bar{x}_2 \end{array} \right) = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$$

This means that the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is always centered at the value of $\mu_1 - \mu_2$, so $\bar{x}_1 - \bar{x}_2$ is an unbiased statistic for estimating $\mu_1 - \mu_2$.

$$2. \sigma_{\bar{x}_1 - \bar{x}_2}^2 = \left(\begin{array}{l} \text{variance of} \\ \bar{x}_1 - \bar{x}_2 \end{array} \right) = \sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

and

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \left(\begin{array}{l} \text{standard deviation} \\ \text{of } \bar{x}_1 - \bar{x}_2 \end{array} \right) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

3. If n_1 and n_2 are both large or the population distributions are (at least approximately) normal, \bar{x}_1 and \bar{x}_2 each have (at least approximately) a normal distribution. This implies that the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is also normal or approximately normal.

Properties 1 and 2 follow from the following general results:

1. The mean value of a difference in means is the difference of the two individual mean values.
2. The variance of a difference of *independent* quantities is the *sum* of the two individual variances.

When the sample sizes are large or when the population distributions are approximately normal, the properties of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ imply that $\bar{x}_1 - \bar{x}_2$ can be standardized to obtain a variable with a sampling distribution that is approximately the standard normal (z) distribution. This leads to the following result.

When two random samples are independently selected and when n_1 and n_2 are both large or the population distributions are normal, the statistic

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has a standard normal (z) distribution.

Although it is possible to base a test procedure and confidence interval on this result, the values of σ_1^2 and σ_2^2 are usually not known. When σ_1^2 and σ_2^2 are not known, we estimate them using the corresponding sample variances, s_1^2 and s_2^2 . This leads to the result given in the accompanying box.

When two random samples are independently selected and when n_1 and n_2 are both large or when the population distributions are normal, the standardized variable

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

has approximately a t distribution with

$$\text{df} = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} \quad \text{where} \quad V_1 = \frac{s_1^2}{n_1} \quad \text{and} \quad V_2 = \frac{s_2^2}{n_2}$$

The calculated value of df should be truncated (rounded down) to obtain an integer value of df.

If one or both sample sizes are small, we can use normal probability plots or boxplots to evaluate whether it is reasonable to consider the population distributions to be approximately normal.

Hypothesis Tests

In a test designed to compare two population means, the null hypothesis is of the form

$$H_0: \mu_1 - \mu_2 = \text{hypothesized value}$$

Often the hypothesized value is 0, indicating that there is no difference between the population means. The alternative hypothesis involves the same hypothesized value but uses one of three inequalities (less than, greater than, or not equal to), depending on the research question of interest. As an example, suppose μ_1 and μ_2 denote the average fuel efficiencies (in miles per gallon, mpg) for cars with 4-cylinder engines and cars with 6-cylinder engines, respectively. The hypotheses under consideration might be

$$H_0: \mu_1 - \mu_2 = 5 \quad \text{versus} \quad H_a: \mu_1 - \mu_2 > 5$$

The null hypothesis is equivalent to the claim that the mean fuel efficiency for the 4-cylinder engine cars exceeds the mean fuel efficiency for the 6-cylinder engine cars by 5 mpg. The alternative hypothesis states that the difference between the mean fuel efficiencies is more than 5 mpg.

A test statistic is obtained by replacing $\mu_1 - \mu_2$ in the standardized t variable (given in the box above) with the hypothesized value that appears in H_0 . For example, the t statistic for testing $H_0: \mu_1 - \mu_2 = 5$ is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 5}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

When the sample sizes are large or when the population distributions are normal, the sampling distribution of the test statistic is approximately a t distribution when H_0 is true. The P -value for the test is obtained by first calculating the appropriate number of degrees of freedom and then using Appendix Table 4, a graphing calculator, or a statistical software package. The following box gives a general description of the test procedure.

Summary of the Two-Sample t Test for Comparing Two Population Means

Null hypothesis: $H_0: \mu_1 - \mu_2 = \text{hypothesized value}$

Test statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - \text{hypothesized value}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

The appropriate df for the two-sample t test is

$$\text{df} = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} \quad \text{where} \quad V_1 = \frac{s_1^2}{n_1} \quad \text{and} \quad V_2 = \frac{s_2^2}{n_2}$$

The calculated number of degrees of freedom should be truncated (rounded down) to an integer.

Alternative hypothesis:

$H_a: \mu_1 - \mu_2 > \text{hypothesized value}$

$H_a: \mu_1 - \mu_2 < \text{hypothesized value}$

$H_a: \mu_1 - \mu_2 \neq \text{hypothesized value}$

P-value:

Area under appropriate t curve to the right of t

Area under appropriate t curve to the left of t

2(area to the right of t) if t is positive

or

2(area to the left of t) if t is negative

Assumptions: 1. The two samples are *independently selected random samples* from the populations of interest.

2. The *sample sizes are large* (generally 30 or larger)
or the population distributions are (at least approximately) normal.

Example 11.1 Facebook and Grades

In a study of the ways in which college students who use Facebook differ from college students who do not use Facebook, each person in a sample of 141 college students who use Facebook was asked to report his or her college grade point average (GPA). College GPA was also reported by each person in a sample of 68 students who do not use Facebook (["Facebook and Academic Performance," Computers in Human Behavior \[2010\]: 1237–1245](#)). One question that the researchers were hoping to answer is whether the mean college GPA for students who use Facebook is lower than the mean college GPA of students who do not use Facebook.

The two samples (141 students who were Facebook users and 68 students who were not Facebook users) were independently selected from students at a large, public

university. Although the samples were not selected at random, they were selected to be representative of the two populations (students who use Facebook and students who do not use Facebook at this university). Data from these samples were used to calculate sample means and standard deviations.

Population	Population Mean	Sample Size	Sample Mean	Sample Standard Deviation
Students at the university who use Facebook	μ_1 = mean college GPA for students who use Facebook	$n_1 = 141$	$\bar{x}_1 = 3.06$	$s_1 = 0.95$
Students at the university who do not use Facebook	μ_2 = mean college GPA for students who do not use Facebook	$n_2 = 68$	$\bar{x}_2 = 3.82$	$s_2 = 0.41$

Do these data provide convincing evidence that the mean GPA for students who use Facebook is lower than the mean GPA for students who do not use Facebook? We can answer this question by testing the relevant hypotheses using a 0.05 level of significance.

Understand the context ➤

1. μ_1 = mean GPA for students who use Facebook
2. μ_2 = mean GPA for students who do not use Facebook
3. $\mu_1 - \mu_2 =$ difference in mean GPA
4. $H_0: \mu_1 - \mu_2 = 0$ (no difference in mean GPA)
5. $H_a: \mu_1 - \mu_2 < 0$ (mean GPA is lower for Facebook users)
6. $\alpha = 0.05$

Formulate a plan ➤

$$5. \text{ Test statistic: } t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

6. Assumptions: From the study description, we know that the samples were independently selected. We also know that the samples were selected to be representative of the two populations of interest. Both samples sizes are large ($n_1 = 141 \geq 30$ and $n_2 = 68 \geq 30$) so it is reasonable to proceed with the two-sample t test.

Do the work ➤

$$7. \text{ Calculation: } t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(3.06 - 3.82) - 0}{\sqrt{\frac{(0.95)^2}{141} + \frac{(0.41)^2}{68}}} \\ = \frac{-0.76}{0.094} \\ = -8.08$$

8. P -value: We first calculate the df for the two-sample t test:

$$V_1 = \frac{s_1^2}{n_1} = 0.0064 \quad V_2 = \frac{s_2^2}{n_2} = 0.0025 \\ \text{df} = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} \\ = \frac{(0.0064 + 0.0025)^2}{\frac{(0.0064)^2}{140} + \frac{(0.0025)^2}{67}} \\ = \frac{0.0000792}{0.00000038} \\ = 208.421$$

We truncate the number of df to 208. This is a lower-tailed test (the inequality in H_a is $<$), so the P -value is the area under the t curve with $df = 208$ and to the left of -8.08 . Because the t curve is symmetric and centered at 0, this area is equal to the area under the t curve and to the right of $+8.08$. Because $t = 8.08$ is greater than the largest tabled value in the ∞ df row of Appendix Table 4, the area to the right is approximately 0. This means that

$$P\text{-value} \approx 0$$

Interpret the results ►

- 9.** Conclusion: Because the P -value is less than the selected significance level, we reject the null hypothesis. Based on the sample data, there is convincing evidence that the mean college GPA for students at the university who use Facebook is less than the mean college GPA for students at the university who do not use Facebook.

Based on this hypothesis test, we conclude that the sample mean GPA for students who use Facebook is enough less than the sample mean GPA for students that do not use Facebook that we don't think that this could have occurred just by chance due to sample-to-sample variability when there is no difference in the population means.

It is also possible to use statistical software or a graphing calculator to carry out the calculation step in a hypothesis test. For example, Minitab output for the test is shown here.

Two-Sample T-Test					
Sample	N	Mean	StDev	SE Mean	
1	141	3.060	0.950	0.080	
2	68	3.820	0.410	0.050	
Difference = mu (1) - mu (2)					
Estimate for difference: -0.760000					
T-Test of difference = 0 (vs <): T-Value = -8.07 P-Value = 0.000					
DF = 205					

From the Minitab output, we see that $t = -8.07$, the associated degrees of freedom is $df = 205$, and the P -value is 0.000. These values are slightly different from those we calculated by hand only because Minitab uses greater decimal accuracy in the calculations leading to the value of the test statistic and the degrees of freedom. The Minitab values can be used in the calculation and P -value steps, but just remember—these are only two of the steps in carrying out a hypothesis test. Even when you use technology, you still need to complete the other steps.

Example 11.2 | Rental Frogs?

- A frog jumping competition described in a short story written by Mark Twain inspired the real annual Calaveras County Frog Jumping Jubilee. In this competition, people enter bullfrogs into a contest to see which frog can jump the farthest. Some serious competitors have frogs that they have trained, and these contestants are known as “professional frogs.” Amateurs also compete with frogs that they can rent from the contest organizers, and these contestants are known as “rental frogs.” The authors of the paper [“Chasing Maximal Performance: A Cautionary Tale from the Celebrated Jumping Frogs of Calaveras County”](#) (*Journal of Experimental Biology* [2013]: 3947–3953) wanted to compare the performance of rental frogs and professional frogs. The authors of the paper used a two-sample t test to compare the mean jump distance (in meters) for the two groups.

Data consistent with summary quantities given in the paper are shown in the accompanying table (the actual sample sizes in the study were much larger).

Jump Distance (in meters)															
Rental Frog	0.9	0.9	1.3	0.8	1.1	1.2	0.8	0.9	1.4	1.1	1.1	1.0	0.9	1.2	1.2
Professional Frog	1.0	1.5	2.2	1.2	1.4	1.3	1.5	1.9	1.7	1.6	0.7	1.2	1.3	1.0	0.5

The authors of the paper hypothesized that rental frogs would not perform as well as professional frogs. Assuming that the frogs in the two samples are representative of rental frogs and professional frogs in general, the sample data can be used to determine if there is convincing evidence that the mean jumping distance for rental frogs is less than the mean distance for professional frogs.

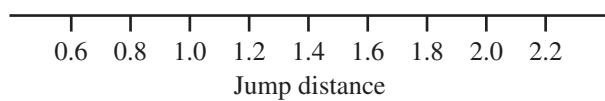
Understand the context ➤

1. μ_1 = population mean jump distance for rental frogs
- μ_2 = population mean jump distance for professional frogs
- $\mu_1 - \mu_2$ = difference in mean jump distance
2. $H_0: \mu_1 - \mu_2 = 0$
3. $H_a: \mu_1 - \mu_2 < 0$
4. Significance level: $\alpha = 0.05$

Formulate a plan ➤

5. Test statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - \text{hypothesized value}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

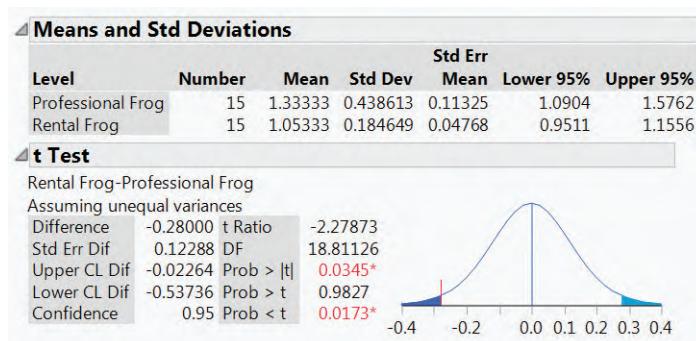
6. Assumptions: The samples were independently selected, and we will assume that the samples are representative of the populations of interest. Because both of the sample sizes are small, it is also necessary to assume that the jump distance distribution is approximately normal for each of the two populations. Boxplots constructed using the sample data are shown here:



Because the boxplots are reasonably symmetric and because there are no outliers, it is reasonable to assume that the two population distributions are approximately normal. With all of the assumptions met, it is appropriate to proceed with the two-sample t -test.

Do the work ➤

7. Calculation: JMP was used to do the calculations, resulting in the following output. From the JMP output, the value of the test statistic is -2.28 (rounded from -2.27873 in the output labeled t Ratio).



8. From the JMP output given in Step 7, $df = 18$ (rounded down from 18.81126 in the output labeled DF). The associated P -value is 0.0173 (from the Prob < t entry in the output because this is a lower tail test).

Interpret the results ➤

9. Conclusion: Because the P -value of 0.0173 is less than 0.05, we reject H_0 . The sample data provide convincing evidence that the mean jump distance for rental frogs is less than the mean jump distance for professional frogs.

Suppose the value of the test statistic in Step 7 had been -1.28 rather than -2.28 . Then the P -value would have been 0.105 (the area to the left of -1.3 under the t curve with 18 df) and the decision would have been to not reject the null hypothesis. We then would have concluded that there was not convincing evidence that the mean jump distance for rental frogs was less than the mean for professional frogs. Notice that when we fail to reject the null hypothesis of no difference between the population means, we are not saying that there is convincing evidence that the means are equal—we can only say that we were not convinced that they were different.

Comparing Treatments

When an experiment is carried out to compare two treatments (or to compare a single treatment with a control), we are interested in the effect of the treatments on some response variable. The treatments are “applied” to individuals (as in an experiment to compare two different medications for decreasing blood pressure) or objects (as in an experiment to compare two different baking temperatures on the density of bread). The value of some response variable (for example, blood pressure or density) is recorded. Based on the resulting data, we can determine whether there is a significant difference in the mean response for the two treatments.

In many actual experimental situations, the individuals or objects to which the treatments will be applied are not selected at random from some larger population. A consequence of this is that it is not possible to generalize the results of the experiment to some larger population. However, *if individuals or objects are randomly assigned to treatments (or treatments are randomly assigned to individuals or objects), it is possible to test hypotheses about treatment differences.*

It is common practice to use the two-sample t test statistic previously described if the experiment employs random assignment and if either the sample sizes are large or it is reasonable to think that the treatment response distributions (the distributions of response values that would result if the treatments were applied to a very large number of individuals or objects) are approximately normal.

Two-Sample t Test for Comparing Two Treatment Means

When

1. *individuals or objects are randomly assigned to treatments (or treatments are randomly assigned to individuals or objects), and*
2. *the sample sizes are large* (generally 30 or larger)

or the treatment response distributions are approximately normal,

the two-sample t test can be used to test

$$H_0: \mu_1 - \mu_2 = \text{hypothesized value}$$

where μ_1 and μ_2 represent the mean response for treatments 1 and 2, respectively.

In this case, these two conditions replace the assumptions previously stated for comparing two population means. Whether the assumption of normality of the treatment response distributions is reasonable can be assessed by constructing normal probability plots or boxplots of the response values in each sample.

When the two-sample t test is used to compare two treatments when the individuals or objects used in the experiment are not randomly selected from some population, it is only an approximate test (the reported P -values are only approximate). However, this is still the most common way to analyze such data.



Example 11.3 Fitness Trackers and Weight Loss

The article “**Activity Trackers May Undermine Weight Loss Efforts**” (*The New York Times*, September 20, 2016) describes a study published in the *Journal of the American Medical Association* (“The Effect of Wearable Technology Combined with a Lifestyle Intervention on Long-Term Weight Loss.” [2016]: 1161–1171). In this study, subjects followed a low-calorie diet and exercise program for 6 months. After 6 months, the subjects were randomly assigned to one of two groups. The people in one group were provided with a website they could use to self-monitor diet and physical activity. The people in the second group were provided with a wearable fitness tracker with an accompanying web interface to monitor diet and physical activity.

The researchers were interested in learning if the mean weight loss (in kilograms) at the end of 2 years was different for the two treatments (self-monitoring and fitness tracker monitoring). Data from this experiment are summarized in the accompanying table.

Group	Sample Size	Mean Weight Loss	Standard Deviation
Self-Monitoring	170	5.9 kg	6.8 kg
Fitness Tracker Monitoring	181	3.5 kg	6.3 kg

Do the data from this experiment provide evidence that the mean weight loss differs for the two treatments? We will test the relevant hypotheses using a significance level of 0.01.

Understand the context ➤

- Let μ_1 denote the mean weight loss for the self-monitor treatment and let μ_2 denote the mean weight loss for the fitness tracker treatment. Then $\mu_1 - \mu_2$ is the difference in treatment means.
- $H_0: \mu_1 - \mu_2 = 0$
- $H_a: \mu_1 - \mu_2 \neq 0$
- Significance level: $\alpha = 0.01$

Formulate a plan ➤

5. Test statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - \text{hypothesized value}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

- Assumptions: Subjects were randomly assigned to the treatment groups, and both sample sizes are large, so use of the two-sample t test is reasonable.

Do the work ➤

7. Calculation: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(5.9 - 3.5) - 0}{\sqrt{\frac{(6.8)^2}{170} + \frac{(6.8)^2}{181}}} = \frac{2.4}{0.701} = 3.42$

- P-value: We first calculate the df for the two-sample t test:

$$V_1 = \frac{s_1^2}{n_1} = \frac{(6.8)^2}{170} = 0.272 \quad V_2 = \frac{s_2^2}{n_2} = \frac{(6.8)^2}{181} = 0.219$$

$$\text{df} = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} = \frac{(0.272 + 0.219)^2}{\frac{(0.272)^2}{169} + \frac{(0.219)^2}{180}} = \frac{0.241}{0.0007} = 344.286$$

Truncating the calculated degrees of freedom results in $\text{df} = 344$.

This is a two-tailed test, so the P-value is two times the area under the t curve with $\text{df} = 344$ and to the right of 3.42. Since 3.42 is so far out in the upper tail of this t curve, $P\text{-value} \approx 0$.

Interpret the results ➤

- 9.** Conclusion: Because the P -value is less than the selected significance level (0.01), H_0 is rejected. There is convincing evidence that the mean weight loss is not the same for the self-monitoring treatment and the fitness tracker monitoring treatment. Notice that the mean weight loss for the group that used the fitness tracker was less than the mean for the group that self monitored. This is the basis for the headline from *The New York Times* article that said that fitness trackers may undermine weight loss efforts.

A statistical software package or a graphing calculator could also have been used to complete the Calculate step. For example, Minitab output is shown here. From the Minitab output, the value of the test statistic is $t = 3.42$. The reported degrees of freedom is 342 and the P -value is reported as 0.001. The small difference in the value of the degrees of freedom is due to differences in rounding.

Two-Sample T-Test and CI					
Sample	N	Mean	StDev	SE Mean	
1	170	5.90	6.80	0.52	
2	181	3.50	6.30	0.47	
Difference = $\mu(1) - \mu(2)$					
Estimate for difference: 2.400					
95% CI for difference: (1.021, 3.779)					
T-Test of difference = 0 (vs ≠): T-Value = 3.42 P-Value = 0.001 DF = 342					

You have probably noticed that evaluating the formula for number of degrees of freedom for the two-sample t test involves quite a bit of arithmetic. An alternative approach is to calculate a conservative estimate of the P -value—one that is close to but larger than the actual P -value. If H_0 is rejected using this conservative estimate, then it will also be rejected if the actual P -value is used.

A conservative estimate of the P -value for the two-sample t test can be found by using the t curve with the number of degrees of freedom equal to the smaller of $(n_1 - 1)$ and $(n_2 - 1)$.

The Pooled t Test

The two-sample t test just described is appropriate when it is reasonable to assume that the population distributions are approximately normal. If it is also known that the variances of the two populations are equal ($\sigma_1^2 = \sigma_2^2$), an alternative test known as the *pooled t test* can be used. This test combines information from both samples to obtain a “pooled” estimate of the common variance and then uses this pooled estimate of the variance in place of s_1^2 and s_2^2 in the t test statistic.

The pooled t test was widely used in the past, but it has fallen into some disfavor because it is quite sensitive to departures from the assumption of equal population variances. If the population variances are equal, the pooled t test has a slightly better chance of detecting departures from H_0 than does the two-sample t test of this section. However, P -values based on the pooled t test can be seriously in error if the population variances are not equal, so, in general, the two-sample t test is a better choice than the pooled t test.

Comparisons and Causation

If the assignment of treatments to the individuals or objects used in a comparison of treatments is not made by the investigators, the study is observational. As an example, the article **“Lead and Cadmium Absorption Among Children near a Nonferrous Metal Plant”** (*Environmental Research* [1978]: 290–308) reported data on blood lead concentrations for

two different samples of children. The first sample was drawn from a population residing within 1 km of a lead smelter. Those in the second sample were selected from a rural area much farther from the smelter.

It was the parents of the children, rather than the investigators, who determined whether the children would be in the close-to-smelter group or the far-from-smelter group. As a second example, a letter in the *Journal of the American Medical Association* (May 19, 1978) reported on a comparison of doctors' longevity after medical school graduation for those with an academic affiliation and those in private practice. (The letter writer's stated objective was to see whether "publish or perish" really meant "publish and perish.") Here again, an investigator did not start out with a group of doctors, assigning some to academic and others to nonacademic careers. The doctors themselves selected their groups.

The difficulty with drawing conclusions based on an observational study is that a statistically significant difference may be due to some underlying factors that have not been controlled rather than to conditions that define the groups. Does the type of medical practice itself have an effect on longevity, or is the observed difference in mean lifetime caused by other factors, which themselves led graduates to choose academic or nonacademic careers? Similarly, is the observed difference in blood lead concentration levels due to proximity to the smelter? Perhaps other physical and socioeconomic factors are related both to choice of living area and to blood lead concentration.

In general, rejection of $H_0: \mu_1 - \mu_2 = 0$ in favor of $H_a: \mu_1 - \mu_2 > 0$ suggests that, on average, higher values of the variable are *associated* with individuals in the first population or receiving the first treatment than with those in the second population or receiving the second treatment. But *association does not imply causation*.

Strong statistical evidence for a causal relationship can be built up over time through many different comparative studies that point to the same conclusions (as in the many investigations linking smoking to lung cancer). A randomized controlled experiment, in which investigators assign subjects at random to the treatments or conditions being compared, is particularly effective in suggesting causality. With random assignment, the investigator and other interested parties can be more easily convinced that an observed difference is caused by the difference in treatments or conditions.

A Confidence Interval

A confidence interval for $\mu_1 - \mu_2$ is easily obtained from the basic *t* variable developed in this section. Both the derivation of and the formula for the interval are similar to those of the one-sample *t* interval introduced in Chapter 9.

The Two-Sample *t* Confidence Interval for the Difference In Two Population or Treatment Means

The general formula for a confidence interval for $\mu_1 - \mu_2$ when

1. the two samples are *independently chosen random samples*
2. the *sample sizes are both large* (generally $n_1 \geq 30$ and $n_2 \geq 30$)
or
the population distributions are approximately normal

is

$$(\bar{x}_1 - \bar{x}_2) \pm (t \text{ critical value}) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The *t* critical value is based on

$$df = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} \quad \text{where} \quad V_1 = \frac{s_1^2}{n_1} \quad \text{and} \quad V_2 = \frac{s_2^2}{n_2}$$

(continued)

df should be truncated (rounded down) to an integer. The t critical values for the usual confidence levels are given in Appendix Table 3.

For a comparison of two treatments, when

1. *individuals or objects are randomly assigned to treatments (or vice versa)*
2. *the sample sizes are large* (generally 30 or larger)
or

the treatment response distributions are approximately normal,

the two-sample t confidence interval formula can be used to estimate $\mu_1 - \mu_2$.

EXAMPLE 11.4 Freshman Year Weight Gain

Understand the context ➤

- The paper “Predicting the ‘Freshman 15’: Environmental and Psychological Predictors of Weight Gain in First-Year University Students” (*Health Education Journal* [2010]: 321–332) described a study conducted by researchers at Carleton University in Canada. The researchers studied a random sample of first-year students who lived on campus and a random sample of first-year students who lived off campus. Data on weight gain (in kg) during the first year, consistent with summary quantities given in the paper, are given below. A negative weight gain represents a weight loss. The researchers believed that the mean weight gain of students living on campus was greater than the mean weight gain for students living off campus and were interested in estimating the difference in means for these two groups using a 95% confidence interval.

On Campus	Off Campus
2.0	1.6
2.3	3.1
1.1	-2.8
-2.0	0.0
-1.9	0.2
5.6	2.9
2.6	-0.9
1.1	3.8
5.6	0.7
8.2	-0.1

For these samples:

$$\text{On campus: } n_1 = 10 \quad \bar{x}_1 = 2.46 \quad s_1 = 3.26$$

$$\text{Off campus: } n_2 = 10 \quad \bar{x}_2 = 0.85 \quad s_2 = 2.03$$

Formulate a plan ➤

We want to estimate

$$\mu_1 - \mu_2 = \text{mean difference in weight gain}$$

where

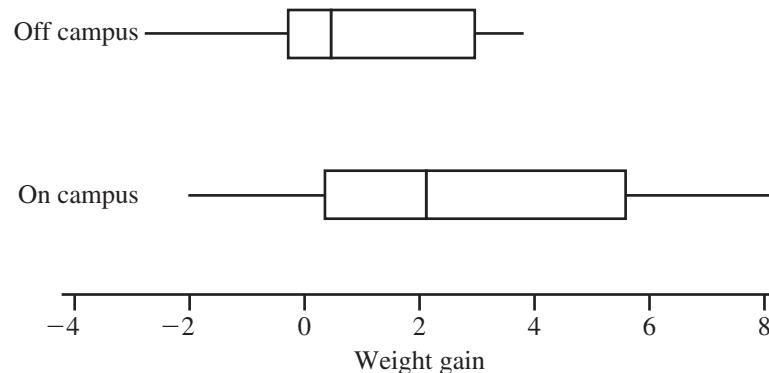
$$\mu_1 = \text{mean weight gain for first-year students living on campus}$$

and

$$\mu_2 = \text{mean weight gain for first-year students living off campus}$$

The samples were random samples from the two populations of interest (first-year students who live on campus and first-year students who live off campus), so the samples were independently selected. Because the sample sizes were not large, we need to be willing to assume that the population weight gain distributions are at least

approximately normal. Boxplots constructed using the data from the two samples are shown below. There are no outliers in either data set and the boxplots are reasonably symmetric, suggesting that the assumption of approximate normality is appropriate for each of the populations.



Do the work ➤ To estimate $\mu_1 - \mu_2$, the difference in mean weight gain for the two treatments, we will calculate a 95% confidence interval.

$$V_1 = \frac{s_1^2}{n_1} = \frac{(3.26)^2}{10} = 1.06 \quad V_2 = \frac{s_2^2}{n_2} = \frac{(2.03)^2}{10} = 0.41$$

$$df = \frac{\frac{(V_1 + V_2)^2}{V_1^2 + V_2^2}}{\frac{n_1 - 1}{n_1} + \frac{n_2 - 1}{n_2}} = \frac{\frac{(1.06 + 0.41)^2}{(1.06)^2 + (0.41)^2}}{9 + 9} = \frac{2.16}{0.14} = 15.43$$

Truncating to an integer gives $df = 15$. In the 15-df row of Appendix Table 3, the t critical value for a 95% confidence level is 2.13. The interval is then

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) &\pm (t \text{ critical value}) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\ &= (2.46 - 0.85) \pm (2.13) \sqrt{\frac{(3.26)^2}{10} + \frac{(2.03)^2}{10}} \\ &= 1.61 \pm (2.13)(1.214) \\ &= 1.61 \pm 2.586 \\ &= (-0.976, 4.196) \end{aligned}$$

Interpret the results ➤ This interval is rather wide, because the two sample sizes are small. Notice that the interval includes 0, so 0 is one of the plausible values for $\mu_1 - \mu_2$. Because 0 is included in this interval, it is possible that there is no real difference in the mean weight gain for students who live on campus and those who live off campus. The 95% confidence level means that we used a method to produce this estimate that correctly captures the actual value of $\mu_1 - \mu_2$ 95% of the time in repeated sampling.

Suppose that the confidence interval had been $(0.976, 4.196)$, so that the interval did not include 0. This interval would have been interpreted by saying that we are 95% confident that the mean weight gain for those who live on campus is greater than the mean weight gain for those who live off campus by somewhere between 0.976 and 4.196 kg.

Most statistical computer packages can compute the two-sample t confidence interval. Minitab was used to construct a 95% confidence interval using the data of this example; the resulting output is shown here:

Two-sample T for On Campus vs. Off Campus				
	N	Mean	StDev	SE Mean
On Campus	10	2.46	3.26	1.0
Off Campus	10	0.85	2.03	0.64
Difference = mu (On Campus) - mu (Off Campus)				
Estimate for difference: 1.61000				
95% CI for difference: (-0.97629, 4.19629)				

EXERCISES 11.1 - 11.21

● Data set available online

- 11.1** Consider two populations for which $\mu_1 = 30$, $\sigma_1 = 2$, $\mu_2 = 25$, and $\sigma_2 = 3$. Suppose that two independent random samples of sizes $n_1 = 40$ and $n_2 = 50$ are selected. Describe the approximate sampling distribution of $\bar{x}_1 - \bar{x}_2$ (center, variability, and shape).
- 11.2** An individual can take either a scenic route to work or a nonscenic route. She decides that taking the nonscenic route can be justified only if it reduces the mean travel time by more than 10 minutes.
- If μ_1 is the mean travel time for the scenic route and μ_2 is the mean travel time for the nonscenic route, what hypotheses should be tested?
 - If μ_1 is the mean travel time for the nonscenic route and μ_2 is the mean travel time for the scenic route, what hypotheses should be tested?
- 11.3** Many people now turn to the Internet to get information on health-related topics. The paper “[An Examination of Health, Medical and Nutritional Information on the Internet: A Comparative Study of Wikipedia, WebMD and the Mayo Clinic Websites](#)” (*The International Journal of Communication and Health* [2015]: 30–38) used Flesch reading ease scores (a measure of reading difficulty based on factors such as sentence length and number of syllables in the words used) to score pages on Wikipedia and on WebMD. Higher Flesch scores correspond to more difficult reading levels. The paper reported that for a representative sample of health-related pages on Wikipedia, the mean Flesch score was 26.7 and

the standard deviation of the Flesch scores was 14.1. For a representative sample of pages from WebMD, the mean score was 43.9 and the standard deviation was 19.4. Suppose that these means and standard deviations were based on samples of 40 pages from each site. Is there convincing evidence that the mean reading level for health-related pages differs for Wikipedia and WebMD? Test the relevant hypotheses using a significance level of 0.05.

- 11.4** Use the information in the previous exercise to answer the following questions.
- Construct a 90% confidence interval estimate of the difference in mean Flesch reading ease score for health-related pages on Wikipedia and health-related pages on WebMD.
 - What does this confidence interval imply about the readability of health-related information from these two sources? Is this consistent with the conclusion in the hypothesis test of the previous exercise?
- 11.5** ● The article “[Plugged In, but Tuned Out](#)” (*USA TODAY*, January 20, 2010) summarizes data from two surveys of kids age 8 to 18. One survey was conducted in 1999 and the other was conducted in 2009. Data on number of hours per day spent using electronic media that are consistent with summary quantities given in the article are given below (the actual sample sizes for the two surveys were much larger). For this exercise, assume that it is reasonable to regard the two samples as representative of kids age 8 to 18 in each of the 2 years that the surveys were conducted.

2009	5	9	5	8	7	6	7	9	7	9	6	9	10	9	8
1999	4	5	7	7	5	7	5	6	5	6	7	8	5	6	6

- a. Because the given sample sizes are small, in order for the two-sample t test to be appropriate, what assumption must be made about the distributions of electronic media use times? Use the given data to construct graphical displays that would be useful in determining whether this assumption is reasonable. Do you think it is reasonable to use these data to carry out a two-sample t test?
- b. Do the given data provide convincing evidence that the mean number of hours per day spent using electronic media was greater in 2009 than in 1999? Test the relevant hypotheses using a significance level of 0.01. (Hint: See Example 11.2.)
- c. Construct and interpret a 98% confidence interval estimate of the difference in the means for the number of hours per day spent using electronic media in 2009 and 1999. (Hint: See Example 11.4.)
- 11.6** The National Sleep Foundation surveyed representative samples of adults in six different countries to ask questions about sleeping habits ([“2013 International Bedroom Poll Summary of Findings,” sleepfoundation.org/sites/default/files/RPT495a.pdf](#)). Each person in a representative sample of 250 adults in each of these countries was asked how much sleep they get on a typical work night. For the United States, the sample mean was 391 minutes and for Mexico the sample mean was 426 minutes. Suppose that the sample standard deviations were 30 minutes for the U.S. sample and 40 minutes for the Mexico sample. The report concludes that on average, adults in the United States get less sleep on work nights than adults in Mexico. Is this a reasonable conclusion? Support your answer with an appropriate hypothesis test.
- 11.7** The report referenced in the previous exercise also gave data for representative samples of 250 adults in Canada and in England. The sample mean amount of sleep on a work night was 423 minutes for Canada and 409 minutes for England. Suppose that the sample standard deviations were 35 minutes for the Canada sample and 42 minutes for the England sample.
- a. Construct and interpret a 95% confidence interval estimate of the difference in the mean amount of sleep on a work night for adults in Canada and adults in England.
- b. Based on the confidence interval from Part (a), would you conclude that there is evidence of a difference in the mean amount of sleep on a work night for the two countries? Explain why or why not.

- 11.8** The paper “[If It’s Hard to Read, It’s Hard to Do” \(*Psychological Science* \[2008\]: 986–988\)](#) described an interesting study of how people perceive the effort required to do certain tasks. Each of 20 students was randomly assigned to one of two groups. One group was given instructions for an exercise routine that were printed in an easy-to-read font (Arial). The other group received the same set of instructions, but printed in a font that is considered difficult to read (*Brush*).

After reading the instructions, subjects estimated the time (in minutes) they thought it would take to complete the exercise routine. Summary statistics are given below.

	Easy font	Difficult font
n	10	10
\bar{x}	8.23	15.10
s	5.61	9.28

The authors of the paper used these data to carry out a two-sample t test, and concluded that at the 0.10 significance level, there was convincing evidence that the mean estimated time to complete the exercise routine was less when the instructions were printed in an easy-to-read font than when printed in a difficult-to-read font. Discuss the appropriateness of using a two-sample t test in this situation.

- 11.9** The paper “[Facebook Use and Academic Performance Among College Students: A Mixed-Methods Study with a Multi-Ethnic Sample” \(*Computers in Human Behavior* \[2015\]: 265–272\)](#) describes a survey of a sample of 66 male students and a sample of 195 female students at a large university in Southern California. The authors of the paper believed that these samples were representative of male and female college students in Southern California. For the sample of males, the mean time spent per day on Facebook was 102.31 minutes. For the sample of females, the mean time was 159.61 minutes. The sample standard deviations were not given in the paper, but for purposes of this exercise, suppose that the sample standard deviations were both 100 minutes.

- a. Do the data provide convincing evidence that the mean time spent on Facebook is not the same for males and for females? Test the relevant hypotheses using $\alpha = 0.05$.
- b. Do you think it is reasonable to generalize the conclusion from the hypothesis test in Part (a) to the populations of all male college students in the United States and all female college students in the United States? Explain why you think this.

- 11.10** Use the information in the previous exercise to answer the following questions.
- Construct a 95% confidence interval estimate of the difference in mean time spent on Facebook for male college students and female college students in Southern California.
 - What does this confidence interval imply about the mean time spent on Facebook for these two populations of students? Is this consistent with the conclusion in the hypothesis test of the previous exercise?
- 11.11** Do male college students spend more time than female college students using a computer? This was one of the questions investigated by the authors of the paper “**An Ecological Momentary Assessment of the Physical Activity and Sedentary Behaviour Patterns of University Students**” (*Health Education Journal* [2010]: 116–125). Each student in a random sample of 46 male students at a university in England and each student in a random sample of 38 female students from the same university kept a diary of how he or she spent time over a three-week period.
- For the sample of males, the mean time spent using a computer per day was 45.8 minutes and the standard deviation was 63.3 minutes. For the sample of females, the mean time spent using a computer was 39.4 minutes and the standard deviation was 57.3 minutes. Is there convincing evidence that the mean time male students at this university spend using a computer is greater than the mean time for female students? Test the appropriate hypotheses using $\alpha = 0.05$
- 11.12** The paper “**Mood Food: Chocolate and Depressive Symptoms in a Cross-Sectional Analysis**” (*Archives of Internal Medicine* [2010]: 699–703) describes a study that investigated the relationship between depression and chocolate consumption. Participants in the study were 931 adults who were not currently taking medication for depression. These participants were screened for depression using a widely used screening test. The participants were then divided into two samples based on the score on the screening test. One sample consisted of people who screened positive for depression, and the other sample consisted of people who did not screen positive for depression. Each of the study participants also completed a food frequency survey.
- The researchers believed that the two samples were representative of the two populations of interest—adults who would screen positive for depression and adults who would not screen positive. The paper reported that the mean number of servings of chocolate for the sample of people that screened positive for depression was 8.39 servings per month and the sample standard deviation was 14.83. For the sample of people who did not screen positive for depression, the mean number of servings per month was 5.39 and the standard deviation was 8.76. The paper did not say how many individuals were in each sample, but for the purposes of this exercise, you can assume that the 931 study participants were divided into 311 who screened positive for depression and 620 who did not screen positive.
- Estimate the difference in the mean number of servings of chocolate per month in the population of people who would screen positive for depression and the mean number of chocolate servings per month in the population of people who would not screen positive for depression. Use a confidence level of 90% and be sure to interpret the interval in context.
- 11.13** The authors of the paper “**Influence of Biofeedback Weight Bearing Training in Sit to Stand to Sit and the Limits of Stability on Stroke Patients**” (*The Journal of Physical Therapy Science* [2016]: 3011–3014) randomly selected two samples of patients admitted to the hospital after suffering a stroke. One sample was selected from patients who received biofeedback weight training for 8 weeks, and the other sample was selected from patients who did not receive this training. At the end of 8 weeks, the time it took (in seconds) to stand from a sitting position and then to sit down again (called sit-stand-sit time) was measured for the people in each sample. Data consistent with summary quantities given in the paper are given below. For purposes of this exercise, you can assume that the samples are representative of the population of stroke patients who receive the biofeedback training and the population of stroke patients who do not receive this training. Use the given data to construct and interpret a 95% confidence interval for the difference in mean sit-stand-sit time for these two populations.
- | | |
|---|---|
| Biofeedback Group | No Biofeedback Group |
| 1.9 2.6 4.3 2.1 2.7 4.1 3.2 4.0 3.2 3.5 2.8 3.5 3.5 2.3 3.1 | 5.1 4.7 3.9 4.2 4.7 4.3 4.2 5.1 3.4 4.2 5.1 4.4 4.0 3.4 3.9 |
- 11.14** Example 11.1 looked at a study comparing students who use Facebook and students who do not use Facebook (“**Facebook and Academic Performance, Computers in Human Behavior** [2010]: 1237–1245). In addition to asking the students in the samples about GPA, each student was also asked how many hours he or she spent studying each day. The two samples (141 students who were Facebook users and 68 students who were not Facebook users) were independently selected from students at a large, public university. Although the samples were not

selected at random, they were selected to be representative of the two populations.

For the sample of Facebook users, the mean number of hours studied per day was 1.47 hours and the standard deviation was 0.83 hours. For the sample of students who do not use Facebook, the mean was 2.76 hours and the standard deviation was 0.99 hours. Do these sample data provide convincing evidence that the mean time spent studying for Facebook users at this university is less than the mean time spent studying by students at the university who do not use Facebook? Use a significance level of 0.01.

- 11.15** Internet addiction has been described as excessive and uncontrolled Internet use. The authors of the paper “**Gender Difference in the Relationship Between Internet Addiction and Depression**” (*Computers in Human Behavior* [2016]: 463–470) used a score designed to measure the extent and severity of Internet addiction in a study of 836 male and 879 female sixth-grade students in China. Internet addiction was measured using Young’s Internet Addiction Diagnostic Test. The lowest possible score on this test is zero, and higher scores indicate higher levels of Internet addiction. For the sample of males, the mean Internet Addiction score was 1.51 and the standard deviation was 2.03. For the sample of females, the mean was 1.07 and the standard deviation was 1.63. For purposes of this exercise, you can assume that it is reasonable to regard these two samples as representative of the population of male Chinese sixth-grade students and the population of female Chinese sixth-grade students, respectively.
- The standard deviation is greater than the mean for each of these samples. Explain why it is not reasonable to think that the distribution of Internet Addiction scores would be approximately normal for either the population of male Chinese sixth-grade students or the population of female Chinese sixth-grade students.
 - Given your response to Part (a), would it be appropriate to use the two-sample t test to test the null hypothesis that there is no difference in the mean Internet Addiction score for male

Chinese sixth-grade students and female Chinese sixth-grade students? Explain why or why not.

- If appropriate, carry out a test to determine if there is convincing evidence that the mean Internet Addiction score is greater for male Chinese sixth-grade students than for female Chinese sixth-grade students. Use $\alpha = 0.05$.

- 11.16** The paper “**Does the Color of the Mug Influence the Taste of the Coffee?**” (*Flavour* [2014]: 1–7) describes an experiment in which subjects were assigned at random to one of two treatment groups. The 12 people in one group were served coffee in a white mug and were asked to rate the quality of the coffee on a scale from 0 to 100. The 12 people in the second group were served the same coffee in a clear glass mug, and they also rated the coffee. The mean quality rating for the 12 people in the white mug group was 50.35 and the standard deviation was 20.17. The mean quality rating for the 12 people in the clear glass mug group was 61.48 and the standard deviation was 16.69. For this exercise, you may assume that the distribution of quality ratings for each of the two treatments is approximately normal.

- Use the given information to construct and interpret a 95% confidence interval for the difference in mean quality rating for this coffee when served in a white mug and when served in a glass mug.
- Based on the interval from Part (a), are you convinced that the color of the mug makes a difference in terms of mean quality rating? Explain.

- 11.17** A newspaper story headline reads “**Gender Plays Part in Monkeys’ Toy Choices, Research Finds—Like Humans, Male Monkeys Choose Balls and Cars, While Females Prefer Dolls and Pots**” (*Knight Ridder Newspapers*, December 8, 2005). The article goes on to summarize findings published in the paper “**Sex Differences in Response to Children’s Toys in Nonhuman Primates**” (*Evolution and Human Behavior* [2002]: 467–479). Forty-four male monkeys and 44 female monkeys were each given a variety of toys, and the time spent playing with each toy was recorded. The table below gives means and standard deviations (approximate values

Table for Exercise 11.17

	Toy	Percent of Time					
		Female Monkeys			Male Monkeys		
		<i>n</i>	Sample Mean	Sample Standard Deviation	<i>n</i>	Sample Mean	Sample Standard Deviation
	Police Car	44	8	4	44	18	5
	Doll	44	20	4	44	9	2
	Furry Dog	44	20	5	44	25	5

read from graphs in the paper) for the percentage of time that a monkey spent playing with a particular toy. Assume that it is reasonable to regard these two samples of 44 monkeys as representative of the populations of male monkeys and female monkeys. Use a 0.05 significance level for any hypothesis tests that you carry out when answering the various parts of this exercise.

- a. The police car was considered a “masculine toy.” Do these data provide convincing evidence that the mean percentage of the time spent playing with the police car is greater for male monkeys than for female monkeys?
- b. The doll was considered a “feminine toy.” Do these data provide convincing evidence that the mean percentage of time spent playing with the doll is greater for female monkeys than for male monkeys?
- c. The furry dog was considered a “neutral toy.” Do these data provide convincing evidence that the mean percentage of time spent playing with the furry dog is not the same for male and female monkeys?
- d. Based on the conclusions from the hypothesis tests of Parts (a)–(c), is the quoted newspaper story headline a reasonable summary of the findings? Explain.
- e. Explain why it would be inappropriate to use the two-sample t test to decide if there was evidence that the mean percentage of time spent playing with the police car and the mean percentage of the time spent playing with the doll is not the same for female monkeys.

- 11.18** The authors of the paper “**The Empowering (Super) Heroine? The Effects of Sexualized Female Characters in Superhero Films on Women**” (*Sex Roles* [2015]: 211–220) were interested in the effect of watching movies that portrayed female heroines in roles that focus on their sex appeal on female viewers. They carried out an experiment in which female college students were assigned at random to one of two experimental groups. The 23 women in one group watched 13 minutes of scenes from the *X-Men* film series and then responded to a questionnaire designed to measure body esteem. Lower scores on this measure correspond to lower body satisfaction. The 29 women in the other group (the control group) did not watch any video prior to responding to the questionnaire measuring body esteem. For the women who watched the *X-Men* video, the mean body esteem score was 4.43 and the standard deviation was 1.02. For the women in the control group, the mean body esteem score was 5.08 and the standard deviation was 0.98. You may assume that the distribution of body esteem scores for each of

the two treatments (video and control) is approximately normal.

- a. Construct and interpret a 90% confidence interval for the difference in mean body esteem score for women who watch the video and those who do not.
- b. Do you think that watching the video has an effect on body esteem? Explain.

- 11.19** Fumonisins are environmental toxins produced by a type of mold and have been found in corn and in products made from raw corn. The **Center for Food Safety and Applied Nutrition** provided recommendations on allowable fumonisin levels in human food and in animal feed based on a study of corn meal. The study compared corn meal made from partially degerned corn (corn that has had the germ, the part of the kernel located at the bottom center of the kernel that is used to produce corn oil, partially removed) and corn meal made from corn that has not been degerned. Specimens of corn meal were analyzed and the total fumonisin level (ppm) was determined for each specimen. Summary statistics for total fumonisin level from the U.S. Food and Drug Administration’s web site are given here.

	\bar{x}	s
Partially Degermed	0.59	1.01
Not Degermed	1.21	1.71

- a. If the given means and standard deviations had been based on a random sample of 10 partially degerned specimens and a random sample of 10 specimens made from corn that was not degerned, explain why it would not be appropriate to carry out a two-sample t test to determine if there is a significant difference in the mean fumonisin level for the two types of corn meal.
- b. Suppose instead that each of the random samples had included 50 corn meal specimens. Explain why it would now be reasonable to carry out a two-sample t test.
- c. Assuming that each random sample size was 50, carry out a test to determine if there is a significant difference in mean fumonisin level for the two types of corn meal. Use a significance level of 0.01.

- 11.20** A researcher at the Medical College of Virginia conducted a study of 60 randomly selected male soccer players and concluded that frequently “heading” the ball in soccer lowers players’ IQs (*USA TODAY*, August 14, 1995). The soccer players were divided into two groups, based on whether they averaged 10 or more headers per game.

Mean IQs were reported in the article, but the sample sizes and standard deviations were not

given. Suppose that these values were as given in the accompanying table.

	n	Sample Mean	Sample SD
Fewer Than 10 Headers	35	112	10
10 or More Headers	25	103	8

- a. Do these data support the researcher's conclusion? Test the relevant hypotheses using $\alpha = 0.05$.
- b. Can you conclude that heading the ball *causes* lower IQ?
- 11.21** Do certain behaviors result in a severe drain on energy resources because a great deal of energy is expended in comparison to energy intake? The article "**The Energetic Cost of Courtship and Aggression in a Plethodontid Salamander**" (*Ecology* [1983]: 979–983) reported on one of the few studies concerned with behavior and energy expenditure. The accompanying table gives oxygen consumption (mL/g/hr) for male-female salamander pairs.

Behavior	Sample Size	Sample Mean	Sample SD
Noncourting	11	0.072	0.0066
Courting	15	0.099	0.0071

- a. The pooled t test is a test procedure for testing $H_0: \mu_1 - \mu_2 = \text{hypothesized value}$ when it is reasonable to assume that the two population distributions are normal with equal standard deviations ($\sigma_1 = \sigma_2$). The test statistic for the pooled t test is obtained by replacing both s_1 and s_2 in the two-sample t test statistic with s_p where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

When the population distributions are normal with equal standard deviations and H_0 is true, the resulting pooled t statistic has a t distribution with $df = n_1 + n_2 - 2$. For the reported data, the two sample standard deviations are similar. Use the pooled t test with $\alpha = 0.05$ to determine whether the mean oxygen consumption for courting pairs is higher than the mean oxygen consumption for noncourting pairs.

- b. Would the conclusion in Part (a) have been different if the two-sample t test had been used rather than the pooled t test?

SECTION 11.2

Inferences Concerning the Difference Between Two Population or Treatment Means Using Paired Samples

Two samples are said to be **independent** if the selection of the individuals or objects that make up one of the samples has no bearing on the selection of individuals or objects in the other sample. In some situations, a study with independent samples is not the best way to obtain information about a possible difference between two populations. For example, suppose that an investigator wants to determine whether regular aerobic exercise affects blood pressure. A random sample of people who jog regularly and a second random sample of people who do not exercise regularly are selected independently of one another.

The researcher might use the two-sample t test to conclude that a significant difference exists between the mean blood pressures for joggers and nonjoggers. But is it reasonable to think that the difference in mean blood pressure is the result of jogging? It is known that blood pressure is related to both diet and body weight. Maybe the joggers in the sample tend to be leaner and adhere to a healthier diet than the nonjoggers. This might account for the observed difference. On the basis of this study, the researcher would not be able to rule out the possibility that the observed difference in blood pressure is explained by weight differences between the people in the two samples and that aerobic exercise itself has no effect.

One way to avoid this difficulty is to match subjects by weight. The researcher could find pairs of subjects where the jogger and nonjogger in each pair were similar in weight (although weights for different pairs might vary widely). The factor *weight* could then be ruled out as a possible explanation for an observed difference in mean blood pressure between the two groups. Matching the subjects by weight results in two samples for which

each observation in the first sample is paired in a meaningful way with a particular observation in the second sample. Such samples are said to be **paired**.

Paired samples occur in a number of different ways. Some studies involve using the same group of individuals with measurements recorded both before and after some intervening treatment. Other studies might use naturally occurring pairs, such as twins or husbands and wives. Others construct pairs by matching on factors with effects that might otherwise make differences (or the lack of them) between the two populations difficult to detect (such as weight in the jogging example).

Example 11.5 illustrates why it is important to consider whether two samples are independent or paired.

Example 11.5 Document Word Clouds

Understand the context ➤



- When people are searching for information online, most people quickly scan a document and attempt to determine if it is relevant before taking the time to read the whole document. The authors of the paper **“Document Word Clouds: Visualising Web Documents as Tag Clouds to Aid Users in Relevance Decisions”** (*Research and Advanced Technology for Digital Libraries* [2009]: 94–105) wondered if people would have an easier time determining if a document was relevant if they first saw a word cloud representation of the document. (A word cloud is a visual representation of a text passage that uses characteristics such as font size, color, and color density to indicate the importance of a particular word.)

The researchers chose 10 documents and created a word cloud representation for each of them. The text document version was shown to one group of people and the word cloud representation was shown to another group of people. The average time (in seconds) it took the people in the group to make a relevance decision was recorded for each version of the 10 documents is shown in the accompanying table.

Document	Time to Relevance Decision	Time to Relevance Decision
	Text Version	Word Cloud Version
1	3.55	2.95
2	2.94	3.04
3	3.72	2.72
4	2.63	1.97
5	2.39	2.17
6	5.36	3.64
7	1.86	1.96
8	2.49	2.04
9	2.76	2.46
10	2.77	2.63

You can view these data as consisting of two samples—a sample of times for the relevance decision for text documents and a sample of times for the relevance decision for word cloud representations.

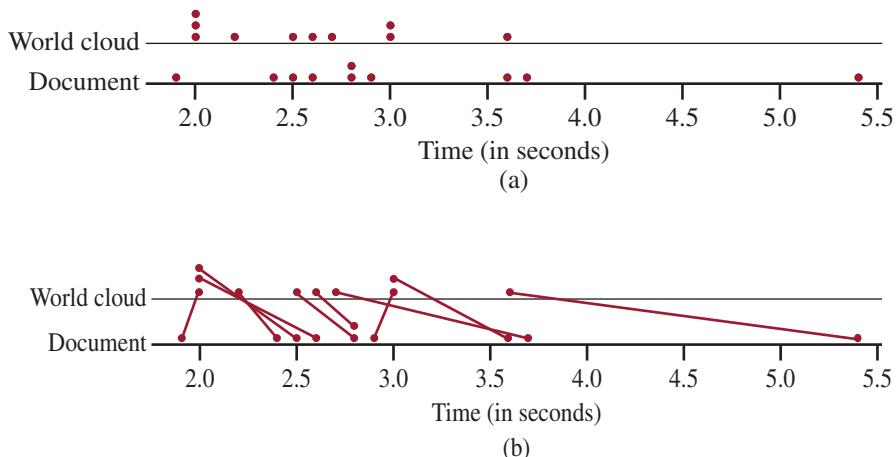
In this case, the samples are paired rather than independent because each time in the text document sample is paired with a particular time in the word cloud sample for the same document. Notice that in 8 of the 10 data pairs, the time to relevance decision is greater for the text version of the document than for the word cloud representation of the same document. This suggests that the two population means may be different.

Both the text document times and the word cloud version times vary from document to document. It is this variability that may make any difference difficult to see if the paired nature of the samples is ignored. To see how this might be the case, consider the two plots in Figure 11.1. The first plot (Figure 11.1 (a)) ignores the pairing, and the two samples look quite similar. However, the plot in which pairs are identified (Figure 11.1 (b)) does suggest that there is a difference because for eight of the ten pairs the text document time to relevance decision is greater than the time for the word cloud representation (these are the pairs linked by line segments that slant to the right). And in the two cases where the

- Data set available online

FIGURE 11.1

Two plots of the paired data from Example 11.5:
 (a) pairing ignored;
 (b) pairs identified.



time for the text document is less than the time for the word cloud, the line segments are close to vertical, indicating that the two times for those documents were not very different.

Example 11.5 suggests that when samples are paired, methods that ignore the pairing are not an appropriate way to analyze the data. This is why it is important to determine if the samples are paired or independent before carrying out a hypothesis test or finding a confidence interval estimate.

Example 11.5 suggests that the methods of inference developed for independent samples are not adequate for dealing with paired samples. When sample observations from the first population are paired in some meaningful way with sample observations from the second population, inferences should be based on the differences between the two observations within each sample pair. The n sample differences can then be regarded as having been selected from a large population of differences.

For example, in Example 11.5, we can think of the 10 (text – word cloud) differences as having been selected from an entire population of differences.

Consider the following notation:

$$\mu_d = \text{mean value of the difference population}$$

and

$$\sigma_d = \text{standard deviation of the difference population}$$

The relationship between the two individual population means and the mean difference is

$$\mu_d = \mu_1 - \mu_2$$

This means that when the samples are paired, inferences about $\mu_1 - \mu_2$ are equivalent to inferences about μ_d . Since inferences about μ_d can be based on the n observed sample differences, the original two-sample problem becomes a familiar one-sample problem.

Paired t Test

To compare two population or treatment means when the samples are paired, we first translate the hypotheses of interest from ones about the value of $\mu_1 - \mu_2$ to equivalent hypotheses involving μ_d :

Hypothesis

- $H_0: \mu_1 - \mu_2 = \text{hypothesized value}$
- $H_a: \mu_1 - \mu_2 > \text{hypothesized value}$
- $H_a: \mu_1 - \mu_2 < \text{hypothesized value}$
- $H_a: \mu_1 - \mu_2 \neq \text{hypothesized value}$

Equivalent Hypothesis

When Samples Are Paired

- $H_0: \mu_d = \text{hypothesized value}$
- $H_a: \mu_d > \text{hypothesized value}$
- $H_a: \mu_d < \text{hypothesized value}$
- $H_a: \mu_d \neq \text{hypothesized value}$

Sample differences (Sample 1 value – Sample 2 value) are then calculated and used as the basis for testing hypotheses about μ_d . When the number of differences is large or when it is reasonable to assume that the population of differences is approximately normal, the one-sample t test based on the differences is the recommended test procedure. In general, if each of the two individual populations is normal, the population of differences is also normal. A normal probability plot or boxplot of the differences can be used to decide if the assumption of normality is reasonable.

Summary of the Paired t Test for Comparing Two Population or Treatment Means

Null hypothesis: $H_0: \mu_d = \text{hypothesized value}$

Test statistic: $t = \frac{\bar{x}_d - \text{hypothesized value}}{\frac{s_d}{\sqrt{n}}}$

where n is the number of sample differences and \bar{x}_d and s_d are the mean and standard deviation of the sample differences. This test is based on $df = n - 1$.

Alternative hypothesis:

$H_a: \mu_d > \text{hypothesized value}$

$H_a: \mu_d < \text{hypothesized value}$

$H_a: \mu_d \neq \text{hypothesized value}$

P-value:

Area under the appropriate t curve to the right of t

Area under the appropriate t curve to the left of t

2(area to the right of t) if t is positive

or

2(area to the left of t) if t is negative

Assumptions:

1. The samples are *paired*.
2. The n sample differences can be viewed as a *random sample* from a population of differences.
3. The *number of sample differences is large* (generally at least 30) or the *population distribution of differences is approximately normal*.

Example 11.6 Word Clouds Revisited

We can use the time to relevance decision data of Example 11.5 to test the claim that the mean time is greater for text documents than for word cloud representations. Because the samples are paired, the first thing to do is calculate the sample differences. These are the text – word cloud time differences for the 10 documents in the sample.

The sample data and the calculated differences are shown in the accompanying table. A positive difference indicates that the time to make a relevance decision was greater for the text document than for the word cloud representation.

Document	Time (in seconds) to Relevance Decision Text Version	Time (in seconds) to Relevance Decision Word Cloud Version	Difference (Text – Word Cloud)
1	3.55	2.95	0.60
2	2.94	3.04	-0.10
3	3.72	2.72	1.00
4	2.63	1.97	0.66
5	2.39	2.17	0.22
6	5.36	3.64	1.72
7	1.86	1.96	-0.10
8	2.49	2.04	0.45
9	2.76	2.46	0.30
10	2.77	2.63	0.14

For this example,

μ_1 = mean time to make a relevance decision for text documents

μ_2 = mean time to make a relevance decision for word cloud representations

and

$\mu_d = \mu_1 - \mu_2$ = mean difference in time (text – word cloud)

The question of interest can be answered by testing the hypothesis

$$H_0: \mu_d = 0 \quad \text{versus} \quad H_a: \mu_d > 0$$

The mean and standard deviation of the sample differences are $\bar{x}_d = 0.489$ and $s_d = 0.552$. Do these data provide evidence that the mean time to make a relevance decision is greater for text documents than for word cloud representations of documents?

We can use the paired *t* test with a significance level of 0.05 to carry out the hypothesis test.

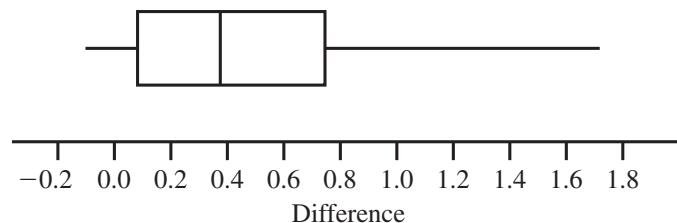
Understand the context ➤

1. μ_d = mean difference in time (text – word cloud)
2. $H_0: \mu_d = 0$
3. $H_a: \mu_d > 0$
4. Significance level: $\alpha = 0.05$

Formulate a plan ➤

$$5. \text{ Test statistic: } t = \frac{\bar{x}_d - \text{hypothesized value}}{\frac{s_d}{\sqrt{n}}}$$

6. Assumptions: Although the sample of 10 documents was not a random sample, the researchers selected the documents to be representative of the population of documents. For this reason, it is reasonable to view the 10 sample differences as a random sample of all such differences. A boxplot of the differences is approximately symmetric and does not show any outliers, so the assumption of normality is not unreasonable and we will proceed with the paired *t* test.



Do the work ➤

$$7. \text{ Calculation: } t = \frac{0.489 - 0}{\frac{0.552}{\sqrt{10}}} = 2.79$$

8. *P*-value: This is an upper-tailed test, so the *P*-value is the area to the right of the calculated *t* value. The appropriate df for this test is $df = 10 - 1 = 9$. From the 9-df column of Appendix Table 4, we find that *P*-value = area to the right of $2.79 \approx 0.010$.

Interpret the results ➤

9. Conclusion: Because the *P*-value (0.010) is less than α (0.05), we reject H_0 . There is convincing evidence that the mean time to make a relevance decision is greater for text documents than the mean time for word cloud representations.

Statistical software could also be used to calculate the value of the test statistic, degrees of freedom, and the P -value. Minitab output is shown here:

Paired T-Test				
Paired T for Text - Word Cloud				
	N	Mean	StDev	SE Mean
Text	10	3.047	0.975	0.308
Word Cloud	10	2.558	0.550	0.174
Difference	10	0.489	0.552	0.175
T-Test of mean difference = 0 (vs > 0): T-Value = 2.80 P-Value = 0.010				

Using the two-sample t test (for independent samples) for the data in Example 11.6 would have been incorrect, because the samples are not independent. Inappropriate use of the two-sample t test would have resulted in a calculated test statistic value of 1.38. The conclusion would have been to not reject the hypothesis of equal mean times. This illustrates why it is important to recognize when samples are paired.

Example 11.7 Charitable Chimps

- The authors of the paper “Chimpanzees Are Indifferent to the Welfare of Unrelated Group Members” (*Nature* [2005]: 1357–1359) concluded that “chimpanzees do not take advantage of opportunities to deliver benefits to individuals at no cost to themselves.” This conclusion was based on data from an experiment in which a sample of chimpanzees was trained to use an apparatus that would deliver food just to the subject chimpanzee when one lever was pushed and would deliver food to both the subject chimpanzee and another chimpanzee in an adjoining cage when another lever was pushed. After training, the chimps were observed when there was no chimp in the adjoining cage and when there was another chimp in the adjoining cage.

The researchers hypothesized that if chimpanzees were motivated by the welfare of others, they would choose the option that provided food to both chimpanzees more often when there was a chimpanzee in the adjoining cage. Data on the number of times the “feed both” option was chosen out of 36 opportunities (approximate values read from a graph in the paper) are given in the accompanying table.

Chimp	Number of Times “Feed Both” Option Was Chosen	
	No Chimp in Adjoining Cage	Chimp in Adjoining Cage
1	21	23
2	22	22
3	23	21
4	21	23
5	18	19
6	16	19
7	19	19

Most statistical software packages will perform a paired t test, and we will use Minitab to carry out a test to determine if there is convincing evidence that the mean number of times the “feed both” option is selected is higher when another chimpanzee is present in the adjoining cage than when the subject chimpanzee is alone.

Understand the context ➤

- μ_d = difference between mean number of “feed both” selections for chimpanzees who are alone and for chimpanzees who have company in the adjoining cage
- $H_0: \mu_d = 0$
- $H_a: \mu_d < 0$

• Data set available online

4. Significance level: $\alpha = 0.05$

Formulate a plan ➤

5. Test statistic: $t = \frac{\bar{x}_d - \text{hypothesized value}}{\frac{s_d}{\sqrt{n}}}$

6. Assumptions: Although the chimpanzees in this study were not randomly selected, the authors considered them to be representative of the population of chimpanzees. A boxplot of the differences is approximately symmetric and does not show any outliers, so the assumption of approximate normality is reasonable and we will proceed with the paired t test.

Do the work ➤

7. Calculation: From the given Minitab output, $t = -1.35$.

Paired T-Test and CI: Alone, Companion

Paired T for Alone - Companion

	N	Mean	StDev	SE Mean
Alone	7	20.0000	2.4495	0.9258
Companion	7	20.8571	1.8645	0.7047
Difference	7	-0.857143	1.676163	0.633530
95% CI for mean difference: (-2.407335, 0.693050)				
T-Test of mean difference = 0 (vs not = 0): T-Value = -1.35				
P-Value = 0.225				

8. P -value: From the Minitab output, P -value = 0.225.

Interpret the results ➤

9. Conclusion: The P -value (0.225) is greater than α (0.05), so the null hypothesis is not rejected. The data do not provide evidence that the mean number of times that the “feed both” option is chosen is greater when there is a chimpanzee in the adjoining cage. This is the basis for the statement quoted at the beginning of this example.

Notice that the numerator of the paired t test statistic \bar{x}_d , and the numerator of the two-sample t test statistic, $\bar{x}_1 - \bar{x}_2$, are always equal. The difference is in the denominator. The variability in differences is usually smaller than the variability in each sample separately (because measurements in a pair tend to be similar). As a result, the value of the paired t test statistic is usually greater in magnitude than the value of the two-sample t test statistic. Pairing usually reduces variability, making differences easier to detect.

A Confidence Interval

The one-sample t confidence interval for μ given in Chapter 9 is easily adapted to obtain an interval estimate for μ_d .

Paired t Confidence Interval for μ_d

When

1. the samples are *paired*.
2. the n sample differences can be viewed as a *random sample* from a population of differences.
3. the *number of sample differences is large* (generally at least 30) or the *population distribution of differences is approximately normal*,

the paired t confidence interval for μ_d is

$$\bar{x}_d \pm (t \text{ critical value}) \cdot \frac{s_d}{\sqrt{n}}$$

For a specified confidence level, the $(n - 1)$ df row of Appendix Table 3 gives the appropriate t critical values.

Example 11.8 Word Clouds One Last Time

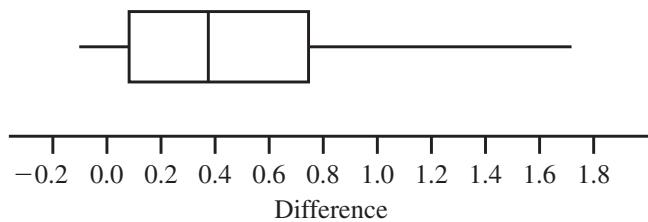
Example 11.5 provided data from a study to assess whether the time it takes to make a decision about the relevance of an online document is greater when the document is in text form than when it is represented by a word cloud. The conclusion in the hypothesis test in Example 11.6 was that there is convincing evidence that the mean time to make a relevance decision is greater for text documents than for word cloud representations. Once you have reached this conclusion, it may also be of interest to estimate how much greater the mean time is for text documents than for word clouds.

The sample data and calculated differences from Example 11.6 are shown again in the accompanying table.

Document	Time (in seconds) to Relevance Decision Text Version	Time (in seconds) to Relevance Decision Word Cloud Version	Difference (Text – Word Cloud)
1	3.55	2.95	0.60
2	2.94	3.04	-0.10
3	3.72	2.72	1.00
4	2.63	1.97	0.66
5	2.39	2.17	0.22
6	5.36	3.64	1.72
7	1.86	1.96	-0.10
8	2.49	2.04	0.45
9	2.76	2.46	0.30
10	2.77	2.63	0.14

Formulate a plan ➤

We will use these data to estimate the mean difference in the time it takes to make a relevance decision using a 95% confidence interval. The accompanying boxplot of the 10 sample differences is not inconsistent with a difference population that is approximately normal, so the paired t confidence interval is appropriate.



Do the work ➤

The mean and standard deviation calculated using the 10 sample differences are $\bar{x}_d = 0.489$ and $s_d = 0.552$. The t critical value for $df = 9$ and a 95% confidence level is 2.26. Substituting these values into the formula for the paired t confidence interval gives

$$\begin{aligned} &= 0.489 \pm (2.26) \frac{0.552}{\sqrt{10}} \\ &= 0.489 \pm 0.395 \\ &= (0.094, 0.884) \end{aligned}$$

interpret the results ➤

Based on the sample data, we can be 95% confident that the actual difference in mean time to make a relevance decision is somewhere between 0.094 seconds and 0.884 seconds. Because both endpoints of the interval are positive, this means we would estimate that the mean time to make a relevance decision is greater for text documents than for word cloud representations by somewhere between 0.094 seconds and 0.884 seconds. The method used to construct this interval is successful in capturing the actual difference in population means approximately 95% of the time.

A statistical software package or a graphing calculator could also have been used to determine the endpoints of the confidence interval. Minitab output is shown below.

Paired T-Test and CI				
Paired T for Text - Word Cloud				
	N	Mean	StDev	SE Mean
Text	10	3.047	0.975	0.308
Word Cloud	10	2.558	0.550	0.174
Difference	10	0.489	0.552	0.175
95% CI for mean difference: (0.094, 0.884)				

When two populations must be compared to draw a conclusion on the basis of sample data, a researcher might choose to use independent samples or paired samples. In many situations, paired data allow for a better comparison by screening out the effects of extraneous variables that might make differences between the two populations difficult to detect or that might suggest a difference when none exists.

EXERCISES 11.22 - 11.40

● Data set available online

11.22 Suppose that you were interested in investigating the effect of a drug that is to be used in the treatment of patients who have glaucoma in both eyes. A comparison between the mean reduction in eye pressure for this drug and for a standard treatment is desired. Both treatments are applied directly to the eye.

- a. Describe how you would go about collecting data for your investigation.
- b. Does your method result in paired data?
- c. Can you think of a reasonable method of collecting data that would not result in paired samples? Would such an experiment be as informative as a paired experiment? Comment.

11.23 ● Many runners believe that listening to music while running enhances their performance. The authors of the paper “Effects of Synchronous Music on Treadmill Running Among Elite Triathletes” (*Journal of Science and Medicine in Sport* [2012]: 52–57) wondered if this was true for experienced runners. They recorded time to exhaustion for 11 triathletes while running on a treadmill at a speed determined to be near their peak running velocity. The time to exhaustion was recorded for each participant on two different days. On one day, each participant ran while listening to music that the runner selected as motivational. On a different day, each participant ran with no music playing.

You can assume that it is reasonable to regard these 11 triathletes as representative of the population of experienced triathletes. Only summary quantities were given in the paper, but the data in

the table below are consistent with the means and standard deviations given in the paper. Do the data provide convincing evidence that the mean time to exhaustion for experienced triathletes is greater when they run while listening to motivational music? Test the relevant hypotheses using a significance level of 0.05.

Runner	Time to Exhaustion (seconds)										
	1	2	3	4	5	6	7	8	9	10	11
Motivational Music	535	533	527	524	431	498	555	396	539	542	523
No Music	467	446	482	573	562	592	473	496	552	500	524

11.24 The study described in the previous exercise also measured time to exhaustion for the 11 triathletes on a day when they listened to music that the runners had classified as neutral as compared to motivational. The researchers calculated the difference between the time to exhaustion while running to motivational music and while running to neutral music. The mean difference in (motivational – neutral) was –7 seconds (the sample mean time to exhaustion was actually lower when listening to music the runner viewed as motivational than when listening to music the runner viewed as neutral). Suppose that the standard deviation of the differences was $s_d = 80$. You can assume that it is reasonable to regard these 11 triathletes as representative of the population of experienced triathletes and that the population difference distribution is approximately normal. Is there convincing evidence that the

mean time to exhaustion for experienced triathletes running to motivational music differs from the mean time to exhaustion when running to neutral music? Carry out a hypothesis test using $\alpha = 0.05$.

- 11.25** In Exercise 11.23, a hypothesis test leads to the conclusion that there is not convincing evidence that the mean time to exhaustion for experienced triathletes is greater when they run while listening to motivational music. Use the information given in that exercise to construct and interpret a 95% confidence interval for the difference in mean time to exhaustion for experienced triathletes when running to motivational music and the mean time when running with no music.

- 11.26** The article “**Puppy Love? It’s Real, Study Says**” (*USA TODAY*, April 17, 2015) describes a study into how people communicate with their pets. The conclusion expressed in the title of the article was based on research published in *Science* (“**Oxytocin-Gaze Positive Loop and the Coevolution of Human-Dog Bonds**,” April 17, 2015). Researchers measured the oxytocin levels (in picograms per milligram, pg/mg) of 22 dog owners before and again after a 30-minute interaction with their dogs. (Oxytocin is a hormone known to play a role in parent-child bonding.) The difference in oxytocin level (before – after) was calculated for each of the 22 dog owners. Suppose that the mean and standard deviation of the differences (approximate values based on a graph in the paper) were $\bar{x}_d = 27$ pg/mg and $s_d = 30$ pg/mg.

- a. Explain why the two samples (oxytocin levels before interaction and oxytocin levels after interaction) are paired.
- b. Assume that it is reasonable to regard the 22 dog owners who participated in this study as representative of dog owners in general. Do the data from this study provide convincing evidence that there is an increase in mean oxytocin level of dog owners after 30 minutes of interaction with their dogs? State and test the appropriate hypotheses using a significance level of 0.05.

- 11.27** The authors of the paper “**Concordance of Self-Report and Measured Height and Weight of College Students**” (*Journal of Nutrition, Education and Behavior* [2015]: 94–98) used a paired-samples t test to reach the conclusion that male college students tend to over-report both weight and height. This conclusion was based on a sample of 634 male college students selected from eight different universities. The sample mean difference between the reported weight and actual measured weight was 1.2 pounds and the standard deviation of the differences was 5.71 pounds. You can assume that the sample was representative of male college students.

- a. Carry out a hypothesis test to determine if there is a significance difference in the mean reported weight and the mean actual weight for male college students.
- b. For height, the mean difference between the reported height and actual measured height was 0.6 inches and the standard deviation of the differences was 0.8 inches. Carry out a hypothesis test to determine if there is a significance difference in the mean reported height and the mean actual height for male college students.
- c. Do the conclusions reached in the hypothesis tests of Parts (a) and (b) support the given conclusion that male college students tend to over-report both height and weight? Explain.

- 11.28** The paper referenced in the previous exercise also compared the reported heights and weights to actual measured heights and weights for a sample of 1052 female college students selected from eight different universities. The resulting data are summarized in the accompanying table. You can assume that this sample is representative of female college students.

	Sample Mean Difference (Reported – Actual)	Sample Standard Deviation of Differences
Weight	−0.6	4.8
Height	−0.2	0.8

- a. Carry out a hypothesis test to determine if there is a significance difference in the mean reported weight and the mean actual weight for female college students.
- b. Carry out a hypothesis test to determine if there is a significance difference in the mean reported height and the mean actual height for female college students.
- c. Do the conclusions reached in the hypothesis tests of Parts (a) and (b) support the conclusion that female college students tend to under-report both height and weight? Explain.

- 11.29** The paper “**Driving Performance While Using a Mobile Phone: A Simulation Study of Greek Professional Drivers**” (*Transportation Research Part F* [2016]: 164–170) describes a study in which 50 Greek male taxi drivers drove in a driving simulator. In the simulator, they were asked to drive following a lead car. On one drive, they had no distractions and the average distance between the driver’s car and the lead car was recorded. In a second drive, the drivers talked on a mobile phone while driving. The authors of the paper used a paired-sampled t test to determine if the mean following distance is greater when the driver has no distractions than when the driver is talking on a mobile phone. The mean of

the 50 sample differences (no distraction – talking on mobile phone) was 0.47 meters and the standard deviation of the sample differences was 1.22 meters. The authors concluded that there was evidence to support the claim that the mean following distance for Greek taxi drivers is greater when there are no distractions than when the driver is talking on a mobile phone. Do you agree with this conclusion? Carry out a hypothesis test to support your answer. You may assume that this sample of 50 drivers is representative of Greek taxi drivers.

- 11.30** The paper referenced in the previous exercise also had the 50 taxi drivers drive in the simulator while sending and receiving text messages. The mean of the 50 sample differences (no distraction – reading text messages) was 1.3 meters and the standard deviation of the sample differences was 1.54 meters. The authors concluded that there was evidence to support the claim that the mean following distance for Greek taxi drivers is greater when there are no distractions than when the driver is texting. Do you agree with this conclusion? Carry out a hypothesis test to support your answer. You can assume that this sample of 50 drivers is representative of Greek taxi drivers.

- 11.31** Use the information given in the previous exercise to construct and interpret a 95% confidence interval for the difference in mean following distance for Greek taxi drivers while driving with no distractions and while driving and texting.

- 11.32** To determine if chocolate milk was as effective as other carbohydrate replacement drinks, nine male cyclists performed an intense workout followed by a drink and a rest period. At the end of the rest period, each cyclist performed an endurance trial in which he exercised until exhausted and time to exhaustion was measured. Each cyclist completed the entire regimen on two different days. On one day the drink provided was chocolate milk and on the other day the drink provided was a carbohydrate replacement drink.

Data consistent with summary quantities appearing in the paper “[The Efficacy of Chocolate Milk as a Recovery Aid](#)” (*Medicine and Science in Sports and Exercise* [2004]: S126) appear in the table below. Is there evidence that the mean time to exhaustion is greater after chocolate milk than

after carbohydrate replacement drink? Use a significance level of 0.05. (Hint: See Examples 11.6 and 11.7.)

- 11.33** The humorous paper “[Will Humans Swim Faster or Slower in Syrup?](#)” (*American Institute of Chemical Engineers Journal* [2004]: 2646–2647) investigates the fluid mechanics of swimming. Twenty swimmers each swam a specified distance in a water-filled pool and in a pool in which the water was thickened with food grade guar gum to create a syrup-like consistency. Velocity, in meters per second, was recorded. Values estimated from a graph that appeared in the paper are given.

The authors of the paper concluded that swimming in guar syrup does not change mean swimming speed. Are the given data consistent with this conclusion? Carry out a hypothesis test using a 0.01 significance level.

Swimmer	Velocity (m/s)	
	Water	Guar Syrup
1	0.90	0.92
2	0.92	0.96
3	1.00	0.95
4	1.10	1.13
5	1.20	1.22
6	1.25	1.20
7	1.25	1.26
8	1.30	1.30
9	1.35	1.34
10	1.40	1.41
11	1.40	1.44
12	1.50	1.52
13	1.65	1.58
14	1.70	1.70
15	1.75	1.80
16	1.80	1.76
17	1.80	1.84
18	1.85	1.89
19	1.90	1.88
20	1.95	1.95

- 11.34** The study described in the paper “[Marketing Actions Can Modulate Neural Representation of Experienced Pleasantness](#)” (*Proceedings of the National Academy of Science* [2008]: 1050–1054) investigated whether price affects people’s judgment.

Table for Exercise 11.32

Cyclist	Time to Exhaustion (minutes)								
	1	2	3	4	5	6	7	8	9
Chocolate Milk	24.85	50.09	38.30	26.11	36.54	26.14	36.13	47.35	35.08
Carbohydrate Replacement	10.02	29.96	37.40	15.52	9.11	21.58	31.23	22.04	17.02

Twenty people each tasted six cabernet sauvignon wines and rated how they liked them on a scale of 1 to 6. Prior to tasting each wine, participants were told the price of the wine. Of the six wines tasted, two were actually the same wine, but for one tasting the participant was told that the wine cost \$10 per bottle and for the other tasting the participant was told that the wine cost \$90 per bottle. The participants were randomly assigned either to taste the \$90 wine first and the \$10 wine second, or the \$10 wine first and the \$90 wine second.

Differences were calculated by subtracting the rating for the tasting in which the participant thought the wine cost \$10 from the rating for the tasting in which the participant thought the wine cost \$90. The differences that follow are consistent with summary quantities given in the paper.

Difference (\$90 – \$10)

2 4 1 2 1 0 0 3 0 2 1 3 3 1 4 1 2 2 1 -1

Carry out a hypothesis test to determine if the mean rating assigned to the wine when the cost is described as \$90 is greater than the mean rating assigned to the wine when the cost is described as \$10. Use $\alpha = 0.01$.

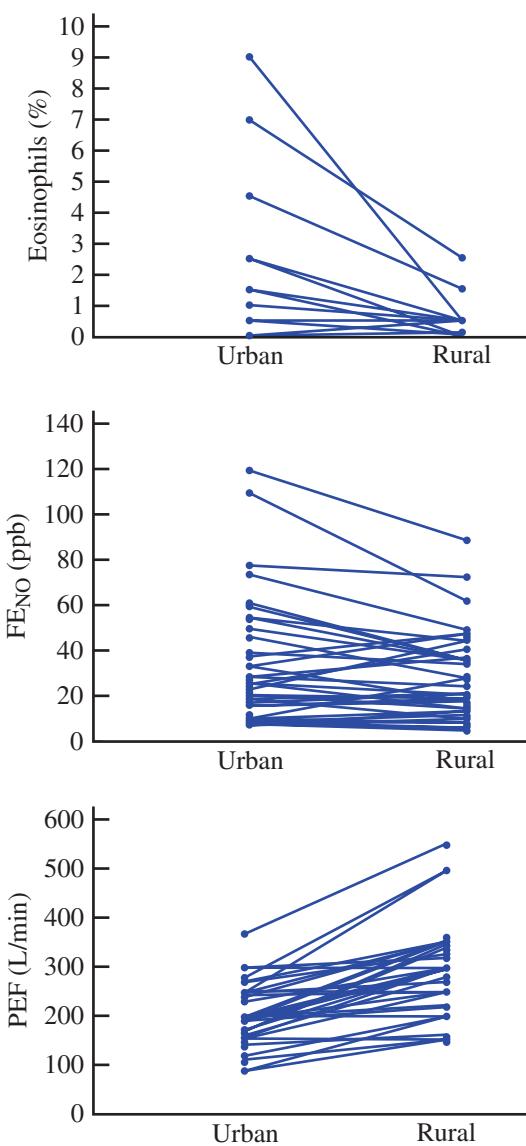
- 11.35** Head movement evaluations are important because disabled individuals may be able to operate communications aids using head motion. The paper “**Constancy of Head Turning Recorded in Healthy Young Humans**” (*Journal of Biomedical Engineering* [2008]: 428–436) reported the accompanying data on neck rotation (in degrees) for 14 subjects both in the clockwise direction (CL) and in the counterclockwise direction (CO). For purposes of this exercise, you can assume that the 14 subjects are representative of the population of adult Americans. Based on these data, is it reasonable to conclude that mean neck rotation is greater in the clockwise direction than in the counterclockwise direction? Carry out a hypothesis test using a significance level of 0.01. (Hint: See Examples 11.6 and 11.7.)

Subject:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
CL:	57.9	35.7	54.5	56.8	51.1	70.8	77.3	51.6	54.7	63.6	59.2	59.2	55.8	38.5
CO:	44.2	52.1	60.2	52.7	47.2	65.6	71.4	48.8	53.1	66.3	59.8	47.5	64.5	34.5

- 11.36** The paper “**Less Air Pollution Leads to Rapid Reduction of Airway Inflammation and Improved Airway Function in Asthmatic Children**” (*Pediatrics* [2009]: 1051–1058) describes a study in which children with mild asthma who live in a polluted urban environment were relocated to a less polluted rural environment for 7 days. Various measures of respiratory function were recorded first in the urban environment and then again after 7 days in the rural environment. The accompanying graphs show the urban and rural

values for three of these measures: nasal eosinophils, exhaled FE_{NO} concentration, and peak expiratory flow (PEF). Urban and rural values for the same child are connected by a line.

The authors of the paper used paired *t* tests to determine that there was a significant difference in the urban and rural means for each of these three measures. One of these tests resulted in a *P*-value less than 0.001, one resulted in a *P*-value between 0.001 and 0.01, and one resulted in a *P*-value between 0.01 and 0.05.



- Which measure (Eosinophils, FE_{NO} , or PEF) do you think resulted in a test with the *P*-value that was less than 0.001? Explain your reasoning.
- Which measure (Eosinophils, FE_{NO} , or PEF) do you think resulted in the test with the largest *P*-value? Explain your reasoning.

- 11.37** The paper “**The Truth About Lying in Online Dating Profiles**” (*Proceedings, Computer-Human Interactions*

[2007]: 1–4) describes an investigation in which 40 men and 40 women with online dating profiles agreed to participate in a study. Each participant's height (in inches) was measured and the actual height was compared to the height given in that person's online profile. The differences between the online profile height and the actual height (profile – actual) were used to calculate the values in the accompanying table.

Men	Women
$\bar{x}_d = 0.57$	$\bar{x}_d = 0.03$
$s_d = 0.81$	$s_d = 0.75$
$n = 40$	$n = 40$

You can assume it is reasonable to regard the two samples in this study as being representative of male online daters and female online daters. (Although the authors of the paper believed that their samples were representative of these populations, participants were volunteers recruited through newspaper advertisements, so we should be a bit hesitant to generalize results to all online daters.)

- a. Use the paired t test to determine if there is convincing evidence that, on average, male online daters overstate their height in online dating profiles. Use $\alpha = 0.05$.
- b. Construct and interpret a 95% confidence interval for the difference between the mean online dating profile height and mean actual height for female online daters. (Hint: See Example 11.8.)

Data for Exercise 11.38

Player	Position Player Dominant Arm	Position Player Nondominant Arm	Pitcher	Pitcher Dominant Arm	Pitcher Nondominant Arm
1	30.31	32.54	1	27.63	24.33
2	44.86	40.95	2	30.57	26.36
3	22.09	23.48	3	32.62	30.62
4	31.26	31.11	4	39.79	33.74
5	28.07	28.75	5	28.50	29.84
6	31.93	29.32	6	26.70	26.71
7	34.68	34.79	7	30.34	26.45
8	29.10	28.87	8	28.69	21.49
9	25.51	27.59	9	31.19	20.82
10	22.49	21.01	10	36.00	21.75
11	28.74	30.31	11	31.58	28.32
12	27.89	27.92	12	32.55	27.22
13	28.48	27.85	13	29.56	28.86
14	25.60	24.95	14	28.64	28.58
15	20.21	21.59	15	28.58	27.15
16	33.77	32.48	16	31.99	29.46
17	32.59	32.48	17	27.16	21.26
18	32.60	31.61			
19	29.30	27.46			

- c. Use the two-sample t test of Section 11.1 to test $H_0: \mu_m - \mu_f = 0$ versus $H_a: \mu_m - \mu_f > 0$, where μ_m is the mean height difference (profile – actual) for male online daters and μ_f is the mean height difference (profile – actual) for female online daters.
- d. Explain why a paired t test was used in Part (a) but a two-sample t test was used in Part (c).

- 11.38** • The paper “Quantitative Assessment of Glenohumeral Translation in Baseball Players” (*American Journal of Sports Medicine* [2004]: 1711–1715) considered various aspects of shoulder motion for a sample of pitchers and another sample of position players. The authors kindly supplied the data shown below on anteroposterior translation (mm), a measure of the extent of anterior and posterior motion, both for the dominant arm and the nondominant arm.
- a. Estimate the true average difference in translation between dominant and nondominant arms for pitchers using a 95% confidence interval.
 - b. Estimate the true average difference in translation between dominant and nondominant arms for position players using a 95% confidence interval.
 - c. The authors asserted that pitchers have greater difference in mean anteroposterior translation of their shoulders than do position players. Do you agree? Explain.
- 11.39** Two proposed computer mouse designs were compared by recording wrist extension in degrees for 24 people who each used both mouse types

(“Comparative Study of Two Computer Mouse Designs,” Cornell Human Factors Laboratory Technical Report RP7992). The difference in wrist extension was calculated by subtracting extension for mouse type B from the wrist extension for mouse type A for each student. The mean difference was reported to be 8.82 degrees. Assume that it is reasonable to regard this sample of 24 people as representative of the population of computer users.

- a. Suppose that the standard deviation of the differences was 10 degrees. Is there convincing evidence that the mean wrist extension for mouse type A is greater than for mouse type B? Use a 0.05 significance level.
 - b. Suppose that the standard deviation of the differences was 26 degrees. Is there convincing evidence that the mean wrist extension for mouse type A is greater than for mouse type B? Use a 0.05 significance level.
 - c. Briefly explain why different conclusions were reached in the hypothesis tests of Parts (a) and (b).
- 11.40** ● The authors of the paper “[Ultrasound Techniques Applied to Body Fat Measurement in Male and Female Athletes](#)” (*Journal of Athletic Training* [2009]: 142–147) compared two different methods for measuring body fat percentage. One method uses ultrasound, and the other method uses X-ray

technology. Body fat percentages using each of these methods for 16 athletes (a subset of the data given in a graph that appeared in the paper) are given in the accompanying table. You can assume that the 16 athletes who participated in this study are representative of the population of athletes. Use these data to estimate the difference in mean body fat percentage measurement for the two methods. Use a confidence level of 95% and interpret the interval in context.

Athlete	X-ray	Ultrasound
1	5.00	4.75
2	7.00	3.75
3	9.25	9.00
4	12.00	11.75
5	17.25	17.00
6	29.50	27.50
7	5.50	6.50
8	6.00	6.75
9	8.00	8.75
10	8.50	9.50
11	9.25	9.50
12	11.00	12.00
13	12.00	12.25
14	14.00	15.50
15	17.00	18.00
16	18.00	18.25

SECTION 11.3 Large-Sample Inferences Concerning the Difference Between Two Population or Treatment Proportions

Large-sample methods for estimating and testing hypotheses about a single population proportion were presented in Chapters 9 and 10. The symbol p was used to represent the proportion of individuals in the population who possess some characteristic (the successes). Inferences about the value of p were based on \hat{p} , the corresponding sample proportion of successes.

Many investigations are carried out to compare the proportion of successes in two populations or for two treatments. As was the case for means, we use the subscripts 1 and 2 to distinguish between the two population proportions, sample sizes, and sample proportions.

Notation

Population or Treatment 1: Proportion of “successes” = p_1

Population or Treatment 2: Proportion of “successes” = p_2

Sample Size	Proportion of Successes
Sample from Population or Treatment 1	n_1
Sample from Population or Treatment 2	n_2

When comparing two populations or treatments on the basis of “success” proportions, we use $p_1 - p_2$, the difference between the two proportions. Because \hat{p}_1 provides an estimate of p_1 and \hat{p}_2 provides an estimate of p_2 , the obvious choice for an estimate of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$.

Because \hat{p}_1 and \hat{p}_2 each vary in value from sample to sample, so does the difference $\hat{p}_1 - \hat{p}_2$. For example, a first sample from each of two populations might yield

$$\hat{p}_1 = 0.69 \quad \hat{p}_2 = 0.70 \quad \hat{p}_1 - \hat{p}_2 = 0.01$$

A second sample from each might result in

$$\hat{p}_1 = 0.79 \quad \hat{p}_2 = 0.67 \quad \hat{p}_1 - \hat{p}_2 = 0.12$$

and so on. Because the statistic $\hat{p}_1 - \hat{p}_2$ is the basis for drawing inferences about $p_1 - p_2$, we need to know something about its behavior.

Properties of the Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

For independently selected random samples, the following properties hold:

1. $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$

This means that the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is centered at $p_1 - p_2$, so $\hat{p}_1 - \hat{p}_2$ is an unbiased statistic for estimating $p_1 - p_2$.

2. $\sigma_{\hat{p}_1 - \hat{p}_2}^2 = \sigma_{\hat{p}_1}^2 + \sigma_{\hat{p}_2}^2 = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}$

and

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

3. If both n_1 and n_2 are large (if $n_1 p_1 \geq 10$, $n_1(1 - p_1) \geq 10$, $n_2 p_2 \geq 10$, and $n_2(1 - p_2) \geq 10$), then \hat{p}_1 and \hat{p}_2 each have a sampling distribution that is approximately normal. The difference $\hat{p}_1 - \hat{p}_2$ also has a sampling distribution that is approximately normal.

The properties in the box imply that when the samples are independently selected and when both sample sizes are large, the statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}}$$

has a distribution that is approximately standard normal.

A Large-Sample Test Procedure

Comparisons of p_1 and p_2 are often based on large, independently selected samples, and we restrict ourselves to this case. The most general null hypothesis of interest has the form

$$H_0: p_1 - p_2 = \text{hypothesized value}$$

However, when the hypothesized value is something other than 0, the appropriate test statistic differs somewhat from the test statistic used for $H_0: p_1 - p_2 = 0$. Because $H_0: p_1 - p_2 = 0$ is almost always the relevant null hypothesis in applied problems, we only consider this case. The test procedures we have seen so far control the probability of a Type I error at the desired level α . This requires using a test statistic with a sampling distribution that is known when H_0 is true. That is, the test statistic should be developed under the assumption that $p_1 = p_2$ (as specified by the null hypothesis $p_1 - p_2 = 0$). If the two

population or treatment proportions are equal, they have a common value. This common value is denoted by p . The z statistic obtained by standardizing $\hat{p}_1 - \hat{p}_2$ then simplifies to

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}}$$

Unfortunately, this z statistic cannot be used as a test statistic, because to calculate the denominator we need to know the value of p . H_0 says that there is a common value p , but it does not specify what that value is. Fortunately, a test statistic can be obtained by first estimating p from the sample data and then using this estimate in the denominator of z .

When $p_1 = p_2$, both \hat{p}_1 and \hat{p}_2 are estimates of the common proportion p . However, a better estimate than either \hat{p}_1 or \hat{p}_2 is a weighted average of the two, with more weight given to the sample proportion based on the larger sample.

DEFINITION

Combined estimate of a common population proportion:

$$\hat{p}_c = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{\text{total number of } S\text{'s in the two samples}}{\text{total of the two sample sizes}}$$

The test statistic for testing $H_0: p_1 - p_2 = 0$ results from using \hat{p}_c , the combined estimate, in place of p in the standardized z statistic given previously. This z statistic has approximately a standard normal distribution when H_0 is true, so a P -value can be calculated using the z table.

Summary of Large-Sample z Tests for $p_1 - p_2 = 0$

Null hypothesis: $H_0: p_1 - p_2 = 0$

Test statistic:
$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2}}}$$

Alternative hypothesis: **P-value:**

$H_a: p_1 - p_2 > 0$

Area under the z curve to the right of z

$H_a: p_1 - p_2 < 0$

Area under the z curve to the left of z

$H_a: p_1 - p_2 \neq 0$

2(area to the right of z) if z is positive
or

2(area to the left of z) if z is negative

Assumptions: 1. The samples are *independently chosen random samples*, or
treatments were assigned at random.

2. Both sample sizes are large:

$$n_1 \hat{p}_1 \geq 10 \quad n_1(1 - \hat{p}_1) \geq 10 \quad n_2 \hat{p}_2 \geq 10 \quad n_2(1 - \hat{p}_2) \geq 10$$

Example 11.9 Electric Cars

The article “[Americans Say No to Electric Cars Despite Gas Prices](#)” (*USA TODAY*, May 25, 2011) describes a survey of public opinion on issues related to rising gas prices. The survey was conducted by Gallup, a national polling organization. Each person in a representative sample of low-income adult Americans (annual income less than \$30,000) and each person

in an independently selected representative sample of high-income adult Americans (annual income greater than \$75,000) was asked whether he or she would consider buying an electric car if gas prices continue to rise.

In the low-income sample, 65% said that they would not buy an electric car no matter how high gas prices were to rise. In the high-income sample, 59% responded this way. The article did not give the sample sizes, but for purposes of this example, we will use sample sizes of 300.

One question of interest is whether the proportion who would never consider buying an electric car is different for the two income groups. The following table summarizes what we know so far.

Population	Population Proportion	Sample Size	Sample Proportion
Low-income adult Americans	p_1 = proportion of all low-income adult Americans who would not buy an electric car	$n_1 = 300$	$\hat{p}_1 = 0.65$
High-income adult Americans	p_2 = proportion of all high-income adult Americans who would not buy an electric car	$n_2 = 300$	$\hat{p}_2 = 0.59$

Notice that the two sample proportions are not equal. But even if the two population proportions were equal, we wouldn't expect the two sample proportions to be equal because of sampling variability—the differences that occur from one sample to another just by chance. The important question is whether chance is a believable explanation for the observed difference in the two sample proportions, or whether this difference is large enough that we don't think it would have occurred just by chance. A hypothesis test will help us to make this determination.

Understand the context ➤

1. p_1 = proportion of low-income adult Americans who would never consider buying an electric car
- p_2 = proportion of high-income adult Americans who would never consider buying an electric car
- $p_1 - p_2$ is the difference between the proportion who would never consider buying an electric car for low-income and high-income adult Americans

2. $H_0: p_1 - p_2 = 0$ ($p_1 = p_2$)
3. $H_a: p_1 - p_2 \neq 0$

Formulate a plan ➤

4. Significance level: $\alpha = 0.05$
5. Test Statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_c(1 - \hat{p}_c)}{n_1} + \frac{\hat{p}_c(1 - \hat{p}_c)}{n_2}}}$$

6. Assumptions: There are two conditions that need to be met in order for the large sample test for a difference in population proportions to be appropriate.

The large samples condition is easily verified. The sample sizes are large enough because

$$\begin{aligned} n_1 \hat{p}_1 &= 300(0.65) = 195 & n_1(1 - \hat{p}_1) &= 300(0.35) = 105 \\ n_2 \hat{p}_2 &= 300(0.59) = 177 & n_2(1 - \hat{p}_2) &= 300(0.41) = 183 \end{aligned}$$

are all greater than 10.

From the study description, we know that the samples were independently selected. We also know that the samples were selected in a way that Gallup believed would result in samples that were representative of adult Americans in the two income groups.

Do the work ➤

7. Calculations:

To calculate the value of the test statistic, we need to know the values of the sample proportions and the value of \hat{p}_c , the combined estimate of the common population proportion.

$$n_1 = 300 \quad \hat{p}_1 = 0.65$$

$$n_2 = 300 \quad \hat{p}_2 = 0.59$$

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{300(0.65) + 300(0.59)}{600} = 0.62$$

We can now calculate the value of the test statistic:

$$\begin{aligned} z &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_c(1 - \hat{p}_c)}{n_1} + \frac{\hat{p}_c(1 - \hat{p}_c)}{n_2}}} \\ &= \frac{0.65 - 0.59}{\sqrt{\frac{(0.62)(0.38)}{300} + \frac{(0.62)(0.38)}{300}}} \\ &= \frac{0.06}{0.04} = 1.50 \end{aligned}$$

- 8. P-value:** This is a two-tailed test (the inequality in H_a is \neq), so the P -value is twice the area under the z curve and to the right of the calculated z value:

$$\begin{aligned} P\text{-Value} &= 2(\text{area under } z \text{ curve to the right of 1.50}) \\ &= 2 \cdot P(z > 1.50) \\ &= 2(0.0668) \\ &= 0.1336 \end{aligned}$$

Interpret the results ➤

- 9. Because the P -value (0.1336) is greater than the selected significance level (0.05), we fail to reject the null hypothesis.** Even though the two sample proportions were different (0.65 and 0.59), this difference could have occurred just by chance when there is no difference in the population proportions. Based on the sample data, we are *not convinced* that there is a difference in the two population proportions.

It is also possible to use statistical software or a graphing calculator to calculate the value of the test statistic and the P -value. For example, Minitab output for this example is shown on the next page.

```
Test and CI for Two Proportions
Sample   X     N   Sample p
1       195   300  0.650000
2       177   300  0.590000
Difference = p(1) - p(2)
Estimate for difference: 0.06
95% CI for difference: (-0.0175281, 0.137528)
Test for difference = 0 (vs not = 0): Z = 1.51   P-Value = 0.130
```

From the Minitab output, we see that $z = 1.51$ and the associated P -value is 0.130. These values are slightly different from those calculated by hand only because Minitab uses greater decimal accuracy in the calculations leading to the value of the test statistic.

Example 11.10 Not Enough Sleep?

Do people who work long hours have more trouble sleeping? This question was examined in the paper “**Long Working Hours and Sleep Disturbances: The Whitehall II Prospective Cohort Study**” (*Sleep* [2009]: 737–745). The data in the accompanying table are from two independently selected samples of British civil service workers, all of whom were employed full-time and worked at least 35 hours per week. The authors of the paper believed that these samples were representative of full-time British civil service workers who work 35 to 40 hours per week and of British civil service workers who work more than 40 hours per week.

	n	Number who usually get less than 7 hours of sleep a night
Work over 40 hours per week	1501	750
Work 35–40 hours per week	958	407

Do these data support the theory that the proportion who usually get less than 7 hours of sleep a night for those who work more than 40 hours per week is greater than the proportion for those who work between 35 and 40 hours per week? We will carry out a hypothesis test with $\alpha = 0.01$. For these samples

$$\text{Over 40 hours per week} \quad n_1 = 1501 \quad \hat{p}_1 = \frac{750}{1501} = 0.500$$

$$\text{Between 35 and 40 hours per week} \quad n_2 = 958 \quad \hat{p}_2 = \frac{407}{958} = 0.425$$

Understand the context ➤

1. p_1 = proportion of British civil service workers who work more than 40 hours per week who get less than 7 hours of sleep
 p_2 = proportion of British civil service workers who work between 35 and 40 hours per week who get less than 7 hours of sleep
2. $H_0: p_1 - p_2 = 0$
3. $H_a: p_1 - p_2 > 0$
4. Significance level: $\alpha = 0.01$

Formulate a plan ➤

5. Test statistic:
$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}}$$

6. Assumptions: The two samples were independently selected. It is reasonable to regard the samples as representative of the two populations of interest. Checking to make sure that the sample sizes are large enough by using $n_1 = 1501$, $\hat{p}_1 = 0.500$, $n_2 = 958$, and $\hat{p}_2 = 0.425$, we have

$$n_1\hat{p}_1 = 750.50 \geq 10$$

$$n_1(1 - \hat{p}_1) = 750.50 \geq 10$$

$$n_2\hat{p}_2 = 407.15 \geq 10$$

$$n_2(1 - \hat{p}_2) = 550.85 \geq 10$$

Do the work ➤

7. Calculations: Minitab output is shown below. From the output, $z = 3.64$.

```
Test for Two Proportions
Sample   X     N   Sample p
1       750   1501  0.499667
2       407   958   0.424843
Difference = p (1) - p (2)
Estimate for difference: 0.0748235
Test for difference = 0 (vs > 0): Z = 3.64 P-Value = 0.000
```

8. *P*-value: From the computer output, *P*-value = 0.000

Interpret the results ➤

9. Conclusion: Because *P*-value (0.000) is less than α (0.01), the null hypothesis is rejected. There is convincing evidence that the proportion who get less than 7 hours of sleep a night is greater for British civil service workers who work more than 40 hours per week than it is for those who work between 35 and 40 hours per week.

Notice that because the data were from an observational study, we are not able to conclude that there is a cause and effect relationship between work hours and sleep. Although we can conclude that a greater proportion of those who work long hours get less than 7 hours of sleep a night, we can't conclude that working long hours is the cause of shorter sleep. We should also note that the sample was selected from British civil service workers, so it would not be a good idea to generalize this conclusion to all workers.

A Confidence Interval

A large-sample confidence interval for $p_1 - p_2$ is a special case of the general *z* interval formula

point estimate \pm (*z* critical value)(estimated standard deviation)

The statistic $\hat{p}_1 - \hat{p}_2$ gives a point estimate of $p_1 - p_2$, and the standard deviation of this statistic is

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

An estimated standard deviation is obtained by using the sample proportions \hat{p}_1 and \hat{p}_2 in place of p_1 and p_2 under the square-root symbol. Notice that this estimated standard deviation differs from the one used previously in the test statistic. When constructing a confidence interval, there isn't a null hypothesis that claims $p_1 = p_2$, so there is no assumed common value of p to estimate.

A Large-Sample Confidence Interval for $p_1 - p_2$

When

1. the samples are *independently selected random samples or treatments were assigned at random to individuals or objects* (or vice versa), and
2. both *sample sizes are large*:

$$n_1\hat{p}_1 \geq 10 \quad n_1(1-\hat{p}_1) \geq 10 \quad n_2\hat{p}_2 \geq 10 \quad n_2(1-\hat{p}_2) \geq 10$$

a large-sample confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm (\text{z critical value}) \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example 11.11 Cell Phone Etiquette

Understand the context ➤

As part of a survey conducted by Pew Research ([“Americans’ Views on Mobile Etiquette,” August 26, 2015, pewinternet.org, retrieved December 12, 2016](#)), people in a representative sample of 708 American adults age 18 to 29 were asked if they thought it was OK to use a cellphone while at a restaurant. The same question was asked of a representative sample of 1029 adult Americans age 30 to 49. You might expect that the proportion who think it is OK to use a cellphone at a restaurant is higher for the 18 to 29 age group than for the 30 to 49 age group, but how much higher? Based on these sample data, what can we learn about the actual difference in proportions for these two populations?

Formulate a plan ➤

To answer this question, we can construct a 90% confidence interval for a difference in population proportions. We will estimate the value of $p_1 - p_2$, where p_1 is the proportion who think it is OK to use a cell phone in a restaurant for the 18 to 29 age group and p_2 is the proportion for the 30 to 49 age group.

There are two conditions that need to be met in order to use the large sample confidence interval for a difference in population proportions. The large sample condition is easily verified. With “success” denoting a person who thinks it is OK to use a cell phone in a restaurant, the sample sizes are large enough because there are more than 10 successes (354 in sample 1 and 412 in sample 2) and more than 10 failures ($708 - 354 = 354$ in sample 1 and $1029 - 412 = 617$ in sample 2) in each sample. This is equivalent to checking that $n_1\hat{p}_1, n_1(1 - \hat{p}_1), n_2\hat{p}_2$, and $n_2(1 - \hat{p}_2)$ are all greater than 10.

The requirement of independent random samples or samples that are representative of the corresponding populations is more difficult. The researchers who collected these data indicate in the report that they believe that the samples are representative of the two populations of interest.

Do the work ➤

A 90% confidence interval for $p_1 - p_2$ can now be calculated:

$$\begin{aligned} n_1 &= 708 & n_2 &= 1029 \\ \hat{p}_1 &= \frac{354}{708} = 0.50 & \hat{p}_2 &= \frac{412}{1029} = 0.40 \end{aligned}$$

For a confidence level of 90%, the appropriate z critical value is 1.645. Substituting the values for n_1, n_2, \hat{p}_1 , and \hat{p}_2 into the 90% confidence interval formula results in

$$\begin{aligned} (\hat{p}_1 - \hat{p}_2) \pm (z \text{ critical value}) \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \\ (0.50 - 0.40) \pm (1.645) \sqrt{\frac{(0.50)(1 - 0.50)}{708} + \frac{(0.40)(1 - 0.40)}{1029}} \\ 0.10 \pm (1.645)\sqrt{0.0006} \\ 0.10 \pm (1.645)(0.024) \\ 0.10 \pm 0.039 \\ (0.061, 0.139) \end{aligned}$$

Statistical software or a graphing calculator could also have been used to compute the end points of the confidence interval. Minitab output is shown here. (Minitab has used more decimal accuracy in computing the endpoints of the confidence interval.)

CI for Two Proportions

Sample	X	N	Sample p
1	354	708	0.500000
2	412	1029	0.400389

Difference = $p(1) - p(2)$

Estimate for difference: 0.0996113

90% CI for difference: (0.0597794, 0.139443)

Interpret the results ➤

We can be 90% confident that the actual difference in the proportion of people who think it is OK to use a cell phone in a restaurant for 18- to 29-year-olds and for 30- to 49-year-olds is between 0.061 and 0.139. This means that we can be confident that the proportion is greater for 18- to 29-year-olds than for 30- to 49-year-olds by somewhere between 0.061 and 0.139. The method used to construct this interval estimate is successful in capturing the actual value of the difference in population proportions about 90% of the time.

EXERCISES 11.41 - 11.59

- 11.41** Some people seem to believe that you can fix anything with duct tape. Even so, many were skeptical when researchers announced that duct tape may be a more effective and less painful alternative than liquid nitrogen, which doctors routinely use to freeze warts. The article “[What a Fix-It: Duct Tape Can Remove Warts](#)” (*San Luis Obispo Tribune*, October 15, 2002) described a study conducted at Madigan Army Medical Center. Patients with warts were randomly assigned to either the duct tape treatment or the more traditional freezing treatment.

Those in the duct tape group wore duct tape over the wart for 6 days, then removed the tape, soaked the area in water, and used an emery board to scrape the area. This process was repeated for a maximum of 2 months or until the wart was gone. Data consistent with values in the article are summarized in the following table:

Treatment	n	Number with Wart Successfully Removed
Liquid nitrogen freezing	100	60
Duct tape	104	88

Do these data suggest that freezing is less successful than duct tape in removing warts? Test the relevant hypotheses using a significance level of 0.01. (Hint: See Example 11.9.)

- 11.42** The report titled “[Digital Democracy Survey](#)” (2016, deloitte.com/us/tmttrends, retrieved December 16, 2016) stated that 31% of the people in a representative sample of adult Americans ages 33 to 49 rated a landline telephone among the three most important services that they purchase for their home. In a representative sample of adult Americans age 50 to 68, 48% rated a landline telephone as one of the top three services they purchase for their home. Suppose that the samples were independently selected and that the sample size was 600 for the 33 to 49 age group sample and 650 for the 50 to 68 age group sample. Does this data provide convincing evidence that the proportion of adult Americans age 33 to 49 who rate a landline phone in the top three is less than this proportion for adult Americans age 50 to 68? Test the relevant hypotheses using $\alpha = 0.05$.

- 11.43** After the 2010 earthquake in Haiti, many charitable organizations conducted fundraising campaigns to raise money for emergency relief. Some of these campaigns allowed people to donate by sending a text message using a cell phone to have the donated

amount added to their cell-phone bill. The report “[Early Signals on Mobile Philanthropy: Is Haiti the Tipping Point?](#)” (*Edge Research*, 2010) describes the results of a national survey of 1526 people that investigated the ways in which people made donations to the Haiti relief effort.

The report states that 17% of Gen Y respondents (those born between 1980 and 1988) and 14% of Gen X respondents (those born between 1968 and 1979) said that they had made a donation to the Haiti relief effort via text message. The percentage making a donation via text message was much lower for older respondents. The report did not say how many respondents were in the Gen Y and Gen X samples, but for purposes of this exercise, suppose that both sample sizes were 400 and that it is reasonable to regard the samples as representative of the Gen Y and Gen X populations.

- Is there convincing evidence that the proportion of those in Gen Y who donated to Haiti relief via text message is greater than the proportion for Gen X? Use $\alpha = 0.01$.
- Estimate the difference between the proportion of Gen Y and the proportion of Gen X that made a donation via text message using a 99% confidence interval. Provide an interpretation of both the interval and the associated confidence level.

- 11.44** The article “[Most Women Oppose Having to Register for the Draft](#)” (February 10, 2016, rasmussenreports.com, retrieved December 15, 2016) describes a survey of likely voters in the United States. The article states that 36% of those in a representative sample of male likely voters and 21% of those in a representative sample of female likely voters said that they thought the United States should have a military draft. Suppose that these percentages were based on random samples of 500 men and 500 women.

Use a significance level of 0.01 to determine if there is convincing evidence that the proportion of male likely voters who think the United States should have a military draft is different from this proportion for female likely voters.

- 11.45** The article referenced in the previous exercise also reported that 53% of the Republicans surveyed indicated that they were opposed to making women register for the draft. Would you use the large-sample test for a difference in population proportions to test the hypothesis that a majority of Republicans are opposed to making women register for the draft? Explain why or why not.

11.46 The report “**Audience Insights: Communicating to Teens (Aged 12–17)**” (cdc.gov, 2009) described teens’ attitudes about traditional media, such as TV, movies, and newspapers. In a representative sample of American teenage girls, 41% said newspapers were boring. In a representative sample of American teenage boys, 44% said newspapers were boring. Sample sizes were not given in the report.

- a. Suppose that the percentages reported had been based on a sample of 58 girls and 41 boys. Is there convincing evidence that the proportion of those who think that newspapers are boring is different for teenage girls and boys? Carry out a hypothesis test using $\alpha = 0.05$.
- b. Suppose that the percentages reported had been based on a sample of 2000 girls and 2500 boys. Is there convincing evidence that the proportion of those who think that newspapers are boring is different for teenage girls and boys? Carry out a hypothesis test using $\alpha = 0.05$.
- c. Explain why the hypothesis tests in Parts (a) and (b) resulted in different conclusions.

11.47 According to the U.S. Census Bureau, the percentage of U.S. residents living in poverty in 2015 was 12.2% for men and 14.8% for women. These percentages were estimates based on data from large representative samples of men and women. Suppose that the sample sizes were 1200 for the sample of men and 1000 for the sample of women. Use the survey data to calculate and interpret a 90% confidence interval for the difference in the proportion living in poverty for men and women.

11.48 The article “**Fish Oil Staves Off Schizophrenia**” ([USA TODAY](http://USA Today), February 2, 2010) describes a study in which 81 patients age 13 to 25 who were considered at-risk for mental illness were randomly assigned to one of two groups. Those in one group took four fish oil capsules daily. The other group took a placebo. After 1 year, 5% of those in the fish oil group and 28% of those in the placebo group had become psychotic.

Is it appropriate to use the two-sample z test of this section to test hypotheses about the difference in the proportions of patients receiving the fish oil and the placebo treatments who became psychotic? Explain why or why not.

11.49 The report “**Young People Living on the Edge**” (Greenberg Quinlan Rosner Research, 2008) summarizes a survey of people in two independent random samples. One sample consisted of 600 young adults (age 19 to 35) and the other sample consisted of 300 parents of children age 19 to 35. The young adults were presented with a variety of situations (such as getting married or buying a house) and were asked if they thought that their parents were

likely to provide financial support in that situation. The parents of young adults were presented with the same situations and asked if they would be likely to provide financial support to their child in that situation.

- a. When asked about getting married, 41% of the young adults said they thought parents would provide financial support and 43% of the parents said they would provide support. Carry out a hypothesis test to determine if there is convincing evidence that the proportion of young adults who think parents would provide financial support and the proportion of parents who say they would provide support are different.
- b. The report stated that the proportion of young adults who thought parents would help with buying a house or apartment was 0.37. For the sample of parents, the proportion who said they would help with buying a house or an apartment was 0.27. Based on these data, can you conclude that the proportion of parents who say they would help with buying a house or an apartment is significantly less than the proportion of young adults who think that their parents would help?

11.50 The report “**Raising Kids and Running a Household: How Working Parents Share the Load**” (November 4, 2015, Pew Research Center, pewresearch.org, retrieved December 12, 2016) described a survey of parents of children under the age of 18. Each person in a representative sample of 825 working fathers and a sample of 586 working mothers was asked if balancing the responsibilities of a job and a family was difficult. It was reported that 429 (52%) of the fathers surveyed and 352 (60%) of the mothers surveyed said that it was difficult. The two samples were independently selected and were thought to be representative of working fathers and mothers of children under 18 years old. Use this information to calculate and interpret a 95% confidence interval estimate of the difference between the proportion of working fathers who find balancing work and family difficult, p_1 , and this proportion for working mothers, p_2 .

11.51 The Bureau of Labor Statistics (bls.gov/opub/ted/2014/ted_20141112.htm, retrieved December 13, 2016) reported that 3.8% of college graduates were unemployed in October 2013 and 3.1% of college graduates were unemployed in October 2014. Suppose that the reported percentages were based on independently selected representative samples of 500 college graduates in each of these two years. Construct and interpret a 95% confidence interval for the difference in the proportion of college graduates who were unemployed in these two years.

- 11.52** The Bureau of Labor Statistics report referenced in the previous exercise also indicated that 7.3% of high school graduates were unemployed in October 2013 and 5.7% of high school graduates were unemployed in October 2014. Suppose that the reported percentages were based on independently selected representative samples of 400 high school graduates in each of these two years.
- Construct and interpret a 99% confidence interval for the difference in the proportion of high school graduates who were unemployed in these two years.
 - Is the confidence interval from Part (a) wider or narrower than the confidence interval calculated in the previous exercise? What are two reasons why it is wider or narrower?
- 11.53** The report “[The New Food Fights: U.S. Public Divides Over Food Science](#)” ([December 1, 2016, pewinternet.org, retrieved December 10, 2016](#)) states that younger adults are more likely to see foods with genetically modified ingredients as being bad for their health than older adults. This statement is based on a representative sample of 178 adult Americans age 18 to 29 and a representative sample of 427 adult Americans age 50 to 64. Of those in the 18 to 29 age group, 48% said they believed these foods were bad for their health, while only 38% of those in the 50 to 64 age group believed this.
- Are the sample sizes large enough to use the large-sample confidence interval to estimate the difference in the population proportions?
 - Estimate the difference in the proportion of adult Americans age 18 to 29 who believe the foods made with genetically modified ingredients are bad for their health and the corresponding proportion for adult Americans age 50 to 64. Use a 90% confidence interval.
 - Is zero in the confidence interval? What does this suggest about the difference in the two population proportions?
- 11.54** A survey of high school students is described in the report “[Students on STEM](#)” ([changetheequation.org/students-stem, retrieved December 12, 2016](#)). The report states that 14% of those in a sample of students in low-income households (defined as a household income less than \$50,000 per year) and 24% of those in a sample of students in higher income households (defined as a household income of \$50,000 or more) participated in a science club or group. Suppose that these samples are representative of high school students in the two income groups and that the two sample sizes were both 500. Use a 95% confidence interval to estimate the difference in the proportion participating in a science club for students in the two income groups.
- 11.55** The article “[Spray Flu Vaccine May Work Better Than Injections for Tots](#)” ([San Luis Obispo Tribune, May 2, 2006](#)) described a study that compared flu vaccine administered by injection and flu vaccine administered as a nasal spray. Each of the 8000 children under the age of 5 who participated in the study received both a nasal spray and an injection, but only one was the real vaccine and the other was salt water. At the end of the flu season, it was determined that 3.9% of the 4000 children receiving the real vaccine by nasal spray got sick with the flu and 8.6% of the 4000 receiving the real vaccine by injection got sick with the flu.
- Why would the researchers give every child both a nasal spray and an injection?
 - Use the given data to estimate the difference in the proportion of children who get sick with the flu after being vaccinated with an injection and the proportion of children who get sick with the flu after being vaccinated with the nasal spray using a 99% confidence interval.
 - Based on the confidence interval, would you conclude that the proportion of children who get the flu is different for the two vaccination methods?
- 11.56** The following quote is from the article “[Canadians Are Healthier Than We Are](#)” ([Associated Press, May 31, 2006](#)): “The Americans also reported more heart disease and major depression, but those differences were too small to be statistically significant.” This statement was based on the responses of a sample of 5183 Americans and a sample of 3505 Canadians. The proportion of Canadians who reported major depression was given as 0.082.
- Assuming that the researchers used a one-sided test with a significance level of 0.05, could the sample proportion of Americans reporting major depression have been as great as 0.09? Explain why or why not.
 - Assuming that the researchers used a significance level of 0.05, could the sample proportion of Americans reporting major depression have been as great as 0.10? Explain why or why not.
- 11.57** A Harris Poll press release dated November 1, 2016, summarized results of a survey of 2463 adults and 510 teens age 13 to 17 (“[American Teens No Longer More Likely than Adults to Believe in God, Miracles, Heaven, Jesus, Angels, or the Devil](#),” [theharrispoll.com, retrieved December 12, 2016](#)). It was reported that 19% of the teens surveyed and 26% of the adults surveyed indicated that they believe in reincarnation. The samples were selected to be representative of American adults and teens. Use the data from this survey to estimate the difference in the proportion of adults who believe in reincarnation and the proportion of teens who believe in

reincarnation. Be sure to interpret your interval in context.

- 11.58** Women diagnosed with breast cancer whose tumors have not spread may be faced with a decision between two surgical treatments—mastectomy (removal of the breast) or lumpectomy (only the tumor is removed). In a long-term study of the effectiveness of these two treatments, 701 women with breast cancer were randomly assigned to one of two treatment groups. One group received mastectomies and the other group received lumpectomies and radiation. Both groups were followed for 20 years after surgery. It was reported that there was no statistically significant difference in the proportion surviving for 20 years for the two treatments (*Associated Press, October 17, 2002*).

What hypotheses do you think the researchers tested in order to reach the given conclusion?

Did the researchers reject or fail to reject the null hypothesis?

- 11.59** Gallup surveyed adult Americans about their consumer debt ("[Americans' Big Debt Burden Growing, Not Evenly Distributed](#)," February 4, 2016, [gallup.com](#), retrieved December 15, 2016). The article reported that 47% of millennials (those born between 1980 and 1996) and 61% of Gen Xers (those born between 1965 and 1971) did not pay off their credit cards each month, and therefore carried a balance from month to month. Suppose that these percentages were based on representative samples of 450 millennials and 300 Gen Xers. Is there convincing evidence that the proportion of Gen Xers who do not pay off their credit cards each month is greater than this proportion for millennials? Test the appropriate hypotheses using a significance level of 0.05.

SECTION 11.4 Interpreting and Communicating the Results of Statistical Analyses

Many different types of research involve comparing two populations or treatments. It is easy to find examples of the two-sample hypothesis tests introduced in this chapter in published sources in a wide variety of disciplines.

Communicating the Results of Statistical Analyses

As was the case with one-sample hypothesis tests, it is important to include a description of the hypotheses, the test procedure used, the value of the test statistic, the *P*-value, and a conclusion in context when summarizing the results of a two-sample test.

Correctly interpreting confidence intervals in the two-sample case is more difficult than in the one-sample case, so take particular care when providing a two-sample confidence interval interpretation. Because the two-sample confidence intervals of this chapter estimate a difference ($\mu_1 - \mu_2$ or $p_1 - p_2$), the most important thing to note is whether or not the interval includes 0. If both endpoints of the interval are positive, then it is correct to say that, based on the interval, you believe that μ_1 is greater than μ_2 (or that p_1 is greater than p_2 if you are working with proportions) and then the interval provides an estimate of how much greater. Similarly, if both interval endpoints are negative, you would say that μ_1 is less than μ_2 (or that p_1 is less than p_2), with the interval providing an estimate of the size of the difference. If 0 is included in the interval, it is plausible that μ_1 and μ_2 (or p_1 and p_2) are equal.

Interpreting the Results of Statistical Analyses

As with one-sample tests, it is common to find only the value of the test statistic and the associated *P*-value (or sometimes only the *P*-value) in published reports. You may have to think carefully about the missing steps to determine whether or not the conclusions are justified.

What to Look For in Published Data

Here are some questions to consider when you are reading a report that contains the result of a two-sample hypothesis test or confidence interval:

- Are only two groups being compared? If more than two groups are being compared two at a time, then a different type of analysis is preferable (see Chapter 15).

- Were the samples selected independently, or were the samples paired? If the samples were paired, was the analysis that was performed appropriate for paired samples?
- If a confidence interval is reported, is it correctly interpreted as an estimate of a difference in population or treatment means or proportions?
- What hypotheses are being tested? Is the test one- or two-tailed?
- Does the validity of the test performed depend on any assumptions about the sampled populations (such as normality)? If so, do the assumptions appear to be reasonable?
- What is the P -value associated with the test? Does the P -value lead to rejection of the null hypothesis?
- Are the conclusions consistent with the results of the hypothesis test? In particular, if H_0 was rejected, does this indicate practical significance or only statistical significance?

For example, the paper “**Ginkgo for Memory Enhancement**” (*Journal of the American Medical Association* [2003]: 835–840) included the following statement in the summary of conclusions from an experiment where participants were randomly assigned to receive ginkgo or a placebo:

Figure 2 shows the 95% confidence intervals (CIs) for differences (treatment group minus control) for performance on each test in the modified intent-to-treat analysis. Each interval contains a zero, indicating that none of the differences are statistically significant.

Because participants were assigned at random to the two treatments and the sample sizes were large (115 in each sample), use of the two-sample t confidence interval was appropriate. The 95% confidence intervals included in the paper (for example, $(-1.71, 0.65)$ and $(-2.25, 0.20)$ for two different measures of logical memory) did all include 0 and were interpreted correctly in the quoted conclusion.

As another example, we consider a study reported in the article “**The Relationship Between Distress and Delight in Males’ and Females’ Reactions to Frightening Films**” (*Human Communication Research* [1991]: 625–637). The investigators measured emotional responses of 50 males and 60 females after the subjects viewed a segment from a horror film. The article included the following statement:

“Females were much more likely to express distress than were males. While males did express higher levels of delight than females, the difference was not statistically significant.”

The following summary information was also contained in the article:

Gender	Distress Index Mean	Delight Index Mean
Males	31.2	12.02
Females	40.4	9.09
P -value < 0.001		Not significant (P -value > 0.05)

The P -values are the only evidence of the hypothesis tests that support the given conclusions. The P -value < 0.001 for the distress index means that the hypothesis $H_0: \mu_F - \mu_M = 0$ was rejected in favor of $H_a: \mu_F - \mu_M > 0$, where μ_F and μ_M are the mean distress indexes for females and males.

The nonsignificant P -value (P -value > 0.05) reported for the delight index means that the hypothesis $H_0: \mu_F - \mu_M = 0$ (where μ_F and μ_M now refer to mean delight index for females and males, respectively) could not be rejected. Chance sample-to-sample variability is a plausible explanation for the observed difference in sample means. We would want to be cautious about the author’s statement that males express higher levels of delight than females, because it is based only on the fact that $12.02 > 9.09$, which could be explained by sampling variability alone.

The article describes the samples as consisting of undergraduates selected from the student body of a large Midwestern university. The authors generalize their results to all American men and women. If this type of generalization is considered unreasonable, we

could be more conservative and view the sampled populations as male and female university students or male and female Midwestern university students or even male and female students at this particular university.

The comparison of males and females was based on two independently selected samples (not paired). Because the sample sizes were large, the two-sample t test for means could reasonably have been used, and this would have required no specific assumptions about the shape of the two underlying population distributions.

In a newspaper article, you may find even less information than in a journal article. For example, the article “[Prayer Is Little Help to Some Heart Patients, Study Shows](#)” (*Chicago Tribune*, March 31, 2006) included the following paragraphs:

Bypass patients who consented to take part in the experiment were divided randomly into three groups. Some patients received prayers but were not informed of that. In the second group the patients got no prayers, and also were not informed one way or the other. The third group got prayers and were told so.

There was virtually no difference in complication rates between the patients in the first two groups. But the third group, in which patients knew they were receiving prayers, had a complication rate of 59 percent—significantly more than the rate of 52 percent in the no-prayer group.

Earlier in the article, the total number of participants in the experiment was given as 1800. The author of this article has done a good job of describing the important aspects of the experiment. The final comparison in the quoted paragraph was probably based on a two-sample z test for proportions, comparing the sample proportion with complications for the 600 patients in the no-prayer group with the sample proportion with complications for the 600 participants who knew that someone was praying for them.

For the reported sample sizes and sample proportions, the test statistic for testing $H_0: p_1 - p_2 = 0$ versus $H_a: p_1 - p_2 < 0$ (where p_1 represents the complication proportion for patients who did not receive prayers and p_2 represents the complication proportion for patients who knew they were receiving prayers) is $z = -2.10$. The associated P -value is 0.036, supporting the conclusion stated in the article.

A Word to the Wise: Cautions and Limitations

The three cautions that appeared at the end of Chapter 10 apply here as well. They were (see Chapter 10 for more detail):

1. Remember that the result of a hypothesis test can never show strong support for the null hypothesis. In two-sample situations, this means that we shouldn’t be *convinced* that there is no difference between population means or proportions based on the outcome of a hypothesis test.
2. If you have complete information (a census) of both populations, there is no need to carry out a hypothesis test or to construct a confidence interval—in fact, it would be inappropriate to do so.
3. Don’t confuse statistical significance and practical significance. In the two-sample setting, it is possible to be convinced that two population means or proportions are not equal even in situations where the actual difference between them is small enough that it is of no practical interest. After rejecting a null hypothesis of no difference (statistical significance), it is useful to look at a confidence interval estimate of the difference to get a sense of practical significance.

And here’s one new caution to keep in mind for two-sample tests:

4. Be sure to think carefully about how the data were collected, and make sure that an appropriate test procedure or confidence interval is used. A common mistake is to overlook pairing and to analyze paired samples as if they were independent. It is usually easy to tell if the samples are paired—you just have to remember to think about how the samples were selected.

EXERCISES 11.60 - 11.62

- 11.60** The paper “[The Psychological Consequences of Money](#)” (*Science* [2006]: 1154–1156) describes several experiments designed to investigate the way in which money can change behavior. In one experiment, participants completed one of two versions of a task in which they were given lists of five words and were asked to rearrange four of the words to create a sensible phrase. For one group, half of the 30 unscrambled phrases related to money, whereas the other half were phrases that were unrelated to money. For the second group (the control group), none of the 30 unscrambled phrases related to money. Participants were 44 students at Florida State University.

Participants received course credit and \$2 for their participation. The following description of the experiment is from the paper:

Participants were randomly assigned to one of two conditions, in which they descrambled phrases that primed money or neutral concepts. Then participants completed some filler questionnaires, after which the experimenter told them that the experiment was finished and gave them a false debriefing. This step was done so that participants would not connect the donation opportunity to the experiment. As the experimenter exited the room, she mentioned that the lab was taking donations for the University Student Fund and that there was a box by the door if the participant wished to donate. Amount of money donated was the measure of helping. We found that participants primed with money donated significantly less money to the student fund than participants not primed with money [$t(38) = 2.13, P < 0.05$].

The paper also gave the following information on amount donated for the two experimental groups.

Group	Mean	Standard Deviation
Money primed	\$0.77	\$0.74
Control	\$1.34	\$1.02

- Explain why the random assignment of participants to experimental groups is important in this experiment.
- Use the given information to verify the values of the test statistic and degrees of freedom (38, given in parentheses just after the t in the quote from the paper) and the statement about the P -value. Assume that both sample sizes are 22.
- Do you think that use of the two-sample t test was appropriate in this situation? Hint: Are the assumptions required for the two-sample t test reasonable?

- 11.61** An experiment to determine if an online intervention can reduce references to sex and substance abuse on social networking web sites of adolescents is described in the paper “[Reducing At-Risk Adolescents' Display of Risk Behavior on a Social Networking Web Site](#)” (*Archives of Pediatrics and Adolescent Medicine* [2009]: 35–41). Researchers selected public MySpace profiles of people who described themselves as between 18 and 20 years old and who referenced sex or substance use (alcohol or drugs) in their profiles. The selected subjects were assigned at random to an intervention group or a control group.

Those in the intervention group were sent an e-mail from a physician about the risks associated with having a public profile and of referencing sex or substance use in their profile. Three months later, networking sites were revisited to see if any changes had been made. The following excerpt is from the paper:

At baseline, 54.2% of subjects referenced sex and 85.3% referenced substance use on their social networking site profiles. The proportion of profiles in which references decreased to 0 was 13.7% in the intervention group vs. 5.3% in the control group for sex ($P = .05$) and 26% vs. 22% for substance use ($P = .61$). The proportion of profiles set to “private” at follow-up was 10.5% in the intervention group and 7.4% in the control group ($P = .45$). The proportion of profiles in which any of these three protective changes were made was 42.1% in the intervention group and 29.5% in the control group ($P = .07$).

- The quote from the paper references four hypothesis tests. For each test, indicate what hypotheses you think were tested and whether or not the null hypothesis was rejected.
- Based on the information provided by the hypothesis tests, what conclusions can be drawn about the effectiveness of the e-mail intervention?

- 11.62** The paper “[Ready or Not? Criteria for Marriage Readiness among Emerging Adults](#)” (*Journal of Adolescent Research* [2009]: 349–375) surveyed emerging adults (defined as age 18 to 25) from five different colleges in the United States. Several questions on the survey were used to construct a scale designed to measure endorsement of cohabitation. The paper states that “on average, emerging adult men ($M = 3.75, SD = 1.21$) reported higher levels of cohabitation endorsement than emerging adult women ($M = 3.39, SD = 1.17$).” The sample sizes were 481 for women and 307 for men.

- a. Carry out a hypothesis test to determine if the reported difference in sample means provides convincing evidence that the mean cohabitation endorsement for emerging adult women is significantly less than the mean for emerging adult men for students at these five colleges.
- b. What additional information would you want in order to determine whether it is reasonable to generalize the conclusion of the hypothesis test from Part (a) to all college students?

SECTION 11.5 Simulation-Based Inference for Two Means (Optional)

Simulation-based methods for inference about one population mean were introduced in Chapter 10. In this section, simulation-based methods that allow us to test hypotheses and construct confidence intervals for a difference in means are introduced. These methods are especially useful when the conditions of the two-sample methods of Sections 11.1 and 11.2 are not met (when the sample sizes are small and it is not clear that the population distributions are normal).

Simulation-Based Inference About the Difference in Two Treatment Means

Simulation-based randomization tests and bootstrap confidence intervals may be used to learn about the difference in two treatment means using data from an experiment.

Example 11.12 Blue Light Exposure and Blood Glucose Level

- The article “[Bright Light at Night Time Can Seriously Mess with Your Metabolism, Study Finds](#)” (Science Alert, May 20, 2016, sciencealert.com/checking-your-phone-at-night-could-be-messing-with-your-metabolism, retrieved May 23, 2017) describes research conducted to examine the effects of blue light exposure (the type of light emitted by smart phones and computer screens) on a variety of measures, including blood glucose levels and sleepiness. The study was published in the [Public Library of Science \(“Morning and Evening Blue-Enriched Light Exposure Alters Metabolic Function in Normal Weight Adults,” PLOS One \[2016\]: e0155601\)](#). Adult volunteers of normal weight were randomly assigned to one of two groups. The first group was exposed to blue light for 3 hours in the morning (30 minutes after waking), and the second group was exposed to blue light for 3 hours in the evening (10 hours and 30 minutes after waking).

For each subject, a baseline blood glucose level was obtained 30 minutes before the blue light exposure began. Then blood glucose level was tracked every half hour for 4 hours during and immediately following the blue light exposure, including when the subject ate a meal. One outcome measure of the study was the peak change from baseline in blood glucose level. A negative peak change means that the subject’s blood glucose level remained below the baseline measurement during the entire time period. We can perform a hypothesis test using a significance level of 0.05 to determine whether the mean peak change in blood glucose level differs for the A.M. and P.M. blue light exposure groups.

Data for the 9 subjects in the A.M. blue light treatment group and the 10 subjects in the P.M. blue light treatment group are given in the following table.

Subject ID	Group	Glucose Peak Change from Baseline (mg/dL)
1	A.M.	4
2	A.M.	-22
3	A.M.	7
4	A.M.	4
5	A.M.	-8

● Data set available online

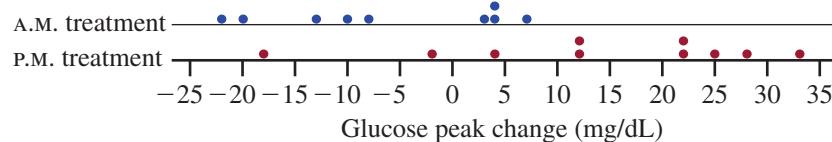
(continued)

Subject ID	Group	Glucose Peak Change from Baseline (mg/dL)
6	A.M.	-13
7	A.M.	3
8	A.M.	-20
9	A.M.	-10
10	P.M.	28
11	P.M.	12
12	P.M.	12
13	P.M.	33
14	P.M.	25
15	P.M.	22
16	P.M.	4
17	P.M.	22
18	P.M.	-2
19	P.M.	-18

The dotplots in Figure 11.2 indicate the assumption that the change distributions are approximately normal is questionable. Because both sample sizes are small, the two-sample t test and confidence interval might not be appropriate choices for analyzing the data from this experiment.

FIGURE 11.2

Dotplots of peak glucose change for the A.M. and P.M. treatments.



The two treatment group means can be represented as follows:

$$\begin{aligned}\mu_1 &= \text{Mean change in glucose level for the A.M. treatment} \\ \mu_2 &= \text{Mean change in glucose level for the P.M. treatment}\end{aligned}$$

Translating the research question regarding whether mean change in blood glucose level differs for the A.M. and P.M. blue light exposure treatments results in the following hypotheses:

$$\begin{aligned}H_0: \mu_1 - \mu_2 &= 0 \\ H_a: \mu_1 - \mu_2 &\neq 0\end{aligned}$$

Summary statistics for the data in the two samples are

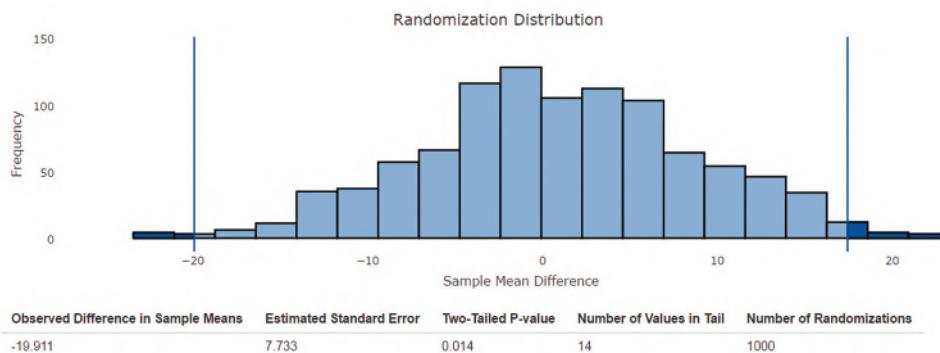
	A.M. Treatment	P.M. Treatment
Sample size	9	10
Sample mean	-6.1	13.8
Sample standard deviation	11.02	15.61

The observed difference in means for change in glucose level, A.M. – P.M., is $\bar{x}_1 - \bar{x}_2 = -19.9$.

A randomization test can be used to determine if there is convincing evidence that the two treatment means differ. In a randomization test using data from an experiment, alternative random assignments of the subjects in a study to two (or more) groups are considered. If the null hypothesis is true, there is no difference in the effect of the two treatments. This means that each subject would have had the same observed change whether he or she was in the A.M. treatment group or the P.M. treatment group. For example, subject 1, who was in the A.M. treatment group and had a change of 4, would have had a change of 4 even if that subject had been in the P.M. treatment group. If this is the case for every subject, the observed difference in sample means could be due to chance in the random assignment of subjects to treatment groups.

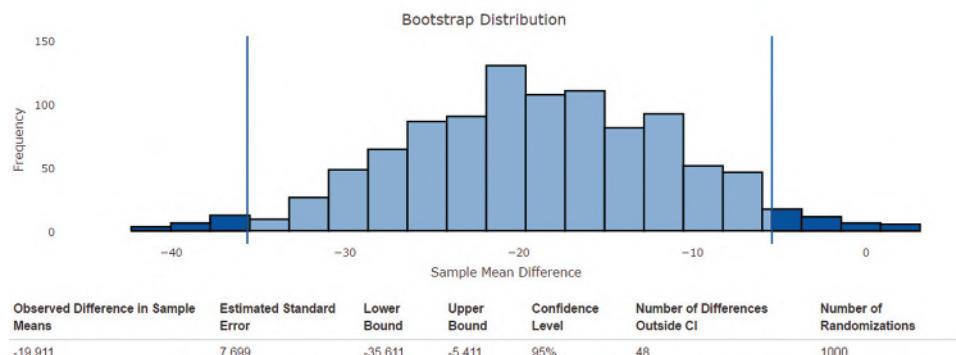
To decide if the observed difference is consistent with what is expected due to chance alone or whether it is evidence of a real difference in treatment means, we begin by exploring what chance differences look like. Simulation is carried out by taking the original 19 observations and randomly assigning them into two groups (one of size 9 and one of size 10) and then calculating a simulated difference in means. This process is repeated many times to form a randomization distribution.

The following figure shows the simulation results from the Shiny app “Randomization Test for Difference in Two Treatment Means” for the difference in means from many different random assignments of the observed data values into two independent groups. This app can be found in the collection that accompanies this text at statistics.cengage.com/PSO6e/Apps.html.



The observed difference in means was -19.9 . Locating -19.9 in the distribution of simulated differences shows that it would be unusual to observe a difference this extreme if the null hypothesis of no difference in treatment means were true. Based on the randomization distribution, the probability of observing a difference at least as extreme as -19.9 is approximately 0.014. Because this two-sided P -value is less than the specified significance level of 0.05, the null hypothesis is rejected. Among many random assignments of the changes in blood glucose levels to hypothetical A.M. and P.M. blue light exposure groups, less than 1.4% produce a difference in means that is at least as inconsistent with the null hypothesis as the observed difference in the means of -19.9 . This is evidence that the treatment means are not equal.

The Shiny app “Bootstrap Confidence Interval for Difference in Two Treatment Means” (found in the collection at statistics.cengage.com/PSO6e/Apps.html) can be used to obtain a 95% confidence interval for the difference in the treatment means by examining the distribution of differences in simulated means calculated from hypothetical randomizations of the combined sample of changes in blood glucose levels into two groups.



The given Shiny app output provides the distribution for 1000 simulated differences in the means for alternative random assignments of the changes in blood glucose levels to two groups. Identifying the 2.5% of differences in means on both the lowest and highest ends of the distributions provides the endpoints for a 95% bootstrap confidence interval.

For this simulation, the 95% bootstrap confidence interval for the difference in treatment means is $(-35.6, -5.4)$. We can be 95% confident that the actual difference in the mean change in blood glucose levels, A.M. – P.M., falls between -35.6 and -5.4 . Notice that this confidence interval does not include zero, and this is consistent with rejecting the null hypothesis that the difference in the mean change in glucose level for the two treatments is zero. With both endpoints negative, we can say that the mean change is greater for the P.M. blue light treatment than for the A.M. blue light treatment by somewhere between 5.4 and 35.6 mg/dL.

Simulation-Based Inference for the Difference in Two Population Means Using Independent Samples

We may wish to compare the means for two populations using independent samples selected at random from the populations, in contrast to using data from an experiment where subjects are assigned at random to treatment groups. The simulation-based methods for comparing two means using independent random samples are a little different from the methods used with data from an experiment.

Example 11.13 Freshman Year Weight Gain Revisited

- Example 11.4 described a study of freshman year weight gain (“Predicting the ‘Freshman 15’: Environmental and Psychological Predictors of Weight Gain in First-Year University Students,” *Health Education Journal* [2016]:321–332). In that study, weight gain (in kilograms) during the freshman year was recorded for independent random samples of first-year students who lived on campus and first-year students who lived off campus. The data from Example 11.4 are also given here.

On Campus	Off Campus
2.0	1.6
2.3	3.1
1.1	-2.8
-2.0	0.0
-1.9	0.2
5.6	2.9
2.6	-0.9
1.1	3.8
5.6	0.7
8.2	-0.1

These samples are quite small, but the boxplots of the data from the two samples were reasonably asymmetric and there were no outliers, so a two-sample t confidence interval was used to estimate the difference in population means. However, we might still choose to use a simulation-based method that does not depend on the assumption of normal population distributions.

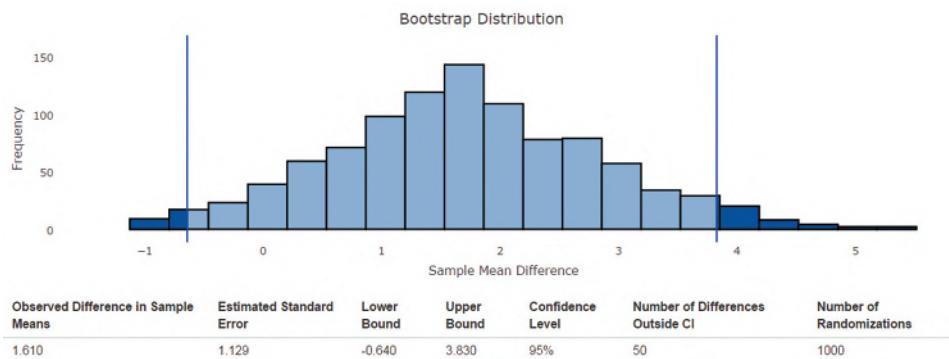
The means for the populations of first-year students living on and off campus were represented by

$$\begin{aligned}\mu_1 &= \text{population mean weight gain for first-year students living on campus} \\ \mu_2 &= \text{population mean weight gain for first-year students living off campus}\end{aligned}$$

In the case of sampling from two populations, we begin by assuming that each sample is representative of the population from which it was selected. Then, as was the case in the one-sample situation, a bootstrap sample from each population is selected using random sampling, with replacement, from each of the two original samples. The difference in the means of these two bootstrap samples is then calculated. This process is then repeated many times to form a bootstrap distribution for the difference in sample means.

After a bootstrap distribution of simulated differences in sample means is constructed, it can be used to produce a 95% confidence interval for the difference in the population means. The endpoints of the confidence interval are the value with 2.5% of the simulated differences in the bootstrap distribution below and the value with 2.5% of the simulated differences in the bootstrap distribution above. Locating these differences in the bootstrap distribution that follows gives a 95% confidence interval for the difference in mean weight gain for students living on campus and students living off campus of $(-0.64, 3.83)$. We can be 95% confident that the difference in population mean weight gains, On Campus – Off Campus, falls between -0.64 and 3.83 kg. These values are not very different from the endpoints of the two-sample t confidence interval from Example 11.4, which was $(-0.98, 4.20)$.

Here is output from the Shiny app “Bootstrap Confidence Interval for the Difference in Two Population Means Using Independent Samples.” This app can be found in the collection at statistics.cengage.com/PSO6e/Apps.html.



It is also possible to compare the population mean weight gains for students living on and off campus using a simulation-based hypothesis test, but the simulation is carried out differently than for the confidence interval because the null hypothesis states only that the population mean weights are equal without specifying their values. Suppose that we want to test the following hypotheses:

$$\begin{aligned} H_0: \mu_1 - \mu_2 &= 0 \\ H_a: \mu_1 - \mu_2 &\neq 0 \end{aligned}$$

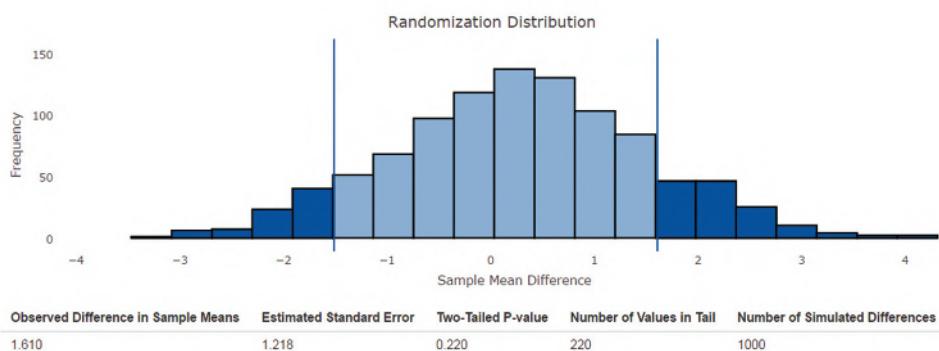
Summary statistics for the weight gains (in kg) calculated using the given data for students living on campus and for those living off campus follow.

	On Campus	Off Campus
Sample size	10	10
Sample mean	2.46	0.85
Sample standard deviation	3.26	2.03

The observed difference in the means for On Campus – Off Campus is $\bar{x}_1 - \bar{x}_2 = 2.46 - 0.85 = 1.61$ kg.

To create a randomization distribution, the two samples are combined into one, and then random samples are selected with replacement from the combined sample. This creates a randomization distribution of the difference in sample means that is consistent with the null hypothesis of no difference in population means.

One example of a randomization distribution based on 1000 simulated differences in means using the Shiny app “Randomization Test for the Difference in Two Population Means Using Independent Samples” is shown on the next page. This app can be found in the collection at statistics.cengage.com/PSO6e/Apps.html.



The randomization distribution in this case represents 1000 differences in two independent sample means, each calculated from two random samples of size 10 taken with replacement from the combined sample of 20 weight gains from students living on and off campus.

Using the randomization distribution from this simulation, the probability of observing a difference in the sample means at least as extreme as 1.61 is 0.220. This two-tailed *P*-value is larger than the 0.05 significance level. We fail to reject the null hypothesis, and conclude that there is not convincing evidence of a difference in the population means for weight gains of students living on campus and students living off campus.

Simulation-Based Inference About Two Population Means Using Paired Samples

In Section 11.2, when the samples are paired, hypothesis tests and confidence intervals for the difference in two population means involved calculating the sample differences and then using one-sample methods with the sample of differences. This is also the case for simulation-based methods—we calculate the sample differences and then use the one-sample bootstrap confidence interval and simulation-based randomization test that were introduced in Chapter 10.

Example 11.14 | Charitable Chimps Revisited

Example 11.7 described a study in which chimpanzees could deliver food to themselves when one lever was pushed and to both themselves and a chimp in an adjoining cage when another lever was pushed. Researchers recorded the number of times out of 16 that each of seven chimps chose the option to “feed both” when there was another chimp in the adjoining cage and when there was not a chimp in the adjoining cage. Because the samples were paired, the differences in the number of times the “feed both” option was chosen by chimpanzees were calculated.

The pairing produced the following set of differences for the seven chimps in the sample:

Chimp	1	2	3	4	5	6	7
Difference	2	0	-2	2	1	3	0

The sample size is small, and, despite the fairly symmetric boxplot presented in Example 11.7, we may still be hesitant to use the one-sample *t* test for the paired samples analysis. Instead, we may choose to use the simulation-based methods presented in Sections 10.7 and 10.8.

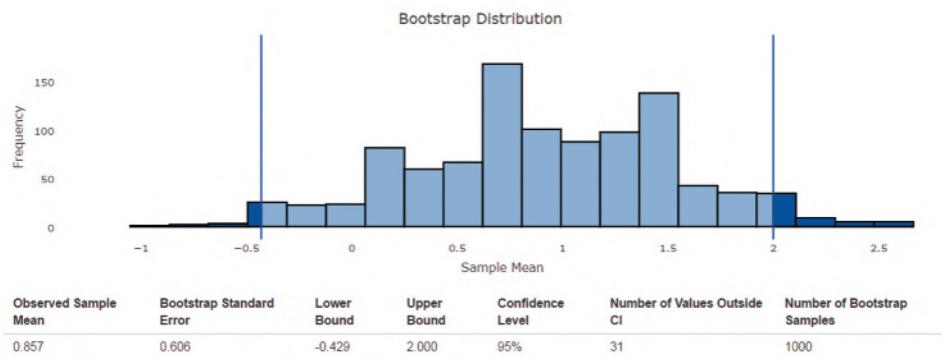
To construct a confidence interval, we begin by selecting bootstrap samples, with replacement, from the original sample of differences. Here is one bootstrap sample:

Resampled Chimp ID	5	5	5	5	1	3	6
Difference	1	1	1	1	2	-2	3

The sample mean for this bootstrap sample is $\bar{x}_d = 1.00$. This process is repeated many times, and the resulting bootstrap distribution of \bar{x}_d values provides information about sampling variability that can be used to find a confidence interval for the population mean difference, μ_d , where μ_d is the mean difference

$$\left(\begin{array}{c} \text{mean number of} \\ \text{charitable responses} \\ \text{when there is a chimp} \\ \text{in the adjoining cage} \end{array} \right) - \left(\begin{array}{c} \text{mean number of} \\ \text{charitable responses} \\ \text{when there is not a chimp} \\ \text{in the adjoining cage} \end{array} \right)$$

The Shiny app “Bootstrap Confidence Interval for One Mean” produces a bootstrap distribution that is used to obtain a bootstrap confidence interval for the mean difference, as shown here. This app can be found in the collection at statistics.cengage.com/PSO6e/Apps.html.



For this particular simulation, the bootstrap method produces a 95% confidence interval for the population mean difference, μ_d , of -0.43 to 2.00 . We can be 95% confident that this interval contains the population mean difference.

We can also perform a hypothesis test to test the claim that the population mean difference is different from zero.

To do this, the data values in the original sample can be shifted to represent a sample from a hypothetical population with mean $\mu_d = 0$ by subtracting 0.86 (the mean of the original sample of differences) from each difference, as follows:

Chimp	1	2	3	4	5	6	7
Difference	2	0	-2	2	1	3	0
Difference - 0.86	1.14	-0.86	-2.86	1.14	0.14	2.14	-0.86

Then the relevant hypotheses are

$$\begin{aligned} H_0: \mu_d &= 0 \\ H_a: \mu_d &\neq 0 \end{aligned}$$

The simulation approach begins with resampling from the original sample shifted to have a mean of 0. Here is one simulated sample from the shifted sample:

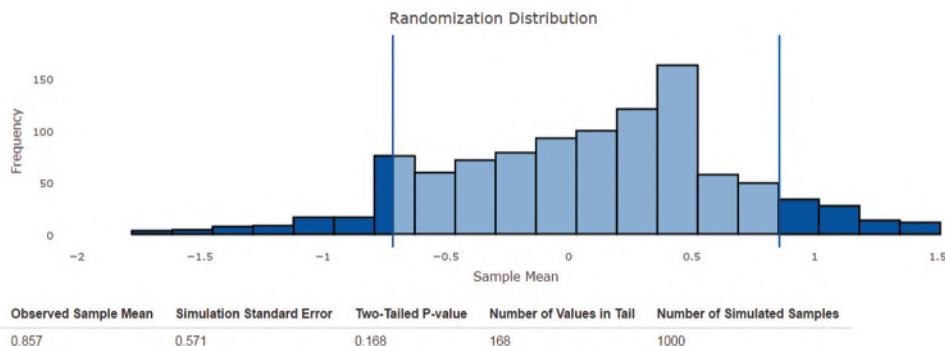
Chimp	4	4	2	1	7	3	4
Difference - 0.86	1.14	1.14	-0.86	1.14	-0.86	-2.86	1.14

The sample mean for this simulated sample is $\bar{x}_d = -0.02$.

A randomization distribution is generated by selecting many simulated samples from the shifted sample. This distribution can be viewed as the distribution of the sample mean difference if $\mu_d = 0$ is true. A P -value is the probability of obtaining a value at least as inconsistent with H_0 as what actually resulted. This means that the P -value for testing the null hypothesis $H_0: \mu_d = 0$ against the two-sided alternative $H_a: \mu_d \neq 0$ is the two times the probability of observing $\bar{x}_d \geq 0.86$ when H_0 is true. This probability can be approximated

using the proportion of simulated values of \bar{x}_d that fall at or above 0.86 in the distribution generated using the simulated samples.

The following randomization distribution was produced using the Shiny app “Randomization Test for One Mean.” This app can be found in the collection at statistics.cengage.com/PSO6e/Apps.html.



The two-tailed P -value based on this simulation is 0.168. This P -value is greater than the 0.05 significance level, and so we fail to reject H_0 . There is not convincing evidence that the population mean difference is different from zero.

To use the one-sample Shiny apps with paired samples, we need to first calculate the sample differences. There are also two apps (one for confidence intervals and one for hypothesis testing) available that permit entering the data from the two samples without first having to calculate the differences. These apps are “Bootstrap Confidence Interval for the Difference in Two Population Means Using Paired Samples” and “Randomization Test for Difference in Two Population Means Using Paired Samples.”

For example, the following output was produced by the Shiny app “Randomization Test for Difference in Two Population Means Using Paired Samples” using the data from Example 11.14.

Randomization Test for Difference in Two Population Means Using Paired Samples

Choose File to Upload:

Data entry windows below must be empty.

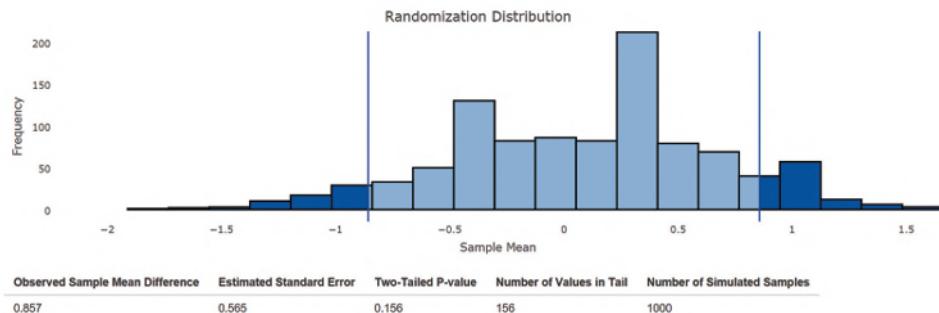
No file selected

Or Enter Data:

Data must be separated by a new line.

Sample 1 Sample 2

23	21
22	22
21	23
23	21
19	18
19	16
19	19



Notice that the P -value based on this randomization distribution is 0.156, compared to the P -value of 0.168 in Example 13.12. Because simulation-based methods such as randomization tests are based on random sampling, the randomization distribution will vary from simulation to simulation, depending on the outcomes of the sampling process. But usually the estimated P -values (or the bootstrap confidence interval endpoints) are similar from one simulation to another (as was the case in this example) as long as the number of samples used in the simulation is large.

EXERCISES 11.63 - 11.70

• Data set available online

- 11.63** • The Sheboygan (Wisconsin) Fire Department received a report on the potential effects of reductions in the number of firefighters it employs (["Study of Fire Department Causes Controversy," USA TODAY NETWORK-Wisconsin, December 22, 2016](#)).

In one section of the report, the average working heart rate percentage (the percentage value of the observed maximum heart rate during firefighting drills divided by an age-adjusted maximum heart rate for each firefighter) was reported for the driver of the first-arriving fire engine when only two firefighters (including the driver) were present, and for the driver of the first-arriving fire engine when more than two firefighters (up to five) were present.

The average working heart rate percentages were based on an earlier study, which included data from a sample of six drills using only two firefighters and from a sample of 18 drills using more than two firefighters. For purposes of this exercise, assume that these samples are representative of all drills with two firefighters and all drills with more than two firefighters.

The following data values are consistent with summary statistics given in the paper. Do these data support the claim that the mean average working heart rate percentage for the driver in a two-firefighter team is greater than the mean average working heart rate percentage for the driver in teams containing from three to five firefighters? Use a 0.05 significance level to carry out a randomization test of the given claim. You can make use of the Shiny apps in the collection at [statistics.cengage.com/PSO6e/Apps.html](#).

Working Heart Rate Percentage	
Two Firefighters	More than two Firefighters
97.7	73.6
87.7	63.7
89.5	72.0
97.3	83.1
81.9	54.2
79.7	73.1
	75.2
	84.5
	72.8
	70.0
	87.1
	81.1
	57.5
	68.4
	80.6
	57.8
	79.5
	75.0

- 11.64** Use the information given in the previous exercise to construct a 95% bootstrap confidence interval to estimate the difference in mean average working heart rates for the driver in teams of two firefighters and the driver in teams of from three to five firefighters. Interpret the interval in context. You can make use of the Shiny apps in the collection at [statistics.cengage.com/PSO6e/Apps.html](#).

- 11.65** • Studies have been conducted to evaluate the effectiveness of psilocybin mushrooms on improving the

quality of life for patients with cancer (“[A Dose of a Hallucinogen from a ‘Magic Mushroom,’ and Then Lasting Peace](#),” *The New York Times*, December 1, 2016). In one study, patients were randomly assigned to either a low-dose psilocybin treatment or to a high-dose treatment. One outcome that was measured was a “Personal Meaning” score, collected five weeks after the psilocybin treatment. Higher scores indicated greater “Personal Meaning.” The following data were estimated from a graphical display in the article. Do these data support the claim that the mean Personal Meaning score for patients with cancer taking a high dose of psilocybin is greater than the mean Personal Meaning score for patients with cancer taking a low dose? Use a randomization test to answer this question. You can make use of the Shiny apps in the collection at [statistics.cengage.com/PSO6e/Apps.html](#).

Personal Meaning		Personal Meaning	
Low Dose	High Dose	Low Dose	High Dose
3	1	1	7
2	5	2	7
6	7	4	8
4	7	4	8
4	5	2	8
2	5		7
3	5		6
7	6		4
7	6		7
6	7		7
6	6		7
7	7		6
5	5		

- 11.66** Use the information in the previous exercise to construct a 95% bootstrap confidence interval to estimate the difference in mean Personal Meaning scores for patients with cancer in the high-dose and low-dose psilocybin groups. Interpret the interval in context. You can make use of the Shiny apps in the collection at [statistics.cengage.com/PSO6e/Apps.html](#).

- 11.67** A new set of cognitive training modules called ONTRAC was developed to help children with attention deficit hyperactivity disorder (ADHD) to improve focus and to more easily dismiss distractions (“[Training Sensory Signal-to-Noise Resolution in Children with ADHD in a Global Mental Health Setting](#),” *Translational Psychiatry*, April 12, 2016, [nature.com/tp/journal/v6/n4/full/tp201645a.html](#), retrieved May 23, 2017). Eighteen children with ADHD were randomly assigned to one of two treatment groups. One group of 11 children received the ONTRAC treatment and another group of 7 children received a control treatment.

Values for 1-year improvement in ADHD Severity Score consistent with graphs and summary statistics in the research article appear in the following table.

ONTRAC	Control
11.0	4.3
8.8	3.8
12.1	8.5
10.9	4.7
8.0	1.5
8.2	7.7
11.9	8.0
11.1	
11.8	
10.8	
12.0	

- a. Explain why it is wise to be wary of using the two-sample t methods to analyze the data from this study.
 b. Do these data support the claim that the mean 1-year improvement in ADHD Severity Score for the ONTRAC treatment is different from the mean 1-year improvement in ADHD Severity Score for the control treatment? Use a randomization test with significance level 0.05 to answer this question. You can make use of the Shiny apps in the collection at [statistics.cengage.com/PSO6e/Apps.html](#).

- 11.68** Use the information in the previous exercise to construct a 95% bootstrap confidence interval to estimate the difference in mean 1-year improvement in ADHD Severity Score for the ONTRAC treatment and the control treatment. Interpret the interval in context. You can make use of the Shiny apps in the collection at [statistics.cengage.com/PSO6e/Apps.html](#).

- 11.69** New “closed loop” (CL) devices have been developed to help to suppress overactive brain activity in patients with conditions such as Parkinson’s disease and epilepsy (“[Conceptualization and Validation of an Open-Source Closed-Loop Deep Brain Stimulation System in Rats](#),” *Scientific Reports*, April 21, 2015, [nature.com/articles/srep09921](#), retrieved May 23, 2017). The CL device is implanted directly into a specific area of the brain, and one of the key advantages is that it can immediately apply a treatment in reaction to heightened brain activity. This may help to reduce the duration of seizures and other periods of uncontrolled movement in patients.

A study was conducted on the effectiveness of the CL device on brain activity in rats. First, each of seven rats was observed for 15 minutes with the CL device implanted but not activated (OFF). The percentage of the time that each rat was moving was recorded. Then, after each CL device was activated (CL), the percentage of the time that

the rat was moving was recorded over another 15-minute period.

The OFF and CL data values are approximated from a graph in the research article and appear in the table below, along with the calculated difference in movement, OFF – CL.

Rat ID	OFF	CL	Difference (OFF – CL)
1	98.8	56.6	42.2
2	56.9	25.5	31.4
3	56.7	40.0	16.7
4	58.1	68.1	-10.0
5	50.8	45.1	5.7
6	53.1	35.7	17.4
7	56.8	44.5	12.3

- a. Explain why two-sample t methods may not be appropriate in this context.

- b. For purposes of this exercise, assume that these rats are representative of rats in general. Do these data support the claim that the mean difference in movement, OFF – CL, is greater than zero? Carry out a randomization test to answer this question. You can make use of the Shiny apps in the collection at statistics.cengage.com/PSO6e/Apps.html.

- 11.70** Use the information given in the previous exercise to calculate a 95% bootstrap confidence interval to estimate the mean difference in movement, OFF – CL. Interpret the interval in context. You can make use of the Shiny apps in the collection at statistics.cengage.com/PSO6e/Apps.html.

SECTION 11.6 Simulation-Based Inference for Two Proportions (Optional)

The large-sample methods for estimating the difference between two proportions and for testing hypotheses about the difference between two proportions (Section 11.3) require independent random samples and sample sizes that are large enough to ensure that the sampling distribution of the differences in the sample proportions, $\hat{p}_1 - \hat{p}_2$, is approximately normal. When one or both sample sizes are not large enough, simulation-based methods can be used to estimate the difference in two proportions or to test hypotheses about the difference in two proportions.

Bootstrap Confidence Intervals for the Difference Between Two Population Proportions or Two Treatment Proportions

Example 11.15 Anti-Clotting Medications After Hip or Knee Surgery

In the study described in the article “[A Pilot Study Comparing Hospital Readmission Rates in Patients Receiving Rivaroxaban or Enoxaparin After Orthopedic Surgery](#),” (ptcommunity.com/system/files/pdf/ptj4106376.pdf, 2016, retrieved May 6, 2017), researchers identified patients who had received hip or knee replacement surgery and who had been given one of two types of medication to prevent blood clotting afterward. This was an observational study, conducted by reviewing the medical charts for patients who had received hip or knee replacement surgery. The outcome of interest was whether the patient was readmitted to the hospital within 30 days after leaving the hospital after surgery. The researchers were interested in estimating the difference in the proportions of patients readmitted for the two drugs (rivaroxaban and enoxaparin).

The study found that 8 of 213 patients who had been given rivaroxaban to prevent blood clots were readmitted to the hospital within 30 days after their surgeries, and that 1 out of 27 patients who had taken enoxaparin to prevent clotting was readmitted within 30 days. Based on the sample data, what can we learn about the difference in the proportion of patients who receive rivaroxaban who are readmitted and the corresponding proportion for patients who receive enoxaparin?

We want to estimate the difference in the proportion of patients readmitted to the hospital for those who receive rivaroxaban and the proportion readmitted for those who receive enoxaparin. If p_1 is the proportion of readmissions for patients who take rivaroxaban and p_2 is the proportion of readmissions for patients who take enoxaparin, then we will estimate the difference, $p_1 - p_2$.

Only 8 patients who received rivaroxaban were readmitted, and only 1 of the patients who received enoxaparin was readmitted. The sample sizes are too small to use the large-sample z confidence interval for a difference in population proportions. Instead, we can estimate the difference in population proportions, $\hat{p}_1 - \hat{p}_2$, using a bootstrap simulation-based method. For this example, a 95% confidence level will be used.

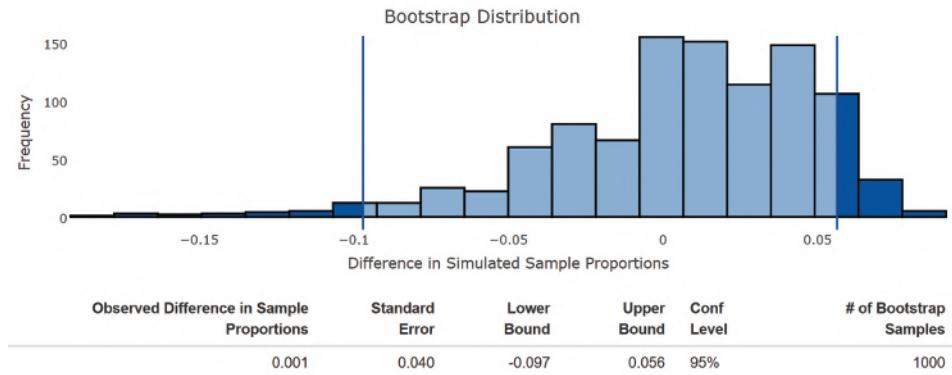
The sample of patient records was not randomly selected, but the researchers believed it to be representative of the population of patients undergoing hip and knee replacements who receive rivaroxaban and the population of patients who received enoxaparin. Based on the given information, we know

$$\begin{aligned} n_1 &= 213 & n_2 &= 27 \\ \hat{p}_1 &= 8/213 = 0.038 & \hat{p}_2 &= 1/27 = 0.037 \end{aligned}$$

To generate bootstrap simulated differences, $\hat{p}_1 - \hat{p}_2$, first consider a hypothetical population in which the proportion of success is 0.038 and take a sample of size 213 from this hypothetical population. This simulated sample produces a simulated value of \hat{p}_1 . Then consider a hypothetical population in which the proportion of success is 0.037 and take a sample of size 27. This simulated sample produces a simulated value of \hat{p}_2 . This pair of values, \hat{p}_1 and \hat{p}_2 , produces a bootstrap simulated value of $\hat{p}_1 - \hat{p}_2$.

For example, a simulated value of $\hat{p}_1 = 11/213 = 0.052$ was generated from a hypothetical population with population proportion of successes $p_1 = 0.038$. A simulated value of $\hat{p}_2 = 1/27 = 0.037$ was generated from a hypothetical population with population proportion of successes $p_2 = 0.037$. The simulated difference in sample proportions is $\hat{p}_1 - \hat{p}_2 = 0.052 - 0.037 = 0.015$.

A graph produced by the Shiny app “Bootstrap Confidence Interval for Difference in Two Proportions” of 1000 simulated differences in sample proportions is shown below. (This Shiny app can be found in the collection at statistics.cengage.com/PSO6e/Apps.html.)



The smallest 2.5% of the simulated differences in sample proportions were -0.097 or less, and the largest 2.5% of the simulated differences were 0.056 or greater. The 95% confidence interval based on the bootstrap resampling is $(-0.097, 0.056)$.

If the samples are representative of the two populations of interest (patients that take rivaroxaban and patients who take enoxaparin), we can be 95% confident that the actual difference in the proportion of patients readmitted to the hospital after taking rivaroxaban and the proportion readmitted after taking enoxaparin is between -0.097 and 0.056 . Zero is contained in this interval, indicating that it is plausible that there is no real difference in the readmission proportions for the two medications.

Randomization Tests for the Difference Between Two Population Proportions or Two Treatment Proportions

In some cases, we want to determine if two population proportions or two treatment proportions differ, or if one is smaller or larger than the other. In these situations, we would test hypotheses about the difference in the proportions.

Example 11.16 Oxytocin Nasal Spray and Social Interaction

Oxytocin is a synthetic hormone that may improve social interaction for young children with autism. An experiment was conducted to evaluate if the use of oxytocin delivered through a nasal spray improves social interaction experiences ([“The Effect of Oxytocin Nasal Spray on Social Interaction Deficits Observed in Young Children with Autism,” Molecular Psychology \[2016\]: 1225–1231.](#))

In many studies, participants drop out for one reason or another. Sometimes participants move to a new location, sometimes they change their minds about participating, and they always have the right to just discontinue participation in the study.

One concern in the oxytocin nasal spray and autism study was that young children with different diagnoses might drop out at different rates. In fact, 4 out of 23 young children with autism spectrum disorder dropped out of the study, and 4 out of 16 young children with pervasive developmental disorder dropped out. The researchers assumed that the young children in the study were representative of the population of all young children who might experience improved social interactions after taking oxytocin nasal spray. Do these data provide evidence that the proportion who drop out is different for the two different types of autism?

Here is a summary of the information in this example:

Diagnosis	Population Proportion	Sample Size	Sample Proportion
Autism spectrum disorder	p_1 = proportion of all young children with autism spectrum disorder who would drop out of the oxytocin study	$n_1 = 23$	$\hat{p}_1 = 0.174$
Pervasive developmental disorder	p_2 = proportion of all young children with pervasive developmental disorder who would drop out of the oxytocin study	$n_2 = 16$	$\hat{p}_2 = 0.250$

To answer the question of interest, we can test the null hypothesis that the two population proportions are equal. The population proportions to be compared are p_1 = proportion of all young children with autism spectrum disorder who would drop out of the oxytocin study and p_2 = proportion of all young children with pervasive developmental disorder who would drop out of the oxytocin study. The null hypothesis is $H_0: p_1 - p_2 = 0$. Initially, the researchers did not know which diagnosis group would have a larger dropout rate, and so the alternative hypothesis is two-sided, $H_a: p_1 - p_2 \neq 0$.

A large-sample z test for a difference in population proportions is not appropriate because the sample sizes are small. The number of young children who dropped out of the study was 4 in each of the two diagnosis groups. Because the sample size conditions for the two-sample z test are not met, rather than use a large-sample test, a simulation-based randomization test can be used. For this example, a significance level of $\alpha = 0.05$ will be used.

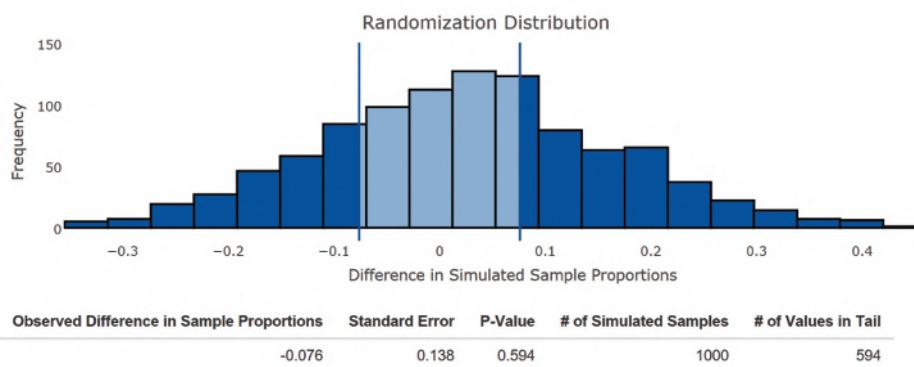
The observed difference in the sample proportions is $\hat{p}_1 - \hat{p}_2 = 0.174 - 0.250 = -0.076$. Recall that in the simulation-based test for a difference in two population means (Section 11.5), data from the two samples were combined into one group to represent the possibility that the two samples might have been taken from the same population. A similar process is used for the simulation-based test for the difference in two population proportions. The two samples are combined into one, and then random samples are selected with replacement from the combined samples (which represents the common population) to create a simulation-based distribution of the differences in sample proportions that is consistent with the null hypothesis of no difference in population or treatment proportions.

This resampling approach to a test for two proportions involves pooling the two original samples together into one hypothetical collection of $23 + 16 = 39$ young children in which $4 + 4 = 8$ are identified as dropping out of the study. The proportion of children who would drop out based on this hypothetical collection is $\hat{p} = 8/39 = 0.205$. This is known as the pooled sample proportion.

A sample of $n_1 = 23$ is drawn from the combined, or “pooled,” collection, with replacement, and a simulated value of \hat{p}_1 is calculated. A sample of $n_2 = 16$ is also drawn from the pooled collection, again with replacement, and a simulated value of \hat{p}_2 is calculated. These are the values needed to calculate a simulated difference $\hat{p}_1 - \hat{p}_2$. This process is repeated many times in order to construct the randomization distribution.

For example, suppose a sample of $n_1 = 23$ young children drawn from a hypothetical population with pooled proportion $\hat{p} = 0.205$ results in $\hat{p}_1 = 6/23 = 0.261$, and a separate sample of $n_2 = 16$ children drawn from the same population with pooled proportion $\hat{p} = 0.205$ results in $\hat{p}_2 = 4/16 = 0.250$. The simulated difference in the sample proportions is $\hat{p}_1 - \hat{p}_2 = 0.261 - 0.250 = 0.011$.

A randomization distribution for the difference in two proportions was produced using the Shiny app “Randomization Test for Two Proportions” and is shown below. (This Shiny app can be found in the collection at statistics.cengage.com/PSO6e/Apps.html.)



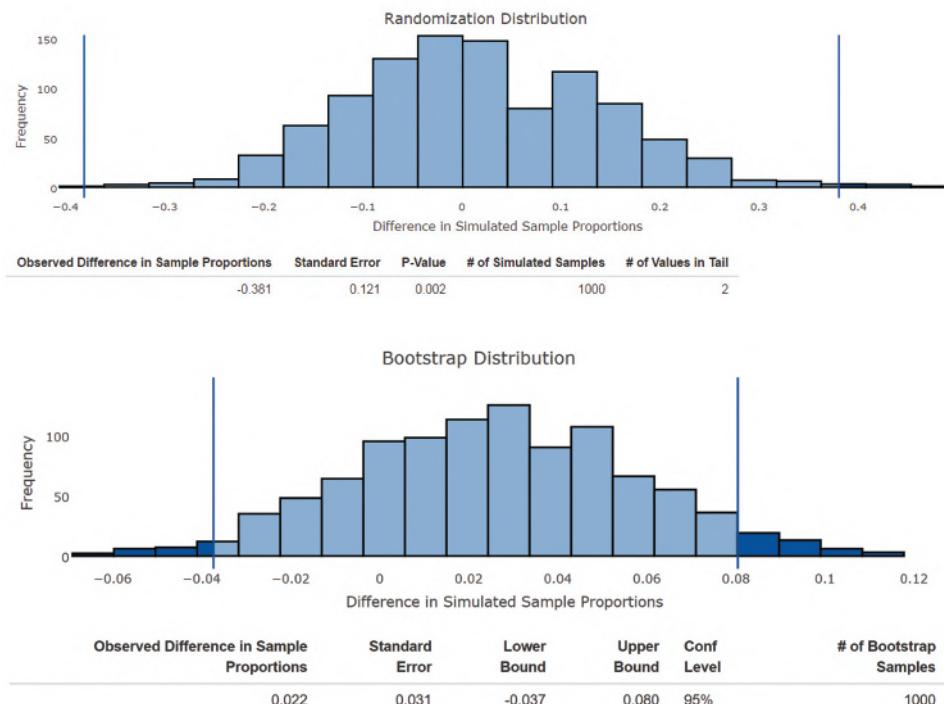
Because the alternative hypothesis in this example is two-sided, the P -value is the sum of the proportion of simulated differences in the two regions representing simulated differences that are as extreme or more extreme than the observed difference in sample proportions. In this example, the observed difference in sample proportions was -0.076 , so the P -value is the sum of the proportion of simulated differences that are less than -0.076 and the proportion of simulated differences that are greater than 0.076 . For the randomization distribution for this particular simulation, this is 0.594 . Because this P -value is greater than the specified significance level $\alpha = 0.05$, we fail to reject the null hypothesis.

Based on the sample data, there is not convincing evidence that there is a difference in dropout proportions for young children with a diagnosis of autism spectrum disorder and for children with a diagnosis of pervasive developmental disorder.

EXERCISES 11.71 - 11.78

- 11.71** The article “[Rapid Evolutionary Response to a Transmissible Cancer in Tasmanian Devils](#)” (nature.com/articles/ncomms12684, retrieved December 20, 2016) describes the spread of devil facial tumor disease (DFTD), which is a fatal form of cancer that swept through the Tasmanian devil population near the beginning of the 21st century. Researchers studied the genetic reaction of the Tasmanian devils by comparing the rates of occurrence of specific genetic markers of interest before and after DFTD swept across the island.

One region of Tasmania is called West Pencil Pine. Analysis of 21 tissue specimens taken from a representative sample of Tasmanian devils living in West Pencil Pine in 2006, before DFTD swept through, revealed that 5% had a specific genetic marker. Also analyzed were 42 tissue specimens from a representative sample of devils living in the same region in 2013 and 2014, after DFTD. In this sample, 43% had the same genetic marker. A significant and substantial change in these rates would indicate a remarkably fast evolution in the genetic code of the Tasmanian devils to protect against DFTD.

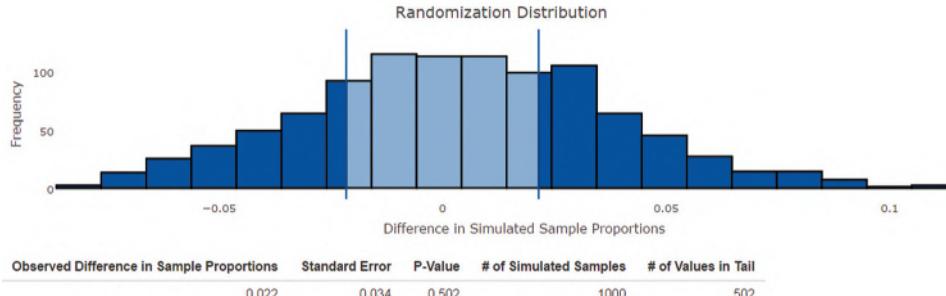
Output for Exercise 11.71

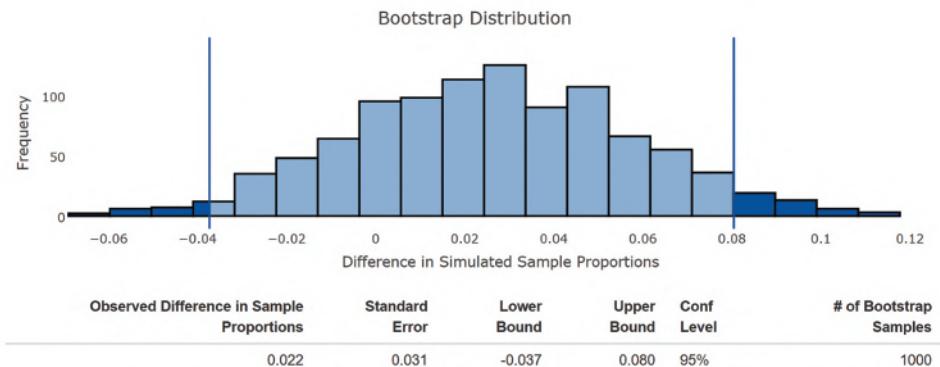
- Explain why the data from this study should not be analyzed using a large-sample hypothesis test for the difference in two population proportions.
- Use the output at the top of the page from the Shiny app “Randomization Test for Two Proportions” to carry out a hypothesis test to determine if there is convincing evidence that the proportion of Tasmanian devils with the genetic marker was greater after DFTD than before DFTD.
- Use the output from the Shiny app “Bootstrap Confidence Interval for Difference in Two Proportions” to identify a 95% confidence interval for the difference in the rates of occurrence of the specific genetic marker in the genes of Tasmanian devils, before and after DFTD. Interpret the confidence interval in context.

11.72 The report “[The 2016 Inside Higher Ed Survey of Faculty Attitudes on Technology](#)” ([insidehighered.com/booklet/2016-survey-faculty-attitudes-technology](#),

(retrieved December 14, 2016) describes the results of a survey of 1129 full-time college faculty and 293 part-time college faculty. Survey participants were asked if they require undergraduate students to submit papers through plagiarism-detection software; 40% of the full-time faculty and 38% of the part-time faculty said “yes.” Notice that the sample sizes for the two groups—full-time college faculty and part-time college faculty—are large enough to satisfy the conditions for a large-sample test and a large-sample confidence interval for a difference in two population proportions. Even though the sample sizes are large, simulation-based methods can still be used.

- Use the output at the bottom of the page from the Shiny app “Randomization Test for Two Proportions” to carry out a randomization test to determine if the proportion who require students to submit papers through plagiarism-detection software is different for full-time faculty and part-time faculty.

Output for Exercise 11.72(a)

Output for Exercise 11.72(b)

- b.** Use the output at the top of the page from the Shiny app “Bootstrap Confidence Interval for Difference in Two Proportions” to identify a 95% confidence interval for the difference in the population proportions of full-time college faculty and part-time college faculty who require students to submit papers through plagiarism-detection software. Interpret the confidence interval in context.

11.73 Researchers were interested in comparing regular-intensity exercise and high-intensity exercise for patients recovering from hospitalization due to chronic obstructive pulmonary disease (COPD). The researchers followed patients in Denmark who were enrolled in each of the two types of exercise programs ([“Increased Mortality in Patients with Severe COPD Associated with High-Intensity Exercise: A Preliminary Cohort Study,” Journal of Chronic Obstructive Pulmonary Disease \[2016\]: 2329–2334](#)). Each exercise program lasted for 8 weeks. The patients were followed for a total of 1.5 years. The researchers observed that 5 out of the 15 patients in the high-intensity group died within a year and a half, but that none of the 16 patients in the regular-intensity group died within a year and a half.

- a. Explain why the data from this study should not be analyzed using a large-sample hypothesis test for a difference in two population proportions.
- b. Carry out a hypothesis test to determine if there is convincing evidence of a difference in the population proportions who die within 1.5 years for the two exercise programs. Use the Shiny app “Randomization Test for Two Proportions” (in the collection at [statistics.cengage.com/PSO6e/Apps.html](#)) to report an approximate P -value and then use it to reach a decision in the hypothesis test. Remember to interpret the results of the test in context.
- c. Use the Shiny app “Bootstrap Confidence Interval for Difference in Two Proportions”

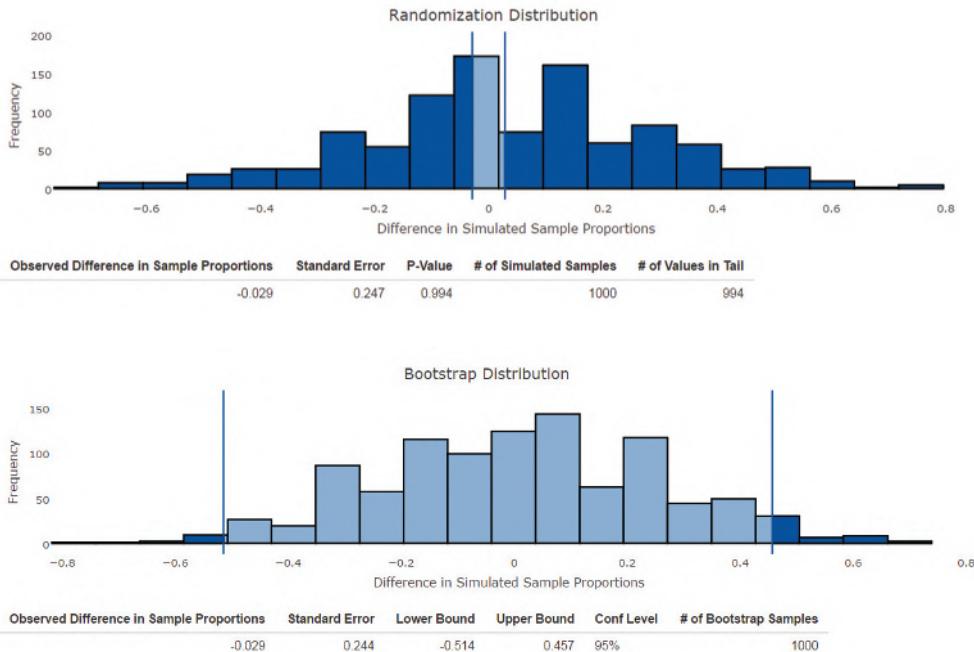
(in the collection at [statistics.cengage.com/PSO6e/Apps.html](#)) to obtain a 95% bootstrap confidence interval for the difference in the population proportions in patients who die within 1.5 years for the two exercise programs. Interpret the interval in the context of the research.

11.74 An article titled “[TCU Horned Frogs Game Preview \(Part 1\)](#)” ([uwdawgpound.com/2016/11/26/13710900/washington-huskies-tcu-horned-frogs-game-preview-part-1](#), retrieved December 20, 2016) previews a college basketball game between the University of Washington Huskies and the TCU Horned Frogs. The profile for TCU player Kenrich Williams states, “He has the statistical oddity of shooting better from behind the arc than from the free throw line but that’s a marker of small sample sizes.” (Note that “behind the arc” means three-point shots, which are taken behind an arc relatively far from the basket.)

At that point in the season, TCU had played five games. Kenrich Williams had made 4 of the 10 free throws that he had attempted (40%), and 3 of the 7 three-point shots that he had attempted (42.9%). Suppose that it is reasonable to consider these shots to be representative sample of Williams’s abilities to make free throws and three-point shots for the 2016–2017 college basketball season.

- a. Explain why the data in this exercise should not be analyzed using a large-sample hypothesis test for the difference in two population proportions.
- b. Use the output at the top of the next page from the Shiny app “Randomization Test for Two Proportions” to carry out a hypothesis test to determine if there is evidence that supports the statement that Williams is a better shooter from behind the arc than from the free throw line.
- c. Use the output to identify a 95% confidence interval for the difference in the proportions of made free throws and made three-point shots by Kenrich Williams for the 2016–2017 season. Interpret the confidence interval in context.

Output for Exercise 11.74

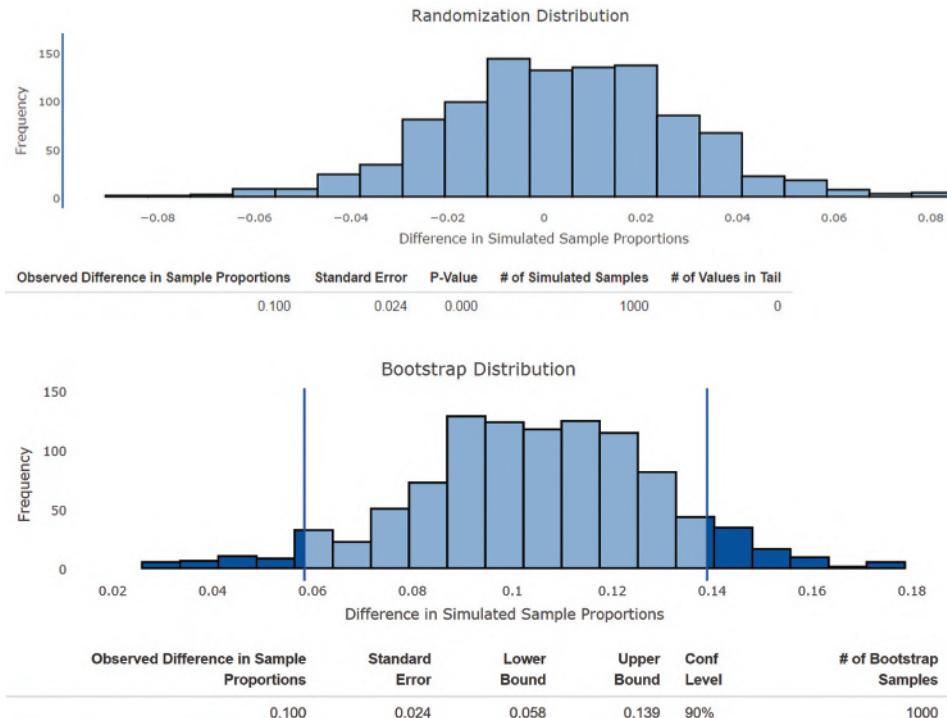


11.75 Example 11.11 describes a study in which 354 of 708 people in the sample of 18- to 29-year-olds and 412 of the 1029 people in the sample of 30- to 49-year-olds said that they thought it was OK to use a cell phone in a restaurant. Notice that the sample sizes for the two groups—people age 18 to 29 and those age 30 to 39—are large enough to satisfy the conditions for a large-sample test and

a large-sample confidence interval for the difference in two population proportions. Even though the sample sizes are large enough, simulation-based methods can be used.

- Use the following output from the Shiny app “Randomization Test for Two Proportions,” to carry out a randomization test to determine if there is convincing evidence that the proportion

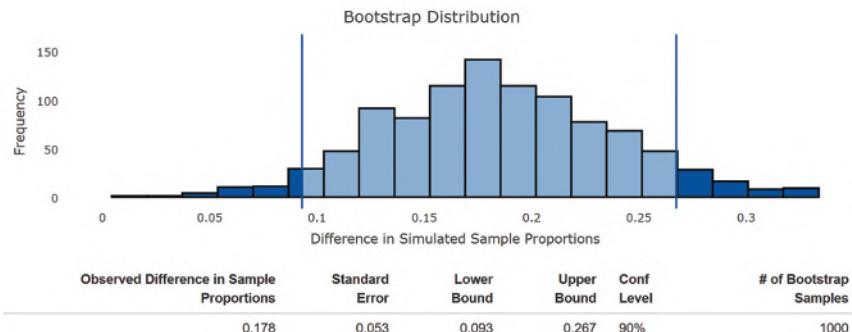
Output for Exercise 11.75



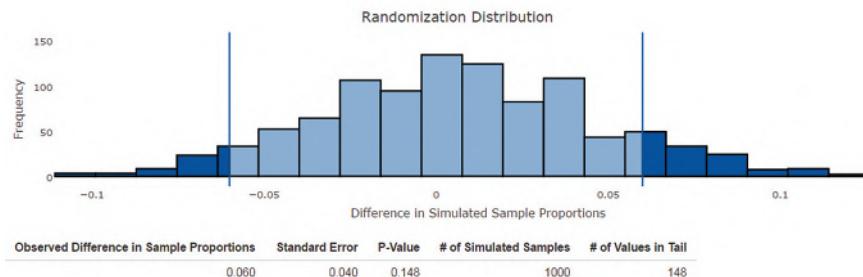
who think it is OK to use a cell phone at a restaurant is higher for the 18 to 29 age group than for the 30 to 49 age group.

- b. Use the output from the Shiny app “Bootstrap Confidence Interval for Difference in Two Proportions” to identify a 90% bootstrap confidence interval for the difference in the population proportions of people age 18 to 29 and those age 30 to 39 who said that they think that it is acceptable to use a cell phone in a restaurant.
 - c. Compare results in Part (b) to the confidence interval reported in Example 11.11 “Cell Phone Etiquette.” Would the interpretation change using the bootstrap confidence interval compared with the large-sample confidence interval? Explain.
- 11.76** A report in *USA TODAY* described an experiment to explore the accuracy of wearable devices designed to measure heart rate ([“Wearable Health Monitors Not Always Reliable, Study Shows,” USA TODAY, October 12, 2016](#)).
- The researchers found that when 50 volunteers wore an Apple Watch to track heart rate as they walked, jogged, and ran quickly on a treadmill for 3 minutes, the results were accurate compared with an EKG 91% of the time. When 50 volunteers wore a Fitbit Charge, the heart rate results were accurate 84% of the time.
- a. Explain why the data from this study should not be analyzed using a large-sample hypothesis test for a difference in two population proportions.
 - b. Carry out a hypothesis test to determine if there is convincing evidence that the proportion of accurate results for people wearing an Apple Watch is greater than this proportion for those wearing a Fitbit Charge. Use the Shiny app “Randomization Test for Two Proportions” (in the collection at [statistics.cengage.com/PSO6e/Apps.html](#)) to report an approximate *P*-value and use it to reach a decision in the hypothesis test. Remember to interpret the results of the test in context.
- c.** Use the Shiny app “Bootstrap Confidence Interval for Difference in Two Proportions” (in the collection at [statistics.cengage.com/PSO6e/Apps.html](#)) to obtain a 95% bootstrap confidence interval for the difference in the population proportions of accurate results for people wearing an Apple Watch and those wearing a Fitbit Charge. Interpret the interval in the context of the research.
- 11.77** As part of a study described in the report [“I Can’t Get My Work Done!” \(harmon.ie/blog/i-can’t-get-my-work-done-how-collaboration-social-tools-drain-productivity, 2011, retrieved May 6, 2017\)](#), people in a sample of 258 cell phone users age 20 to 39 were asked if they use their cell phones to stay connected while they are in bed, and 168 said “yes.” The same question was also asked of each person in a sample of 129 cell phone users age 40 to 49, and 61 said “yes.” We might expect the proportion who stay connected while in bed to be higher for the 20 to 39 age group than for the 40 to 49 age group, but how much higher?
- a. Construct and interpret a 90% large-sample confidence interval for the difference in the population proportions of cell phone users age 20 to 39 and those age 40 to 49 who say that they sleep with their cell phones. Interpret the confidence interval in context.
 - b. Note that the sample sizes in the two groups—cell phone users age 20 to 39 and those age 40 to 49—are large enough to satisfy the conditions for a large-sample test and a large-sample confidence interval for two population proportions. Even though the sample sizes are large enough, simulation-based methods can still be used. Use the output at the bottom of the page to identify a 90% bootstrap confidence interval for the difference in the population proportions of cell phone users age 20 to 39 and those age 40 to 49 who say that they sleep with their cell phones.

Output for Exercise 11.77



Output for Exercise 11.78



- c. Compare the confidence intervals calculated in Parts (a) and (b). Would the interpretation change using the bootstrap confidence interval compared with the large-sample confidence interval? Explain.

11.78 The article “**Americans Say No to Electric Cars Despite Gas Prices**” (*USA TODAY*, May 25, 2011) describes a survey of public opinion on issues related to rising gas prices. The survey was conducted by Gallup, a national polling organization. Each person in a representative sample of low-income adult Americans (annual income less than \$30,000) and each person in an independently selected representative sample of high-income adult Americans (annual income greater than \$75,000) was asked whether he or she would consider buying an electric car if gas prices continue to rise. In the low-income sample, 65% said that they would not buy an electric car no matter how high gas prices were to rise. In the high-income sample, 59% responded this way. The article did not give the sample sizes, but for the purposes of this exercise, suppose the sample sizes were both

300. One question of interest is whether the proportion who would never consider buying an electric car is different for the two income groups.

Note that the sample sizes in the two groups—low-income adult Americans and high-income adult Americans—are large enough to satisfy the conditions for a large-sample test and a large-sample confidence interval for two population proportions. Even though the sample sizes are large enough, we can still use simulation-based methods.

- a. Based on these data, is it reasonable to conclude that the proportions of low-income and high-income adults who would never consider buying an electric car differ? Use large-sample methods to test the appropriate hypotheses using a significance level of 0.05.
- b. Use the output at the top of the page to carry out a randomization test of the same hypotheses tested in Part (a).
- c. Compare the conclusions in Parts (a) and (b). Would the same conclusion be reached in either case? Explain.

CHAPTER ACTIVITIES

● Data set available online

ACTIVITY 11.1 HELIUM-FILLED FOOTBALLS?

Technology activity: Requires Internet access.

Background: Do you think that a football filled with helium will travel farther than a football filled with air? Two researchers at the Ohio State University investigated this question by performing an experiment in which 39 people each kicked a helium-filled football and an air-filled football. Half were assigned to kick the air-filled football first and then the helium-filled ball, whereas the other half kicked the helium-filled ball first followed by the air-filled ball. Distance (in yards) was measured for each kick.

In this activity, you will use data from this experiment (from the [Data and Story Library, lib.stat.cmu.edu/DASL](#), retrieved January, 2000) and then carry out a hypothesis

test to determine whether the mean distance is greater for helium-filled footballs than for air-filled footballs.

1. Do you think that helium-filled balls will tend to travel farther than air-filled balls when kicked? Before looking at the data, write a few sentences indicating what you think the outcome of this experiment was and describing the reasoning that supports your prediction.
2. ● Download the data set for Activity 11.1 from the student resource website or the WebAssign resources tab.
3. There are two samples in this data set. One consists of distances traveled for the 39 kicks of the air-filled football, and the other consists of the 39 distances for

the helium-filled football. Are these samples independent or paired? Explain.

- Carry out an appropriate hypothesis test to determine whether there is convincing evidence that the mean distance traveled is greater for a helium-filled football than for an air-filled football.

- Is the conclusion in the test of Step 4 consistent with your initial prediction of the outcome of this experiment? Explain.
- Write a paragraph for the sports section of your school newspaper describing this experiment and the conclusions that can be drawn from it.

ACTIVITY 11.2 THINKING ABOUT DATA COLLECTION

Background: In this activity you will design two experiments that would allow you to investigate whether people tend to have quicker reflexes when reacting with their dominant hand than with their nondominant hand.

- Working in a group, design an experiment to investigate the given research question that would result in independent samples. Be sure to describe how you plan to measure quickness of reflexes, what extraneous variables will be directly controlled, and the role that randomization plays in your design.
- How would you modify the design from Step 1 so that the resulting data are paired? Is the way in which

randomization is incorporated into the new design different from the way it is incorporated in the design from Step 1? Explain.

- Which of the two proposed designs would you recommend, and why?
- If assigned to do so by your instructor, carry out one of your experiments and analyze the resulting data. Write a brief report that describes the experimental design, includes both graphical and numerical summaries of the resulting data, and communicates the conclusions that follow from your data analysis.

ACTIVITY 11.3 A MEANINGFUL PARAGRAPH

Write a meaningful paragraph that includes the following six terms: **paired samples, significantly different, P-value, sample, population, alternative hypothesis.**

A “meaningful paragraph” is a coherent piece of writing in an appropriate context that uses all of the listed words. The paragraph should show that you understand the

meaning of the terms and their relationship to one another. A sequence of sentences that just define the terms is *not* a meaningful paragraph. When choosing a context, think carefully about the terms you need to use. Choosing a good context will make writing a meaningful paragraph easier.

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT
Independent samples	Two samples where the individuals or objects in the first sample are selected independently from those in the second sample.
Paired samples	Two samples for which each observation in one sample is paired in a meaningful way with a particular observation in a second sample.
$t = \frac{(\bar{x}_1 - \bar{x}_2) - \text{hypothesized value}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	The test statistic for testing $H_0: \mu_1 - \mu_2 = \text{hypothesized value}$ when the samples are independently selected and the sample sizes are large or it is reasonable to assume that both population distributions are normal.
$(\bar{x}_1 - \bar{x}_2) \pm (\text{t critical value}) \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	A formula for constructing a confidence interval for $\mu_1 - \mu_2$ when the samples are independently selected and the sample sizes are large or it is reasonable to assume that the population distributions are normal.
$\text{df} = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}}$ where $V_1 = \frac{s_1^2}{n_1}$ and $V_2 = \frac{s_2^2}{n_2}$	The formula for determining df for the two-sample t test and confidence interval.

TERM OR FORMULA	COMMENT
\bar{x}_d	The sample mean difference.
s_d	The standard deviation of the sample differences.
μ_d	The mean value for the population of differences.
σ_d	The standard deviation for the population of differences.
$t = \frac{\bar{x}_d - \text{hypothesized value}}{\frac{s_d}{\sqrt{n}}}$	The paired t test statistic for testing $H_0: \mu_d = \text{hypothesized value.}$
$\bar{x}_d \pm (t \text{ critical value}) \frac{s_d}{\sqrt{n}}$	The paired t confidence interval formula.
$\hat{p}_c = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$	\hat{p}_c is the statistic for estimating the common population proportion when $p_1 = p_2$.
$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_c(1 - \hat{p}_c)}{n_1} + \frac{\hat{p}_c(1 - \hat{p}_c)}{n_2}}}$	The test statistic for testing $H_0: p_1 - p_2 = 0$ when the samples are independently selected and both sample sizes are large.
$(\hat{p}_1 - \hat{p}_2) \pm (z \text{ critical value}) \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$	A formula for constructing a confidence interval for $p_1 - p_2$ when both sample sizes are large.

CHAPTER REVIEW Exercises 11.79 - 11.91

● Data set available online

- 11.79** ● A deficiency of the trace element selenium in the diet can negatively impact growth, immunity, muscle and neuromuscular function, and fertility. The introduction of selenium supplements to dairy cows is justified when pastures have low selenium levels. Authors of the paper “Effects of Short-Term Supplementation with Selenised Yeast on Milk Production and Composition of Lactating Cows” (*Australian Journal of Dairy Technology*, [2004]: 199–203) supplied the following data on milk selenium concentration (mg/L) for a sample of cows given a selenium supplement (the treatment group) and a control sample given no supplement, both initially and after a 9-day period.

Initial Measurement		After 9 Days	
Treatment	Control	Treatment	Control
11.4	9.1	138.3	9.3
9.6	8.7	104.0	8.8
10.1	9.7	96.4	8.8
8.5	10.8	89.0	10.1
10.3	10.9	88.0	9.6
10.6	10.6	103.8	8.6
11.8	10.1	147.3	10.4

(continued)

Initial Measurement		After 9 Days	
Treatment	Control	Treatment	Control
9.8	12.3	97.1	12.4
10.9	8.8	172.6	9.3
10.3	10.4	146.3	9.5
10.2	10.9	99.0	8.4
11.4	10.4	122.3	8.7
9.2	11.6	103.0	12.5
10.6	10.9	117.8	9.1
10.8		121.5	
8.2		93.0	

- Use the given data for the treatment group to determine if there is sufficient evidence to conclude that the mean selenium concentration is greater after 9 days of the selenium supplement.
- Are the data for the cows in the control group (no selenium supplement) consistent with the hypothesis of no significant change in mean selenium concentration over the 9-day period?
- Would you use the paired t test to determine if there was a significant difference in the initial mean selenium concentration for the control group and the treatment group? Explain why or why not.

- 11.80** The article “A ‘White’ Name Found to Help in Job Search” (*Associated Press*, January 15, 2003) described an experiment to investigate if it helps to have a “white-sounding” first name when looking for a job. Researchers sent 5000 resumes in response to ads that appeared in the *Boston Globe* and *Chicago Tribune*. The resumes were identical except that 2500 of them had “white-sounding” first names, such as Brett and Emily, whereas the other 2500 had “black-sounding” names such as Tamika and Rasheed. Resumes of the first type elicited 250 responses and resumes of the second type only 167 responses.

Do these data support the theory that the proportion receiving responses is greater for those resumes with “white-sounding first” names?

- 11.81** In a study of a proposed treatment for diabetes prevention, 339 people under the age of 20 who were thought to be at high risk of developing type I diabetes were assigned at random to two groups. One group received twice-daily injections of a low dose of insulin. The other group (the control) did not receive any insulin, but was closely monitored. Summary data (from the article “**Diabetes Theory Fails Test**,” *USA TODAY*, June 25, 2001) follow.

Group	n	Number Developing Diabetes
Insulin	169	25
Control	170	24

- a. Use the given data to construct a 90% confidence interval for the difference in the proportion that develop diabetes for the control group and the insulin group.
- b. Give an interpretation of the confidence interval and the associated confidence level.
- c. Based on your interval from Part (a), write a few sentences commenting on the effectiveness of the proposed prevention treatment.

- 11.82** When a surgeon repairs injuries, sutures (stitched knots) are used to hold together and stabilize the injured area. If these knots elongate and loosen through use, the injury may not heal properly because the tissues would not be optimally positioned. Researchers at the University of California, San Francisco, tied a series of different types of knots with two types of suture material, Maxon and Ticron.

Suppose that 112 tissue specimens were available and that for each specimen the type of knot and suture material were randomly assigned. The investigators tested the knots to see how much the loops elongated. The elongations (in mm) were measured and the resulting data are summarized here.

For purposes of this exercise, assume it is reasonable to regard the elongation distributions as approximately normal.

Types of knot	Maxon		
	n	\bar{x}	sd
Square (control)	10	10.0	0.1
Duncan Loop	15	11.0	0.3
Overhand	15	11.0	0.9
Roeder	10	13.5	0.1
Snyder	10	13.5	2.0

Types of knot	Ticron		
	n	\bar{x}	sd
Square (control)	10	2.5	0.06
Duncan Loop	11	10.9	0.40
Overhand	11	8.1	1.00
Roeder	10	5.0	0.04
Snyder	10	8.1	0.06

- a. Is there a significant difference in mean elongation between the square knot and the Duncan loop for Maxon thread?
- b. Is there a significant difference in mean elongation between the square knot and the Duncan loop for Ticron thread?
- c. For the Duncan loop knot, is there a significant difference in mean elongation between the Maxon and Ticron threads?

- 11.83** The article “**Trial Lawyers and Testosterone: Blue-Collar Talent in a White-Collar World**” (*Journal of Applied Social Psychology* [1998]: 84–94) compared trial lawyers and nontrial lawyers on the basis of mean testosterone level. Random samples of 35 male trial lawyers, 31 male nontrial lawyers, 13 female trial lawyers, and 18 female nontrial lawyers were selected for study. The article includes the following statement:

Trial lawyers had higher testosterone levels than did nontrial lawyers. This was true for men, $t(64) = 3.75$, $p < .001$, and for women, $t(29) = 2.26$, $p < .05$.

- a. Based on the information given, is the mean testosterone level for male trial lawyers significantly greater than for male nontrial lawyers?
- b. Based on the information given, is the mean testosterone level for female trial lawyers significantly greater than for female nontrial lawyers?
- c. Do you have enough information to carry out a test to determine whether there is a significant difference in the mean testosterone levels of male and female trial lawyers? If so, carry out such a

test. If not, what additional information would you need to be able to conduct the test?

- 11.84** In a study of memory recall, eight students from a large psychology class were selected at random and given 10 minutes to memorize a list of 20 non-sense words. Each was asked to list as many of the words as he or she could remember both 1 hour and 24 hours later. The data are as shown in the accompanying table.

Is there evidence to suggest that the mean number of words recalled after 1 hour exceeds the mean recall after 24 hours by more than 3? Use a level 0.01 test.

Subject	1	2	3	4	5	6	7	8
1 hour later	14	12	18	7	11	9	16	15
24 hour later	10	4	14	6	9	6	12	12

- 11.85** As part of a study to determine the effects of allowing the use of credit cards for alcohol purchases in Canada (“Changes in Alcohol Consumption Patterns Following the Introduction of Credit Cards in Ontario Liquor Stores,” *Journal of Studies on Alcohol* [1999]: 378–382), randomly selected individuals were given a questionnaire asking them how many drinks they had consumed during the previous week. A year later (after liquor stores started accepting credit cards for purchases), these same individuals were again asked how many drinks they had consumed in the previous week. The values in the accompanying table are consistent with summary values presented in the article.

	1994		1995		
	n	Mean	Mean	\bar{x}_d	s _d
Credit-Card Shoppers	96	6.72	6.34	0.38	5.52
Non-Credit-Card Shoppers	850	4.09	3.97	0.12	4.58

- The standard deviations of the differences were quite large. Explain how this could be the case.
- Calculate a 95% confidence interval for the mean difference in drink consumption for credit-card shoppers between 1994 and 1995. Is there evidence that the mean number of drinks decreased?
- Test the hypothesis that there was no change in the mean number of drinks between 1994 and 1995 for the non-credit-card shoppers. Be sure to calculate and interpret the P-value for this test.

- 11.86** The article “Truth and DARE: Tracking Drug Education to Graduation” (*Social Problems* [1994]: 448–456) compared the drug use of 288 randomly selected high school seniors exposed to a drug education program (DARE) and 335 randomly selected high school seniors who were not exposed to such

a program. Data for marijuana use are given in the accompanying table. Is there evidence that the proportion using marijuana is lower for students exposed to the DARE program? Use $\alpha = 0.05$.

	Number Who Use Marijuana	
n	Exposed to DARE	Not Exposed to DARE
Exposed to DARE	288	141
Not Exposed to DARE	335	181

- 11.87** The article “Softball Sliding Injuries” (*American Journal of Diseases of Children* [1988]: 715–716) provided a comparison of breakaway bases (designed to reduce injuries) and stationary bases. Consider the accompanying data (which agree with summary values given in the paper).

	Number of Games Played	Number of Games Where a Player Suffered a Sliding Injury
Stationary Bases	1,250	90
Breakaway Bases	1,250	20

- Is the proportion of games with a player suffering a sliding injury significantly lower for games using breakaway bases? Answer by performing a level 0.01 test.
- What did you have to assume in order for your conclusion to be valid? Do you think it is likely that this assumption was satisfied in this study?

- 11.88** Wayne Gretzky was one of ice hockey’s most prolific scorers when he played for the Edmonton Oilers. During his last season with the Oilers, Gretzky played in 41 games and missed 17 games due to injury. The article “The Great Gretzky” (*Chance* [1991]: 16–21) looked at the number of goals scored by the Oilers in games with and without Gretzky, as shown in the accompanying table. If we view the 41 games with Gretzky as a random sample of all Oiler games in which Gretzky played and the 17 games without Gretzky as a random sample of all Oiler games in which Gretzky did not play, is there evidence that the mean number of goals scored by the Oilers is greater for games in which Gretzky played? Use $\alpha = 0.01$.

	n	Sample Mean	Sample SD
Games with Gretzky	41	4.73	1.29
Games without Gretzky	17	3.88	1.18

- 11.89** Here’s one to sink your teeth into: The authors of the article “Analysis of Food Crushing Sounds During Mastication: Total Sound Level Studies”

(*Journal of Texture Studies* [1990]: 165–178) studied the nature of sounds generated during eating. Peak loudness (in decibels at 20 cm away) was measured for both open-mouth and closed-mouth chewing of potato chips and of tortilla chips. Forty subjects participated, with ten assigned at random to each combination of conditions (such as closed-mouth potato chip, and so on). We are not making this up! Summary values taken from plots given in the article appear in the accompanying table. For purposes of this exercise, suppose that it is reasonable to regard the peak loudness distributions as approximately normal.

	<i>n</i>	\bar{x}	<i>s</i>
Potato Chip			
Open mouth	10	63	13
Closed mouth	10	54	16
Tortilla Chip			
Open mouth	10	60	15
Closed mouth	10	53	16

- a.** Construct a 95% confidence interval for the difference in mean peak loudness between open-mouth and closed-mouth chewing of potato chips. Interpret the resulting interval.
- b.** For closed-mouth chewing (the recommended method!), is there sufficient evidence to indicate that there is a difference between potato chips and tortilla chips with respect to mean peak loudness? Test the relevant hypotheses using $\alpha = 0.01$.
- c.** The means and standard deviations given here were actually for stale chips. When ten measurements of peak loudness were recorded for closed-mouth chewing of fresh tortilla chips, the resulting mean and standard deviation were 56 and 14, respectively. Is there sufficient evidence to conclude that chewing fresh tortilla chips is louder than chewing stale chips? Use $\alpha = 0.05$.
- 11.90** Dentists make many people nervous. To see whether such nervousness elevates blood pressure, the blood

pressure and pulse rates of 60 subjects were measured in a dental setting and in a medical setting (“The Effect of the Dental Setting on Blood Pressure Measurement,” *American Journal of Public Health* [1983]: 1210–1214). For each subject, the difference (dental-setting blood pressure minus medical-setting blood pressure) was calculated. The analogous differences were also calculated for pulse rates. Summary data follow.

	Mean Difference	Standard Deviation of Differences
Systolic Blood Pressure	4.47	8.77
Pulse (beats/min)	−1.33	8.84

- a.** Do the data strongly suggest that true mean blood pressure is greater in a dental setting than in a medical setting? Use a level 0.01 test.
- b.** Is there sufficient evidence to indicate that true mean pulse rate in a dental setting differs from the true mean pulse rate in a medical setting? Use a significance level of 0.05.

- 11.91** • Both surface soil and subsoil specimens were taken from eight randomly selected agricultural locations in a particular county. The soil specimens were analyzed to determine both surface pH and subsoil pH, with the results shown in the accompanying table.

Location	1	2	3	4	5	6	7	8
Surface pH	6.55	5.98	5.59	6.17	5.92	6.18	6.43	5.68
Subsoil pH	6.78	6.14	5.80	5.91	6.10	6.01	6.18	5.88

- a.** Calculate a 90% confidence interval for the mean difference between surface and subsoil pH for agricultural land in this county.
- b.** What assumptions are necessary for the interval in Part (a) to be valid?

TECHNOLOGY NOTES

Two-sample *t* Test for $\mu_1 - \mu_2$

JMP

1. Input the raw data for both groups into the first column
2. Input the group information into the second column
3. Click **Analyze** and select **Fit Y by X**
4. Click and drag the first column's name from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the second column's name from the box under **Select Columns** to the box next to **X, Factor**

6. Click **OK**

7. Click the red arrow next to **Oneway Analysis of...** and select **t Test**

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **2-sample t...**
2. Click the radio button next to **Summarized data**

3. In the boxes next to **First**: For the first sample, type the value for n , the sample size in the box under **Sample size**; and type the value for the sample mean in the box under **Mean**; and finally type the value for the sample standard deviation in the box under **Standard deviation**:
4. In the boxes next to **Second**: For the second sample, type the value for n , the sample size in the box under **Sample size**; and type the value for the sample mean in the box under **Mean**; and finally type the value for the sample standard deviation in the box under **Standard deviation**:
5. Click **Options...**
6. Input the appropriate hypothesized value for the difference of the population means in the box next to **Test difference**:
7. Select the appropriate alternative from the drop-down menu next to **Alternative**:
8. Click **OK**
9. Click **OK**

Note: You may also run this test with the assumption of equal variances by clicking the checkbox next to **Assume equal variances** after Step 8 in the above sequence.

Raw data

1. Input the raw data into two separate columns
2. Click **Stat** then click **Basic Statistics** then click **2-sample t...**
3. Click the radio button next to **Samples in different columns**:
4. Click in the box next to **First**:
5. Double-click the column name where the first group's data is stored
6. Click in the box next to **Second**:
7. Double-click the column name where the second group's data is stored
8. Click **Options...**
9. Input the appropriate hypothesized value for the difference of the population means in the box next to **Test difference**:
10. Select the appropriate alternative from the drop-down menu next to **Alternative**:
11. Click **OK**
12. Click **OK**

Note: You may also run this test with the assumption of equal variances by clicking the checkbox next to **Assume equal variances** after Step 11 in the above sequence.

SPSS

1. Input the raw data for BOTH groups into the first column
2. Input the data for groups into the second column (input A for the first group and B for the second group)
3. Click **Analyze** then click **Compare Means** then click **Independent-Samples T Test**
4. Click the name of the column containing the raw data and click the arrow to move this variable to the **Test Variable(s)** box
5. Click the name of the column containing the group information and click the arrow to move this variable to the **Grouping Variable** box

6. Click the **Define Groups...** button
7. In the box next to **Group 1**: type A
8. In the box next to **Group 2**: type B
9. Click **Continue**
10. Click **OK**

Note: This procedure produces two-sample t tests under the assumption of equal variances AND also when equal variances are not assumed. It also outputs a two-tailed P -value.

Excel

1. Input the raw data for each group into two separate columns
2. Click on the **Data** ribbon
3. Click **Data Analysis** in the **Analysis** group

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.

If you are performing a test where you are assuming equal variances, continue with Steps 4–9 below. If you are performing a test where you are not assuming equation variances, skip to Step 10.

4. Select **t-Test: Two Samples Assuming Equal Variances** from the dialog box and click **OK**
5. Click in the box next to **Variable 1 Range**: and select the first column of data
6. Click in the box next to **Variable 2 Range** and select the second column of data (if you have used and selected column titles for BOTH variables, select the check box next to **Labels**)
7. Click in the box next to **Hypothesized Mean Difference** and type your hypothesized value (in general, this will be 0)
8. Click in the box next to **Alpha**: and type in the significance level
9. Click **OK**
10. Select **t-Test: Two Samples Assuming Unequal Variances** from the dialog box and click **OK**
11. Click in the box next to **Variable 1 Range**: and select the first column of data
12. Click in the box next to **Variable 2 Range** and select the second column of data (if you have used and selected column titles for BOTH variables, select the check box next to **Labels**)
13. Click in the box next to **Hypothesized Mean Difference** and type your hypothesized value (in general, this will be 0)
14. Click in the box next to **Alpha**: and type in the significance level
15. Click **OK**

Note: This procedure outputs P -values for both a one-sided and two-sided test.

TI-83/84

Summarized data

1. Press **STAT**
2. Highlight **TESTS**
3. Highlight **2-SampT-Test...**
4. Highlight **Stats** and press **ENTER**

5. Next to \bar{x}_1 input the value for the sample mean from the first sample
6. Next to sx_1 input the value for the sample standard deviation from the first sample
7. Next to n_1 input the value for the sample size from the first sample
8. Next to \bar{x}_2 input the value for the sample mean from the second sample
9. Next to sx_2 input the value for the sample standard deviation from the second sample
10. Next to n_2 input the value for the sample size from the second sample
11. Next to μ_1 highlight the appropriate alternative hypothesis and press **ENTER**
12. Highlight **Calculate** and press **ENTER**

Raw data

1. Enter the data into **L1** and **L2** (In order to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**
4. Highlight **2-SampT-Test...**
5. Highlight **Data** and press **ENTER**
6. Next to μ_1 highlight the appropriate alternative hypothesis and press **ENTER**
7. Highlight **Calculate** and press **ENTER**

TI-Nspire**Summarized data**

1. Enter the Calculate Scratchpad
2. Press the **menu** key and select **6:Statistics** then **7:Stat Tests** then **4:2-Sample t test...** and press **enter**
3. From the drop-down menu select **Stats**
4. Press **OK**
5. Next to \bar{x}_1 input the value for the sample mean for the first sample
6. Next to sx_1 input the value for the sample standard deviation for the first sample
7. Next to n_1 input the value for the sample size for the first sample
8. Next to \bar{x}_2 input the value for the sample mean for the second sample
9. Next to sx_2 input the value for the sample standard deviation for the second sample
10. Next to n_2 input the value for the sample size for the second sample
11. Next to **Alternate Hyp** select the appropriate alternative hypothesis from the drop-down menu
12. Press **OK**

Raw data

1. Enter the data into a data list (In order to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the list by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then **4:Stat Tests** then **4:2-Sample t test...** and press **enter**
3. From the drop-down menu select **Data**
4. Press **OK**
5. Next to **List 1** select the list containing your data from the first sample
6. Next to **List 2** select the list containing your data from the second sample
7. Next to **Alternate Hyp** select the appropriate alternative hypothesis from the drop-down menu
8. Press **OK**

Paired t Test for Difference of Population Means**JMP**

1. Enter the data for one sample in the first column
2. Enter the paired data from the second sample in the second column
3. Click **Analyze** and select **Matched Pairs**
4. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Paired Response**
5. Click and drag the second column name from the box under **Select Columns** to the box next to **Y, Paired Response**
6. Click **OK**

Minitab**Summarized data**

1. Click **Stat** then click **Basic Statistics** then click **Paired t...**
2. Click the radio button next to **Summarized data**
3. In the box next to **Sample size:** type the sample size, n
4. In the box next to **Mean:** type the sample mean for the DIFFERENCE of each pair of data values
5. In the box next to **Standard deviation:** type the sample standard deviation for the DIFFERENCE of each pair of data values
6. Click **Options...**
7. Input the appropriate hypothesized value for the difference of the paired population means in the box next to **Test mean:**
8. Select the appropriate alternative from the drop-down menu next to **Alternative:**
9. Click **OK**
10. Click **OK**

Raw data

1. Input the raw data into two separate columns
2. Click **Stat** then click **Basic Statistics** then click **Paired t...**
3. Click in the box next to **First:**
4. Double click the column name where the first group's data are stored
5. Click in the box next to **Second:**
6. Double click the column name where the second group's data are stored
7. Click **Options...**
8. Input the appropriate hypothesized value for the difference of the paired population means in the box next to **Test mean:**

9. Select the appropriate alternative from the drop-down menu next to **Alternative:**
10. Click **OK**
11. Click **OK**

SPSS

1. Input the data for each group into two separate columns
2. Click **Analyze** then click **Compare Means** then click **Paired-Samples T Test**
3. Select the first column and click the arrow to move it into the Pair 1 row, Variable 1 column
4. Select the second column and click the arrow to move it to the Pair 1 row, Variable 2 column
5. Click **OK**

Note: This procedure produces a two-sided P -value.

Excel

1. Input the raw data for each group into two separate columns
2. Click on the **Data** ribbon
3. Click **Data Analysis** in the **Analysis** group

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.

4. Select **t-Test: Paired Two Sample for Means** from the dialog box and click **OK**
5. Click in the box next to **Variable 1 Range:** and select the first column of data
6. Click in the box next to **Variable 2 Range** and select the second column of data (if you have used and selected column titles for BOTH variables, select the check box next to **Labels**)
7. Click in the box next to **Hypothesized Mean Difference** and type your hypothesized value (in general, this will be 0)
8. Click in the box next to **Alpha:** and type in the significance level
9. Click **OK**

Note: This procedure outputs P -values for both a one-sided and two-sided test.

TI-83/84

The TI-83/84 does not provide the option for a paired t test. However, one can be found by entering the difference data into a list and following the procedures in Chapter 12.

TI-Nspire

The TI-Nspire does not provide the option for a paired t test. However, one can be found by entering the difference data into a list and following the procedures in Chapter 12.

Confidence Interval for $\mu_1 - \mu_2$

1. Input the raw data for both groups into the first column
2. Input the group information into the second column

	Column 1	Column 2
1	1	A
2	2	A
3	3	A
4	4	A
5	5	A
6	6	A
7	7	A
8	8	A
9	9	A
10	2	A
11	3	A
12	5	B
13	4	B
14	6	B
15	7	B
16	8	B
17	4	B
18	5	B
19	3	B
20	7	B
	8	B

3. Click **Analyze** and select **Fit Y by X**
4. Click and drag the first column's name from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the second column's name from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Oneway Analysis of...** and select **t Test**

Note: The 95% confidence interval is automatically displayed. To change the confidence level, click the red arrow next to **Oneway Analysis of...** and select **Set α Level** then click the appropriate α -level or select **Other** and type the appropriate level.

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **2-sample t...**
2. Click the radio button next to **Summarized data**
3. In the boxes next to **First:** For the first sample, type the value for n , the sample size in the box under **Sample size:** and type the value for the sample mean in the box under **Mean:** and finally type the value for the sample standard deviation in the box under **Standard deviation:**
4. In the boxes next to **Second:** For the second sample, type the value for n , the sample size in the box under **Sample size:** and type the value for the sample mean in the box under **Mean:** and finally type the value for the sample standard deviation in the box under **Standard deviation:**
5. Click **Options...**
6. Input the appropriate confidence level in the box next to **Confidence Level**
7. Click **OK**
8. Click **OK**

Raw data

1. Input the raw data for each group into a separate column
2. Click **Stat** then click **Basic Statistics** then click **2-sample t...**
3. Click the radio button next to **Samples in different columns:**
4. Click in the box next to **First:**
5. Double click the column name where the first group's data are stored
6. Click in the box next to **Second:**
7. Double-click the column name where the second group's data are stored
8. Click **Options...**
9. Input the appropriate confidence level in the box next to **Confidence Level**
10. Click **OK**
11. Click **OK**

SPSS

1. Input the raw data for BOTH groups into the first column
2. Input the data for groups into the second column (input A for the first group and B for the second group)
3. Click **Analyze** then click **Compare Means** then click **Independent-Samples T Test**
4. Click the name of the column containing the raw data and click the arrow to move this variable to the **Test Variable(s):** box
5. Click the name of the column containing the group information and click the arrow to move this variable to the **Grouping Variable:** box
6. Click the **Define Groups...** button
7. In the box next to **Group 1:** type A
8. In the box next to **Group 2:** type B
9. Click **Continue**
10. Click **Options...**
11. Input the confidence level in the box next to **Confidence Interval Percentage:**
12. Click **Continue**
13. Click **OK**

Note: This procedure produces confidence intervals under the assumption of equal variances AND also when equal variances are not assumed.

Excel

Excel does not have the functionality to produce a confidence interval automatically for the difference of two population means. However, you can manually type the formulas for the lower and upper limits into two separate cells and have Excel calculate the results for you. You may also use Excel to find the *t* critical value based on the confidence level using the following steps:

1. Click in an empty cell
2. Click **Formulas**
3. Click **Insert Function**
4. Select **Statistical** from the drop-down box for the category
5. Select **TINV** and click **OK**
6. In the box next to **Probability** type in the value representing one minus your selected confidence level

7. In the box next to **Deg_freedom** type in the degrees of freedom ($n - 1$)
8. Click **OK**

TI-83/84**Summarized data**

1. Press **STAT**
2. Highlight **TESTS**
3. Highlight **2-SampTInt...** and press **ENTER**
4. Highlight **Stats** and press **ENTER**
5. Next to $\bar{x}1$ input the value for the sample mean from the first sample
6. Next to $sx1$ input the value for the sample standard deviation from the first sample
7. Next to $n1$ input the value for the sample size from the first sample
8. Next to $\bar{x}2$ input the value for the sample mean from the second sample
9. Next to $sx2$ input the value for the sample standard deviation from the second sample
10. Next to $n2$ input the value for the sample size from the second sample
11. Next to **C-Level** input the appropriate confidence level
12. Highlight **Calculate** and press **ENTER**

Raw data

1. Enter the data into **L1** and **L2** (In order to access lists press the **STAT** key, then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**
4. Highlight **2-SampTInt...** and press **ENTER**
5. Highlight **Data** and press **ENTER**
6. Next to **C-Level** input the appropriate confidence level
7. Highlight **Calculate** and press **ENTER**

TI-Nspire**Summarized data**

1. Enter the Calculate Scratchpad
2. Press the **menu** key and select **6:Statistics** then **6:Confidence Intervals** then **4:2-Sample t Interval...** and press **enter**
3. From the drop-down menu select **Stats**
4. Press **OK**
5. Next to $\bar{x}1$ input the value for the sample mean from the first sample
6. Next to $sx1$ input the value for the sample standard deviation from the first sample
7. Next to $n1$ input the value for the sample size from the first sample
8. Next to $\bar{x}2$ input the value for the sample mean from the second sample
9. Next to $sx2$ input the value for the sample standard deviation from the second sample
10. Next to $n2$ input the value for the sample size from the second sample
11. Next to **C Level** input the appropriate confidence level
12. Press **OK**

Raw data

- Enter the data into two separate data lists (In order to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the lists by selecting the top row of the column and typing a title.

- Press the **menu** key and select **4:Statistics** then **3:Confidence Intervals** then **2:t Interval...** and press **enter**
- From the drop-down menu select **Data**
- Press **OK**
- Next to **List 1** select the list containing the first data sample from the drop-down menu
- Next to **List 2** select the list containing the second data sample from the drop-down menu
- Next to **C-Level** input the appropriate confidence level
- Press **OK**

Confidence Interval for Paired Data**JMP**

- Enter the data for one group in the first column
- Enter the paired data from the second group in the second column

	Column 1	Column 2
1	1	5
2	4	4
3	6	6
4	4	7
5	7	8
6	8	4
7	6	5
8	9	3
9	2	7
10	3	8

- Click **Analyze** and select **Matched Pairs**
- Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Paired Response**
- Click and drag the second column name from the box under **Select Columns** to the box next to **Y, Paired Response**
- Click **OK**

Note: The 95% confidence interval is automatically displayed. To change the confidence level, click the red arrow next to **Oneway Analysis of...** and select **Set α Level** then click the appropriate α -level or select **Other** and type the appropriate level.

Minitab**Summarized data**

- Click **Stat** then click **Basic Statistics** then click **Paired t...**
- Click the radio button next to **Summarized data**
- In the box next to **Sample size:** type the sample size, n
- In the box next to **Mean:** type the sample mean for the DIFFERENCE of each pair of data values
- In the box next to **Standard deviation:** type the sample standard deviation for the DIFFERENCE of each pair of data values

- Click **Options...**

- Input the appropriate confidence level in the box next to **Confidence Level**
- Click **OK**
- Click **OK**

Raw data

- Input the raw data into two separate columns
- Click **Stat** then click **Basic Statistics** then click **Paired t...**
- Click in the box next to **First sample:**
- Double click the column name where the first group's data are stored
- Click in the box next to **Second sample:**
- Double click the column name where the second group's data are stored
- Click **Options...**
- Input the appropriate confidence level in the box next to **Confidence Level**
- Click **OK**
- Click **OK**

SPSS

- Input the data for each group into two separate columns
- Click **Analyze** then click **Compare Means** then click **Paired-Samples T Test**
- Select the first column and click the arrow to move it into the Pair 1 row, Variable 1 column
- Select the second column and click the arrow to move it to the Pair 1 row, Variable 2 column
- Click **Options...**
- Input the appropriate confidence level in the box next to **Confidence Interval Percentage:**
- Click **Continue**
- Click **OK**

Excel

Excel does not have the functionality to produce a confidence interval automatically for the difference of two population means. However, you can manually type the formulas for the lower and upper limits into two separate cells and have Excel calculate the results for you. You may also use Excel to find the t critical value based on the confidence level using the following steps:

- Click in an empty cell
- Click **Formulas**
- Click **Insert Function**
- Select **Statistical** from the drop-down box for the category
- Select **TINV** and click **OK**
- In the box next to **Probability** type in the value representing one minus your selected confidence level
- In the box next to **Deg_freedom** type in the degrees of freedom ($n - 1$)
- Click **OK**

TI-83/84

The TI-83/84 does not provide the option for a paired t confidence interval. However, one can be found by entering the difference data into a list and following the procedures in Chapter 12.

TI-Nspire

The TI-Nspire does not provide the option for a paired *t* confidence interval. However, one can be found by entering the difference data into a list and following the procedures in Chapter 12.

 z Test for $p_1 - p_2$ **JMP**

JMP does not have the functionality to automatically provide the results of a *z* test for the difference of two proportions.

Minitab**Summarized data**

1. Click **Stat** then click **Basic Statistics** then click **2 Proportions...**
2. Click the radio button next to **Summarized data**
3. In the boxes next to **First:** type the value for *n*, the total sample size in the box under the **Trials:** column and the number of successes in the box under **Events:**
4. In the boxes next to **Second:** type the value for *n*, the total sample size in the box under the **Trials:** column and the number of successes in the box under **Events:**
5. Click **Options...**
6. Input the appropriate hypothesized value in the box next to **Test difference:** (this is usually 0)
7. Check the box next to **Use pooled estimate of p for test**
8. Click **OK**
9. Click **OK**

Raw data

1. Input the raw data two separate columns
2. Click **Stat** then click **Basic Statistics** then click **2 Proportion...**
3. Select the radio button next to **Samples in different columns**
4. Click in the box next to **First:**
5. Double-click the column name where the first group's raw data is stored
6. Click in the box next to **Second:**
7. Double-click the column name where the second group's raw data are stored
8. Click **Options...**
9. Input the appropriate hypothesized value in the box next to **Test difference:** (this is usually 0)
10. Check the box next to **Use test and interval based on normal distribution**
11. Click **OK**
12. Click **OK**

SPSS

SPSS does not have the functionality to automatically produce a *z* test for the difference of two proportions.

Excel

Excel does not have the functionality to automatically produce a *z*-test for the difference of two proportions. However, you can type the formulas into a cell for the test statistic in order to have Excel calculate this for you. Then use the methods from Chapter 6 to find the *P*-value using the Normal distribution.

TI-83/84

1. Press the **STAT** key
2. Highlight **TESTS**
3. Highlight **2-PropZTest...** and press **ENTER**
4. Next to **x1** type the number of successes from the first sample
5. Next to **n1** type the sample size from the first sample
6. Next to **x2** type the number of successes from the second sample
7. Next to **n2** type the sample size from the second sample
8. Next to **p1**, highlight the appropriate alternative hypothesis
9. Highlight **Calculate** and press **ENTER**

TI-Nspire

1. Enter the Calculate Scratchpad
2. Press the **menu** key then select **6:Statistics** then select **7:Stat Tests** then **6:2-Prop z Test...** then press **enter**
3. In the box next to **Successes, x1** type the number of successes from the first sample
4. In the box next to **n1** type the number of trials from the first sample
5. In the box next to **Successes, x2** type the number of successes from the second sample
6. In the box next to **n2** type the sample size from the second sample
7. In the box next to **Alternate Hyp** choose the appropriate alternative hypothesis from the drop-down menu
8. Press **OK**

Confidence Interval for $p_1 - p_2$ **JMP****Summarized data**

1. Input the data into the JMP data table with categories for one variable in the first column, categories for the second variable in the second column, and counts for each combination in the third column

	Column 1	Column 2	Column 3
1	A	Yes	15
2	A	No	10
3	B	Yes	25
4	B	No	5

2. Click **Analyze** and select **Fit Y by X**
3. Click and drag the first column containing the response variable from the box under **Select Columns** to the box next to **X, Factor**
4. Click and drag the second column containing the group information from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the third column containing the counts for each combination from the box under **Select Columns** to the box next to **Freq**
6. Click **OK**
7. Click the red arrow next to **Contingency Analysis of...** and select **Two Sample Test for Proportions**

Note: You can change the response of interest (i.e., “Yes” instead of “No” or “Success” instead of “Failure”) by clicking the radio button at the bottom of the **Two Sample Test for Proportions** section.

Raw data

1. Input the raw data into two separate columns: one containing the response variable and one containing the group information
2. Click **Analyze** and select **Fit Y by X**

	Column 1	Column 2
1	Yes	A
2	No	A
3	Yes	A
4	Yes	A
5	Yes	A
6	Yes	A
7	No	A
8	Yes	A
9	Yes	A
10	No	A
11	Yes	B
12	No	B
13	No	B
14	No	B
15	No	B
16	No	B

3. Click and drag the first column containing the response variable from the box under **Select Columns** to the box next to **Y, Response**
4. Click and drag the second column containing the group information from the box under **Select Columns** to the box next to **X, Factor**
5. Click **OK**
6. Click the red arrow next to **Contingency Analysis of...** and select **Two Sample Test for Proportions**

Note: You can change the response of interest (i.e., “Yes” instead of “No” or “Success” instead of “Failure”) by clicking the radio button at the bottom of the **Two Sample Test for Proportions** section.

Minitab

Summarized data

1. Click **Stat** then click **Basic Statistics** then click **2 Proportions...**
2. Click the radio button next to **Summarized data**
3. In the boxes next to **First:** type the value for n , the total sample size in the box under the **Trials:** column and the number of successes in the box under **Events:**
4. In the boxes next to **Second:** type the value for n , the total sample size in the box under the **Trials:** column and the number of successes in the box under **Events:**
5. Click **Options...**
6. Input the appropriate confidence level in the box next to **Confidence Level**

7. Click **OK**

8. Click **OK**

Raw data

1. Input the raw data two separate columns
2. Click **Stat** then click **Basic Statistics** then click **2 Proportion...**
3. Select the radio button next to **Samples in different columns**
4. Click in the box next to **First:**
5. Double click the column name where the first group’s raw data is stored
6. Click in the box next to **Second:**
7. Double click the column name where the second group’s raw data is stored
8. Click **Options...**
9. Input the appropriate confidence level in the box next to **Confidence Level**
10. Click **OK**
11. Click **OK**

SPSS

SPSS does not have the functionality to automatically produce a confidence interval for the difference of two proportions.

Excel

Excel does not have the functionality to automatically produce a confidence interval for the difference of two proportions. However, you can type the formulas into two separate cells for the lower and upper limit to have Excel calculate these results for you.

TI-83/84

1. Press the **STAT** key
2. Highlight **TESTS**
3. Highlight **2-PropZInt...** and press **ENTER**
4. Next to **x1** type the number of successes from the first sample
5. Next to **n1** type the sample size from the first sample
6. Next to **x2** type the number of successes from the second sample
7. Next to **n2** type the sample size from the second sample
8. Next to **C-Level** type the appropriate confidence level
9. Highlight **Calculate** and press **ENTER**

TI-Nspire

1. Enter the Calculate Scratchpad
2. Press the **menu** key then select **6:Statistics** then select **6:Confidence Intervals** then **6:2-Prop z Interval...** then press enter
3. In the box next to **Successes**, **x1** type the number of successes from the first sample
4. In the box next to **n1** type the number of trials from the first sample
5. In the box next to **Successes**, **x2** type the number of successes from the second sample
6. In the box next to **n2** type the number of trials from the second sample
7. In the box next to **C Level** input the appropriate confidence level
8. Press **OK**

12

The Analysis of Categorical Data and Goodness-of-Fit Tests



LStockStudio/Shutterstock.com

It is often the case that information is collected on categorical variables, such as political affiliation, gender, or college major. As with numerical data, categorical data sets can be univariate (consisting of observations on a single categorical variable), bivariate (observations on two categorical variables), or even multivariate. In this chapter, we will first consider methods for analyzing univariate categorical data sets and then turn to methods appropriate for use with bivariate categorical data.

LEARNING OBJECTIVES

Students will understand:

- The differences between goodness-of-fit tests, tests for homogeneity, and tests of independence.

Students will be able to:

- Test hypotheses about the distribution of a categorical variable using a goodness-of-fit test and interpret the results in context.
- Test the hypothesis that the distribution of a categorical variable is the same for two or more populations or treatments using a test of homogeneity and interpret the results in context.
- Test hypotheses about association between two categorical variables using a test of independence and interpret the results in context.

SECTION 12.1 Chi-Square Tests for Univariate Data

Univariate categorical data sets arise in a variety of settings. If each student in a sample of 100 is classified according to whether he or she is enrolled full-time or part-time, data on a categorical variable with two categories result. Each registered voter in a sample of 100 selected from those registered in a particular city might be asked which of the five city council members he or she favors for mayor. This would yield observations on a categorical variable with five categories.

Univariate categorical data are conveniently summarized in a **one-way frequency table**. For example, many colleges now allow students to pay their tuition using a credit card ([“Credit Card Tuition Payment Survey 2014,” creditcards.com/credit-card-news/tuition-charge-fee-survey.php, retrieved May 27, 2017](https://www.creditcards.com/credit-card-news/tuition-charge-fee-survey.php)). Suppose that 100 randomly selected students at one of these colleges participated in a survey, with possible responses being definitely will use a credit card to pay tuition next year, probably will use a credit card, probably won’t use a credit card, and definitely won’t use a credit card. The first few observations might be

Probably will	Definitely will not	Probably will not
Probably will not	Definitely will	Definitely will not

Counting the number of observations of each type might then result in the following one-way table:

	Outcome			
	Definitely Will	Probably Will	Probably Will Not	Definitely Will Not
Frequency	14	12	24	50

For a categorical variable with k possible values (k different categories), sample data are summarized in a one-way frequency table consisting of k cells, which can be displayed either horizontally or vertically.

In this section, we consider testing hypotheses about the proportions of the population that fall into the possible categories. For example, a college that allows students to pay tuition by credit card might be interested in determining whether the four possible responses to the credit card question occur equally often. If this is indeed the case, the long-run proportion of responses falling into each of the four categories is $1/4$, or 0.25. The test procedure to be presented shortly would allow the college to decide whether it is plausible that all four category proportions are equal to 0.25.

Notation and Hypotheses

k = number of categories of a categorical variable

p_1 = population proportion for Category 1

p_2 = population proportion for Category 2

\vdots

p_k = population proportion for Category k

(Note: $p_1 + p_2 + \dots + p_k = 1$)

The hypotheses to be tested have the form

H_0 : p_1 = hypothesized proportion for Category 1

p_2 = hypothesized proportion for Category 2

\vdots

p_k = hypothesized proportion for Category k

H_a : H_0 is not true, so at least one of the population category proportions differs from the corresponding hypothesized value.

For the example using responses to the tuition survey, we could use

- p_1 = the proportion of all students who will definitely pay by credit card
- p_2 = the proportion of all students who will probably pay by credit card
- p_3 = the proportion of all students who will probably not pay by credit card

and

- p_4 = the proportion of all students who will definitely not pay by credit card

The null hypothesis of interest is then

$$H_0: p_1 = 0.25, p_2 = 0.25, p_3 = 0.25, p_4 = 0.25$$

A null hypothesis of the type just described can be tested by first selecting a random sample of size n and then classifying each sample response into one of the k possible categories. To decide whether the sample data are compatible with the null hypothesis, we compare the observed cell counts (frequencies) to the cell counts that would have been expected when the null hypothesis is true. The expected cell counts are

$$\text{Expected cell count for Category 1} = np_1$$

$$\text{Expected cell count for Category 2} = np_2$$

and so on. The expected cell counts when H_0 is true result from using the corresponding hypothesized proportions to calculate the expected counts.

Example 12.1 *Jeopardy!* Nerds

The article “[Memo to Alex Trebek: Your Viewers Are Nerds](http://www.yougov.com/news/2016/10/21/jeopardy-fans-are-nerds/)” (October 21, 2016, today.yougov.com/news/2016/10/21/jeopardy-fans-are-nerds/, retrieved May 26, 2017) reports that viewers of the popular game show *Jeopardy!* tend to label themselves as nerds and that *Jeopardy!* viewers are more educated than the average American. These conclusions were based on a survey of a representative sample of people who said that they had watched the show in the past year. Data on the highest level of education completed for the sample of *Jeopardy!* viewers that is consistent with percentages given in the article are given in the accompanying table. The article did not give the sample size, so the numbers in the table are based on assuming a sample size of 1000.



Amanda Edwards/Getty Images

Highest Level of Education Completed	Observed Frequency
Less than high school	10
High school	270
Some college	180
2-year degree	70
4-year college degree	300
Post-graduate study	170
Total	1,000

The article also indicated that for the general population in the United States, the percentage of the population falling into each of the education categories is as shown in the following table.

Highest Level of Education Completed	Percentage of U.S. Population
Less than high school	7%
High school	36%
Some college	22%
2-year degree	9%
4-year college degree	17%
Post-graduate study	9%
Total	100%

The given information can be used to investigate whether the claim that *Jeopardy!* viewers tend to be more educated than the general population. You can represent the education category proportions as

- p_1 = proportion of *Jeopardy!* viewers with a highest level of education completed of less than high school
- p_2 = proportion of *Jeopardy!* viewers with a highest level of education completed of high school
- p_3 = proportion of *Jeopardy!* viewers with a highest level of education completed of some college
- p_4 = proportion of *Jeopardy!* viewers with a highest level of education completed of 2-year degree
- p_5 = proportion of *Jeopardy!* viewers with a highest level of education completed of 4-year degree
- p_6 = proportion of *Jeopardy!* viewers with a highest level of education completed of post-graduate study

If *Jeopardy!* viewers are like the population in general in terms of education level, you would expect that the *Jeopardy!* viewer proportions would match those of the general population. This means that they would be

$$\begin{aligned}P_1 &= 0.07 \\P_2 &= 0.36 \\P_3 &= 0.22 \\P_4 &= 0.09 \\P_5 &= 0.17 \\P_6 &= 0.09\end{aligned}$$

The hypotheses of interest are then

$$\begin{aligned}H_0: p_1 &= 0.07, \quad p_2 = 0.36, \quad p_3 = 0.22, \quad p_4 = 0.09, \quad p_5 = 0.17, \quad p_6 = 0.09 \\H_a: H_0 &\text{ is not true.}\end{aligned}$$

There were a total of 1000 *Jeopardy!* viewers in the sample. If the null hypothesis is true, the expected counts for the first two categories are

$$\begin{aligned}\left(\begin{array}{l} \text{expected count for} \\ \text{less than high school} \end{array} \right) &= n \left(\begin{array}{l} \text{hypothesized proportion} \\ \text{for less than high school} \end{array} \right) = 1000(0.07) = 70 \\ \left(\begin{array}{l} \text{expected count for} \\ \text{high school} \end{array} \right) &= n \left(\begin{array}{l} \text{hypothesized proportion} \\ \text{for high school} \end{array} \right) = 1000(0.36) = 360\end{aligned}$$

Expected counts for the other four categories are calculated in a similar way. The observed and expected counts are given in the following table.

Highest Level of Education Completed	Observed Count	Expected Count
Less than high school	10	70
High school	270	360
Some college	180	220
2-year degree	70	90
4-year college degree	300	170
Post-graduate study	170	90
Total	1,000	1,000

Because the observed counts are based on a *sample* of *Jeopardy!* viewers, it would be surprising to see *exactly* 7% of the sample falling into the first education category, exactly 36% falling into the second category, and so on, even when H_0 is true. If the differences between the observed and expected cell counts can reasonably be attributed to sampling variability, the data are considered compatible with H_0 . On the other hand, if the differences between the observed and the expected cell counts are too large to be explained by

chance differences from one sample to another, H_0 should be rejected in favor of H_a . To make a decision, we need to assess how different the observed and expected counts are.

The goodness-of-fit statistic, denoted by X^2 , is a measure of the extent to which the observed counts differ from those expected when H_0 is true. (The Greek letter χ is often used in place of X . The symbol X^2 is referred to as the chi-square [χ^2] statistic. Using X^2 rather than χ^2 follows the convention of denoting sample statistics by Roman letters.)

For a sample of size n ,

$$\left(\frac{\text{expected cell}}{\text{count}} \right) = n \left(\frac{\text{hypothesized value of corresponding}}{\text{population proportion}} \right)$$

The **goodness-of-fit statistic**, X^2 , results from first calculating the quantity

$$\frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

for each cell.

The X^2 statistic is the sum of these quantities for all k cells:

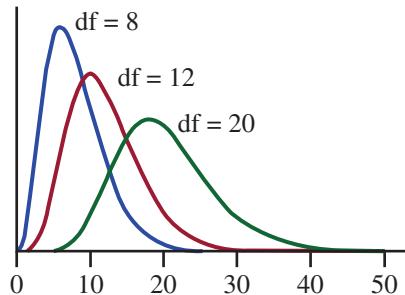
$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

The value of the X^2 statistic reflects the magnitude of the differences between observed and expected cell counts. When the differences are large, the value of X^2 tends to be large. This means that large values of X^2 suggest rejection of H_0 . A small value of X^2 (it can never be negative) occurs when the observed cell counts are quite similar to those expected when H_0 is true, and would be consistent with H_0 .

As with previous test procedures, a conclusion is reached by comparing a P -value to the significance level for the test. The P -value is calculated as the probability of observing a value of X^2 at least as large as the observed value when H_0 is true. This requires information about the sampling distribution of X^2 when H_0 is true.

When the null hypothesis is true and the sample size is large, the behavior of X^2 is described approximately by a **chi-square distribution**. A chi-square distribution curve has no area associated with negative values and is not symmetric, with a longer tail on the right. There are many chi-square distributions, and each one has a different number of degrees of freedom. Curves corresponding to several chi-square distributions are shown in Figure 12.1.

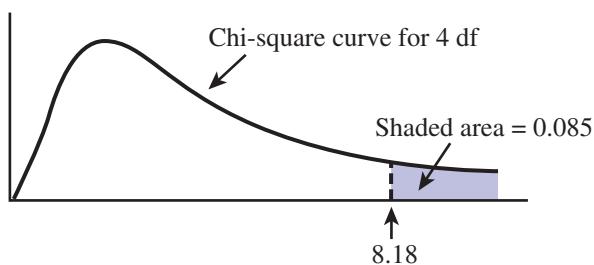
FIGURE 12.1
Chi-square curves.



For a test procedure based on the X^2 statistic, the associated P -value is the area under the appropriate chi-square curve and to the right of the calculated X^2 value. Appendix Table 8 gives upper-tail areas for chi-square distributions with up to 20 df.

FIGURE 12.2

A chi-square upper-tail area.



To find the area to the right of a particular X^2 value, locate the appropriate df column in Appendix Table 8. Determine which listed value is closest to the X^2 value of interest, and read the right-tail area corresponding to this value from the left-hand column of the table. For example, for a chi-square distribution with $df = 4$, the area to the right of $X^2 = 8.18$ is 0.085, as shown in Figure 12.2. For this same chi-square distribution ($df = 4$), the area to the right of 9.70 is approximately 0.045 (the area to the right of 9.74, the closest entry in the table for $df = 4$).

It is also possible to use a statistical software package or a graphing calculator to determine areas under a chi-square distribution curve.

Goodness-of-Fit Tests

When H_0 is true, the X^2 goodness-of-fit statistic has approximately a chi-square distribution with $df = (k - 1)$, as long as none of the expected cell counts are too small. When expected counts are small, and especially when an expected count is less than 1, the

value of $\frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$ can be inflated because it involves dividing by a small number. *The use of the chi-square distribution is appropriate when the sample size is large enough for every expected cell count to be at least 5.* If any of the expected cell frequencies are less than 5, categories can be combined in a sensible way to create acceptable expected cell counts. If you do this, remember to calculate the number of degrees of freedom based on the reduced number of categories.

Goodness-of-Fit Test

Hypotheses: $H_0: p_1 = \text{hypothesized proportion for Category 1}$

⋮

$p_k = \text{hypothesized proportion for Category } k$

$H_a: H_0$ is not true

Test Statistic:
$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

P-values: When H_0 is true and all expected counts are at least 5, X^2 has approximately a chi-square distribution with $df = k - 1$. The P -value associated with the calculated test statistic value is the area to the right of X^2 under the $df = k - 1$ chi-square curve. Upper-tail areas for chi-square distributions are found in Appendix Table 8.

Assumptions: 1. Observed cell counts are based on a *random sample*.

2. The *sample size is large*. The sample size is large enough for the chi-square test to be appropriate as long as every expected cell count is at least 5.

Example 12.2 *Jeopardy!* Viewers Revisited

You can use the *Jeopardy!* data in Example 12.1 to test the hypothesis that *Jeopardy!* viewers are like the population in general in terms of education level. A significance level of 0.05 will be used. The observed and expected counts calculated in Example 12.1 were

Highest Level of Education Completed	Observed Count	Expected Count
Less than high school	10	70
High school	270	360
Some college	180	220
2-year degree	70	90
4-year college degree	300	170
Post-graduate study	170	90
Total	1,000	1,000

Understand the context ➤

- Let p_1, p_2, p_3, p_4, p_5 , and p_6 denote the proportions falling into the six education categories as defined in Example 12.1.

If *Jeopardy!* viewers are like the population in terms of level of education, then you would expect the proportions falling into the education categories to match those for the general population. This leads to the null hypothesis given in Example 12.1:

- H_0 : $p_1 = 0.07, p_2 = 0.36, p_3 = 0.22, p_4 = 0.09, p_5 = 0.17, p_6 = 0.09$
- H_a : H_0 is not true. At least one of the education category proportions is different from the value specified in the null hypothesis.
- Significance level: $\alpha = 0.05$.

Formulate a plan ➤

- Test statistic: $X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$

Do the work ➤

- Assumptions: The expected cell counts (from Example 12.1) are all greater than 5. The viewers were a representative sample of *Jeopardy!* viewers.

- Calculation:

$$\begin{aligned} X^2 &= \frac{(10 - 70)^2}{70} + \frac{(270 - 360)^2}{360} + \frac{(180 - 220)^2}{220} + \frac{(70 - 90)^2}{90} + \frac{(300 - 170)^2}{170} \\ &\quad + \frac{(170 - 90)^2}{90} \\ &= 51.429 + 22.500 + 7.273 + 4.444 + 99.412 + 71.111 \\ &= 256.169 \end{aligned}$$

Interpret the result ➤

- P-value: The P-value is based on a chi-square distribution with $df = 6 - 1 = 5$. The calculated value of X^2 is greater than 20.51 (the largest entry in the $df = 5$ column of Appendix Table 8), so $P\text{-value} < 0.001$.

- Conclusion: Because the P-value is less than α , H_0 is rejected. There is convincing evidence that *Jeopardy!* viewers are not like the population in general with respect to education level. Notice that the observed counts are much greater than the expected counts for the 4-year degree and post-graduate study categories. This is consistent with the conclusion in the article that *Jeopardy!* viewers are “more educated than the average American.”

Statistical software can be used to perform a chi-square goodness-of-fit test. Minitab output for the data and hypothesized proportions of this example is shown here.

```
Chi-Square Goodness-of-Fit Test for Observed Counts in Variable:
Number of Viewers
Using category names in Education Level
```

Category	Observed	Proportion	Expected	Test to Chi-Sq	Contribution
Less than high school	10	0.07	70	51.4286	
High school	270	0.36	360	22.5000	

(continued)

Category	Observed	Proportion	Test		Contribution to Chi-Sq
			Expected		
Some college	180	0.22	220		7.2727
2-year degree	70	0.09	90		4.4444
4-year degree	300	0.17	170		99.4118
Post-graduate study	170	0.09	90		71.1111
		N	DF	Chi-Sq	P-Value
		1000	5	256.169	0.000

Example 12.3 Tasty Dog Food?

The article “[Can People Distinguish Pâté from Dog Food?” \(American Association of Wine Economists, April 2009, wine-economics.org\)](#) describes a study that investigated whether people can tell the difference between dog food, pâté (a spread made of finely chopped liver, meat, or fish), and processed meats (such as Spam and liverwurst).

Researchers used a food processor to make spreads that had the same texture and consistency as pâté from Newman’s Own dog food and from the processed meats. Each participant in the study tasted five spreads (duck liver pâté, Spam, dog food, pork liver pâté, and liverwurst). After tasting all five spreads, each participant was asked to choose the one that they thought was the dog food. The researchers recorded which of the five spreads was selected as the one the participant thought was dog food. Consider the data in the following table.

		Spread Chosen as Dog Food				
		Duck liver pâté	Spam	Dog food	Pork liver pâté	Liverwurst
Frequency		3	11	8	6	22

(Note: The frequencies in the table are consistent with summary values given in the paper. However, the sample size in the study was not actually 50.)

We can use the dog food taste data to test the hypothesis that the five different spreads (duck liver pâté, Spam, dog food, pork liver pâté, and liverwurst) are chosen equally often when people who have tasted all five spreads are asked to identify the one that they think is the dog food. If this is the case, each category proportion would be $1/5 = 0.20$. Because the sample size was 50, the expected count for the duck liver category is $50(0.20) = 10$. Because the other hypothesized category proportions are also 0.20, all of the expected counts are equal to 10.

Expected counts:

$$\begin{aligned} \text{Category 1: } np_1 &= 50(0.20) = 10 \\ \text{Category 2: } np_2 &= 50(0.20) = 10 \\ \text{Category 3: } np_3 &= 50(0.20) = 10 \\ \text{Category 4: } np_4 &= 50(0.20) = 10 \\ \text{Category 5: } np_5 &= 50(0.20) = 10 \end{aligned}$$

We are now ready to use the nine-step hypothesis testing procedure to test the hypotheses of interest.

Understand the context ➤

1. p_1 = proportion of all people who would choose duck liver pâté as the dog food
 p_2 = proportion of all people who would choose Spam as the dog food
 p_3 = proportion of all people who would choose dog food as the dog food
 p_4 = proportion of all people who would choose pork liver pâté as the dog food
 p_5 = proportion of all people who would choose liverwurst as the dog food

2. $H_0: p_1 = 0.20, p_2 = 0.20, p_3 = 0.20, p_4 = 0.20, p_5 = 0.20$
3. $H_a:$ At least one of the population proportions is not 0.20.
4. Significance level: $\alpha = 0.05$.

Formulate a plan ➤

5. Test statistic: $X^2 = \sum_{\text{all categories}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$

6. Assumptions: In order to use the chi-square goodness-of-fit test, we must be willing to assume that the participants in this study can be regarded as a random or representative sample. If this assumption is not reasonable, we should be very careful generalizing results from this analysis to any larger population. All expected counts are at least 5, so the sample size is large enough for the chi-square goodness-of-fit test to be appropriate.

Do the work ➤

7. Calculations: From Minitab

Chi-Square Goodness-of-Fit Test

Category	Observed	Proportion	Test	Contribution	to Chi-Sq
			Expected		
1	3	0.2	10	4.9	
2	11	0.2	10	0.1	
3	8	0.2	10	0.4	
4	6	0.2	10	1.6	
5	22	0.2	10	14.4	
		N	DF	Chi-Sq	P-Value
		50	4	21.4	0.000

Interpret the results ➤

8. P-value: From Minitab, the P-value is 0.000.

9. Conclusion: Based on this sample data, there is convincing evidence that the proportion identifying a spread as dog food is not the same for all five spreads. Here, it is interesting to note that the large differences between the observed counts and the counts that would have been expected if the null hypothesis of equal proportions was true are in the duck liver pâté and the liverwurst categories, indicating that fewer than expected chose the duck liver and many more than expected chose the liverwurst as the one they thought was dog food. So, although we reject the hypothesis that proportion choosing each of the five spreads is the same, it is not because people were actually able to identify which one was really dog food!

EXERCISES 12.1 - 12.13

- 12.1** A particular cell phone case is available in a choice of four different colors. A store sells all four colors. To test the hypothesis that sales are equally divided among the four colors, a random sample of 100 purchases is identified.

- If the resulting X^2 value were 6.4, what is the conclusion when using a test with significance level 0.05?
- What conclusion would be appropriate at significance level 0.01 if $X^2 = 15.3$?
- If there were six different covers rather than just four, what is the conclusion if $X^2 = 13.7$ and a test with $\alpha = 0.05$ was used?

- 12.2** From the given information in each case below, use Appendix Table 8 or technology to find the P-value for a chi-square test and give the conclusion for a significance level of $\alpha = 0.01$.

- $X^2 = 7.5$, df = 2
- $X^2 = 13.0$, df = 6
- $X^2 = 18.0$, df = 9
- $X^2 = 21.3$, df = 4
- $X^2 = 5.0$, df = 3

- 12.3** In 2014, the University of Houston carried out a study for the Texas Lottery Commission (“Demographic Survey of Texas Lottery Players”) that gives the age distribution for a representative sample of 375 Texas Lottery players.

Age Group	Frequency
18 to 24	7
25 to 34	32
35 to 44	34
45 to 54	60
55 to 64	106
65 and older	136
Total	375

Using data from the [U.S. Census Bureau \(census.gov\)](#) for 2014, the age distribution of adults in Texas was 14% between age 18 and 24, 20% between age 25 and 34, 19% between age 35 and 44, 18% between age 45 and 54, 14% between age 55 and 64, and 15% age 65 or older. Is it reasonable to conclude that one or more of the age groups buys a disproportionate share of Texas Lottery tickets? Use a chi-square goodness-of-fit test with $\alpha = 0.05$.

(Hint: See Example 12.2.)

- 12.4** The “[Global Automotive 2016 Color Popularity Report](#)” ([Axalta Coating Systems, axaltacs.com](#)) included data on the colors for a sample of new cars sold in North America. They reported that 25% of the cars in the sample were white, 21% were black, 16% were gray, 11% were silver, 10% were red, and 17% were some other color. Suppose that these percentages were based on a random sample of 1200 new cars sold in North America. Is there convincing evidence that the proportions of new cars sold are not the same for all six of the color categories?

- 12.5** A popular urban legend is that more babies than usual are born during certain phases of the lunar cycle, especially near the full moon. The paper “[The Effect of the Gravitation of the Moon on Frequency of Births](#)” ([Environmental Health Insights \[2010\]: 65–69](#)) classified a random sample of 1007 births at a large hospital in Japan according to lunar phase. In each lunar cycle (27.32 days), the moon moves 360° relative to the earth. To determine lunar phase, the researchers divided the 360° in one lunar cycle into 12 phases of 30° . The sample data are summarized in the accompanying frequency table.

Lunar Phase (degrees)	Number of Births
0–30	90
31–60	81
61–90	76
91–120	87
121–150	90
151–180	76
181–210	94
211–240	79

(continued)

Lunar Phase (degrees)	Number of Births
241–270	76
271–300	80
301–330	93
331–360	85

The researchers concluded that the frequency of births is not related to lunar cycle. Carry out a chi-square goodness-of-fit test to determine if the data are consistent with the researchers’ claim. Use a significance level of 0.05 for your test.

- 12.6** The authors of the paper “[External Factors and the Incidence of Severe Trauma: Time, Date, Season and Moon](#)” ([Injury \[2014\]: S93–S99](#)) classified admissions to hospitals in Germany according to season. They wondered if severe trauma injuries were more common in some seasons than others. Assume that there were 1200 trauma cases in the sample and that the sample is representative of severe trauma injuries in Germany. The data in the accompanying table are consistent with summary quantities given in the paper. Do these data support the theory that the proportion of severe trauma cases is not the same for the four seasons? Test the relevant hypotheses using a significance level of 0.05.

Season				
Winter	Spring	Summer	Fall	Total
228	332	352	288	1,200

- 12.7** The authors of the paper “[Is It Really About Me? Message Content in Social Awareness Streams](#)” ([Computer Supported Cooperative Work 2010](#)) studied a random sample of 350 Twitter users. For each Twitter user in the sample, the tweets sent during a particular time period were analyzed and the Twitter user was classified into one of the following categories based on the type of messages they usually sent:

Category	Description
IS	Information sharing
OC	Opinions and complaints
RT	Random thoughts
ME	Me now (what I am doing now)
O	Other

The accompanying table gives the observed counts for the five categories (approximate values read from a graph in the paper).

Twitter Type	IS	OC	RT	ME	O
Observed count	51	61	64	101	73

Carry out a hypothesis test to determine if there is convincing evidence that the proportions of Twitter users falling into each of the five categories are not all the same. Use a significance level of 0.05. (Hint: See Example 12.2.)

- 12.8** The article “*In Bronx, Hitting Home Runs Is A Breeze*” (*USA Today*, June 2, 2009) included a classification of 87 home runs hit at the new Yankee Stadium according to the direction that the ball was hit, resulting in the accompanying data.

Direction	Left Field	Left Center	Right Center	Right Field
Number of Home Runs				
Home Runs	18	10	7	18
				34

- a. Assume that it is reasonable to regard this sample of 87 home runs as representative of home runs hit at Yankee Stadium. Carry out a hypothesis test to determine if there is convincing evidence that the proportion of home runs hit is not the same for all five directions. (Hint: See Example 12.2.)
- b. Write a few sentences describing how the observed counts for the five directions differ from what would have been expected if the proportion of home runs is the same for all five directions.

- 12.9** The authors of the paper “*Racial Stereotypes in Children’s Television Commercials*” (*Journal of Advertising Research* [2008]: 80–93) counted the number of times that characters of different ethnicities appeared in commercials aired on Philadelphia television stations, resulting in the data in the accompanying table.

Ethnicity	African-American	Asian	Caucasian	Hispanic
Observed Frequency	57	11	330	6

Based on the 2000 Census, the proportion of the U.S. population falling into each of these four ethnic groups are 0.177 for African-American, 0.032 for Asian, 0.734 for Caucasian, and 0.057 for Hispanic. Do the data provide sufficient evidence to conclude that the proportions appearing in commercials are not the same as the census proportions? Test the relevant hypotheses using a significance level of 0.01. (Hint: See Example 12.2.)

- 12.10** The paper “*Sociochemosensory and Emotional Functions*” (*Psychological Science* [2009]: 1118–1124) describes an experiment to determine if college students can identify their roommates by smell. Forty-four female college students participated as subjects in the experiment. Each subject was presented with a

set of three t-shirts that were identical in appearance. Each of the three t-shirts had been slept in for at least 7 hours by a person who had not used any scented products (like scented deodorant, soap, or shampoo) for at least 48 hours prior to sleeping in the shirt.

One of the three shirts had been worn by the subject’s roommate. The subject was asked to identify the shirt worn by her roommate. This process was then repeated with another three shirts, and the number of times out of the two trials that the subject correctly identified the shirt worn by her roommate was recorded. The resulting data are summarized in the accompanying table.

Number of Correct Identifications	0	1	2
Observed Count	21	10	13

- a. Can a person identify her roommate by smell? If not, the data from the experiment should be consistent with what we would have expected to see if subjects were just guessing on each trial. That is, we would expect that the probability of selecting the correct shirt would be $1/3$ on each of the two trials.

Calculate the proportions of the time we would expect to see 0, 1, and 2 correct identifications if subjects are just guessing. (Hint: 0 correct identifications occurs if the first trial is incorrect *and* the second trial is incorrect.)

- b. Use the three proportions calculated in Part (a) to carry out a test to determine if the numbers of correct identifications by the students in this study are significantly different from what would have been expected by guessing. Use $\alpha = 0.05$. (Note: One of the expected counts is just a bit less than 5. For purposes of this exercise, assume that it is OK to proceed with a goodness-of-fit test.)

- 12.11** How would you answer the following question: Next Wednesday’s meeting has been moved forward two days. What day is the meeting now that it has been rescheduled?

This question is ambiguous as “moved forward” can be interpreted in two different ways. Did you answer Monday or Friday? The authors of the paper “*Even Abstract Motion Influences the Understanding of Time*” (*Metaphor and Symbol* [2011]: 260–271) wondered if the answers Monday and Friday would be provided an equal proportion of the time. A sample of students at Stanford University were asked this question, and the responses are summarized in the following table.

Response	Frequency
Monday	11
Friday	33

The authors of the paper used a chi-squared goodness-of-fit test to test the null hypothesis $H_0: p_1 = 0.50$, $p_2 = 0.50$, where p_1 is the proportion who would respond Monday and p_2 is the proportion who would respond Friday. They reported $X^2 = 11.00$ and $P\text{-value} < 0.001$. What conclusion can be drawn from this test?

- 12.12** *USA Today* (“**Hybrid Car Sales Rose 81% Last Year**,” April 25, 2005) reported the top five states for sales of hybrid cars in 2004 as California, Virginia, Washington, Florida, and Maryland. Suppose that each car in a sample of 2004 hybrid car sales is classified by state where the sale took place. Sales from states other than the top five were excluded from the sample, resulting in the accompanying table.

State	Observed Frequency
California	250
Virginia	56
Washington	34
Florida	33
Maryland	33
Total	406

(The given observed counts are artificial, but they are consistent with hybrid sales figures given in the article.)

The 2004 population estimates from the Census Bureau website are given in the accompanying table.

The population proportion for each state was calculated by dividing each state population by the total population for all five states. Use the X^2 goodness-of-fit test and a significance level of $\alpha = 0.01$ to test the hypothesis that hybrid sales for these five states are proportional to the 2004 population for these states.

State	2004 Population	Population Proportion
California	35,842,038	0.495
Virginia	7,481,332	0.103
Washington	6,207,046	0.085
Florida	17,385,430	0.240
Maryland	5,561,332	0.077
Total	72,477,178	

- 12.13** A certain genetic characteristic of a particular plant can appear in one of three forms (phenotypes). A researcher has developed a theory, according to which the hypothesized proportions are $p_1 = 0.25$, $p_2 = 0.50$, and $p_3 = 0.25$. A random sample of 200 plants yields $X^2 = 4.63$.
- Carry out a test of the null hypothesis that the theory is correct, using level of significance $\alpha = 0.05$.
 - Suppose that a random sample of 300 plants had resulted in the same value of X^2 . How would the analysis and conclusion differ from those in Part (a)?

SECTION 12.2 Tests for Homogeneity and Independence in a Two-way Table

Data resulting from observations made on two different categorical variables can also be summarized in a table. As an example, suppose that residents of a particular city can watch national news on ABC, CBS, NBC, or PBS. A researcher wishes to know whether there is any relationship between political philosophy (liberal, moderate, or conservative) and preferred news network among those residents who regularly watch the national news. Let x denote the variable *political philosophy* and y the variable *preferred network*. A random sample of 300 regular watchers is selected, and each individual is asked for his or her x and y values (political philosophy and preferred network). The data set is bivariate and might initially be displayed as follows:

Observation	x Value	y Value
1	Liberal	CBS
2	Conservative	ABC
3	Conservative	PBS
:	:	:
299	Moderate	NBC
300	Liberal	PBS

Bivariate categorical data of this sort can most easily be summarized by constructing a **two-way frequency table**, or **contingency table**. This is a rectangular table that consists

of a row for each possible value of x (each category for this variable) and a column for each possible value of y . There is a cell in the table for each possible (x, y) combination.

Once such a table has been constructed, the number of times each particular (x, y) combination occurs in the data set is determined, and these numbers (frequencies) are entered in the corresponding cells of the table. The resulting numbers are called **observed cell counts**. The table for the example relating political philosophy to preferred network contains 3 rows and 4 columns (because x has 3 possible values and y has 4 possible values). Table 12.1 is one possible table.

TABLE 12.1 An Example of a 3×4 Frequency Table

	ABC	CBS	NBC	PBS	Row Marginal Total
Liberal	20	20	25	15	80
Moderate	45	35	50	20	150
Conservative	15	40	10	5	70
Column Marginal Total	80	95	85	40	300

Marginal totals are obtained by adding the observed cell counts in each row and also in each column of the table. The row and column marginal totals, along with the total of all observed cell counts in the table—the **grand total**—have been included in Table 12.1.

The marginal totals provide information about the distribution of observed values for each variable separately. In this example, the row marginal totals reveal that the sample consisted of 80 liberals, 150 moderates, and 70 conservatives. Similarly, column marginal totals indicate how often each of the preferred network categories occurred: 80 preferred ABC, 95 preferred CBS, and so on. The grand total, 300, is the number of observations in the bivariate data set.

Two-way frequency tables are often characterized by the number of rows and columns in the table (specified in that order: rows first, then columns). Table 12.1 is called a 3×4 table. The smallest two-way frequency table is a 2×2 table, which has only two rows and two columns, resulting in four cells.

Two-way tables arise naturally in two different types of investigations. A researcher may be interested in comparing two or more populations or treatments on the basis of a single categorical variable and so might obtain independent samples from each population or treatment. For example, data could be collected at a university to compare students, faculty, and staff on the basis of primary mode of transportation to campus (car, bicycle, motorcycle, bus, or by foot). One random sample of 200 students, another random sample of 100 faculty members, and a third random sample of 150 staff members might be selected, and the individuals in each sample could be interviewed to obtain the necessary transportation information.

Data from such a study could be summarized in a 3×5 two-way frequency table with row categories of student, faculty, and staff and column categories corresponding to the five possible modes of transportation. The observed cell counts could then be used to learn about differences and similarities among the three groups with respect to mode of transportation. This type of bivariate categorical data set is characterized by having one set of marginal totals predetermined (the sample sizes for the different groups). In the 3×5 situation just discussed, the row totals would be fixed at 200, 100, and 150.

A two-way table also arises when the values of two different categorical variables are observed for all individuals or items in a single sample. For example, a sample of 500 registered voters might be selected. Each voter could then be asked both if he or she favored a particular property tax initiative and if he or she was a registered Democrat, Republican, or Independent. This would result in a bivariate data set with x representing the variable *political affiliation* (with categories Democrat, Republican, and Independent) and y representing the variable *response* (favors initiative or opposes initiative). The corresponding 3×2 frequency table could then be used to investigate

any association between position on the tax initiative and political affiliation. This type of bivariate categorical data set is characterized by having only the grand total predetermined (by the sample size).

Comparing Two or More Populations or Treatments: A Test of Homogeneity

When the value of a categorical variable is recorded for members of independent random samples obtained from each population or treatment under study, the question of interest is whether the category proportions are the same for all the populations or treatments. As in Section 12.1, the test procedure uses a chi-square statistic that compares the observed counts to those that would be expected if there were no differences.

Example 12.4 Heart Attacks in High-Rise Buildings Revised

Understand the context ➤



The paper “Out-of-Hospital Cardiac Arrest in High-Rise Buildings: Delays to Patient Care and Effect on Survival” (*Canadian Medical Association Journal* [2016]: 413–419) compared heart attack survival rates for people who lived in a house or townhouse, people who lived on the first or second floor of an apartment building, and people who lived on the third or a higher floor in an apartment building. Table 12.2, a 3×4 two-way frequency table, is the result of classifying each heart attack victim in independently selected representative samples of 5531 heart attacks that occurred in a house or townhouse, 667 heart attacks that occurred in an apartment building on the first or second floor, and 1696 heart attacks that occurred in an apartment building on the third or higher floor into one of two categories (survived and did not survive).

TABLE 12.2 Observed Counts for Example 12.4

	Survived	Did Not Survive	Row Marginal Total
House or Townhouse	217	5,314	5,531
Apartment First or Second Floor	35	632	667
Apartment Third or Higher Floor	46	1,650	1,696
Column Marginal Total	298	7,596	7,894

Notice that there are a total of 7894 heart attack victims in the three samples combined. Of these, 298 survived. The proportion of the total who survived is then

$$\frac{298}{7894} = 0.038$$

If there is no difference in survival for the three different groups, you would expect about 3.8% of the heart attack victims who live in a house or townhouse to survive, about 3.8% of the heart attack victims who live in an apartment on the first or second floor to survive, and about 3.8% of the heart attack victims who live in an apartment on the third or a higher floor to survive. This means that if there is no difference in survival, the expected number surviving for each of the three cells in the “survived” column are

$$\text{Expected count for house or townhouse and survive} = 0.038(5531) = 210.178$$

$$\text{Expected count for apartment on first or second floor and survive} = 0.038(667) = 23.346$$

$$\text{Expected count for apartment on third or higher floor and survive} = 0.038(1696) = 64.448$$

Notice that the expected cell counts do not need to be whole numbers.

The expected cell counts for the remaining cells can be calculated in a similar manner. The proportion of the total who did not survive is

$$\frac{7596}{7894} = 0.962$$

Then

$$\begin{aligned}\text{Expected count for house or townhouse and did not survive} &= 0.962(5531) = 5320.822 \\ \text{Expected count for apartment on first or second floor and did not survive} &= \\ &\quad 0.962(667) = 641.654 \\ \text{Expected count for apartment on third or higher floor and did not survive} &= \\ &\quad 0.962(1696) = 1631.552\end{aligned}$$

It is common practice to display the observed cell counts and the corresponding expected cell counts in the same table, with the expected cell counts enclosed in parentheses. Table 12.3 gives the observed cell counts and the expected cell counts. Notice that each marginal total for the expected cell counts is equal to the corresponding marginal total for the observed cell counts. (Sometimes there will be small differences in the marginal totals due to rounding when calculating the expected cell counts.)

TABLE 12.3 Observed and Expected Counts for Example 12.4

	Survived	Did Not Survive	Row Marginal Total
House or Townhouse	217 (210.178)	5,314 (5,320.822)	5,531
Apartment First or Second Floor	35 (23.346)	632 (641.654)	667
Apartment Third or Higher Floor	46 (64.448)	1,650 (1,631.552)	1,696
Column Marginal Total	298	7,596	7,894

A quick comparison of the observed and expected cell counts in Table 12.3 reveals some large differences, suggesting that the survival proportions may not be the same for the three groups considered. This will be explored further in Example 12.5.

In Example 12.4, the expected count for a cell corresponding to a particular group-response combination was calculated in two steps. First, the response *marginal proportion* was calculated (for example, 298/7894 for the “survived” response). Then this proportion was multiplied by a marginal group total (for example, 5531(298/7894) for the house or townhouse group). Algebraically, this is equivalent to first multiplying the row and column marginal totals and then dividing by the grand total:

$$\frac{(5531)(298)}{7894}$$

To compare two or more populations or treatments on the basis of a categorical variable, calculate an **expected cell count** for each cell by selecting the corresponding row and column marginal totals and then calculating

$$\text{expected cell count} = \frac{(\text{row marginal total})(\text{column marginal total})}{\text{grand total}}$$

These quantities represent what would be expected when there is no difference between the groups under study.

The X^2 statistic, introduced in Section 12.1, can now be used to compare the observed cell counts to the expected cell counts. A large value of X^2 results when there are large differences between the observed and expected counts and suggests that the hypothesis of no differences between the populations should be rejected. A formal test procedure is described in the following box.

X² Test for Homogeneity

Null Hypothesis: H_0 : The category proportions are the same for all the populations or treatments (homogeneity of populations or treatments).

Alternative Hypothesis: H_a : The category proportions are not all the same for all of the populations or treatments.

Test Statistic: $X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$

The expected cell counts are estimated from the sample data (assuming that H_0 is true) using the formula

$$\text{expected cell count} = \frac{(\text{row marginal total})(\text{column marginal total})}{\text{grand total}}$$

P-values: When H_0 is true and the assumptions of the X^2 test are satisfied, X^2 has approximately a chi-square distribution with

$$\text{df} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

The P -value associated with the calculated test statistic value is the area to the right of X^2 under the chi-square curve with the appropriate df. Upper-tail areas for chi-square distributions are found in Appendix Table 8.

- Assumptions:**
1. The data are from *independently selected random samples or from subjects who were assigned at random to treatment groups.*
 2. *The sample size is large:* all expected counts are at least 5. If some expected counts are less than 5, rows or columns of the table may be combined to achieve a table with satisfactory expected counts.

Example 12.5 Heart Attacks in High-Rise Buildings Revised

The following table of observed and expected cell counts appeared in Example 12.4.

	Survived	Did Not Survive	Row Marginal Total
House or Townhouse	217 (210.178)	5,314 (5,320.822)	5,531
Apartment First or Second Floor	35 (23.346)	632 (641.654)	667
Apartment Third or Higher Floor	46 (64.448)	1,650 (1,631.552)	1,696
Column Marginal Total	298	7,596	7,894

We are interested in learning if there is a difference in the survival category proportions for the three populations (heart attack victims who live in a house or townhouse, heart attack victims who live on the first or second floor of an apartment building, and heart attack victims who live on the third or a higher floor in an apartment building).

Understand the context ➤

Hypotheses: H_0 : Proportions in each survival category are the same for all three groups.

H_a : The survival category proportions are not all the same for all three groups.

Significance level: A significance level of $\alpha = 0.05$ will be used.

Formulate a plan ➤

$$\text{Test statistic: } X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

Assumptions: The samples were independently selected and thought to be representative of the three populations of interest. This means that it is appropriate to use the chi-square test if the sample size is large enough. All of the expected cell counts are at least 5, so the sample is large enough to proceed with the test.

Do the work ➤ Calculation:

$$X^2 = \frac{(217 - 208.795)^2}{208.795} + \dots + \frac{(1650 - 1631.976)^2}{1631.976} = 9.59$$

P-value: The two-way table for this example has 3 rows and 2 columns, so the appropriate number of df is $(3 - 1)(2 - 1) = 2$. The calculated value of the test statistic is between 9.21 and 10.59 in the 2-df column of Appendix Table 8, so $0.005 < P\text{-value} < 0.010$.

Interpret the results ➤

Conclusion: The *P*-value is less than α (0.05), so H_0 is rejected. There is convincing evidence that the proportions in the survival categories are not the same for the three groups compared. Notice that there are more people who survived in the house or townhouse and first or second floor apartment categories than would have been expected if the survival proportions were the same for all three groups. This led the researchers who collected these data to conclude that there is a smaller chance of survival for people who suffer a heart attack in an apartment that is on the third or higher floor.

Most statistical software packages can calculate expected cell counts, the value of the X^2 statistic, and the associated *P*-value. This is illustrated in the following example.

Example 12.6 Keeping the Weight Off

The article “[Daily Weigh-ins Can Help You Keep Off Lost Pounds, Experts Say](#)” (*Associated Press, October 17, 2005*) describes an experiment in which 291 people who had lost at least 10% of their body weight in a medical weight loss program were assigned at random to one of three groups for follow-up. One group met monthly in person, one group “met” online monthly in a chat room, and one group received a monthly newsletter by mail. After 18 months, participants in each group were classified according to whether or not they had regained more than 5 pounds, resulting in the data summarized in Table 12.4.

TABLE 12.4 Observed and Expected Counts for Example 12.6

		Amount of Weight Gained		Row Marginal Total
		Regained 5 lb or Less		
	In-Person	52 (41.0)	45 (56.0)	97
	Online	44 (41.0)	53 (56.0)	97
	Newsletter	27 (41.0)	70 (56.0)	97

Does there appear to be a difference in the weight regained proportions for the three follow-up methods? The relevant hypotheses are

Understand the context ➤

H_0 : The proportions for the two weight-regained categories are the same for the three follow-up methods.

H_a : The weight-regained category proportions are not all the same for all three follow-up methods.

Significance level: $\alpha = 0.01$.

Formulate a plan ➤ Test statistic: $X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$

Assumptions: Table 12.4 contains the calculated expected counts, all of which are greater than 5. The subjects in this experiment were assigned at random to the treatment groups.

Do the work ➤ Calculation: Minitab output follows. For each cell, the Minitab output includes the observed cell count, the expected cell count, and the value of $\frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$ for that cell (this is the contribution to the X^2 statistic for this cell). From the output, $X^2 = 13.773$.

Chi-Square Test

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

	<=5	>5	Total
In-person	52	45	97
	41.00	56.00	
	2.951	2.161	
Online	44	53	97
	41.00	56.00	
	0.220	0.161	
Newsletter	27	70	97
	41.00	56.00	
	4.780	3.500	
Total	123	168	291

Chi-Sq = 13.773, DF = 2, P-Value = 0.001

P-value: From the Minitab output, P -value = 0.001.

Interpret the results ➤ Conclusion: Since the P -value is less than α , H_0 is rejected. The data indicate that the proportions who have regained more than 5 pounds are not the same for the three follow-up methods. Comparing the observed and expected cell counts, we can see that the observed number in the newsletter group who had regained more than 5 pounds was greater than would have been expected and the observed number in the in-person group who had regained 5 or more pounds was less than would have been expected if there were no difference in the three follow-up methods.

Testing for Independence of Two Categorical Variables

A chi-square test can also be used to investigate the possibility of an association between two categorical variables in a single population. For example, television viewers in a particular city might be categorized with respect to both preferred network (ABC, CBS, NBC, or PBS) and favorite type of programming (comedy, drama, or information and news). The question of interest is often whether knowledge of the value of one variable provides any information about the value of the other variable—that is, are the two variables independent?

Continuing the example, suppose that those who favor ABC prefer the three types of programming in proportions 0.4, 0.5, and 0.1 and that these proportions are also correct for individuals favoring any of the other three networks. Then, learning an individual's preferred network provides no added information about that individual's favorite type of programming. The categorical variables *preferred network* and *favorite program type* would be independent.

To see how expected counts are obtained in this situation, recall from Chapter 6 that if two outcomes A and B are independent, then

$$P(A \text{ and } B) = P(A)P(B)$$

so the proportion of time that the two outcomes occur together in the long run is the product of the two individual long-run relative frequencies. Similarly, two categorical variables are independent in a population if, for each particular category of the first variable and each particular category of the second variable,

$$\left(\begin{array}{c} \text{proportion of individuals} \\ \text{in a particular category} \\ \text{combination} \end{array} \right) = \left(\begin{array}{c} \text{proportion in} \\ \text{specified category} \\ \text{of first variable} \end{array} \right) \cdot \left(\begin{array}{c} \text{proportion in} \\ \text{specified category} \\ \text{of second variable} \end{array} \right)$$

This means that if 30% of all viewers prefer ABC and the proportions of program type preferences are as previously given, then, assuming that the two variables are independent, the proportion of individuals who both favor ABC and prefer comedy is $(0.3)(0.4) = 0.12$ (or 12%).

Multiplying the right-hand side of the expression above by the sample size gives the expected number of individuals in the sample who are in both specified categories of the two variables when the variables are independent. However, these expected counts cannot be calculated, because the individual population proportions are not known. The solution is to estimate each population proportion using the corresponding sample proportion:

$$\begin{aligned} \left(\begin{array}{c} \text{estimated expected number} \\ \text{in specified categories} \\ \text{of the two variables} \end{array} \right) &= (\text{sample size}) \cdot \frac{\left(\begin{array}{c} \text{observed number} \\ \text{in category of} \\ \text{first variable} \end{array} \right)}{\text{sample size}} \cdot \frac{\left(\begin{array}{c} \text{observed number} \\ \text{in category of} \\ \text{second variable} \end{array} \right)}{\text{sample size}} \\ &= \frac{\left(\begin{array}{c} \text{observed number in} \\ \text{category of first variable} \end{array} \right) \cdot \left(\begin{array}{c} \text{observed number in} \\ \text{category of second variable} \end{array} \right)}{\text{sample size}} \end{aligned}$$

Suppose that the observed counts are displayed in a rectangular table in which rows correspond to the categories of the first variable and columns to the categories of the second variable. Then, the numerator in the expression for estimated expected counts is just the product of the row and column marginal totals. This is exactly how expected counts were calculated in the test for homogeneity of several populations, even though the reasoning that leads to the formula is different.

X² Test for Independence

Null Hypothesis: H_0 : The two variables are independent.

Alternative Hypothesis: H_a : The two variables are not independent.

Test Statistic: $X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$

The expected cell counts are estimated (assuming H_0 is true) using the formula

$$\text{expected cell count} = \frac{(\text{row marginal total})(\text{column marginal total})}{\text{grand total}}$$

P-values: When H_0 is true and the assumptions of the X^2 test are satisfied, X^2 has approximately a chi-square distribution with

$$\text{df} = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

The P -value associated with the calculated test statistic value is the area to the right of X^2 under the chi-square curve with the appropriate df. Upper-tail areas for chi-square distributions are found in Appendix Table 8.

(continued)

- Assumptions:**
1. The observed counts are from a *random sample*.
 2. The *sample size is large*: All expected counts are at least 5. If some expected counts are less than 5, rows or columns of the table can be combined to achieve a table with satisfactory expected counts.

Example 12.7 A Pained Expression

The paper “**Facial Expression of Pain in Elderly Adults with Dementia**” (*Journal of Undergraduate Research* [2006]) examined the relationship between a nurse’s assessment of a patient’s facial expression and his or her self-reported level of pain. Data for 89 patients are summarized in Table 12.5.

TABLE 12.5 Observed Counts for Example 12.7

		Self-Report	
		Facial Expression	No Pain
Facial Expression	No Pain	17	40
	Pain	3	29

The authors were interested in determining if there is evidence of a relationship between a facial expression that reflects pain and self-reported pain because patients with dementia do not always give a verbal indication that they are in pain.

Using a 0.05 significance level, we will test

Understand the context ➤

H_0 : Facial expression and self-reported pain are independent.

H_a : Facial expression and self-reported pain are not independent.

Significance level: $\alpha = 0.05$.

Formulate a plan ➤

Test statistic: $X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$

Assumptions: Before we can check the assumptions we must first calculate the expected cell counts.

Row	Column	Cell	
			Expected Cell Count
1	1		$\frac{(57)(20)}{89} = 12.81$
	2		$\frac{(57)(69)}{89} = 44.19$
2	1		$\frac{(32)(20)}{89} = 7.19$
	2		$\frac{(32)(69)}{89} = 24.81$

All expected cell counts are greater than 5. Although the participants in the study were not randomly selected, they were thought to be representative of the population of nursing home patients with dementia. The observed and expected counts are given together in Table 12.6.

TABLE 12.6 Observed and Expected Counts for Example 12.7

Facial Expression	Self-Report	
	No Pain	Pain
No Pain	17 (12.81)	40 (44.19)
Pain	3 (7.19)	29 (24.81)

Do the work ➤ Calculation: $X^2 = \frac{(17 - 12.81)^2}{12.81} + \dots + \frac{(29 - 24.81)^2}{24.81} = 4.92$

P-value: The table has 2 rows and 2 columns, so $df = (2 - 1)(2 - 1) = 1$. The entry closest to 4.92 in the 1-df column of Appendix Table 8 is 5.02, so the approximate *P*-value for this test is

$$P\text{-value} \approx 0.025$$

- Interpret the results ➤ Conclusion: Since the *P*-value is less than α , we reject H_0 and conclude that there is convincing evidence that a nurse's assessment of facial expression and self-reported pain are not independent.
-

Example 12.8 | Exercise and Sleep Quality

The National Sleep Foundation asked each person in a representative sample of 1000 adult Americans about activity level and sleep quality ("2013 Sleep in America Poll," February 20, 2013, sleepfoundation.org/sites/default/files/RPT336%20Summary%20of%20Findings%202002%202020%202013.pdf, retrieved May 27, 2017). Survey participants were classified into one of four activity levels (none, light, moderate, and vigorous). Each participant was also classified into one of two sleep categories. Data consistent with summary quantities given in the paper are given in Table 12.7. Expected cell counts (calculated under the assumption of no association between activity level and sleep quality) are also shown in Table 12.7.

TABLE 12.7 Observed and Expected Counts for Example 12.8

	Poor Sleep Quality	Good Sleep Quality
No activity	40 (21.96)	50 (68.04)
Light	116 (117.12)	364 (362.88)
Moderate	57 (61.00)	193 (189.00)
Vigorous	31 (43.92)	149 (136.08)

The Sleep Foundation was interested in using these sample data to determine whether there was an association between quality of sleep and activity level.

The X^2 test with a significance level of 0.01 will be used to test the relevant hypotheses:

- Understand the context ➤ H_0 : Quality of sleep and activity level are independent.
 H_a : Quality of sleep and activity level are not independent.

Significance level: $\alpha = 0.01$.

Formulate a plan ➤ Test statistic: $X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$

Assumptions: All expected cell counts are at least 5. Assuming that the sample is representative of adult Americans, the X^2 test can be used.

Do the work ➤ Calculation: Minitab output is shown. From the Minitab output, $X^2 = 24.991$.

Chi-Square Test for Association: Activity Level, Sleep Quality			
Rows: Activity Level		Columns: Sleep Quality	
	Poor Sleep Quality	Good Sleep Quality	All
None	40 21.96	50 68.04	90
Light	116 117.12	364 362.88	480
Moderate	57 61.00	193 189.00	250
Vigorous	31 43.92	149 136.08	180
All	244	756	1000
Cell Contents:		Count	
		Expected count	
Pearson Chi-Square = 24.991, DF = 3, P-Value = 0.000			

P-value: From the Minitab output, *P*-value = 0.000.

Interpret the results ➤ Conclusion: Since the *P*-value is less than α , H_0 is rejected. There is convincing evidence that there is an association between quality of sleep and activity level.

In some investigations, values of more than two categorical variables are recorded for each individual in the sample. For example, in addition to the variables *quality of sleep* and *activity level*, the researchers in the study referenced in Example 12.8 might also have collected information on occupation. A number of interesting questions could then be explored: Are all three variables independent of one another? Is it possible that occupation and quality of sleep are dependent but that the relationship between them does not depend on activity level? For a particular activity level group, is there an association between quality of sleep and occupation?

The X^2 test procedure described in this section for analysis of bivariate categorical data can be extended for use with *multivariate categorical data*. Appropriate hypothesis tests can then be used to provide insight into the relationships between variables. However, the calculations required to determine expected cell counts and to calculate the value of X^2 are quite tedious, so they are seldom done without the aid of a computer. Most statistical software packages can perform this type of analysis.

EXERCISES 12.14 - 12.28

- 12.14** A particular state university system has six campuses. On each campus, a random sample of students will be selected, and each student will be categorized with respect to political philosophy as liberal, moderate, or conservative. The null hypothesis of interest is that the proportion of students falling in each of these three categories is the same at all six campuses.
- On how many degrees of freedom will the resulting X^2 test be based?
 - How does the answer in Part (a) change if there are seven campuses rather than six?
 - How does the answer in Part (a) change if there are four rather than three categories for political philosophy?
- 12.15** A random sample of 1000 registered voters in a certain county is selected, and each voter is categorized with respect to both educational level (four categories) and preferred candidate in an upcoming election for county supervisor (five possibilities). The hypothesis of interest is that educational level and preferred candidate are independent.

- a. If $X^2 = 7.2$, what is the conclusion at significance level 0.10?
- b. If there were only four candidates running for election, what would the conclusion be if $X^2 = 14.5$ and $\alpha = 0.05$?

- 12.16** The report “**Mobile Youth Around the World**” ([The Nielsen Company, December 2010](#)) provided the following information on sex of smartphone users for representative samples of young people age 15 to 24 in several different countries:

	Percent Female	Percent Male
United States	55%	45%
Spain	39%	61%
Italy	38%	62%
India	20%	80%

- a. Suppose the sample sizes were 1000 for the United States and for India and 500 for Spain and Italy. Complete the following two-way table by entering the observed counts.

	Female	Male
United States		
Spain		
Italy		
India		

- b. Carry out a hypothesis test to determine if there is convincing evidence that the sex proportions are not the same for all four countries. Use a significance level of 0.05. (Hint: See Example 12.7.)

- 12.17** Some colleges now allow students to pay their tuition using a credit card. The report “[Credit Card Tuition Payment Survey 2014](#)” ([creditcards.com/credit-card-news/tuition-charge-fee-survey.php](#), retrieved May 27, 2017) includes data from a survey of 100 public 4-year colleges, 100 private 4-year colleges, and 100 community colleges. The accompanying table gives information on credit card acceptance for each of these samples of colleges. For purposes of this exercise, suppose that these three samples are representative of the populations of public 4-year colleges, private 4-year colleges, and community colleges in the United States. Is there convincing evidence that the proportions in each of the two credit card categories are not the same for all three types of colleges? Test the

relevant hypotheses using a 0.05 significance level. (Hint: See Example 12.5.)

	Accepts Credit Cards for Tuition Payment	Does Not Accept Credit Cards for Tuition Payment
Public 4-Year Colleges	92	8
Private 4-Year Colleges	68	32
Community Colleges	100	0

- 12.18** The Knight Foundation asked each person in a representative sample of high school students and in a representative sample of high school teachers which of the rights guaranteed by the First Amendment they thought was the most important (“[Future of the First Amendment 2014 Survey of High School Students and Teachers](#),” [knightfoundation.org/media/uploads/publication_pdbs/Future_of_the_First_Amendment_cx2.pdf](#), retrieved May 27, 2017). Suppose that the sample size for each sample was 1000. Data consistent with summary values given in the paper are summarized in the table at the bottom of the page.

- a. Carry out a hypothesis test to determine if there is convincing evidence that the proportions falling into the five First Amendment right categories are not the same for teachers and students. Use a significance level of 0.01.
 b. Based on your test in Part (a) and a comparison of observed and expected cell counts, write a brief description of how teachers and students differ with respect to what they view as the most important of the First Amendment rights.

- 12.19** The Knight Foundation investigated whether high school students agreed with the statement that people should be allowed to burn or deface the American flag as a political statement. This question was asked in a survey of a representative sample of high school students in 2004 and again in a survey of a representative sample of high school students in 2014 (“[Future of the First Amendment: 2014 Survey of High School Students and Teachers](#),” [knightfoundation.org/media/uploads/publication_pdbs/Future_of_the_First_Amendment_cx2.pdf](#), retrieved May 27, 2017). Suppose that the sample size was 1000 in each of the 2 years. Data consistent with summary values

Table for Exercise 12.18

	Most Important First Amendment Right			
	Freedom of the Press	Freedom of Religion	Freedom to Peacefully Assemble	Freedom to Petition the Government
Students	650	30	250	20
Teachers	400	60	420	50

given in the paper are summarized in the accompanying table. Is there convincing evidence that the proportions falling into each of the response categories were not the same for high school students in 2004 and 2014?

People should be allowed to burn or deface the American flag as a political statement					
	Strongly Agree	Mildly Agree	Mildly Disagree	Strongly Disagree	Don't know
2004	80	80	110	630	100
2014	70	70	110	660	90

- 12.20** Each person in a representative sample of 445 college students age 18 to 24 was classified according to age and to the response to the following question: "How often have you used a credit card to buy items knowing you wouldn't have money to pay the bill when it arrived?" Possible responses were never, rarely, sometimes, or frequently (["Majoring in Money: How American College Students Manage Their Finances," June 28, 2016, salliemae.newshq.businesswire.com/sites/salliemae.newshq.businesswire.com/files/doc_library/file/SallieMae_MajoringinMoney_2016.pdf, retrieved May 27, 2017](#)). The responses are summarized in the accompanying table. Do these data provide evidence that there is an association between age group and the response to the question? Test the relevant hypotheses using $\alpha = 0.01$.

	Age 18 to 20	Age 21 to 22	Age 23 to 24
Never	72	62	29
Rarely	36	34	32
Sometimes	30	42	40
Frequently	12	24	32

- 12.21** The report ["Education Pays 2016"](#) ([The College Board, trends.collegeboard.org/sites/default/files/education-pays-2016-full-report.pdf, retrieved May 27, 2017](#)) provided information on education level and earnings for a sample of adult Americans who are employed full-time. Data consistent with summary percentages given in the report are summarized in the accompanying table. Suppose this data resulted from a representative sample of 1000 working adults whose highest level of education was a high school diploma, an Associate degree, or a Bachelor's degree. Each person in the sample was classified according to education level (high school diploma, Associate degree, or Bachelor's degree) and yearly income with possible categories of less than \$20,000, \$20,000 to \$39,999, \$40,000 to \$59,999 and \$60,000 or more. Is there evidence of an association between income category and education level? Test the appropriate hypotheses using a 0.05 significance level.

	Less than \$20,000	\$20,000 to \$39,999	\$40,000 to \$59,999	\$60,000 or more
High School Diploma	8	68	106	243
Associate Degree	11	56	56	63
Bachelor's Degree	47	160	101	82

- 12.22** The report ["Consumer Revolving Credit and Debt Over the Life Cycle and Business Cycle"](#) describes a study conducted by the [Federal Reserve Bank of Boston \(bostonfed.org, October 2015, retrieved May 27, 2017\)](#). Data consistent with summary values given in the report are summarized in the accompanying table. Suppose that these data resulted from a random sample of 800 adult Americans age 20 to 39 years old who have at least one credit card. Each person in the sample was classified according to age (with possible categories of 20 to 24 years, 25 to 29 years, 30 to 34 years, and 35 to 39 years). The people in the sample were also classified according to whether or not they pay the full balance on their credit cards each month or sometimes carry over a balance from month to month.

	Pay Full Balance Each Month	Carry Balance from Month to Month
Age 20 to 24 years	75	95
Age 25 to 29 years	74	126
Age 30 to 34 years	75	145
Age 35 to 39 years	67	143

- To investigate if whether or not people pay their balance in full each month is related to age, which chi-square test (homogeneity or independence) would be the appropriate test? Explain your choice.
- Carry out an appropriate test to determine if these data provide convincing evidence that whether or not people pay their balance in full each month is related to age.
- To what population would it be reasonable to generalize the conclusion from the test in Part (b)?

- 12.23** The paper ["Contemporary College Students and Body Piercing"](#) ([Journal of Adolescent Health \[2004\]: 58–61](#)) described a survey of 450 undergraduate students at a state university in the southwestern region of the United States. Each student in the sample was classified according to class standing (freshman, sophomore, junior, or senior) and body art category (body piercings only, tattoos only, both tattoos and body piercings, no body art).

Use the data in the accompanying table to determine if there is evidence of an association between class standing and response to the body art question. Assume that it is reasonable to regard the sample of students as representative of the students at this university. Use $\alpha = 0.01$. (Hint: See Example 12.7.)

	Body Piercings Only	Tattoos Only	Both Body Piercing and Tattoos	No Body Art
Freshman	61	7	14	86
Sophomore	43	11	10	64
Junior	20	9	7	43
Senior	21	17	23	54

- 12.24** The authors of the paper “Movie Character Smoking and Adolescent Smoking: Who Matters More, Good Guys or Bad Guys?” (*Pediatrics* [2009]: 135–141) classified characters who were depicted smoking in movies released between 2000 and 2005. The smoking characters were classified according to sex and whether the character type was positive, negative, or neutral. The resulting data are summarized in the accompanying table.

Assume that it is reasonable to consider this sample of smoking movie characters as representative of smoking movie characters. Do the data provide evidence of an association between sex and character type for movie characters who smoke? Use $\alpha = 0.05$. (Hint: See Example 12.7.)

Character Type			
Sex	Positive	Negative	Neutral
Male	255	106	130
Female	85	12	49

- 12.25** The data summarized in the accompanying table are from the paper “Gender Differences in Food Selections of Students at a Historically Black College and University” (*College Student Journal* [2009]: 800–806). Suppose that the data resulted from classifying each person in a random sample of 48 male students and each person in a random sample of 91 female students at a particular college according to their response to a question about whether they usually eat three meals a day or rarely eat three meals a day.

	Usually Eat 3 Meals a Day	Rarely Eat 3 Meals a Day
Male	26	22
Female	37	54

- a. Is there evidence that the proportions falling into each of the two response categories are not the

same for males and females? Use the X^2 statistic to test the relevant hypotheses with a significance level of 0.05.

- b. Are the calculations and conclusions from Part (a) consistent with the accompanying Minitab output?

Expected counts are printed below observed counts

Chi-Square contributions are printed below expected counts

	Usually	Rarely	Total
Male	26	22	48
	21.76	26.24	
	0.828	0.686	
Female	37	54	91
	41.24	49.76	
	0.437	0.362	
Total	63	76	139

$$\text{Chi-Sq} = 2.314, \text{ DF} = 1, \text{ P-Value} = 0.128$$

- c. Because the response variable in this exercise has only two categories (usually and rarely), we could have also answered the question posed in Part (a) by carrying out a two-sample z test of $H_0: p_1 - p_2 = 0$ versus $H_a: p_1 - p_2 \neq 0$, where p_1 is the proportion who usually eat three meals a day for males and p_2 is the proportion who usually eat three meals a day for females. Minitab output from the two-sample z test is shown below. Using a significance level of 0.05, does the two-sample z test lead to the same conclusion as in Part (a)?

Test for Two Proportions

Sample	X	N	Sample p
Male	26	48	0.541667
Female	37	91	0.406593

$$\text{Difference} = p(1) - p(2)$$

$$\text{Test for difference} = 0 \text{ (vs not } = 0\text{)} : \\ Z = 1.53 \text{ P-Value} = 0.127$$

- d. How do the P -values from the tests in Parts (a) and (c) compare? Is this surprising? Explain.

- 12.26** The paper “Credit Card Misuse, Money Attitudes, and Compulsive Buying Behavior: Comparison of Internal and External Locus of Control Consumers” (*College Student Journal* [2009]: 268–275) describes a study that surveyed a sample of college students at two midwestern public universities. Based on the survey responses, students were classified into two “locus of control” groups (internal and external) based on the extent to which they believe that they control what happens to them. Those in the internal locus of control group believe that they are usually in control of what happens to them, whereas those in the external locus of control group

believe that it is usually factors outside their control that determines what happens to them. Each student was also classified according to a measure of compulsive buying.

The resulting data are summarized in the accompanying table. Can the researchers conclude that there is an association between locus of control and compulsive buying behavior? Carry out a χ^2 test using $\alpha = 0.01$. Assume it is reasonable to regard the sample as representative of college students at midwestern public universities.

		Locus of Control	
		Internal	External
Compulsive Buyer?	Yes	3	14
	No	52	57

- 12.27** Each person in a large sample of German adolescents was asked to indicate which of 50 popular movies they had seen in the past year. Based on the response, the amount of time (in minutes) of alcohol use contained in the movies the person had watched was estimated. Each person was then classified into one of four groups based on the amount of movie alcohol exposure (groups 1, 2, 3, and 4, with 1 being the lowest exposure and 4 being the highest exposure). Each person was also classified according to school performance. The resulting data is given in the accompanying table (from “**Longitudinal Study of Exposure to Entertainment Media and Alcohol Use among German Adolescents**,” *Pediatrics* [2009]: 989–995).

Assume it is reasonable to regard this sample as a random sample of German adolescents. Is there evidence that there is an association between school performance and movie exposure to alcohol? Carry out a hypothesis test using $\alpha = 0.05$.

		Alcohol Exposure Group			
School Performance		1	2	3	4
		Excellent	Good	Average/	Poor
		110	93	49	65
		328	325	316	295
		239	259	312	317

- 12.28** Can people tell the difference between a female nose and a male nose? This research question was examined in the article “**You Can Tell by the Nose: Judging Sex from an Isolated Facial Feature**” (*Perception* [1995]: 969–973). Eight Caucasian males and eight Caucasian females posed for nose photos. The article states that none of the volunteers wore nose studs or had prominent nasal hair. Each person placed a black Lycra tube over his or her head in such a way that only the nose protruded through a hole in the material. Photos were then taken from three different angles: front view, three-quarter view, and profile.

These photos were shown to a sample of undergraduate students. Each student in the sample was shown one of the nose photos and asked whether it was a photo of a male or a female. The response was then classified as either correct or incorrect. The accompanying table was constructed using summary values reported in the article.

Is there evidence that the proportion of correct sex identifications differs for the three different nose views?

		View		
Sex ID		Front	Profile	Three-Quarter
		Correct	23	26
		Incorrect	17	14
				11

SECTION 12.3 Interpreting and Communicating the Results of Statistical Analyses

Many studies, particularly those in the social sciences, result in categorical data. The questions of interest in such studies often lead to an analysis that involves using a chi-square test.

Communicating the Results of Statistical Analyses

Three different chi-square tests were introduced in this chapter—the goodness-of-fit test, the test for homogeneity, and the test for independence. They are used in different settings and to answer different questions. When summarizing the results of a chi-square test, be sure to indicate which chi-square test was performed. One way to do this is to be clear about how the data were collected and the nature of the hypotheses being tested.

It is also a good idea to include a table of observed and expected counts in addition to reporting the value of the test statistic and the P -value. And finally, make sure to give a conclusion in context, and that the conclusion is worded appropriately for the type of test

conducted. For example, don't use terms such as *independence* and *association* to describe the conclusion if the test performed was a test for homogeneity.

Interpreting the Results of Statistical Analyses

As with the other hypothesis tests considered, it is common to find the result of a chi-square test summarized by giving the value of the chi-square test statistic and an associated *P*-value. Because categorical data can be summarized compactly in frequency tables, the data are often given in the article (unlike data for numerical variables, which are rarely given).

What to Look For in Published Data

Here are some questions to consider when reading an article that includes the results of a chi-square test:

- Are the variables of interest categorical rather than numerical?
- Are the data given in the article in the form of a frequency table?
- If a two-way frequency table is involved, is the question of interest one of homogeneity or one of independence?
- What null hypothesis is being tested? Are the results of the analysis reported in the correct context (homogeneity, etc.)?
- Is the sample size large enough to make use of a chi-square test reasonable? (Are all expected counts at least 5?)
- What is the value of the test statistic? Is the associated *P*-value given? Should the null hypothesis be rejected?
- Are the conclusions drawn consistent with the results of the test?
- How different are the observed and expected counts? Does the result have practical significance as well as statistical significance?

The authors of the article “[Predicting Professional Sports Game Outcomes from Intermediate Game Scores](#)” (*Chance* [1992]: 18–22) used a chi-square test to determine whether there was any merit to the idea that basketball games are not settled until the last quarter, whereas baseball games are over by the seventh inning. They also considered football and hockey. Data were collected for 189 basketball games, 92 baseball games, 80 hockey games, and 93 football games. The analyzed games were random samples from all games played during a single season. For each game, the late-game leader was determined, and then it was noted whether the late-game leader actually ended up winning the game. The resulting data are summarized in the following table:

	Late-Game Leader Wins	Late-Game Leader Loses
Basketball	150	39
Baseball	86	6
Hockey	65	15
Football	72	21

The authors stated that the

late-game leader is defined as the team that is ahead after three quarters in basketball and football, two periods in hockey, and seven innings in baseball. The chi-square value (with three degrees of freedom) is 10.52 ($P < 0.015$).

They also concluded that

the sports of basketball, hockey, and football have remarkably similar percentages of late-game reversals, ranging from 18.8% to 22.6%. The sport that is an anomaly is baseball. Only 6.5% of baseball games resulted in late reversals. . . . [The chi-square test] is statistically significant due almost entirely to baseball.

In this particular analysis, the authors are comparing four populations (games from each of the four sports) on the basis of a categorical variable with two categories (late-game leader wins and late-game leader loses). The appropriate null hypothesis is then

H_0 : The population proportion in each category (leader wins, leader loses) is the same for all four sports.

Based on the reported value of the chi-square statistic and the associated P -value, this null hypothesis is rejected, leading to the conclusion that the category proportions are not the same for all four sports.

The validity of the chi-square test requires that the sample sizes be large enough so that no expected counts are less than 5. Is this reasonable here? The following Minitab output shows the expected cell counts and the calculation of the X^2 statistic:

Chi-Square Test			
Expected counts are printed below observed counts			
	Leader W	Leader L	Total
1	150	39	189
	155.28	33.72	
2	86	6	92
	75.59	16.41	
3	65	15	80
	65.73	14.27	
4	72	21	93
	76.41	16.59	
Total	373	81	454
Chi-Sq =	0.180 + 0.827 +		
	1.435 + 6.607 +		
	0.008 + 0.037 +		
	0.254 + 1.171 = 10.518		
DF = 3, P-Value = 0.015			

The smallest expected count is 14.27, so the sample sizes are large enough to justify the use of the X^2 test. Notice that the two cells in the table that correspond to baseball contribute a total of $1.435 + 6.607 = 8.042$ to the value of the X^2 statistic of 10.518. This is due to the large discrepancies between the observed and expected counts for these two cells. There is reasonable agreement between the observed and the expected counts in the other cells. This is probably the basis for the authors' conclusion that baseball is unusual and that the other sports were similar.

A Word to the Wise: Cautions and Limitations

Be sure to keep the following in mind when analyzing categorical data using one of the chi-square tests presented in this chapter:

1. Don't confuse tests for homogeneity with tests for independence. The hypotheses and conclusions are different for the two types of test. Tests for homogeneity are used when the individuals in each of two or more independent samples are classified according to a single categorical variable. Tests for independence are used when individuals in a *single* sample are classified according to two categorical variables.
2. As was the case for the hypothesis tests of earlier chapters, remember that we can never say we have strong support for the null hypothesis. For example, if we do not reject the null hypothesis in a chi-square test for independence, we cannot conclude that there is convincing evidence that the variables are independent. We can only say that we were not convinced that there is an association between the variables.

3. Be sure that the assumptions for the chi-square test are reasonable. P -values based on the chi-square distribution are only approximate, and if the large sample conditions are not met, the actual P -value may be quite different from the approximate one based on the chi-square distribution. This can sometimes lead to incorrect conclusions. Also, for the chi-square test of homogeneity, the assumption of *independent* samples is particularly important.
4. Don't jump to conclusions about causation. Just as a strong correlation between two numerical variables does not mean that there is a cause-and-effect relationship between them, an association between two categorical variables does not imply a causal relationship.

EXERCISES 12.29 - 12.31

- 12.29** The following passage is from the paper “Gender Differences in Food Selections of Students at a Historically Black College and University” (*College Student Journal* [2009]: 800–806):

Also significant was the proportion of males and their water consumption (8 oz. servings) compared to females ($X^2 = 8.16$, $P = 0.086$). Males came closest to meeting recommended daily water intake (64 oz. or more) than females (29.8% vs. 20.9%).

This statement was based on carrying out a X^2 test of independence using data in a two-way table where rows corresponded to sex (male, female) and columns corresponded to number of servings of water consumed per day, with categories none, one, two to three, four to five, and six or more.

- What hypotheses did the researchers test? What is the number of degrees of freedom associated with the report value of the X^2 statistic?
- The researchers based their statement that the proportions falling in the water consumption categories were not all the same for males and females on a test with a significance level of 0.10. Would they have reached the same conclusion if a significance level of 0.05 had been used? Explain.
- The paper also included the accompanying data on how often students said they had consumed fried potatoes (fries or potato chips) in the past week.

	Number of Times Consumed Fried Potatoes in the Past Week						21 or more
	0	1 to 3	4 to 6	7 to 13	14 to 20		
Sex	Male	2	10	15	12	6	3
	Female	15	15	10	20	19	12

Use the Minitab output that follows to carry out a X^2 test of independence. Is the authors' conclusion that there was a significant association between sex and consumption of fried potatoes justified?

Expected counts are printed below observed counts

Chi-Square contributions are printed below expected counts

	0	1-3	4-6	7-13	14-20	21 or more	Total
M	2	10	15	12	6	3	48
	5.87	8.63	8.63	11.05	8.63	5.18	
	2.552	0.216	4.696	0.082	0.803	0.917	
F	15	15	10	20	19	12	91
	11.13	16.37	16.37	20.95	16.37	9.82	
	1.346	0.114	2.477	0.043	0.424	0.484	
Total	17	25	25	32	25	15	139
Chi-Sq =	14.153	DF = 5	P-Value =	0.015			

- 12.30** The article titled “Nap Time” (pewsocialtrends.org/2009/07/29/nap-time/, July 2009, retrieved April 14, 2018) described results from a nationally representative survey of 1488 adult Americans. The survey asked several demographic questions (such as sex, age, and income) and also included a question asking respondents if they had taken a nap in the past 24 hours. The article stated that 38% of the men surveyed and 31% of the women surveyed reported that they had napped in the past 24 hours. For purposes of this exercise, suppose that men and women were equally represented in the sample.
- Use the given information to fill in observed cell counts for the following table:

	Napped	Did Not Nap	Row Total
Men			744
Women			744

- Use the data in the table from Part (a) to carry out a hypothesis test to determine if there is an association between sex and napping.
- The press release states that more men than women nap. Although this is true for the people in the sample, based on the result of the test in Part (b), is it reasonable to conclude that this holds for adult Americans in general? Explain.

12.31 Using data from a national survey, the authors of the paper “[What Do Happy People Do?](#)” (*Social Indicators Research* [2008]: 565–571) concluded that there was convincing evidence of an association between amount of time spent watching television and whether or not a person reported that they were happy. They observed that unhappy people tended to watch more television. The authors write:

This could lead us to two possible interpretations:

1. Television viewing is a pleasurable enough activity with no lasting benefit, and it pushes aside time spent in other activities—ones that might be less immediately pleasurable, but that would provide long-term benefits in one’s condition. In other words, television does cause people to be less happy.

2. Television is a refuge for people who are already unhappy. TV is not judgmental nor difficult, so people with few social skills or resources for other activities can engage in it. Furthermore, chronic unhappiness can be socially and personally debilitating and can interfere with work and most social and personal activities, but even the unhappiest people can click a remote and be passively entertained by a TV. In other words, the causal order is reversed for people who watch television; unhappiness leads to television viewing.

Using only data from this study, is it possible to determine which of these two conclusions is correct? If so, which conclusion is correct and why? If not, explain why it is not possible to decide which conclusion is correct based on the study data.

CHAPTER ACTIVITIES

ACTIVITY 12.1 PICK A NUMBER, ANY NUMBER . . .

Background: There is evidence to suggest that human beings are not very good random number generators. In this activity, you will investigate this phenomenon by collecting and analyzing a set of human-generated “random” digits.

For this activity, work in a group with four or five other students.

1. Each member of the group should complete this step individually. Ask 25 different people to pick a digit from 0 to 9 at random. Record the responses.
2. Combine the responses you collected with those of the other members of your group to form a single sample. Summarize the resulting data in a one-way frequency table.

3. If people are adept at picking digits at random, what would you expect for the proportion of the responses in the sample that were 0? that were 1?
4. State a null hypothesis and an alternative hypothesis that could be tested to determine whether there is evidence that the 10 digits from 0 to 9 are not selected an equal proportion of the time when people are asked to pick a digit at random.
5. Carry out the appropriate hypothesis test, and write a few sentences indicating whether or not the data support the theory that people are not good random number generators.

ACTIVITY 12.2 COLOR AND PERCEIVED TASTE

Background: Does the color of a food or beverage affect the way people perceive its taste? In this activity you will conduct an experiment to investigate this question and analyze the resulting data using a chi-square test.

You will need to recruit at least 30 subjects for this experiment, so it is advisable to work in a large group (perhaps even the entire class) to complete this activity.

Subjects for the experiment will be assigned at random to one of two groups. Each subject will be asked to taste a sample of gelatin (for example, Jell-O) and rate the taste as not very good, acceptable, or very good. Subjects assigned to the first group will be asked to taste and rate a cube of lemon-flavored gelatin. Subjects in the second group will be asked to taste and rate a cube of lemon-flavored gelatin that has been colored an unappealing color by adding food coloring to the gelatin mix before the gelatin sets.

Note: You may choose to use something other than gelatin, such as lemonade. Any food or beverage whose color can be altered using food coloring can be used. You can experiment with the food colors to obtain a color that you think is particularly unappealing!

1. As a class, develop a plan for collecting the data. How will subjects be recruited? How will they be assigned to one of the two treatment groups (unaltered color, altered color)? What extraneous variables will be directly controlled, and how will you control them?
2. After the class is satisfied with the data collection plan, assign members of the class to prepare the gelatin to be used in the experiment.

3. Carry out the experiment, and summarize the resulting data in a two-way table like the one shown:

		Taste Rating		
		Not Very Good	Acceptable	Very Good
Unaltered Color	Unaltered Color			
	Altered Color			

4. The two-way table summarizes data from two independent samples (as long as subjects were assigned *at random* to the two treatments, the samples are independent). Carry out an appropriate test to determine whether the proportion for each of the three taste rating categories is the same when the color is altered as when the color is not altered.

SUMMARY Key Concepts and Formulas

TERM OR FORMULA

One-way frequency table

$$X^2 = \sum_{\text{all cell}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}}$$

X^2 goodness-of-fit test

Two-way frequency table (contingency table)

X^2 test for homogeneity

X^2 test for independence

COMMENT

A compact way of summarizing data on a categorical variable. It gives the number of times each of the possible categories in the data set occurs (the frequencies).

A statistic used to provide a comparison between observed counts and those expected when a given hypothesis is true. When none of the expected counts are too small, X^2 has approximately a chi-square distribution.

A hypothesis test performed to determine whether population category proportions are different from those specified by a given null hypothesis.

A rectangular table used to summarize a categorical data set. Two-way tables are used to compare several populations on the basis of a categorical variable or to determine if an association exists between two categorical variables.

The hypothesis test performed to determine whether category proportions are the same for two or more populations or treatments.

The hypothesis test performed to determine whether an association exists between two categorical variables.

CHAPTER REVIEW Exercises 12.32 - 12.36

- 12.32** Each observation in a random sample of 100 bicycle accidents resulting in death was classified according to the day of the week on which the accident occurred. Data consistent with information given on the web site highwaysafety.com are given in the following table

Day of Week	Frequency
Sunday	14
Monday	13
Tuesday	12
Wednesday	15
Thursday	14
Friday	17
Saturday	15

Based on these data, is it reasonable to conclude that the proportion of accidents is not the same for all days of the week? Use $\alpha = 0.05$.

- 12.33** The color vision of birds plays a role in their foraging behavior: Birds use color to select and avoid certain types of food. The authors of the article “Colour Avoidance in Northern Bobwhites: Effects of Age, Sex, and Previous Experience” (*Animal Behaviour* [1995]: 519–526) studied the pecking behavior of 1-day-old bobwhites. In an area painted white, they inserted four pins with different colored

heads. The color of the pin chosen on the bird’s first peck was noted for each of 33 bobwhites, resulting in the accompanying table.

Color	First Peck Frequency
Blue	16
Green	8
Yellow	6
Red	3

Do the data provide evidence of a color preference? Test using $\alpha = 0.01$.

- 12.34** The report “Majoring in Money: How American College Students Manage Their Finances,” (June 28, 2016, salliemae.newshq.businesswire.com/sites/salliemae.newshq.businesswire.com/files/doc_library/file/SallieMae_MajoringinMoney_2016.pdf, retrieved May 27, 2017) included data from a study in which 792 people in a representative sample of college students age 18 to 24 were asked how they perceive their money management skills. Possible responses were excellent, good, average, not very good, and poor. Each student in the sample was also classified by sex, resulting in the data in the accompanying table. Is there convincing evidence that there is an association between sex and how students

perceive their money management skills? Test the relevant hypotheses using a significance level of 0.05.

Perception of Money Management Skills					
	Excellent	Good	Average	Not Very Good	Poor
Male	103	139	89	17	8
Female	83	192	135	22	4

- 12.35** The report referenced in the previous exercise also provided data on perception of money management skills by age group. Use the data from 788 people, in the accompanying table to determine if there is evidence of an association between age and perception of money management skills. Because some of the expected values are less than 5, construct a new table that combines the not very good and the poor categories. Use a significance level of 0.05.

Perception of Money Management Skills					
	Excellent	Good	Average	Not Very Good	Poor
Age 18 to 20	73	146	104	17	8
Age 21 to 22	54	103	85	10	3
Age 23 to 24	58	84	36	7	0

- 12.36** The authors of the paper “Risk of Malnutrition Is an Independent Predictor of Mortality, Length of Hospital Stay, and Hospitalization Costs in Stroke Patients” (*Journal of Stroke and Cerebrovascular Diseases* [2016]: 799–806) describe a sample of patients admitted to a hospital after suffering a stroke. Each of 537 patients was classified according to a measure of risk of malnutrition (with possible categories low, medium, and high) and whether or not the patient was alive at 6 months following the stroke. The authors concluded that there was an association between survival and the risk of malnutrition. Do you agree? Support your answer with evidence based on that data. For purposes of this exercise, you may assume that the sample of 537 patients is representative of stroke patients.

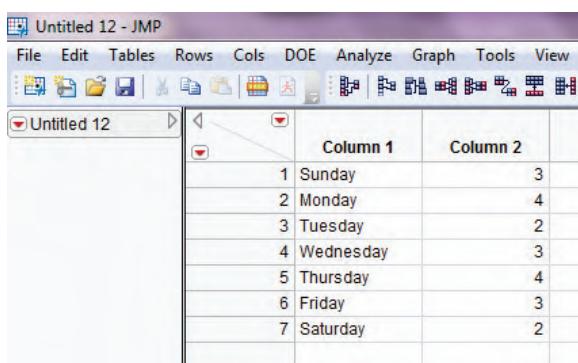
	Survived	Did Not Survive
Low Risk of Malnutrition	322	20
Medium Risk of Malnutrition	29	10
High Risk of Malnutrition	91	65

TECHNOLOGY NOTES

χ^2 Goodness-of-Fit Test

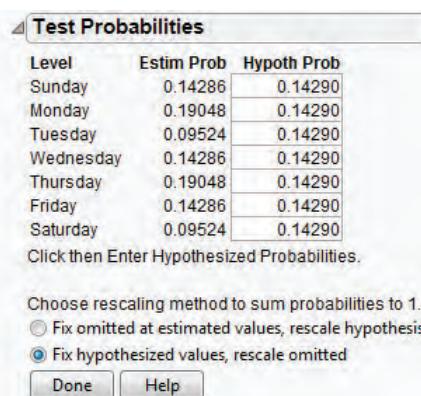
JMP

- Enter each category name into the first column
- Enter the count for each category into the second column



- Click **Analyze** then select **Distribution**
- Click and drag the name for the first column from the box under **Select Columns** to the box next to **Y, Columns**
- Click and drag the name for the second column from the box under **Select Columns** to the box next to **Freq**
- Click **OK**
- Click the red arrow next to the column name and select **Test Probabilities**

- Under **Test Probabilities** input the hypothesized probabilities for each category



- Click **Done**

Note: The test statistic and *P*-value for the chi-squared test will appear in the row called **Pearson**.

Minitab

MINITAB Student Version 14 does not have the functionality to produce a χ^2 goodness-of-fit test.

SPSS

1. Input the observed data into one column
2. Click **Analyze** then click **Nonparametric Tests** then click **One Sample...**
3. Click the **Settings** tab
4. Select the radio button next to **Customize tests**
5. Check the box next to **Compare observed probabilities to hypothesized (Chi-Square test)**
6. Click the **Options...** button
7. Select the appropriate option (to test with equal probabilities for each category or to input expected probabilities manually)
8. Once the appropriate option is selected and input expected probabilities if necessary, click **OK**
9. Click **Run**

Excel

Excel does not have the functionality to produce the X^2 goodness-of-fit test automatically. Use Excel to find the P -value for this test after finding the value of the test statistic by using the following steps.

1. Click on an empty cell
2. Click **Formulas**
3. Click **Insert Function**
4. Select **Statistical** from the drop-down box for category
5. Select **CHIDIST** and press **OK**
6. In the box next to **X**, type the value of the test statistic
7. In the box next to **Deg_freedom** type the value for the degrees of freedom
8. Click **OK**

Note: This outputs the value for $P(X \geq x)$.

TI-83/84

1. Enter the observed cell counts into **L1** and the expected cell counts into **L2** (in order to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**
4. Highlight **X2GOF-Test** and press **ENTER**
5. Next to Observed enter **L1**
6. Next to Expected enter **L2**
7. Next to df enter the appropriate df (this will be the number of categories – 1)
8. Highlight **Calculate** and press **ENTER**

TI-Nspire

1. Enter the observed data into a data list (In order to access data lists select the spreadsheet option and press **enter**)
2. Be sure to title the lists by selecting the top row of the column and typing a title.
3. Enter the expected data into a data list
4. Press **menu** and select **4:Statistics** then **4:Stat Tests** then **7: χ^2 GOF...** and press **enter**
5. For **Observed List** select the title of the list that contains the observed data from the drop-down menu

5. For **Expected List** select the title of the list that contains the expected data from the drop-down menu
6. For **Deg of Freedom** input the degrees of freedom for this test
7. Press **OK**

 X^2 Tests for Independence and Homogeneity**JMP**

1. Input the data table into the JMP data table
2. Click **Analyze** then select **Fit Y by X**
3. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Response**

	Column 1	Column 2	Column 3
1	A	Yes	15
2	A	No	10
3	B	Yes	25
4	B	No	5

4. Click and drag the second column name from the box under **Select Columns** to the box next to **X, Factor**
5. Click and drag the third column name from the box under **Select Columns** to the box next to **Freq**
6. Click **OK**

Minitab

1. Input the data table into MINITAB

Worksheet 2 ***				
	C1-T	C2	C3	C4
	Male	Female		
1	Small	145	165	
2	Medium	120	145	
3	Large	300	75	
4	XL	45	6	
5				
6				

2. Click **Stat** then **Tables** then **Chi-Square Test (Table in Worksheet)**
3. Select all columns containing data (do NOT select the column containing the row labels)
4. Click **OK**

Note: This output returns expected cell counts as well as the chi-square test statistic and P -value.

SPSS

1. Enter the row variable data into one column
2. Enter the column variable data into a second column

	VAR00001	VAR00002	var	var	var	var	
1	M	Y					
2	M	N					
3	M	N					
4	M	Y					
5	M	N					
6	F	Y					
7	F	Y					
8	F	Y					
9	F	Y					
10	F	N					
11							
12							

3. Click **Analyze** then click **Descriptive Statistics** then click **Crosstabs...**
4. Select the name of the row variable and press the arrow to move the variable to the box under **Row(s)**:
5. Select the name of the column variable and press the arrow to move the variable to the box under **Column(s)**:
6. Click **Statistics...**
7. Check the box next to Chi-square
8. Click **Continue**
9. Click **Cells...**
10. Check the box next to Expected
11. Click **Continue**
12. Click **OK**

Note: The *P*-value for this test can be found in the **Chi-Square Tests** table in the Pearson Chi-Square row.

Excel

1. Input the observed contingency table
2. Input the expected table
3. Click on an empty cell
4. Click **Formulas**
5. Click **Insert Function**
6. Select **Statistical** from the drop-down box for category
7. Select **CHITEST** and press **OK**
8. Click in the box next to **Actual_range** and select the data values from the actual table (do NOT select column or row labels or totals)
9. Click in the box next to **Expected_range** and select the data values from the expected table (do NOT select column or row labels or totals)
10. Click **OK**

TI-83/84

1. Input the observed contingency table into matrix A (To access and edit matrices, press **2nd** then **x⁻¹**, then highlight **EDIT** and press **ENTER**. Then highlight [A] and press **ENTER**. Type the value for the number of rows and press **ENTER**; type the value for the number of columns and press **ENTER**. Type the data values into the matrix.)

MATRIX[A] 3 ×3

$$\begin{bmatrix} 33 & 65 & 82 \\ 45 & 79 & 82 \\ 21 & 47 & 82 \end{bmatrix}$$

 $3,3=63$

NAME MATH EDIT
1: [A] 3×3
2: [B] 3×3
3: [C]
4: [D]
5: [E]
6: [F]
7: [G]

Excel Worksheet Content After Step 4 Above

Book1 - Microsoft Excel											
Function Library											
A6	fx										
1	OBSERVED	Yes	No	Total		EXPECTED	Yes	No	Total		
2	Male	14	25	39		Male	20.05714	18.94286	39		
3	Female	40	26	66		Female	33.94286	32.05714	66		
4	Total	54	51	105		Total	54	51	105		
5											
6											
7											
8											
9											
10											

(continued)

Excel Worksheet Content After Step 10 on Page 687

The screenshot shows a Microsoft Excel spreadsheet titled "Book1 - Microsoft Excel". The ribbon tabs are Home, Insert, Page Layout, Formulas, Data, Review, and View. The Formulas tab is selected. The Function Library group contains icons for Insert Function, AutoSum, Recently Used, Financial, Logical, Text, Date & Time, Lookup & Reference, Math & Trig, and More Functions. Below the ribbon, a formula bar displays =CHITEST(B2:C3,G2:H3). The main area of the spreadsheet contains a contingency table:

	A	B	C	D	E	F	G	H	I	J	K
1	OBSERVED	Yes	No	Total		EXPECTED	Yes	No	Total		
2	Male	14	25	39		Male	20.05714	18.94286	39		
3	Female	40	26	66		Female	33.94286	32.05714	66		
4	Total	54	51	105		Total	54	51	105		
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											

A "Function Arguments" dialog box is open over the spreadsheet. It is titled "CHITEST". The "Actual_range" is set to B2:C3, which is highlighted in yellow in the spreadsheet. The "Expected_range" is set to G2:H3. The formula result is shown as 0.014375481. The dialog box has "OK" and "Cancel" buttons.

2. Press STAT
3. Highlight TESTS
4. Highlight χ^2 -Test... and press ENTER
5. Highlight Calculate and press ENTER

TI-Nspire

1. Enter the Calculator Scratchpad
2. Press the menu key then select 7:Matrix & Vector then select 1:Create then select 1:Matrix... and press enter
3. Next to Number of rows enter the number of rows in the contingency table (do not include title rows or total rows)
4. Next to Number of columns enter the number of columns in the contingency table (do not include title columns or total columns)
5. Press OK
6. Input the values into the matrix (pressing tab after entering each value and press enter)
7. Press **ctrl** then press var

8. Type in amat and press enter

The screenshot shows the TI-Nspire Scratchpad window. The top bar says "Scratchpad". The scratchpad area contains the following text and calculations:

```

9-10
-3.16228
2
√40
[33 65 82] →amat
[33 65 82]
[45 79 95]
[21 47 63]

```

The matrix [33 65 82] is labeled →amat. The matrix [33 65 82] is shown again below it. The matrix [45 79 95] and [21 47 63] are also shown. The bottom right corner of the scratchpad shows a progress bar at 9/99.

9. Press the menu key then select 6:Statistics then 7:Stat tests then 8: χ^2 2-way Test...
10. For Observed Matrix, select amat from the drop-down list
11. Press OK

13

Simple Linear Regression and Correlation: Inferential Methods



Monkey Business Images/Shutterstock.com

of a meaningful relationship between these two variables, the regression line could be used as the basis for predicting the GPA for an older student with a specified language processing score.

In this chapter, we develop inferential methods for bivariate numerical data, including a hypotheses test to determine if there is a useful linear relationship in the entire population of (x, y) pairs.

LEARNING OBJECTIVES

Students will understand:

- That the simple linear regression model provides a basis for making inferences about linear relationships.

Students will be able to:

- Interpret the parameters of the simple linear regression model in context.
- Use scatterplots, residual plots, and normal probability plots to assess whether the assumptions of the simple linear regression model are reasonable.
- Construct and interpret a confidence interval for the slope of the population regression line.
- Test hypotheses about the slope of the population regression line and interpret the results in context.

Regression and correlation were introduced in Chapter 5 as methods for describing and summarizing bivariate numerical data consisting of (x, y) pairs. For example, data on

$$y = \text{College grade point average (GPA)}$$

and

$$x = \text{Language processing score}$$

for a sample of older college students (age 50 to 79) were used in an investigation of predictors of academic success for older students ([“Age Is No Barrier: Predictors of Academic Success in Older Learners,” Nature Partner Journals, nature.com/articles/s41539-017-0014-5.pdf, retrieved April 17, 2018](#)). The data were used to construct a scatterplot. The scatterplot showed a linear pattern. The sample correlation coefficient was $r = 0.323$, and the equation of the least-squares line had a positive slope, indicating that older students with higher language processing scores tended to have higher GPAs.

Could the pattern observed in the scatterplot be plausibly explained by chance, or does the sample provide convincing evidence that there really is a linear relationship between these two variables for older college students? If there is evidence

SECTION 13.1 Simple Linear Regression Model

A **deterministic relationship** is one in which the value of y is completely determined by the value of an independent variable x . A deterministic relationship can be described using traditional mathematical notation $y = f(x)$, where $f(x)$ is a specified function of x . For example, we might have

$$y = f(x) = 10 + 2x$$

or

$$y = f(x) = 4 - (10)^{2x}$$

However, in many situations, the variables of interest are not deterministically related. For example, the value of y = first-year college grade point average is not an exact function of x = high school grade point average, and y = crop yield is determined partly by factors other than x = amount of fertilizer used.

A relationship between two variables x and y that are not deterministically related is described by a **probabilistic model**. The general form of an **additive probabilistic model** allows y to be larger or smaller than $f(x)$ by a random amount e . The **model equation** is of the form

$$y = \text{deterministic function of } x + \text{random deviation} = f(x) + e$$

For example, the graph of the function $y = 50 - 10x + x^2$ is shown as the orange curve in Figure 13.1. The observed point $(4, 30)$ is also shown in the figure. Because

$$f(4) = 50 - 10(4) + 4^2 = 50 - 40 + 16 = 26$$

for the point $(4, 30)$, we can write $y = f(x) + e$, where $e = 4$. The point $(4, 30)$ falls 4 above the graph of the function $y = 50 - 10x + x^2$.

Thinking geometrically, if $e > 0$, the corresponding point will lie above the graph of $y = f(x)$. If $e < 0$ the corresponding point will fall below the graph. If $f(x)$ is a function used in a probabilistic model relating y to x and if observations on y are made for various values of x , the resulting (x, y) points will be distributed around the graph of $f(x)$, some falling above it and some falling below it.

For example, consider the probabilistic model

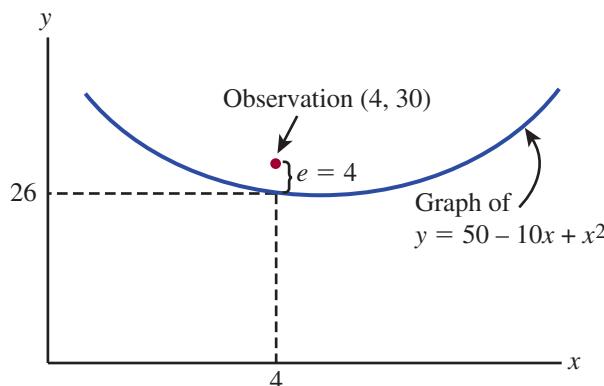
$$y = \underbrace{50 - 10x + x^2}_{f(x)} + e$$

Simple Linear Regression

The simple linear regression model is a special case of the general probabilistic model in which the deterministic function $f(x)$ is linear (so its graph is a straight line).

FIGURE 13.1

A deviation from the deterministic part of a probabilistic model.



DEFINITIONS

Simple linear regression model: A probabilistic model in which the deterministic part is a line with vertical or y intercept α and slope β . This line is called the **population regression line**.

When a value of the independent variable x is fixed and an observation on the dependent variable y is made,

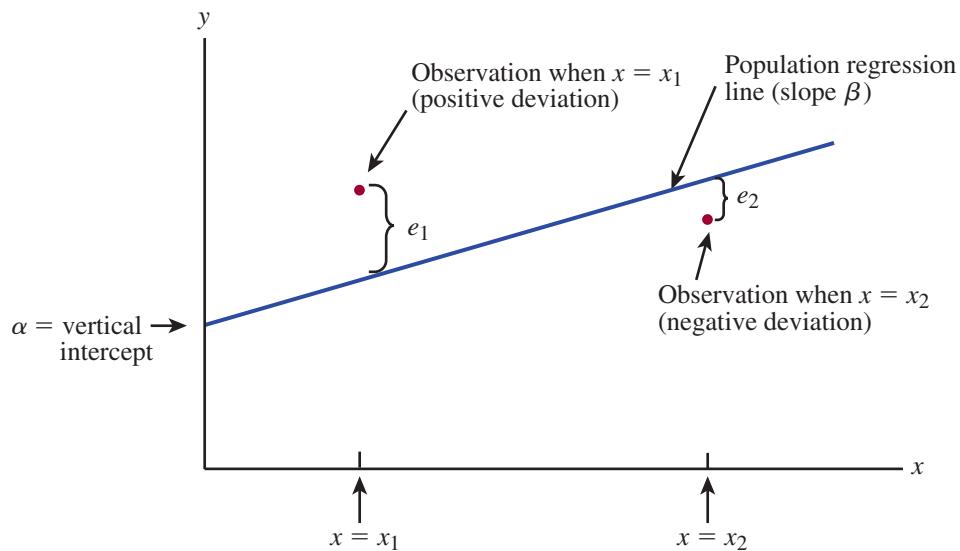
$$y = \alpha + \beta x + e$$

Without the random deviation e , all observed (x, y) points would fall exactly on the population regression line. The inclusion of e in the model equation recognizes that points will deviate from the line by a random amount.

Figure 13.2 shows two observations in relation to the population regression line.

FIGURE 13.2

Two observations and deviations from the population regression line.



Before we make an observation of y for any particular value of x , we are uncertain about the value of e . It could be negative, positive, or even 0. Also, it might be quite large in magnitude (a point far from the population regression line) or quite small (a point very close to the line). In this chapter, we make some assumptions about the distribution of e in repeated sampling at any particular x value.

Basic Assumptions of the Simple Linear Regression Model

1. The distribution of e at any particular x value has mean value 0. That is, $\mu_e = 0$.
2. The standard deviation of e (which describes the variability in its distribution) is the same for any particular value of x . This standard deviation is denoted by σ .
3. The distribution of e at any particular x value is normal.
4. The random deviations e_1, e_2, \dots, e_n associated with different observations are independent of one another.

These assumptions about the e term in the simple linear regression model also imply that there is variability in the y values observed at any particular value of x . Consider y when x has some fixed value x^* , so that

$$y = \alpha + \beta x^* + e$$

Because α and β are fixed numbers, $\alpha + \beta x^*$ is also a fixed number. The sum of a fixed number and a normally distributed variable (e) is also a normally distributed variable (the bell-shaped curve is simply relocated). This means that y itself has a normal distribution. Furthermore, $\mu_e = 0$ implies that the mean value of y is $\alpha + \beta x^*$, the height of the population regression line above (or below) the value x^* . Finally, because there is no variability in the fixed number $\alpha + \beta x^*$, the standard deviation of y is the same as the standard deviation of e . These properties are summarized in the following box.

At any fixed x value, y has a normal distribution, with

$$\begin{pmatrix} \text{mean } y \text{ value} \\ \text{for fixed } x \end{pmatrix} = \begin{pmatrix} \text{height of the population} \\ \text{regression line above } x \end{pmatrix} = \alpha + \beta x$$

and

$$(\text{standard deviation of } y \text{ for a fixed } x) = \sigma$$

The slope β of the population regression line is the *average* change in y associated with a 1-unit increase in x .

The y intercept α is the height of the population line when $x = 0$.

The value of σ determines the extent to which (x, y) observations deviate from the population line. When σ is small, most observations will be quite close to the line, but when σ is large, there are likely to be some large deviations.

The key features of the model are illustrated in Figures 13.3 and 13.4. Notice that the three normal curves in Figure 13.3 have the same standard deviation (equal variability). This is a consequence of $\sigma_e = \sigma$, which implies that the variability in the y values at a particular x does not depend on the value of x .

FIGURE 13.3

Illustration of the simple linear regression model.

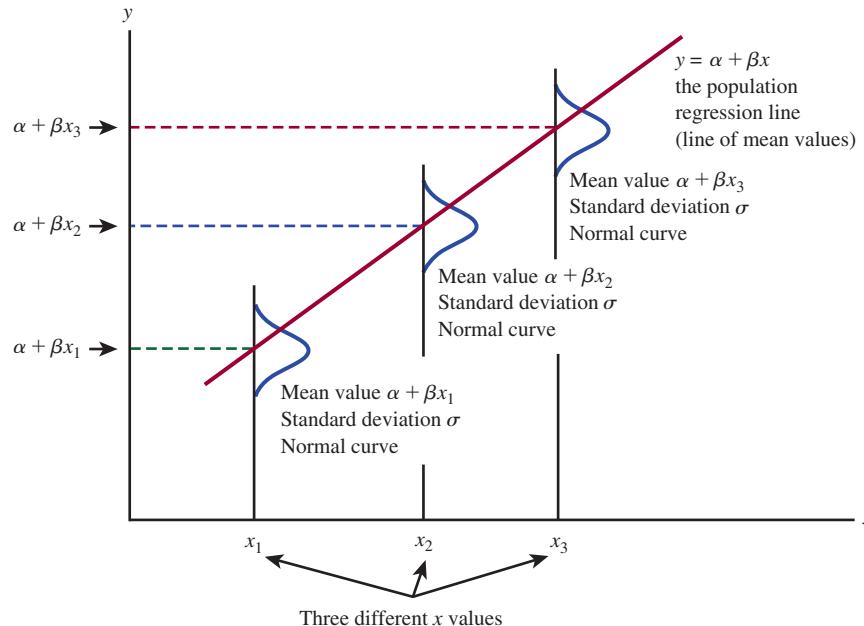
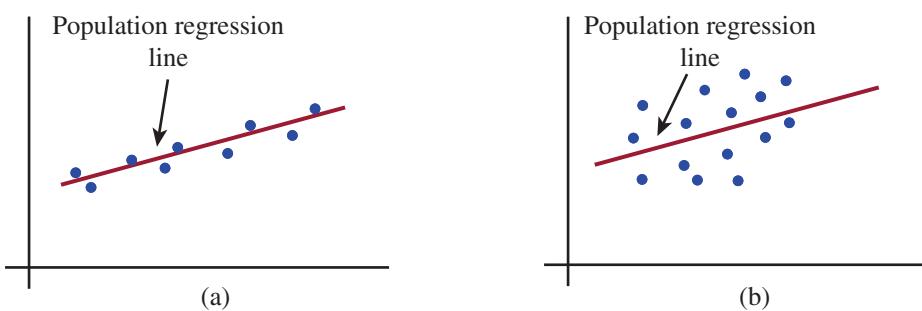


FIGURE 13.4

Data from the simple linear regression model:
 (a) small σ ;
 (b) large σ .



Understand the context ➤



ImageState Royalty Free/Alamy Stock Photo

EXAMPLE 13.1 Stand on Your Head to Lose Weight?

The authors of the article “[On Weight Loss by Wrestlers Who Have Been Standing on Their Heads](#)” ([paper presented at the Sixth International Conference on Statistics, Combinatorics, and Related Areas, Forum for Interdisciplinary Mathematics, 1999](#), with the data also appearing in *A Quick Course in Statistical Process Control*, Mick Norton, Pearson Prentice Hall, 2005) stated that “amateur wrestlers who are overweight near the end of the weight certification period, but just barely so, have been known to stand on their heads for a minute or two, get on their feet, step back on the scale, and establish that they are in the desired weight class. Using a headstand as the method of last resort has become a fairly common practice in amateur wrestling.”

Does this really work? Data were collected in an experiment in which weight loss was recorded for each wrestler after exercising for 15 minutes and then doing a headstand for 1 minute 45 seconds. Based on these data, the authors of the article concluded that there was in fact a demonstrable weight loss that was greater than that for a control group that exercised for 15 minutes but did not do the headstand. (The authors give a plausible explanation for why this might be the case based on the way blood and other body fluids collect in the head during the headstand and the effect of weighing while these fluids are draining immediately after standing.)

The authors also concluded that a simple linear regression model was a reasonable way to describe the relationship between the variables

$$y = \text{Weight loss (in pounds)}$$

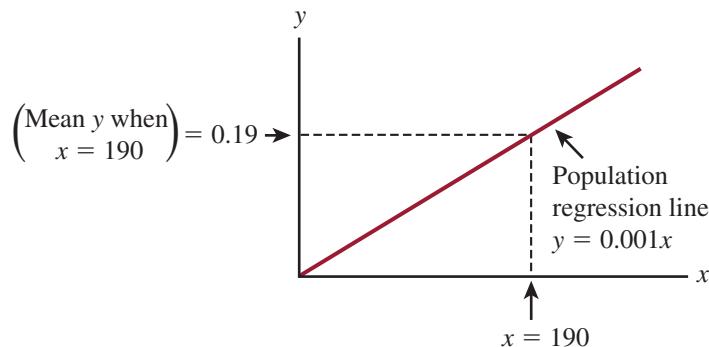
and

$$x = \text{Body weight prior to exercise and headstand (in pounds)}$$

Suppose that the actual model equation has $\alpha = 0$, $\beta = 0.001$, and $\sigma = 0.09$ (these values are consistent with the findings in the article). The population regression line is shown in Figure 13.5.

FIGURE 13.5

The population regression line for Example 13.1.



If the distribution of the random errors at any fixed weight (x value) is normal, then the variable $y = \text{weight loss}$ is normally distributed with

$$\begin{aligned}\mu_y &= \alpha + \beta x = 0 + 0.001x \\ \sigma_y &= \sigma = 0.09\end{aligned}$$

For example, when $x = 190$ (corresponding to a 190-pound wrestler), weight loss has mean value

$$\mu_y = 0 + 0.001(190) = 0.19 \text{ pounds}$$

Because the standard deviation of y is $\sigma = 0.09$, the interval $0.19 \pm 2(0.09) = (0.01, 0.37)$ includes y values that are within 2 standard deviations of the mean value for y when $x = 190$. Roughly 95% of the weight loss observations made for 190-pound wrestlers will be in this range.

The slope $\beta = 0.001$ is the average change in weight loss associated with each additional pound of body weight.

More insight into model properties can be gained by thinking of the population of all (x, y) pairs as consisting of many smaller populations. Each one of these smaller populations contains pairs for which x has a fixed value. For example, suppose that in a large population of college students the relationship between the variables

x = Grade point average in major courses

and

y = Starting salary after graduation

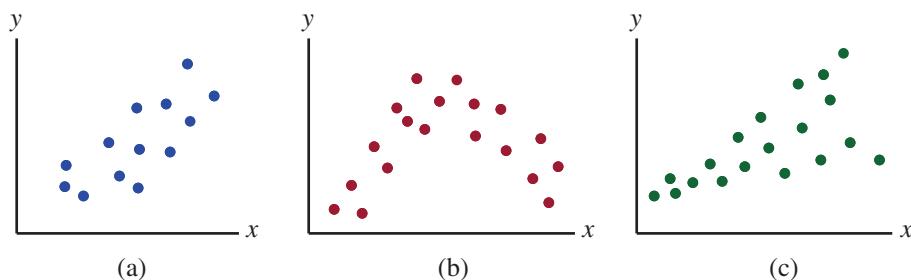
can be described by the simple linear regression model. Then there is the population of all pairs with $x = 3.20$ (corresponding to all students with a grade point average of 3.20 in major courses), the population of all pairs having $x = 2.75$, and so on. The model assumes that for each such population, y is normally distributed with the same standard deviation, and the *mean y value* (rather than y itself) is a linear function of x .

In practice, the initial judgment of whether the simple linear regression model is appropriate is based on how the data were collected and on a scatterplot of the data. The sample observations should be independent of one another. In addition, the scatterplot should show a linear rather than a curved pattern, and the vertical variability of points should be relatively homogeneous throughout the range of x values. Figure 13.6 shows plots with three different patterns, but only the first is consistent with the model assumptions.

FIGURE 13.6

Some commonly encountered patterns in scatterplots:

- (a) consistent with the simple linear regression model;
- (b) suggests a nonlinear probabilistic model;
- (c) suggests that variability in y changes with x .



Estimating the Slope and Intercept of the Population Regression Line

For the remainder of this chapter, we will presume that the basic assumptions of the simple linear regression model are reasonable. The values of α and β (the y intercept and the slope of the population regression line) will almost never be known to an investigator. Instead, these values must be estimated from the sample data.

We will use a and b to denote the point estimates of α and β , respectively. These estimates are based on the method of least squares introduced in Chapter 5. The sum of squared vertical deviations is smaller for the least-squares line than for any other line.

The point estimates of β , the slope, and α , the y intercept of the population regression line, are the slope and y intercept of the least-squares line:

$$b = \text{point estimate of } \beta = \frac{S_{xy}}{S_{xx}}$$

$$a = \text{point estimate of } \alpha = \bar{y} - b\bar{x}$$

where

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} \quad \text{and} \quad S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

The estimated population regression line is the least-squares line

$$\hat{y} = a + bx$$

Suppose x^* denotes a specified value of the predictor variable x . Then $a + bx^*$ has two different interpretations:

1. It is a point estimate of the mean y value when $x = x^*$.
2. It is a point prediction of an individual y value to be observed when $x = x^*$.

Example 13.2 Mother's Age and Baby's Birth Weight

Understand the context ➤

- Medical researchers have noted that adolescent females are much more likely to deliver low-birth-weight babies than are adult females. Because low-birth-weight babies have higher mortality rates, a number of studies have examined the relationship between birth weight and mother's age for babies born to young mothers.

One such study is described in the article “**Body Size and Intelligence in 6-Year-Olds: Are Offspring of Teenage Mothers at Risk?**” (*Maternal and Child Health Journal* [2009]:847–856). The following data on

x = Maternal age (in years)

and

y = Birth weight of baby (in grams)

are consistent with summary values given in the referenced article and also with data published by the National Center for Health Statistics.

Consider the data ➤

	Observation									
	1	2	3	4	5	6	7	8	9	10
x	15	17	18	15	16	19	17	16	18	19
y	2,289	3,393	3,271	2,648	2,897	3,327	2,970	2,535	3,138	3,573

A scatterplot of the data is given in Figure 13.7. The scatterplot shows a linear pattern, and the variability in the y values appears to be similar across the range of x values. This supports the appropriateness of the simple linear regression model.

Do the work ➤

The summary statistics (calculated from the given sample data) are

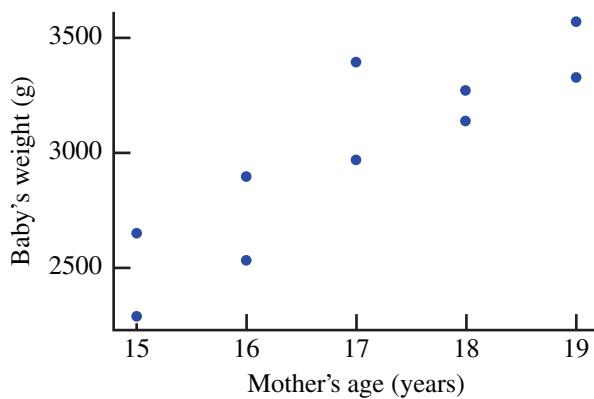
$$n = 10 \quad \Sigma x = 170 \quad \Sigma y = 30,041$$

$$\Sigma x^2 = 2910 \quad \Sigma xy = 515,600 \quad \Sigma y^2 = 91,785,351$$

● Data set available online

FIGURE 13.7

Scatterplot of the data from Example 13.2.



from which

$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 515,600 - \frac{(170)(30,041)}{10} = 4903.0$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 2910 - \frac{(170)^2}{10} = 20.0$$

$$\bar{x} = \frac{170}{10} = 17.0 \quad \bar{y} = \frac{30,041}{10} = 3004.1$$

This gives

$$b = \frac{S_{xy}}{S_{xx}} = \frac{4903.0}{20.0} = 245.15$$

$$a = \bar{y} - b\bar{x} = 3004.1 - (245.1)(17.0) = -1163.45$$

The equation of the estimated regression line is then

$$\hat{y} = a + bx = -1163.45 + 245.15x$$

Interpret the results ➤ A point estimate of the mean birth weight of babies born to 18-year-old mothers results from substituting $x = 18$ into the estimated regression equation:

$$\begin{aligned} (\text{estimated mean } y \text{ when } x = 18) &= a + bx \\ &= -1163.45 + 245.15(18) \\ &= 3249.25 \text{ grams} \end{aligned}$$

Similarly, we would predict the birth weight of a baby to be born to a particular 18-year-old mother to be

$$(\text{predicted } y \text{ value when } x = 18) = a + b(18) = 3249.25 \text{ grams}$$

Interpret concept ➤

The point estimate and the point prediction are identical, because the same x value was used in each calculation. However, the interpretation of each is different. One represents our prediction of the weight of a single baby whose mother is 18, whereas the other represents our estimate of the mean weight of *all* babies born to 18-year-old mothers. This distinction will become important in Section 13.4 (online), when interval estimates and interval predictions are considered.

The least-squares line could have also been estimated using a graphing calculator or a statistical software package. Minitab output for the data of this example is shown on the next page. Notice that Minitab has rounded the values of the estimated coefficients

in the equation of the regression line, which would then result in small differences in predictions based on the line.

Regression Analysis: Birth Weight versus Maternal Age

The regression equation is

$$\text{Birth Weight} = -1163 + 245 \text{ Maternal Age}$$

Predictor	Coef	SE Coef	T	P
Constant	-1163.4	783.1	-1.49	0.176
Maternal Age	245.15	45.91	5.34	0.001
S = 205.308	R-Sq = 78.1%	R-Sq(adj) = 75.4%		

In Example 13.2, the x values in the sample ranged from 15 to 19. An estimate or prediction should not be attempted for any x value much outside this range. Without sample data for such values, there is no evidence that the observed linear relationship continues outside the range from 15 to 19. For this reason, it is not a good idea to make predictions outside this range. Statisticians refer to this potential pitfall as the **danger of extrapolation**.

Estimating σ^2 and σ

The value of σ describes the extent to which observed points (x, y) tend to fall close to or far away from the population regression line. A point estimate of σ is based on

$$\text{SSResid} = \sum(y - \hat{y})^2$$

where $\hat{y}_1 = a + bx_1, \dots, \hat{y}_n = a + bx_n$ are the fitted or predicted y values and the residuals are $y_1 - \hat{y}_1, \dots, y_n - \hat{y}_n$. SSResid is a measure of the extent to which the sample data spread out about the estimated regression line.

The statistic for estimating the variance σ^2 is

$$s_e^2 = \frac{\text{SSResid}}{n - 2}$$

where

$$\text{SSResid} = \sum(y - \hat{y})^2 = \sum y^2 - a \sum y - b \sum xy$$

The subscript e in s_e^2 reminds us that we are estimating the variance of the “errors” or residuals.

The estimate of σ is the **estimated standard deviation**

$$s_e = \sqrt{s_e^2}$$

The number of degrees of freedom associated with estimating σ^2 or σ in simple linear regression is $n - 2$.

The estimates and number of degrees of freedom here have analogs in our previous work involving a single sample x_1, x_2, \dots, x_n . The sample variance s^2 had a numerator of $\sum(x - \bar{x})^2$, a sum of squared deviations (residuals), and denominator $n - 1$, the number of degrees of freedom associated with s^2 and s . The use of \bar{x} as an estimate of μ in the formula for s^2 reduces the number of degrees of freedom by 1, from n to $n - 1$. In simple linear regression, estimation of α and β results in a loss of 2 degrees of freedom, leaving $n - 2$ as the number of degrees of freedom for SSResid, s_e^2 , and s_e .

The coefficient of determination was defined previously (see Chapter 5) as

$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

where

$$\text{SSTo} = \sum(y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = S_{yy}$$

The value of r^2 is interpreted as the proportion of observed variability in y that can be explained by (or attributed to) the model relationship.

The estimate s_e also gives another assessment of model performance. Roughly speaking, the value of σ represents the magnitude of a typical deviation of a point (x, y) in the population from the population regression line. Similarly, s_e is the approximate magnitude of a typical sample deviation (residual) from the least-squares line. The smaller the value of s_e , the closer the points in the sample tend to fall to the line and the better the line does in predicting y from x .

Example 13.3 Predicting Election Outcomes

Understand the context ➤

- The authors of the paper “**Inferences of Competence from Faces Predict Election Outcomes**” (*Science [2005]: 1623–1626*) found that they could successfully predict the outcome of a U.S. congressional election substantially more than half the time based on the facial appearance of the candidates. In the study described in the paper, participants were shown photos of two candidates for a U.S. Senate or House of Representatives election. Each participant was asked to look at the photos and then indicate which candidate he or she thought was more competent.

The two candidates were labeled A and B. If a participant recognized either candidate, data from that participant were not used in the analysis. The proportion of participants who chose candidate A as the more competent was calculated. After the election, the difference in votes (candidate A – candidate B) expressed as a proportion of the total votes cast in the election was also calculated. This difference falls between +1 and –1. It is 0 for an election where both candidates receive the same number of votes, positive for an election where candidate A received more votes than candidate B (with +1 indicating that candidate A received all of the votes), and negative for an election where candidate A received fewer votes than candidate B.

This process was carried out for a large number of congressional races. A subset of the resulting data (approximate values read from a graph that appears in the paper) is given in the accompanying table, which also includes the predicted values and residuals for the least-squares line for these data.

Consider the data ➤

Competent Proportion	Difference in Vote Proportion	Predicted y Value	Residual
0.20	−0.70	−0.389	−0.311
0.23	−0.40	−0.347	−0.053
0.40	−0.35	−0.109	−0.241
0.35	0.18	−0.179	0.359
0.40	0.38	−0.109	0.489
0.45	−0.10	−0.040	−0.060
0.50	0.20	0.030	0.170
0.55	−0.30	0.100	−0.400
0.60	0.30	0.170	0.130
0.68	0.18	0.281	−0.101
0.70	0.50	0.309	0.191
0.76	0.22	0.393	−0.173

The scatterplot (Figure 13.8) suggests a positive linear relationship between

and

y = Difference in vote proportion.

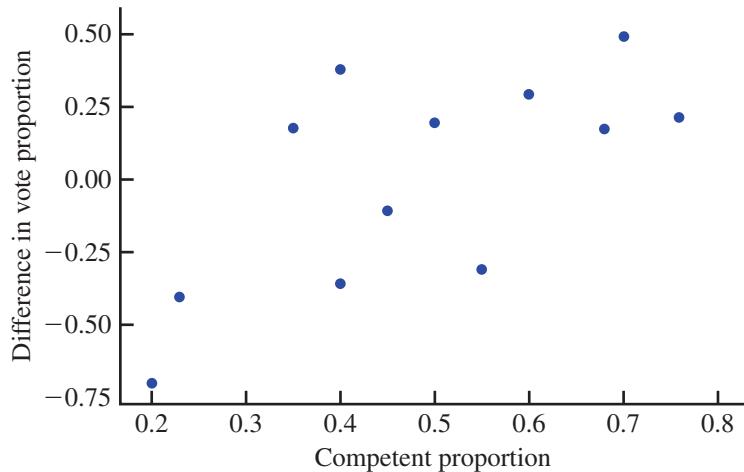
Do the work ➤ The summary statistics are

$$\begin{aligned} n &= 12 & \Sigma x &= 5.82 & \Sigma y &= 0.11 \\ \Sigma x^2 &= 3.1804 & \Sigma xy &= 0.5526 & \Sigma y^2 &= 1.5101 \end{aligned}$$

from which we calculate

$$\begin{aligned} b &= 1.3957 & a &= -0.6678 \\ \text{SSResid} &= 0.81228 & \text{SSTo} &= 1.50909 \\ r^2 &= 1 - \frac{\text{SSResid}}{\text{SSTo}} = 1 - \frac{0.81228}{1.50909} = 1 - 0.538 = 0.462 \\ s_e^2 &= \frac{\text{SSResid}}{n - 2} = \frac{0.81228}{10} = 0.081 \\ s_e &= \sqrt{0.081} = 0.285 \end{aligned}$$

FIGURE 13.8
Minitab scatterplot for Example 13.3.



Interpret the results ➤

Approximately 46.2% of the observed variability in the difference in vote proportion y can be attributed to the linear relationship with proportion of participants who judged the candidate to be more competent based on facial appearance alone. The magnitude of a typical sample deviation from the least-squares line is about 0.285, which is not very large in comparison to the y values themselves. The model may be useful for estimation and prediction. In Section 13.2, we show how a model utility test can be used to judge whether this is actually the case.

A key assumption of the simple linear regression model is that the random deviation e in the model equation is normally distributed. In Section 13.3, we will see how the residuals can be used to determine whether this is plausible.

EXERCISES 13.1 - 13.13

● Data set available online

- 13.1** Let x be the size of a house (in square feet) and y be the amount of natural gas used (therms) during a specified period. Suppose that for a particular community, x and y are related according to the simple linear regression model with

$$\beta = \text{slope of population regression line} = 0.017$$

$$\alpha = \text{y intercept of population regression line} = -5.0$$

Houses in this community range in size from 1000 to 3000 square feet.

- a. What is the equation of the population regression line?
 - b. Graph the population regression line by first finding the point on the line corresponding to $x = 1000$ and then the point corresponding to $x = 2000$, and drawing a line through these points.
- 13.2** Consider the variables and population regression line given in the previous exercise.
- a. What is the mean value of gas usage for houses with 2100 sq. ft. of space?
 - b. What is the average change in usage associated with a 1 sq. ft. increase in size?
 - c. What is the average change in usage associated with a 100 sq. ft. increase in size?
 - d. Should the model be used to predict mean usage for a 500 sq. ft. house? Why or why not? (Hint: See the discussion following Example 13.2.)

- 13.3** The flow rate in a device used for air quality measurement depends on the pressure drop x (inches of water) across the device's filter. Suppose that for x values between 5 and 20, these two variables are related according to the simple linear regression model with population regression line

$$y = -0.12 + 0.095x.$$

- a. What is the mean flow rate for a pressure drop of 10 inches? A drop of 15 inches?
- b. What is the average change in flow rate associated with a 1-inch increase in pressure drop? Explain.

- 13.4** The paper “Predicting Yolk Height, Yolk Width, Albumen Length, Eggshell Weight, Egg Shape Index, Eggshell Thickness, Egg Surface Area of Japanese Quails Using Various Egg Traits as Regressors” (*International Journal of Poultry Science* [2008]: 85–88) suggests that the simple linear regression model is reasonable for describing the relationship between y = Eggshell thickness (in micrometers) and x = Egg length (mm) for quail eggs. Suppose that the population regression line is $y = 0.135 + 0.003x$ and that $\sigma = 0.005$. Then, for a fixed x value, y has a normal distribution with mean $0.135 + 0.003x$ and standard deviation 0.005.

- a. What is the mean eggshell thickness for quail eggs that are 15 mm in length? For quail eggs that are 17 mm in length?
- b. What is the probability that a quail egg with a length of 15 mm will have a shell thickness that is greater than $0.18 \mu\text{m}$?
- c. Approximately what proportion of quail eggs of length 14 mm have a shell thickness of greater than 0.175 ? Less than 0.178 ? (Hint: The distribution of y at a fixed x is approximately normal.)

- 13.5** A sample of small cars was selected, and the values of x = horsepower and y = fuel efficiency (mpg) were determined for each car. Fitting the simple linear regression model gave the estimated regression equation $\hat{y} = 44.0 - 0.150x$.
- a. How would $b = -0.150$ be interpreted?
 - b. Substituting $x = 100$ gives $\hat{y} = 29.0$. Give two different interpretations of this number.
 - c. What happens if the efficiency for a car with a 300-horsepower engine is predicted? Why does this occur?

- 13.6** Use the information given in the previous exercise to answer the following questions.
- a. The value of r^2 was found to be 0.680. Interpret r^2 in the context of this problem.
 - b. The value of s_e was found to be 3.0. Interpret s_e in the context of this problem.

- 13.7** Suppose that a simple linear regression model is appropriate for describing the relationship between y = House price (in dollars) and x = House size (in square feet) for houses in a large city. The population regression line is $y = 23,000 + 47x$ and $\sigma = 5000$.
- a. What is the average change in price associated with one extra square foot of space? With an additional 100 sq. ft. of space?
 - b. Approximately what proportion of 1800 sq. ft. homes would be priced over \$110,000? Under \$100,000?

- 13.8** a. Explain the difference between the line $y = \alpha + \beta x$ and the line $\hat{y} = a + bx$.
- b. Explain the difference between β and b .
 - c. Let x^* denote a particular value of the independent variable. Explain the difference between $\alpha + \beta x^*$ and $a + bx^*$.
 - d. Explain the difference between σ and s_e .

- 13.9** The paper “Depression, Body Mass Index, and Chronic Obstructive Pulmonary Disease—A Holistic Approach” (*International Journal of COPD* [2016]: 239–249) gives data on

$$x = \text{Change in body mass index (BMI in kg/m}^2)$$

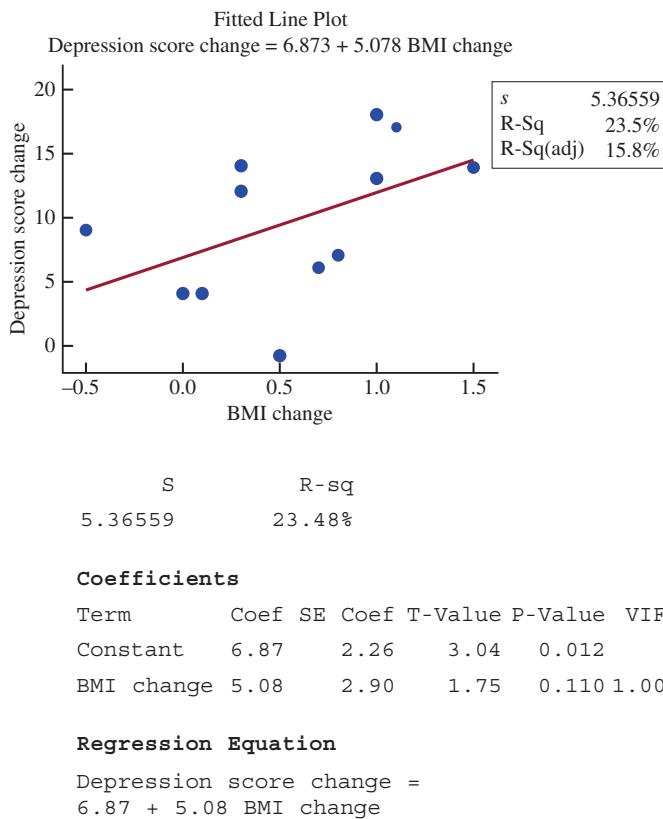
and

$$y = \text{Change in a measure of depression}$$

for patients suffering from depression who participated in a pulmonary rehabilitation program. The table below contains a subset of the data given in the paper and are approximate values read from a scatterplot in the paper.

BMI Change (kg/m ²)	Depression Score Change
0.5	-1
-0.5	9
0.0	4
0.1	4
0.7	6
0.8	7
1.0	13
1.5	14
1.1	17
1.0	18
0.3	12
0.3	14

The accompanying computer output is from Minitab.



- a. What percentage of observed variation in depression score change at age 27 can be explained by the simple linear regression model? (Hint: See Example 13.3.)

- b. Give an estimate of σ and interpret this estimate.
 c. Give an estimate of the average change in depression score change associated with a 1 kg/m² increase in BMI change.
 d. Calculate a point estimate of the mean depression score change for a patient whose BMI change was 1.2 kg/m².

- 13.10** • Hormone replacement therapy (HRT) is thought to increase the risk of breast cancer. The accompanying data on

$$x = \text{Percent of women using HRT}$$

and

$$y = \text{Breast cancer incidence (cases per 100,000 women) for a region in Germany}$$

for 5 years appeared in the paper “Decline in Breast Cancer Incidence after Decrease in Utilisation of Hormone Replacement Therapy” (*Epidemiology* [2008]: 427–430). The authors of the paper used a simple linear regression model to describe the relationship between HRT use and breast cancer incidence.

HRT Use	Breast Cancer Incidence
46.30	103.30
40.60	105.00
39.50	100.00
36.60	93.80
30.00	83.50

- a. What is the equation of the estimated regression line?
 b. What is the estimated average change in breast cancer incidence associated with a 1 percentage point increase in HRT use?
 c. What breast cancer incidence would be predicted in a year when HRT use was 40%?
 d. Should this regression model be used to predict breast cancer incidence for a year when HRT use was 20%? Explain.

- 13.11** Consider the data and estimated regression line from the previous exercise.

- a. Calculate and interpret the value of r^2 .
 b. Calculate and interpret the value of s_e .

- 13.12** A simple linear regression model was used to describe the relationship between y = Hardness of molded plastic and x = Amount of time elapsed since the end of the molding process. Summary quantities included $n = 15$, $\text{SSResid} = 1235.470$, and $\text{SSTo} = 25,321.368$.

- a. Calculate an estimate of σ . What value for degrees of freedom is associated with this estimate?

- b. What percentage of observed variation in hardness can be explained by the linear relationship between hardness and elapsed time?

13.13 ● Consider the accompanying data on

x = Advertising share

and

y = Market share

for a particular brand of soft drink during 10 randomly selected years.

x	0.103	0.072	0.071	0.077	0.086	0.047	0.060	0.050	0.070	0.052
y	0.135	0.125	0.120	0.086	0.079	0.076	0.065	0.059	0.051	0.039

- Construct a scatterplot for these data. Is the simple linear regression model appropriate for describing the relationship between x and y ?
- Calculate the equation of the estimated regression line and use it to obtain the predicted market share when the advertising share is 0.09.
- Calculate the value of r^2 . Interpret this value.
- Calculate a point estimate of σ . What value for degrees of freedom is associated with this estimate?

SECTION 13.2 Inferences About the Slope of the Population Regression Line

The slope coefficient β in the simple linear regression model represents the average or expected change in the dependent variable y that is associated with a 1-unit increase in the value of the independent variable x . For example, consider x = the size of a house (in square feet) and y = selling price of the house. If we assume that the simple linear regression model is appropriate for the population of houses in a particular city, β would be the average increase in selling price associated with a 1-square foot increase in size. As another example, if x = number of hours per week a computer system is used and y = the annual maintenance expense, then β would be the expected change in expense associated with using the computer system one additional hour per week.

Because the value of β is almost always unknown, it must be estimated from the sample data. The slope b of the least-squares line gives a point estimate. As with any point estimate, though, it is desirable to have some indication of how accurately b estimates β .

In some situations, the value of the statistic b may vary greatly from sample to sample, so b calculated from a single sample may be quite different from the actual population slope β . In other situations, almost all possible samples yield b values that are close to β , so the error of estimation is likely to be small.

To proceed further, we need to know about the sampling distribution of b . In particular, we need to know the shape of the sampling distribution, where the sampling distribution is centered relative to β , and how much it spreads out about its center.

Properties of the Sampling Distribution of b

When the four basic assumptions of the simple linear regression model are satisfied, the following statements are true:

- The mean value of b is β . Because $\mu_b = \beta$, the sampling distribution of b is always centered at the value of β . This means that b is an unbiased statistic for estimating β .
- The standard deviation of the statistic b is

$$\sigma_b = \frac{\sigma}{\sqrt{S_{xx}}}$$

- The statistic b has a normal distribution (this is a consequence of the model assumption that the random deviation e is normally distributed).

The fact that b is unbiased means that the sampling distribution is centered at the right place. The standard deviation of b describes variability in the values of b . If σ_b is large, then the sampling distribution of b will be quite spread out around β and an estimate far from the actual value of β could result. For σ_b to be small, the numerator σ should be small

(representing little variability about the population line) and/or the denominator $\sqrt{S_{xx}}$ or, equivalently, $S_{xx} = \sum(x - \bar{x})^2$ should be large. Because $\sum(x - \bar{x})^2$ is a measure of how much the observed x values spread out, β tends to be more precisely estimated when the x values in the sample are spread out rather than when they are close together.

The normality of the sampling distribution of b implies that the standardized variable

$$z = \frac{b - \beta}{\sigma_b}$$

has a standard normal distribution. However, inferential methods cannot be based on this statistic, because the value of σ_b is not known (since the unknown σ appears in the numerator of σ_b). One way to proceed is to estimate σ with s_e .

The estimated standard deviation of the statistic b is

$$s_b = \frac{s_e}{\sqrt{S_{xx}}}$$

When the four basic assumptions of the simple linear regression model are satisfied, the statistic

$$t = \frac{b - \beta}{s_b}$$

has a t distribution with $df = (n - 2)$.

In the same way that $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ was used in Chapter 9 to develop a confidence interval for μ , the t variable in the preceding box can be used to obtain a confidence interval for β .

Confidence Interval for β

When the four basic assumptions of the simple linear regression model are satisfied, a **confidence interval for β** , the slope of the population regression line, has the form

$$b \pm (t \text{ critical value}) \cdot s_b$$

where the t critical value is based on $df = (n - 2)$. Appendix Table 3 gives critical values corresponding to the most frequently used confidence levels.

The interval estimate of β is centered at b and extends out from the center by an amount that depends on the sampling variability of b . When s_b is small, the interval is narrow, implying relatively precise knowledge of the value of β .

Example 13.4 The Bison of Yellowstone Park

Understand the context ➤

- The dedicated work of conservationists for over 100 years has brought the bison in Yellowstone National Park from near extinction to a herd of over 3000 animals. This recovery is a mixed blessing. Many bison have been exposed to the bacteria that cause brucellosis, a disease that infects domestic cattle, and there are many domestic cattle herds near Yellowstone. Because of concerns that free-ranging bison can infect nearby cattle, it is important to monitor and manage the size of the bison population, and if possible, keep bison from transmitting this bacteria to ranch cattle.

● Data set available online

The article “**Reproduction and Survival of Yellowstone Bison**” (*The Journal of Wildlife Management* [2007]: 2365–2372) describes a large multiyear study of the factors that influence bison movement and herd size. The researchers studied a number of environmental factors to better understand the relationship between bison reproduction and the environment. One factor thought to influence reproduction is stress due to accumulated snow, making foraging more difficult for pregnant bison. Data on

$$y = \text{Spring calf ratio (SCR)}$$

and

$$x = \text{Previous fall snow-water equivalent (SWE)}$$

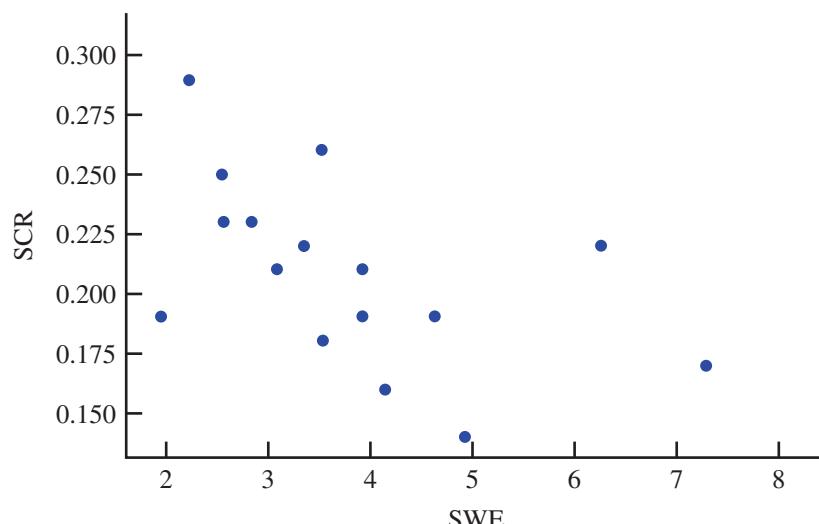
for 17 years are shown in the accompanying table. The spring calf ratio is the ratio of calves to adults, a measure of reproductive success. Snow water equivalent is the depth of water that would result if the snow pack melted. The SWE measurements in the table are in thousands of centimeters (so 1.933 represents 1933 cm). The researchers were interested in estimating the mean change in spring calf ratio associated with each additional 1000 centimeters in snow-water equivalent.

Consider the data ►

SCR	SWE	SCR	SWE
0.19	1.933	0.22	3.317
0.14	4.906	0.22	3.332
0.21	3.072	0.18	3.511
0.23	2.543	0.21	3.907
0.26	3.509	0.25	2.533
0.19	3.908	0.19	4.611
0.29	2.214	0.22	6.237
0.23	2.816	0.17	7.279
0.16	4.128		

To proceed, we would need to assume that these years are representative of yearly circumstances at Yellowstone, and that each year’s reproduction and snowfall is independent of previous years.

A scatterplot of the data is shown here.



The plot shows a linear pattern, and the vertical variability of points does not appear to be changing over the range of x values in the sample. If we assume that the distribution of errors at any given x value is approximately normal, then the simple linear regression model would be appropriate.

The slope β in this context is the average change in spring calf ratio associated with each additional 1000 cm in snow water equivalent. The scatterplot shows a negative linear relationship, so the point estimate of β will be negative.

Do the work ➤

Based on the given data

$$n = 17 \quad \Sigma x = 63.756 \quad \Sigma y = 3.56$$

$$\Sigma x^2 = 270.425 \quad \Sigma xy = 12.923 \quad \Sigma y^2 = 0.7682$$

from which

$$b = -0.0137 \quad a = 0.261$$

$$\text{SSResid} = 0.016847 \quad \text{SSTo} = 0.022694$$

$r^2 = 0.258$ (25.8% of the observed variability in spring calf ratio can be explained by the simple linear regression model)

$$s_e^2 = 0.0011 \quad s_e = 0.0335$$

$$s_b = \frac{s_e}{\sqrt{S_{xx}}} = \frac{0.0335}{\sqrt{31.3175}} = 0.006$$

Calculation of the 95% confidence interval for β requires a t critical value based on $df = n - 2 = 17 - 2 = 15$, which (from Appendix Table 3) is 2.13. The resulting interval is then

$$\begin{aligned} b \pm (t \text{ critical value}) \cdot s_b &= -0.0137 \pm (2.13)(0.006) \\ &= -0.0137 \pm 0.0128 \\ &= (-0.027, -0.001) \end{aligned}$$

Interpret the results ➤

We interpret this interval as follows: Based on the sample data, we are 95% confident that the average change in spring calf ratio associated with a 1000-cm increase in snow-water equivalent is between -0.027 and -0.001 . This means that we think that spring calf ratio decreases by somewhere between 0.001 and 0.027 with each additional 1000 cm of snow-water equivalent.

Output from any of the standard statistical software packages routinely includes the calculated values of a , b , SSResid, SSTo, and s_b . Figure 13.9 (on the next page) displays partial Minitab output for the data of Example 13.4. The format from other software packages is similar. Rounding occasionally leads to small discrepancies between hand-calculated and computer-calculated values, but there are no such discrepancies in this example.

Hypothesis Tests Concerning β

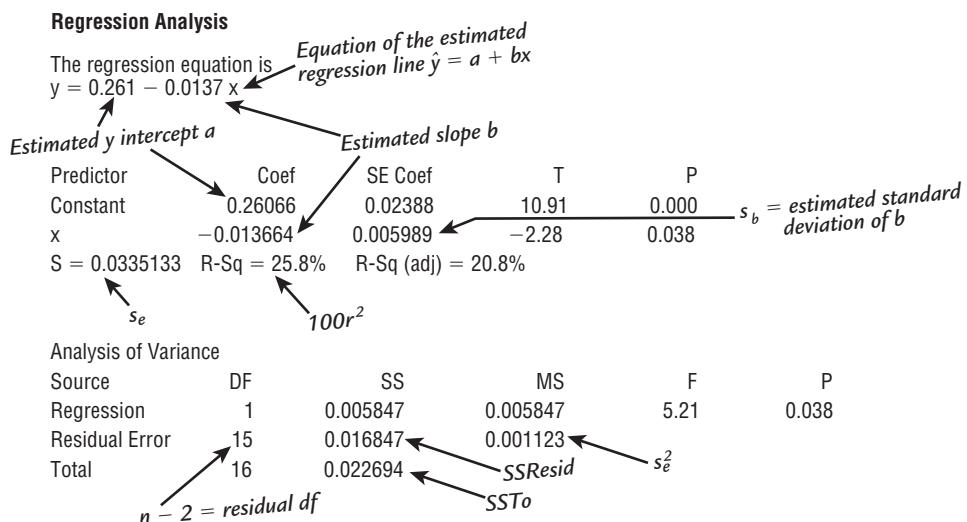
Hypotheses about β can be tested using a t test similar to the t tests introduced in Chapters 10 and 11. The null hypothesis states that β has a specified hypothesized value. The t statistic results from standardizing b , the point estimate of β , under the assumption that H_0 is true. When H_0 is true, the sampling distribution of this statistic is the t distribution with $df = (n - 2)$.

Frequently, the null hypothesis of interest is $\beta = 0$. If this is the case, the population regression line is a horizontal line, and the value of y in the simple linear regression model does not depend on x . That is,

$$y = \alpha + 0 \cdot x + e$$

FIGURE 13.9

Partial Minitab output for the data of Example 13.4.



Summary of Hypothesis Tests Concerning β

Null hypothesis: $H_0: \beta = \text{hypothesized value}$

Test statistic: $t = \frac{b - \text{hypothesized value}}{s_b}$

The test is based on $df = (n - 2)$.

Alternative hypothesis: **P-value:**

$H_a: \beta > \text{hypothesized value}$ Area to the right of the calculated t under the appropriate t curve

$H_a: \beta < \text{hypothesized value}$ Area to the left of the calculated t under the appropriate t curve

$H_a: \beta \neq \text{hypothesized value}$ (1) 2(area to the right of t) if t is positive or
(2) 2(area to the left of the t) if t is negative

Assumptions: For this test to be appropriate, the four basic assumptions of the simple linear regression model must be satisfied:

1. The distribution of e at any particular x value has mean value 0 (that is $\mu_e = 0$).
2. The standard deviation of e is σ , which does not depend on x .
3. The distribution of e at any particular x value is normal.
4. The random deviations e_1, e_2, \dots, e_n associated with different observations are independent of one another.

or equivalently,

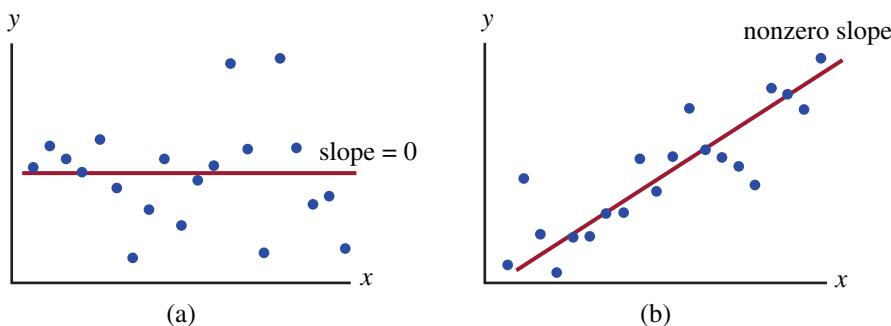
$$y = \alpha + e$$

In this situation, knowledge of x is not useful for predicting y . On the other hand, if $\beta \neq 0$, there is a linear relationship between x and y , and knowledge of x is useful for predicting y . This is illustrated by the scatterplots in Figure 13.10.

The test of $H_0: \beta = 0$ versus $H_a: \beta \neq 0$ is called the model utility test for simple linear regression.

FIGURE 13.10

- (a) $\beta = 0$;
 (b) $\beta \neq 0$.



The Model Utility Test for Simple Linear Regression

The **model utility test for simple linear regression** is the test of

$$H_0: \beta = 0$$

versus

$$H_a: \beta \neq 0$$

The null hypothesis specifies that there is *no* useful linear relationship between x and y , whereas the alternative hypothesis specifies that there *is* a useful linear relationship between x and y . If H_0 is rejected, we conclude that the simple linear regression model is useful for predicting y .

The test procedure in the previous box (with hypothesized value = 0) is used to carry out the model utility test. The test statistic is $t = \frac{b}{s_b}$.

If a scatterplot and the r^2 value do not provide clear evidence for a useful linear relationship, we recommend that the model utility test be carried out before using the regression line to make predictions.

Example 13.5 University Graduation Rates

- The accompanying data on 6-year graduation rate (%), student-related expenditure per full-time student, and median SAT score for a random sample of the primarily undergraduate public universities and colleges in the United States with enrollments between 10,000 and 20,000 were taken from [College Results Online, The Education Trust](#).

Median SAT	Expenditure	Graduation Rate
1,065	7,970	49
950	6,401	33
1,045	6,285	37
990	6,792	49
950	4,541	22
970	7,186	38
980	7,736	39
1,080	6,382	52
1,035	7,323	53
1,010	6,531	41
1,010	6,216	38
930	7,375	37
1,005	7,874	45
1,090	6,355	57
1,085	6,261	48

● Data set available online

Let's first investigate the relationship between graduation rate and median SAT score. With y = graduation rate and x = median SAT score, summary statistics are as follows:

$$\begin{array}{lll} n = 15 & \Sigma x = 15,195 & \Sigma y = 638 \\ \Sigma x^2 = 15,430,725 & \Sigma xy = 651,340 & \Sigma y^2 = 28,294 \end{array}$$

from which

$$\begin{array}{lll} b = 0.132 & a = -91.31 & \text{SSResid} = 491.01 \\ s_e = 6.146 & r^2 = 0.576 & S_{xx} = 38,190 \end{array}$$

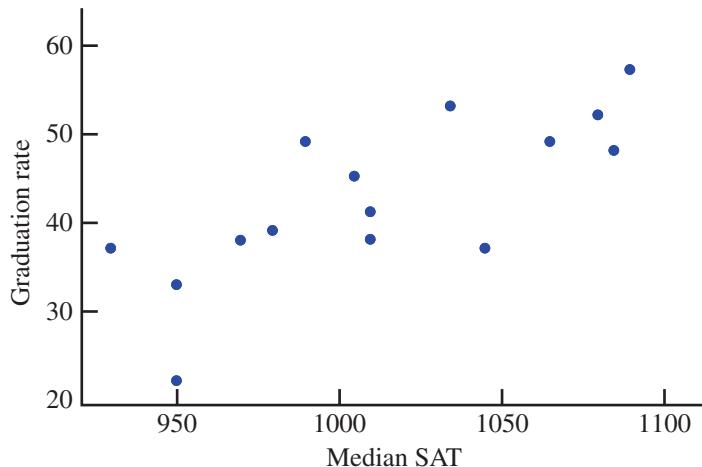
Because $r^2 = 0.576$, about 57.6% of observed variation in graduation rates can be explained by the simple linear regression model. This suggests that there is a useful linear relation between the two variables, and this can be confirmed by the model utility test. We will use a significance level of 0.05 to carry out this test.

Understand the context ►

1. β = the average change in graduation rate associated with an increase of 1 unit in median SAT score
2. $H_0: \beta = 0$
3. $H_a: \beta \neq 0$
4. $\alpha = 0.05$

Formulate a plan ►

5. Test statistic: $t = \frac{b - \text{hypothesized value}}{s_b} = \frac{b - 0}{s_b} = \frac{b}{s_b}$
6. Assumptions: The data are from a random sample, so the observations are independent. The accompanying scatterplot of the data shows a linear pattern and the vertical variability of points does not appear to be changing with x :



Assuming that the distribution of errors at any given x value is approximately normal, the assumptions of the simple linear regression model are satisfied.

Do the work ►

7. Calculation: The calculation of t requires

$$s_b = \frac{s_e}{\sqrt{S_{xx}}} = \frac{6.146}{\sqrt{38,190}} = 0.031$$

yielding

$$t = \frac{0.132 - 0}{0.031} = 4.26$$

8. *P*-value: Appendix Table 4 shows that for a *t* test with 13 df, $P(t > 4.26) < 0.001$. The inequality in H_a requires a two-tailed test, so
- $$P\text{-value} < 2(0.001) = 0.002.$$

Interpret the results ➤

9. Conclusion: Since P -value < 0.002 is smaller than the significance level 0.05, H_0 is rejected. We conclude that there is a useful linear relationship between graduation rate and median SAT score.

Figure 13.11 shows partial Minitab output from a simple linear regression analysis. The Coef column gives $b = 0.13213$; $s_b = 0.03145$ is in the SE Coef column; the T column (for *t* ratio) contains the value of the test statistic for testing $H_0: \beta = 0$; and the *P*-value for the model utility test is given in the last column as 0.001 (slightly different from the values given in Step 8 because of rounding and because the use of the table produces only approximate *P*-values). Other commonly used statistical software packages also include this information in their output.

FIGURE 13.11

Minitab output for the data of Example 13.5.

Regression Analysis: Graduation Rate versus Median SAT

The regression equation is

$$\text{Graduation Rate} = -91.3 + 0.132 \text{ Median SAT}$$

Predictor	Coef	SE Coef	T	P
Constant	-91.31	31.90	-2.86	0.013
Median SAT	0.13213	0.03145	4.20	0.001

$$S = 6.14574 \quad R\text{-Sq} = 57.6\% \quad R\text{-Sq(adj)} = 54.3\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	666.72	666.72	17.65	0.001
Residual Error	13	491.01	37.77		
Total	14	1157.73			

Let's next consider the relationship between graduation rate and expenditure per full-time student. Figure 13.12 shows partial Minitab output from a simple linear regression with expenditure as the predictor variable.

FIGURE 13.12

Minitab output using expenditure as the predictor.

The regression equation is

$$\text{Graduation Rate} = 10.9 + 0.00468 \text{ Expenditure}$$

Predictor	Coef	SE Coef	T	P
Constant	10.95	17.51	0.63	0.543
Expenditure	0.004680	0.002574	1.82	0.092

$$S = 8.42608 \quad R\text{-Sq} = 20.3\% \quad R\text{-Sq(adj)} = 14.1\%$$

The value of the test statistic for the model utility test in this case is $t = 1.82$ and the associated *P*-value is 0.092. For a 0.05 level of significance, we would not reject the null hypothesis of $H_0: \beta = 0$. There is not convincing evidence of a linear relationship between graduation rate and expenditure per full-time student.

When $H_0: \beta = 0$ cannot be rejected by the model utility test at a reasonably small significance level, other models might be considered. One possibility is to use a nonlinear model—an appropriate strategy if the scatterplot shows curvature. Alternatively, a multiple regression model using more than one predictor variable could be considered. Multiple regression models are introduced in Chapter 14.

EXERCISES 13.14 - 13.30

- 13.14** What is the difference between σ and σ_b ? What is the difference between σ_b and s_b ?
- 13.15** The largest commercial fishing enterprise in the southeastern United States is the harvest of shrimp. In a study described in the paper “[Long-term Trawl Monitoring of White Shrimp, *Litopenaeus setiferus* \(Linnaeus\), Stocks within the ACE Basin National Estuarine Research Reserve, South Carolina](#)” (*Journal of Coastal Research* [2008]: 193–199), researchers monitored variables thought to be related to the abundance of white shrimp.
- One variable the researchers thought might be related to abundance is the amount of oxygen in the water. The relationship between the mean number of white shrimp caught in a single outing and oxygen saturation was described by fitting a regression line using data from 10 randomly selected offshore sites. Computer output is shown below.
- The regression equation is
Mean number caught = - 5859 + 97.2 Saturation
- | Predictor | Coef | SE Coef | T | P |
|------------|-------|---------|-------|-------|
| Constant | -5859 | 2394 | -2.45 | 0.040 |
| Saturation | 97.22 | 34.63 | 2.81 | 0.023 |
- $S = 481.632$ R-Sq = 49.6% R-Sq(adj) = 43.3%
- a. Is there convincing evidence of a useful linear relationship between the mean number of shrimp caught and oxygen saturation? Explain. (Hint: See Example 13.5.)
- b. Is the relationship strong? Explain why or why not.
- c. Construct a 95% confidence interval for β and interpret it in context. (Hint: See Example 13.4.)
- 13.16** Refer back to Example 13.3 in which the simple linear regression model was fit to data on x = Proportion who judged candidate A as more competent and y = Vote difference proportion. For the purpose of estimating β as accurately as possible, would it have been preferable to have observations with x values 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95 and 0.98? Explain.
- 13.17** Exercise 13.12 summarized data on y = Hardness of molded plastic and x = Time elapsed since the molding was completed.
- Summary quantities included
- $$n = 15 \quad b = 2.50 \quad \text{SSResid} = 1235.470$$
- $$\Sigma(x - \bar{x})^2 = 4024.20$$
- a. Calculate the estimated standard deviation of the statistic b .
- 13.18** A simple linear regression model was used to describe the relationship between sales revenue y (in thousands of dollars) and advertising expenditure x (also in thousands of dollars) for fast-food outlets during a 3-month period. A random sample of 15 outlets resulted in the accompanying summary quantities.
- $$\Sigma x = 14.10 \quad \Sigma y = 1438.50 \quad \Sigma x^2 = 13.92$$
- $$\Sigma y^2 = 140,354 \quad \Sigma xy = 1387.20$$
- $$\Sigma(y - \bar{y})^2 = 2401.85 \quad \Sigma(y - \hat{y})^2 = 561.46$$
- a. What proportion of observed variation in sales revenue can be attributed to the linear relationship between revenue and advertising expenditure?
- b. Calculate s_e and s_b .
- c. Calculate a 90% confidence interval for β , the average change in revenue associated with a \$1000 (that is, 1 unit) increase in advertising expenditure.
- 13.19** An experiment to study the relationship between x = Time spent exercising (minutes) and y = Amount of oxygen consumed during the exercise period resulted in the following summary statistics.
- $$n = 20 \quad \Sigma x = 50 \quad \Sigma y = 16,705 \quad \Sigma x^2 = 150$$
- $$\Sigma y^2 = 14,194,231 \quad \Sigma xy = 44,194$$
- a. Estimate the slope and y intercept of the population regression line.
- b. One sample observation on oxygen usage was 757 for a 2-minute exercise period. What amount of oxygen consumption is predicted for this exercise period, and what is the corresponding residual?
- c. Calculate a 99% confidence interval for the average change in oxygen consumption associated with a 1-minute increase in exercise time.
- 13.20** The paper “[The Effects of Split Keyboard Geometry on Upper Body Postures](#)” (*Ergonomics* [2009]: 104–111) describes a study to determine the effects of several keyboard characteristics on typing speed. One of the variables considered was the front-to-back surface angle of the keyboard. Minitab output resulting from fitting the simple linear regression model with x = Surface angle (degrees) and y = Typing speed (words per minute) is given below.

● Data set available online

Regression Analysis: Typing Speed versus Surface Angle

The regression equation is

$$\text{Typing Speed} = 60.0 + 0.0036 \text{ Surface Angle}$$

Predictor	Coef	SE Coef	T	P
Constant	60.0286	0.2466	243.45	0.000
Surface Angle	0.00357	0.03823	0.09	0.931
S	0.511766	R-Sq = 0.3%	R-Sq(adj) = 0.0%	

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.0023	0.0023	0.01	0.931
Residual Error	3	0.7857	0.2619		
Total	4	0.7880			

- Assuming that the basic assumptions of the simple linear regression model are reasonably met, carry out a hypothesis test to decide if there is a useful linear relationship between x and y . (Hint: See Example 13.5.)
- Are the values of s_e and r^2 consistent with the conclusion from Part (a)? Explain.

- 13.21** The authors of the paper “Decreased Brain Volume in Adults with Childhood Lead Exposure” (*Public Library of Science Medicine* [May 27, 2008]: e112) studied the relationship between childhood environmental lead exposure and a measure of brain volume change in a particular region of the brain. Data were given for x = Mean childhood blood lead level ($\mu\text{g}/\text{dL}$) and y = Brain volume change (BVC, in percent). A subset of data read from a graph that appeared in the paper was used to produce the accompanying Minitab output.

Regression Analysis: BVC versus Mean Blood Lead Level

The regression equation is

$$\text{BVC} = -0.00179 - 0.00210 \text{ Mean Blood Lead Level}$$

Predictor	Coef	SE Coef	T	P
Constant	-0.001790	0.008303	-0.22	0.830
Mean Blood Lead Level	-0.0021007	0.0005743	-3.66	0.000

Carry out a hypothesis test to decide if there is convincing evidence of a useful linear relationship between x and y . Assume that the basic assumptions of the simple linear regression model are reasonably met.

- 13.22** Do taller adults make more money? The authors of the paper “Stature and Status: Height, Ability, and Labor Market Outcomes” (*Journal of Political Economics* [2008]: 499–532) investigated the association between height and earnings. They used the simple linear regression model to describe the relationship between

x = Height (in inches) and y = log(Weekly gross earnings in dollars) in a very large sample of men. The logarithm of weekly gross earnings was used because this transformation resulted in a relationship that was approximately linear.

The paper reported that the slope of the estimated regression line was $b = 0.023$ and the standard deviation of b was $s_b = 0.004$. Carry out a hypothesis test to decide if there is convincing evidence of a useful linear relationship between height and the logarithm of weekly earnings. Assume that the basic assumptions of the simple linear regression model are reasonably met.

- 13.23** Researchers studying pleasant touch sensations measured the firing frequency (impulses per second) of nerves that were stimulated by a light brushing stroke on the forearm and also recorded the subject’s numerical rating of how pleasant the sensation was. The accompanying data was read from a graph in the paper “Coding of Pleasant Touch by Unmyelinated Afferents in Humans” (*Nature Neuroscience*, April 12, 2009).

Firing Frequency	Pleasantness Rating
23	0.2
24	1.0
22	1.2
25	1.2
27	1.0
28	2.0
34	2.3
33	2.2
36	2.4
34	2.8

- Estimate the mean change in pleasantness rating associated with an increase of 1 impulse per second in firing frequency using a 95% confidence interval. Interpret the resulting interval.
- Carry out a hypothesis test to decide if there is convincing evidence of a useful linear relationship between firing frequency and pleasantness rating.

- 13.24** The accompanying data are a subset of data from the report “Great Jobs, Great Lives” (*Gallup-Purdue Index 2015 Report*, gallup.com/reports/197144/gallup-purdue-index-report-2015.aspx, retrieved May 27, 2017). The data are approximate values read from a scatterplot. University students were asked if they agreed that their education was worth the cost. One variable in the table is the *U.S. News and World Report* ranking of the university in 2015. The other variable in the table is the percentage of students at the university who responded “strongly agree.”

University Ranking	Percentage of Alumni Who Strongly Agree
28	53
29	58
30	62
37	55
45	54
47	62
52	55
54	62
57	70
60	58
65	66
66	55
72	65
75	57
82	67
88	59
98	75

- 13.25** Acrylamide is a chemical that is sometimes found in cooked starchy foods and which is thought to increase the risk of certain kinds of cancer. The paper “A Statistical Regression Model for the Estimation of Acrylamide Concentrations in French Fries for Excess Lifetime Cancer Risk Assessment” (*Food and Chemical Toxicology* [2012]: 3867–3876) describes a study to investigate the effect of frying time (in seconds) and acrylamide concentration (in micrograms per kilogram) in French fries. The data in the accompanying table are approximate values read from a graph that appeared in the paper.

Frying Time	Acrylamide Concentration
150	155
240	120
240	190
270	185
300	140
300	270

- a. For these data, the estimated regression line for predicting $y = \text{Acrylamide concentration}$ based on $x = \text{Frying time}$ is $\hat{y} = 87 + 0.359x$. What is an estimate of the average change in acrylamide

concentration associated with a 1-second increase in frying time?

- b. What would you predict for acrylamide concentration for a frying time of 250 seconds?
 c. Use the given Minitab output to decide if there is convincing evidence of a useful linear relationship between acrylamide concentration and frying temperature.

S	R-sq	R-sq(adj)	R-sq(pred)
54.7108	14.38%	0.00%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	87	112	0.78	0.480	
x	0.359	0.438	0.82	0.459	1.00

Regression Equation

$$y = 87 + 0.359x$$

- 13.26** The paper referenced in Exercise 13.9 (“Depression, Body Mass Index, and Chronic Obstructive Pulmonary Disease—A Holistic Approach,” *International Journal of COPD* [2016]: 239–249) gave data on

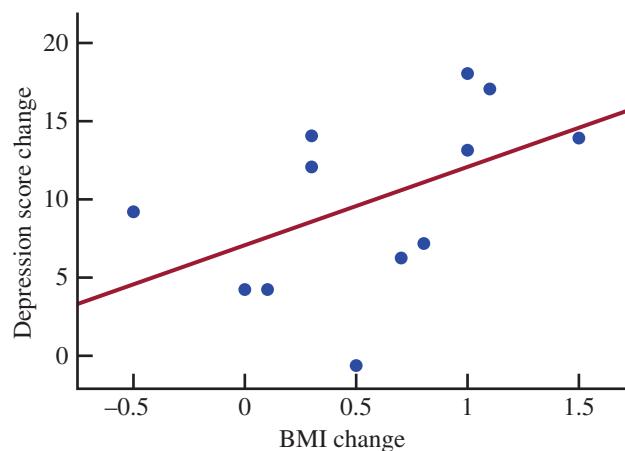
x = Change in body mass index (BMI in kg/m^2)

and

y = Change in a measure of depression

for patients suffering from depression who participated in a pulmonary rehabilitation program. JMP output for these data is shown below.

Bivariate Fit of Depression score change By BMI Change



Linear Fit

$$\text{Depression score change} = 6.8725681 + 5.077821 * \text{BMI Change}$$

Summary of Fit

RSquare	0.234828
RSquare Adj	0.158311
Root Mean Square Error	5.365593
Mean of Response	9.75
Observations (or Sum Wgts)	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	88.35409	88.3541	3.0690
Error	10	287.89591	28.7896	Prob > F
C. Total	11	376.25000		0.1104

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob > t
Intercept	6.8725681	2.257651	3.04	0.0124*
BMI Change	5.077821	2.898557	1.75	0.1104

- a. What does the scatterplot suggest about the relationship between depression score change and BMI change?
- b. What is the equation of the estimated regression line?
- c. Is there a useful linear relationship between the two variables? Carry out an appropriate test using a significance level of $\alpha = 0.05$.

13.27 Exercise 13.18 described a regression analysis with y = Sales revenue and x = Advertising expenditure. Summary quantities given there result in

$$n = 15 \quad b = 52.27 \quad s_b = 8.05$$

- a. Test the hypothesis $H_0: \beta = 0$ versus $H_a: \beta \neq 0$ using a significance level of 0.05. What does the conclusion say about the nature of the relationship between x and y ?
- b. Consider the hypothesis $H_0: \beta = 40$ versus $H_a: \beta > 40$. The null hypothesis states that the average change in sales revenue associated with a 1-unit increase in advertising expenditure is (at most) \$40,000. Carry out a test using significance level 0.01.

13.28 ● Consider the accompanying data on x = Research and development expenditure (thousands of dollars)

and y = Growth rate (% per year) for eight different industries.

x 2,024 5,038 905 3,572 1,157 327 378 191
 y 1.90 3.96 2.44 0.88 0.37 -0.90 0.49 1.01

- a. Would a simple linear regression model provide useful information for predicting growth rate from research and development expenditure? Test the appropriate hypotheses using a 0.05 significance level.
- b. Use a 90% confidence interval to estimate the average change in growth rate associated with a \$1000 increase in expenditure. Interpret the resulting interval.

13.29 Suppose that a single y observation is made at each of the x values 5, 10, 15, 20, and 25.

- a. If $\sigma = 4$, what is the standard deviation of the statistic b ?
- b. Now suppose that a second observation is made at every x value listed in Part (a) (for a total of 10 observations). Is the resulting value of σ_b half of what it was in Part (a)?
- c. How many observations at each x value in Part (a) are required to yield a σ_b value that is half the value calculated in Part (a)?

13.30 ● In anthropological studies, an important characteristic of fossils is cranial capacity. Frequently skulls are at least partially decomposed, so it is necessary to use other characteristics to obtain information about cranial capacity. One measure that has been used is the length of the lambda-opisthion chord. The article "[Vertesszollos and the Presapiens Theory](#)" (*American Journal of Physical Anthropology* [1971]) reported the accompanying data for $n = 7$ *Homo erectus* fossils.

x (Chord length in mm) 78 75 78 81 84 86 87
 y (Capacity in cm^3) 850 775 750 975 915 1,015 1,030

Suppose that from previous evidence, anthropologists had believed that for each 1-mm increase in chord length, cranial capacity would be expected to increase by 20 cm^3 . Do the data given here strongly contradict prior belief?

SECTION 13.3 | Checking Model Adequacy

The simple linear regression model equation is

$$y = \alpha + \beta x + e$$

where e represents the random deviation of an observed y value from the population regression line $\alpha + \beta x$. The inferential methods presented in Section 13.2 required some assumptions about e , including these:

1. At any particular x value, the distribution of e is a normal distribution.

2. At any particular x value, the standard deviation of e is σ , which is the same for all values of x (meaning that σ does not depend on x).

Inferences based on the simple linear regression model continue to be reliable when model assumptions are slightly violated (for example, mild nonnormality of the random deviation distribution). However, using an estimated model in the face of substantially violated assumptions can result in misleading conclusions. In this section, we consider methods for identifying such serious violations of model assumptions.

Residual Analysis

If the deviations e_1, e_2, \dots, e_n from the population line were available, they could be examined for any inconsistencies with model assumptions. For example, a normal probability plot would suggest whether or not normality was reasonable. But, because

$$\begin{aligned}e_1 &= y_1 - (\alpha + \beta x_1) \\&\vdots \\e_n &= y_n - (\alpha + \beta x_n)\end{aligned}$$

these deviations can be calculated only if the equation of the population line is known. In practice, this is not the case. Instead, diagnostic checks are based on the residuals

$$\begin{aligned}y_1 - \hat{y}_1 &= y_1 - (a + b x_1) \\&\vdots \\y_n - \hat{y}_n &= y_n - (a + b x_n)\end{aligned}$$

which are the deviations from the *estimated* line.

When all model assumptions are met, the mean value of the residuals at any particular x value is 0. Any observation that gives a large positive or negative residual should be examined carefully for any unusual circumstances, such as a recording error or exceptional experimental conditions. Identifying residuals with unusually large magnitudes is made easier by considering the **standardized residuals**.

Recall that a quantity is standardized by subtracting its mean value (0 in this case) and dividing by its estimated standard deviation. So, to obtain standardized residuals, we calculate

$$\text{standardized residual} = \frac{\text{residual}}{\text{estimated standard deviation of residual}}$$

The value of a standardized residual tells how many standard deviations the corresponding residual is from its expected value, 0.

Because residuals at different x values have different standard deviations* (depending on the value of x for that observation), calculating the standardized residuals can be tedious. Fortunately, many statistical software packages provide standardized residuals as part of the output.

In Chapter 7, the normal probability plot was introduced as a technique for deciding whether the n observations in a random sample could plausibly have come from a normal population distribution. To assess whether the assumption that e_1, e_2, \dots, e_n all come from the same normal distribution is reasonable, we can construct a normal probability plot of the standardized residuals. This is illustrated in the following example.

*The estimated standard deviation of the i th residual, $(y_i - \hat{y}_i)$ is $s_e \sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}$.

Example 13.6 Political Faces

Understand the context ➤ Example 13.3 introduced data on

x = Proportion who judged candidate A as the more competent of two candidates based on facial appearance

and

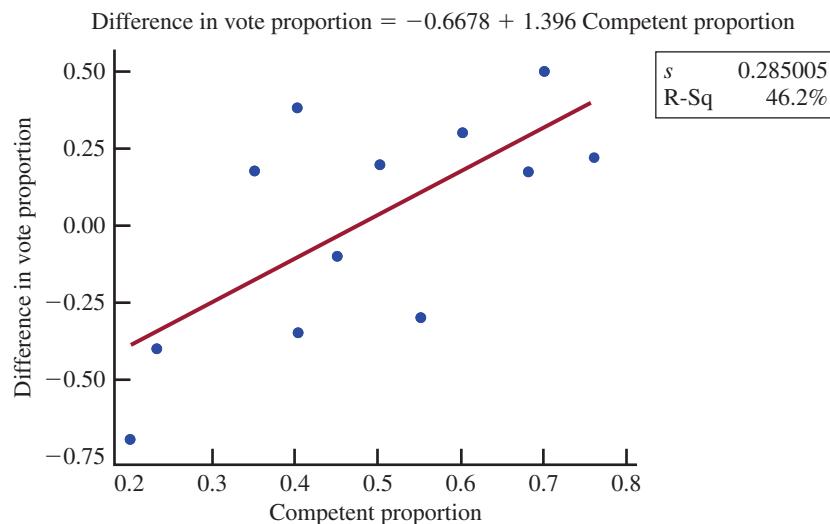
y = Vote difference (candidate A – candidate B) expressed as a proportion of the total number of votes cast

for a sample of 12 congressional elections. (See Example 13.3 for a more detailed description of the study.)

The scatterplot in Figure 13.13 is consistent with the assumptions of the simple linear regression model.

FIGURE 13.13

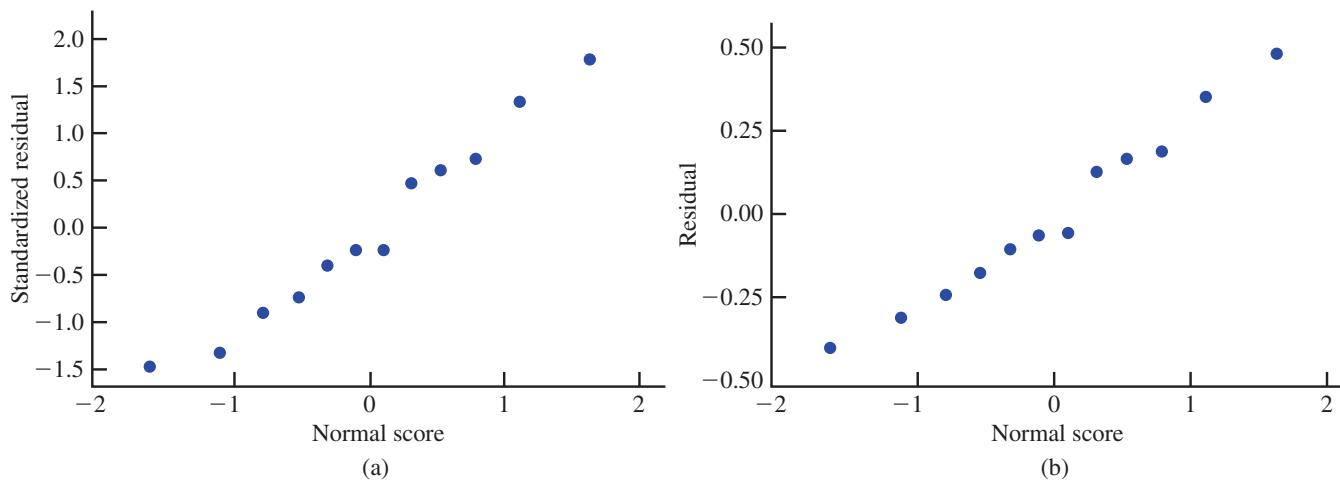
Minitab output for the data of Example 13.6.



The residuals, their standard deviations, and the standardized residuals (computed using Minitab) are given in Table 13.1. For the residual with the largest magnitude, 0.49, the standardized residual is 1.81. This means that this residual is approximately 1.8 standard deviations above its expected value of 0, which is not particularly unusual in a sample of this size. On the standardized scale, no residual here is surprisingly large.

TABLE 13.1 Data, Residuals, and Standardized Residuals for Example 13.6

Observation	Competent Proportion x	Difference in Vote Proportion y	\hat{y}	Residual	Estimated Standard Deviation of Residual	Standardized Residual
1	0.20	-0.70	-0.39	-0.31	0.24	-1.32
2	0.23	-0.40	-0.35	-0.05	0.24	-0.22
3	0.40	-0.35	-0.11	-0.24	0.27	-0.89
4	0.35	0.18	-0.18	0.36	0.27	1.35
5	0.40	0.38	-0.11	0.49	0.27	1.81
6	0.45	-0.10	-0.04	-0.06	0.27	-0.22
7	0.50	0.20	0.03	0.17	0.27	0.62
8	0.55	-0.30	0.10	-0.40	0.27	-1.48
9	0.60	0.30	0.17	0.13	0.27	0.49
10	0.68	0.18	0.28	-0.10	0.26	-0.39
11	0.70	0.50	0.31	0.19	0.25	0.75
12	0.76	0.22	0.39	-0.17	0.24	-0.72

**FIGURE 13.14**

Normal probability plots for Example 13.6 (from Minitab):
 (a) standardized residuals;
 (b) residuals.

Figure 13.14 displays a normal probability plot of the standardized residuals and also a separate normal probability plot of the residuals. Notice that in this case the plots are nearly identical. It is usually the case that the two plots are similar. Although it is preferable to work with the standardized residuals, without access to a statistical software package or calculator that will produce standardized residuals, a plot of the unstandardized residuals can be used. Both of the normal probability plots in Figure 13.14 are approximately linear. The plots would not cause us to question the assumption of normality.

In addition to a normal probability plot, a boxplot of the residuals might also be used to assess whether the assumption of normality is reasonable. A boxplot that is reasonably symmetric and that does not show any outliers is consistent with the assumption of normality.

Plotting the Residuals

A plot of the $(x, \text{residual})$ pairs is called a **residual plot**, and a plot of the $(x, \text{standardized residual})$ pairs is a **standardized residual plot**. Residual and standardized residual plots typically exhibit the same general shapes. With access to a statistical software package or graphing calculator that calculates standardized residuals, we recommend using the standardized residual plot. If this is not possible, it is acceptable to use the residual plot instead.

A standardized residual plot or a residual plot is often helpful in identifying unusual or highly influential observations and in checking for violations of model assumptions. A desirable plot is one that exhibits no particular pattern (such as curvature or a much greater vertical variability in one part of the plot than in another) and that has no point that is far removed from all the others.

A point falling far above or far below the horizontal line $\text{standardized residual} = 0$ corresponds to a large standardized residual, which can indicate some kind of unusual behavior, such as a recording error, a nonstandard experimental condition, or an unusual experimental subject. A point that has an x value that differs greatly from others in the data set could have exerted excessive influence in determining the fitted line.

A standardized residual plot like the one pictured in Figure 13.15(a) is desirable, because no point lies much outside the horizontal band between -2 and 2 (so there is no unusually large standardized residual corresponding to an outlying observation). There is also no point far to the left or right of the others (which could indicate an observation that might greatly influence the fit). Finally there is no pattern to indicate that the model should somehow be modified.

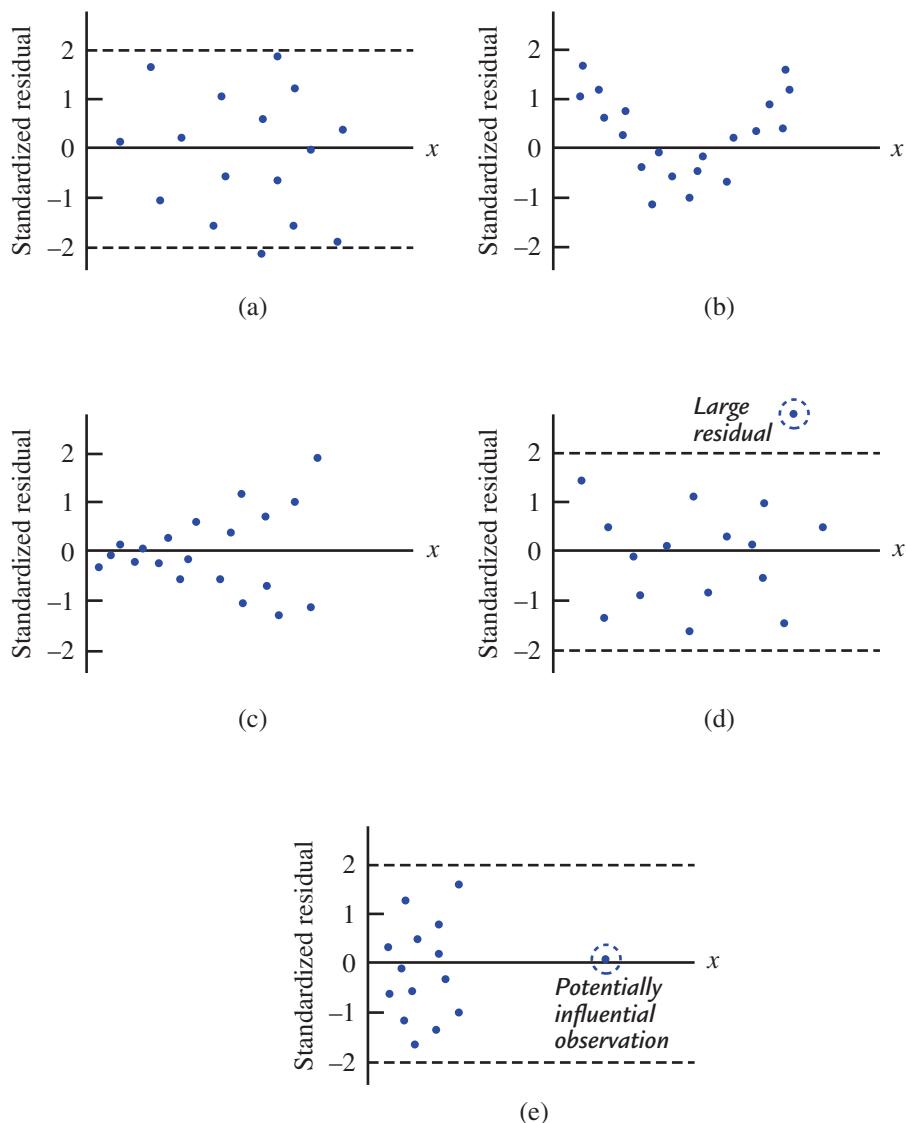
When the plot has the appearance of Figure 13.15(b), the fitted model should be changed to incorporate curvature (a nonlinear model). The increasing variability from left to right in Figure 13.15(c) suggests that the variance of y is not the same at each

x value but rather increases with x . A linear model may still be appropriate, but the best-fit line should be obtained by using *weighted least squares* rather than ordinary least squares. This involves giving more weight to observations in the region exhibiting low variability and less weight to observations in the region exhibiting high variability. A specialized regression analysis textbook or a statistician should be consulted for more information on using weighted least squares.

The standardized residual plots of Figures 13.15(d) and 13.15(e) show an **outlier** (a point with a large standardized residual) and a **potentially influential observation**, respectively. Consider deleting the potentially influential observation from the data set and refitting the same model. Substantial changes in estimates warn of instability in the data. The investigator should certainly carry out a more careful analysis and perhaps collect more data before drawing any firm conclusions.

FIGURE 13.15

- Examples of residual plots:
 (a) satisfactory plot;
 (b) plot suggesting that a curvilinear regression model is needed;
 (c) plot indicating nonconstant variance;
 (d) plot showing a large residual;
 (e) plot showing a potentially influential observation.



Example 13.7 A New Pediatric Tracheal Tube

Understand the context ➤

The article “Appropriate Placement of Intubation Depth Marks in a New Cuffed, Paediatric Tracheal Tube” (*British Journal of Anaesthesia* [2004]: 80–87) describes a study of the use of tracheal tubes in newborns and infants. Newborns and infants have small trachea, and there is little margin for error when inserting tracheal tubes.

Using X-rays of a large number of children age 2 months to 14 years, the researchers examined the relationships between appropriate trachea tube insertion depth and other variables such as height, weight, and age. A scatterplot and a standardized residual plot constructed using data on the insertion depth and height of the children (both measured in cm) are shown in Figure 13.16.

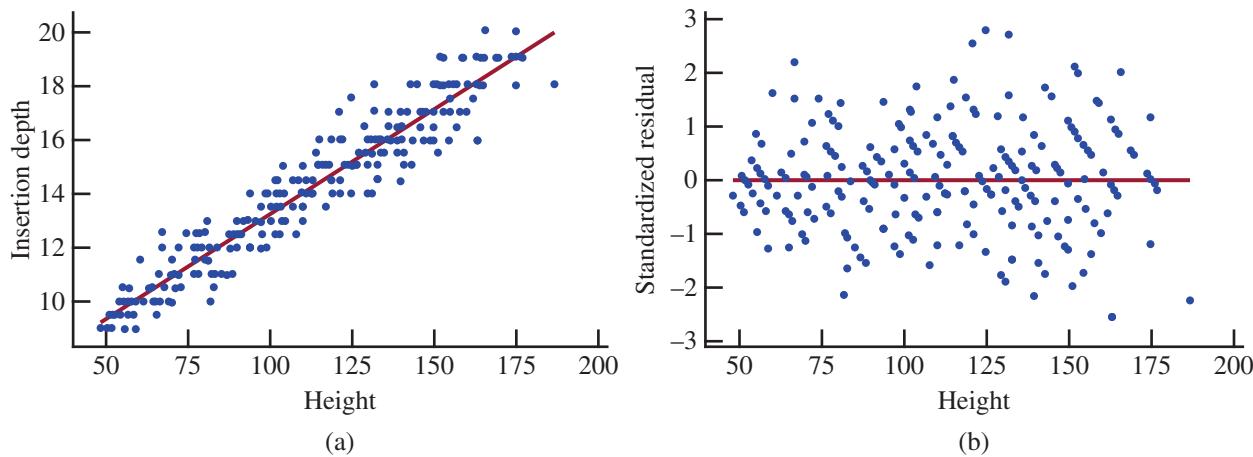


FIGURE 13.16

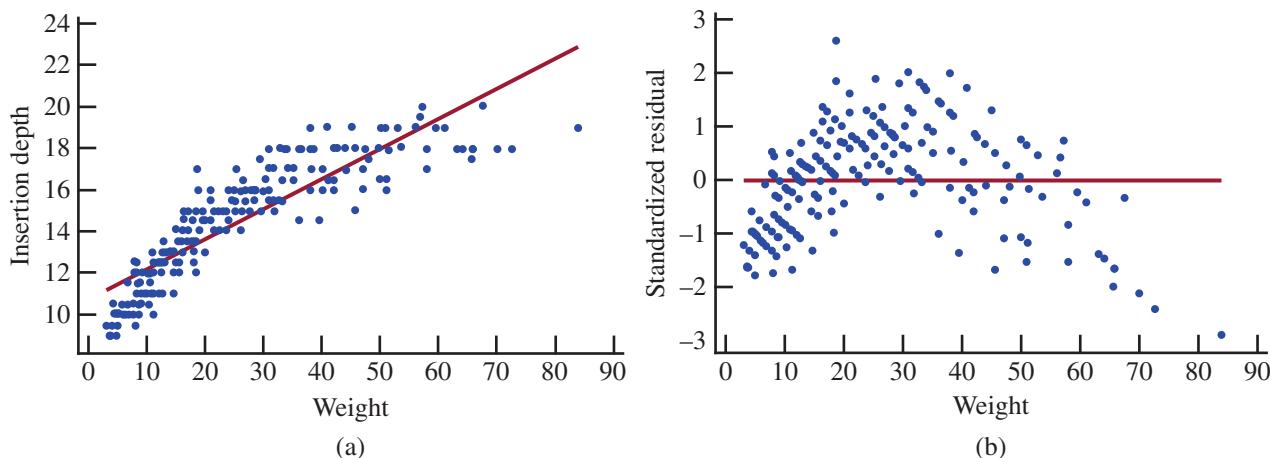
- (a) Scatterplot for insertion depth vs. height data of Example 13.7;
- (b) standardized residual plot.

Residual plots like the ones pictured in Figure 13.16(b) are desirable. No point lies outside the horizontal band between -3 and 3 , and most are between -2 and 2 (so there are no unusually large residuals corresponding to outliers). There is no point far to the left or right of the others (that is, there are no observations that might be influential), and there is no pattern of curvature or differences in the variability of the residuals for different height values to indicate that the model assumptions are not reasonable.

But consider what happens when the relationship between insertion depth and weight is examined. A scatterplot of insertion depth and weight (kg) is shown in Figure 13.17(a), and a standardized residual plot in Figure 13.17(b). While some curvature is evident in the original scatterplot, it is even more clearly visible in the standardized residual plot. A careful inspection of these plots suggests that along with curvature, the residuals may be more variable at larger weights. When plots have this curved appearance and increasing variability in the residuals, the linear regression model is not appropriate.

FIGURE 13.17

- (a) Scatterplot for insertion depth vs. weight data of Example 13.7;
- (b) standardized residual plot.



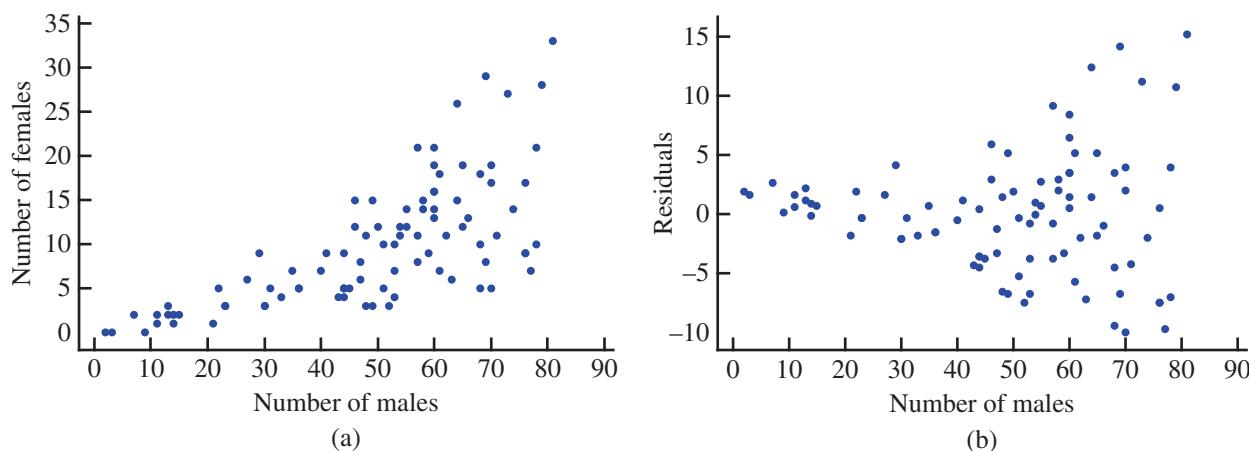
Example 13.8 Looking for Love in All the Right... Trees?

Understand the context ➤

Treefrogs' search for mating partners was the examined in the article, "The Cause of Correlations Between Nightly Numbers of Male and Female Barking Treefrogs (*Hyla gratiosa*) Attending Choruses" (*Behavioral Ecology* [2002]: 274–281). A "lek," in the world of animal behavior, is a cluster of males gathered in a relatively small area to exhibit courtship displays. The "female preference" hypothesis asserts that females will prefer larger leks over smaller leks, presumably because there are more males to choose from. The scatterplot and residual plot in Figure 13.18 show the relationship between the number of females and the number of males in observed leks of barking treefrogs. The unequal vertical variability is noticeable in the scatterplot and is even more evident in the residual plot. This indicates that the assumptions of the linear regression model are not reasonable in this situation.

FIGURE 13.18

(a) Scatterplot for treefrog data of Example 13.8;
(b) residual plot.



Example 13.9 The Business of Baseball

The article "The Business of Baseball" (forbes.com/mlb-valuations/list/#tab:overall, retrieved May 27, 2017) ranked the 30 major league baseball teams based on their 2016 value (in millions of dollars). Also included in the article are data on annual operating income (in millions of dollars). A positive value for operating income indicates a profit for the year and a negative operating income represents a loss for the year.

Team	2016 Value (millions of dollars)	Operating Income (millions of dollars)
New York Yankees	3,400	13.0
Los Angeles Dodgers	2,500	-73.2
Boston Red Sox	2,300	43.2
San Francisco Giants	2,250	72.6
Chicago Cubs	2,200	50.8
New York Mets	1,650	46.8
St. Louis Cardinals	1,600	59.8
Los Angeles Angels of Anaheim	1,340	41.7
Washington Nationals	1,300	22.5
Philadelphia Phillies	1,235	-8.9
Texas Rangers	1,225	-4.7
Seattle Mariners	1,200	16.8
Atlanta Braves	1,175	27.8
Detroit Tigers	1,150	11.0

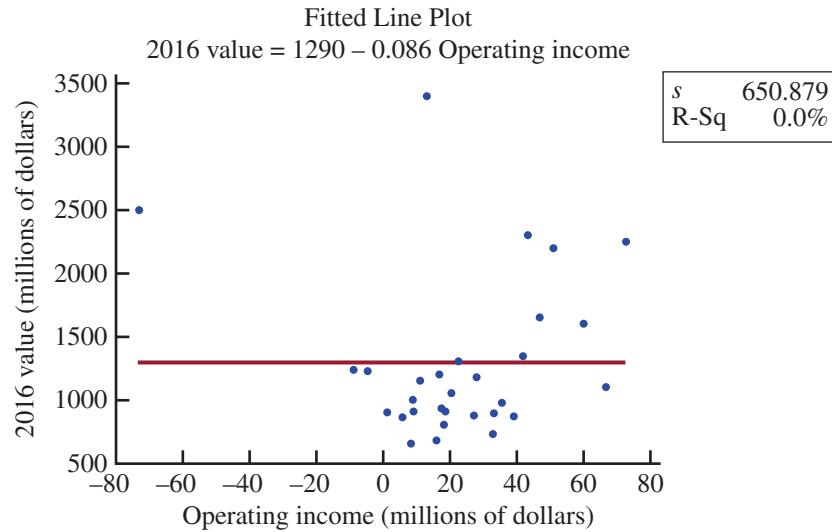
(continued)

Team	2016 Value (millions of dollars)	Operating Income (millions of dollars)
Houston Astros	1,100	66.6
Chicago White Sox	1,050	20.2
Baltimore Orioles	1,000	8.8
Pittsburgh Pirates	975	35.3
Arizona Diamondbacks	925	17.4
Minnesota Twins	910	18.5
Cincinnati Reds	905	9.0
Toronto Blue Jays	900	1.2
San Diego Padres	890	32.9
Milwaukee Brewers	875	27.0
Kansas City Royals	865	39.0
Colorado Rockies	860	5.5
Cleveland Indians	800	18.0
Oakland Athletics	725	32.7
Miami Marlins	675	15.8
Tampa Bay Rays	650	8.2

To investigate whether there is a relationship between $y = 2016$ value and $x = \text{Annual operating income}$, a simple linear regression model was fit. Figure 13.19 shows a scatterplot of the data and the least-squares regression line. Notice that there are two teams that stand out in the plot. One has an unusually low operating income (the Dodgers, with an operating income of -73.2 million dollars). The other team that stands out is a team with an unusually high 2016 value (the Yankees, with a 2016 value of 3400 million dollars [3.4 billion dollars]).

FIGURE 13.19

Scatterplot of 2016 value versus annual operating income for 30 major league baseball teams.

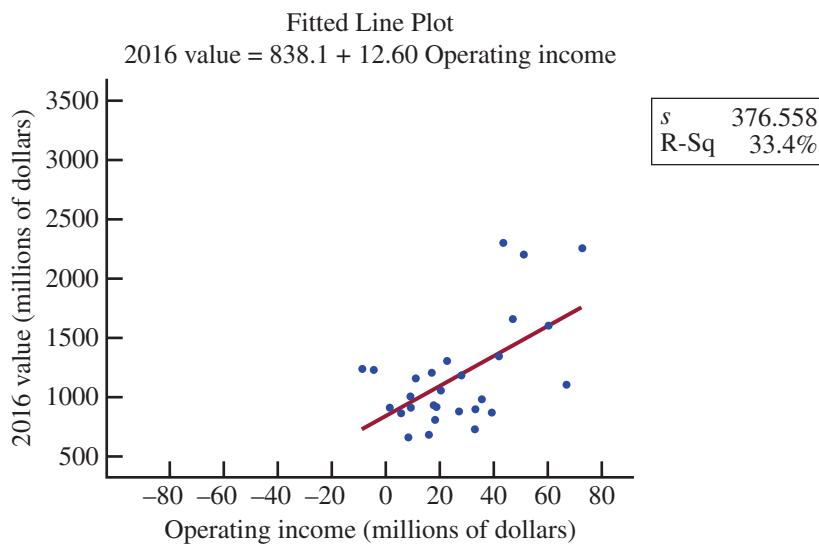


For this model, r^2 is approximately 0 and the null hypothesis that $\beta = 0$ is not rejected. This would lead us to think that there is not a useful linear relationship between 2016 value and operating income. But looking back at the scatterplot, we can see that if we were to ignore the two unusual teams, there does appear to be a positive linear relationship between 2016 value and operating income for the data set that consists of the remaining 28 teams. To investigate the potential influence of these two teams on the model, we can delete these two teams from the data set and then fit a regression model to the remaining data.

Figure 13.20 shows a scatterplot and the least-squares regression line for the 28 major league baseball teams that remain after the Dodgers and the Yankees are excluded from the data set. Notice that the slope of line has changed dramatically (from 0.086 to 12.60) and that the r^2 value is now 0.334. The model utility test confirms that there is a useful linear relationship between 2016 value and operating income for these 28 teams.

FIGURE 13.20

Scatterplot of 2016 value versus annual operating income for 28 major league baseball teams (Dodgers and Yankees excluded).



Occasionally, a residual plot or a standardized residual plot will display \hat{y} plotted on the horizontal axis rather than x . Because \hat{y} is just a linear function of x , using \hat{y} rather than x changes the scale of the horizontal axis but does not change the pattern of the points in the plot. As a consequence, residual plots that use \hat{y} on the horizontal axis can be interpreted in the same manner as residual plots that use x .

When the distribution of the random deviation e has heavier tails than the normal distribution, observations with large standardized residuals are not that unusual. Such observations can have a substantial effect on the equation of the estimated regression line when the least-squares approach is used. Statisticians have proposed a number of alternative methods—called **robust**, or **resistant**, methods—for fitting a line. These methods give less weight to outlying observations than does the least-squares method without deleting the outliers from the data set. The most widely used robust procedures require a substantial amount of computation, so an appropriate computer program is necessary.

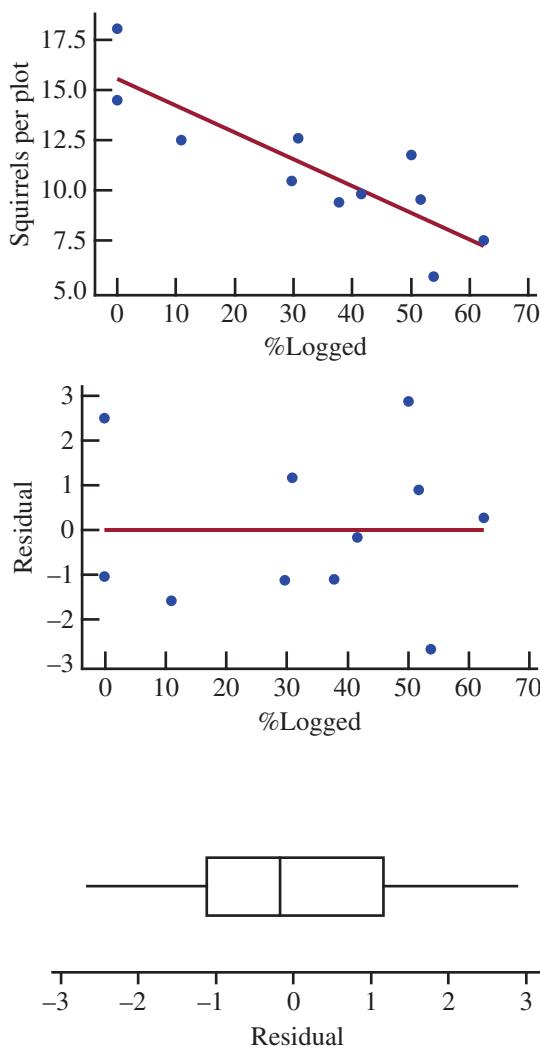
EXERCISES 13.31 - 13.36

● Data set available online

- 13.31** The graphs accompanying this exercise are based on data from an experiment to assess the effects of logging on a squirrel population in British Columbia (“Effects of Logging Pattern and Intensity on Squirrel Demography,” *The Journal of Wildlife Management* [2007]: 2655–2663). Plots of land, each 9 hectares in area, were subjected to different percentages of

logging, and the squirrel population density for each plot was measured after 3 years. The scatterplot, residual plot, and a boxplot of the residuals are shown on the next page.

Does it appear that the assumptions of the simple linear regression model are plausible? Explain.

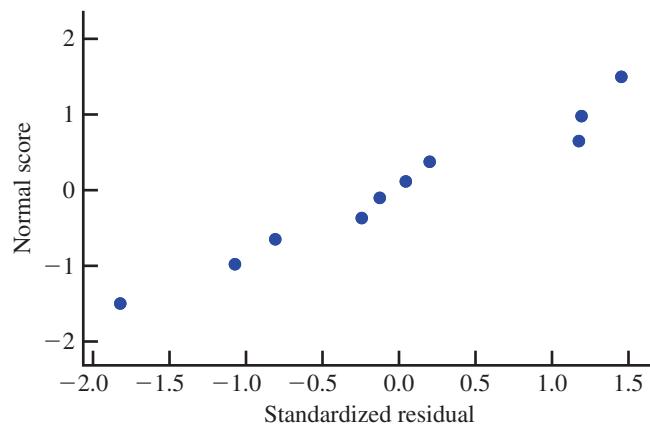


- 13.32** • Exercise 13.23 gave data on x = Nerve firing frequency and y = Pleasantness rating when nerves were stimulated by a light brushing stroke on the forearm. The x values and the corresponding residuals from a simple linear regression are as follows:

Firing Frequency, x	Standardized Residual
23	-1.83
24	0.04
22	1.45
25	0.20
27	-1.07
28	1.19
34	-0.24
33	-0.13
36	-0.81
34	1.17

- Construct a standardized residual plot. Does the plot exhibit any unusual features?
- A normal probability plot of the standardized residuals follows. Based on this plot, is it

reasonable to assume that the error distribution is approximately normal? Explain.



- 13.33** • Sea bream are a type of fish that are often raised in large fish farming enterprises. These fish are usually fed a diet consisting primarily of fish meal. The authors of the paper "Growth and Economic Profit of Gilthead Sea Bream (*Sparus aurata*, L.) Fed Sunflower Meal" (*Aquaculture* [2007]: 528–534) describe a study to investigate whether it would be more profitable to substitute plant protein in the form of sunflower meal for some of the fish meal in the sea bream's diet. The accompanying data are consistent with summary quantities given in the paper for x = Percentage of sunflower meal in the diet and y = Average weight (in grams) of fish after 248 days.

Sunflower Meal (%)	Average Fish Weight
0	432
6	450
12	455
18	445
24	427
30	422
36	421

The estimated regression line for these data is $\hat{y} = 448.536 - 0.696x$ and the standardized residuals are as given.

Sunflower Meal (%), x	Standardized Residual
0	-1.96
6	0.58
12	1.42
18	0.84
24	-0.46
30	-0.58
36	-0.29

Construct a standardized residual plot. What does the plot suggest about the adequacy of the simple linear regression model?

- 13.34** ● The article “*Vital Dimensions in Volume Perception: Can the Eye Fool the Stomach?*” (*Journal of Marketing Research* [1999]: 313–326) gave the accompanying data on the dimensions of 27 representative food products (Gerber baby food, Cheez Whiz, Skippy Peanut Butter, and Ahmed’s tandoori paste, to name a few).

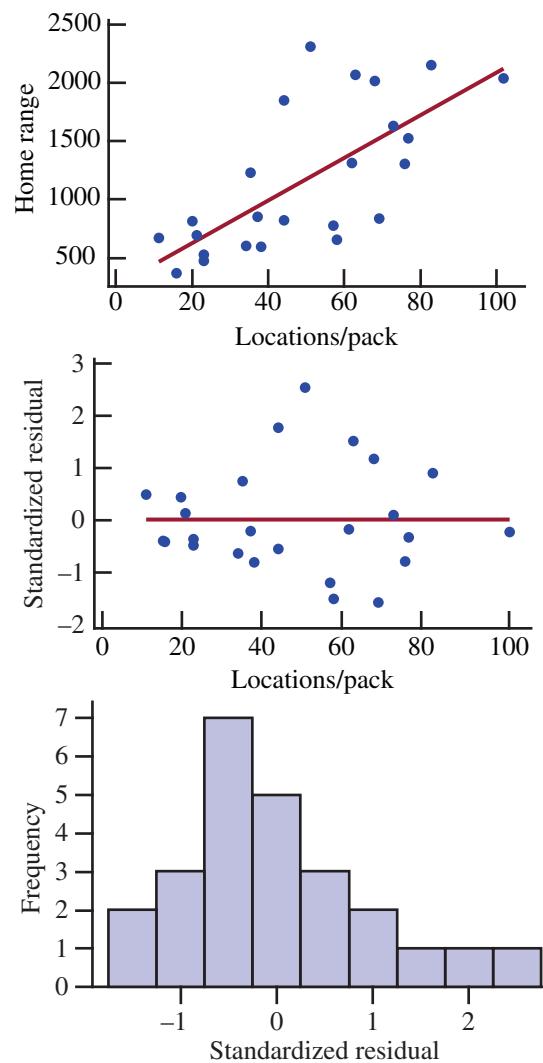
Product	Maximum Width (cm)	Minimum Width (cm)
1	2.50	1.80
2	2.90	2.70
3	2.15	2.00
4	2.90	2.60
5	3.20	3.15
6	2.00	1.80
7	1.60	1.50
8	4.80	3.80
9	5.90	5.00
10	5.80	4.75
11	2.90	2.80
12	2.45	2.10
13	2.60	2.20
14	2.60	2.60
15	2.70	2.60
16	3.10	2.90
17	5.10	5.10
18	10.20	10.20
19	3.50	3.50
20	2.70	1.20
21	3.00	1.70
22	2.70	1.75
23	2.50	1.70
24	2.40	1.20
25	4.40	1.20
26	7.50	7.50
27	4.25	4.25

- Fit the simple linear regression model that would allow prediction of the maximum width of a food container based on its minimum width.
- Calculate the standardized residuals (or just the residuals if a computer program that doesn’t give standardized residuals is used) and make a residual plot to determine whether there are any outliers.
- The data point with the largest residual is for a 1-liter Coke bottle. Delete this data point and determine the equation of the regression line. Did deletion of this point result in a large change in the equation of the estimated regression line?
- For the regression line of Part (c), interpret the estimated slope and, if appropriate, the intercept.

- For the data set with the Coke bottle deleted, are the assumptions of the simple linear regression model reasonable? Give statistical evidence.

- 13.35** Investigators in northern Alaska periodically monitored radio-collared wolves in 25 wolf packs over 4 years, keeping track of the packs’ home ranges (“*Population Dynamics and Harvest Characteristics of Wolves in the Central Brooks Range, Alaska*,” *Wildlife Monographs*, [2008]: 1–25). The home range of a pack is the area typically covered by its members in a specified amount of time. The investigators noticed that wolf packs with larger home ranges tended to be located more often by monitoring equipment. The investigators decided to explore the relationship between home range and the number of locations per pack. A scatterplot and standardized residual plot of the data are shown below, as well as a histogram of the standardized residuals.

Does it appear that the assumptions of the simple linear regression model are plausible? Explain in a few sentences.



- 13.36** • An investigation of the relationship between x = Traffic flow (thousands of cars per 24 hours) and y = Lead content of bark on trees near the highway (mg/g dry weight) yielded the accompanying data. A simple linear regression model was fit, and the resulting estimated regression line was $\hat{y} = 28.7 + 33.3x$. Both residuals and standardized residuals are also given.

x	8.3	8.3	12.1	12.1	17.0
y	227	312	362	521	640
Residual	-78.1	6.9	-69.6	89.4	45.3
St. resid.	-0.99	0.09	-0.81	1.04	0.51

x	17.0	17.0	24.3	24.3	24.3
y	539	728	945	738	759
Residual	-55.7	133.3	107.2	-99.8	-78.8
St. resid.	-0.63	1.51	1.35	-1.25	-0.99

- a. Plot the $(x, \text{residual})$ pairs. Does the resulting plot suggest that a simple linear regression model is an appropriate choice? Explain.
- b. Construct a standardized residual plot. Does the plot differ significantly in general appearance from the plot in Part (a)?

CHAPTER ACTIVITIES

ACTIVITY 13.1 ARE TALL WOMEN FROM “BIG” FAMILIES?

In this activity, work with a partner (or in a small group).

Consider the following data on height (in inches) and number of siblings for a random sample of 10 female students at a large university.

1. Construct a scatterplot of the given data. Does there appear to be a linear relationship between y = Height and x = Number of siblings?

Height (y)	Number of Siblings (x)	Height (y)	Number of Siblings (x)
64.2	2	65.5	1
65.4	0	67.2	2
64.6	2	66.4	2
66.1	6	63.3	0
65.1	3	61.7	1

2. Calculate the value of the correlation coefficient. Is the value of the correlation coefficient consistent with the answer from Step 1? Explain.
3. What is the equation of the least-squares line for these data?
4. Is the slope of the least-squares regression line from Step 3 equal to 0? Does this necessarily mean that there is a meaningful relationship between height and

number of siblings in the population of female students at this university? Discuss this, and then write a few sentences of explanation.

5. For the population of all female students at the university, is it reasonable to assume that the distribution of heights at each particular x value is approximately normal and that the standard deviation of the height distribution at each particular x value is the same? Is it reasonable to assume that the distribution of heights for female students with zero siblings is approximately normal and that the distribution of heights for female students with one sibling is approximately normal with the same standard deviation as for female students with no siblings, and so on? Discuss, and then write a few sentences of explanation.
6. Carry out the model utility test ($H_0: \beta = 0$). Explain why the conclusion from this test is consistent with the explanation in Step 4.
7. Is using the least-squares regression line a useful way to predict heights for women at this university? Explain.
8. Write a paragraph explaining why it is a good idea to include a model utility test ($H_0: \beta = 0$) as part of a regression analysis.

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
Simple linear regression model, $y = \alpha + \beta x + e$	This model assumes that there is a line with slope β and y intercept α , called the population regression line, such that an observation deviates from the line by a random amount e . The random deviation is assumed to have a normal distribution with mean zero and standard deviation σ , and random deviations for different observations are assumed to be independent of one another.	$t = \frac{b - \text{hypothesized value}}{s_b}$	The test statistic for testing hypotheses about β . The test is based on $(n - 2)$ degrees of freedom.
Estimated regression line, $\hat{y} = a + bx$	The least-squares line introduced in Chapter 5.	Model utility test, with test statistic $t = \frac{b}{s_b}$	A test of $H_0: \beta = 0$, which asserts that there is no useful linear relationship between x and y , versus $H_a: \beta \neq 0$, the claim that there is a useful linear relationship.
$s_e = \sqrt{\frac{\text{SSResid}}{n - 2}}$	The point estimate of the standard deviation σ , with associated degrees of freedom $(n - 2)$.	Residual analysis	Methods based on the residuals or standardized residuals for checking the assumptions of a regression model.
$s_b = \frac{s_e}{\sqrt{S_{xx}}}$	The estimated standard deviation of the statistic b .	Standardized residual	A residual divided by its standard deviation.
$b \pm (t \text{ critical value})s_b$	A confidence interval for the slope β of the population regression line, where the t critical value is based on $(n - 2)$ degrees of freedom.	Standardized residual plot	A plot of the $(x, \text{standardized residual})$ pairs. A pattern in this plot suggests that the simple linear regression model may not be appropriate.

TECHNOLOGY NOTES

Test for Slope of Regression Line

TI-83/84

- Enter the data for the independent variable into **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
- Enter the data for the dependent variable into **L2**
- Press **STAT**
- Highlight **TESTS**
- Highlight **LinRegTTest...** and press **ENTER**
- Next to β & ρ select the appropriate alternative hypothesis
- Highlight **Calculate**

TI-Nspire

- Enter the data into two separate data lists (In order to access data lists select the spreadsheet option and press **enter**) **Note:** Be sure to title the lists by selecting the top row of the column and typing a title.
- Press the **menu** key and select **4:Stat Tests** then **4:Stats Tests** then **A:Linear Reg t Test...** and press **enter**
- In the box next to **X List** choose the list title where the independent data are stored from the drop-down menu
- In the box next to **Y List** choose the list title where the dependent data are stored from the drop-down menu
- In the box next to **Alternate Hyp** choose the appropriate alternative hypothesis from the drop-down menu
- Press **OK**

JMP

- Input the data for the dependent variable into the first column
- Input the data for the independent variable into the second column
- Click **Analyze** and select **Fit Y by X**
- Select the dependent variable (Y) from the box under **Select Columns** and click on **Y, Response**
- Select the independent variable (X) from the box under **Select Columns** and click on **X, Factor**
- Click the red arrow next to **Bivariate Fit of...** and select **Fit Line**

Minitab

- Input the data for the dependent variable into the first column
- Input the data for the independent variable into the second column
- Select **Stat** then **Regression** then **Regression...**
- Highlight the name of the column containing the dependent variable and click **Select**
- Highlight the name of the column containing the independent variable and click **Select**
- Click **OK**

Note: Scroll up in the Session window to view the t -test results for the regression analysis.

SPSS

1. Input the data for the dependent variable into one column
2. Input the data for the independent variable into a second column
3. Click **Analyze** then click **Regression** then click **Linear...**
4. Select the name of the dependent variable and click the arrow to move the variable to the box under **Dependent:**
5. Select the name of the independent variable and click the arrow to move the variable to the box under **Independent(s):**
6. Click **OK**

Note: The p -value for the regression test can be found in the **Coefficients** table in the row with the independent variable name.

Excel

1. Input the data for the dependent variable into the first column
2. Input the data for the independent variable into the second column

CUMULATIVE REVIEW EXERCISES**CR13.1 - CR13.18**

● Data set available online

CR13.1 The article “**You Will Be Tested on This**” (*The Chronicle of Higher Education*, June 8, 2007) describes an experiment to investigate the effect of quizzes on student learning. The goal of the experiment was to determine if students who take daily quizzes have better end-of-semester retention than students who attend the same lectures and complete the same homework assignments but who do not take the daily quizzes.

Describe how an experiment using the 400 students enrolled in an introductory psychology course as subjects might be carried out.

CR13.2 The paper “**Pistachio Nut Consumption and Serum Lipid Levels**” (*Journal of the American College of Nutrition* [2007]: 141–148) describes a study to determine if eating pistachio nuts can have an effect on blood cholesterol levels in people with high cholesterol. Fifteen subjects followed their regular diet for 4 weeks and then followed a diet in which 15% of the daily caloric intake was from pistachio nuts for 4 weeks. Total blood cholesterol was measured for each subject at the end of each of the two 4-week periods, resulting in two samples (one for the regular diet and one for the pistachio diet).

- a. Are the two samples independent or paired? Explain.
- b. The mean difference in total cholesterol (regular diet—pistachio diet) was 11 mg/dL. The standard deviation of the differences was 24 mg/dL. Assume that it is reasonable to regard the 15 study participants as representative of adults with high cholesterol and that total cholesterol differences are approximately normally distributed.

Do the data support the claim that eating the pistachio diet for 4 weeks is effective in reducing total cholesterol level? Test the relevant hypotheses using $\alpha = 0.01$.

CR13.3 ● The article “**Fines Show Airline Problems**” (*USA TODAY*, February 2, 2010) gave the accompanying data on the number of fines for violating FAA maintenance regulations

3. Select **Analyze** then choose **Regression** then choose **Linear...**
4. Highlight the name of the column containing the dependent variable
5. Click the arrow button next to the **Dependent** box to move the variable to this box
6. Highlight the name of the column containing the independent variable
7. Click the arrow button next to the **Independent** box to move the variable to this box
8. Click **OK**

Note: The test statistic and p -value for the regression test for the slope can be found in the third table of output. These values are listed in the row titled with the independent variable name and the columns entitled *t Stat* and *P-value*.

assessed against each of the 25 U.S. airlines from 2004 to 2009.

1	12	3	7	23	36	6	14	1	3
4	10	6	2	2	2	2	2	3	1
2	2	1	0	0					

- a. Construct a boxplot of these data. Are any of the observations in the data set outliers? If so, which ones?
- b. Explain why it may not be reasonable to assume that the two airlines with the highest number of fines assessed are the worst airlines in terms of maintenance violations.

CR13.4 The article “**Odds Are, It’s Wrong**” (*Science News*, March 27, 2010) poses the following scenario:

Suppose that a test for steroid use among baseball players is 95% accurate—that is, it correctly identifies actual steroid users 95% of the time, and misidentifies non-users as users 5 percent of the time. . . . Now suppose, based on previous testing, that experts have established that about 5 percent of professional baseball players use steroids.

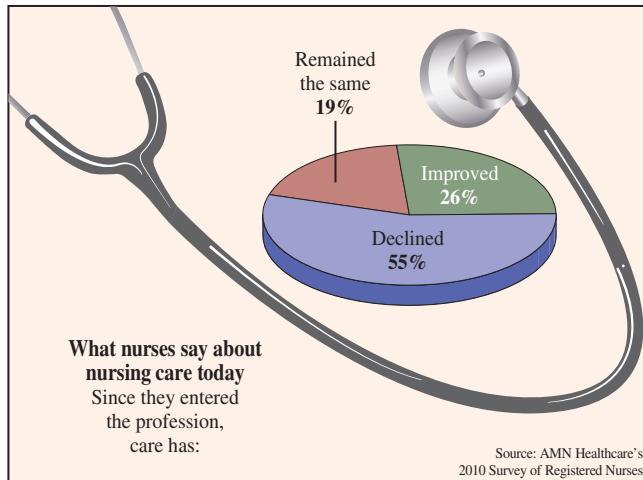
Answer the following questions for this scenario.

- a. If 400 professional baseball players are selected at random, how many would be expected to be steroid users and how many would be expected to be nonusers?
- b. How many of the steroid users would be expected to test positive for steroid use?
- c. How many of the players who do not use steroids would be expected to test positive for steroid use (a false positive)?
- d. Use the answers to Parts (b) and (c) to estimate the proportion of those who test positive for steroid use who actually do use steroids.
- e. Write a few sentences explaining why, in this scenario, the proportion of those who test positive for steroid use who actually use steroids is not 0.95.

CR13.5 The press release “**Luxury or Necessity? The Public Makes a U-Turn**” (Pew Research Center, April 23, 2009) summarizes results from a survey of a nationally representative sample of $n = 1003$ adult Americans.

- One question in the survey asked participants if they think of a landline phone as a necessity or as a luxury that they could do without. Sixty-eight percent said they thought a landline phone was a necessity. Estimate the proportion of adult Americans who view a landline phone as a necessity using a 95% confidence interval.
- In the same survey, 52% said they viewed a television set as a necessity. Is there convincing evidence that a majority of adult Americans view a television set as a necessity? Test the relevant hypotheses using $\alpha = 0.05$.
- The press release also described a survey conducted in 2003. When asked about a microwave oven, 68% of the 2003 sample regarded a microwave oven as a necessity, whereas only 47% of the 2009 sample said they thought a microwave oven was a necessity. Assume that the sample size for the 2003 survey was also 1003. Is there convincing evidence that the proportion of adult Americans who regard a microwave oven as a necessity decreased between 2003 and 2009? Test the appropriate hypotheses using $\alpha = 0.01$.

CR13.6 The accompanying graphical display is similar to one that appeared in **USA TODAY** (February 19, 2010). It is meant to be a pie chart, but an oval rather than a circle is used to represent the whole pie. Does this graph do a good job of conveying the proportion falling into each of the three response categories? Explain why or why not.



CR13.7 The following quote describing 18- to 29-year-olds is from the article “**Study: Millennial Generation More Educated, Less Employed**” (**USA TODAY**, February 23, 2010): “38% have a tattoo (and half of those with tattoos have two to five; 18% have six or more).” These percentages were based on a representative sample of 830 Americans age 18 to 29, but for purposes of this exercise, suppose that they hold for the population of all Americans in this age group. Define the random variable $x = \text{Number of tattoos for a}$

randomly selected American age 18 to 29. Find the following probabilities:

- $P(x = 0)$
- $P(x = 1)$
- $P(2 \leq x \leq 5)$
- $P(x > 5)$

CR13.8 To raise revenues, many airlines now charge fees to check luggage. Suppose that the number of checked bags was recorded for each person in a random sample of 100 airline passengers selected before fees were imposed and also for each person in a random sample of 100 airline passengers selected after fees were imposed, resulting in the accompanying data. Do the data provide convincing evidence that the proportions in each of the number of checked bags categories are not all the same before and after fees were imposed? Test the appropriate hypotheses using a significance level of 0.05.

Number of Checked Bags			
	0	1	2 or more
Before fees	7	70	23
After fees	22	64	14

CR13.9 Consider the following data on $y = \text{Number of songs stored on an MP3 player}$ and $x = \text{Number of months the user has owned the MP3 player}$ for a sample of 15 owners of MP3 players.

x	y
23	486
35	747
2	81
28	581
5	117
32	728
23	445
10	128
4	61
26	476
1	35
8	121
13	266
9	126
5	141

- Construct a scatterplot of the data. Does the relationship between x and y look approximately linear?
- What is the equation of the estimated regression line?
- Are the assumptions of the simple linear regression model reasonable? Justify the answer using appropriate graphs.
- Is the simple linear regression model useful for describing the relationship between x and y ? Test the relevant hypotheses using a significance level of 0.05.

CR13.10 Many people take ginkgo supplements advertised to improve memory. Are these over-the-counter supplements effective? In a study reported in the paper “**Ginkgo for Memory Enhancement**” (*Journal of the American Medical Association* [2002]: 835–840), elderly adults were assigned at random to either a treatment group or a control group. The 104 participants who were assigned to the treatment group took 40 mg of ginkgo three times a day for 6 weeks. The 115 participants assigned to the control group took a placebo pill three times a day for 6 weeks. At the end of 6 weeks, the Wechsler Memory Scale (a test of short-term memory) was administered. Higher scores indicate better memory function. Summary values are given in the following table.

	<i>n</i>	\bar{x}	<i>s</i>
Ginkgo	104	5.6	0.6
Placebo	115	5.5	0.6

Based on these results, is there evidence that taking 40 mg of ginkgo three times a day is effective in increasing mean performance on the Wechsler Memory Scale? Test the relevant hypotheses using $\alpha = 0.05$.

CR13.11 • The **Harvard University Institute of Politics** surveys undergraduates across the United States annually. Responses to the question “When it comes to voting, do you consider yourself to be affiliated with the Democratic Party, the Republican Party, or are you Independent or unaffiliated with a major party?” for the surveys conducted in 2003, 2004, and 2005 are summarized in the given table. The samples for each year were independently selected and are considered to be representative of the population of undergraduate students in the year the survey was conducted.

Is there evidence that the distribution of political affiliation is not the same for all three years for which data are given?

Political Affiliation	Year		
	2005	2004	2003
Democrat	397	409	325
Republican	301	349	373
Independent/unaffiliated	458	397	457
Other	60	48	48

CR13.12 • The survey described in the previous exercise also asked the following question: “Please tell me whether you trust the President to do the right thing all of the time, most of the time, some of the time, or never.” Use the data in the table on the next page and an appropriate hypothesis test to determine if there is evidence that trust in the president was not the same in 2005 as it was in 2002.

Response	Year	
	2005	2002
All of the time	132	180
Most of the time	337	528
Some of the time	554	396
Never	169	96

CR13.13 • The report “**Undergraduate Students and Credit Cards in 2004**” (Nellie Mae, May 2005) included information collected from individuals in a random sample of undergraduate students in the United States. Students were classified according to region of residence and whether or not they have one or more credit cards, resulting in the accompanying two-way table.

Carry out a test to determine if there is evidence that region of residence and having a credit card are not independent. Use $\alpha = 0.05$.

Region	Credit Card?	
	At Least One Credit Card	No Credit Cards
Northeast	401	164
Midwest	162	36
South	408	115
West	104	23

CR13.14 • The report described in the previous exercise also classified students according to region of residence and whether or not they had a credit card with a balance of more than \$7000. Do these data support the conclusion that there is an association between region of residence and whether or not the student has a balance exceeding \$7000? Test the relevant hypotheses using a 0.01 significance level.

Region	Balance Over \$7000?	
	No	Yes
Northeast	28	537
Midwest	162	182
South	42	481
West	9	118

CR13.15 The discharge of industrial wastewater into rivers affects water quality. To assess the effect of a particular power plant on water quality, 24 water specimens were taken 16 km upstream and 4 km downstream of the plant. Alkalinity (mg/L) was determined for each specimen, resulting in the summary quantities in the accompanying table. Do the data suggest that the actual mean alkalinity is higher downstream than upstream by more than 50 mg/L? Use a 0.05 significance level.

Location	<i>n</i>	Mean	Standard Deviation
Upstream	24	75.9	1.83
Downstream	24	183.6	1.70

CR13.16 • The report of a European commission on radiation protection titled “[Cosmic Radiation Exposure of Aircraft Crew](#)” (2004) measured the exposure to radiation on eight international flights from Madrid using several different methods for measuring radiation. Data for two of the methods are given in the accompanying table. Use these data to test the hypothesis that there is no significant difference in mean radiation measurement for the two methods.

Flight	Method 1	Method 2
1	27.5	34.4
2	41.3	38.6
3	3.5	3.5
4	24.3	21.9
5	27.0	24.4
6	17.7	21.4
7	12.0	11.8
8	20.9	24.1

CR13.17 • It is hypothesized that when homing pigeons are disoriented in a certain manner, they will exhibit no preference for any direction of flight after takeoff. To test this, 120 pigeons are disoriented and released, and the direction of flight of each is recorded. The resulting data are given in the accompanying table.

Direction	Frequency
0° to < 45°	12
45° to < 90°	16
90° to < 135°	17
135° to < 180°	15
180° to < 225°	13
225° to < 270°	20
270° to < 315°	17
315° to < 360°	10

Use the goodness-of-fit test with significance level 0.10 to determine whether the data are consistent with this hypothesis.

CR13.18 The authors of the paper “[Inadequate Physician Knowledge of the Effects of Diet on Blood Lipids and Lipoproteins](#)” (*Nutrition Journal* [2003]: 19–26) summarized the responses to a questionnaire on basic knowledge of nutrition that was mailed to 6000 physicians selected at random from a list of physicians licensed in the United States. Sixteen percent of those who received the questionnaire completed and returned it. The authors report that 26 of 120 cardiologists and 222 of 419 internists did not know that carbohydrate was the diet component most likely to raise triglycerides.

- a. Estimate the difference between the proportion of cardiologists and the proportion of internists who did not know that carbohydrate was the diet component most likely to raise triglycerides using a 95% confidence interval.
- b. What potential source of bias might limit the ability to generalize the estimate from Part (a) to the populations of all cardiologists and all internists?

14

Multiple Regression Analysis



pics721/Shutterstock.com

The general objective of regression analysis is to model the relationship between a dependent variable y and one or more independent variables (also sometimes called predictor or explanatory variables). The simple linear regression model $y = \alpha + \beta x + e$, introduced in Chapter 13, relates y to a single independent variable x . In many situations, the relationship between y and any single independent variable is not strong, but knowing the values of several independent variables may considerably reduce uncertainty about the associated y value.

For example, some variability in house prices in a large city can be attributed to house size, but knowledge of size by itself would not usually allow a bank appraiser to accurately predict a home's value. Price is also determined to some extent by other variables, such as the age of the house, lot size, number of bedrooms and bathrooms, and distance from schools.

In this chapter, we extend the regression methodology developed in the simple linear regression model introduced in the previous chapter to *multiple regression models*, which include at least two independent variables. Fortunately, many of the concepts developed in the context of simple linear regression carry over to multiple regression with little or no modification.

However, the calculations required to fit a multiple regression model are *much* more time consuming than those for simple linear regression, so a computer is an indispensable tool.

LEARNING OBJECTIVES

Students will understand:

- How a multiple regression model can be used to model the relationship between a dependent variable and two or more independent variables.
- How numerical and categorical independent variables can be incorporated into a multiple regression model.
- The concept of interaction in a multiple regression setting.

Students will be able to:

- Interpret the parameters of the multiple linear regression model in context.
- Use both numerical and categorical variables in a multiple regression model.
- Estimate the parameters in a multiple regression model and assess the usefulness of the model.

SECTION 14.1 Multiple Regression Models

The relationship between a dependent variable y and two or more independent variables is deterministic if the value of y is completely determined, with no uncertainty, once values of the independent variables have been specified. For example, consider a school district in which teachers with no prior teaching experience and no college credits beyond a bachelor's degree start at an annual salary of \$58,000. Suppose that for each year of teaching experience up to 20 years, a teacher receives an additional \$800 per year and that each unit of postcollege coursework up to 75 units results in an extra \$60 per year. Consider the following three variables:

y = Salary of a teacher who has at most 20 years of teaching experience and at most 75 postcollege units

x_1 = Number of years of teaching experience

x_2 = Number of postcollege units

Previously, x_1 and x_2 denoted the first two observations on the single variable x . In the usual notation for multiple regression, however, x_1 and x_2 represent two different variables.

For these variables, the value of y is entirely determined by values of x_1 and x_2 through the equation

$$y = 58,000 + 800x_1 + 60x_2$$

If $x_1 = 10$ and $x_2 = 30$ then

$$\begin{aligned} y &= 58,000 + 800(10) + 60(30) \\ &= 58,000 + 8000 + 1800 \\ &= 67,800 \end{aligned}$$

If two different teachers both have the same x_1 values and the same x_2 values, they will also have identical y values.

In practice, y is rarely deterministically related to predictors x_1, \dots, x_k . A probabilistic model is more realistic in most situations. A probabilistic model results from adding a random deviation e to a deterministic function of the x_i 's.

DEFINITION

General additive multiple regression model: A model that relates a dependent variable y to k independent variables x_1, x_2, \dots, x_k .

The model is specified by the model equation

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

Model assumptions for the multiple regression model are given in the following box:

The random deviation e is assumed to be normally distributed with mean value 0 and standard deviation σ for any particular values of x_1, \dots, x_k .

This implies that for fixed x_1, x_2, \dots, x_k values, y has a normal distribution with standard deviation σ and

$$\left(\begin{array}{c} \text{mean } y \text{ value for fixed} \\ x_1, \dots, x_k \text{ values} \end{array} \right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The β_i 's in the multiple regression model are called **population regression coefficients**. Each β_i can be interpreted as the mean change in y when the predictor x_i increases by 1 unit and the values of all the other predictors remain fixed.

$\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is called the **population regression function**.

As in simple linear regression, if σ (the standard deviation of the random error distribution) is quite close to 0, any particular observed y will tend to be quite near its mean value. When σ is large, y observations may deviate substantially from their mean y values.

Example 14.1 Sophomore Success

Understand the context ➤

What factors contribute to the academic success of college sophomores? Data collected in a survey of approximately 1000 second-year college students suggest that GPA at the end of the second year is related to the student's level of interaction with faculty and staff and to the student's commitment to his or her major ("An Exploration of the Factors That Affect the Academic Success of College Sophomores," *College Student Journal* [2005] 367–376). Consider the variables

y = GPA at the end of the sophomore year

x_1 = Level of faculty and staff interaction (measured on a scale from 1 to 5)

x_2 = Level of commitment to major (measured on a scale from 1 to 5)

One possible population model might be

$$y = 1.4 + 0.33x_1 + 0.16x_2 + e$$

with

$$\sigma = 0.15$$

The population regression function is

$$(\text{mean } y \text{ value for fixed } x_1, x_2) = 1.4 + 0.33x_1 + 0.16x_2$$

Interpret the results ➤

For sophomore students whose level of interaction with faculty and staff is rated at 4.2 and whose level of commitment to their major is rated as 2.1,

$$(\text{mean value of GPA}) = 1.4 + 0.33(4.2) + 0.16(2.1) = 3.12$$

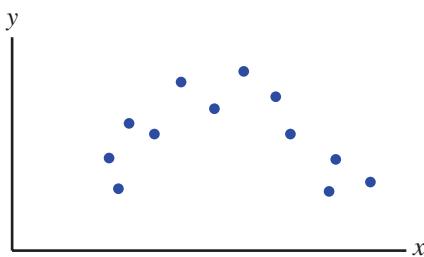
Because $2\sigma = 2(0.15) = 0.30$, it is likely that an actual y value will be within 0.30 of the mean value. This means that an individual student with $x_1 = 4.2$ and $x_2 = 2.1$ would be predicted to have a GPA between 2.82 and 3.42.

A Special Case: Polynomial Regression

Consider again the case of a single independent variable x , and suppose that a scatterplot of the n sample (x, y) pairs has the appearance of Figure 14.1. The simple linear regression model is clearly not appropriate, but it does look as though a parabola (quadratic function) with equation $y = \alpha + \beta_1 x + \beta_2 x^2$ would provide a good fit to the data.

FIGURE 14.1

A scatterplot that suggests the appropriateness of a quadratic probabilistic model.



Just as the inclusion of the random deviation e in simple linear regression allowed an observation to deviate from the population regression line by a random amount, adding e to this quadratic function yields a probabilistic model in which an observation is allowed to fall above or below the parabola. The quadratic regression model equation is

$$y = \alpha + \beta_1 x + \beta_2 x^2 + e$$

We can rewrite the model equation by using x_1 to denote x and x_2 to denote x^2 . The model equation then becomes

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$$

This is a special case of the general multiple regression model with $k = 2$. At first it may seem odd to allow one predictor variable to be a mathematical function of another predictor—here, $x_2 = (x_1)^2$. However, there is nothing in the general multiple regression model that prevents this. *In the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + e$ it is possible to have predictors that are mathematical functions of other predictors.*

For example, starting with the two independent variables x_1 and x_2 , we could create a model with $k = 4$ predictors in which x_1 and x_2 themselves are the first two predictor variables and $x_3 = (x_1)^2$, $x_4 = x_1 x_2$. (We will soon discuss the consequences of using a predictor such as x_4 .)

The general polynomial regression model begins with a single independent variable x and creates predictors $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, ..., $x_k = x^k$ for some specified value of k .

The k th-degree polynomial regression model

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + e$$

is a special case of the general multiple regression model with

$$x_1 = x \quad x_2 = x^2 \quad x_3 = x^3 \quad \dots \quad x_k = x^k$$

The **population regression function** (the mean value of y for fixed values of the predictors) is

$$\alpha + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k$$

The most important special case other than simple linear regression ($k = 1$) is the **quadratic regression model**

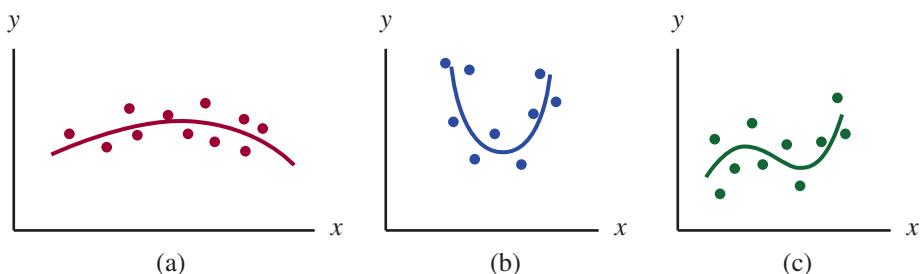
$$y = \alpha + \beta_1 x + \beta_2 x^2 + e$$

This model replaces the line of mean values $\alpha + \beta x$ in simple linear regression with a parabolic curve of mean values $\alpha + \beta_1 x + \beta_2 x^2$. If $\beta_2 > 0$, the curve opens upward, and if $\beta_2 < 0$, the curve opens downward.

A less frequently encountered special case is cubic regression, in which $k = 3$. (See Figure 14.2.)

FIGURE 14.2

Polynomial regression:
 (a) quadratic regression model with $\beta_2 < 0$;
 (b) quadratic regression model with $\beta_2 > 0$;
 (c) cubic regression model with $\beta_3 > 0$.



Example 14.2 Increased Risk of Heart Attack

Understand the context ➤

Many researchers have examined factors that are believed to contribute to the risk of heart attack. The authors of the paper “**Obesity and the Risk of Myocardial Infarction in 27,000 Participants from 53 Countries: A Case-Control Study**” (*The Lancet* [2005]: 1640–1649) found that hip-to-waist ratio was a better predictor of heart attacks than was body mass index.

The dependent variable in this study was a measure of heart attack risk (y). Larger values of y indicate greater risk of heart attack. The independent variable was hip-to-waist ratio (x). A scatterplot exhibited a curved relationship. A model consistent with summary values given in the paper is

$$y = 1.023 + 0.024x + 0.060x^2 + e$$

The population regression function is

$$\left(\begin{array}{l} \text{mean value of} \\ \text{heart attack risk measure} \end{array} \right) = 1.023 + 0.024x + 0.060x^2$$

Interpret the results ► For example, if $x = 1.3$

$$\left(\begin{array}{l} \text{mean value of} \\ \text{heart attack risk measure} \end{array} \right) = 1.023 + 0.024(1.3) + 0.060(1.3)^2 = 1.16$$

If $\sigma = 0.25$, then $2\sigma = 0.50$. This implies that it is likely that the heart attack risk measure for a person with a hip-to-waist ratio of 1.3 would be between 0.66 and 1.66.

The interpretation of β_i previously given for the general multiple regression model is not appropriate for polynomial regression. This is because all predictors are functions of the single variable x , so $x_i = x^i$ cannot be increased by 1 unit without changing the values of all the other predictor variables as well. *The interpretation of regression coefficients requires extra care when some predictor variables are mathematical functions of other variables.*

Interaction Between Variables

Suppose that an industrial chemist is interested in the relationship between

y = Product yield from a certain chemical reaction

and two independent variables,

x_1 = Reaction temperature

and

x_2 = Pressure at which the reaction is carried out.

The chemist initially suggests that for temperature values between 80 and 110 in combination with pressure values ranging from 50 to 70, the relationship can be well described by the multiple regression model

$$y = 1200 + 15x_1 - 35x_2 + e$$

The regression function, which gives the mean y value for any specified values of x_1 and x_2 , is then $1200 + 15x_1 - 35x_2$. Consider this mean y value for three different particular temperature values:

$$\begin{aligned} x_1 = 90: & \quad \text{mean } y \text{ value} = 1200 + 15(90) - 35x_2 = 2550 - 35x_2 \\ x_1 = 95: & \quad \text{mean } y \text{ value} = 2625 - 35x_2 \\ x_1 = 100: & \quad \text{mean } y \text{ value} = 2700 - 35x_2 \end{aligned}$$

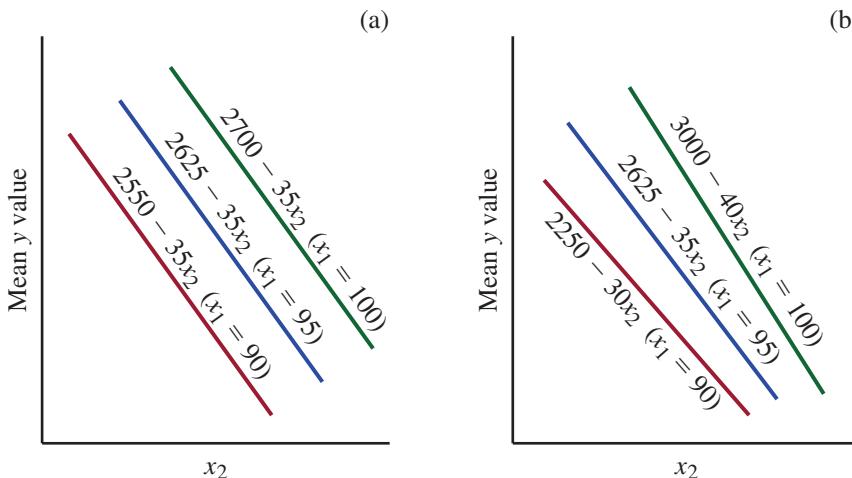
Graphs of these three mean value functions (each a function only of pressure x_2 , because a particular temperature value has been specified) are shown in Figure 14.3(a). Each graph is a straight line, and the three lines are parallel, each one having slope -35 . Because of this, the average change in yield when pressure x_2 is increased by 1 unit is -35 , regardless of the fixed temperature value.

Because chemical theory suggests that the decline in average yield when pressure x_2 increases should be more rapid for a high temperature than for a low temperature, the chemist now has reason to doubt the appropriateness of the proposed model. Rather than

FIGURE 14.3

Graphs of mean y value for two different models:

- (a) $1200 + 15x_1 - 35x_2$;
- (b) $-4500 + 75x_1 + 60x_2 - x_1x_2$.



the lines being parallel, the line for a temperature of 100 should be steeper than the line for a temperature of 95, and that line should be steeper than the one for $x_1 = 90$.

A model that has this property includes a third independent variable, $x_3 = x_1x_2$, in addition to x_1 and x_2 separately. One such model is

$$y = -4500 + 75x_1 + 60x_2 - x_1x_2 + e$$

which has regression function $-4500 + 75x_1 + 60x_2 - x_1x_2$. Then

$$\begin{aligned} (\text{mean } y \text{ when } x_1 = 100) &= -4500 + 75(100) + 60x_2 - 100x_2 \\ &= 3000 - 40x_2 \end{aligned}$$

whereas

$$\begin{aligned} (\text{mean } y \text{ when } x_1 = 95) &= 2625 - 35x_2 \\ (\text{mean } y \text{ when } x_1 = 90) &= 2250 - 30x_2 \end{aligned}$$

These functions are graphed in Figure 14.3(b), where it is clear that the three slopes are different. In fact, each different value of x_1 yields a different slope, so the average change in yield associated with a 1-unit increase in x_2 depends on the value of x_1 . When this is the case, the two variables are said to *interact*.

DEFINITION

Interaction: If the change in the mean y value associated with a 1-unit increase in one independent variable depends on the value of a second independent variable, there is **interaction** between these two independent variables.

When the variables are denoted by x_1 and x_2 , interaction can be modeled by including x_1x_2 , the product of the variables that interact, as an independent variable in the model.

The general equation for a multiple regression model based on two independent variables x_1 and x_2 that also includes an interaction term is

$$y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2 + e$$

When x_1 and x_2 do interact, this model usually gives a much better fit to the resulting sample data—and thus explains more variability in y —than does the no interaction model. Failure to consider a model with interaction may lead to an incorrect conclusion that there is no strong relationship between y and a set of independent variables.

More than one interaction term can be included in the model when there are more than two independent variables. For example, if there are three independent variables x_1 , x_2 , and x_3 , one possible model is

$$y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \beta_6x_6 + e$$

where

$$x_4 = x_1 x_2 \quad x_5 = x_1 x_3 \quad x_6 = x_2 x_3$$

We could even include a three-way interaction variable $x_7 = x_1 x_2 x_3$ (the product of all three independent variables), although this is not done very often in practice.

In applied work, quadratic terms, such as x_1^2 and x_2^2 are often included to model a curved relationship between y and several independent variables. A frequently used model involving just two independent variables x_1 and x_2 but $k = 5$ predictors is the *full quadratic or complete second-order model*

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + e$$

This model replaces the straight lines of Figure 14.3 with parabolas (each one is the graph of the regression function for different values of x_2 when x_1 has a fixed value).

With four independent variables, a model containing four quadratic terms and six two-way interaction terms is possible. Clearly, with just a few independent variables, there are a great many different multiple regression models. In Section 14.5 (online), we briefly discuss methods for selecting one model from a number of competing models.

When developing a multiple regression model, scatterplots of y with each potential predictor can be informative. This is illustrated in Example 14.3, which considers a model that includes a term that is a function of one of the independent variables and also an interaction term.

Example 14.3 Wind Chill Factor

Understand the context ➤

- The wind chill index, often included in winter weather reports, combines information on air temperature and wind speed to describe how cold it really feels. In 2001, the National Weather Service announced that it would begin using a new wind chill formula beginning in the fall of that year ([USA TODAY, August 13, 2001](#)). The following table gives the wind chill index for various combinations of air temperature and wind speed.

Wind Speed (mph)	Temperature (°F)														
	35	30	25	20	15	10	5	0	-5	-10	-15	-20	-25	-30	-35
5	31	25	19	13	7	1	-5	-11	-16	-22	-28	-34	-40	-46	-52
10	27	21	15	9	3	-4	-10	-16	-22	-28	-35	-41	-47	-53	-59
15	25	19	13	6	0	-7	-13	-19	-26	-32	-39	-45	-51	-58	-64
20	24	17	11	4	-2	-9	-15	-22	-29	-35	-42	-48	-55	-61	-68
25	23	16	9	3	-4	-11	-17	-24	-31	-37	-44	-51	-58	-64	-71
30	22	15	8	1	-5	-12	-19	-26	-33	-39	-46	-53	-60	-67	-73
35	21	14	7	0	-7	-14	-21	-27	-34	-41	-48	-55	-62	-69	-76
40	20	13	6	-1	-8	-15	-22	-29	-36	-43	-50	-57	-64	-71	-78
45	19	12	5	-2	-9	-16	-23	-30	-37	-44	-51	-58	-65	-72	-79

Figure 14.4(a) shows a scatterplot of wind chill index versus air temperature with different wind speeds denoted by different colors. It appears that the wind chill index increases linearly with air temperature at each of the wind speeds, but the linear patterns for the different wind speeds are not quite parallel. This suggests that to model the relationship between

$$y = \text{Wind chill index}$$

and the two variables

$$x_1 = \text{Air temperature}$$

and

$$x_2 = \text{Wind speed},$$

● Data set available online

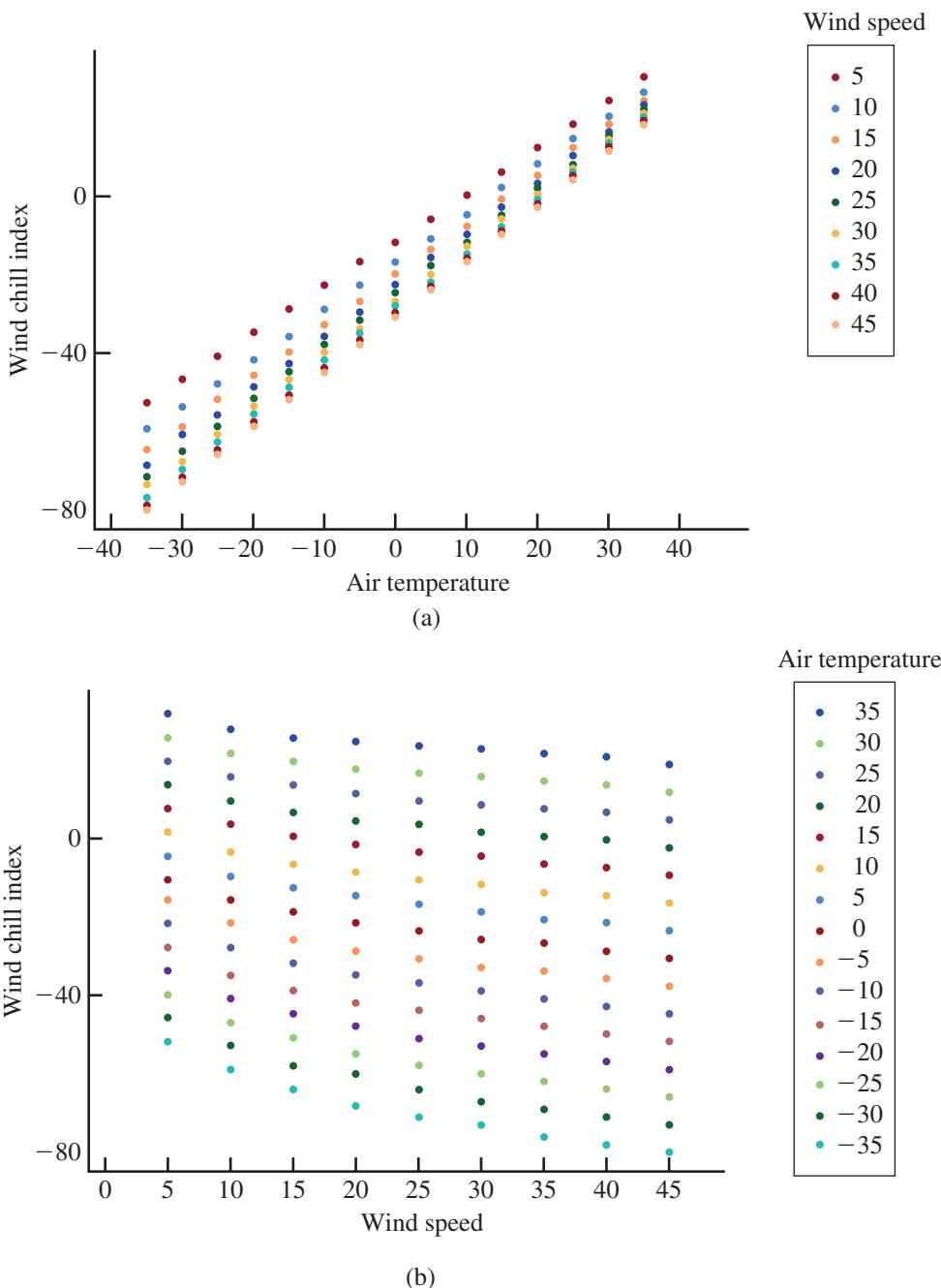
we should include both x_1 and an interaction term that involves x_2 .

Figure 14.4(b) shows a scatterplot of wind chill index versus wind speed with different temperatures denoted by different colors. This plot reveals that the relationship between wind chill index and wind speed is nonlinear at each of the different temperatures. Also, because the pattern is more markedly curved at some temperatures than at others, an interaction is suggested.

FIGURE 14.4

Scatterplots of wind chill index data of Example 14.3:

- (a) wind chill index versus air temperature;
- (b) wind chill index versus wind speed.



These observations are consistent with the new model used by the National Weather Service for relating wind chill index to air temperature and wind speed. The model used is

$$(\text{mean } y) = 35.74 + 0.621x_1 - 35.75(x'_2) + 0.4275x_1x'_2$$

where

$$x'_2 = x_2^{0.16}$$

which incorporates a variable that is a transformation of x_2 (to model the nonlinear relationship between wind chill index and wind speed) and an interaction term.

Qualitative Predictor Variables

Up to this point, we have only considered quantitative (numerical) predictor variables in a multiple regression model. Using a simple numerical coding, qualitative (categorical) variables can also be incorporated into a model. Let's focus first on a dichotomous variable, one with just two possible categories, such as married or not married, male or female, a house with or without a view, and so on. With a dichotomous variable, we associate a numerical variable x whose possible values are 0 and 1, where 0 is identified with one category (for example, married) and 1 is identified with the other possible category (for example, not married). This 0-1 variable is often called an **indicator variable**.

Example 14.4 Predictors of Writing Competence

Understand the context ➤

The article “**Grade Level and Gender Differences in Writing Self-Beliefs of Middle School Students**” (*Contemporary Educational Psychology* [1999]: 390–405) considered relating writing competence score to a number of predictor variables, including perceived value of writing and sex. Both writing competence and perceived value of writing are numerical variables, but sex is a qualitative predictor.

Consider the following variables:

$$y = \text{Writing competence score}$$

$$x_1 = \text{Sex} \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases}$$

$$x_2 = \text{Perceived value of writing}$$

One possible multiple regression model is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$$

Considering the mean y value first when $x_1 = 0$ and then when $x_1 = 1$ results in

$$(\text{average writing competence score}) = \alpha + \beta_2 x_2 \quad \text{when } x_1 = 0 \text{ (male)}$$

$$(\text{average writing competence score}) = \alpha + \beta_1 + \beta_2 x_2 \quad \text{when } x_1 = 1 \text{ (female)}$$

The coefficient β_1 is the difference in mean writing competence score between males and females when perceived value of writing is held fixed.

A second possibility is a model with an interaction term:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

The regression function for this model is $\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ where $x_3 = x_1 x_2$. Now the two cases $x_1 = 0$ and $x_1 = 1$ result in

$$(\text{average writing competence score}) = \alpha + \beta_2 x_2 \quad \text{when } x_1 = 0 \text{ (males)}$$

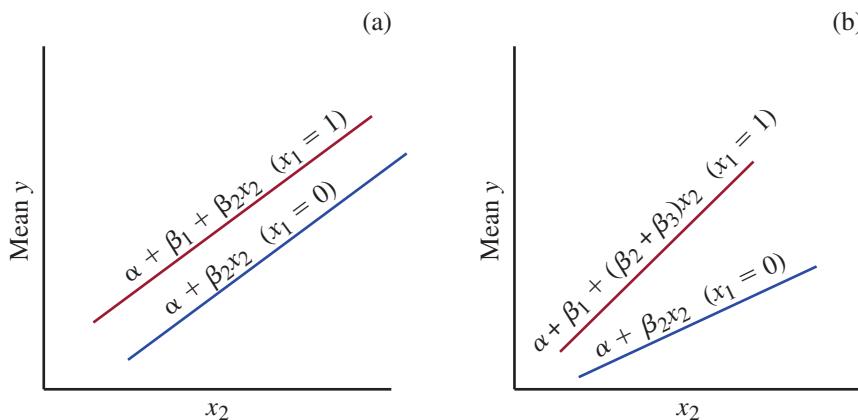
$$(\text{average writing competence score}) = \alpha + \beta_1 + (\beta_2 + \beta_3) x_2 \quad \text{when } x_1 = 1 \text{ (females)}$$

For each model, the graph of the average writing competence score, when regarded as a function of perceived value of writing, is a line for males and a line for females (Figure 14.5).

In the no-interaction model, the coefficient of x_2 is β_2 both when $x_1 = 0$ and when $x_1 = 1$. This means that the two lines are parallel, although their intercepts are different (unless $\beta_1 = 0$). With interaction, the lines not only have different intercepts but also have different slopes (unless $\beta_3 = 0$). For the interaction model, the change in average writing competence score when perceived value of writing increases by 1 unit depends on sex. The two variables *perceived value* and *sex* interact.

FIGURE 14.5

Regression functions for models with one qualitative variable (x_1) and one quantitative variable (x_2):
 (a) no interaction;
 (b) interaction.



One might think that the way to incorporate a qualitative variable with three categories is to define a single numerical variable with coded values such as 0, 1, and 2 corresponding to the three categories. This is incorrect because it imposes an ordering on the categories that may not be appropriate. The correct approach to modeling a categorical variable with three categories is to define *two* different indicator variables, as illustrated in Example 14.5.

Example 14.5 Location, Location, Location

Understand the context ➤



Rich Reid/National Geographic/Getty Images

One of the factors that has an effect on the price of a house is location. We might want to incorporate location, as well as numerical predictors such as size and age, into a multiple regression model for predicting house price.

Suppose that in a California beach community houses can be classified by location into three categories—ocean view and beachfront, ocean view but not beachfront, and no ocean view. Let

$$x_1 = \begin{cases} 1 & \text{if the house has an ocean view and is beachfront} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if the house has an ocean view but is not beachfront} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \text{House size}$$

$$x_4 = \text{House age}$$

This means that $x_1 = 1, x_2 = 0$ indicates a beachfront ocean-view house; $x_1 = 0, x_2 = 1$ indicates a house with an ocean view but not beachfront; and $x_1 = x_2 = 0$ indicates a house that does not have an ocean view. ($x_1 = x_2 = 1$ is not possible.) We could then consider a multiple regression model of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

This model allows individual adjustments to the predicted price for a house with no ocean view for the other two location categories. For example, β_1 is the amount that would be added to the predicted price for a home with no ocean view to adjust for a beachfront location (assuming that age and size were the same).

In general, incorporating a categorical variable with c possible categories into a regression model requires the use of $c - 1$ indicator variables. Even one such categorical variable can add many predictors to a model.

EXERCISES 14.1 - 14.15

- 14.1** a. Explain the difference between a deterministic and a probabilistic model.
 b. Give an example of a dependent variable y and two or more independent variables that might be related to y in a deterministic way.
 c. Give an example of a dependent variable y and two or more independent variables that might be related to y in a probabilistic way.

- 14.2** The authors of the paper “**Weight-Bearing Activity during Youth Is a More Important Factor for Peak Bone Mass than Calcium Intake**” (*Journal of Bone and Mineral Density* [1994]: 1089–1096) used a multiple regression model to describe the relationship between

y = Bone mineral density (g/cm^3)

x_1 = Body weight (kg)

x_2 = Measure of weight-bearing activity, with higher values indicating greater activity

- a. The authors concluded that both body weight and weight-bearing activity were important predictors of bone mineral density and that there was no significant interaction between body weight and weight-bearing activity. What multiple regression function is consistent with this description?
 b. The value of the coefficient of body weight in the multiple regression function given in the paper is 0.587. Interpret this value.

- 14.3** A number of investigations have focused on the problem of assessing weights that can be lifted in a safe manner. The article “**Anthropometric, Muscle Strength, and Spinal Mobility Characteristics as Predictors in the Rating of Acceptable Loads in Parcel Sorting**” (*Ergonomics* [1992]: 1033–1044) proposed using a regression model to relate the dependent variable

y = Individual’s rating of acceptable weight (kg)
 to $k = 3$ independent (predictor) variables:

x_1 = Extent of left lateral bending (cm)

x_2 = Dynamic hand grip endurance (seconds)

x_3 = Trunk extension ratio (N/kg)

Suppose that the model equation is

$$y = 30 + 0.90x_1 + 0.08x_2 - 4.50x_3 + e$$

and that $\sigma = 5$.

- a. What is the population regression function?
 b. What are the values of the population regression coefficients?
 c. Interpret the value of β_1 .
 d. Interpret the value of β_3 .

- 14.4** Consider the regression model introduced in the previous exercise.

- a. What is the mean rating of acceptable weight when extent of left lateral bending is 25 cm, dynamic hand grip endurance is 200 seconds, and trunk extension ratio is 10 N/kg?
 b. If repeated observations on rating are made on different individuals, all of whom have the values of x_1 , x_2 , and x_3 specified in Part (a), in the long run approximately what percentage of ratings will be between 13.5 kg and 33.5 kg?

- 14.5** The authors of the paper “**Predicting Yolk Height, Yolk Width, Albumen Length, Eggshell Weight, Egg Shape Index, Eggshell Thickness, Egg Surface Area of Japanese Quails Using Various Egg Traits as Regressors**” (*International Journal of Poultry Science* [2008]: 85–88) used a multiple regression model with two independent variables where

y = Quail egg weight (g)

x_1 = Egg width (mm)

x_2 = Egg length (mm)

The regression function suggested in the paper is $y = -21.658 + 0.828x_1 + 0.373x_2$.

- a. What is the mean egg weight for quail eggs that have a width of 20 mm and a length of 50 mm?
 b. Interpret the values of β_1 and β_2 .

- 14.6** According to the paper “**Assessing the Validity of the Post-Materialism Index**” (*American Political Science Review* [1999]: 649–664), it is possible to predict an individual’s level of support for ecology based on demographic and ideological characteristics. The multiple regression model proposed by the authors was

$$y = 3.60 - 0.01x_1 + 0.01x_2 - 0.07x_3 + 0.12x_4 + 0.02x_5 - 0.04x_6 - 0.01x_7 - 0.04x_8 - 0.02x_9 + e$$

where the variables are defined as follows:

y = Ecology score (higher values indicate a greater concern for ecology)

x_1 = Age times 10

x_2 = Income (in thousands of dollars)

x_3 = Sex (1 = male, 0 = female)

x_4 = Race (1 = white, 0 = nonwhite)

x_5 = Education (in years)

x_6 = Ideology (4 = conservative, 3 = right of center, 2 = middle of the road, 1 = left of center, and 0 = liberal)

x_7 = Social class (4 = upper, 3 = upper middle, 2 = middle, 1 = lower middle, and 0 = lower)

$x_8 = \text{Postmaterialist}$ (1 if postmaterialist, 0 otherwise)

$x_9 = \text{Materialist}$ (1 if materialist, 0 otherwise)

- Suppose you knew a person with the following characteristics: a 25-year-old, white female with a college degree (16 years of education), who has a \$32,000-per-year job, is from the upper middle class, and considers herself left of center, but who is neither a materialist nor a postmaterialist. Predict her ecology score.
- If the woman described in Part (a) were Hispanic rather than white, how would the prediction change? (Hint: See Example 14.5.)
- Given that the other variables are the same, what is the estimated mean difference in ecology score for men and women?
- How would you interpret the coefficient of x_2 ?
- Comment on the numerical coding of the ideology and social class variables. Can you suggest a better way of incorporating these two variables into the model? (Hint: See Example 14.5.)

- 14.7** The authors of the paper “[Power-Load Prediction Based on Multiple Linear Regression Model](#)” (*Boletin Tecnico* [2017]: 390–397) were interested in predicting the load on the electric power system in China using data on y = Power consumption (in hundreds of millions of kWh), x_1 = Population (in millions), and x_2 = Gross domestic product (in billions of dollars), for 21 years. The model equation proposed in the paper is

$$y = -113,527 + 0.974x_1 + 0.057x_2 + e$$

- According to this model, what is the mean power consumption for a year if the population was 130,000 million and the gross domestic product was 400,000 billion dollars?
- Interpret the value of β_1 in this model.

- 14.8** Groundwater is the main source of water in many countries. In a study of the quality of ground water in Yemen, the authors of the paper “[Multiple Linear Regression Model for Chloride Estimation of Groundwater in Ash-Shihr Town and Its Outskirts Hadhramout-Yemen](#)” (*International Journal of Environmental Sciences* [2016]: 881–888) used a multiple regression model with two independent variables, where

y = Chloride concentration (mg/L)

x_1 = Total alkalinity

x_2 = Electrical conductivity

The model equation suggested in the paper is

$$y = -183.560 + 1.589x_1 + 0.069x_2 + e$$

- Based on this model, what is the mean value of y when $x_1 = 300$ and $x_2 = 2500$?

- Based on this model, what mean chloride concentration is associated with a total alkalinity of 230 and an electrical conductivity of 2700?

- 14.9** The article “[Effects of Age on the Variability and Stability of Gait: A Cross-Sectional Treadmill Study in Healthy Individuals Between 20 and 69 Years of Age](#)” (*Gait and Posture* [2014]: 170–174) proposed a quadratic regression model to describe the relationship between a measure of gait instability, y , and age. The model suggested in the paper is

$$y = 0.92 - 0.005x + 0.00007x^2 + e$$

- Graph the function $y = 0.92 - 0.005 + 0.00007x^2$ over x values between 20 and 70. (Hint: Substitute $x = 20, 30, 40, 50, 60$, and 70 into the function to find points on the graph and connect them with a smooth curve, or use appropriate technology.)
- Would the mean gait instability be higher for age 40 or age 60?
- What is the change in mean gait instability when age increases from 20 to 30? From 50 to 60?

- 14.10** The relationship between yield of maize (a type of corn), date of planting, and planting density was investigated in the article “[Development of a Model for Use in Maize Replant Decisions](#)” (*Agronomy Journal* [1980]: 459–464). Let

y = Maize yield (percent)

x_1 = Planting date (days after April 20)

x_2 = Planting density (10,000 plants/ha)

The regression model with both quadratic terms ($y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + e$ where $x_3 = x_1^2$ and $x_4 = x_2^2$) provides a good description of the relationship between y and the independent variables.

- If $\alpha = 21.09$, $\beta_1 = 0.653$, $\beta_2 = 0.0022$, $\beta_3 = 2.0206$, and $\beta_4 = 0.4$, what is the population regression function?
- Use the regression function in Part (a) to determine the mean yield for a plot planted on May 6 with a density of 41,180 plants/ha.
- Would the mean yield be higher for a planting date of May 6 or May 22 (for the same density)?
- Is it appropriate to interpret $\beta_1 = 0.653$ as the average change in yield when planting date increases by one day and the values of the other three predictors are held fixed? Why or why not?

- 14.11** Suppose that the variables y , x_1 , and x_2 are related by the regression model

$$y = 1.8 + 0.1x_1 + 0.8x_2 + e$$

- Construct a graph (similar to that of Figure 14.5) showing the relationship between mean y and x_2 for fixed values 10, 20, and 30 of x_1 .

- b.** Construct a graph depicting the relationship between mean y and x_1 for fixed values 50, 55, and 60 of x_2 .
- c.** What aspect of the graphs in Parts (a) and (b) can be attributed to the lack of an interaction between x_1 and x_2 ?
- d.** Suppose the interaction term $0.03x_3$, where $x_3 = x_1x_2$ is added to the regression model equation. Using this new model, construct the graphs described in Parts (a) and (b). How do they differ from those of Parts (a) and (b)?
- 14.12** A manufacturer of wood stoves collected data on y = Particulate matter concentration and x_1 = Flue temperature for three different air intake settings (low, medium, and high).
- Write a model equation that includes indicator variables to incorporate intake setting, and interpret each of the β coefficients.
 - What additional predictors would be needed to incorporate interaction between temperature and intake setting?
- 14.13** Consider a regression analysis with three independent variables x_1 , x_2 , and x_3 . Give the equation for the following regression models:
- The model that includes as predictors all independent variables but no quadratic or interaction terms.
 - The model that includes as predictors all independent variables and all quadratic terms.
 - All models that include as predictors all independent variables, no quadratic terms, and exactly one interaction term.
 - The model that includes as predictors all independent variables, all quadratic terms, and all interaction terms (the full quadratic model).

- 14.14** The article “The Value and the Limitations of High-Speed Turbo-Exhausters for the Removal of Tar-Fog from Carburetted Water-Gas” (*Society of Chemical Industry Journal* [1946]: 166–168) presented data on y = Tar content (grains/100 ft³) of a gas stream as a function of x_1 = Rotor speed (rev/minute) and x_2 = Gas inlet temperature (°F). The following regression model using x_1 , x_2 , $x_3 = x_2^2$ and $x_4 = x_1x_2$ was suggested:

$$\begin{aligned} \text{(mean } y \text{ value)} &= 86.8 - 0.123x_1 + 5.09x_2 \\ &\quad - 0.0709x_3 + 0.001x_4 \end{aligned}$$

- According to this model, what is the mean y value if $x_1 = 3200$ and $x_2 = 57$?
- For this particular model, does it make sense to interpret the value of β_2 as the average change in tar content associated with a 1-degree increase in gas inlet temperature when rotor speed is held constant? Explain.

- 14.15** Consider the dependent variable y = Fuel efficiency of a car (mpg).
- Suppose that you want to incorporate type of car, with four categories (subcompact, compact, midsize, and large), into a regression model that also includes x_1 = Age of car and x_2 = Engine size. Define the necessary indicator variables, and write out the complete model equation.
 - Suppose that you want to incorporate interaction between age and type of car. What additional predictors would be needed to accomplish this?

SECTION 14.2 Fitting a Model and Assessing Its Utility

In Section 14.1, multiple regression models containing several different types of predictors were introduced. Now suppose that a particular set of k predictor variables x_1, x_2, \dots, x_k has been selected for inclusion in the model

$$y = \alpha + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k + e$$

The next steps are to estimate the model coefficients $\alpha, \beta_1, \dots, \beta_k$ and the regression function $\alpha + \beta_1x_1 + \cdots + \beta_kx_k$ (the mean y value for specified values of the predictors), assess the model’s usefulness, and if appropriate, use the estimated model to make predictions. This requires sample data. As before, n denotes the number of observations in the sample. With just one independent variable, the sample consisted of n (x, y) pairs. Now, each observation consists of $(k + 1)$ numbers: a value of x_1 , a value of x_2, \dots , a value of x_k and the associated value of y . The n observations are assumed to have been selected independently of one another.

Example 14.6 Graduation Rates at Small Colleges

Understand the context ➤

- One way colleges measure success is by graduation rates. **The Education Trust** publishes 6-year graduation rates along with other college characteristics on its website (collegeresults.org). We will consider the following variables:

$$y = \text{6-year graduation rate}$$

x_1 = Median SAT score of students accepted to the college

x_2 = Student-related expense per full-time student (in dollars)

$$x_3 = \begin{cases} 1 & \text{if college has only female students or only male students} \\ 0 & \text{if college has both male and female students} \end{cases}$$

The following data represent a random sample of 22 colleges selected from the 1037 colleges in the United States with enrollments under 5000 students. The data consist of 22 observations on each of these four variables.

College	y	x_1	x_2	x_3
Cornerstone University	0.391	1,065	9,482	0
Barry University	0.389	950	13,149	0
Wilkes University	0.532	1,090	9,418	0
Colgate University	0.893	1,350	26,969	0
Lourdes College	0.313	930	8,489	0
Concordia University at Austin	0.315	985	8,329	0
Carleton College	0.896	1,390	29,605	0
Letourneau University	0.545	1,170	13,154	0
Ohio Valley College	0.288	950	10,887	0
Chadron State College	0.469	990	6,046	0
Meredith College	0.679	1,035	14,889	1
Tougaloo College	0.495	845	11,694	0
Hawaii Pacific University	0.410	1,000	9,911	0
University of Michigan-Dearborn	0.497	1,065	9,371	0
Whittier College	0.553	1,065	14,051	0
Wheaton College	0.845	1,325	18,420	0
Southampton College of Long Island	0.465	1,035	13,302	0
Keene State College	0.541	1,005	8,098	0
Mount St Mary's College	0.579	918	12,999	1
Wellesley College	0.912	1,370	35,393	1
Fort Lewis College	0.298	970	5,518	0
Bowdoin College	0.891	1,375	35,669	0

One possible model that could be considered to describe the relationship between y and these three independent variables is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

We will return to this example after we see how sample data are used to estimate model coefficients.

As in simple linear regression, the principle of least squares is used to estimate the coefficients $\alpha, \beta_1, \dots, \beta_k$. For specified estimates a, b_1, \dots, b_k

$$y - (a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k)$$

- Data set available online

is the deviation between the observed y value for a particular observation and the predicted value using the estimated regression function $a + b_1x_1 + \dots + b_kx_k$. For example, the first observation in the data set of Example 14.6 is

$$(y, x_1, x_2, x_3) = (0.391, 1065, 9482, 0)$$

The resulting deviation between observed and predicted y values is

$$0.391 - [a + b_1(1065) + b_2(9482) + b_3(0)]$$

Deviations corresponding to other observations are found in a similar manner. Using the principle of least squares, estimates of α , β_1 , β_2 , and β_3 are the values of a , b_1 , b_2 , and b_3 that minimize the sum of these squared deviations.

According to the principle of least squares, the fit of a particular estimated regression function $a + b_1x_1 + \dots + b_kx_k$ to the observed data is measured by the sum of the squared deviations between the observed y values and the y values predicted by the estimated regression function:

$$\sum [y - (a + b_1x_1 + \dots + b_kx_k)]^2$$

The **least-squares estimates** of α , β_1, \dots, β_k are those values of a, b_1, \dots, b_k that make this sum of squared deviations as small as possible.

The least-squares estimates for a given data set are obtained by solving a system of $(k + 1)$ equations in the $(k + 1)$ unknowns a, b_1, \dots, b_k (called the *normal equations*). In the case $k = 1$ (simple linear regression), there are only two equations, and we gave their general solution—the expressions for b and a —in Chapter 5. For $k \geq 2$, it is not as easy to write general expressions for the estimates without using advanced mathematical notation. Fortunately all the commonly used statistical software packages can calculate these estimates using sample data.

Example 14.7 More on Graduation Rates at Small Colleges

Understand the context ►

Figure 14.6 displays Minitab output for the model $y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + e$ based on the small college data of Example 14.6. Focus on the column labeled Coef (for coefficient) in the table near the top of the figure. The four numbers in this column are the estimated model coefficients:

$$\begin{aligned} a &= -0.3906 \text{ (the estimate of the constant term } \alpha) \\ b_1 &= 0.0007602 \text{ (the estimate of the coefficient } \beta_1) \\ b_2 &= 0.00000697 \text{ (the estimate of the coefficient } \beta_2) \\ b_3 &= 0.12495 \text{ (the estimate of the coefficient } \beta_3) \end{aligned}$$

Interpret the results ►

We estimate that the average change in 6-year graduation rate associated with a \$1 increase in expenditure per full-time student while type of institution (same sex or coed) and median SAT score remains fixed is 0.00000697. A similar interpretation applies to b_1 .

The variable x_3 is an indicator variable that takes on a value of 1 for colleges that have either all female students or all male students. We would interpret the estimated value of $b_3 = 0.125$ as the “correction” that we would make to the predicted 6-year graduation rate of a coed college with the same median SAT and expenditure per full-time student to incorporate the difference associated with having only female or only male students.

The estimated regression function is

$$\begin{aligned} \left(\begin{array}{l} \text{estimated mean value of } y \\ \text{for specified } x_1, x_2, \text{ and } x_3 \text{ values} \end{array} \right) &= -0.3906 + 0.0007602x_1 \\ &\quad + 0.00000697x_2 + 0.12495x_3 \end{aligned}$$

FIGURE 14.6

Minitab output for the regression analysis of Example 14.7.

Regression Analysis: y versus x_1, x_2, x_3

The regression equation is

$$y = -0.391 + 0.000760 x_1 + 0.000007 x_2 + 0.125 x_3$$

Predictor	Coef	SE Coef	T	P
Constant	-0.3906	0.1976	-1.98	0.064
x_1	0.0007602	0.0002300	3.30	0.004
x_2	0.00000697	0.00000451	1.55	0.139
x_3	0.12495	0.05943	2.10	0.050

$$S = 0.0844346 \quad R-Sq = 86.1\% \quad R-Sq(adj) = 83.8\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	0.79486	0.26495	37.16	0.000
Residual Error	18	0.12833	0.00713		
Total	21	0.92318			

Coefficient
of multiple
determination = 0.861

P-value for
model utility test

Substituting $x_1 = 1000$, $x_2 = 11,000$, and $x_3 = 0$ gives

$$-0.3906 + 0.0007602(1000) + 0.00000697(11,000) + 0.12495(0) = 0.4462$$

which can be interpreted either as an estimate for the mean 6-year graduation rate of coed colleges with a median SAT of 1000 and an expenditure per full-time student of \$11,000. This is also the predicted graduation rate for a single college with these same characteristics.

Is the Model Useful?

Whether an estimated model is useful can be assessed by examining the extent to which predicted y values based on the estimated regression function are close to the y values actually observed.

The first predicted value \hat{y}_1 is obtained by taking the values of the independent variables x_1, x_2, \dots, x_k for the first sample observation and substituting these values into the estimated regression function.

Doing this successively for the remaining observations results in the **predicted values** $\hat{y}_2, \dots, \hat{y}_n$.

The **residuals** are then the differences between the observed and predicted y values $(y_1 - \hat{y}_1), (y_2 - \hat{y}_2), \dots, (y_n - \hat{y}_n)$.

In multiple regression, the predicted values and residuals are defined in the same way as in simple linear regression, but calculation of the values is more involved because there is more than one predictor. Fortunately, statistical software can be used to find the predicted values (the \hat{y} 's) and the residuals (the $(y - \hat{y})$'s).

Consider again the college data from Examples 14.6 and 14.7. Because the first y observation, $y_1 = 0.391$, was for a college with $x_1 = 1065$, $x_2 = 9482$, and $x_3 = 0$, the first predicted value is

$$\hat{y} = -0.3906 + 0.0007602(1065) + 0.00000697(9482) + 0.12495(0) = 0.485$$

The first residual is then

$$(y_1 - \hat{y}_1) = (0.391 - 0.485) = -0.094$$

The other predicted values and residuals are calculated in a similar way. The sum of residuals from a least-squares fit should be 0, except for small differences due to rounding.

As in simple linear regression, the sum of squared residuals is the basis for several important summary quantities that tell us about the usefulness of a model.

The **residual (or error) sum of squares**, **SSResid**, and **total sum of squares**, **SSTo**, are given by

$$\text{SSResid} = \sum(y - \hat{y})^2 \quad \text{SSTo} = \sum(y - \bar{y})^2$$

where \bar{y} is the mean of the y observations in the sample.

The number of degrees of freedom associated with SSResid is $n - (k + 1)$, because $(k + 1)$ df are lost in estimating the $k + 1$ coefficients $\alpha, \beta_1, \dots, \beta_k$.

An estimate of the random error variance σ^2 is given by

$$s_e^2 = \frac{\text{SSResid}}{n - (k + 1)}$$

and $s_e = \sqrt{s_e^2}$ is an estimate of σ .

The **coefficient of multiple determination**, R^2 , interpreted as the proportion of the variability in observed y values that is explained by the fitted model, is

$$R^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$$

Example 14.8 Small Colleges Revisited

Figure 14.6 gives Minitab output for the college data for a three-predictor model. The residual sum of squares is found in the Residual Error row and SS column of the table headed Analysis of Variance: SSResid = 0.12833. The associated number of degrees of freedom is $n - (k + 1) = 22 - (3 + 1) = 18$, which appears in the DF column just to the left of SSResid.

The sample mean y value is $\bar{y} = 0.5544$, and SSTo = $\sum(y - 0.5544)^2 = 0.92318$ appears in the Total row and SS column of the Analysis of Variance table just under the value of SSResid. The values of s_e , s_e^2 , and R^2 are then

$$s_e^2 = \frac{\text{SSResid}}{n - (k + 1)} = \frac{0.12833}{18} = 0.007$$

(also found in the MS column of the Minitab output)

$$s_e = \sqrt{s_e^2} = \sqrt{0.007} = 0.084$$

(which appears in the Minitab output just above the Analysis of Variance table)

$$R^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}} = 1 - \frac{0.12833}{0.92318} = 1 - 0.139 = 0.861$$

This means that the percentage of variation explained is $100R^2 = 86.1\%$, which appears on the Minitab output as R-Sq = 86.1%.

Because the value of R^2 is large and the value of s_e is not too large, the values of R^2 and s_e suggest that the chosen model is successful in relating y to the predictors.

In general, a useful model is one that results in both a large R^2 value and a small s_e value. However, there is a catch. These two conditions can be achieved by fitting a model that contains a large number of predictors. Such a model might be successful in explaining the variability in y for the data in our sample, but it almost always specifies a relationship that cannot be generalized to the population and that may be unrealistic and difficult to interpret. What we really want is a model that has relatively few predictors whose roles are easily interpreted and that also explains much of the variability in y .

All statistical software packages include R^2 and s_e in their output, and most also give SSResid. In addition, some packages compute the quantity called the **adjusted R^2** :

$$(\text{adjusted } R^2) = 1 - \left[\frac{n-1}{n-(k+1)} \right] \left(\frac{\text{SSResid}}{\text{SSTo}} \right)$$

Because the quantity in square brackets is greater than 1, the number subtracted from 1 is larger than SSResid/SSTo, so the adjusted R^2 is smaller than R^2 . The value of R^2 must be between 0 and 1, but the adjusted R^2 can, on rare occasions, be negative.

If a large R^2 has been achieved by using just a few independent variables, the adjusted R^2 and R^2 values will not differ greatly. However, the adjustment can be substantial when a large number of predictors (relative to the number of observations) have been used or when R^2 itself is small to moderate (which could happen even when there is no relationship between y and the independent variables). In Example 14.7, the adjusted $R^2 = 0.838$, which is not much less than $R^2 = 0.861$ because the model included only two independent variables and the sample size was 22.

F Distributions

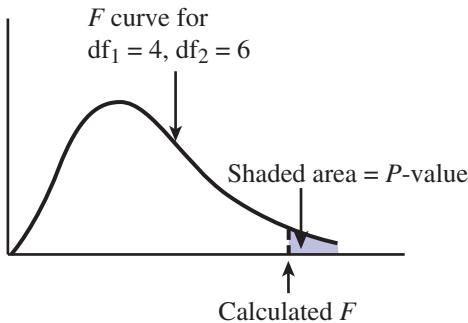
The model utility test in simple linear regression was based on a test statistic that has a t distribution when $H_0: \beta = 0$ is true. The model utility test for multiple regression is based on a test statistic that has a probability distribution called an *F distribution*.

An *F* distribution arises in connection with a ratio in which the numerator is based on one sum of squares and the denominator is based on a second sum of squares. Each sum of squares has a specified number of degrees of freedom associated with it, so a particular *F* distribution is determined by specifying values of df_1 = numerator degrees of freedom and df_2 = denominator degrees of freedom. There is a different *F* distribution for each different df_1 and df_2 combination.

For example, there is an *F* distribution with 4 numerator degrees of freedom and 12 denominator degrees of freedom, another *F* distribution with 3 numerator degrees of freedom and 20 denominator degrees of freedom, and so on. A typical *F* curve for specified numerator and denominator degrees of freedom appears in Figure 14.7.

All *F* tests introduced in this textbook are upper-tailed. The *P*-value for an upper-tailed *F* test is the area under the associated *F* curve to the right of the calculated *F*. Figure 14.7 illustrates this for a test with $df_1 = 4$ and $df_2 = 6$.

FIGURE 14.7
A *P*-value for an upper-tailed *F* test.

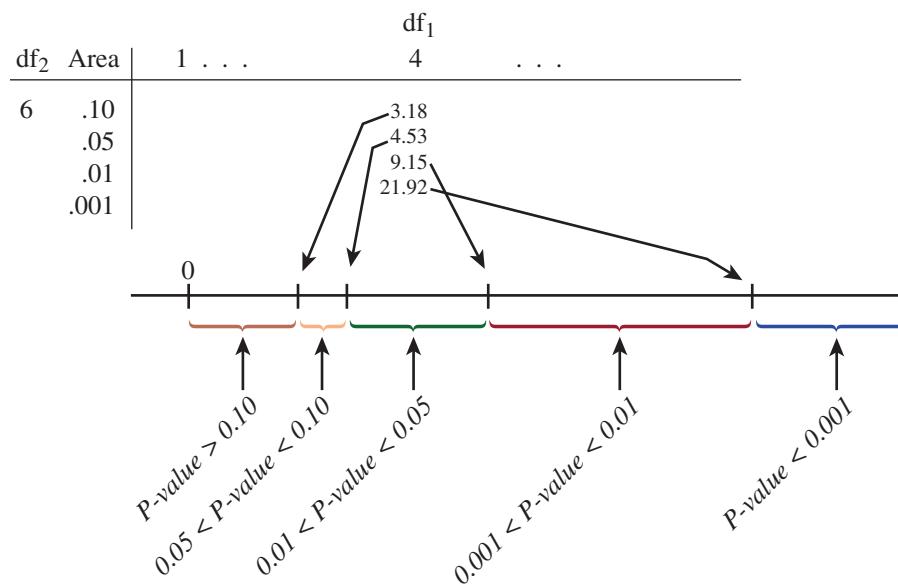


Unfortunately, tables of these upper-tail areas are more complicated than are tables for *t* distributions, because two *df* are involved. The *F* table (Appendix Table 6) gives only four numbers: the values that capture tail areas 0.10, 0.05, 0.01, and 0.001. Different columns in

the table correspond to different values of df_1 , and each different group of rows is for a different value of df_2 . Figure 14.8 shows how this table is used to obtain P -value information.

FIGURE 14.8

Obtaining P -value information from the F table.



For example, for a test with $df_1 = 4$ and $df_2 = 6$,

calculated $F = 5.70 \rightarrow 0.01 < P\text{-value} < 0.05$

calculated $F = 2.16 \rightarrow P\text{-value} > 0.10$

calculated $F = 25.03 \rightarrow P\text{-value} < 0.001$

Only if calculated F equals a tabulated value do we obtain an exact P -value (for example, if calculated $F = 4.53$, then $P\text{-value} = 0.05$). If $0.01 < P\text{-value} < 0.05$, we should reject the null hypothesis at a significance level of 0.05 but not at a level of 0.01. When $P\text{-value} < 0.001$, H_0 would be rejected at any reasonable significance level. Statistical software packages such as Minitab, and some graphing calculators can also be used to find P -values for F distributions.

The F Test for Model Utility

In the simple linear model with regression function $y = \alpha + \beta x$, if $\beta = 0$, there is no useful linear relationship between y and the single independent variable x . Similarly, if all k coefficients $\beta_1, \beta_2, \dots, \beta_k$ are 0 in the general k -predictor multiple regression model, there is no useful linear relationship between y and *any* of the independent variables x_1, x_2, \dots, x_k included in the model. Before using an estimated multiple regression model to make further inferences (for example, predictions or estimates of mean values), you should confirm that the model is useful using a formal test procedure.

Recall that SSTo is a measure of total variability in the observed y values and that SSResid measures the amount of total variability that has not been explained by the fitted model. The difference between total and error sums of squares is itself a sum of squares, called the **regression sum of squares**, which is denoted by SSRegr:

$$\text{SSRegr} = \text{SSTo} - \text{SSResid}$$

SSRegr is interpreted as the amount of total variability that *has* been explained by the model. Intuitively, the model should be judged useful if SSRegr is large relative to SSResid and the model uses a small number of predictors relative to the sample size. The number of degrees of freedom associated with SSRegr is k , the number of model predictors, and the number of degrees of freedom for SSResid is $n - (k + 1)$.

The model utility F test is based on the following result:

When all $k \beta_i$'s are 0 in the model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$ and when the distribution of e is normal with mean 0 and variance σ^2 for any particular values of x_1, x_2, \dots, x_k , the statistic

$$F = \frac{\text{SSRegr}/k}{\text{SSResid}/(n - (k + 1))}$$

has an F distribution with numerator $\text{df}_1 = k$ and denominator $\text{df}_2 = n - (k + 1)$.

The value of F tends to be larger when at least one β_i is not 0 than when all the β_i 's are 0, because more variability is typically explained by the model. An F statistic value far out in the upper tail of the associated F distribution can be interpreted as evidence that at least one β_i is not equal to 0. This is why the model utility F test is upper-tailed.

The Model Utility F Test for Multiple Regression

Null hypothesis: $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

(There is no useful linear relationship between y and *any* of the predictors.)

Alternative hypothesis: $H_a: \text{At least one among } \beta_1, \dots, \beta_k \text{ is not zero.}$ (There is a useful linear relationship between y and *at least one* of the predictors.)

Test statistic: $F = \frac{\text{SSRegr}/k}{\text{SSResid}/(n - (k + 1))}$

where $\text{SSRegr} = \text{SSTo} - \text{SSResid}$.

An equivalent formula is

$$F = \frac{R^2/k}{(1 - R^2)/(n - (k + 1))}$$

The test is upper-tailed, and the information in Appendix Table 6 can be used to obtain a bound or bounds on the P -value using numerator $\text{df}_1 = k$ and denominator $\text{df}_2 = n - (k + 1)$. A statistical software package could also be used to obtain P -values.

Assumptions: For any particular combination of predictor variable values, the distribution of e , the random deviation, is *normal* with mean 0 and *constant variance*, σ^2 .

For the model utility test, the null hypothesis is the claim that the model is not useful. Unless H_0 can be rejected, the model has not demonstrated that it is useful.

Example 14.9 Small Colleges One Last Time

The model fit to the college data introduced in Example 14.6 included $k = 3$ predictors. The Minitab output in Figure 14.6 contains the relevant information for carrying out the model utility test.

Understand the context ➤

1. The model is $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$ where $y = 6$ -year graduation rate, $x_1 = \text{Median SAT score}$, $x_2 = \text{Expenditure per full-time student}$, and x_3 is an indicator variable that is equal to 1 for a college that has only female or only male students and is equal to 0 if the college is coed.
2. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ (the model is not useful)

3. H_a : At least one of the three β_i 's is not zero.

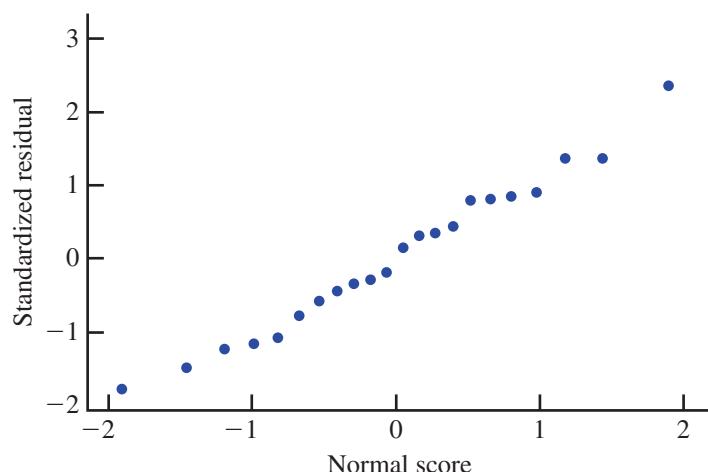
4. Significance level: $\alpha = 0.05$

Formulate a plan ➤ 5. Test statistic: $F = \frac{\text{SSRegr}/k}{\text{SSResid}/[n - (k + 1)]}$

6. Assumptions: The accompanying table gives the residuals and standardized residuals (from Minitab) for the model under consideration.

Observation	<i>y</i>	<i>x</i> ₁	<i>x</i> ₂	<i>x</i> ₃	Residual	Standardized Residual
1	0.391	1,065	9,482	0	-0.094	-1.166
2	0.389	950	13,149	0	-0.034	-0.442
3	0.532	1,090	9,418	0	0.028	0.358
4	0.893	1,350	26,969	0	0.069	0.908
5	0.313	930	8,489	0	-0.062	-0.779
6	0.315	985	8,329	0	-0.101	-1.244
7	0.896	1,390	29,605	0	0.024	0.319
8	0.545	1,170	13,154	0	-0.045	-0.575
9	0.288	950	10,887	0	-0.119	-1.497
10	0.469	990	6,046	0	0.065	0.812
11	0.679	1,035	14,889	1	0.054	0.806
12	0.495	845	11,694	0	0.162	2.388
13	0.410	1,000	9,911	0	-0.029	-0.350
14	0.497	1,065	9,371	0	0.013	0.158
15	0.553	1,065	14,051	0	0.036	0.441
16	0.845	1,325	18,420	0	0.100	1.381
17	0.465	1,035	13,302	0	-0.024	-0.292
18	0.541	1,005	8,098	0	0.111	1.371
19	0.579	918	12,999	1	0.056	0.857
20	0.912	1,370	35,393	1	-0.110	-1.793
21	0.298	970	5,518	0	-0.087	-1.091
22	0.891	1,375	35,669	0	-0.012	-0.189

A normal probability plot of the standardized residuals is shown below. The plot is reasonably straight, indicating that the assumption of normality of the random deviation distribution is plausible.



Do the work ➤

7. Directly from the table in Figure 14.6, the SS column gives SSRegr = 0.79486 and SSResid = 0.12833. Then

$$F = \frac{0.79486/3}{0.12833/18} = \frac{0.26495}{0.00713} = 37.16 \quad \left(\begin{array}{l} \text{also found in the column labeled } F \\ \text{in Figure 14.6} \end{array} \right)$$

Interpret the results ➤

8. Appendix Table 6 shows that for a test with $df_1 = k = 3$ and $df_2 = n - (k + 1) = 22 - (3 + 1) = 18$, the value 8.49 captures upper-tail F curve area 0.001. Since calculated $F = 37.16 > 8.49$, it follows that $P\text{-value} < 0.001$.
9. Because $P\text{-value} < 0.001$, which is less than the significance level of 0.05, H_0 should be rejected. The conclusion would be the same using $\alpha = 0.01$ or $\alpha = 0.001$. The usefulness of the multiple regression model is confirmed. The P -value for the model utility test ($P\text{-value} = 0.000$) can also be found in the Minitab output to the right of the value of the F test statistic in the column labeled P.

Example 14.10 | Compression Strength of Wood

Understand the context ➤

Wood is sometimes processed using a heat treatment to improve durability and resistance to fungi. A multiple regression analysis presented in the article “[An Artificial Neural Network Model for Predicting Compression Strength of Heat Treated Woods and Comparison with a Multiple Linear Regression Model](#)” (*Construction and Building Materials* [2014]: 102–108) considered a model in which the dependent variable was

$$y = \text{Compression strength of heat-treated wood (N/mm}^2\text{)}$$

and the predictors were

$$\begin{aligned} x_1 &= \text{Heat treatment temperature } (\text{°C}) \\ x_2 &= \text{Heat treatment time (hours)} \end{aligned}$$

Summary quantities included $n = 48$ and $R^2 = 0.83$. Does this model specify a useful relationship between y and the two predictors? To answer this question, we carry out a model utility test:

1. The fitted model was $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$
2. $H_0: \beta_1 = \beta_2 = 0$
3. $H_a: \text{at least one of the } \beta_i\text{'s is not zero}$
4. Significance level: $\alpha = 0.01$

Formulate a plan ➤

$$5. \text{ Test statistic: } F = \frac{R^2/k}{(1 - R^2)/(n - (k + 1))}$$

6. Assumptions: The raw data were not given in this article, so we are unable to compute standardized residuals or construct a normal probability plot. For this test to be valid, we must be willing to assume that the random deviation distribution is normal.

Do the work ➤

$$7. F = \frac{0.83/2}{(0.17)/(48 - (2 + 1))} = \frac{0.415}{0.004} = 103.75$$

Interpret the results ➤

8. The test has $df_1 = k = 2$ and $df_2 = n - (k + 1) = 45$. This latter df is not included in the F table. However, the 0.001 cutoff value for $df_2 = 40$ is 8.25, and for $df_2 = 60$ it is 7.77. The value of the test statistic ($F = 103.75$) is much greater than these values, implying that the $P\text{-value} < 0.001$.
9. Since $P\text{-value} < 0.001$ which is less than the significance level of 0.01, H_0 is rejected at significance level 0.01. There appears to be a useful linear relationship between y and at least one of the two predictors.

In Section 14.3 (online), we will see how a model that has been judged to be useful can be used to draw further conclusions. However, it is important to realize that in many applications, more than one model's utility could be confirmed by the F test. Also, just because the model utility test indicates that the multiple regression model is useful does not necessarily mean that all the predictors included in the model contribute to the usefulness of the model. This is illustrated in Example 14.11, and strategies for selecting a model are considered later in Section 14.4 (online).

Example 14.11 The Cost of Energy Bars

Understand the context ➤

- What factors contribute to the price of energy bars promoted to provide endurance and increase muscle power? The article “**Energy Bars, Unwrapped**” (*Consumer Reports* [June 2003] 19–21) included the following data on price, calorie content, protein content (in grams), and fat content (in grams) for a sample of 19 energy bars.

Price	Calories	Protein	Fat
1.40	180	12	3.0
1.28	200	14	6.0
1.31	210	16	7.0
1.10	220	13	6.0
2.29	220	17	11.0
1.15	230	14	4.5
2.24	240	24	10.0
1.99	270	24	5.0
2.57	320	31	9.0
0.94	110	5	30.0
1.40	180	10	4.5
0.53	200	7	6.0
1.02	220	8	5.0
1.13	230	9	6.0
1.29	230	10	2.0
1.28	240	10	4.0
1.44	260	6	5.0
1.27	260	7	5.0
1.47	290	13	6.0

Figure 14.9 displays Minitab output for the model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

where

$$y = \text{Price} \quad x_1 = \text{Calorie content} \quad x_2 = \text{Protein content} \quad x_3 = \text{Fat content}$$

FIGURE 14.9

Minitab output for the energy bar data of Example 14.11.

The regression equation is
 $\text{Price} = 0.252 + 0.00125 \text{ Calories} + 0.0485 \text{ Protein} + 0.0444 \text{ Fat}$

Predictor	Coef	SE Coef	T	P
Constant	0.2511	0.3524	0.71	0.487
Calories	0.001254	0.001724	0.73	0.478
Protein	0.04849	0.01353	3.58	0.003
Fat	0.04445	0.03648	1.22	0.242

S = 0.2789 R-Sq = 74.7% R-Sq(adj) = 69.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	3.4453	1.1484	14.76	0.000
Residual Error	15	1.1670	0.0778		
Total	18	4.6122			

Interpret the results ➤ From the Minitab output, $F = 14.76$, with an associated P -value of 0.000, indicating that the null hypothesis in the model utility test, $H_0: \beta_1 = \beta_2 = \beta_3 = 0$, should be rejected. We would conclude that there is a useful linear relationship between y and at least one of x_1 , x_2 , and x_3 . However, consider the Minitab output shown in Figure 14.10, which resulted from fitting a model that uses only x_2 = Protein content as a predictor. Notice that the F test would also confirm the usefulness of this model and also that the R^2 and adjusted R^2 values of 71.1% and 69.4% are quite similar to those of the model that included all three predictors (74.7% and 69.6% from the Minitab output of Figure 14.9). This suggests that protein content alone explains about the same amount of the variability in price as all three variables together, and so the simpler model with just one predictor may be preferred over the more complicated model with three independent variables.

FIGURE 14.10

Minitab output for the energy bar data of Example 14.11 when only x_2 = protein content is included as a predictor.

The regression equation is
Price = 0.607 + 0.0623 Protein
Predictor Coef SE Coef T P
Constant 0.6072 0.1419 4.28 0.001
Protein 0.062256 0.009618 6.47 0.000
S = 0.279843 R-Sq = 71.1% R-Sq(adj) = 69.4%
Analysis of Variance
Source DF SS MS F P
Regression 1 3.2809 3.2809 41.90 0.000
Residual Error 17 1.3313 0.0763
Total 18 4.6122

EXERCISES 14.16 - 14.35

● Data set available online

- 14.16** Because it is difficult to measure body fat, researchers have considered different ways to predict body fat based on other characteristics. One model proposed in the paper “[A Comparison Between Multiple Regression Models an CUNBAE Equation to Predict Body Fat in Adults](#)” (*PLOS ONE* [2015]: 1–15) included three predictor variables. The variables in the model were

$$\begin{aligned}y &= \text{Percentage of body fat mass} \\x_1 &= \log \text{ of body mass index } (\log(\text{kg}/\text{m}^2)) \\x_2 &= \log \text{ of age } (\log(\text{years})) \\x_3 &= \text{Sex } (0 = \text{male}, 1 = \text{female})\end{aligned}$$

The estimated regression equation based on a sample of $n = 3200$ was

$$\hat{y} = -96.07 + 76.91x_1 + 6.65x_2 + 11.40x_3$$

Suppose that $\text{SSRegr} = 540,000$ and $\text{SSResid} = 180,000$ (these values were not given in the paper but are consistent with other given information).

- Interpret the values of b_2 and b_3 .
- What proportion of observed variability in percentage of body fat mass can be explained by the model relationship?

- Estimate the value of σ .
- Calculate the value of adjusted R^2 . How does it compare to the value of R^2 calculated in Part (b)?

- 14.17** State as much information as you can about the P -value for an upper-tailed F test in each of the following situations. (Hint: See the section on F distributions.)

- $df_1 = 3$, $df_2 = 15$, calculated $F = 4.23$
- $df_1 = 4$, $df_2 = 18$, calculated $F = 1.95$
- $df_1 = 5$, $df_2 = 20$, calculated $F = 4.10$
- $df_1 = 4$, $df_2 = 35$, calculated $F = 4.58$

- 14.18** Give as much information as you can about the P -value for the model utility F test in each of the following situations:

- $k = 2$, $n = 21$, calculated $F = 2.47$
- $k = 8$, $n = 25$, calculated $F = 5.98$
- $k = 5$, $n = 26$, calculated $F = 3.00$
- The full quadratic model based on x_1 and x_2 is fit, $n = 20$, and calculated $F = 8.25$.
- $k = 5$, $n = 100$, calculated $F = 2.33$

- 14.19** Data from a sample of $n = 150$ quail eggs were used to fit a multiple regression model relating

y = Eggshell surface area (mm^2)

x_1 = Egg weight (g)

x_2 = Egg width (mm)

x_3 = Egg length (mm)

(“Predicting Yolk Height, Yolk Width, Albumen Length, Eggshell Weight, Egg Shape Index, Eggshell Thickness, Egg Surface Area of Japanese Quails Using Various Egg Traits as Regressors,” *International Journal of Poultry Science* [2008]: 85–88).

The resulting estimated regression function was

$$\hat{y} = 10.561 + 1.535x_1 - 0.178x_2 - 0.045x_3$$

and $R^2 = 0.996$.

- Carry out a model utility test to determine if this multiple regression model is useful. (Hint: See Example 14.10.)
- A simple linear regression model was also used to describe the relationship between y and x_1 , resulting in the estimated regression function $\hat{y} = 6.254 + 1.387x_1$. The P -value for the associated model utility test was reported to be less than 0.01, and $r^2 = 0.994$. Is the linear model useful? Explain. (Hint: See Example 14.11.)
- Based on your answers to Parts (a) and (b), which of the two models would you recommend for predicting eggshell surface area? Explain the rationale for your choice.

- 14.20** • This exercise requires the use of a statistical software package. The paper “Habitat Selection by Black Bears in an Intensively Logged Boreal Forrest” (*Canadian Journal of Zoology* [2008]: 1307–1316) gave the accompanying data on $n = 11$ female black bears.

Age (years)	Weight (kg)	Home Range Size (km^2)
10.5	54	43.1
6.5	40	46.6
28.5	62	57.4
6.5	55	35.6
7.5	56	62.1
6.5	62	33.9
5.5	42	39.6
7.5	40	32.2
11.5	59	57.2
9.5	51	24.4
5.5	50	68.7

- Fit a multiple regression model to describe the relationship between y = Home range size and the predictors x_1 = Age and x_2 = Weight.
- Construct a normal probability plot of the 11 standardized residuals. Based on the plot, does it seem reasonable to regard the random deviation distribution as approximately normal? Explain. (Hint: See Example 14.9.)

- If appropriate, carry out a model utility test with a significance level of 0.05 to determine if at least one of the predictors age and $weight$ is useful for predicting home range size.

- 14.21** The ability of ecologists to identify regions of greatest species richness could have an impact on the preservation of genetic diversity, a major objective of the World Conservation Strategy. The article “Prediction of Rarities from Habitat Variables: Coastal Plain Plants on Nova Scotian Lakeshores” (*Ecology* [1992]: 1852–1859) used a sample of $n = 37$ lakes to obtain the estimated regression equation

$$\begin{aligned}\hat{y} = & 3.89 + 0.033x_1 + 0.024x_2 + 0.023x_3 \\ & + 0.008x_4 - 0.13x_5 - 0.72x_6\end{aligned}$$

where y = Species richness, x_1 = Watershed area, x_2 = Shore width, x_3 = Drainage (%), x_4 = Water color (total color units), x_5 = Sand (%), and x_6 = Alkalinity. The coefficient of multiple determination was reported as $R^2 = 0.83$. Use a test with significance level 0.01 to decide whether the chosen model is useful.

- 14.22** The article “Impacts of On-Campus and Off-Campus Work on First-Year Cognitive Outcomes” (*Journal of College Student Development* [1994]: 364–370) reported on a study in which y = spring math comprehension score was thought to be related to predictors x_1 = previous fall test score, x_2 = Previous fall academic motivation, x_3 = Age, x_4 = Number of credit hours, x_5 = Residence (1 if on campus, 0 otherwise), x_6 = Hours worked on campus, and x_7 = Hours worked off campus.

The sample size was $n = 210$, and $R^2 = 0.543$. Test to see whether there is a useful relationship between y and at least one of the predictors.

- 14.23** The paper referenced in Exercise 14.16 used a multiple regression model to describe the relationship between y = Percentage of body fat mass and three predictor variables.

x_1 = log of body mass index ($\log(\text{kg}/\text{m}^2)$)

x_2 = log of age ($\log(\text{years})$)

x_3 = Sex (0 = male, 1 = female)

Exercise 14.16 also provided the following information: $n = 3200$, $\text{SSRegr} = 540,000$, and $\text{SSresid} = 180,000$. Is the model proposed useful? Carry out a test using a significance level of 0.10.

- 14.24** The accompanying Minitab output results from fitting the model described in Exercise 14.14.

Predictor	Coef	Stdev	t-ratio
Constant	86.85	85.39	1.02
X1	-0.12297	0.03276	-3.75
X2	5.090	1.969	2.58
X3	-0.07092	0.01799	-3.94
X4	0.00015380	0.0005560	2.77
S	4.784	R-sq	= 90.8% R-sq(adj) = 89.4%

Analysis of Variance

	DF	SS	MS
Regression	4	5896.6	1474.2
Error	26	595.1	22.9
Total	30	6491.7	

- What is the estimated regression equation?
- Using a 0.01 significance level, carry out the model utility test. (Hint: See Example 14.9.)
- Interpret the values of R^2 and s_e given in the output.

14.25 For the multiple regression model in Exercise 14.6, the value of R^2 was 0.06 and the adjusted R^2 was 0.06. The model was based on a data set with 1136 observations. Carry out a model utility F test.

14.26 *This exercise requires the use of a statistical software package.* The article “Movement and Habitat Use by Lake Whitefish During Spawning in a Boreal Lake: Integrating Acoustic Telemetry and Geographic Information Systems” (*Transactions of the American Fisheries Society* [1999]: 939–952) included the accompanying data on 17 fish caught in 2 consecutive years.

Year	Fish Number	Weight (g)	Length (mm)	Age (years)
Year 1	1	776	410	9
	2	580	368	11
	3	539	357	15
	4	648	373	12
	5	538	361	9
	6	891	385	9
	7	673	380	10
	8	783	400	12
Year 2	9	571	407	12
	10	627	410	13
	11	727	421	12
	12	867	446	19
	13	1,042	478	19
	14	804	441	18
	15	832	454	12
	16	764	440	12
	17	727	427	12

- Fit a multiple regression model to describe the relationship between weight and the predictors *length* and *age*.
- Carry out the model utility test to determine whether at least one of the predictors *length* and *age* are useful for predicting weight.

14.27 *This exercise requires the use of a statistical software package.* The authors of the article “Absolute Versus per Unit Body Length Speed of Prey as an Estimator of Vulnerability to Predation” (*Animal Behaviour* [1999]: 347–352) found that the speed of a

prey (twips/s) and the length of a prey (twips $\times 100$) are good predictors of the time (seconds) required to catch the prey. (A twip is a measure of distance used by programmers.) Data were collected in an experiment in which subjects were asked to “catch” an animal of prey moving across his or her computer screen by clicking on it with the mouse. The investigators varied the length of the prey and the speed with which the prey moved across the screen.

The following data are consistent with summary values and a graph given in the article. Each value represents the average catch time over all subjects. The order of the various speed-length combinations was randomized for each subject.

Prey Length	Prey Speed	Catch Time
7	20	1.10
6	20	1.20
5	20	1.23
4	20	1.40
3	20	1.50
3	40	1.40
4	40	1.36
6	40	1.30
7	40	1.28
7	80	1.40
6	60	1.38
5	80	1.40
7	100	1.43
6	100	1.43
7	120	1.70
5	80	1.50
3	80	1.40
6	100	1.50
3	120	1.90

- Fit a multiple regression model for predicting catch time using prey length and speed as predictors.
- Predict the catch time for an animal of prey whose length is 6 and whose speed is 50.
- Is the multiple regression model useful for predicting catch time? Test the relevant hypotheses using $\alpha = 0.05$.
- The authors of the article suggest that a simple linear regression model with the single predictor

$$x = \frac{\text{length}}{\text{speed}}$$

might be a better model for predicting catch time. Calculate these x values and use them to fit a simple linear regression model.

- Which of the two models considered (the multiple regression model from Part (a) or the simple linear regression model from Part (d)) would you recommend for predicting catch time? Justify your choice.

14.28 • This exercise requires the use of a statistical software package. The article “[Vital Dimensions in Volume Perception: Can the Eye Fool the Stomach?](#)” (*Journal of Marketing Research* [1999]: 313–326) gave the data, shown in the table below, on dimensions of 27 representative food products.

- Fit a multiple regression model for predicting the volume (in ml) of a package based on its minimum width, maximum width, and elongation score.
- Why should we consider adjusted R^2 instead of R^2 when evaluating this model?
- Carry out a model utility F test.

14.29 • The article “[The Undrained Strength of Some Thawed Permafrost Soils](#)” (*Canadian Geotechnical Journal* [1979]: 420–427) contained the accompanying data on y = Shear strength of sandy soil (kPa), x_1 = Depth (m), and x_2 = Water content (%). The predicted values and residuals were calculated using the estimated regression equation

$$\hat{y} = -151.36 - 16.22x_1 + 13.48x_2 + 0.094x_3 - 0.253x_4 + 0.492x_5$$

where $x_3 = x_1^2$, $x_4 = x_2^2$, and $x_5 = x_1x_2$.

y	x_1	x_2	Predicted y	Residual
14.7	8.9	31.5	23.35	-8.65
48.0	36.6	27.0	46.38	1.62
25.6	36.8	25.9	27.13	-1.53
10.0	6.1	39.1	10.99	-0.99
16.0	6.9	39.2	14.10	1.90
16.8	6.9	38.3	16.54	0.26
20.7	7.3	33.9	23.34	-2.64
38.8	8.4	33.8	25.43	13.37
16.9	6.5	27.9	15.63	1.27
27.0	8.0	33.1	24.29	2.71
16.0	4.5	26.3	15.36	0.64
24.9	9.9	37.8	29.61	-4.71
7.3	2.9	34.6	15.38	-8.08
12.8	2.0	36.4	7.96	4.84

- Use the given information to calculate SSResid, SSTo, and SSRegr.
- Calculate R^2 for this regression model. How would you interpret this value?
- Use the value of R^2 from Part (b) and a 0.05 level of significance to carry out a model utility F test.

14.30 The paper referenced in Exercise 14.9 used a quadratic regression model to describe the relationship

Table for exercise 14.28

Product	Material	Height	Maximum Width	Minimum Width	Elongation	Volume
1	glass	7.7	2.50	1.80	1.50	125
2	glass	6.2	2.90	2.70	1.07	135
3	glass	8.5	2.15	2.00	1.98	175
4	glass	10.4	2.90	2.60	1.79	285
5	plastic	8.0	3.20	3.15	1.25	330
6	glass	8.7	2.00	1.80	2.17	90
7	glass	10.2	1.60	1.50	3.19	120
8	plastic	10.5	4.80	3.80	1.09	520
9	plastic	3.4	5.90	5.00	0.29	330
10	plastic	6.9	5.80	4.75	0.59	570
11	tin	10.9	2.90	2.80	1.88	340
12	plastic	9.7	2.45	2.10	1.98	175
13	glass	10.1	2.60	2.20	1.94	240
14	glass	13.0	2.60	2.60	2.50	240
15	glass	13.0	2.70	2.60	2.41	360
16	glass	11.0	3.10	2.90	1.77	310
17	cardboard	8.7	5.10	5.10	0.85	635
18	cardboard	17.1	10.20	10.20	0.84	1,250
19	glass	16.5	3.50	3.50	2.36	650
20	glass	16.5	2.70	1.20	3.06	305
21	glass	9.7	3.00	1.70	1.62	315
22	glass	17.8	2.70	1.75	3.30	305
23	glass	14.0	2.50	1.70	2.80	245
24	glass	13.6	2.40	1.20	2.83	200
25	plastic	27.9	4.40	1.20	3.17	1,205
26	tin	19.5	7.50	7.50	1.30	2,330
27	tin	13.8	4.25	4.25	1.62	730

between y = Gait instability and x = Age. The sample size in this study was $n = 100$ and $R^2 = 0.15$. Carry out a model utility test for the quadratic model. Use $\alpha = 0.05$.

- 14.31** The article “**Effect of Manual Defoliation on Pole Bean Yield**” (*Journal of Economic Entomology* [1984]: 1019–1023) used a quadratic regression model to describe the relationship between y = Yield (kg/plot) and x = Defoliation level (a proportion between 0 and 1). The estimated regression equation based on $n = 24$ was $\hat{y} = 12.39 + 6.67x_1 - 15.25x_2$ where $x_1 = x$ and $x_2 = x^2$. The article also reported that R^2 for this model was 0.902.

Does the quadratic model specify a useful relationship between y and x ? Carry out the appropriate test using a 0.01 level of significance.

- 14.32** Suppose that a multiple regression data set consists of $n = 15$ observations. For what values of k , the number of model predictors, would the corresponding model with $R^2 = 0.90$ be judged useful at significance level 0.05? Does such a large R^2 value necessarily imply a useful model? Explain.

- 14.33** This exercise requires the use of a statistical software package. Use the data given in Exercise 14.29 to verify that the regression function

$$\text{(mean } y \text{ value)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

is estimated by

$$\hat{y} = -151.36 - 16.22x_1 + 13.48x_2 + 0.094x_3 - 0.253x_4 + 0.492x_5$$

- 14.34** • This exercise requires the use of a statistical software package. The accompanying data resulted from a study of the relationship between y = Brightness of finished paper and the independent variables x_1 = Hydrogen peroxide (% by weight), x_2 = Sodium hydroxide (% by weight), x_3 = Silicate (% by weight), and x_4 = Process temperature (“**Advantages of CE-HDP Bleaching for High Brightness Kraft Pulp Production**,” *TAPPI* [1964]: 107A–173A).

x_1	x_2	x_3	x_4	y
0.2	0.2	1.5	145	83.9
0.4	0.2	1.5	145	84.9
0.2	0.4	1.5	145	83.4
0.4	0.4	1.5	145	84.2
0.2	0.2	3.5	145	83.8
0.4	0.2	3.5	145	84.7
0.2	0.4	3.5	145	84.0
0.4	0.4	3.5	145	84.8
0.2	0.2	1.5	175	84.5
0.4	0.2	1.5	175	86.0
0.2	0.4	1.5	175	82.6
0.4	0.4	1.5	175	85.1
0.2	0.2	3.5	175	84.5

x_1	x_2	x_3	x_4	y
0.4	0.2	3.5	175	86.0
0.2	0.4	3.5	175	84.0
0.4	0.4	3.5	175	85.4
0.1	0.3	2.5	160	82.9
0.5	0.3	2.5	160	85.5
0.3	0.1	2.5	160	85.2
0.3	0.5	2.5	160	84.5
0.3	0.3	0.5	160	84.7
0.3	0.3	4.5	160	85.0
0.3	0.3	2.5	130	84.9
0.3	0.3	2.5	190	84.0
0.3	0.3	2.5	160	84.5
0.3	0.3	2.5	160	84.7
0.3	0.3	2.5	160	84.6
0.3	0.3	2.5	160	84.9
0.3	0.3	2.5	160	84.9
0.3	0.3	2.5	160	84.5
0.3	0.3	2.5	160	84.6

- a. Find the estimated regression equation for the model that includes all independent variables, all quadratic terms, and all interaction terms.
- b. Using a 0.05 significance level, carry out the model utility test.
- c. Interpret the values of the following quantities: SSResid, R^2 , and s_e .

- 14.35** • This exercise requires the use of a statistical software package. The cotton aphid poses a threat to cotton crops. The accompanying data on

y = Infestation rate (aphids/100 leaves)

x_1 = Mean temperature (°C)

x_2 = Mean relative humidity

appeared in the article “**Estimation of the Economic Threshold of Infestation for Cotton Aphid**” (*Mesopotamia Journal of Agriculture* [1982]: 71–75). Use the data to find the estimated regression equation and assess the utility of the multiple regression model $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e$

y	x_1	x_2	y	x_1	x_2
61	21.0	57.0	77	24.8	48.0
87	28.3	41.5	93	26.0	56.0
98	27.5	58.0	100	27.1	31.0
104	26.8	36.5	118	29.0	41.0
102	28.3	40.0	74	34.0	25.0
63	30.5	34.0	43	28.3	13.0
27	30.8	37.0	19	31.0	19.0
14	33.6	20.0	23	31.8	17.0
30	31.3	21.0	25	33.5	18.5
67	33.0	24.5	40	34.5	16.0
6	34.3	6.0	21	34.3	26.0
18	33.0	21.0	23	26.5	26.0
42	32.0	28.0	56	27.3	24.5
60	27.8	39.0	59	25.8	29.0
82	25.0	41.0	89	18.5	53.5
77	26.0	51.0	102	19.0	48.0
108	18.0	70.0	97	16.3	79.5

(continued)

CHAPTER ACTIVITY

ACTIVITY 14.1 EXPLORING THE RELATIONSHIP BETWEEN NUMBER OF PREDICTORS AND SAMPLE SIZE

This activity requires the use of a statistical software package capable of fitting multiple regression models.

Background: The given data on y , x_1 , x_2 , x_3 , and x_4 were generated using a software package capable of producing random observations from any specified normal distribution. Because the data were generated at random, there is no reason to believe that y is related to any of the proposed predictor variables x_1 , x_2 , x_3 , and x_4 .

y	x_1	x_2	x_3	x_4
20.5	18.6	22.0	17.1	18.5
20.1	23.9	19.1	21.1	21.3
20.0	20.9	20.7	19.4	20.6
21.7	18.7	18.1	20.9	18.1
20.7	21.1	21.7	23.7	17.0

1. Construct four scatterplots—one of y versus each of x_1 , x_2 , x_3 , and x_4 . Do the scatterplots look the way we

would expect based on the way the data were generated? Explain.

2. Fit each of the following regression models:
 - i. y with x_1
 - ii. y with x_1 and x_2
 - iii. y with x_1 and x_2 and x_3
 - iv. y with x_1 and x_2 and x_3 and x_4
3. Make a table that gives the R^2 , the adjusted R^2 , and s_e values for each of the models fit in Step 2. Describe what happens to each of these three quantities as additional predictor variables are added to the multiple regression model.
4. Given the manner in which these data were generated, what is the implication of what was observed in Step 3? What does this suggest about the relationship between number of predictors and sample size?

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT	TERM OR FORMULA	COMMENT
Additive multiple regression model, $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$	This equation specifies a general probabilistic relationship between y and k independent variables x_1 , x_2, \dots, x_k , where $\alpha, \beta_1, \dots, \beta_k$ are population regression coefficients and $\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is the population regression function (the mean value of y for fixed values of x_1, x_2, \dots, x_k).	Adjusted R^2	A downward adjustment of R^2 that depends on the number of predictors k and the sample size n .
Estimated regression function, $\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$	The estimates a, b_1, \dots, b_k of $\alpha, \beta_1, \dots, \beta_k$ result from applying the principle of least squares.	F distribution	A probability distribution used in the multiple regression model utility test. An F distribution is specified by a numerator df_1 and a denominator df_2 .
Coefficient of multiple determination, $R^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}}$	The proportion of observed y variation that can be explained by the model relationship, where SSResid is defined as it is in simple linear regression but is now based on $n - (k + 1)$ degrees of freedom.	$F = \frac{\text{SSRegr}/k}{\text{SSResid}/(n - (k + 1))}$ or $F = \frac{R^2/k}{(1 - R^2)/(n - (k + 1))}$	The test statistic for testing $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$, which states that there is no useful linear relationship between y and any of the model predictors. The F test is upper-tailed and is based on numerator $df_1 = k$ and denominator $df_2 = n - (k + 1)$.

15 Analysis of Variance



James Woodson/Digital Vision/Getty Images

In Chapter 11 methods were introduced for testing H_0 : $\mu_1 - \mu_2 = 0$ (that is, $\mu_1 = \mu_2$), where μ_1 and μ_2 are the means of two different populations or the mean responses when two different treatments are applied. Many investigations involve a comparison of more than two population or treatment means.

For example, an investigation was carried out to study possible consequences of the high incidence of head injuries among soccer players (“**No Evidence of Impaired Neurocognitive Performance in Collegiate Soccer Players**,” *The American Journal of Sports Medicine* [2002]: 157–162). Three groups of college students (soccer athletes, nonsoccer athletes, and a control group consisting of students who did not participate in intercollegiate sports) were considered in the study. Scores on the Hopkins Verbal Learning Test (which measures immediate memory recall) were used to calculate the summary statistics given in the following table.

	Soccer Athletes	Nonsoccer Athletes	Control
Sample Size	86	95	53
Sample Mean Score	29.90	30.94	29.32
Sample Standard Deviation	3.73	5.14	3.78

Let μ_1 , μ_2 , and μ_3 denote the population mean scores on the Hopkins test for soccer athletes, nonsoccer athletes, and students who do not participate in collegiate athletics, respectively. Do the data support the claim that $\mu_1 = \mu_2 = \mu_3$, or does it appear that at least two of the μ 's are different from one another? This is an example of a **single-factor analysis of variance (ANOVA)** problem, in which the objective is to decide whether the means for two or more populations or treatments are equal.

The first two sections of this chapter discuss various aspects of single-factor ANOVA. In Sections 15.3 and 15.4 (which can be found online), we consider additional ANOVA methods.

LEARNING OBJECTIVES

Students will understand:

- How analysis of variance is used to judge whether there is evidence that two or more population or treatment means are not all equal.

Students will be able to:

- Carry out a hypothesis test to determine whether there is evidence that two or more population or treatment means are not all equal and interpret the results in context.
- Use a multiple comparison procedure to identify differences in population or treatment means.

SECTION 15.1 Single-Factor ANOVA and the F Test

When two or more populations or treatments are being compared, the characteristic that distinguishes the populations or treatments from one another is called the **factor** under investigation. For example, an experiment might be carried out to compare three different methods for teaching reading (three different treatments), in which case the factor of interest would be *teaching method*, a qualitative factor. If the growth of fish raised in water having different salinity levels—0%, 10%, 20%, and 30%—is of interest, then the factor *salinity level* is quantitative.

A **single-factor analysis of variance (ANOVA)** problem involves a comparison of k population or treatment means $\mu_1, \mu_2, \dots, \mu_k$. The objective is to test

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

against

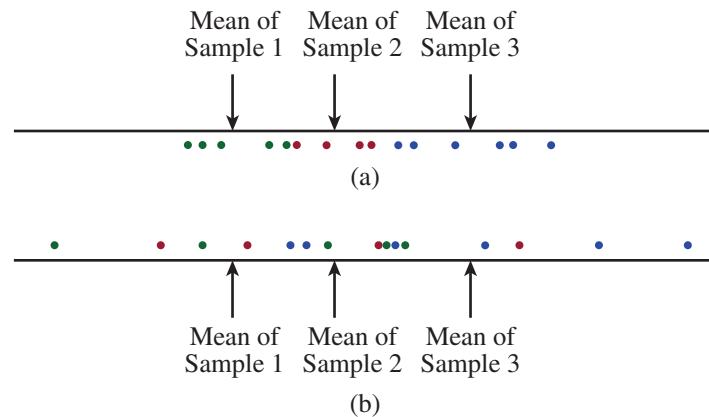
$$H_a: \text{At least two of the } \mu\text{'s are different from each other}$$

When comparing populations, the analysis is based on independently selected random samples, one from each population. When comparing treatment means, the data typically result from an experiment and the analysis assumes random assignment of the experimental units (subjects or objects) to treatments. If, in addition, the experimental units are chosen at random from a population of interest, it is also possible to generalize the results of the analysis to this population. (See Chapter 2 for a more detailed discussion of conclusions that can reasonably be drawn based on data from an experiment.)

Whether the null hypothesis in a single-factor ANOVA should be rejected depends on how substantially the samples from the different populations or treatments differ from one another. Figure 15.1 displays observations that might result when random samples are selected from each of three populations. Each dotplot displays five observations from the first population (using green dots), four observations from the second population (using red dots), and six observations from the third population (using blue dots). For both displays, the three sample means are located by arrows. The means of the two samples from Population 1 are identical, and a similar statement holds for the two samples from Population 2 and those from Population 3.

FIGURE 15.1

Two possible ANOVA data sets when three populations are under investigation:
 green dot = observation from Population 1;
 red dot = observation from Population 2;
 blue dot = observation from Population 3.



After looking at the data in Figure 15.1(a), many people would agree that the claim $\mu_1 = \mu_2 = \mu_3$ appears to be false. Not only are the three sample means different, but the three samples are also clearly separated. In this situation, differences between the three sample means are quite large relative to the variability within each sample.

The situation pictured in Figure 15.1(b) is much less clear-cut. The sample means are as different as they were in for the samples shown in Figure 15.1(a), but now there is considerable overlap among the three samples. The separation between sample means here might plausibly be attributed to substantial variability in the populations (and therefore the samples) rather than to differences between μ_1, μ_2 , and μ_3 .

The phrase *analysis of variance* comes from the idea of analyzing variability in the data to see how much can be attributed to differences in the μ 's and how much is due to variability in the individual populations. In Figure 15.1(a), the within-sample variability is small relative to the between-sample variability, whereas in Figure 15.1(b), a great deal more of the total variability is due to variability within each sample. If differences between the sample means can be explained by within-sample variability, there is no compelling reason to reject H_0 .

Notation and Assumptions

Notation in single-factor ANOVA is a natural extension of the notation used in Chapter 11 for comparing two population or treatment means.

ANOVA Notation

k = number of populations or treatments being compared

Population or treatment	1	2	...	k
Population or treatment mean	μ_1	μ_2	...	μ_k
Population or treatment variance	σ_1^2	σ_2^2	...	σ_k^2
Sample size	n_1	n_2	...	n_k
Sample mean	\bar{x}_1	\bar{x}_2	...	\bar{x}_k
Sample variance	s_1^2	s_2^2	...	s_k^2

$$N = n_1 + n_2 + \cdots + n_k \quad (\text{the total number of observations in the data set})$$

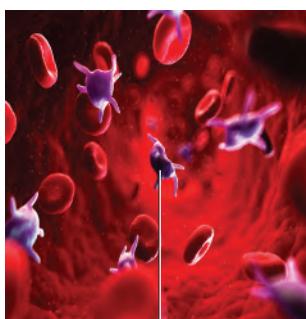
$$\begin{aligned} T &= \text{grand total} = \text{sum of all } N \text{ observations in the data set} \\ &= n_1 \bar{x}_1 + n_2 \bar{x}_2 + \cdots + n_k \bar{x}_k \end{aligned}$$

$$\bar{\bar{x}} = \text{grand mean} = \frac{T}{N}$$

A decision between H_0 and H_a is based on examining the \bar{x} values to see whether observed differences are small enough to be attributable simply to sampling variability or whether an alternative explanation for the differences is more plausible.

Example 15.1 An Indicator of Heart Attack Risk

Understand the context ➤



Science Photo Library/Alamy Stock Photo

Activated platelet

The article “Could Mean Platelet Volume Be a Predictive Marker for Acute Myocardial Infarction?” (*Medical Science Monitor* [2005]: 387–392) describes an experiment in which four groups of patients seeking treatment for chest pain were compared with respect to mean platelet volume (MPV, measured in fL). The four groups considered were based on the clinical diagnosis and were (1) noncardiac chest pain, (2) stable angina pectoris, (3) unstable angina pectoris, and (4) myocardial infarction (heart attack). The purpose of the study was to determine if the mean MPV differed for the patients in the four groups, and in particular if the mean MPV was different for the heart attack group. If this is the case, MPV could be used as an indicator of heart attack risk.

To carry out this study, patients seen for chest pain were divided into groups according to diagnosis. The researchers then selected a random sample of 35 from each of the resulting $k = 4$ groups. The researchers believed that this sampling process would result in samples that were representative of the four populations of interest and that could be regarded as if they were random samples from these four populations. Table 15.1 presents summary values given in the paper.

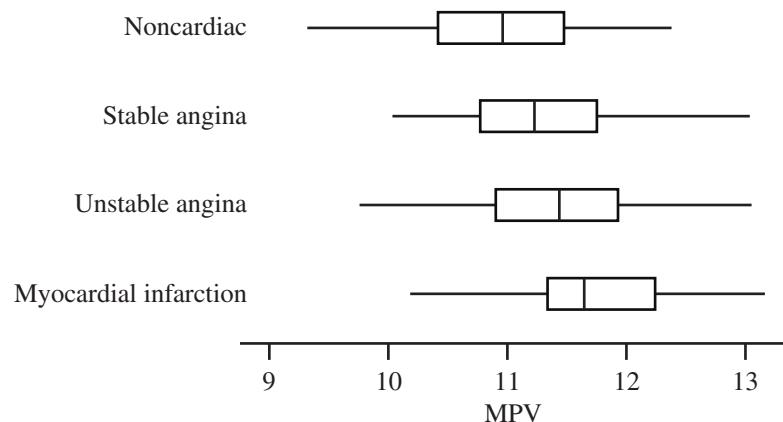
With μ_i denoting the true mean MPV for group i ($i = 1, 2, 3, 4$), consider the null hypothesis $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$. Figure 15.2 shows a comparative boxplot for the four samples (based on data consistent with summary values given in the paper). The mean

TABLE 15.1 Summary Values for MPV Data of Example 15.1

Group Number	Group Description	Sample Size	Sample Mean	Sample Standard Deviation
1	Noncardiac chest pain	35	10.89	0.69
2	Stable angina pectoris	35	11.25	0.74
3	Unstable angina pectoris	35	11.37	0.91
4	Myocardial infarction (heart attack)	35	11.75	1.07

FIGURE 15.2

Boxplots for Example 15.1.



MPV for the heart attack sample is larger than for the other three samples, and the boxplot for the heart attack sample appears to be shifted a bit higher than the boxplots for the other three samples. However, because the four boxplots show substantial overlap, it is not obvious whether H_0 is plausible or should be rejected. In situations like this, a formal test procedure is helpful.

As with the inferential methods of previous chapters, the validity of the ANOVA test for $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ requires some assumptions.

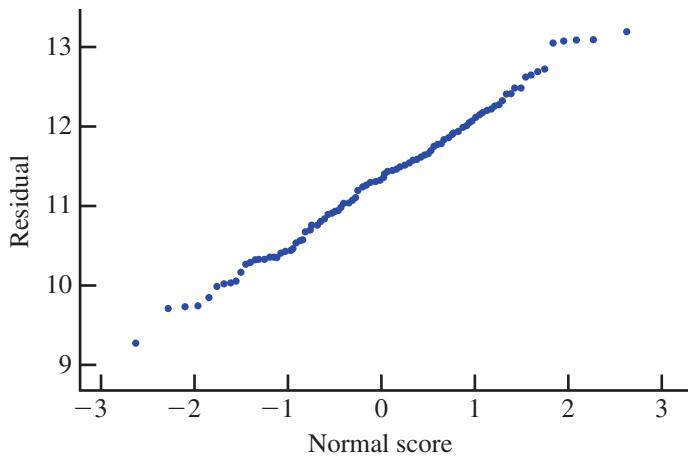
Assumptions for ANOVA

1. Each of the k population or treatment response distributions is normal.
2. $\sigma_1 = \sigma_2 = \dots = \sigma_k$ (The k normal distributions have equal standard deviations.)
3. The observations in the sample from any particular one of the k populations or treatments are independent of one another.
4. When comparing population means, the samples are independent random samples. When comparing treatment means, experimental units are assigned at random to treatments.

In practice, the test based on these assumptions works well as long as the assumptions are not too badly violated. If the sample sizes are reasonably large, normal probability plots or boxplots of the data in each sample are helpful for checking the assumption of normality. Often, however, sample sizes are so small that a separate normal probability plot or boxplot for each sample is of little value in checking normality. In this case, a single combined plot can be constructed by first subtracting \bar{x}_1 from each observation in the first sample, \bar{x}_2 from each value in the second sample, and so on and then constructing a normal probability or boxplot of all N deviations from their respective means. The plot should be reasonably straight. Figure 15.3 shows such a normal probability plot for the data of Example 15.1.

FIGURE 15.3

A normal probability plot using the combined data of Example 15.1.



There is a formal procedure for testing the equality of population standard deviations. Unfortunately, it is quite sensitive to even a small departure from the normality assumption, so we do not recommend its use. Instead, we suggest that the ANOVA *F* test (to be described later in this section) can safely be used if the largest of the sample standard deviations is at most twice the smallest one. The largest standard deviation in Example 15.1 is $s_4 = 1.07$, which is only about 1.5 times the smallest standard deviation ($s_1 = 0.69$).

The analysis of variance test procedure is based on the following measures of variability in the data.

A measure of differences among the sample means is the **treatment sum of squares**, denoted by **SSTr** and given by

$$\text{SSTr} = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_k(\bar{x}_k - \bar{\bar{x}})^2$$

A measure of variability within the k samples, called **error sum of squares** and denoted by **SSE**, is

$$\text{SSE} = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2$$

Each sum of squares has an associated df:

$$\text{treatment df}_1 = k - 1 \quad \text{error df}_2 = N - k$$

A **mean square** is a sum of squares divided by its df. In particular,

$$\text{mean square for treatments} = \text{MStr} = \frac{\text{SSTr}}{k - 1}$$

$$\text{mean square for error} = \text{MSE} = \frac{\text{SSE}}{N - k}$$

The number of error degrees of freedom comes from adding the number of degrees of freedom associated with each of the sample variances:

$$\begin{aligned} (n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1) &= n_1 + n_2 + \cdots + n_k - 1 - 1 - \cdots - 1 \\ &= N - k \end{aligned}$$

Example 15.2 Heart Attack Calculations

For the mean platelet volume (MPV) data of Example 15.1, the grand mean \bar{x} was calculated to be 11.315. Notice that because the sample sizes are all equal, the grand mean is just the average of the four sample means (this will not usually be the case when the sample sizes are unequal).

With $\bar{x}_1 = 10.89$, $\bar{x}_2 = 11.25$, $\bar{x}_3 = 11.37$, $\bar{x}_4 = 11.75$, and $n_1 = n_2 = n_3 = n_4 = 35$,

$$\begin{aligned} \text{SSTr} &= n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_k(\bar{x}_k - \bar{\bar{x}})^2 \\ &= 35(10.89 - 11.315)^2 + 35(11.25 - 11.315)^2 + 35(11.37 - 11.315)^2 \\ &\quad + 35(11.75 - 11.315)^2 \\ &= 6.322 + 0.148 + 0.106 + 6.623 \\ &= 13.199 \end{aligned}$$

Because $s_1 = 0.69$, $s_2 = 0.74$, $s_3 = 0.91$, and $s_4 = 1.07$

$$\begin{aligned} \text{SSE} &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \\ &= (35 - 1)(0.69)^2 + (35 - 1)(0.74)^2 + (35 - 1)(0.91)^2 + (35 - 1)(1.07)^2 \\ &= 101.888 \end{aligned}$$

The numbers of degrees of freedom are

treatment $\text{df}_1 = k - 1 = 3$ error $\text{df}_2 = N - k = 35 + 35 + 35 + 35 - 4 = 136$
from which

$$\text{MSTr} = \frac{\text{SSTr}}{k - 1} = \frac{13.199}{3} = 4.400$$

$$\text{MSE} = \frac{\text{SSE}}{N - k} = \frac{101.888}{136} = 0.749$$

Both MSTr and MSE are statistics whose values can be calculated once sample data are available. Each of these statistics varies in value from data set to data set. Both statistics MSTr and MSE have sampling distributions, and these sampling distributions have mean values. The following box describes the key relationship between the mean values of MSTr and MSE.

When H_0 is true ($\mu_1 = \mu_2 = \cdots = \mu_k$),

$$\mu_{\text{MSTr}} = \mu_{\text{MSE}}$$

However, when H_0 is false,

$$\mu_{\text{MSTr}} > \mu_{\text{MSE}}$$

and the greater the differences among the μ 's, the larger μ_{MSTr} will be relative to μ_{MSE} .

According to this result, when H_0 is true, we expect the values of the two mean squares to be close. However, we expect MSTr to be substantially greater than MSE when some μ 's differ greatly from others. This means that a calculated MSTr that is much greater than MSE is inconsistent with H_0 .

In Example 15.2, $\text{MSTr} = 4.400$ and $\text{MSE} = 0.749$, so MSTr is about 6 times as large as MSE. Can this difference be attributed solely to sampling variability, or is the ratio MSTr/MSE large enough to suggest that H_0 should be rejected? Before a formal

test procedure is introduced, it is necessary to revisit F distributions, which were first introduced in Chapter 14.

F distributions arise in connection with ratios of mean squares. A particular F distribution is obtained by specifying both numerator degrees of freedom (df_1) and denominator degrees of freedom (df_2).

Figure 15.4 shows an F curve for a particular choice of df_1 and df_2 . All F tests in this text are upper-tailed, so P -values are areas under the F curve to the right of the calculated values of F .

Tabulation of these upper-tail areas is cumbersome, because there are two degrees of freedom rather than just one (as in the case of t distributions). For selected (df_1, df_2) pairs, the F table (Appendix Table 6) gives only the four numbers that capture tail areas 0.10, 0.05, 0.01, and 0.001, respectively. For example, here are the four numbers for $df_1 = 4$, $df_2 = 10$ along with the statements that can be made about the P -value:

Tail area	0.10	0.05	0.01	0.001	
Value	2.61	3.48	5.99	11.28	
	↑ a	↑ b	↑ c	↑ d	↑ e

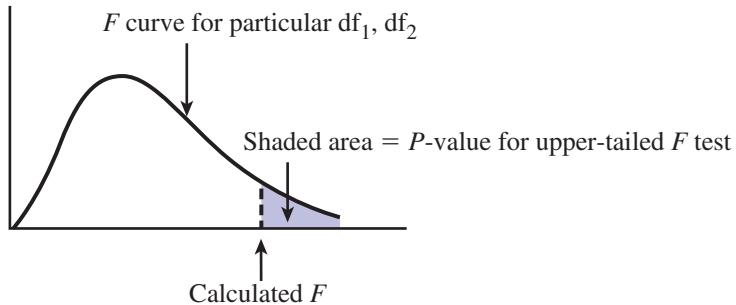
- a. $F < 2.61 \rightarrow$ tail area = P -value > 0.10
- b. $2.61 < F < 3.48 \rightarrow 0.05 < P$ -value < 0.10
- c. $3.48 < F < 5.99 \rightarrow 0.01 < P$ -value < 0.05
- d. $5.99 < F < 11.28 \rightarrow 0.001 < P$ -value < 0.01
- e. $F > 11.28 \rightarrow P$ -value < 0.001

For example, if $F = 7.12$, then $0.001 < P$ -value < 0.01. If a test with $\alpha < 0.05$ is used, H_0 would be rejected, because P -value $\leq \alpha$.

Many statistical software packages can provide exact P -values for F tests.

FIGURE 15.4

An F curve and P -value for an upper-tailed test.



The Single-Factor ANOVA F Test

Null hypothesis: $H_0: \mu_1 = \mu_2 = \dots = \mu_k$

Test statistic: $F = \frac{MSTr}{MSE}$

When H_0 is true and the ANOVA assumptions are reasonable, F has an F distribution with $df_1 = k - 1$ and $df_2 = N - k$.

The P -value is the area captured in the upper tail of the corresponding F curve. Appendix Table 6, a statistical software package, or a graphing calculator can be used to determine P -values for F tests.

Example 15.3 Heart Attacks Revisited

The two mean squares for the MPV data given in Example 15.1 were calculated in Example 15.2 to be

$$\text{MSTr} = 4.400 \quad \text{MSE} = 0.749$$

The value of the F statistic is then

$$F = \frac{\text{MSTr}}{\text{MSE}} = \frac{4.400}{0.749} = 5.87$$

with $\text{df}_1 = k - 1 = 3$ and $\text{df}_2 = N - k = 140 - 4 = 136$.

Using $\text{df}_1 = 3$ and $\text{df}_2 = 120$ (the closest value to 136 that appears in the table), Appendix Table 6 shows that 5.78 captures tail area 0.001. Since $5.87 > 5.78$, it follows that $P\text{-value} = (\text{area to the right of } 5.87) < 0.001$.

Interpret the results ►

The P -value is smaller than any reasonable α , so the null hypothesis is rejected. There is convincing evidence that the mean MPV is not the same for all four patient populations. Techniques for determining which means differ are introduced in Section 15.2.

Example 15.4 Hormones and Body Fat

Understand the context ►

- The article “Growth Hormone and Sex Steroid Administration in Healthy Aged Women and Men” (*Journal of the American Medical Association* [2002]: 2282–2292) described an experiment to investigate the effect of four treatments on various body characteristics. In this double blind experiment, each of 57 female subjects age 65 or older was assigned at random to one of the following four treatments:

- placebo “growth hormone” and placebo “steroid” (denoted by P + P);
- placebo “growth hormone” and the steroid estradiol (denoted by P + S);
- growth hormone and placebo “steroid” (denoted by G + P); and
- growth hormone and the steroid estradiol (denoted by G + S).

The following table gives data on change in body fat mass over the 26-week period following the treatments that are consistent with summary quantities given in the article.

Treatment	Change in Body Fat Mass (kg)			
	P + P	P + S	G + P	G + S
	0.1	-0.1	-1.6	-3.1
	0.6	0.2	-0.4	-3.2
	2.2	0.0	0.4	-2.0
	0.7	-0.4	-2.0	-2.0
	-2.0	-0.9	-3.4	-3.3
	0.7	-1.1	-2.8	-0.5
	0.0	1.2	-2.2	-4.5
	-2.6	0.1	-1.8	-0.7
	-1.4	0.7	-3.3	-1.8
	1.5	-2.0	-2.1	-2.3
	2.8	-0.9	-3.6	-1.3
	0.3	-3.0	-0.4	-1.0
	-1.0	1.0	-3.1	-5.6
	-1.0	1.2		-2.9
				-1.6
				-0.2
n	14	14	13	16
\bar{x}	0.064	-0.286	-2.023	-2.250
s	1.545	1.218	1.264	1.468
s^2	2.387	1.484	1.598	2.155

● Data set available online

Also, for this data set $N = 57$, the grand total = -65.4 , and $\bar{x} = \frac{-65.4}{57} = -1.15$.

Formulate a plan ➤

We can carry out an F test to see whether actual mean change in body fat mass differs for the four treatments.

1. Let μ_1, μ_2, μ_3 , and μ_4 denote the actual mean change in body fat for treatments P + P, P + S, G + P, and G + S, respectively.

2. $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

3. H_a : At least two among μ_1, μ_2, μ_3 , and μ_4 are different.

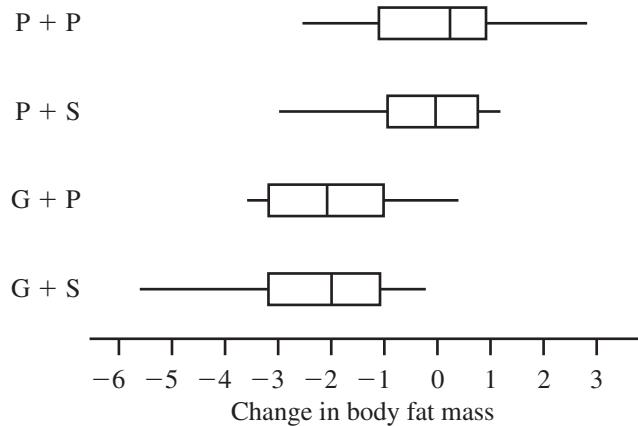
4. Significance level: $\alpha = 0.01$

5. Test statistic: $F = \frac{\text{MSTr}}{\text{MSE}}$

6. Assumptions: Figure 15.5 shows boxplots of the data from each of the four samples. The boxplots are approximately symmetric, and there are no outliers. The largest standard deviation ($s_1 = 1.545$) is not more than twice as large as the smallest ($s_2 = 1.264$). The subjects were randomly assigned to treatments. The assumptions of ANOVA are reasonable.

FIGURE 15.5

Boxplots for the data of Example 15.4.



Do the work ➤

7. Calculation:

$$\begin{aligned}\text{SSTr} &= n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_k(\bar{x}_k - \bar{\bar{x}})^2 \\ &= 14(0.064 - (-1.15))^2 + 14(-0.286 - (-1.15))^2 \\ &\quad + 13(-2.023 - (-1.15))^2 + 16(-2.250 - (-1.15))^2 \\ &= 60.37\end{aligned}$$

$$\text{treatment df}_1 = k - 1 = 3$$

$$\begin{aligned}\text{SSE} &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \\ &= 13(2.387) + 13(1.484) + 12(1.598) + 15(2.155) \\ &= 101.81\end{aligned}$$

$$\text{error df}_2 = N - k = 57 - 4 = 53$$

Then,

$$F = \frac{\text{MSTr}}{\text{MSE}} = \frac{\text{SSTr/treatment df}_1}{\text{SSE/error df}_2} = \frac{60.37/3}{101.81/53} = \frac{20.12}{1.92} = 10.48$$

8. P -value: Appendix Table 6 shows that for $\text{df}_1 = 3$ and $\text{df}_2 = 53$ (the closest tabled df to $\text{df} = 53$), the value 6.17 captures upper-tail area 0.001. Because $F = 10.48 > 6.17$, it follows $P\text{-value} < 0.001$.

Interpret the results ➤

9. Conclusion: Since $P\text{-value} \leq \alpha$, we reject H_0 . There is convincing evidence that mean change in body fat mass is not the same for all four treatments.

Summarizing an ANOVA

ANOVA calculations are often summarized in a table called an ANOVA table. To understand such a table, one more sum of squares must be defined.

Total sum of squares, denoted by $SSTo$, is given by

$$SSTo = \sum_{\text{all } N \text{ obs.}} (x - \bar{x})^2$$

with associated $df = N - 1$.

The relationship between the three sums of squares $SSTo$, $SSTr$, and SSE is

$$SSTo = SSTr + SSE$$

which is called the *fundamental identity for single-factor ANOVA*.

The quantity $SSTo$, the sum of squared deviations about the grand mean, is a measure of total variability in the data set consisting of all k samples. The quantity SSE results from measuring variability separately within each sample and then combined. Within-sample variability is present regardless of whether or not H_0 is true. The magnitude of $SSTr$, on the other hand, does depend on whether the null hypothesis is true or false. The more the μ 's differ from one another, the larger $SSTr$ will tend to be. This means that $SSTr$ represents variability that can (at least to some extent) be explained by differences among the means. An informal paraphrase of the fundamental identity for single-factor ANOVA is

$$\text{total variability} = \text{explained variability} + \text{unexplained variability}$$

Once any two of the sums of squares have been calculated, the remaining one is easily obtained from the fundamental identity. Often $SSTo$ and $SSTr$ are calculated first (using computational formulas given in the online appendix to this chapter), and then SSE is obtained by subtraction: $SSE = SSTo - SSTr$. All the degrees of freedom, sums of squares, and mean squares are entered in an ANOVA table, as displayed in Table 15.2. The P -value usually appears to the right of F when the analysis is done by a statistical software package.

TABLE 15.2 General Format for a Single-Factor ANOVA Table

Source of Variation	df	Sum of Squares	Mean Square	F
Treatments	$k - 1$	$SSTr$	$MStr = \frac{SSTr}{k - 1}$	$F = \frac{MStr}{MSE}$
Error	$N - k$	SSE	$MSE = \frac{SSE}{N - k}$	
Total	$N - 1$	$SSTo$		

An ANOVA table from Minitab for the change in body fat mass data of Example 15.4 is shown in Table 15.3. The reported P -value is 0.000, and this is consistent with our previous conclusion that $P\text{-value} < 0.001$.

TABLE 15.3 An ANOVA Table from Minitab for the Data of Example 15.4

Source	DF	SS	MS	F	P
Factor	3	60.37	20.12	10.48	0.000
Error	53	101.81	1.92		
Total	56	162.18			

EXERCISES 15.1 - 15.13

● Data set available online

- 15.1** Give as much information as you can about the *P*-value for an upper-tailed *F* test in each of the following situations:

- a. $df_1 = 4, df_2 = 15, F = 5.37$
- b. $df_1 = 4, df_2 = 15, F = 1.90$
- c. $df_1 = 4, df_2 = 15, F = 4.89$
- d. $df_1 = 3, df_2 = 20, F = 14.48$
- e. $df_1 = 3, df_2 = 20, F = 2.69$
- f. $df_1 = 4, df_2 = 50, F = 3.24$

- 15.2** Give as much information as you can about the *P*-value of the single-factor ANOVA *F* test in each of the following situations:

- a. $k = 5, n_1 = n_2 = n_3 = n_4 = n_5 = 4, F = 5.37$
- b. $k = 5, n_1 = n_2 = n_3 = 5, n_4 = n_5 = 4, F = 2.83$
- c. $k = 3, n_1 = 4, n_2 = 5, n_3 = 6, F = 5.02$
- d. $k = 3, n_1 = n_2 = 4, n_3 = 6, F = 15.90$
- e. $k = 4, n_1 = n_2 = 15, n_3 = 12, n_4 = 10, F = 1.75$

- 15.3** Employees of a state university system can choose from among four different health plans. Each plan differs somewhat from the others in terms of hospitalization coverage. Four random samples of recently hospitalized individuals were selected, each sample consisting of people covered by a different health plan. The length of the hospital stay (number of days) was determined for each individual selected.

- a. What hypotheses would you test to decide whether the mean lengths of stay are not the same for all four health plans? (Note: Carefully define the population characteristics of interest.)
- b. If each sample consisted of eight individuals and the value of the ANOVA *F* statistic was $F = 4.37$, what conclusion would be appropriate for a test with $\alpha = 0.01$? (Hint: See Example 15.4.)
- c. Answer the question posed in Part (b) if the *F* value given there resulted from sample sizes $n_1 = 9, n_2 = 8, n_3 = 7$, and $n_4 = 8$.

- 15.4** The accompanying summary statistics for a measure of social marginality for samples of youths, young adults, adults, and seniors appeared in the paper “Perceived Causes of Loneliness in Adulthood”

(*Journal of Social Behavior and Personality* [2000]: 67–84). The social marginality score measured actual and perceived social rejection, with higher scores indicating greater social rejection.

For purposes of this exercise, assume that it is reasonable to regard the four samples as representative of the U.S. population in the corresponding age groups and that the distributions of social marginality scores for these four groups are approximately normal with the same standard deviation.

Is there evidence that the mean social marginality scores are not the same for all four age groups? Test the relevant hypotheses using $\alpha = 0.01$. (Hint: See Example 15.4.)

Age Group	Youths	Young Adults		
		Adults	Seniors	Sample Size
\bar{x}	2.00	3.40	3.07	2.84
s	1.56	1.68	1.66	1.89

- 15.5** ● The authors of the paper “Age and Violent Content Labels Make Video Games Forbidden Fruits for Youth” (*Pediatrics* [2009]: 870–876) carried out an experiment to determine if restrictive labels on video games actually increased the attractiveness of the game for young game players. Participants read a description of a new video game and were asked how much they wanted to play the game. The description also included an age rating. Some participants read the description with an age restrictive label of 7+, indicating that the game was not appropriate for children under the age of 7. Others read the same description, but with an age restrictive label of 12+, 16+, or 18+.

The data on the following page for 12- to 13-year-old boys are consistent with summary statistics given in the paper. (The sample sizes in the actual experiment were larger.) For purposes of this exercise, you can assume that the boys were assigned at random to one of the four age label treatments (7+, 12+, 16+, and 18+). Data shown are the boys’ ratings of how much they wanted to play the game on a scale of 1 to 10.

Do the data provide convincing evidence that the means of the ratings associated with the game descriptions by 12- to 13-year-old boys are not the same for all four restrictive rating labels? Test the appropriate hypotheses using a significance level of 0.05. (Hint: See Example 15.4.)

7+ label	12+ label	16+ label	18+ label
6	8	7	10
6	7	9	9
6	8	8	6
5	5	6	8
4	7	7	7
8	9	4	6
6	5	8	8
1	8	9	9
2	4	6	10
4	7	7	8

- 15.6** The paper referenced in the previous exercise also gave data for 12- to 13-year-old girls. Data consistent with summary values in the paper are shown below. Do the data provide convincing evidence that the mean rating associated with the game description for 12- to 13-year-old girls is not the same for all four age restrictive rating labels? Test the appropriate hypotheses using $\alpha = 0.05$.

7+ label	12+ label	16+ label	18+ label
4	4	6	8
7	5	4	6
6	4	8	6
5	6	6	5
3	3	10	7
6	5	8	4
4	3	6	10
5	8	6	6
10	5	8	8
5	9	5	7

- 15.7** Do people feel hungrier after sampling a healthy food? The authors of the paper “[When Healthy Food Makes You Hungry](#)” (*Journal of Consumer Research* [2010]: S34–S44) carried out a study to answer this question. They randomly assigned volunteers into one of three groups. The people in the first group were asked to taste a snack that was described as a new health bar containing high levels of protein, vitamins, and fiber. The people in the second group were asked to taste the same snack, but were told it was a tasty chocolate bar with a raspberry center. After tasting the snack, participants were asked to rate their hunger level on a scale from 1 (not at all hungry) to 7 (very hungry). The people in the third group were asked to rate their hunger but were not given a snack.

The data in the accompanying table are consistent with summary quantities given in the paper (although the sample sizes in the actual study were larger).

- Do these data provide evidence that the mean hunger rating is not the same for all three treatments (“healthy” snack, “tasty” snack, no snack)? Test the relevant hypotheses using a significance level of 0.05.
- Is it reasonable to conclude that the mean hunger rating is greater for people who do not get a snack? Explain.

Treatment Group	Hunger Rating								Sample Mean	Standard Deviation	
Healthy	4	7	7	4	5	3	4	7	6	5.2	1.56
Tasty	4	1	3	2	6	2	5	3	4	3.3	1.58
No Snack	3	4	5	6	5	4	2	4	4	4.1	1.17
Overall mean = 4.2											

- 15.8** The authors of the paper “[Reading Subtitles and Taking Notes While Learning Scientific Materials in a Multimedia Environment](#)” (*Educational Technology and Society* [2016]: 47–58) were interested in determining if including subtitles and providing opportunities to take electronic notes while listening to online materials would enhance learning for students whose first language is not English. Students were randomly assigned to one of four groups. In the first group, subtitles were included in the online materials that the students were asked to study, but the ability to take electronic notes was not provided. For the second group, no subtitles were provided, but students were able to take electronic notes. For the third group, both subtitles and the ability to take electronic notes were available. Students in the fourth group did not have access to either subtitles or the ability to take electronic notes. After studying the online materials, all students took a 14-question test on the material studied.

The Minitab output based on data consistent with summary quantities in the paper is shown below. Is there evidence to conclude that the mean test score is not the same for the four different treatments? Use the given computer output to test the appropriate hypotheses with a significance level of 0.05.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	3	37.76	12.588	2.39	0.077
Error	63	332.36	5.276		
Total	66	370.13			

- 15.9** ● The paper “**Women’s and Men’s Eating Behavior Following Exposure to Ideal-Body Images and Text**” (*Communication Research* [2006]: 507–529) describes an experiment in which 74 men were assigned at random to one of four treatments:

1. Viewed slides of fit, muscular men
2. Viewed slides of fit, muscular men accompanied by diet and fitness-related text
3. Viewed slides of fit, muscular men accompanied by text not related to diet and fitness
4. Did not view any slides

The participants then went to a room to complete a questionnaire. In this room, bowls of pretzels were set out on the tables. A research assistant noted how many pretzels were consumed by each participant while completing the questionnaire. Data consistent with summary quantities given in the paper are given in the accompanying table.

Do these data provide convincing evidence that the means for the numbers of pretzels consumed are not the same for all four treatments? Test the relevant hypotheses using a significance level of 0.05.

Treatment 1	Treatment 2	Treatment 3	Treatment 4
8	6	1	5
7	8	5	2
4	0	2	5
13	4	0	7
2	9	3	5
1	8	0	2
5	6	3	0
8	2	4	0
11	7	4	3
5	8	5	4
1	8	5	2
0	5	7	4
6	14	8	1
4	9	4	1
10	0	0	
7	6	6	
0	3	3	
12	12		
	5		
	6		
	10		
	8		
	6		
	2		
	10		

- 15.10** Can use of an online plagiarism-detection system reduce plagiarism in student research papers? The paper “**Plagiarism and Technology: A Tool for Coping with Plagiarism**” (*Journal of Education for Business* [2005]: 149–152) describes a study in

which randomly selected research papers submitted by students during five semesters were analyzed for plagiarism. For each paper, the percentage of plagiarized words in the paper was determined by an online analysis. In each of the five semesters, students were told during the first two class meetings that they would have to submit an electronic version of their research papers and that the papers would be reviewed for plagiarism. Suppose that the number of papers sampled in each of the five semesters and the means and standard deviations for percentage of plagiarized words are as given in the accompanying table.

For purposes of this exercise, assume that the conditions necessary for the ANOVA *F* test are reasonable. Do these data provide evidence to support the claim that mean percentage of plagiarized words is not the same for all five semesters? Test the appropriate hypotheses using $\alpha = 0.05$.

Semester	<i>n</i>	Mean	Standard deviation
1	39	6.31	3.75
2	42	3.31	3.06
3	32	1.79	3.25
4	32	1.83	3.13
5	34	1.50	2.37

- 15.11** It is common for baseball pitchers to use stretching to prepare for a game. But does this make a difference? The authors of the paper “**The Acute Effects of Upper Extremity Stretching on Throwing Velocity in Baseball Throwers**” (*Journal of Sports Medicine* [2013]: 1–7) carried out an experiment to compare two different types of stretching and a control treatment consisting of no stretching. Participants were adult males with varying levels of baseball throwing experience and who were not professional or collegiate baseball players. Participants in the two stretching treatments went through a warm-up that included 8 minutes of stretching. Each participant then threw 10 pitches, and the average speed (km/hour) was calculated.

- Explain why it is important that the participants be assigned at random to the three different treatment groups (Stretching Method 1, Stretching Method 2, and No Stretching).
- The following computer output and summary values are based on simulated data that are consistent with information and conclusions given in the paper. Use the given output to determine if there is evidence to support the claim that the means for average speed are not the same for all three treatments. Use a significance level of 0.05 for your test.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	2	1097	548.7	2.44	0.094
Error	78	17570	225.3		
Total	80	18667			

Means

Factor	N	Mean	StDev	95% CI
Stretching	27	86.26	15.95	(80.51, 92.01)
Method 1				
Stretching	27	90.30	17.06	(84.55, 96.05)
Method 2				
No Stretching	27	95.26	11.42	(89.51, 101.01)

- c. Previous research on the effect of stretching on performance in other sports, such as running, has concluded that stretching can improve performance. Why do you think that the authors of this paper were surprised by the results of this study?

- 15.12** In the introduction to this chapter, we considered a study comparing three groups of college students (soccer athletes, nonsoccer athletes, and a control group consisting of students who did not participate in intercollegiate sports). The following table gives information on scores from the Hopkins Verbal Learning Test (which measures immediate memory recall).

	Soccer Athletes	Nonsoccer Athletes	Control
Sample size	86	95	53
Sample mean score	29.90	30.94	29.32
Sample standard deviation	3.73	5.14	3.78

In addition, $\bar{x} = 30.19$. Suppose that it is reasonable to regard these three samples as random samples from the three student populations of interest.

Is there sufficient evidence to conclude that the means for Hopkins score are not the same for the three student populations? Use $\alpha = 0.05$.

- 15.13** In an experiment to investigate the performance of four different brands of spark plugs intended for use on a 125-cc motorcycle, five plugs of each brand were tested, and the number of miles (at a constant speed) until failure was observed. A partially completed ANOVA table is given. Fill in the missing entries, and test the relevant hypotheses using a 0.05 level of significance. (Hint: See Table 15.2.)

Source of Variation	df	Sum of Squares	Mean Square	F
Treatments				
Error		235,419.04		
Total		310,500.76		

SECTION 15.2 Multiple Comparisons

When $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ is rejected by the F test, we believe that there are differences among the k population or treatment means. A natural question to ask then is, Which means differ? For example, with $k = 4$, it might be the case that $\mu_1 = \mu_2 = \mu_4$, with μ_3 different from the other three means. Another possibility is that $\mu_1 = \mu_4$ and $\mu_2 = \mu_3$. Still another possibility is that all four means are different from one another.

A **multiple comparisons procedure** is a method for identifying differences among the μ 's once the hypothesis that all the means are equal has been rejected. We present one such method, the **Tukey-Kramer** (T-K) multiple comparisons procedure.

The T-K procedure is based on calculating confidence intervals for the difference between each possible pair of μ 's. For example, for $k = 3$, there are three differences to consider:

$$\mu_1 - \mu_2 \quad \mu_1 - \mu_3 \quad \mu_2 - \mu_3$$

(The difference $\mu_2 - \mu_1$ is not considered, because the interval for $\mu_1 - \mu_2$ provides the same information. Similarly, intervals for $\mu_3 - \mu_1$ and $\mu_3 - \mu_2$ are not necessary.)

After all confidence intervals have been calculated, each is examined to determine whether the interval includes 0. If a particular interval does not include 0, the two means are declared “significantly different” from one another. An interval that does include 0 supports the conclusion that there is no significant difference between the means involved.

Suppose, for example, that $k = 3$ and that the three confidence intervals are

Difference	T-K Confidence Interval
$\mu_1 - \mu_2$	(−0.9, 3.5)
$\mu_1 - \mu_3$	(2.6, 7.0)
$\mu_2 - \mu_3$	(1.2, 5.7)

Because the interval for $\mu_1 - \mu_2$ includes 0, we say that μ_1 and μ_2 do not differ significantly. The other two intervals do not include 0, so we conclude that $\mu_1 \neq \mu_3$ and $\mu_2 \neq \mu_3$.

The T-K intervals are based on critical values for a probability distribution called the *Studentized range distribution*. These critical values appear in Appendix Table 7. To find a critical value, enter the table at the column corresponding to the number of populations or treatments being compared, move down to the row corresponding to the number of error degrees of freedom, and select either the value for a 95% confidence level or the one for a 99% level.

The Tukey-Kramer (T-K) Multiple Comparison Procedure

When there are k populations or treatments being compared, $\frac{k(k-1)}{2}$ confidence intervals are calculated.

Denoting the relevant Studentized range critical value (from Appendix Table 7) by q , the intervals are as follows:

$$\text{For } \mu_i - \mu_j: (\bar{x}_i - \bar{x}_j) \pm q \sqrt{\frac{\text{MSE}}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Two means are judged to be significantly different if the corresponding interval does not include zero.

If the sample sizes are all equal, we can use n to denote the common value of n_1, \dots, n_k . In this case, the \pm term for all of the intervals is the same quantity

$$q \sqrt{\frac{\text{MSE}}{n}}$$

Example 15.5 Hormones and Body Fat Revisited

Understand the context ➤

- Example 15.4 introduced the accompanying data on change in body fat mass resulting from a double-blind experiment designed to compare the following four treatments:

- placebo “growth hormone” and placebo “steroid” (denoted by P + P);
- placebo “growth hormone” and the steroid estradiol (denoted by P + S);
- growth hormone and placebo “steroid” (denoted by G + P); and
- growth hormone and the steroid estradiol (denoted by G + S).

From Example 15.4, $\text{MSTr} = 20.12$, $\text{MSE} = 1.92$, and $F = 10.48$ with an associated P -value < 0.001 . We concluded that the mean change in body fat mass is not the same for all four treatments. The data are presented once more in the following table.

● Data set available online

Treatment	Change in Body Fat Mass (kg)			
	P + P	P + S	G + P	G + S
0.1	-0.1	-1.6	-3.1	
0.6	0.2	-0.4	-3.2	
2.2	0.0	0.4	-2.0	
0.7	-0.4	-2.0	-2.0	
-2.0	-0.9	-3.4	-3.3	
0.7	-1.1	-2.8	-0.5	
0.0	1.2	-2.2	-4.5	
-2.6	0.1	-1.8	-0.7	
-1.4	0.7	-3.3	-1.8	
1.5	-2.0	-2.1	-2.3	
2.8	-0.9	-3.6	-1.3	
0.3	-3.0	-0.4	-1.0	
-1.0	1.0	-3.1	-5.6	
-1.0	1.2		-2.9	
			-1.6	
			-0.2	
<i>n</i>	14	14	13	16
\bar{x}	0.064	-0.286	-2.023	-2.250
<i>s</i>	1.545	1.218	1.264	1.468
<i>s</i> ²	2.387	1.484	1.598	2.155

Do the work ➤ Appendix Table 7 gives the 95% Studentized range critical value $q = 3.74$ (using $k = 4$ and error df = 60, the closest tabled value to df = $N - k = 53$). The first two T-K intervals are

$$\mu_1 - \mu_2: (0.064 - (-0.286)) \pm 3.74 \sqrt{\left(\frac{1.92}{2}\right)\left(\frac{1}{14} + \frac{1}{14}\right)}$$

$$= 0.35 \pm 1.39$$

$\xleftarrow{\text{Includes } 0}$

$$= (-1.04, 1.74)$$

$$\mu_1 - \mu_3: (0.064 - (-2.023)) \pm 3.74 \sqrt{\left(\frac{1.92}{2}\right)\left(\frac{1}{14} + \frac{1}{13}\right)}$$

$$= 2.09 \pm 1.41$$

$\xleftarrow{\text{Does not include } 0}$

$$= (0.68, 3.50)$$

The remaining intervals are

$\mu_1 - \mu_4$	$(0.97, 3.66)$	\leftarrow Does not include 0
$\mu_2 - \mu_3$	$(0.32, 3.15)$	\leftarrow Does not include 0
$\mu_2 - \mu_4$	$(0.62, 3.31)$	\leftarrow Does not include 0
$\mu_3 - \mu_4$	$(-1.145, 1.60)$	\leftarrow Includes 0

Interpret the results ➤ We would conclude that μ_1 is not significantly different from μ_2 and that μ_3 is not significantly different from μ_4 . We would also conclude that μ_1 and μ_2 are significantly different from both μ_3 and μ_4 . Notice that Treatments 1 and 2 were treatments that administered a placebo in place of the growth hormone and Treatments 3 and 4 were treatments that included the growth hormone. This analysis was the basis of the researchers' conclusion that growth hormone, with or without steroids, decreased mean body fat mass.

Minitab can be used to construct T-K intervals if raw data are available. Typical output (based on Example 15.5) is shown in Figure 15.6. From the output, we see that the confidence interval for $\mu_1 (P + P) - \mu_2 (P + S)$ is $(-1.039, 1.739)$, that for $\mu_2 (P + S) - \mu_4 (G + S)$ is $(0.619, 3.309)$, and so on.

FIGURE 15.6

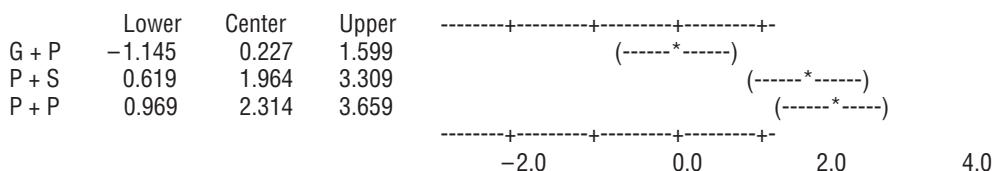
The T-K intervals for Example 15.5
(from Minitab).

Tukey 95% Simultaneous Confidence Intervals

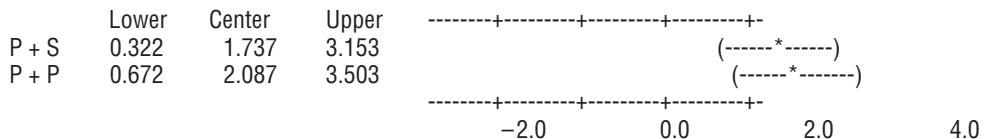
All Pairwise Comparisons

Individual confidence level = 98.95%

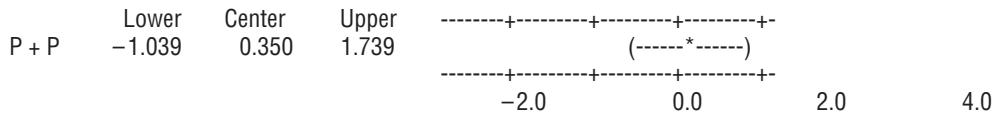
G + S subtracted from:



G + P subtracted from:



P + S subtracted from:



Why calculate the T-K intervals rather than use the t confidence interval for a difference between two means introduced in Chapter 11? The answer is that the T-K intervals control the **simultaneous confidence level** at approximately 95% (or 99%). That is, if the procedure is used repeatedly on many different data sets, in the long run only about 5% (or 1%) of the time would *at least one* of the intervals not include the value of what the interval is estimating.

Consider using separate 95% t intervals, each one having a 5% error rate. When this is done, the chance that at least one interval would make an incorrect statement about a difference in μ 's increases dramatically with the number of intervals calculated. The Minitab output in Figure 15.6 shows that to achieve a simultaneous confidence level of about 95% (experimentwise or “family” error rate of 5%) when $k = 4$ and error df = 76, the confidence level for individual intervals must be 98.95% (individual error rate 1.05%).

An effective display for summarizing the results of any multiple comparisons procedure involves listing the \bar{x} 's and underscoring pairs judged to be not significantly different. The process for constructing such a display is described in the box at the top of page 776.

To illustrate this summary procedure, suppose that four samples with $\bar{x}_1 = 19$, $\bar{x}_2 = 27$, $\bar{x}_3 = 24$, and $\bar{x}_4 = 10$ are used to test $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ and that this hypothesis is rejected. Suppose the T-K confidence intervals indicate that μ_2 is significantly different from both μ_1 and μ_4 , and that there are no other significant differences. The resulting summary display would then be

Population	4	1	3	2
Sample mean	10	19	24	27

Summarizing the Results of the Tukey-Kramer Procedure

1. List the sample means in increasing order, identifying the corresponding population just above the value of each \bar{x} .
2. Use the T-K intervals to determine the group of means that do not differ significantly from the first mean in the list. Draw a horizontal line extending from the smallest mean to the last mean in the group identified. For example, if there are five means, arranged in order,

Population	3	2	1	4	5
Sample mean	\bar{x}_3	\bar{x}_2	\bar{x}_1	\bar{x}_4	\bar{x}_5

and μ_3 is judged to be not significantly different from μ_2 or μ_1 , but is judged to be significantly different from μ_4 and μ_5 , draw the following line:

Population	3	2	1	4	5
Sample mean	\bar{x}_3	\bar{x}_2	\bar{x}_1	\bar{x}_4	\bar{x}_5

3. Use the T-K intervals to determine the group of means that are not significantly different from the second smallest mean. (You need consider only means that appear to the right of the mean under consideration.) If there is already a line connecting the second smallest mean with all means in the new group identified, no new line need be drawn. If this entire group of means is not underscored with a single line, draw a line extending from the second smallest mean to the last mean in the new group. Continuing with our example, if μ_2 is not significantly different from μ_1 but is significantly different from μ_4 and μ_5 , no new line need be drawn. However, if μ_2 is not significantly different from either μ_1 or μ_4 but is judged to be different from μ_5 , a second line is drawn as shown:

Population	3	2	1	4	5
Sample mean	\bar{x}_3	\bar{x}_2	\bar{x}_1	\bar{x}_4	\bar{x}_5

4. Continue considering the means in the order listed, adding new lines as needed.

Example 15.6 Sleep Time

Understand the context ➤

- Suppose that a biologist studied the effects of ethanol on sleep time. A sample of 20 rats of the same age was selected, and each rat was given an oral injection having a particular concentration of ethanol per body weight. The rapid eye movement (REM) sleep time for each rat was then recorded for a 24-hour period, with the results shown in the following table:

Treatment	Observations					\bar{x}
1. 0 (control)	88.6	73.2	91.4	68.0	75.2	79.28
2. 1 g/kg	63.0	53.9	69.2	50.1	71.5	61.54
3. 2 g/kg	44.9	59.5	40.2	56.3	38.7	47.92
4. 4 g/kg	31.0	39.6	45.3	25.2	22.7	32.76

Table 15.4 (an ANOVA table from SAS) leads to the conclusion that actual mean REM sleep time is not the same for all four treatments; the P -value for the F test is 0.0001.

TABLE 15.4 SAS ANOVA Table for Example 15.6

Analysis of Variance Procedure Dependent Variable: TIME					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5882.35750	1960.78583	21.09	0.0001
Error	16	1487.40000	92.96250		
Total	19	7369.75750			

Do the work ➤ The T-K intervals are

Difference	Interval	Includes 0?
$\mu_1 - \mu_2$	17.74 ± 17.446	no
$\mu_1 - \mu_3$	31.36 ± 17.446	no
$\mu_1 - \mu_4$	46.24 ± 17.446	no
$\mu_2 - \mu_3$	13.08 ± 17.446	yes
$\mu_2 - \mu_4$	28.78 ± 17.446	no
$\mu_3 - \mu_4$	15.16 ± 17.446	yes

Interpret the results ➤ The only T-K intervals that include zero are those for $\mu_2 - \mu_3$ and $\mu_3 - \mu_4$. The corresponding underscoring pattern is

$$\begin{array}{cccc} \bar{x}_4 & \bar{x}_3 & \bar{x}_2 & \bar{x}_1 \\ 32.76 & 47.92 & 61.54 & 79.28 \end{array}$$

Figure 15.7 displays the SAS output that agrees with this underscoring. In the SAS output, letters are used to indicate groupings in place of the underscoring.

FIGURE 15.7
SAS output for Example 15.6.

```
Alpha = 0.05 df = 16 MSE = 92.9625
Critical Value of Studentized Range = 4.046
Minimum Significant Difference = 17.446
Means with the same letter are not significantly different.

Tukey Grouping      Mean      N      Treatment
          A        79.280     5      0 (control)
          B        61.540     5      1 g/kg
C       B        47.920     5      2 g/kg
          C        32.760     5      4 g/kg
```

Example 15.7 Mug Color and the Taste of Coffee

The paper “[Does the Colour of the Mug Influence the Taste of the Coffee?” \(Flavour \[2014\]: 1–7\)](#) describes an experiment to investigate whether the color of a coffee mug has an effect on how people perceive the flavor intensity of coffee. In this experiment, 18 volunteers were randomly assigned to three groups. All were served the same coffee, but those in first group received their coffee in a white mug, those in the second group received their coffee in a blue mug, and those in the third group received their coffee in a clear glass mug. After tasting the coffee, each person rated the flavor intensity of the coffee. The means for intensity rating in the three treatment groups (approximate values from a graph in the paper) were 50 for the white mug group, 42 for the blue mug group, and 23 for the clear glass mug group.

The researchers used an ANOVA F test to test the null hypothesis of no difference in the mean intensity rating for the three treatments. They reported an F statistic value of 4.78 and a P -value of 0.025. Based on this P -value, they rejected the null hypothesis and concluded that the means for intensity rating were not equal for all three mug colors. A multiple comparison procedure was used to identify differences in treatment means, and these conclusions were reached:

- Coffee was rated as significantly more intense when served in a white mug than when served in a clear glass mug.
- None of the other differences were found to be significant.

These results are summarized in the following underscoring pattern. The treatment group means are arranged in order from smallest to largest.

$$\begin{array}{ccc} \text{Clear} & \text{Blue} & \text{White} \\ 23 & 42 & 50 \end{array}$$

Based on this analysis, we would conclude that there is a difference between the means for intensity rating for coffee served in a white mug and coffee served in a clear glass mug. The effect of a blue mug is unclear, and there is not convincing evidence of a difference between the means for intensity rating of coffee served in a blue mug and coffee served in a white mug. There is also no convincing evidence of a difference for coffee served in a blue mug and coffee served in a clear glass mug.

EXERCISES 15.14 - 15.22

● Data set available online

- 15.14** Leaf surface area is an important variable in plant gas-exchange rates. Dry matter per unit surface area (mg/cm^3) was measured for trees raised under three different growing conditions. Let μ_1 , μ_2 , and μ_3 represent the mean dry matter per unit surface area for the growing conditions 1, 2, and 3, respectively. Suppose that the 95% T-K confidence intervals are

Difference	Interval
$\mu_1 - \mu_2$:	(-3.11, -1.11)
$\mu_1 - \mu_3$:	(-4.06, -2.06)
$\mu_2 - \mu_3$:	(-1.95, 0.05)

Which of the following four statements do you think describes the relationship between μ_1 , μ_2 , and μ_3 ? Explain your choice. (Hint: See Example 15.5.)

- a. $\mu_1 = \mu_2$, and μ_3 differs from μ_1 and μ_2 .
- b. $\mu_1 = \mu_3$, and μ_2 differs from μ_1 and μ_3 .
- c. $\mu_2 = \mu_3$, and μ_1 differs from μ_2 and μ_3 .
- d. All three μ 's are different from one another.

- 15.15** The paper “[Trends in Blood Lead Levels and Blood Lead Testing among U.S. Children Aged 1 to 5 Years](#)” (*Pediatrics* [2009]: e376–e385) gave data on blood lead levels (in $\mu\text{g}/\text{dL}$) for samples of children living in homes that had been classified either at low, medium, or high risk of lead exposure based on when the home was constructed. After using a multiple comparison procedure, the authors reported the following:

1. The difference in mean blood lead level between low-risk housing and medium-risk housing was significant.
2. The difference in mean blood lead level between low-risk housing and high-risk housing was significant.
3. The difference in mean blood lead level between medium-risk housing and high-risk housing was significant.

Which of the following sets of T-K intervals (Set 1, 2, or 3) is consistent with the authors’ conclusions? Explain your choice. (Hint: See Example 15.5.)

$$\begin{aligned}\mu_L &= \text{mean blood lead level for children living in low-risk housing} \\ \mu_M &= \text{mean blood lead level for children living in medium-risk housing} \\ \mu_H &= \text{mean blood lead level for children living in high-risk housing}\end{aligned}$$

Difference	Set 1	Set 2	Set 3
$\mu_L - \mu_M$	(-0.6, 0.1)	(-0.6, -0.1)	(-0.6, -0.1)
$\mu_L - \mu_H$	(-1.5, -0.6)	(-1.5, -0.6)	(-1.5, -0.6)
$\mu_M - \mu_H$	(-0.9, -0.3)	(-0.9, 0.3)	(-0.9, -0.3)

- 15.16** The accompanying underscoring pattern appears in the article “[Women’s and Men’s Eating Behavior Following Exposure to Ideal-Body Images and Text](#)” (*Communications Research* [2006]: 507–529). Women either viewed slides depicting images of thin female models with no text (treatment 1), viewed the same slides accompanied by diet and exercise-related text (treatment 2), or viewed the same slides accompanied by text that was unrelated to diet and exercise (treatment 3). A fourth group of women did not view any slides (treatment 4). Participants were assigned at random to the four treatments. Participants were then asked to complete a questionnaire in a room where pretzels were set out on the tables. An observer recorded how many pretzels participants ate while completing the questionnaire.

Write a few sentences interpreting this underscoring pattern. (Hint: See Example 15.7.)

Treatment:	2	1	4	3
Mean number of pretzels consumed:	0.97	1.03	2.20	2.65

- 15.17** The paper referenced in the previous exercise also gave the following underscoring pattern for men:

Treatment:	2	1	3	4
Mean number of pretzels consumed:	6.61	5.96	3.38	2.70

- Write a few sentences interpreting this underscoring pattern.
- Using your answers from Part (a) and from the previous exercise, write a few sentences describing the differences between how men and women respond to the treatments.

- 15.18** • The paper referenced in Exercise 15.5 described an experiment to determine if restrictive age labeling on video games increased the attractiveness of the game for boys age 12 to 13. In that exercise, the null hypothesis $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$, where μ_1 is the population mean attractiveness rating for the game with the 7+ age label, and μ_2 , μ_3 , and μ_4 are the population mean attractiveness scores for the 12+, 16+, and 18+ age labels, respectively, was rejected. The sample data are given in the accompanying table.

7 + label	12 + label	16 + label	18 + label
6	8	7	10
6	7	9	9
6	8	8	6
5	5	6	8
4	7	7	7
8	9	4	6
6	5	8	8
1	8	9	9
2	4	6	10
4	7	7	8

- Calculate the 95% T-K intervals and then use the underscoring procedure described in this section to identify significant differences among the age labels.
- Based on your answer to Part (a), write a few sentences commenting on the theory that the more restrictive the age label on a video game, the more attractive the game is to 12- to 13-year-old boys.

- 15.19** In an experiment to investigate the effect of the portrayal of female characters in superhero movies, researchers randomly assigned female college students to one of three groups (["The Empowering \(Super\) Heroine? The Effects of Sexualized Female Characters in Superhero Films on Women," Sex Roles \[2015\]: 211–220](#)). One group was a control group, one group watched 13 minutes of video scenes from the movie *Spider-Man* (where a sexy female character was portrayed as a victim), and one group watched 13 minutes of video scenes from the movie *X-Men* (where a sexy female character was portrayed as a heroine). The women in the control group did not watch a video. The women in all three groups then completed a questionnaire and their answers were used to calculate a measure of gender stereotyping, with lower values indicating attitudes more accepting of equality of women and men.

The researchers used a one-way ANOVA to analyze the data. The following Minitab ANOVA output and summary statistics are based on data consistent with information and conclusions from the paper.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	2	3.264	1.6321	3.34	0.041
Error	79	38.632	0.4890		
Total	81	41.896			

Means

Factor	N	Mean	StDev	95% CI
Control	29	2.668	0.673	(2.409, 2.926)
Spider Man	30	3.120	0.800	(2.866, 3.374)
X-Men	23	3.020	0.580	(2.730, 3.310)

- Use the given output to test the null hypothesis of no difference in the means for gender stereotyping score for the three different treatment groups. Use a significance level of 0.05.
- Use the Tukey-Kramer procedure to calculate 95% simultaneous confidence intervals for all relevant differences between means and give the corresponding underscore pattern.
- Describe what you learn from the results of the Tukey-Kramer procedure.

- 15.20** • The accompanying data resulted from a flammability study in which specimens of five different fabrics were tested to determine burn times.

Fabric	1	17.8	16.2	15.9	15.5
	2	13.2	10.4	11.3	
	3	11.8	11.0	9.2	10.0
	4	16.5	15.3	14.1	15.0
	5	13.9	10.8	12.8	11.7

$$MSTr = 23.67$$

$$MSE = 1.39$$

$$F = 17.08$$

$$P\text{-value} = 0.000$$

The accompanying output gives the T-K intervals as calculated by Minitab. Identify significant differences and give the underscoring pattern.

Individual error rate = 0.00750
 Critical value = 4.37
 Intervals for (column level mean) - (row level mean)

	1	2	3	4
2	1.938			
3	7.495	3.278	-1.645	
4	8.422	8.422	3.912	-1.050
5	-1.050	-5.983	-6.900	3.830
	1.478	-0.670	-2.020	1.478
	6.622	2.112	0.772	6.622

- 15.21** Do lizards play a role in spreading plant seeds? Some research carried out in South Africa would suggest so (“**Dispersal of Namaqua Fig [Ficus cordata cordata] Seeds by the Augrabies Flat Lizard [Platysaurus broadleyi]**,” *Journal of Herpetology* [1999]: 328–330). The researchers collected 400 seeds of this particular type of fig, 100 of which were from each treatment: lizard dung, bird dung, rock hyrax dung, and uneaten figs. They planted these seeds in batches of 5, and for each group of 5 they recorded how many of the seeds germinated. This resulted in 20 observations for each treatment. The treatment means and standard deviations are given in the accompanying table.

Treatment	n	\bar{x}	s
Uneaten figs	20	2.40	0.30
Lizard dung	20	2.35	0.33
Bird dung	20	1.70	0.34
Hyrax dung	20	1.45	0.28

- a. Construct the appropriate ANOVA table, and test the hypothesis that there is no difference between the means for the number of seeds germinating for the four treatments.

- b. Is there evidence that seeds eaten and then excreted by lizards germinate at a higher rate, on average, than those eaten and then excreted by birds? Give statistical evidence to support your answer.

- 15.22** ● Samples of six different brands of diet or imitation margarine were analyzed to determine the level of physiologically active polyunsaturated fatty acids (PAPUFA, in percent), resulting in the data shown in the accompanying table. (The data are fictitious, but the sample means agree with data reported in *Consumer Reports*.)

Imperial	14.1	13.6	14.4	14.3	
Parkay	12.8	12.5	13.4	13.0	12.3
Blue Bonnet	13.5	13.4	14.1	14.3	
Chiffon	13.2	12.7	12.6	13.9	
Mazola	16.8	17.2	16.4	17.3	18.0
Fleischmann's	18.1	17.2	18.7	18.4	

- a. Test for differences among the true means for PAPUFA percentage among the different brands. Use $\alpha = 0.05$.
- b. Use the T-K procedure to calculate 95% simultaneous confidence intervals for all differences between means and give the corresponding underscoring pattern.

CHAPTER ACTIVITY

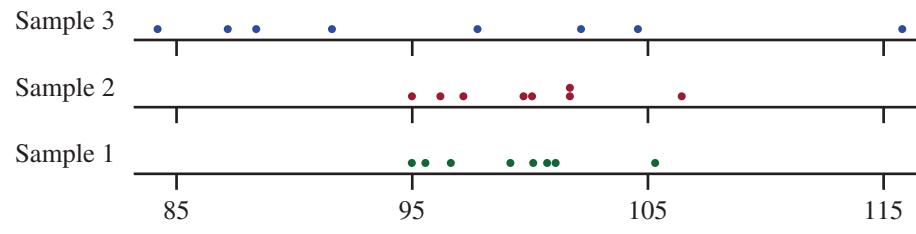
ACTIVITY 15.1 EXPLORING SINGLE-FACTOR ANOVA

Working with a partner, consider the following:

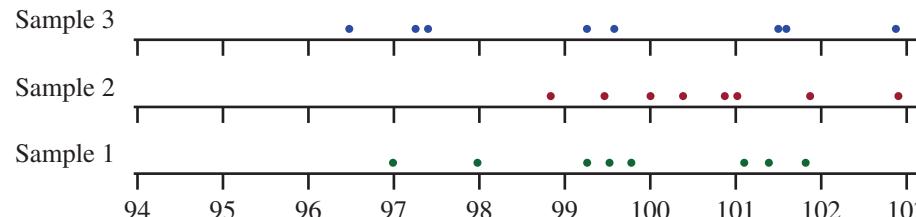
1. Each of the four accompanying graphs shows a dotplot of data from three separate random samples. For each

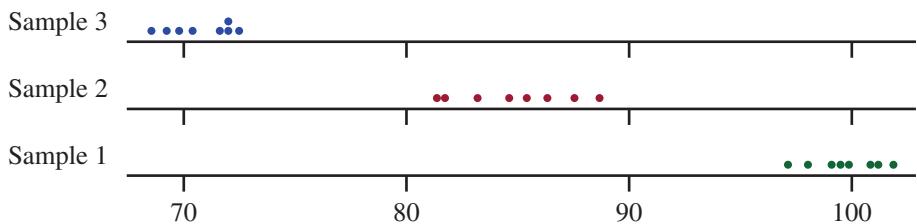
of the four graphs, indicate whether you think that the basic assumptions for single-factor ANOVA are plausible. Write a sentence or two justifying your answer.

Graph 1

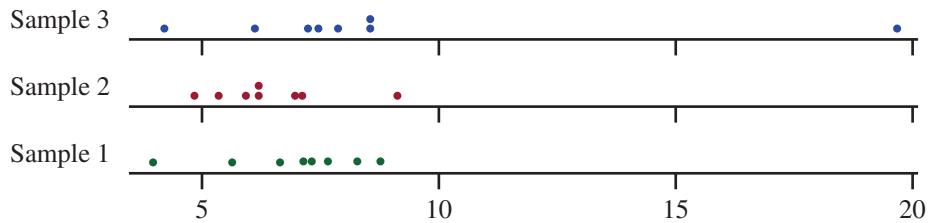


Graph 2





Graph 3

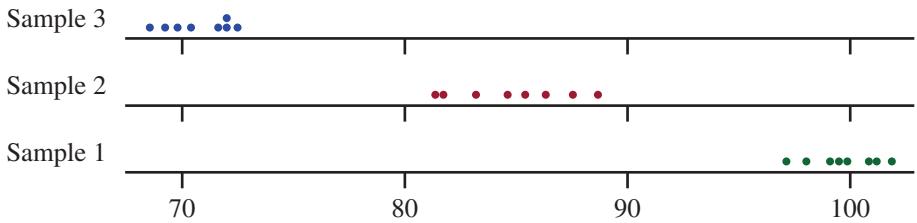


Graph 4

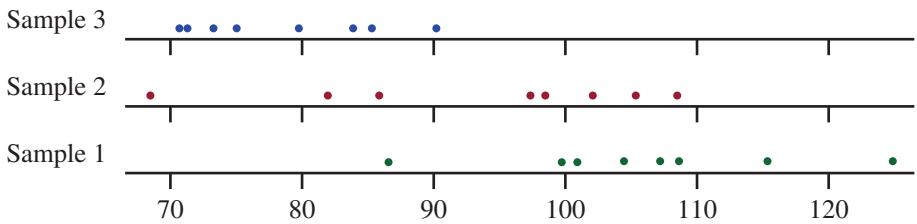
2. Each of the three accompanying graphs shows a dotplot of data from three separate random samples. For each of the three graphs, indicate whether you think that the three population means are probably not all the same, you think that the three population means might be the same, or you are unsure whether the population means could be

the same. Write a sentence or two explaining your reasoning.

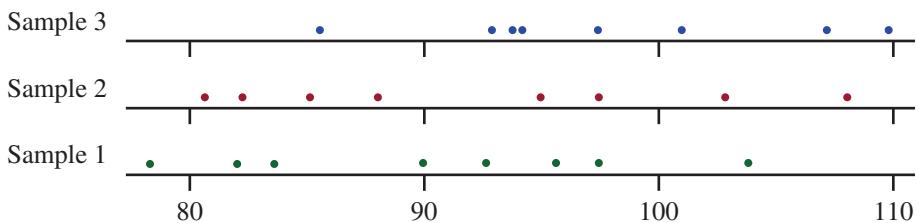
3. Sample data for each of the three graphs in Step 2 are shown in the table on the following page. For each of the three graphs, carry out a single-factor ANOVA. Are the results of the F tests consistent with your answers in Step 2? Explain.



Graph A



Graph B



Graph C

GRAPH A			GRAPH B			GRAPH C		
Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
99.7	91.3	69.3	104.2	81.9	71.7	82.3	82.4	94.2
98.0	82.0	72.1	107.0	105.4	79.7	97.4	87.5	109.8
101.4	83.6	71.7	88.6	98.4	70.9	83.7	97.3	94.9
99.2	84.8	69.9	99.6	108.4	76.6	103.6	102.6	85.8
101.0	86.5	68.8	124.3	102.1	85.3	78.6	94.8	97.4
101.8	91.5	70.7	100.7	68.9	90.3	90.1	81.3	101.0
99.5	81.8	72.7	108.3	85.8	84.2	92.8	85.2	93.0
97.0	85.5	72.2	116.5	97.5	74.6	95.5	107.8	107.1

SUMMARY Key Concepts and Formulas

TERM OR FORMULA

Single-factor analysis of variance (ANOVA)

COMMENT

A test procedure for determining whether there are significant differences among k population or treatment means. The hypotheses tested are $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ versus H_a : at least two μ 's differ.

Treatment sum of squares:

$$\text{SSTr} = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + \dots + n_k(\bar{x}_k - \bar{\bar{x}})^2$$

A measure of how different the k sample means $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ are from one another; associated $\text{df}_1 = k - 1$.

Error sum of squares:

$$\text{SSE} = (n_1 - 1)s_1^2 + \dots + (n_k - 1)s_k^2$$

A measure of the amount of variability within the individual samples; associated $\text{df}_2 = N - k$, where $N = n_1 + \dots + n_k$.

Mean square

A sum of squares divided by its df. For single-factor ANOVA, $\text{MSTr} = \text{SSTr}/(k - 1)$ and $\text{MSE} = \text{SSE}/(N - k)$.

$$F = \frac{\text{MSTr}}{\text{MSE}}$$

The test statistic for testing $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ in a single-factor ANOVA. When H_0 is true, F has an F distribution with numerator $\text{df}_1 = k - 1$ and denominator $\text{df}_2 = N - k$.

$$\text{SSTo} = \text{SSTr} + \text{SSE}$$

The fundamental identity in single-factor ANOVA, where $\text{SSTo} = \text{total sum of squares} = \sum(x - \bar{\bar{x}})^2$.

Tukey-Kramer multiple comparison procedure

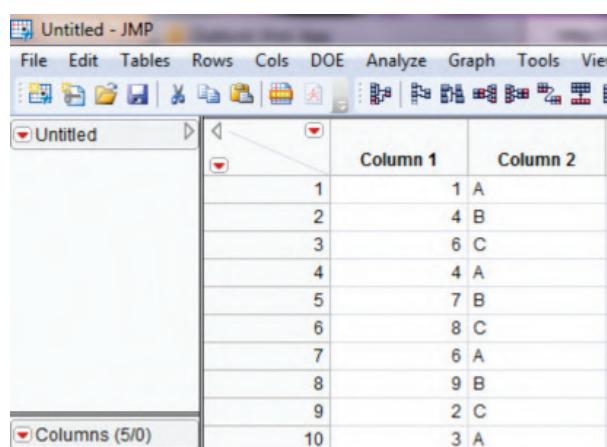
A procedure for identifying significant differences among the μ 's once the hypothesis $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ has been rejected by the ANOVA F test.

TECHNOLOGY NOTES

ANOVA

JMP

1. Input the raw data into the first column
2. Input the group information into the second column



3. Click **Analyze** then select **Fit Y by X**
4. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Response**
5. Click and drag the second column name from the box under **Select Columns** to the box next to **X, Factor**
6. Click **OK**
7. Click the red arrow next to **Oneway Analysis of...** and select **Means/ANOVA**

Minitab

Data stored in separate columns

1. Input each group's data in a separate column

Worksheet 3 ***				
	C1	C2	C3	C4
	Group 1	Group 2	Group 3	
1	1	4	1	
2	2	5	5	
3	3	6	6	
4	4	1	7	
5	5	7	5	
6	6	8	6	
7	7	9	4	
8	8	5	5	
9	9	6	6	
10	10	4	8	

2. Click **Stat** then **ANOVA** then **One-Way (Unstacked)...**
3. Click in the box under **Responses (in separate columns):**
4. Double-click the column name containing each group's data
5. Click **OK**

Data stored in one column

1. Input the data into one column
2. Input the group information into a second column

Worksheet 3 ***		
	C1	C2
	Data	Group
7	7	1
8	8	1
9	9	1
10	10	1
11	4	2
12	5	2
13	6	2
14	1	2
15	7	2
16	8	2
17	9	2

3. Click **Stat** then **ANOVA** then **One-Way...**
4. Click in the box next to **Response:** and double-click the column name containing the raw data values
5. Click in the box next to **Factor:** and double-click the column name containing the group information
6. Click **OK**

SPSS

1. Input the raw data for all groups into one column
2. Input the group information into a second column (use group numbers)

	VAR00001	VAR00002	var	var
1	1.00	1.00		
2	2.00	1.00		
3	3.00	1.00		
4	4.00	1.00		
5	5.00	1.00		
6	6.00	1.00		
7	7.00	1.00		
8	8.00	1.00		
9	9.00	1.00		
10	10.00	1.00		
11	2.00	2.00		
12	4.00	2.00		
13	3.00	2.00		
14	5.00	2.00		

3. Click **Analyze** then click **Compare Means** then click **One-Way ANOVA...**
4. Click the name of the column containing the raw data and click the arrow to move it to the box under **Dependent List:**
5. Click the name of the column containing the group data and click the arrow to move it to the box under **Factor:**
6. Click **OK**

Excel

1. Input the raw data for each group into a separate column
2. Click the **Data** ribbon
3. Click **Data Analysis** in the **Analysis** group

Note: If you do not see **Data Analysis** listed on the Ribbon, see the Technology Notes for Chapter 2 for instructions on installing this add-on.

4. Select **Anova: Single Factor** and click **OK**
5. Click on the box next to **Input Range** and select ALL columns of data (if you typed and selected column titles, click the box next to **Labels in First Row**)
6. Click in the box next to **Alpha** and type the significance level
7. Click **OK**

Note: The test statistic and p-value can be found in the first row of the table under *F* and *P-value*, respectively.

TI-83/84

1. Enter the data for each group into a separate list starting with **L1** (In order to access lists press the **STAT** key, highlight the option called **Edit...** then press **ENTER**)
2. Press **STAT**
3. Highlight **TESTS**
4. Highlight **ANOVA** and press **ENTER**
5. Press **2nd** then **1**
6. Press,
7. Press **2nd** then **2**
8. Press,
9. Continue to input lists where data is stored separated by commas until you input the final list
10. When you are finished entering all lists, press **)**
11. Press **ENTER**

TI-Nspire**Summarized Data**

1. Enter the summary information for the first group in a list in the following order: the value for n followed by a comma then the value of \bar{x} followed by a comma then the value of s (In order to access data lists select the spreadsheet option and press **enter**)

Note: Be sure to title the lists by selecting the top row of the column and typing a title.

2. Enter the summary information for the first group in a list in the following order: the value for n followed by a comma then the value of \bar{x} followed by a comma then the value of s
3. Continue to enter summary information for each group in this manner

4. When you are finished entering data for each group, press **menu** then **4:Statistics** then **4:Stat Tests** then **C:ANOVA...** then press **enter**
5. For **Data Input Method** choose **Stats** from the drop-down menu
6. For **Number of Groups** enter the number of groups, k
7. In the box next to **Group 1 Stats** select the list containing group one's summary statistics
8. In the box next to **Group 2 Stats** select the list containing group one's summary statistics
9. Continue entering summary statistics in this manner for all groups
10. Press **OK**

Raw data

1. Enter each group's data into separate data lists (In order to access data lists, select the spreadsheet option and press **enter**)

Note: Be sure to title the lists by selecting the top row of the column and typing a title.

2. Press the **menu** key and select **4:Statistics** then **4:Stat Tests** then **C:ANOVA...** and press **enter**
3. For **Data Input Method** choose **Data** from the drop-down menu
4. For **Number of Groups** input the number of groups, k
5. Press **OK**
6. For **List 1** select the list title that contains group one's data from the drop-down menu
7. For **List 2** select the list title that contains group two's data from the drop-down menu
8. Continue to select the appropriate lists for all groups
9. When you are finished inputting lists press **OK**

Appendix

Appendix

Statistical Tables

TABLE 1 Random Numbers 786

TABLE 2 Standard Normal Probabilities (Cumulative z Curve Areas) 788

TABLE 3 t Critical Values 790

TABLE 4 Tail Areas for t Curves 791

TABLE 5 Curves of $\beta = P(\text{Type II Error})$ for t Tests 794

TABLE 6 Values That Capture Specified Upper-Tail F Curve Areas 795

TABLE 7 Critical Values of q for the Studentized Range Distribution 799

TABLE 8 Upper-Tail Areas for Chi-Square Distributions 800

TABLE 9 Binomial Probabilities 802

Appendix: Statistical Tables

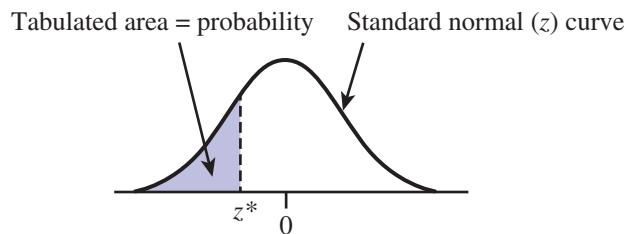
TABLE 1 Random Numbers

Row	4	5	1	8	5	0	3	3	7	1	2	8	4	5	1	1	0	9	5	7
1	4	5	1	8	5	0	3	3	7	1	2	8	4	5	1	1	0	9	5	7
2	4	2	5	5	8	0	4	5	7	0	7	0	3	6	6	1	3	1	3	1
3	8	9	9	3	4	3	5	0	6	3	9	1	1	8	2	6	9	2	0	9
4	8	9	0	7	2	9	9	0	4	7	6	7	4	7	1	3	4	3	5	3
5	5	7	3	1	0	3	7	4	7	8	5	2	0	1	3	7	7	6	3	6
6	0	9	3	8	7	6	7	9	9	5	6	2	5	6	5	8	4	2	6	4
7	4	1	0	1	0	2	2	0	4	7	5	1	1	9	4	7	9	7	5	1
8	6	4	7	3	6	3	4	5	1	2	3	1	1	8	0	0	4	8	2	0
9	8	0	2	8	7	9	3	8	4	0	4	2	0	8	9	1	2	3	3	2
10	9	4	6	0	6	9	7	8	8	2	5	2	9	6	0	1	4	6	0	5
11	6	6	9	5	7	4	4	6	3	2	0	6	0	8	9	1	3	6	1	8
12	0	7	1	7	7	7	2	9	7	8	7	5	8	8	6	9	8	4	1	0
13	6	1	3	0	9	7	3	3	6	6	0	4	1	8	3	2	6	7	6	8
14	2	2	3	6	2	1	3	0	2	2	6	6	9	7	0	2	1	2	5	8
15	0	7	1	7	4	2	0	0	0	1	3	1	2	0	4	7	8	4	1	0
16	6	6	5	1	6	1	8	1	5	5	2	6	2	0	1	1	5	2	3	6
17	9	9	6	2	5	3	5	9	8	3	7	5	0	1	3	9	3	8	0	8
18	9	9	9	6	1	2	9	3	4	6	5	6	4	6	5	8	2	7	4	0
19	2	5	6	3	1	9	8	1	1	0	3	5	6	7	9	1	4	5	2	0
20	5	1	1	9	8	1	2	1	1	6	9	8	1	8	1	9	9	1	2	0
21	1	9	8	0	7	4	6	8	4	0	3	0	8	1	1	0	6	2	3	2
22	9	7	0	9	6	3	8	9	9	7	0	6	5	4	3	6	5	0	3	2
23	1	7	6	4	8	2	0	3	9	6	3	6	2	1	0	7	7	3	1	7
24	6	2	5	8	2	0	7	8	6	4	6	6	8	9	2	0	6	9	0	4
25	1	5	7	1	1	1	9	5	1	4	5	2	8	3	4	3	0	7	3	5
26	1	4	6	6	5	6	0	1	9	4	0	5	2	7	6	4	3	6	8	8
27	1	8	5	0	2	1	6	8	0	7	7	2	6	2	6	7	5	4	8	7
28	7	8	7	4	6	5	4	3	7	9	3	9	2	7	9	5	4	2	3	1
29	1	6	3	2	8	3	7	3	0	7	2	4	8	0	9	9	9	4	7	0
30	2	8	9	0	8	1	6	8	1	7	3	1	3	0	9	7	2	5	7	9
31	0	7	8	8	6	5	7	5	5	4	0	0	3	4	1	2	7	3	7	9
32	8	4	0	1	4	5	1	9	1	1	2	1	5	3	2	8	5	5	7	5
33	7	3	5	9	7	0	4	9	1	2	1	3	2	5	1	9	3	3	8	3
34	4	7	2	6	7	6	9	9	2	7	8	7	5	5	5	2	4	4	3	4
35	9	3	3	7	0	7	0	5	7	5	6	9	5	4	3	1	4	6	6	8
36	0	2	4	9	7	8	1	6	3	8	7	8	0	5	6	7	2	7	5	0
37	7	1	0	1	8	4	7	1	2	9	3	8	0	0	8	7	9	2	8	6
38	9	7	9	4	4	5	3	1	9	3	4	5	0	6	3	5	9	6	9	8
39	0	4	2	5	0	0	9	9	6	4	0	6	9	0	3	8	3	5	7	2
40	0	7	1	2	3	6	1	7	9	3	9	5	4	6	8	4	8	8	0	6
41	3	5	6	6	2	4	4	5	6	3	7	8	7	6	5	2	0	4	3	2
42	6	6	8	5	5	2	9	7	9	3	3	1	6	9	5	9	7	1	1	2
43	9	5	0	4	3	1	1	7	3	9	2	7	7	4	7	0	3	1	2	8
44	5	1	7	8	9	4	7	2	9	2	8	9	9	8	0	6	3	7	2	1
45	1	6	3	9	4	1	3	2	1	1	8	5	6	3	4	1	9	3	1	7
46	4	4	8	6	4	0	3	8	3	8	3	5	9	5	9	4	8	3	9	4
47	7	7	6	6	4	5	4	4	8	4	4	0	3	9	8	5	2	0	2	3
48	2	5	6	6	3	7	0	6	5	6	9	0	1	9	5	2	6	9	1	2

(Continued)

TABLE 1 Random Numbers (*Continued*)

Row	9	4	0	4	7	5	3	2	8	7	2	7	4	9	3	9	6	5	5	6
49	9	4	0	4	7	5	3	2	8	7	2	7	4	9	3	9	6	5	5	6
50	7	3	1	5	6	6	5	0	3	5	3	7	2	8	6	2	4	1	8	7
51	7	5	8	2	8	8	8	7	6	4	1	1	0	2	3	1	9	3	6	0
52	3	3	6	0	9	1	1	0	3	2	7	8	2	0	5	3	4	8	9	8
53	0	2	9	6	9	8	9	3	8	1	5	3	9	9	7	0	7	7	1	6
54	8	5	9	6	2	9	6	8	2	1	2	4	7	0	6	8	3	4	6	1
55	5	4	7	6	1	0	0	1	0	4	6	1	4	1	5	0	9	6	5	5
56	5	0	3	6	4	1	9	8	4	4	1	2	0	2	5	1	8	1	2	1
57	0	2	6	3	7	5	1	1	6	6	0	5	8	1	2	3	3	6	1	3
58	3	8	1	6	3	8	1	4	5	2	9	4	2	5	7	3	2	3	1	8
59	9	1	5	6	0	6	5	6	6	3	6	2	3	0	0	0	1	8	5	9
60	5	3	5	6	3	9	5	4	7	3	6	6	7	5	0	1	5	6	7	3
61	9	6	6	4	5	7	7	6	1	5	4	4	8	0	6	5	7	6	3	0
62	6	3	0	6	7	9	5	5	4	6	2	2	8	4	4	0	0	9	9	8
63	8	5	8	3	5	2	0	6	6	0	0	6	0	6	3	0	1	7	0	5
64	3	8	2	4	9	0	9	2	6	2	9	5	1	9	1	9	0	8	3	3
65	1	4	4	1	1	7	4	6	3	6	5	6	5	5	7	7	0	3	5	8
66	5	9	9	5	3	7	2	5	1	7	1	1	0	7	1	0	9	2	8	8
67	8	7	1	7	5	2	5	6	8	7	9	9	1	3	9	6	4	9	3	0
68	6	7	2	3	1	4	9	2	1	7	0	8	6	7	8	9	9	4	7	4
69	2	3	2	8	7	0	9	7	1	1	1	2	8	2	9	1	0	6	7	7
70	2	9	5	7	8	4	7	9	0	3	6	9	2	0	6	0	6	2	6	8
71	4	8	9	8	3	2	7	6	9	1	9	8	6	9	5	2	4	9	9	9
72	1	5	6	5	7	7	5	4	3	4	3	8	1	8	9	9	4	4	1	1
73	1	8	1	1	7	2	8	5	5	8	9	9	9	6	2	0	1	6	6	7
74	5	7	7	0	9	5	5	6	8	6	8	2	2	6	0	5	5	1	8	7
75	1	8	6	0	5	4	8	3	4	5	3	5	8	7	7	8	5	7	0	
76	2	6	6	7	9	4	2	2	8	7	4	3	4	9	6	1	9	4	3	9
77	3	6	6	4	5	7	8	3	0	2	8	4	6	7	2	1	4	5	2	3
78	0	7	8	0	1	2	1	1	3	4	2	1	6	9	3	3	5	4	0	4
79	8	3	6	0	5	7	7	9	1	5	8	8	4	9	5	7	2	2	7	6
80	5	3	6	9	0	6	3	8	7	5	9	5	9	7	4	2	5	6	2	9
81	0	9	3	7	7	2	8	6	4	3	2	9	4	8	2	9	9	6	9	9
82	9	4	7	4	0	0	0	3	5	4	6	6	2	6	2	3	6	1	1	4
83	5	5	4	1	7	8	6	4	2	3	2	9	8	4	6	3	8	3	0	5
84	5	3	0	0	5	4	8	0	7	4	7	6	2	1	1	2	1	2	6	9
85	3	3	0	9	3	2	9	4	0	5	5	4	8	7	5	7	5	3	8	8
86	3	0	5	7	1	9	5	8	0	0	4	5	3	0	3	0	2	7	6	7
87	5	0	8	6	0	8	1	6	2	0	8	6	5	4	0	7	2	9	1	0
88	3	6	4	7	8	2	3	5	7	9	8	5	2	7	6	9	0	2	4	9
89	9	0	4	4	9	1	6	8	5	2	8	9	0	7	5	7	2	5	1	8
90	9	5	2	6	9	3	9	6	5	1	8	8	7	8	2	0	4	4	7	9
91	9	4	5	7	0	3	4	6	4	2	5	4	8	6	1	1	9	1	8	8
92	8	1	1	8	0	5	4	2	8	5	3	3	3	0	1	1	4	4	8	3
93	6	9	4	7	8	3	3	9	1	2	5	0	1	2	3	0	1	1	2	5
94	0	0	6	8	8	7	2	4	4	7	6	6	0	3	4	7	5	6	8	2
95	5	3	3	9	3	8	4	9	1	9	1	7	8	4	5	2	2	5	4	4
96	2	5	6	2	7	6	0	3	8	1	4	4	2	6	8	3	6	3	2	8
97	7	4	3	7	9	6	8	6	2	8	3	8	4	2	2	0	7	0	5	3
98	1	9	0	8	8	0	1	2	2	2	7	5	6	5	5	7	8	7	2	6
99	2	4	8	0	2	5	2	7	0	5	9	6	6	1	5	8	7	9	7	5
100	4	1	7	8	6	7	1	1	5	8	9	4	8	9	8	3	0	9	0	7

**TABLE 2** Standard Normal Probabilities (Cumulative z Curve Areas)

z^*	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.8	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
-3.7	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-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)

Tabulated area = probability Standard normal (z) curve

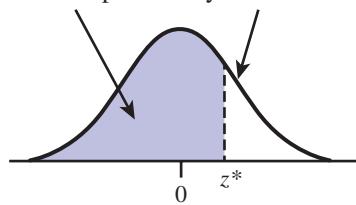
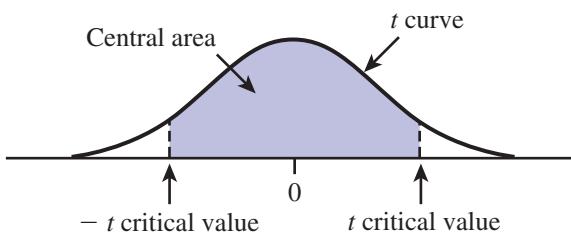
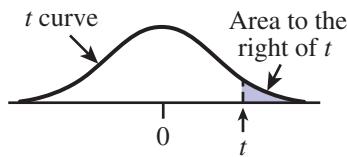


TABLE 2 Standard Normal Probabilities (Cumulative z Curve Areas) (*Continued*)

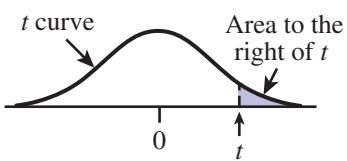
**TABLE 3** *t* Critical Values

Central area captured: Confidence level:	.80 80%	.90 90%	.95 95%	.98 98%	.99 99%	.998 99.8%	.999 99.9%	
Degrees of freedom	1	3.08	6.31	12.71	31.82	63.66	318.31	636.62
	2	1.89	2.92	4.30	6.97	9.93	23.33	31.60
	3	1.64	2.35	3.18	4.54	5.84	10.21	12.92
	4	1.53	2.13	2.78	3.75	4.60	7.17	8.61
	5	1.48	2.02	2.57	3.37	4.03	5.89	6.86
	6	1.44	1.94	2.45	3.14	3.71	5.21	5.96
	7	1.42	1.90	2.37	3.00	3.50	4.79	5.41
	8	1.40	1.86	2.31	2.90	3.36	4.50	5.04
	9	1.38	1.83	2.26	2.82	3.25	4.30	4.78
	10	1.37	1.81	2.23	2.76	3.17	4.14	4.59
	11	1.36	1.80	2.20	2.72	3.11	4.03	4.44
	12	1.36	1.78	2.18	2.68	3.06	3.93	4.32
	13	1.35	1.77	2.16	2.65	3.01	3.85	4.22
	14	1.35	1.76	2.15	2.62	2.98	3.79	4.14
	15	1.34	1.75	2.13	2.60	2.95	3.73	4.07
	16	1.34	1.75	2.12	2.58	2.92	3.69	4.02
	17	1.33	1.74	2.11	2.57	2.90	3.65	3.97
	18	1.33	1.73	2.10	2.55	2.88	3.61	3.92
	19	1.33	1.73	2.09	2.54	2.86	3.58	3.88
	20	1.33	1.73	2.09	2.53	2.85	3.55	3.85
	21	1.32	1.72	2.08	2.52	2.83	3.53	3.82
	22	1.32	1.72	2.07	2.51	2.82	3.51	3.79
	23	1.32	1.71	2.07	2.50	2.81	3.49	3.77
	24	1.32	1.71	2.06	2.49	2.80	3.47	3.75
	25	1.32	1.71	2.06	2.49	2.79	3.45	3.73
	26	1.32	1.71	2.06	2.48	2.78	3.44	3.71
	27	1.31	1.70	2.05	2.47	2.77	3.42	3.69
	28	1.31	1.70	2.05	2.47	2.76	3.41	3.67
	29	1.31	1.70	2.05	2.46	2.76	3.40	3.66
	30	1.31	1.70	2.04	2.46	2.75	3.39	3.65
	40	1.30	1.68	2.02	2.42	2.70	3.31	3.55
	60	1.30	1.67	2.00	2.39	2.66	3.23	3.46
	120	1.29	1.66	1.98	2.36	2.62	3.16	3.37
<i>z</i> critical values	∞	1.28	1.645	1.96	2.33	2.58	3.09	3.29

**TABLE 4** Tail Areas for t Curves

$t \setminus df$	1	2	3	4	5	6	7	8	9	10	11	12
0.0	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500
0.1	.468	.465	.463	.463	.462	.462	.462	.461	.461	.461	.461	.461
0.2	.437	.430	.427	.426	.425	.424	.424	.423	.423	.423	.423	.422
0.3	.407	.396	.392	.390	.388	.387	.386	.386	.386	.385	.385	.385
0.4	.379	.364	.358	.355	.353	.352	.351	.350	.349	.349	.348	.348
0.5	.352	.333	.326	.322	.319	.317	.316	.315	.315	.314	.313	.313
0.6	.328	.305	.295	.290	.287	.285	.284	.283	.282	.281	.280	.280
0.7	.306	.278	.267	.261	.258	.255	.253	.252	.251	.250	.249	.249
0.8	.285	.254	.241	.234	.230	.227	.225	.223	.222	.221	.220	.220
0.9	.267	.232	.217	.210	.205	.201	.199	.197	.196	.195	.194	.193
1.0	.250	.211	.196	.187	.182	.178	.175	.173	.172	.170	.169	.169
1.1	.235	.193	.176	.167	.162	.157	.154	.152	.150	.149	.147	.146
1.2	.221	.177	.158	.148	.142	.138	.135	.132	.130	.129	.128	.127
1.3	.209	.162	.142	.132	.125	.121	.117	.115	.113	.111	.110	.109
1.4	.197	.148	.128	.117	.110	.106	.102	.100	.098	.096	.095	.093
1.5	.187	.136	.115	.104	.097	.092	.089	.086	.084	.082	.081	.080
1.6	.178	.125	.104	.092	.085	.080	.077	.074	.072	.070	.069	.068
1.7	.169	.116	.094	.082	.075	.070	.066	.064	.062	.060	.059	.057
1.8	.161	.107	.085	.073	.066	.061	.057	.055	.053	.051	.050	.049
1.9	.154	.099	.077	.065	.058	.053	.050	.047	.045	.043	.042	.041
2.0	.148	.092	.070	.058	.051	.046	.043	.040	.038	.037	.035	.034
2.1	.141	.085	.063	.052	.045	.040	.037	.034	.033	.031	.030	.029
2.2	.136	.079	.058	.046	.040	.035	.032	.029	.028	.026	.025	.024
2.3	.131	.074	.052	.041	.035	.031	.027	.025	.023	.022	.021	.020
2.4	.126	.069	.048	.037	.031	.027	.024	.022	.020	.019	.018	.017
2.5	.121	.065	.044	.033	.027	.023	.020	.018	.017	.016	.015	.014
2.6	.117	.061	.040	.030	.024	.020	.018	.016	.014	.013	.012	.012
2.7	.113	.057	.037	.027	.021	.018	.015	.014	.012	.011	.010	.010
2.8	.109	.054	.034	.024	.019	.016	.013	.012	.010	.009	.009	.008
2.9	.106	.051	.031	.022	.017	.014	.011	.010	.009	.008	.007	.007
3.0	.102	.048	.029	.020	.015	.012	.010	.009	.007	.007	.006	.006
3.1	.099	.045	.027	.018	.013	.011	.009	.007	.006	.006	.005	.005
3.2	.096	.043	.025	.016	.012	.009	.008	.006	.006	.005	.005	.004
3.3	.094	.040	.023	.015	.011	.008	.007	.005	.005	.004	.004	.003
3.4	.091	.038	.021	.014	.010	.007	.006	.005	.004	.003	.003	.003
3.5	.089	.036	.020	.012	.009	.006	.005	.004	.003	.003	.002	.002
3.6	.086	.035	.018	.011	.008	.006	.004	.004	.003	.002	.002	.002
3.7	.084	.033	.017	.010	.007	.005	.004	.003	.002	.002	.002	.002
3.8	.082	.031	.016	.010	.006	.004	.003	.003	.002	.002	.001	.001
3.9	.080	.030	.015	.009	.006	.004	.003	.002	.002	.001	.001	.001
4.0	.078	.029	.014	.008	.005	.004	.003	.002	.002	.001	.001	.001

(Continued)

**TABLE 4** Tail Areas for t Curves (*Continued*)

$t \setminus df$	13	14	15	16	17	18	19	20	21	22	23	24
0.0	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500	.500
0.1	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461	.461
0.2	.422	.422	.422	.422	.422	.422	.422	.422	.422	.422	.422	.422
0.3	.384	.384	.384	.384	.384	.384	.384	.384	.384	.383	.383	.383
0.4	.348	.347	.347	.347	.347	.347	.347	.347	.347	.347	.346	.346
0.5	.313	.312	.312	.312	.312	.312	.311	.311	.311	.311	.311	.311
0.6	.279	.279	.279	.278	.278	.278	.278	.278	.278	.277	.277	.277
0.7	.248	.247	.247	.247	.247	.246	.246	.246	.246	.246	.245	.245
0.8	.219	.218	.218	.218	.217	.217	.217	.217	.216	.216	.216	.216
0.9	.192	.191	.191	.191	.190	.190	.190	.189	.189	.189	.189	.189
1.0	.168	.167	.167	.166	.166	.165	.165	.165	.164	.164	.164	.164
1.1	.146	.144	.144	.144	.143	.143	.143	.142	.142	.142	.141	.141
1.2	.126	.124	.124	.124	.123	.123	.122	.122	.122	.121	.121	.121
1.3	.108	.107	.107	.106	.105	.105	.105	.104	.104	.104	.103	.103
1.4	.092	.091	.091	.090	.090	.089	.089	.089	.088	.088	.087	.087
1.5	.079	.077	.077	.076	.076	.075	.075	.075	.074	.074	.074	.073
1.6	.067	.065	.065	.065	.064	.064	.063	.063	.062	.062	.062	.061
1.7	.056	.055	.055	.054	.054	.053	.053	.052	.052	.052	.051	.051
1.8	.048	.046	.046	.045	.045	.044	.044	.043	.043	.043	.042	.042
1.9	.040	.038	.038	.038	.037	.037	.036	.036	.036	.035	.035	.035
2.0	.033	.032	.032	.031	.031	.030	.030	.030	.029	.029	.029	.028
2.1	.028	.027	.027	.026	.025	.025	.025	.024	.024	.024	.023	.023
2.2	.023	.022	.022	.021	.021	.021	.020	.020	.020	.019	.019	.019
2.3	.019	.018	.018	.018	.017	.017	.016	.016	.016	.016	.015	.015
2.4	.016	.015	.015	.014	.014	.014	.013	.013	.013	.013	.012	.012
2.5	.013	.012	.012	.012	.011	.011	.011	.011	.010	.010	.010	.010
2.6	.011	.010	.010	.010	.009	.009	.009	.009	.008	.008	.008	.008
2.7	.009	.008	.008	.008	.008	.007	.007	.007	.007	.007	.006	.006
2.8	.008	.007	.007	.006	.006	.006	.006	.006	.005	.005	.005	.005
2.9	.006	.005	.005	.005	.005	.005	.005	.004	.004	.004	.004	.004
3.0	.005	.004	.004	.004	.004	.004	.004	.004	.003	.003	.003	.003
3.1	.004	.004	.004	.003	.003	.003	.003	.003	.003	.003	.003	.002
3.2	.003	.003	.003	.003	.003	.002	.002	.002	.002	.002	.002	.002
3.3	.003	.002	.002	.002	.002	.002	.002	.002	.002	.002	.002	.001
3.4	.002	.002	.002	.002	.002	.002	.002	.001	.001	.001	.001	.001
3.5	.002	.002	.002	.001	.001	.001	.001	.001	.001	.001	.001	.001
3.6	.002	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001
3.7	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001	.001
3.8	.001	.001	.001	.001	.001	.001	.001	.001	.001	.000	.000	.000
3.9	.001	.001	.001	.001	.001	.001	.000	.000	.000	.000	.000	.000
4.0	.001	.001	.001	.001	.000	.000	.000	.000	.000	.000	.000	.000

(Continued)

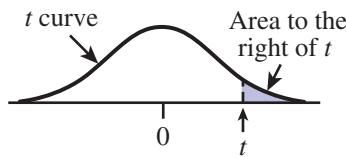
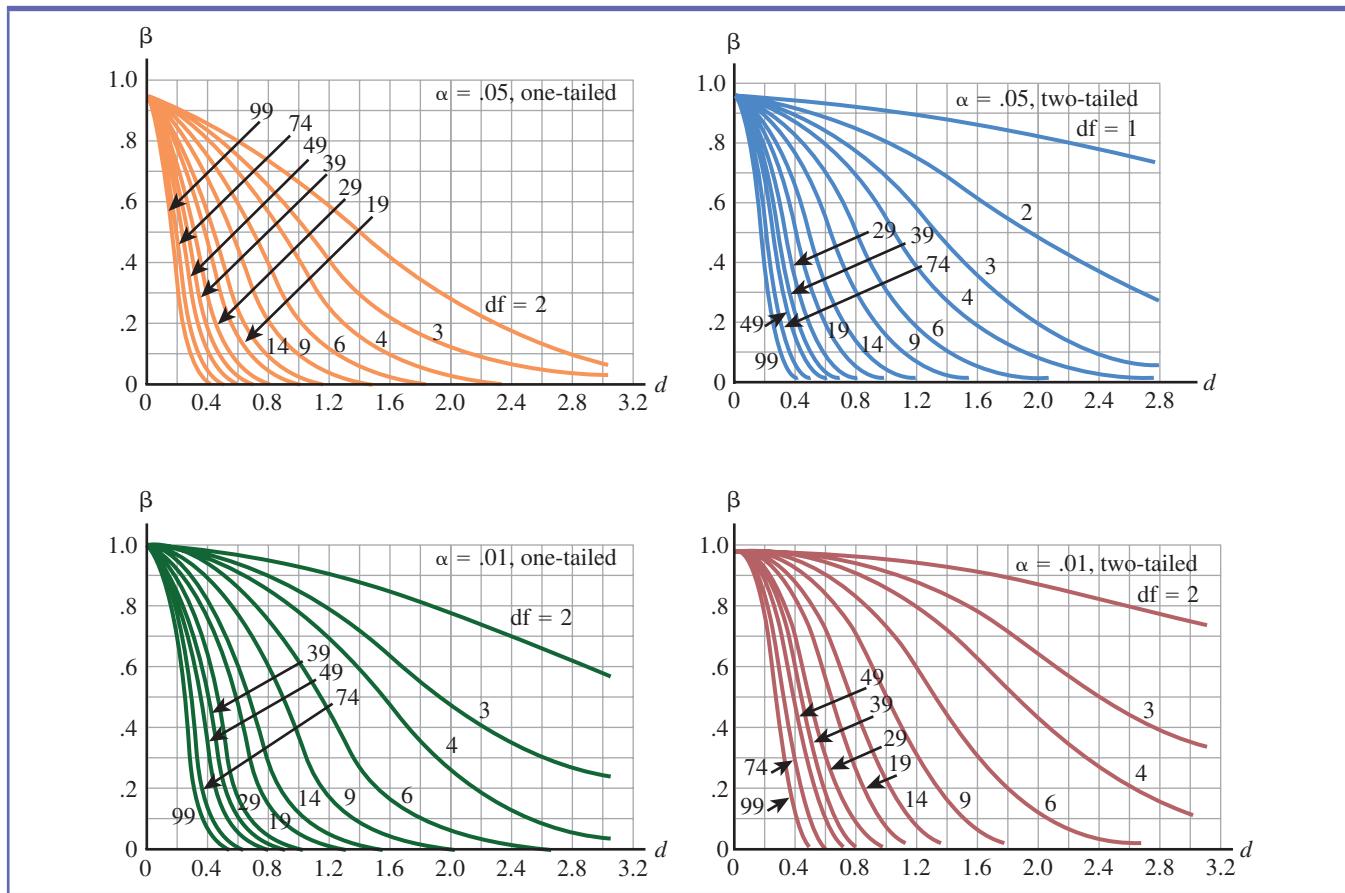
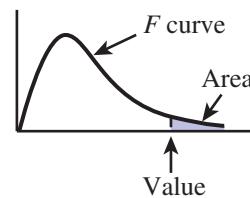


TABLE 4 Tail Areas for t Curves (*Continued*)

TABLE 5 Curves of $\beta = P(\text{Type II Error})$ for t Tests

**TABLE 6** Values That Capture Specified Upper-Tail *F* Curve Areas

df ₂	Area	df ₁									
		1	2	3	4	5	6	7	8	9	10
1	.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19
	.05	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90
	.01	4052.00	5000.00	5403.00	5625.00	5764.00	5859.00	5928.00	5981.00	6022.00	6056.00
2	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
	.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
	.001	998.50	999.00	999.20	999.20	999.30	999.30	999.40	999.40	999.40	999.40
3	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
	.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
	.001	167.00	148.50	141.10	137.10	134.60	132.80	131.60	130.60	129.90	129.20
4	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05
5	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92
6	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41
7	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08
8	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54
9	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89

(Continued)

TABLE 6 Values That Capture Specified Upper-Tail *F* Curve Areas (*Continued*)

df ₂	Area	df ₁									
		1	2	3	4	5	6	7	8	9	10
10	.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75
11	.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92
12	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29
13	.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
	.001	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80
14	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
	.001	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40
15	.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08
16	.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
	.001	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81
17	.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
	.001	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58
18	.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39
19	.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22

(Continued)

TABLE 6 Values That Capture Specified Upper-Tail *F* Curve Areas (*Continued*)

df ₂	Area	df ₁									
		1	2	3	4	5	6	7	8	9	10
20	.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08
21	.10	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
	.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	.01	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95
22	.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83
23	.10	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
	.05	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
	.001	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73
24	.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
	.001	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64
25	.10	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
	.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56
26	.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
	.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
	.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
	.001	13.74	9.12	7.36	6.41	5.80	5.38	5.07	4.83	4.64	4.48
27	.10	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
	.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
	.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
	.001	13.61	9.02	7.27	6.33	5.73	5.31	5.00	4.76	4.57	4.41
28	.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
	.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
	.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
	.001	13.50	8.93	7.19	6.25	5.66	5.24	4.93	4.69	4.50	4.35
29	.10	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
	.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
	.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
	.001	13.39	8.85	7.12	6.19	5.59	5.18	4.87	4.64	4.45	4.29

(Continued)

TABLE 6 Values That Capture Specified Upper-Tail *F* Curve Areas (*Continued*)

df ₂	Area	df ₁									
		1	2	3	4	5	6	7	8	9	10
30	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24
40	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
	.001	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87
60	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	.001	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54
90	.10	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
	.05	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
	.01	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
	.001	11.57	7.47	5.91	5.06	4.53	4.15	3.87	3.65	3.48	3.34
120	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65
	.05	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
	.001	11.38	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38	3.24
240	.10	2.73	2.32	2.10	1.97	1.87	1.80	1.74	1.70	1.65	1.63
	.05	3.88	3.03	2.64	2.41	2.25	2.14	2.04	1.98	1.92	1.87
	.01	6.74	4.69	3.86	3.40	3.09	2.88	2.71	2.59	2.48	2.40
	.001	11.10	7.11	5.60	4.78	4.25	3.89	3.62	3.41	3.24	3.09
∞	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32
	.001	10.83	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10	2.96

TABLE 7 Critical Values of q for the Studentized Range Distribution

Error df	Confidence level	Number of populations, treatments, or levels being compared							
		3	4	5	6	7	8	9	10
5	95%	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99
	99%	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24
6	95%	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49
	99%	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10
7	95%	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16
	99%	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37
8	95%	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92
	99%	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86
9	95%	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74
	99%	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49
10	95%	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60
	99%	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21
11	95%	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49
	99%	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99
12	95%	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39
	99%	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81
13	95%	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32
	99%	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67
14	95%	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25
	99%	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54
15	95%	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20
	99%	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44
16	95%	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15
	99%	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35
17	95%	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11
	99%	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27
18	95%	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07
	99%	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20
19	95%	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04
	99%	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14
20	95%	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01
	99%	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09
24	95%	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92
	99%	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92
30	95%	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82
	99%	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76
40	95%	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73
	99%	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60
60	95%	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65
	99%	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45
120	95%	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56
	99%	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30
∞	95%	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47
	99%	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16

TABLE 8 Upper-Tail Areas for Chi-Square Distributions

Right-tail area	df = 1	df = 2	df = 3	df = 4	df = 5
>0.100	<2.70	<4.60	<6.25	<7.77	<9.23
0.100	2.70	4.60	6.25	7.77	9.23
0.095	2.78	4.70	6.36	7.90	9.37
0.090	2.87	4.81	6.49	8.04	9.52
0.085	2.96	4.93	6.62	8.18	9.67
0.080	3.06	5.05	6.75	8.33	9.83
0.075	3.17	5.18	6.90	8.49	10.00
0.070	3.28	5.31	7.06	8.66	10.19
0.065	3.40	5.46	7.22	8.84	10.38
0.060	3.53	5.62	7.40	9.04	10.59
0.055	3.68	5.80	7.60	9.25	10.82
0.050	3.84	5.99	7.81	9.48	11.07
0.045	4.01	6.20	8.04	9.74	11.34
0.040	4.21	6.43	8.31	10.02	11.64
0.035	4.44	6.70	8.60	10.34	11.98
0.030	4.70	7.01	8.94	10.71	12.37
0.025	5.02	7.37	9.34	11.14	12.83
0.020	5.41	7.82	9.83	11.66	13.38
0.015	5.91	8.39	10.46	12.33	14.09
0.010	6.63	9.21	11.34	13.27	15.08
0.005	7.87	10.59	12.83	14.86	16.74
0.001	10.82	13.81	16.26	18.46	20.51
<0.001	>10.82	>13.81	>16.26	>18.46	>20.51
Right-tail area	df = 6	df = 7	df = 8	df = 9	df = 10
>0.100	<10.64	<12.01	<13.36	<14.68	<15.98
0.100	10.64	12.01	13.36	14.68	15.98
0.095	10.79	12.17	13.52	14.85	16.16
0.090	10.94	12.33	13.69	15.03	16.35
0.085	11.11	12.50	13.87	15.22	16.54
0.080	11.28	12.69	14.06	15.42	16.75
0.075	11.46	12.88	14.26	15.63	16.97
0.070	11.65	13.08	14.48	15.85	17.20
0.065	11.86	13.30	14.71	16.09	17.44
0.060	12.08	13.53	14.95	16.34	17.71
0.055	12.33	13.79	15.22	16.62	17.99
0.050	12.59	14.06	15.50	16.91	18.30
0.045	12.87	14.36	15.82	17.24	18.64
0.040	13.19	14.70	16.17	17.60	19.02
0.035	13.55	15.07	16.56	18.01	19.44
0.030	13.96	15.50	17.01	18.47	19.92
0.025	14.44	16.01	17.53	19.02	20.48
0.020	15.03	16.62	18.16	19.67	21.16
0.015	15.77	17.39	18.97	20.51	22.02
0.010	16.81	18.47	20.09	21.66	23.20
0.005	18.54	20.27	21.95	23.58	25.18
0.001	22.45	24.32	26.12	27.87	29.58
<0.001	>22.45	>24.32	>26.12	>27.87	>29.58

(Continued)

TABLE 8 Upper-Tail Areas for Chi-Square Distributions (*Continued*)

Right-tail area	df = 11	df = 12	df = 13	df = 14	df = 15
>0.100	<17.27	<18.54	<19.81	<21.06	<22.30
0.100	17.27	18.54	19.81	21.06	22.30
0.095	17.45	18.74	20.00	21.26	22.51
0.090	17.65	18.93	20.21	21.47	22.73
0.085	17.85	19.14	20.42	21.69	22.95
0.080	18.06	19.36	20.65	21.93	23.19
0.075	18.29	19.60	20.89	22.17	23.45
0.070	18.53	19.84	21.15	22.44	23.72
0.065	18.78	20.11	21.42	22.71	24.00
0.060	19.06	20.39	21.71	23.01	24.31
0.055	19.35	20.69	22.02	23.33	24.63
0.050	19.67	21.02	22.36	23.68	24.99
0.045	20.02	21.38	22.73	24.06	25.38
0.040	20.41	21.78	23.14	24.48	25.81
0.035	20.84	22.23	23.60	24.95	26.29
0.030	21.34	22.74	24.12	25.49	26.84
0.025	21.92	23.33	24.73	26.11	27.48
0.020	22.61	24.05	25.47	26.87	28.25
0.015	23.50	24.96	26.40	27.82	29.23
0.010	24.72	26.21	27.68	29.14	30.57
0.005	26.75	28.29	29.81	31.31	32.80
0.001	31.26	32.90	34.52	36.12	37.69
<0.001	>31.26	>32.90	>34.52	>36.12	>37.69
Right-tail area	df = 16	df = 17	df = 18	df = 19	df = 20
>0.100	<23.54	<24.77	<25.98	<27.20	<28.41
0.100	23.54	24.77	25.98	27.20	28.41
0.095	23.75	24.98	26.21	27.43	28.64
0.090	23.97	25.21	26.44	27.66	28.88
0.085	24.21	25.45	26.68	27.91	29.14
0.080	24.45	25.70	26.94	28.18	29.40
0.075	24.71	25.97	27.21	28.45	29.69
0.070	24.99	26.25	27.50	28.75	29.99
0.065	25.28	26.55	27.81	29.06	30.30
0.060	25.59	26.87	28.13	29.39	30.64
0.055	25.93	27.21	28.48	29.75	31.01
0.050	26.29	27.58	28.86	30.14	31.41
0.045	26.69	27.99	29.28	30.56	31.84
0.040	27.13	28.44	29.74	31.03	32.32
0.035	27.62	28.94	30.25	31.56	32.85
0.030	28.19	29.52	30.84	32.15	33.46
0.025	28.84	30.19	31.52	32.85	34.16
0.020	29.63	30.99	32.34	33.68	35.01
0.015	30.62	32.01	33.38	34.74	36.09
0.010	32.00	33.40	34.80	36.19	37.56
0.005	34.26	35.71	37.15	38.58	39.99
0.001	39.25	40.78	42.31	43.81	45.31
<0.001	>39.25	>40.78	>42.31	>43.81	>45.31

TABLE 9 Binomial Probabilities

<i>n = 5</i>													
<i>x</i>	<i>p</i>												
	0.05	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	0.95
0	.774	.590	.328	.237	.168	.078	.031	.010	.002	.001	.000	.000	.000
1	.204	.328	.410	.396	.360	.259	.156	.077	.028	.015	.006	.000	.000
2	.021	.073	.205	.264	.309	.346	.313	.230	.132	.088	.051	.008	.001
3	.001	.008	.051	.088	.132	.230	.313	.346	.309	.264	.205	.073	.021
4	.000	.000	.006	.015	.028	.077	.156	.259	.360	.396	.410	.328	.204
5	.000	.000	.000	.001	.002	.010	.031	.078	.168	.237	.328	.590	.774

<i>n = 10</i>													
<i>x</i>	<i>p</i>												
	0.05	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	0.95
0	.599	.349	.107	.056	.028	.006	.001	.000	.000	.000	.000	.000	.000
1	.315	.387	.268	.188	.121	.040	.010	.002	.000	.000	.000	.000	.000
2	.075	.194	.302	.282	.233	.121	.044	.011	.001	.000	.000	.000	.000
3	.010	.057	.201	.250	.267	.215	.117	.042	.009	.003	.001	.000	.000
4	.001	.011	.088	.146	.200	.251	.205	.111	.037	.016	.006	.000	.000
5	.000	.001	.026	.058	.103	.201	.246	.201	.103	.058	.026	.001	.000
6	.000	.000	.006	.016	.037	.111	.205	.251	.200	.146	.088	.011	.001
7	.000	.000	.001	.003	.009	.042	.117	.215	.267	.250	.201	.057	.010
8	.000	.000	.000	.000	.001	.011	.044	.121	.233	.282	.302	.194	.075
9	.000	.000	.000	.000	.000	.002	.010	.040	.121	.188	.268	.387	.315
10	.000	.000	.000	.000	.000	.000	.001	.006	.028	.056	.107	.349	.599

(Continued)

TABLE 9 Binomial Probabilities (*Continued*)

<i>n = 15</i>													
<i>x</i>	<i>p</i>												
	0.05	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	0.95
0	.463	.206	.035	.013	.005	.000	.000	.000	.000	.000	.000	.000	.000
1	.366	.343	.132	.067	.031	.005	.000	.000	.000	.000	.000	.000	.000
2	.135	.267	.231	.156	.092	.022	.003	.000	.000	.000	.000	.000	.000
3	.031	.129	.250	.225	.170	.063	.014	.002	.000	.000	.000	.000	.000
4	.005	.043	.188	.225	.219	.127	.042	.007	.001	.000	.000	.000	.000
5	.001	.010	.103	.165	.206	.186	.092	.024	.003	.001	.000	.000	.000
6	.000	.002	.043	.092	.147	.207	.153	.061	.012	.003	.001	.000	.000
7	.000	.000	.014	.039	.081	.177	.196	.118	.035	.013	.003	.000	.000
8	.000	.000	.003	.013	.035	.118	.196	.177	.081	.039	.014	.000	.000
9	.000	.000	.001	.003	.012	.061	.153	.207	.147	.092	.043	.002	.000
10	.000	.000	.000	.001	.003	.024	.092	.186	.206	.165	.103	.010	.001
11	.000	.000	.000	.000	.001	.007	.042	.127	.219	.225	.188	.043	.005
12	.000	.000	.000	.000	.000	.002	.014	.063	.170	.225	.250	.129	.031
13	.000	.000	.000	.000	.000	.000	.003	.022	.092	.156	.231	.267	.135
14	.000	.000	.000	.000	.000	.000	.000	.005	.031	.067	.132	.343	.366
15	.000	.000	.000	.000	.000	.000	.000	.000	.005	.013	.035	.206	.463
<i>n = 20</i>													
<i>x</i>	<i>p</i>												
	0.05	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	0.95
0	.358	.122	.012	.003	.001	.000	.000	.000	.000	.000	.000	.000	.000
1	.377	.270	.058	.021	.007	.000	.000	.000	.000	.000	.000	.000	.000
2	.189	.285	.137	.067	.028	.003	.000	.000	.000	.000	.000	.000	.000
3	.060	.190	.205	.134	.072	.012	.001	.000	.000	.000	.000	.000	.000
4	.013	.090	.218	.190	.130	.035	.005	.000	.000	.000	.000	.000	.000
5	.002	.032	.175	.202	.179	.075	.015	.001	.000	.000	.000	.000	.000
6	.000	.009	.109	.169	.192	.124	.037	.005	.000	.000	.000	.000	.000
7	.000	.002	.055	.112	.164	.166	.074	.015	.001	.000	.000	.000	.000
8	.000	.000	.022	.061	.114	.180	.120	.035	.004	.001	.000	.000	.000
9	.000	.000	.007	.027	.065	.160	.160	.071	.012	.003	.000	.000	.000
10	.000	.000	.002	.010	.031	.117	.176	.117	.031	.010	.002	.000	.000
11	.000	.000	.000	.003	.012	.071	.160	.160	.065	.027	.007	.000	.000
12	.000	.000	.000	.001	.004	.035	.120	.180	.114	.061	.022	.000	.000
13	.000	.000	.000	.000	.001	.015	.074	.166	.164	.112	.055	.002	.000
14	.000	.000	.000	.000	.000	.005	.037	.124	.192	.169	.109	.009	.000
15	.000	.000	.000	.000	.000	.001	.015	.075	.179	.202	.175	.032	.002
16	.000	.000	.000	.000	.000	.000	.005	.035	.130	.190	.218	.090	.013
17	.000	.000	.000	.000	.000	.000	.001	.012	.072	.134	.205	.190	.060
18	.000	.000	.000	.000	.000	.000	.000	.003	.028	.067	.137	.285	.189
19	.000	.000	.000	.000	.000	.000	.000	.000	.007	.021	.058	.270	.377
20	.000	.000	.000	.000	.000	.000	.000	.000	.001	.003	.012	.122	.358

(Continued)

TABLE 9 Binomial Probabilities (*Continued*)

<i>n</i> = 25													
<i>x</i>	<i>p</i>												
	0.05	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	0.95
0	.277	.072	.004	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
1	.365	.199	.024	.006	.001	.000	.000	.000	.000	.000	.000	.000	.000
2	.231	.266	.071	.025	.007	.000	.000	.000	.000	.000	.000	.000	.000
3	.093	.226	.136	.064	.024	.002	.000	.000	.000	.000	.000	.000	.000
4	.027	.138	.187	.118	.057	.007	.000	.000	.000	.000	.000	.000	.000
5	.006	.065	.196	.165	.103	.020	.002	.000	.000	.000	.000	.000	.000
6	.001	.024	.163	.183	.147	.044	.005	.000	.000	.000	.000	.000	.000
7	.000	.007	.111	.165	.171	.080	.014	.001	.000	.000	.000	.000	.000
8	.000	.002	.062	.124	.165	.120	.032	.003	.000	.000	.000	.000	.000
9	.000	.000	.029	.078	.134	.151	.061	.009	.000	.000	.000	.000	.000
10	.000	.000	.012	.042	.092	.161	.097	.021	.001	.000	.000	.000	.000
11	.000	.000	.004	.019	.054	.147	.133	.043	.004	.001	.000	.000	.000
12	.000	.000	.001	.007	.027	.114	.155	.076	.011	.002	.000	.000	.000
13	.000	.000	.000	.002	.011	.076	.155	.114	.027	.007	.001	.000	.000
14	.000	.000	.000	.001	.004	.043	.133	.147	.054	.019	.004	.000	.000
15	.000	.000	.000	.000	.001	.021	.097	.161	.092	.042	.012	.000	.000
16	.000	.000	.000	.000	.000	.009	.061	.151	.134	.078	.029	.000	.000
17	.000	.000	.000	.000	.000	.003	.032	.120	.165	.124	.062	.002	.000
18	.000	.000	.000	.000	.000	.001	.014	.080	.171	.165	.111	.007	.000
19	.000	.000	.000	.000	.000	.000	.005	.044	.147	.183	.163	.024	.001
20	.000	.000	.000	.000	.000	.000	.002	.020	.103	.165	.196	.065	.006
21	.000	.000	.000	.000	.000	.000	.000	.007	.057	.118	.187	.138	.027
22	.000	.000	.000	.000	.000	.000	.000	.002	.024	.064	.136	.226	.093
23	.000	.000	.000	.000	.000	.000	.000	.000	.007	.025	.071	.266	.231
24	.000	.000	.000	.000	.000	.000	.000	.000	.002	.006	.024	.199	.365
25	.000	.000	.000	.000	.000	.000	.000	.000	.000	.001	.004	.072	.277

Answers to Selected Odd-Numbered Exercises

Chapter 1

1.1 *Descriptive statistics* is the branch of statistics that involves the organization and summary of the values in a data set. *Inferential statistics* is the branch of statistics concerned with reaching conclusions about a population based on the information provided by a sample.

1.3 Sample

1.5 **a.** The population of interest is the set of all 15,000 students at the university. **b.** The sample is the 200 students who are interviewed.

1.7 The population is the set of all 7000 property owners. The sample is the group of 500 owners included in the survey.

1.9 The population is the set of 5000 used bricks. The sample is the set of 100 bricks she checks.

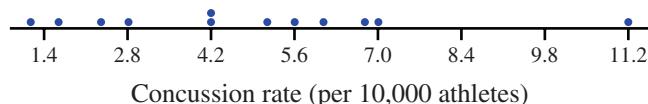
1.11 **a.** The researchers wanted to know if taking a garlic supplement reduces the risk of getting a cold. **b.** We would want to know how many people participated in the study and how the people were assigned to the two groups.

1.13 **a.** Categorical **b.** Categorical **c.** Numerical (discrete) **d.** Numerical (continuous) **e.** Categorical **f.** Numerical (continuous)

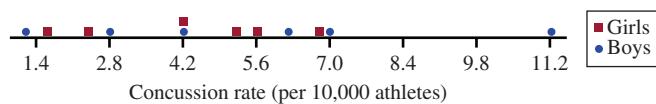
1.15 **a.** Continuous **b.** Continuous **c.** Continuous **d.** Discrete

1.17 **a.** Gender of purchaser, brand of motorcycle, telephone area code **b.** Number of previous motorcycles **c.** Bar chart **d.** Dotplot

1.19 **a.**

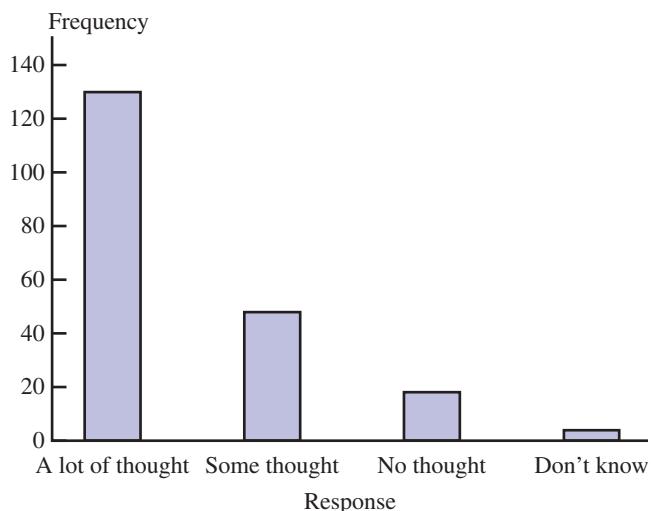


b.



The sport with an unusually high (when compared to all the other sports) concussion rate is football. Without considering football, the concussion rates for girls' sports is essentially the same as the concussion rate for boys' sports.

1.21 **a.**



b. The most common response was “A lot of thought,” accounting for 130 (or 65%) of the students who started college but did not complete a degree. The next two most common responses were “Some thought” and “No thought,” accounting for 48 (or 24%) and 18 (or 9%), respectively, of the students who started college but did not complete a degree. Finally, 4 of the 200 respondents (2%) indicated that don’t know how much thought they have given to going back to school.

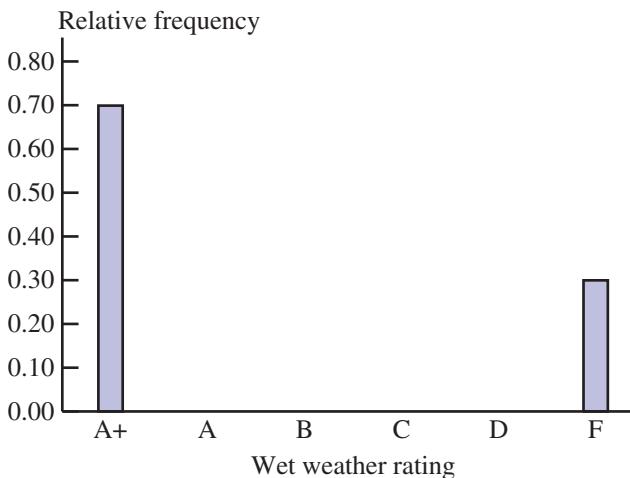
1.23 **a.** There were three sites that received far greater numbers of visits than the remaining six sites. Also, the distribution of the number of visits has the greatest density of points for the smaller numbers of visits, with the density decreasing as the number of visits increases.

b. There were two sites that were used by far greater numbers of individuals (unique visitors) than the remaining seven sites. However, these two sites are used less than the others in terms of the number of unique visitors than they are in terms of the total number of visits. The distribution of the number of unique visitors has the greatest density of points for the smaller numbers of visitors, with the density decreasing as the number of unique visitors increases.

c. The statistic “visits per unique visitor” tells us how heavily the individuals are using the sites. The table tells us that the most popular sites (Facebook and YouTube) in terms of total visits and unique visitors do not have the highest value of this statistic. The dotplot of visits per

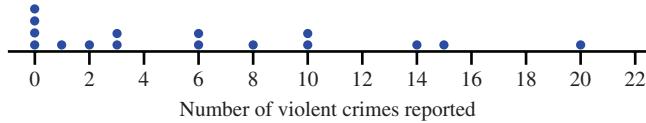
unique visitor shows that there are two individual sites that are far ahead of the rest in this respect (Pinterest and Twitter).

1.25 a.



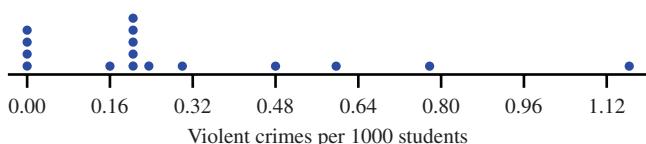
b. Seventy-five percent (75%) of the dry weather ratings are B or higher, and 70% of wet weather ratings are B or higher, indicating that dry weather ratings are higher than wet weather ratings. Note that the wet weather ratings are only A+ or F, so the wet weather ratings are more extreme than dry weather ratings.

1.27 a.



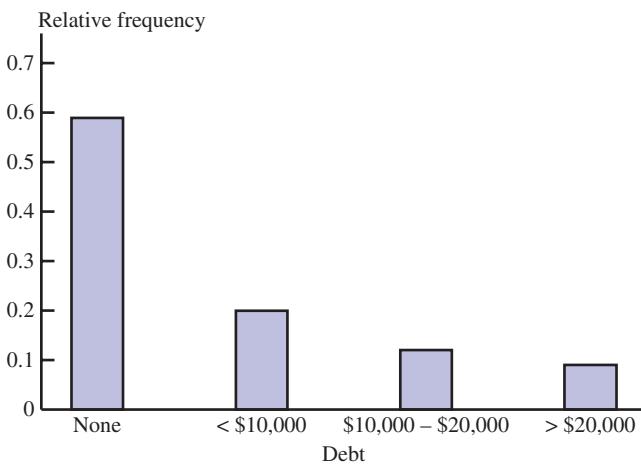
Three schools seem to stand out from the rest (in increasing order of number of crimes): University of Central Florida (14 crimes reported), Florida International University (15 crimes reported), and Florida State University (20 crimes reported).

b.



The colleges that stand out in violent crimes per 1000 students are (in increasing order of crime rate) Florida State University, Florida A&M University, University of West Florida, and New College of Florida. Only Florida State University stands out in both dotplots. **c.** For the number of violent crimes, there are three schools that stand out by having high numbers of crimes, with the majority of the schools having similar, and low (10 or fewer), numbers of crimes. There seems to be greater consistency for crime rate (per 1000 students) among the 16 schools than there is for number of crimes, with four schools standing out as having high crime rates, and four schools with crime rates that stand out as being noticeably low.

1.29 a.

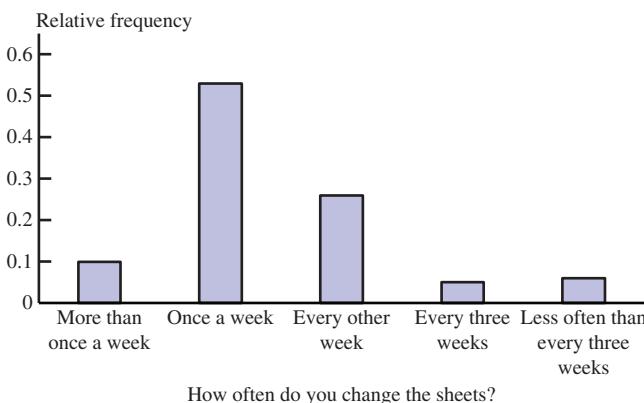


b. Most public community college graduates have no debt at all, and a debt of \$10,000 or less accounts for 79% of the graduates. Among the 21% of the graduates who have a debt of more than \$10,000, nearly 43% (9% of all graduates) have a debt of more than \$20,000.

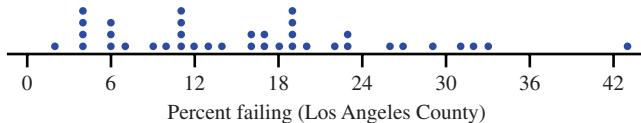
1.31 a.

How Often	Relative Frequency
More than once a week	0.10
Once a week	0.53
Every other week	0.26
Every three weeks	0.05
Less often than every three weeks	0.06

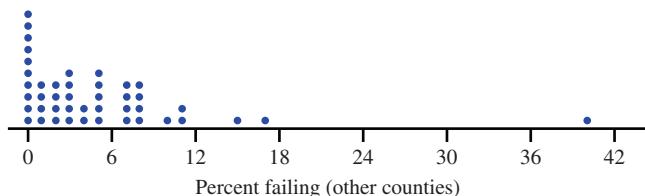
b.



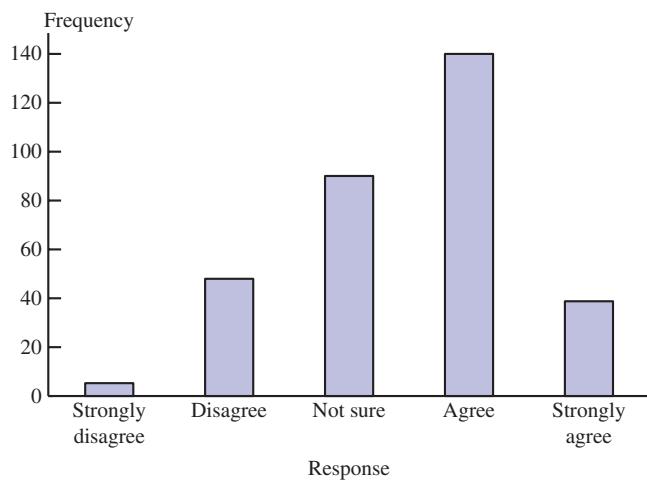
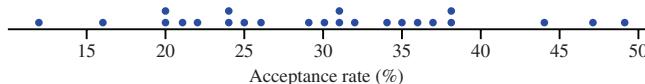
1.33 a.



A typical percent of tests failing for Los Angeles County is around 16. There is one value that is unusually high (43), with the other values ranging from 2 to 33. There is a greater density of points toward the lower end of the distribution than toward the upper end.

b.

A typical percent of tests failing for the other counties is around 3. There is one extreme result at the upper end of the distribution (40); the other values range from 0 to 17. The density of points is highest at the left hand end of the distribution and decreases as the percent failing values increase. **c.** The typical value for Los Angeles County (around 16) is greater than for the other counties (around 3) and, disregarding the one extreme value in each case, there is a greater variability in the values for Los Angeles County than for the other counties. In the distribution for Los Angeles County the points are closer to being uniformly distributed than in the distribution for the other counties.

1.35**1.37 a.**

b. A typical acceptance rate for these top 25 schools is around 30, with the great majority of acceptance rates being between 19 and 39. There are no particularly extreme values. The pattern of the points is roughly symmetrical.

Chapter 2

2.1 Observational study, because the researchers did not assign participants to the length of stay groups.

2.3 Experiment, because the professor determined which students were in each of the two experimental groups.

2.5 a. Experiment, because researchers decided which participants would receive which treatments. **b.** Yes, because the participants were randomly assigned to the treatments.

2.7 a. Experiment. **b.** Yes, because the participants were randomly assigned to the treatments.

2.9 We are told that moderate drinkers, as a group, tended to be better educated, wealthier, and more active than nondrinkers. It is possible the observed reduction in the risk of heart disease among moderate drinkers is caused by one of these attributes and not by the moderate drinking.

2.11 a. The data would need to be collected from a simple random sample of all adult American Internet users. **b.** No

2.13 Number the names on the list from 1 to n , where n is the number of students at the college. Then use a random number generator to select 100 different numbers between 1 and n . Students corresponding to these 100 numbers would be included in the sample.

2.15 Number names on list, use random number generator to select 30 numbers between 1 and 500. Signatures corresponding to the 30 selected numbers constitute the random sample.

2.17 a. All American women **b.** No, only women for three states were included in the sample. **c.** No **d.** The description does not say how the women were selected, so it is difficult to tell. However, selection bias is present because women from most states had no chance of being included in the sample.

2.19 a. Using the list, first number the part-time students 1–3000. Use a random number generator on a calculator or computer to randomly select a whole number between 1 and 3000. The number selected represents the first part-time student to be included in the sample. Repeat the number selection, ignoring repeated numbers, until 10 part-time students have been selected. Then number the full-time students 1–3500 and select 10 full-time students using the same procedure. **b.** No

2.21 a. The pages of the book have already been numbered between 1 and the highest page number in the book. Use a random number generator on a calculator or computer to randomly select a whole number between 1 and the highest page number in the book. The number selected will be the first page to be included in the sample. Repeat the number selection, ignoring repeated numbers, until the required number of pages has been selected. **b.** Pages that include exercises tend to contain more words than pages that do not include exercises. Therefore, it would be sensible to stratify according to this criterion. Assuming that 20 nonexercise pages and 20 exercise pages will be included in the sample, the sample should be selected as follows: Use a random number generator to randomly select a whole number between 1 and the highest page number in the book; the number selected will be the first page to be included in the sample; repeat the number selection, ignoring repeated numbers and keeping track of the number of pages of each type selected, until 20 pages of one type have been selected; then continue in the same way, but ignore numbers corresponding to pages of that

type; when 20 pages of the other type have been selected, stop the process. **c.** Randomly select one page from the first 20 pages in the book. Include in your sample that page and every 20th page from that page onward. **d.** Roughly speaking, in terms of the numbers of words per page, each chapter is representative of the book as a whole. It is therefore sensible for the chapters to be used as clusters. Using a random number generator, randomly choose three chapters. Then count the number of words on each page in those three chapters.

2.23 Answers will vary.

2.25 The researchers should be concerned about nonresponse bias.

2.27 It is not reasonable to consider the participants to be representative of all students with regard to their truthfulness in the various forms of communication. Also, the students knew they were surveying themselves as to the truthfulness of their interactions. This could easily have changed their behavior in particular social contexts and, therefore, could have distorted the results of the study.

2.29 It is quite possible that people who read that newspaper or access this web site differ from the population in some relevant way, particularly considering that they are both New York City-based publications.

2.31 **a.** Yes. It is possible that students of a given class-standing tend to be similar in the amounts of money they spend on textbooks. **b.** Yes. It is possible that students who pursue a certain field of study tend to be similar in the amounts of money they spend on textbooks. **c.** No. It is unlikely that stratifying in this way will produce groups that are homogeneous in terms of the students' spending on textbooks.

2.33 No, because the people who sent their hair for testing did so voluntarily. It is quite possible that people who would choose to participate in a study of this sort differ in their mercury levels from the population as a whole.

2.35 **a.** Binding strength **b.** Type of glue **c.** The extraneous variables mentioned are the number of pages in the book and whether the book is bound as a hardback or a paperback. Further extraneous variables that might be considered include the weight of the material used for the cover and the type of paper used.

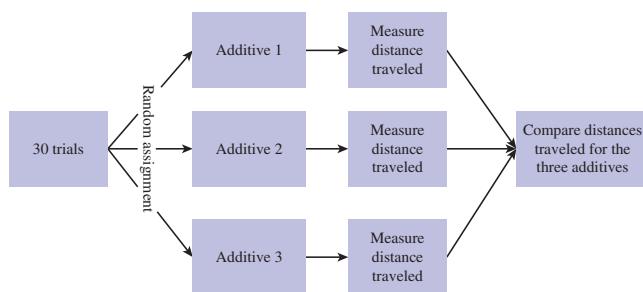
2.37 Answers will vary. One possible answer: Write the names of the 30 subjects on slips of paper. Mix the slips and then select 10. Assign these 10 subjects to hand-drying method 1. Mix the remaining slips and then select 10. Assign these 10 subjects to hand-drying method 2. Assign the remaining 10 subjects to method 3.

2.39 **a.** Blocking **b.** Direct control

2.41 The figure shows that comparable groups in terms of age have been formed.

2.43 We rely on random assignment to produce comparable experimental groups.

2.45



2.47 **a.** Experiment **b.** No **c.** Yes **d.** Yes, because the experiment used random assignment of subjects to treatments. **e.** No, because the subjects were not randomly selected.

2.49 **a.** The treatments are the names—Ann Clark and Andrew Clark—given to the participants. **b.** The response variables are the participants' answers to the questions given.

2.51 **a.** Red wine, yellow onions, black tea **b.** Absorption of flavonol into the blood **c.** Gender, amount of flavonols consumed apart from experimental treatment, tolerance of alcohol in wine

2.53 “Blinding” is ensuring that the experimental subjects do not know which treatment they were given and/or ensuring that the people who measure the response variable do not know who was given which treatment.

2.55 **a.** Allowing study participants to choose which group they want to be in could introduce systematic differences between the two experimental conditions (knee replacement surgery with exercise and exercise therapy alone), resulting in potential confounding. Those who chose knee replacement surgery plus exercise might, in some way, be different from those who chose exercise therapy alone. We would not know if differences in pain relief between the two groups were due to the knee replacement surgery with exercise or due to some inherent differences in the subjects who chose their experimental groups. **b.** The researchers likely did not include a control group because the study participants needed some relief from their pain. Because the purpose of this experiment is to determine whether knee replacement surgery with exercise provided more pain relief than exercise therapy alone, a control group would not allow the study participants to have the opportunity to experience pain relief.

2.57 We will assume that only four colors will be compared, and that only headache sufferers will be included in the study. Prepare a supply of “Regular Strength” Tylenol in four different colors: white (the current color of the medication, and therefore the “control”), red, green, and blue. Recruit 20 volunteers who suffer from headaches. Instruct each volunteer not to take any pain relief medication for a week. After that week is over, issue each volunteer a supply of all four colors. Give each volunteer an order in which to use the colors (this order would be determined randomly for each volunteer). Instruct the volunteers

to use one fixed dose of the medication for each headache over a period of 4 weeks, and to note on a form the color used and the pain relief achieved (on a scale of 0-10, where 0 is no pain relief and 10 is complete pain relief). At the end of the 4 weeks gather the results and compare the pain relief achieved by the four colors.

2.59 Answers will vary. One possible answer: Number the girls from 1 to 700 and then use a random number generator to select 350 of these girls for the book group. Assign the remaining girls to the other group. Then number the boys from 1 to 600 and use a random number generator to select 300 of the boys for the book group. Assign the remaining boys to the other group.

2.61 **a.** If the judges had known which chowder came from which restaurant, then it is unlikely that Denny's chowder would have won the contest, since the judges would probably be conditioned by this knowledge to choose chowders from more expensive restaurants. **b.** In experiments, if the people measuring the response are not blinded, they will often be conditioned to see different responses to some treatments over other treatments, in the same way as the judges would have been conditioned to favor the expensive restaurant chowders. Therefore, it is necessary that the people measuring the response should not know which subject received which treatment, so that the treatments can be compared on their own merits.

2.63 **a.** A placebo group would be necessary if the mere thought of having amalgam fillings could produce kidney disorders. However, since the experimental subjects were sheep, the researchers do not need to be concerned that this would happen. **b.** A resin filling treatment group would be necessary in order to provide evidence that it is the material in the amalgam fillings, rather than the process of filling the teeth, or just the presence of foreign bodies in the teeth, that is the cause of the kidney disorders. If the amalgam filling group developed the kidney disorders and the resin filling group did not, then this would provide evidence that it is some ingredient in the amalgam fillings that is causing the kidney problems. **c.** Since there is concern about the effect of amalgam fillings, it would be considered unethical to use humans in the experiment.

2.65 **a.** This is an observational study. **b.** In order to evaluate the study, we need to know whether the sample was a random sample.

2.67 Answers will vary.

2.69 Answers will vary.

2.71 By randomly selecting the phone numbers, calling back those for which there are no answers, and asking for the adult in the household with the most recent birthday, the researchers are avoiding selection bias. However, selection bias could result from the fact that not all Californians have phones.

2.73 We rely on random assignment to produce comparable experimental groups. If the researchers had hand-picked the treatment groups, they might unconsciously

have favored one group over the other in terms of some variable that affects the ability of the people at the centers to respond to the materials provided.

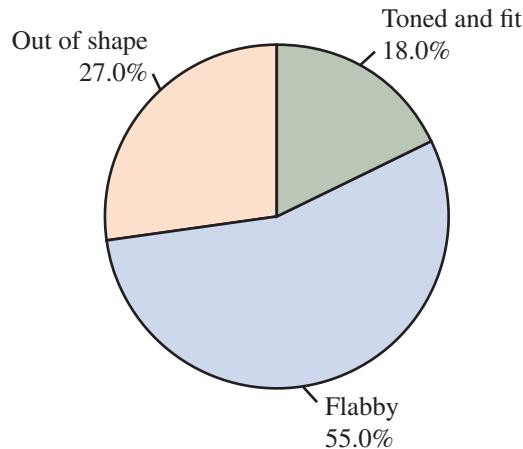
2.75 This is an observational study based on results of a survey (no consumers were assigned to different experimental conditions).

2.77 **a.** The design is good in that it includes several desired experimental design components. Specifically, the design includes a control group, random assignment of the subjects to the treatments, and includes blinding.

b. Rather than taking photos of the tops of the heads of all the women, the expert who determined the change in hair density should have the opportunity to evaluate the women in person. Additionally, there could have been more than one expert doing the change in hair density evaluation.

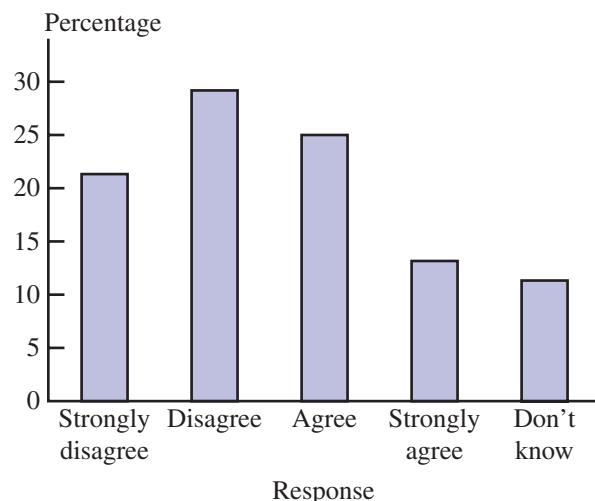
Chapter 3

3.1



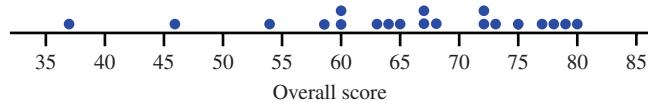
3.3 **a.** The second and third categories ("Permitted for business purposes only" and "Permitted for limited personal use") were combined into one category ("No, but some limits apply"). **c.** Pie chart, regular bar graph

3.5 **a.**



b. Answers will vary. One possible answer: Majority of students not ready for ebooks.

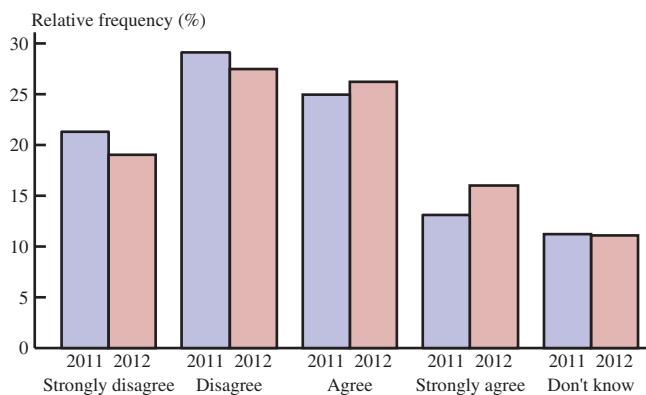
3.7



One possibility would be to require 72 or higher for “top of the class,” 63 or higher for “passing,” 59 or higher for “barely passing,” and below 59 for “failing.” This makes a clear separation between the jurisdictions in the four grade categories. (Since this makes it easier to be rated as “top of the class,” it could be considered more appropriate to leave the grade boundaries as they are.)

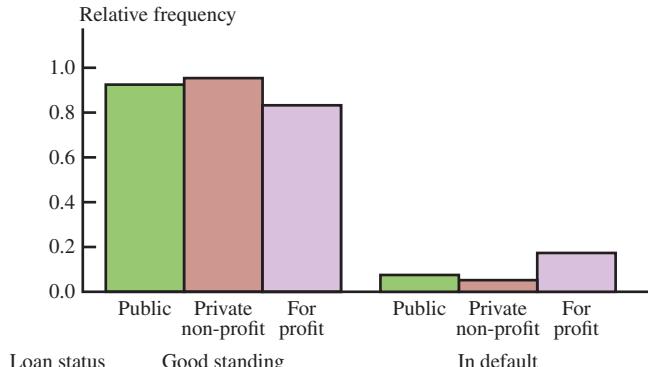
3.9 a. Were the surveys carried out on random samples of married women from those countries? How were the questions worded? **b.** In one country, Japan, the percentage of women who say they never get help from their husbands is far higher than the percentages in any of the other four countries included. The percentages in the other four countries are similar, with Canada showing the lowest percentage of women who say they do not get help from their husbands.

3.11 a.



b. The distribution of responses was very similar for the 2 years. However, there was a small shift toward wanting textbooks in digital form from 2011 to 2012.

3.13 a.



b. The bar for “in default” for the for-profit colleges is higher than the “in default” bars for the other two types of colleges.

3.15

9	345
10	0457
11	01456889
12	00022333334445666677899
13	124689
14	12479
15	3
16	9

Stem: Ones
Leaf: Tenth

A typical number of births per thousand of the population is around 12.4, with most birth rates concentrated in the 11.0 to 13.9 range. The distribution has just one peak (at the 12–13 class). There are no extreme values. The distribution is approximately symmetrical.

3.17 a.

6H	8
7L	0044
7H	67889
8L	01233444
8H	5555667777788
9L	0000012223334444
9H	5777

Stem: Tens
Leaf: Ones

b. There is state-to-state variability in percent of drivers wearing seat belts, and seat belt use percentages tend to be relatively high. A typical seat belt use percentage is around 86 or 87%.

3.19 a.

0H	55567889999
1L	0000111113334
1H	556666666667789
2L	00001122233
2H	5

Stem: Tens
Leaf: Ones

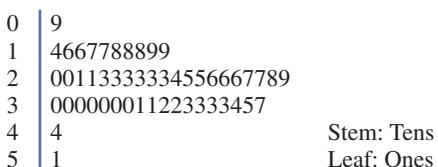
A typical percentage of households with only a wireless phone is around 15.

b.

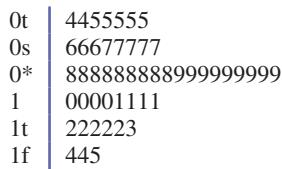
West	998	OH	555789
	110	1L	00011134
	8766	1H	666
	21	2L	00
	5	2H	

Stem: Tens
Leaf: Ones

A typical percentage of households with only a wireless phone for the West is around 16, which is greater than that for the East (around 11). There is a slightly greater spread of values in the West than in the East, with values in the West ranging from 8 to 25 (a range of 17) and values in the East ranging from 5 to 20 (a range of 15). The distribution for the West is roughly symmetrical, while the distribution in the East shows a slightly greater spread to the right of its center than to the left. Neither distribution has any outliers.

3.21 a.

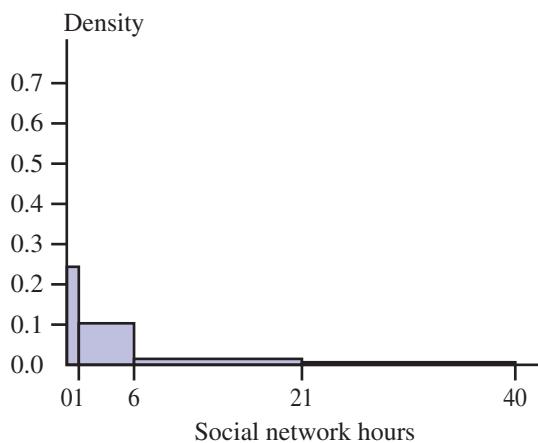
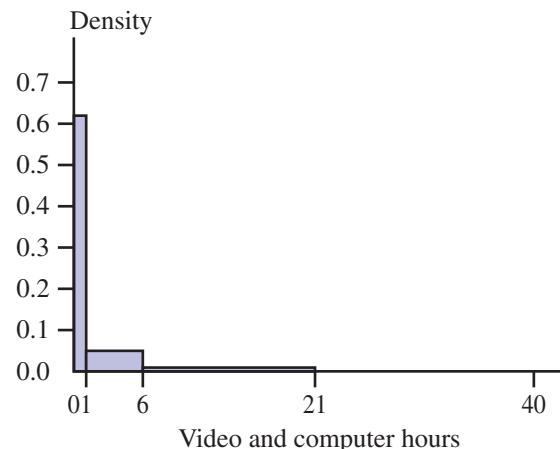
b. The center is approximately 26 cents per gallon, and most states have a tax that is near the center value, with tax values ranging from 9 cents per gallon to 51 cents per gallon. The distribution is approximately symmetric. c. No. The only value that might be considered unusual is the 51 cents per gallon tax in Pennsylvania. Although not an obvious outlier, it is approximately 7 cents per gallon higher than the next lower gasoline tax. There are no other states that make such a big jump to the next higher tax. The lowest gasoline tax is 9 cents per gallon (Alaska) and the highest is 51 cents per gallon (Pennsylvania).

3.23

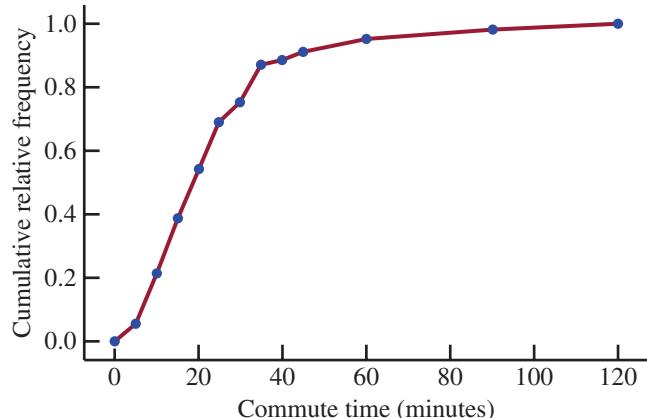
The stem-and-leaf display shows that the distribution of high school dropout rates is roughly symmetrical. A typical dropout rate is 7%. The great majority of rates are between 4% and 9%, inclusive.

3.25 The distribution of maximum wind speeds is positively skewed and is bimodal, with peaks at the 35–40 and 60–65 intervals.

3.27 b. The typical percentage of workers belonging to a union is around 11, with values ranging from 3.5 to 24.9. There are three states with percentages that stand out as being higher than those of the rest of the states. The distribution is positively skewed. c. The dotplot is more informative as it shows where the data points actually lie. For example, in the histogram we can tell that there are three observations in the 20 to 25 interval, but we don't see the actual values and miss the fact that these values are actually considerably higher than the other values in the data set.

3.29 a.**b.**

c. Both histograms are positively skewed, but a typical value for social networking hours is greater than the typical value for video and computer hours.

3.31 a.

b. i. 0.93 ii. $1 - 0.62 = 0.38$ iii. approximately 19 minutes

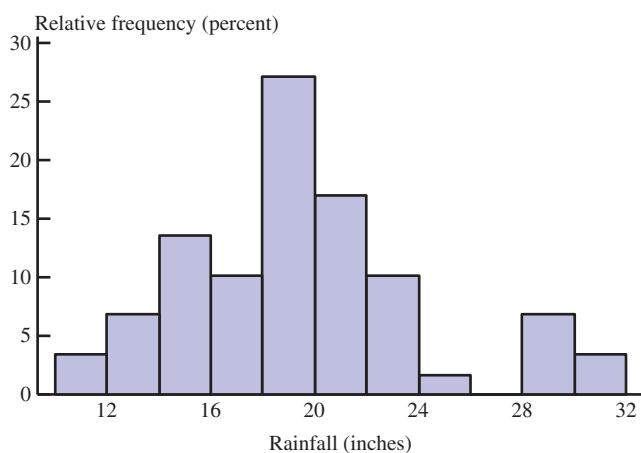
3.33 a. The distribution is likely to be negatively skewed. There will be a high density of scores close to 100, with no score greater than 100, and the density of scores tailing off to the left. b. So long as there is no cutoff at the lower end, the distribution is likely to be roughly symmetrical. For example, the greatest density of points might be around 65, with the density tailing off equally to the left and to the right of that value. c. In this case a bimodal distribution would be likely. There would be a clustering of points around, say, 65 for the students with less math knowledge and another clustering around, say, 95, for those who have had calculus.

3.35 a.

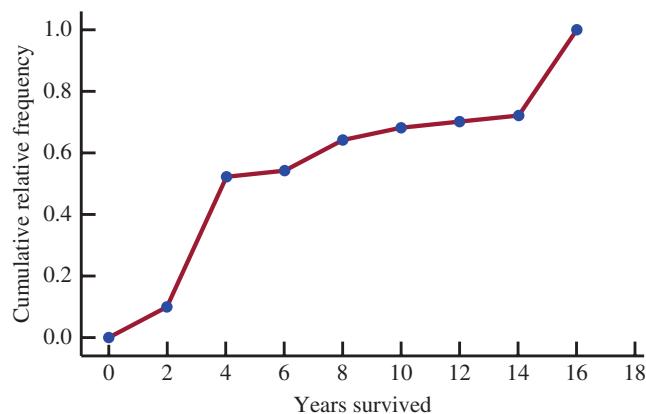
Rainfall (inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
10 to <12	2	0.034	0.034
12 to <14	4	0.068	0.102
14 to <16	8	0.136	0.237
16 to <18	6	0.102	0.339
18 to <20	16	0.271	0.610

(continued)

Rainfall (inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
20 to <22	10	0.169	0.780
22 to <24	6	0.102	0.881
24 to <26	1	0.017	0.898
26 to <28	0	0.000	0.898
28 to <30	4	0.068	0.966
30 to <32	2	0.034	1.000

b.

The histogram shows a distribution that is slightly positively skewed with a trough between 24 and 28.

3.37 a.

b. i. 0.53 ii. 0.62 iii. $1 - 0.68 = 0.32$

3.39 Answers will vary. One possibility for each part is shown below.

a.

Class Interval	100 to <120	120 to <140	140 to <160	160 to <180	180 to <200
Frequency	5	10	40	10	5

b.

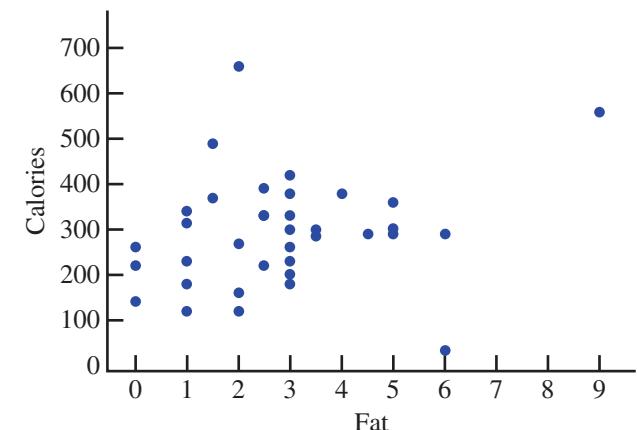
Class Interval	100 to <120	120 to <140	140 to <160	160 to <180	180 to <200
Frequency	20	10	4	25	11

c.

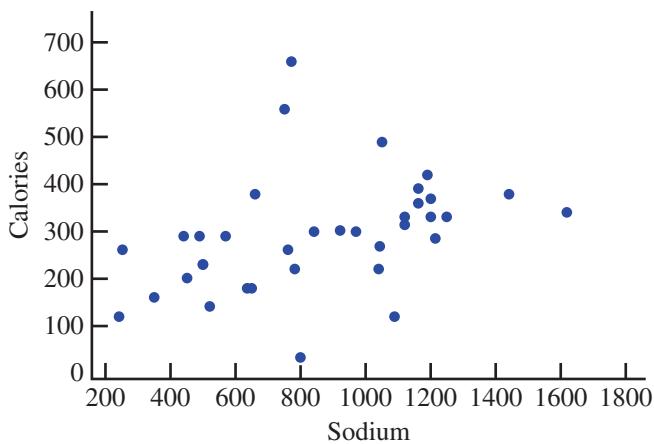
Class Interval	100 to <120	120 to <140	140 to <160	160 to <180	180 to <200
Frequency	33	15	10	7	5

d.

Class Interval	100 to <120	120 to <140	140 to <160	160 to <180	180 to <200
Frequency	5	7	10	15	33

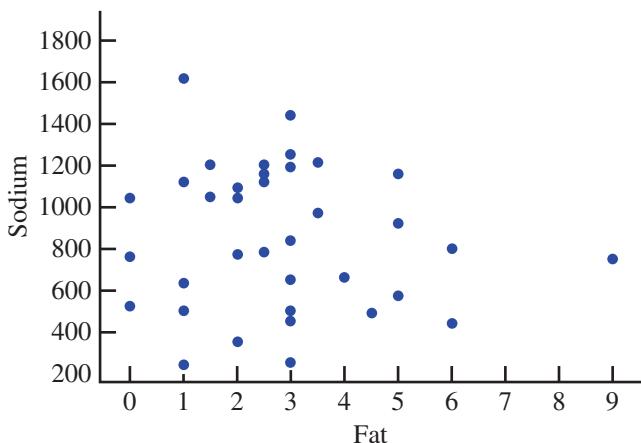
3.41 a.

There does appear to be a relationship between fat and calories. As fat increases, calories also tends to increase, which is not surprising.

b.

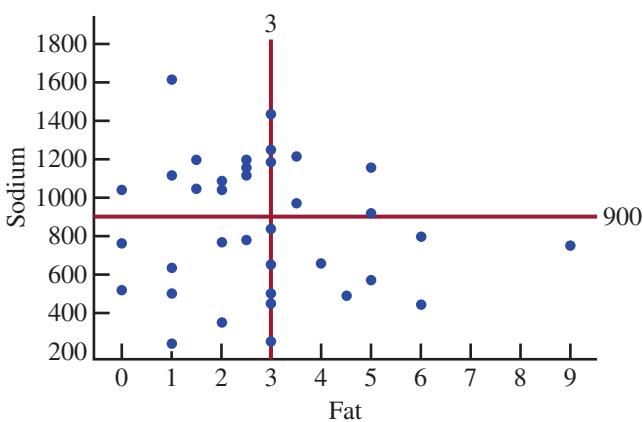
The relationship between sodium and calories is not strong, but it does appear that fast-food items with more sodium also tend to have more calories.

c.



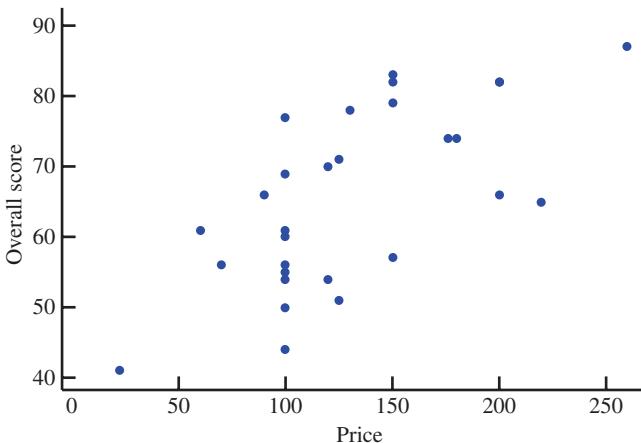
There does not appear to be a relationship between fat and sodium.

d.



The region in the lower left hand corner corresponds to the healthier fast-food choices. These are the choices that are lower in both sodium and fat.

3.43 a.



b. There appears to be a weak positive relationship between price and overall score. The scatterplot only weakly supports the statement.

3.45 The time series plot is consistent with the statement of having seen steady growth in recycling and composting because those trends are both increasing. However, the statement about the amounts by which landfills have generally declined might be somewhat misleading, at least when the conclusion is based on the graph shown, with the given scale. It might be more appropriate to state that the amounts of landfills have remained roughly constant, with a slight decrease in the given time period.

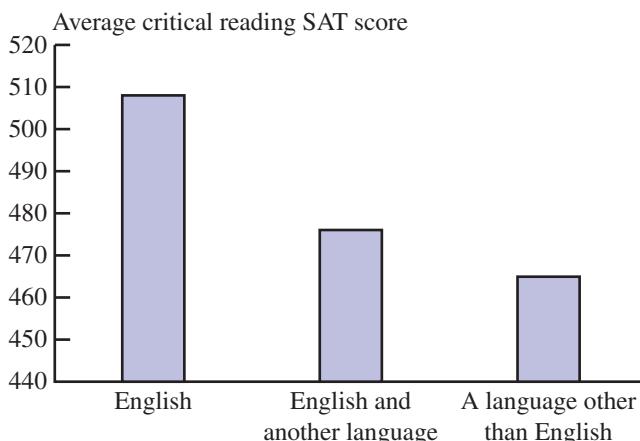
3.47 According to the 2001 and 2002 data, there are seasonal peaks at weeks 4, 9, and 14, and seasonal lows at weeks 2, 6, 10–12, and 18.

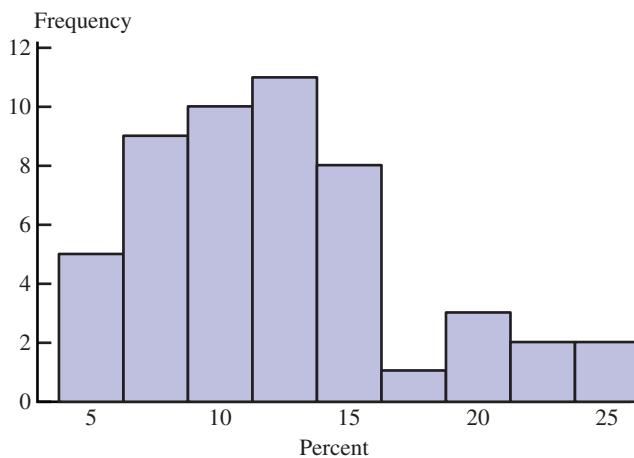
3.49 b. The graphical display created in Part (a) is more informative, since it gives an accurate representation of the proportions of the ethnic groups. **c.** The people who designed the original display possibly felt that the four ethnic groups shown in the segmented bar section might seem to be underrepresented at the college if they used a single pie chart.

3.51 The first graphical display is not drawn appropriately. The Z's have been drawn so that their heights are in proportion to the percentages shown. However, the widths and the perceived depths are also in proportion to the percentages, and so neither the areas nor the perceived volumes of the Z's are proportional to the percentages. The graph is therefore misleading to the reader. In the second graphical display, however, *only* the heights of the cars are in proportion to the percentages shown. The widths of the cars are all equal. Therefore the areas of the cars are in proportion to the percentages, and this is an appropriately drawn graphical display.

3.53 The piles of cocaine have been drawn so that their heights are in proportion to the percentages shown. However, the widths are also in proportion to the percentages, and therefore neither the areas (nor the perceived volumes) are in proportion to the percentages. The graph is therefore misleading to the reader.

3.55



3.57

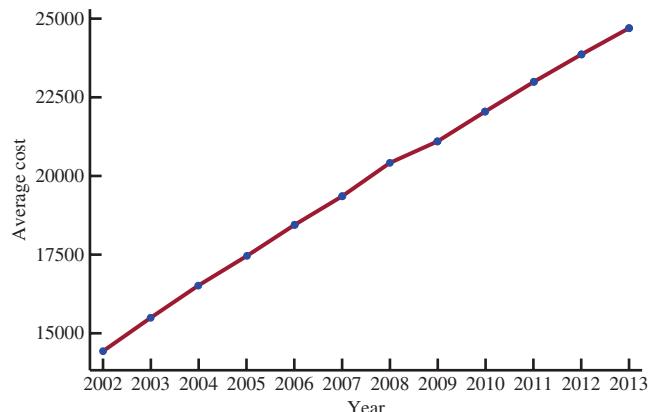
- 3.59** **a.** High graft weight ratios are clearly associated with low body weights (and vice versa), and the relationship is not linear. (In fact, roughly speaking, there seems to be an inverse proportionality between the two variables, apart from a small increase in the graft weight ratios for increasing body weights among those recipients with the greater body weights. This is interesting in that an inverse proportionality between the variables would imply that the actual weights of transplanted livers are chosen independently of the recipients' body weights.) **b.** A likely reason for the negative relationship is that the livers to be transplanted are probably chosen according to whatever happens to be available at the time. Therefore, lighter patients are likely to receive livers that are too large and heavier patients are likely to receive livers that are too small.

3.61 **a.**

Disney		Other
975332100	0	0001259
765	1	156
920	2	0
	3	
	4	
4	5	

Stem: Hundreds
Leaf: Tens

- b.** On average, the total tobacco exposure times for the Disney movies are higher than the others, with a typical value for Disney of about 90 seconds and a typical value for the other companies of about 50 seconds. Both distributions have one peak and are positively skewed. There is one extreme value (548) in the Disney data and no extreme value in the data for the other companies. There is a greater spread in the Disney data, with values ranging from 6 seconds to 540 seconds, than for the other companies, with values ranging from 1 second to 205 seconds.

3.63

There is a strong, positive trend in the average cost of per year for tuition, fees, and room and board for 4-year public institutions in the United States. The average cost has steadily increased from a low of \$14,439 per year in 2002 to a high of \$24,706 per year in 2013.

3.65 **a.**

Skeletal Retention	Frequency
0.15 to <0.20	4
0.20 to <0.25	2
0.25 to <0.30	5
0.30 to <0.35	21
0.35 to <0.40	9
0.40 to <0.45	9
0.45 to <0.50	4
0.50 to <0.55	0
0.55 to <0.60	1

- b.** The histogram is centered at approximately 0.34, with values ranging from 0.15 to 0.5, plus one extreme value in the 0.55–0.6 range. The distribution has a single peak and is slightly positively skewed.

CUMULATIVE REVIEW EXERCISES 3

CR3.1 No. For example, it is quite possible that men who ate a high proportion of cruciferous vegetables generally speaking also had healthier lifestyles than those who did not, and that it was the healthier lifestyles that were causing the lower incidence of prostate cancer, not the eating of cruciferous vegetables.

CR3.3 Very often those who choose to respond generally have a different opinion on the subject of the study from those who do not respond. (In particular, those who respond often have strong feelings against the status quo.) This can lead to results that are not representative of the population that is being studied.

CR3.5 Only a small proportion (around 11%) of the doctors responded, and it is quite possible that those who did respond had different opinions regarding managed care than the majority who did not. Therefore, the results could

have been very inaccurate for the population of doctors in California.

CR3.7 For example, suppose the women had been allowed to choose whether or not they participated in the program. Then it is quite possible, generally speaking, that those women with more social awareness would have chosen to participate, and those with less social awareness would have chosen not to. Then it would be impossible to tell whether the stated results came about as a result of the program or of the greater social awareness among the women who participated. By randomly assigning the women to participate or not, comparable groups of women would have been obtained.

CR3.9 **b.** Between 2002 and 2003 and between 2003 and 2004, the pass rates rose for both the high school and the state, with a particularly sharp rise between 2003 and 2004 for the state. However, the pass rate for the county fell between 2002 and 2003 and then rose between 2003 and 2004.

CR3.11

a.

0	123334555599	
1	00122234688	
2	1112344477	
3	0113338	
4	37	Stem: Thousands
5	23778	Leaf: Hundreds

The stem-and-leaf display shows a positively skewed distribution with a single peak. There are no extreme values. A typical total length is around 2100 and the great majority of total lengths lie in the 100 to 3800 range.

c. The number of subdivisions that have total lengths less than 2000 is $12 + 11 = 23$, and so the proportion of subdivisions that have total lengths less than 2000 is $23/47 = 0.489$. The number of subdivisions that have total lengths between 2000 and 4000 is $10 + 7 = 17$, and so the proportion of subdivisions that have total lengths between 2000 and 4000 is $17/47 = 0.361$.

CR3.13 **a.** The histogram shows a smooth positively skewed distribution with a single peak. **b.** A typical time difference between the two phases of the race is 150 seconds, with the majority of time differences lying between 50 and 350 seconds. There are about three values that could be considered extreme, with those values lying in the 650 to 750 range. **c.** Estimating the frequencies from the histogram we see that approximately 920 runners were included in the study and that approximately eight of those runners ran the late distance more quickly than the early distance (indicated by a negative time difference). Therefore the proportion of runners who ran the late distance more quickly than the early distance is approximately $8/920 = 0.009$.

CR3.15 There is a strong negative linear relationship between racket resonance frequency and sum of peak-to-peak accelerations. There are two rackets with data points

separated from the remaining data points. Those two rackets have very high resonance frequencies and their peak-to-peak accelerations are lower than those of all the other rackets.

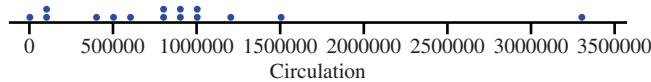
Chapter 4

4.1 **a.** $\bar{x} = \$148.77$; median = $\$142.33$. **b.** The mean is larger than the median since the distribution of these six values is positively skewed. The largest value is greatly separated from the remaining five values. **c.** The median is better as a description of a typical value since it is not influenced by the extreme value.

4.3 The mean caffeine concentration for the brands of coffee listed = 125.417 mg/cup. Therefore, the mean caffeine concentration of the coffee brands in mg/ounce is 15.677. This is significantly greater than the previous mean caffeine concentration of the energy drinks.

4.5 **a.** The large difference between the mean cost and the median cost along with the fact that the mean is greater than the median tells us that there are some large outliers in the distribution of wedding costs in 2012. **b.** The average cost is misleading because 50% of the weddings in 2012 cost less than \$18,086, which is much lower than the mean wedding cost. **c.** Agree with this statement. The mean is strongly influenced by outliers, and large outliers will pull the mean toward larger values. Given that the mean is so much larger than the median (which divides the distribution of wedding costs in half), less than 50% of the wedding costs must be greater than or equal to the mean.

4.7 The dotplot of circulation numbers (reproduced below) shows that the distribution is positively skewed with a large outlier. The mean should be used to describe a typical value of symmetric distributions and therefore should not be used to describe the center of this distribution.



4.9 **a.** The mean is 448.3. **b.** The median is 446. **c.** This sample represents the 20 days with the highest number of speeding-related fatalities, and so it is not reasonable to generalize from this sample to the other 345 days of the year.

4.11 Neither statement is correct. Regarding the first statement, we should note that unless the “fairly expensive houses” constitute a majority of the houses selling, these more costly houses will not have an effect on the median. Turning to the second statement, we point out that the small number of very high or very low prices will have no effect on the median, whatever the number of sales. Both statements can be corrected by replacing the median with the mean.

4.13 The two possible solutions are $x_5 = 32$ and $x_5 = 39.5$.

4.15 The median is 680 hours.

4.17 a. $\bar{x} = 52.111$. Variance = 279.111. $s = 16.707$.

b. The addition of the very expensive cheese would increase both the mean and the standard deviation.

4.19 a. Lower quartile = 4th value = 41. Upper quartile = 12th value = 62. Iqr = 21. b. The iqr for cereals rated good (calculated in Exercise 4.18) is 24. This is greater than the value calculated in Part (a).

4.21 $\bar{x} = 51.33$ $s = 15.22$. A typical amount poured into a tall slender glass is 51.33 ml. A typical deviation from the mean amount poured is 15.22 ml.

4.23 a. Variance = 22,735.7.

Standard deviation = $\sqrt{22735.7} = 150.78$

b. The large value of the standard deviation tells us that there is considerable variation between prices of these highly rated smart phones.

4.25 Variance = 4922.0

$s = \sqrt{4922.0} = 70.2$ milliseconds

4.27 a. Lower quartile = 0. Upper quartile = 195.

Interquartile range = 195. b. The lower quartile equals the minimum value for this data set because there is a large number of equal values (zero in this case) at the lower end of the distribution. This is unusual, and therefore, generally speaking, the lower quartile is not equal to the minimum value.

4.29 a. The standard deviation of percent return is a reasonable measure of unpredictability because it measures how much, on average, individual returns deviate from the mean return of the fund. A smaller standard deviation indicates smaller deviations (on average) from the mean return, and therefore less risk. b. A fund with a small standard deviation can still lose money if the average percent return is small relative to the standard deviation. Recall that the standard deviation is a typical deviation from the mean (either above or below the mean), and if the mean is smaller than the standard deviation, the return could be negative, resulting in the fund losing money.

4.31 a. i b. iii c. iv

4.33 a.

	Mean	Standard Deviation	Coefficient of Variation
Sample 1	7.81	0.398	5.102
Sample 2	49.68	1.739	3.500

b. The values of the coefficient of variation are given in the table in Part (a). The fact that the coefficient of variation is smaller for Sample 2 than for Sample 1 is not surprising since, relative to the actual amount placed in the containers, it is easier to be more accurate when larger amounts are being placed in the containers.

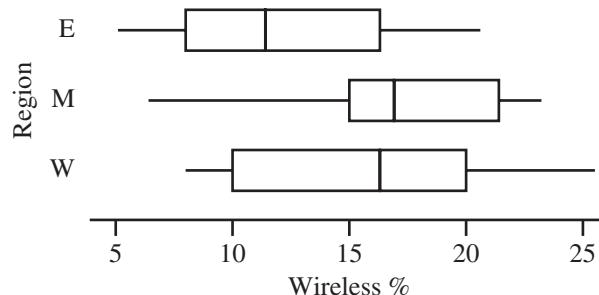
4.35 a. Median = 1.62, lower quartile = 1.28, upper quartile = 1.92 b. Connecticut (5.10%) is an outlier.

c.



The distribution is positively skewed with one outlier on the high end of the scale. The median is 1.62%, and the lower and upper quartiles are 1.28% and 1.92%, respectively. The middle 50% of the data values range between these quartiles and is approximately symmetric. Excluding the outlier, the distribution of the remaining data values is approximately symmetric.

4.37 a.



b. The distributions of wireless percent for the Midwest states and West states are centered at a higher value than for states in the East.

4.39 a. Median = average of 9th and 10th values = $(10 + 10)/2 = 10$

Lower quartile = 5th value = 6

Upper quartile = 14th value = 13

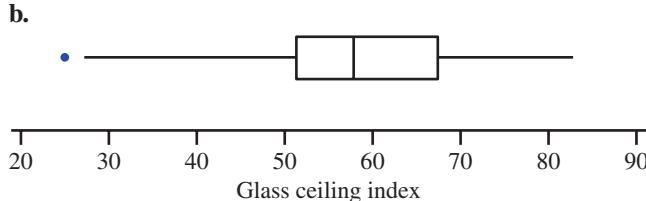
Interquartile range = $13 - 6 = 7$

b. There are no outliers.

4.41 a. Lower quartile: 51.3; Upper quartile: 67.45;

Interquartile range = $67.45 - 51.3 = 16.15$. South Korea's observation (25.0) is an outlier.

b.



c. The Nordic countries are located in the whisker at the right side of the boxplot.

4.43 a. Roughly 68% of speeds would have been between those two values. b. Roughly 16% of speeds would exceed 57 mph.

4.45 Since the mean of the distribution is 27 minutes and the standard deviation is 24 minutes, 0 is just over 1 standard deviation below the mean. Therefore, if the

distribution were approximately normal, then just under $(100 - 68)/2 = 16\%$ of travel times would be less than 0, which is clearly not the case. Thus the distribution cannot be well approximated by a normal curve.

- 4.47** **a.** At least 75% of observations must lie between those two values. **b.** The required interval is $(2.90, 70.94)$. **c.** The distribution cannot be approximately normal.

4.49 For the first test, $z = 1.5$; for the second test, $z = 1.875$. Since the student's z score in the second test is higher, the student did better relative to the other test takers in the second test.

4.51 At least 10% of the students had no debt. At least 25% of the students had no debt. Approximately 50% of the students had a debt of \$11,000 or less. Approximately 75% of the students had a debt of \$24,600 or less.

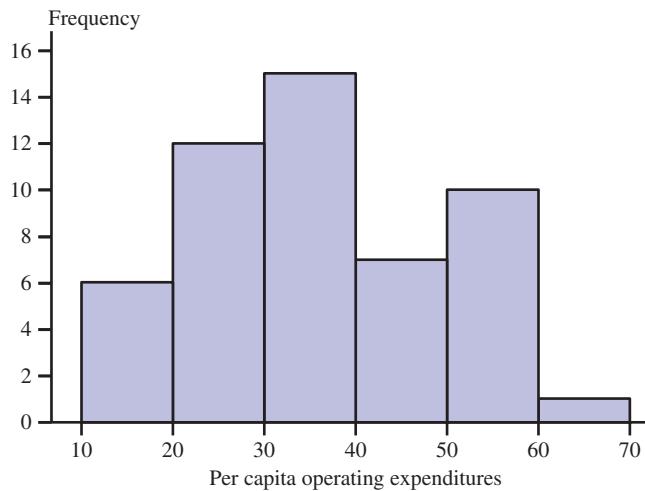
Approximately 90% of the students had a debt of \$39,300 or less.

4.53 The best conclusion we can reach is that at most 16% of weight readings will be between 49.75 and 50.25.

4.55 The value of the standard deviation tells us that a typical deviation of *the number of answers changed from right to wrong* from the mean of this variable is 1.0. However, 0 is only 0.9 below the mean and negative values are not possible, and so for a typical deviation to be 1.0 there must be some values more than 1.0 *above* the mean, that is, values above 1.9. This suggests that the distribution is positively skewed.

The value 3 is the lowest whole number value more than 3 standard deviations above the mean. Therefore, using Chebyshev's Rule, we can conclude that at most $1/3^2 = 1/9$ of students; that is, at most $72/9 = 8$ students changed at least three answers from correct to incorrect.

- 4.57** **a.**



- b. i.** The 50th percentile is between per capita operating expenditures of 30 and 40. **ii.** The 70th percentile is between per capita operating expenditures of 40 and 50. **iii.** The 10th percentile is between per capita operating expenditures of 10 and 20. **iv.** The 90th percentile is between per capita operating expenditures of 50 and 60.

v. The 40th percentile is between per capita operating expenditures of 30 and 40.

4.59 **a.** The minimum value and the lower quartile were both 1. **b.** More than half of the data values were equal to the minimum value. **c.** Between 25% and 50% of patients had unacceptable times to defibrillation.

d. $(\text{Upper quartile}) + 3(\text{iqr}) = 9$. Since 7 is less than 9, 7 must be a mild outlier.

4.61 **a.** $\bar{x} = 299$. The five deviations are: 46, -7, 35, -23, -51

b. The sum of the rounded deviations is 0.

c. Variance = 1630

$$s = \sqrt{1630} = 40.37$$

4.63 **a.** This is a correct interpretation of the median.

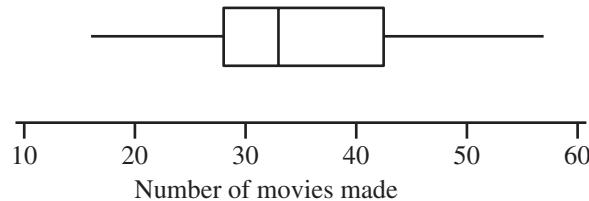
b. Here the word "range" is being used to describe the interval from the minimum value to the maximum value. The statement claims that the median is defined as the midpoint of this interval, which is not true. **c.** If there is no home below \$300,000, then certainly the median will be greater than \$300,000 (unless more than half of the homes cost exactly \$300,000).

4.65 The new mean is $\bar{x} = 38.364$. The new values and their deviations from the mean are shown in the table below.

Value	Deviation
52	13.636
13	-25.364
17	-21.364
46	7.636
42	3.636
24	-14.364
32	-6.364
30	-8.364
58	19.636
35	-3.364

The value of s^2 for the new values is the same as for the old values.

- 4.67**



The distribution of number of movies made is slightly positively skewed. This tells us that there is less variability in the lower 50% of number of movies made than in the upper 50%.

4.69 **a.** $\bar{x} = 22.15$. $s = \sqrt{129.187} = 11.366$.

b. The lower quartile is the average of the 5th and 6th values = $(18 + 18)/2 = 18$. The upper quartile is the average of the 15th and 16th values = $(20 + 21)/2 = 20.5$. Interquartile range = $20.5 - 18 = 2.5$. **c.** The values 25 and 28 are mild outliers and 69 is an extreme outlier.

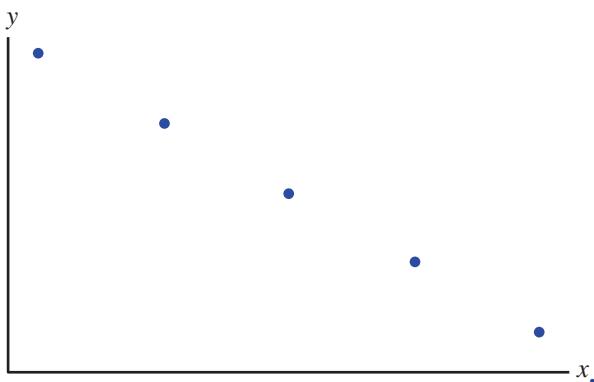
Chapter 5

- 5.1** Scatterplot 1 (i) Yes (ii) Yes (iii) Positive
 Scatterplot 2 (i) Yes (ii) Yes (iii) Negative
 Scatterplot 3 (i) Yes (ii) No (iii) -
 Scatterplot 4 (i) Yes (ii) Yes (iii) Negative

5.3 **a.** Positive. Husbands and wives tend to come from similar backgrounds and therefore have similar expectations in terms of income. **b.** Close to zero. There is no reason to believe that there is an association between height and IQ. **c.** Positive. People with large feet tend to be taller than people with small feet.

5.5 No. A correlation coefficient of 0 implies that there is no *linear* relationship between the variables. There could still be a nonlinear relationship.

5.7 Scatterplot for which $r = -1$:



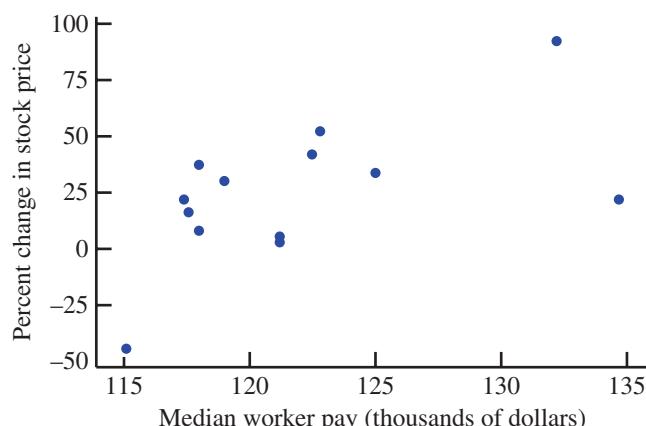
5.9 **a.** $r = 0.204$. There is a weak positive linear relationship between cost per serving and fiber per serving.

b. $r = 0.241$. This correlation coefficient is slightly greater than the correlation coefficient for the per serving data.

5.11 **a.** The value of the correlation coefficient is negative, which suggests that students who use a cell phone for more hours per day tend to have lower GPAs.

b. The relationship between texting and GPA has the same direction as the direction for cell phone use and GPA, so the correlations must have the same sign. The relationship between texting and GPA is not as strong as the relationship between cell phone use and GPA, so the correlation coefficient must be closer to zero. Therefore, $r = -0.10$ is the only option that satisfies both of the criteria. **c.** Since it is reasonable to believe that texts sent would be approximately equal to texts received, there would be a positive association between these items. In addition, given that the two texting items are nearly perfectly correlated, the correlation coefficient must be close to +1 or -1. Therefore, the correlation coefficient must be close to +1.

5.13 **a.**



b. $r = 0.578$; There is a moderately strong, positive association between the percent change in stock price and median worker pay. **c.** The conclusion is justified based on these data because there is a positive association between the variables. In general, as median worker pay increases, so does the percent change in stock price. **d.** It is not reasonable to generalize conclusions based on these data to all U.S. companies because these data were not randomly selected. These are the top 13 highest paying companies in the United States.

5.15 No, because “artist” is a categorical variable.

5.17 Scatterplot 1 seems to show a linear relationship between x and y , while Scatterplot 2 shows a curved relationship between the two variables. It makes sense to use the least-squares regression line to summarize the relationship between x and y for the data set in Scatterplot 1, but not for the data set in Scatterplot 2.

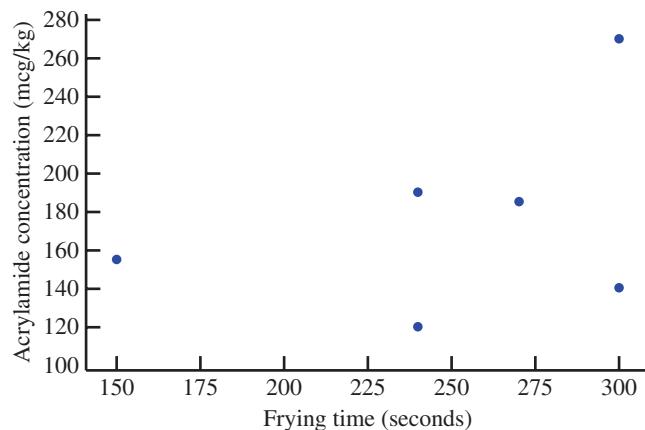
5.19 **a.** $\hat{y} = 51.305 + 0.1633x$, where \hat{y} is the predicted percentage of alumni who would strongly agree and x is the ranking. **b.** 59.47 **c.** Since the ranking of 10 is outside this range, the least-squares regression line should not be used to predict the percentage of alumni who would strongly agree because there is no guarantee that the linear pattern will continue outside this range.

5.21 **a.** The scatterplot and the least-squares line support the fact that, generally speaking, the higher the temperature the greater the proportion of larvae that were captured moving upstream. **b.** Approximately the same number of larvae moving upstream as downstream is represented by a net directionality of zero. According to the least-squares line, this will happen when the mean temperature is approximately 8.8°C .

5.23 **a.** Negative. As the patient-to-nurse ratio increases we would expect nurses’ stress levels to increase and therefore their job satisfaction to decrease. **b.** Negative. As the patient-to-nurse ratio increases we would expect patient satisfaction to decrease. **c.** Negative. As the patient-to-nurse ratio increases we would expect quality of care to decrease.

- 5.25** a. The response variable is the acrylamide concentration, and the predictor variable is the frying time.

b.



There is a weak, positive association ($r = 0.379$) between frying time and acrylamide concentration.

- 5.27** a. There is a strong negative relationship between mean call-to-shock time and survival rate. The relationship is close to linear, particularly if the point with the highest x value is disregarded. If that point is included, then there is the suggestion of a curve. b. $\hat{y} = 101.32847 - 9.29562x$

- 5.29** The slope would be -4000 , since the slope of the least-squares line is the increase in the average y value for each increase of one unit in x . Here the average home price decreases by \$4000 for each increase of 1 mile in the distance east of the bay.

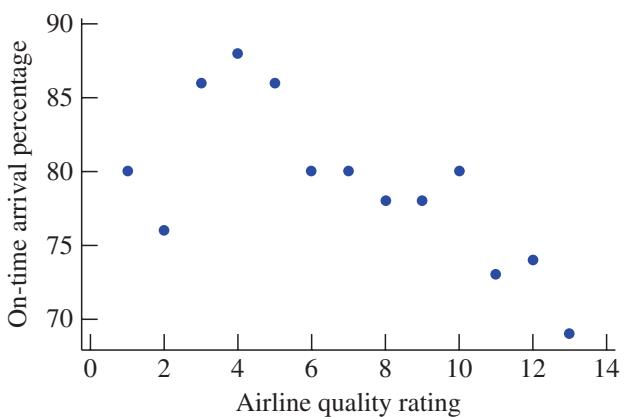
- 5.31** We do not know that the same linear relationship will apply for x values outside this range.

- 5.33** $b = r(s_y/s_x)$, where s_y and s_x are the standard deviations of the y values and the x values, respectively. Since standard deviations are always positive, b and r must have the same sign.

- 5.35** a. The equation of the least squares regression line is $\hat{y} = -9.071 + 1.571x$, where x = years of schooling and y = median hourly wage gain. When $x = 15$, $\hat{y} = 14.5$. The least squares regression line predicts a median hourly wage gain of 14.5 percent for the 15th year of schooling. b. The actual wage gain percent is very close to the value predicted in Part (a).

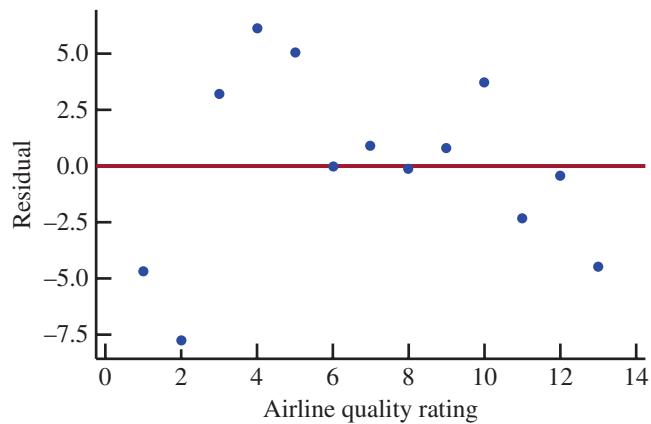
- 5.37** b. The scatterplot for girls shows stronger evidence of a curved relationship. c. $\hat{y} = 479.997 + 12.525x$ e. The decision to use a curve is supported by the curved pattern in the residual plot.

- 5.39** a. No, the pattern in the scatterplot does not look linear. There are two unusual values (airline quality ratings of 1 and 2) that make the pattern look curved. Without those two points, the scatterplot would look linear.



- b. $\hat{y} = 85.615 - 0.9341x$, where \hat{y} is the predicted on-time arrival percentage and x is the airline quality rating.

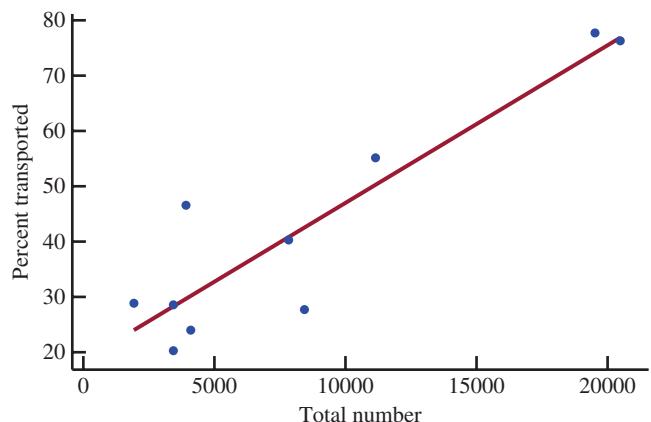
c.



The residual plot shows a curved pattern that would call into question the appropriateness of using a linear model to describe the relationship between airline quality rating and on-time arrival percentage.

- 5.41** a. The observation (150,155) is potentially influential because that point has an x value that is far away from the rest of the data set. b. $\hat{y} = -44 + 0.83(270) = 180.1$; This prediction is smaller than that made in the previous exercise.

- 5.43** a.



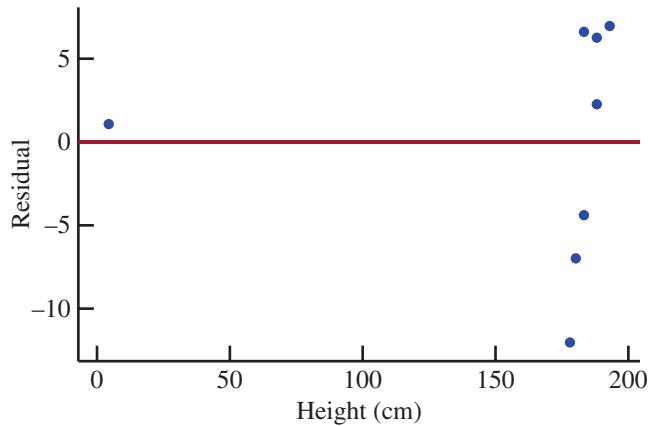
- b. Yes, there appears to be a strong linear relationship between the total number of salmon in the stream and the

percent of salmon killed by bears that are transported away from the stream. **c.** $\hat{y} = 18.483 + 0.00287x$, where x is the total number of salmon in a creek and y is the percent of salmon killed by bears that were transported away from the stream prior to the bear eating.

5.45 a.

Height (x)	Weight (y)	Predicted Weight (\hat{y})	Residuals ($y - \hat{y}$)
188	95	88.66	6.34
4	4	2.916	1.084
188	91	88.66	2.34
178	72	84	-12
183	93	86.33	6.67
180	78	84.932	-6.932

b.



The residual plot has one observation (the Lego Batman, at 4 cm and 4 kg) that is noticeably different from the others. There is a strong pattern in the residual plot.

- c.** This point is influential in determining both the slope and the intercept. That point is far removed from the other points in both the x and y directions. In fact, by removing the Lego Batman point, the equation of the least-squares regression line changes to $\hat{y} = -220.43 + 1.6643x$.
d. No. The least-squares line in Part (a) was calculated using the data for men, and we have no reason to believe that this line will also apply to women.

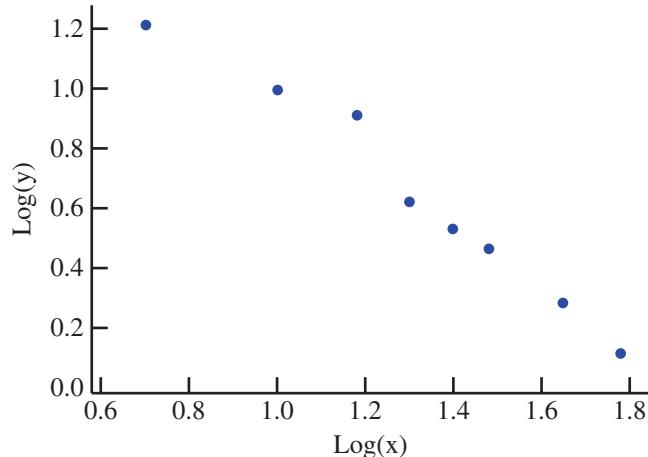
5.47 a. 0.154 **b.** No, since the r^2 value for y = first-year-college GPA and x = SAT II score was 0.16, which is not large. Only 16% of the variation in first-year-college GPA could be attributed to the approximate linear relationship between SAT II score and first-year-college GPA.

5.49 b. $\hat{y} = 85.334 - 0.0000259x$. $r^2 = 0.016$. **c.** The line will not give accurate predictions. **d.** Deleting the point (620, 231, 67), the equation of the least-squares line is now $\hat{y} = 83.402 + 0.0000387x$. Removal of the point does greatly affect the equation of the line.

5.51 $r^2 = 0.951$. 95.1% of the variation in hardness is attributable to the approximate linear relationship between time elapsed and hardness.

5.53 a. $r = 0$, $\hat{y} = \bar{y}$. **b.** For values of r close to 1 or -1, s_e will be much smaller than s_y . **c.** $s_e \approx 1.5$. **d.** $\hat{y} = 7.92 + 0.544x$, $s_e \approx 1.02$

5.55 a.

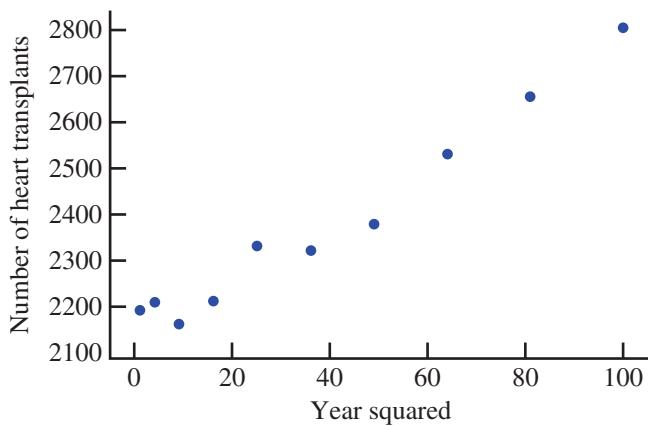


The relationship between $\log(x)$ and $\log(y)$ seems to be linear.
b. The value of r^2 is large (0.973) and the value of s_e is relatively small (0.0657), indicating that a line is a good fit for the data. **c.** 2.477%

5.57 Answers will vary. Scatterplot (b) shows a curve that tends to follow the pattern of points in the scatterplot.

- 5.59 a.** The relationship appears to be nonlinear.
b. There is an obvious pattern in the residual plot.
c. The second transformation ($\ln(x)$ vs. y) seems to be somewhat more successful in straightening the data.
d. $\hat{y} = 62.42 + 43.81x$, where \hat{y} is the predicted success %, and x is the natural logarithm of the energy of shock.
e. 86.94%, 52.64%

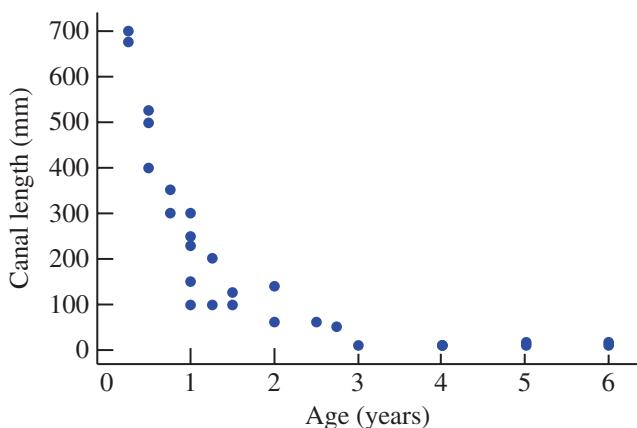
5.61 a. One transformation to straighten the plot is $x = \text{year}^2$ and $y = \text{number of heart transplants}$.



b. The equation of the least-squares regression line is $\hat{y} = 2139.87 + 6.2321x^*$, where \hat{y} is the predicted number of heart transplants, and x^* is the year (since 2006) squared. The predicted number of heart transplants in 2016 (the eleventh year since 2006, or $x^* = 11^2 = 121$) is $\hat{y} = 2139.87 + 6.2321(121) = 2893.9541$. It is predicted that there will be approximately 2894 heart transplants in 2016. **c.** We have to be confident that the pattern observed

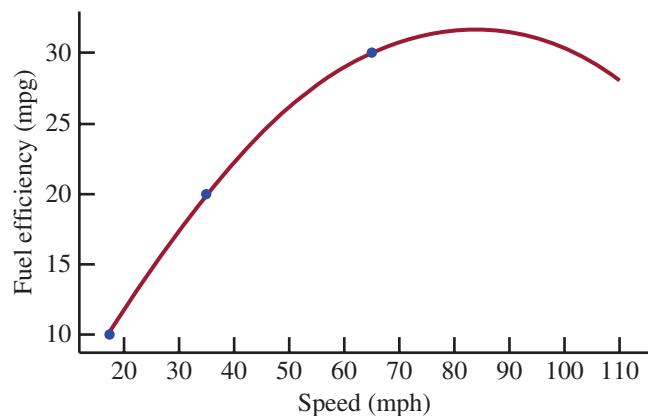
between 2006 and 2015 will continue up to 2016. This is reasonable so long as circumstances remain basically the same. To expect the same pattern to continue to 2026, however, might be unreasonable.

5.63



The relationship between age and canal length is not linear. One transformation that makes the plot roughly linear is $x' = \log(x)$ and $y' = \log(y)$.

5.65



As illustrated in the graph above, it is quite possible that the relationship between speed and fuel efficiency is modeled by a curve (in particular, by a quadratic curve), and that for greater speeds the fuel efficiency is negatively related to the speed.

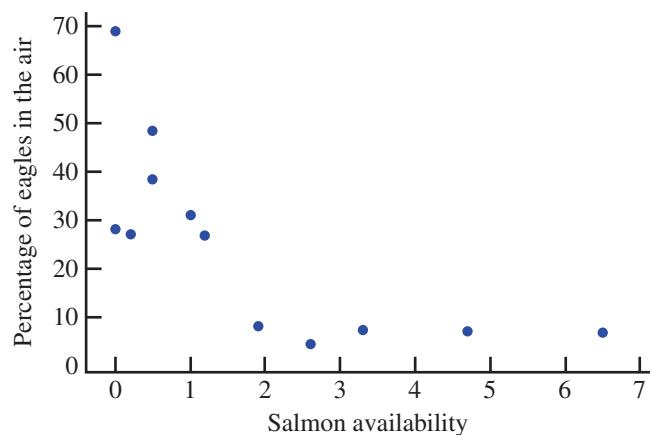
5.67 Consider the scatterplot for the combined group. Between $x = 0$ and $x = 1$ the points form an approximately linear pattern with a slope of 6.93, while between $x = 1$ and $x = 10$ the points form an approximately linear pattern with a slope of just over 3. Clearly this pattern could not be modeled with a single straight line; a curve would be more appropriate.

5.69 **a.** $r = 0.944$. There is a strong, positive linear relationship between depression rate and sugar consumption. **b.** No. Since this was an observational study, no conclusion relating to causation may be drawn. **c.** Yes. Since the set of countries used was not a random sample from the set of all countries (nor do we have any reason to think that these

countries are *representative* of the set of all countries), we cannot generalize the conclusions to countries in general.

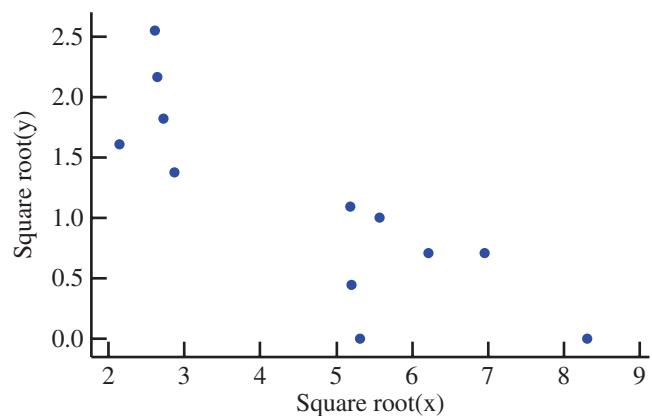
5.71 **a.** $y - \hat{y} = 20.796$ **b.** $r = -0.755$ **c.** $s_e = 11.638$

5.73 **a.**



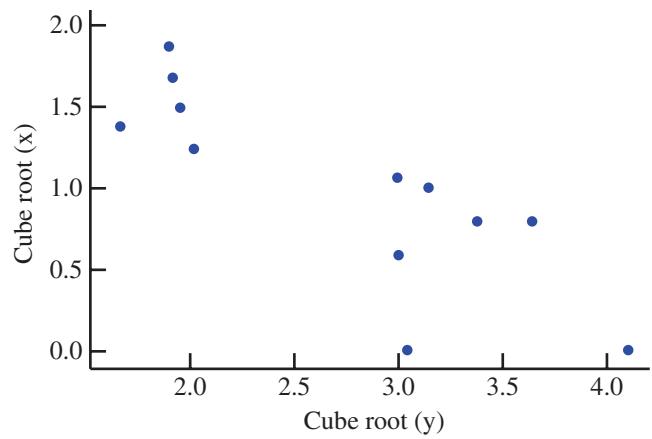
There seems to be a curved relationship between percentage of eagles in the air and salmon availability.

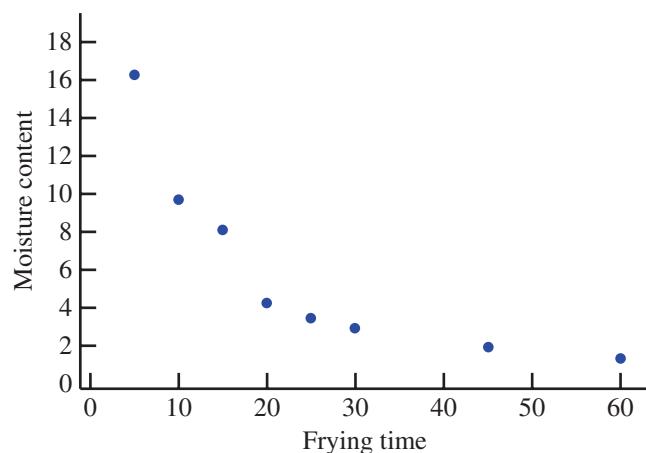
c.



Yes, the pattern in the scatterplot is more nearly straight than the one in Part (a).

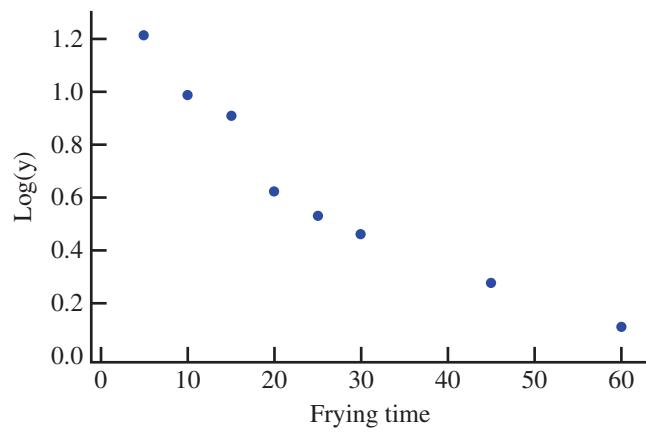
d. Taking the cube roots of both variables might be successful in straightening the plot. (In fact, this transformation does seem to result in a more nearly linear relationship, as shown in the scatterplot.)



5.75 a.

No, the relationship does not seem to be linear.

b.



Yes, the transformed relationship is closer to being linear than the relationship shown in Part (a).

c. $y' = 1.143 - 0.0192x$, where x = frying time and y' is log(moisture content). d. 2.964%

5.77 a. The correlation coefficient of 0.12 indicates a weak relationship between price and weight of the wine bottles. b. $r^2 = 0.0144$. Approximately 1.44% of the variation in price of the wine can be explained by the linear relationship between price and weight of the wine bottles.

5.79 a. $r = 0$ b. For example, adding the point $(6, 1)$ gives $r = 0.510$. (Any y -coordinate greater than 0.973 will work.) c. For example, adding the point $(6, -1)$ gives $r = -0.510$. (Any y -coordinate less than -0.973 will work.)

CUMULATIVE REVIEW EXERCISES 5

CR5.3 The peaks in rainfall do seem to be followed by peaks in the number of *E. coli* cases, with rainfall peaks around May 12, May 17, and May 23 followed by peaks in the number of cases on May 17, May 23, and May 28. (The incubation period seems to be more like 5 days than the 3 to 4 days mentioned in the caption.) Thus, the graph does show a close connection between unusually heavy rainfall and the incidence of the infection. The storms may not be

responsible for the increased illness levels, however, since the graph can only show us association, not causation.

CR5.5 a. $r = 0.394$ b. $r = -0.664$ c. There is a stronger relationship between Happiness Index and response to Statement 2. d. There is a weak, positive association between the Happiness Index and the response to Statement 1. In contrast, there is a moderate, negative association between the Happiness Index and the response to Statement 2.

CR5.7 $r = 0.975$; this value is consistent with the previous answer, because the correlation coefficient is large and positive (close to 1), which indicates a strong positive association between the amount spent on science and the amount spent on pets.

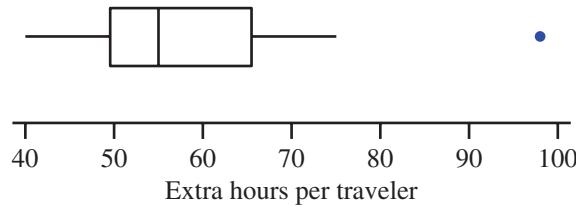
CR5.9 Lower quartile = 7th value = 10478 mg/kg.

Upper quartile = 20th value = 11778 mg/kg.

Interquartile range = 11778 - 10478 = 1300 mg/kg.

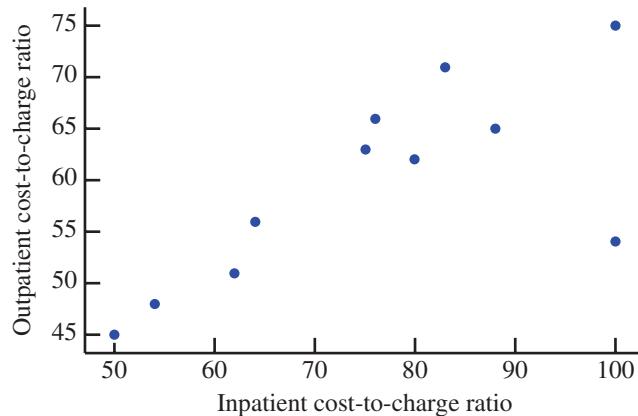
CR5.11 a. Mean = 59.846, Median = 55. Since the mean is greater than the median, the distribution of extra travel hours is likely to be positively skewed.

b.



There is one outlier (Los Angeles, with 98 extra travel hours per traveler). The median number of extra hours of travel per traveler is 55, and, apart from the one outlier, the values range from 40 to 75. The distribution is positively skewed.

CR5.13 a. $r = 0.730$.



Based on the correlation coefficient value of 0.730 and the scatterplot, there is a moderately strong linear relationship between the cost-to-charge ratio for inpatient and outpatient services. Looking at the scatterplot, it is clear that one outlier is affecting the correlation. If that point were removed, there would be a strong linear relationship between the variables.

- b.** There is an outlier (the point for Harney District).
c. If the outlying point were removed, then the linear relationship would be much stronger, and the value of r would therefore be greater.

CR5.15 **a.** 76.64% of the variability in clutch size can be attributed to the approximate linear relationship between snout-vent length and clutch size. **b.** $s_e = 29.250$. This is a typical deviation of an observed clutch size from the clutch size predicted by the least-squares line.

Chapter 6

6.1 A chance experiment is any activity or situation in which there is uncertainty about which of two or more possible outcomes will result.

6.3 **a.** {AA, AM, MA, MM}.

6.5 **a.** $A = \{\text{Head oversize, Prince oversize, Slazenger oversize, Wimbledon oversize, Wilson oversize}\}$ **b.** $B = \{\text{Wimbledon midsize, Wilson midsize, Wimbledon oversize, Wilson oversize}\}$ **c.** $\text{not } B = \{\text{Head midsize, Prince midsize, Slazenger midsize, Head oversize, Prince oversize, Slazenger oversize}\}$

6.7 **b. i.** $A^C = \{(15, 50), (15, 100), (15, 150), (15, 200)\}$

ii. $A \cup B = \{(10, 50), (10, 100), (10, 150), (10, 200), (12, 50), (12, 100), (12, 150), (12, 200), (15, 50), (15, 100)\}$

iii. $A \cap B = \{(10, 50), (10, 100), (12, 50), (12, 100)\}$

c. A and C are not disjoint. B and C are disjoint.

6.9 **b.** $A = \{3, 4, 5\}$ **c.** $C = \{125, 15, 215, 25, 5\}$

6.11 **a.** $A = \{\text{NN, DNN, NDN}\}$ **b.** $B = \{\text{DDNN, DNDN, NDDN}\}$ **c.** The number of outcomes is infinite.

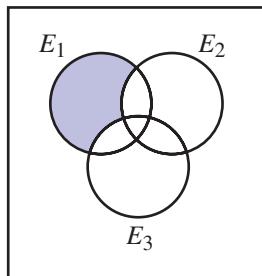
6.13 **a.** $B^C = \{(1,1,1), (1,1,2), (1,1,3), (1,2,1), (1,2,2), (1,3,1), (1,3,3), (2,1,1), (2,1,2), (2,2,1), (2,2,2), (2,2,3), (2,3,2), (2,3,3), (3,1,1), (3,1,3), (3,2,2), (3,2,3), (3,3,1), (3,3,2), (3,3,3)\}$

b. $C^C = \{(1,1,2), (1,2,1), (1,2,2), (1,2,3), (1,3,2), (2,1,1), (2,1,2), (2,1,3), (2,2,1), (2,2,2), (2,2,3), (2,3,1), (2,3,2), (2,3,3), (3,1,2), (3,2,1), (3,2,2), (3,2,3), (3,3,2)\}$

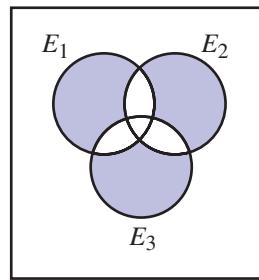
c. $A \cup B = \{(1,1,1), (2,2,2), (3,3,3), (1,2,3), (1,3,2), (2,1,3), (2,3,1), (3,1,2), (3,2,1)\}$ **d.** $A \cap B = \emptyset$

e. $A \cap C = \{(1,1,1), (3,3,3)\}$

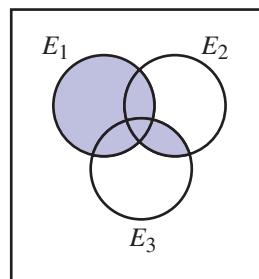
6.15 **a.**



b.



c.



6.17 {expedited overnight delivery, expedited second-business-day delivery, standard delivery, delivery to the nearest store for customer pick-up}

6.19 **a.** 0.8 **b.** 0.8. This probability is the same as the one in Part (a).

6.21 **a.** {fiction hardcover, fiction paperback, fiction digital, fiction audio, nonfiction hardcover, nonfiction paperback, nonfiction digital, nonfiction audio} **b.** No

6.23 **a. i.** $A = \{(C, N), (N, C), (N, N)\}$.

ii. $P(A) = 0.09 + 0.09 + 0.01 = 0.19$.

b. i. $B = \{(C, C), (N, N)\}$. **ii.** $P(B) = 0.81 + 0.01 = 0.82$.

6.25 **a.** 0.225 **b.** 0.775 **c.** 0.06 **d.** 0.715 **e.** 0.234

6.27 **a.** 0.3 **b.** No. The problems to be graded will be selected at random, so the probability of being able to turn in both of the selected problems is the same whatever your choice of three problems to complete. **c.** 0.6

6.29 If the events F and I were mutually exclusive, we could add the probabilities to obtain $P(F \cup I)$. Since $P(F) + P(I) = 0.71 + 0.52 = 1.23$ is greater than 1 (and probabilities cannot be greater than 1), we know that F and I cannot be mutually exclusive.

6.31 **a.** 0.231 **b.** 0.586 **c.** 0.818

6.33 **a.** The 24 outcomes (including the given outcome) are: $(1, 2, 3, 4), (1, 2, 4, 3), (1, 3, 2, 4), (1, 3, 4, 2), (1, 4, 2, 3), (1, 4, 3, 2), (2, 1, 3, 4), (2, 1, 4, 3), (2, 3, 1, 4), (2, 3, 4, 1), (2, 4, 1, 3), (2, 4, 3, 1), (3, 1, 2, 4), (3, 1, 4, 2), (3, 2, 1, 4), (3, 2, 4, 1), (3, 4, 1, 2), (3, 4, 2, 1), (4, 1, 2, 3), (4, 1, 3, 2), (4, 2, 1, 3), (4, 2, 3, 1), (4, 3, 1, 2), (4, 3, 2, 1)$.

b. The event is $\{(1, 2, 4, 3), (1, 4, 3, 2), (1, 3, 2, 4), (4, 2, 3, 1), (3, 2, 1, 4), (2, 1, 3, 4)\}$. The probability is $6/24 = 1/4$.

6.35 **a.** BC, BM, BP, BS, CM, CP, CS, MP, MS, PS
b. 0.1 **c.** 0.4 **d.** 0.3

6.37 a. $P(O_1) = P(O_3) = P(O_5) = 1/9$ and $P(O_2) = P(O_4) = P(O_6) = 2/9$ b. $1/3, 4/9$ c. $9/21, 2/7$

6.39 a. 0.24 b. 0.36 c. 0.12 d. 0.75

6.41 $P(\text{hybrid}|\text{male}) = 0.397$. $P(\text{male}|\text{hybrid}) = 0.694$. These probabilities are not equal. The first is the proportion of males who bought hybrids, and the second is the proportion of hybrid buyers who were males; they are different quantities.

6.43 a. 0.440 b. 0.560 c. 0.570 d. 0.260

6.45 a. Yes. The probabilities given are consistent with the comments in the article that a baby is more likely to have Down Syndrome if it is born to an older mother and that younger women are more fertile. b. No. According to the quote from the article $P(D|Y^c)$ is greater than $P(D|Y)$. In this statement these two probabilities are equal. c. No. According to the quote from the article $P(D|Y^c)$ is greater than $P(D|Y)$. In this statement these two probabilities are equal. d. No. The statement " $P(Y) = 0.4$ " is not consistent with the comment in the article that younger women are more fertile. e. No. The statement " $P(Y) = 0.4$ " is not consistent with the comment in the article that younger women are more fertile. f. No. The statement " $P(Y) = 0.4$ " is not consistent with the comment in the article that younger women are more fertile.

6.47 a. 0.769 b. 0.890 c. Since the conditional probabilities in (a) and (b) are not equal, a prediction that a baby is male and a prediction that a baby is female are not equally reliable.

6.49 The following probabilities can be used to justify the report's conclusion that females are more likely to wear seat belts than males in both urban and rural areas.

$$P(\text{urban wear seat belt}|\text{male}) \approx \frac{871}{871 + 129} = 0.871$$

$$P(\text{urban wear seat belt}|\text{female}) \approx \frac{928}{928 + 72} = 0.928$$

$$P(\text{rural wear seat belt}|\text{male}) \approx \frac{770}{770 + 230} = 0.77$$

$$P(\text{rural wear seat belt}|\text{female}) \approx \frac{837}{837 + 163} = 0.837$$

The differences in percentages of females and males who wear seat belts are below, which support the report's conclusion that the difference in the proportion of females and the proportion of males who wear seat belts is greater for rural areas.

$$\text{Urban: } 0.928 - 0.871 = 0.057$$

$$\text{Rural: } 0.837 - 0.77 = 0.067$$

6.51 a. 0.096 b. 0.384 c. No, the probabilities in Parts (a) and (b) are not equal. Part (a) is the proportion of female drivers who do not use a seat belt, and Part (b) is the proportion of drivers who do not use a seat belt who are female. There is no reason to believe that these proportions should be equal.

6.53 a. 85% b. 0.15

6.55 a.

	Does Not Use Complementary Therapies	Does Use Complementary Therapies	Total
Convention			
Medications	0.758	0.122	0.879
Usually Help			
Convention			
Medications	0.096	0.025	0.121
Usually Do Not Help			
Total	0.853	0.147	1

b. The cell entry 0.122 represents the probability of (conventional medicines usually help *and* does use complementary therapies). The cell entry 0.096 represents the probability of (conventional medicines usually do not help *and* does not use complementary therapies). The cell entry 0.025 represents the probability of (conventional medicines usually do not help *and* does use complementary therapies). c. $P(CH|CT) = 0.829$, and $P(CH) = 0.879$. Since these two probabilities are not equal, the two events are not independent.

6.57 a. Given that the selected graduate finished college with no student debt, the probability that the selected graduate strongly agrees that education was worth the cost is 0.49. b. Given that the selected graduate finished college with high student debt, the probability that the selected graduate strongly agrees that education was worth the cost is 0.18. c. We are given that $P(A) = 0.38$ and $P(A|H) = 0.18$. The events A and H are not independent because $P(A|H) \neq P(A)$.

6.59 a. 0.001. We have to assume that she deals with the three errands independently. b. 0.999 c. 0.009

6.61 a. 0.9931, increases b. 0.9266, decreases

6.63 The reason that $P(T)$ is not the average of the two given conditional probabilities is because there are different numbers (or different proportions) of people in the two given age groups (19 to 36 and 37 or older).

6.65 a. 0.336 b. 0.56 c. 0.06 d. 0.18

6.67 a. 0.00391 b. 0.00383

6.69 a. 0.6, slightly smaller b. 0.556 c. 0.333

6.71 a. 0.55 b. 0.45 c. 0.4 d. 0.25

6.73 a. $P(\text{at least one food allergy and severe reaction}) = (0.08)(0.39) = 0.0312$. b. $P(\text{allergic to multiple foods}) = (0.08)(0.3) = 0.024$.

6.75 a. 0.49 b. 0.3125 c. 0.24 d. 0.1

6.77 a. 0.645 b. 0.676, higher

6.79 a. i. 0.307 ii. 0.693 iii. 0.399 iv. 0.275 v. 0.122 b.

30.7% of faculty members use Twitter; 69.3% of faculty members do not use Twitter; 39.9% of faculty members who use Twitter also use it to communicate with students; 27.5% of faculty members who use Twitter also use it as a

learning tool in the classroom; 12.2% of faculty members use Twitter and use it to communicate with students
c. 0.122 **d.** 0.084

6.81 **a.**

	Uses Alternative Therapies	Does Not Use Alternative Therapies	Total
High School or Less	315	7,005	7,320
College – 1 to 4 years	393	4,400	4,793
College – 5 or more years	120	975	1,095
Total	828	12,380	13,208

b.

	Uses Alternative Therapies	Does Not Use Alternative Therapies	Total
High School or Less	0.024	0.530	0.554
College – 1 to 4 years	0.030	0.333	0.363
College – 5 or more years	0.009	0.074	0.083
Total	0.063	0.937	1.000

c. i. 0.083 ii. 0.063

6.83 The reason that $P(C)$ is not the average of the three conditional probabilities is that there are different numbers of people driving the three different types of vehicles (and that there are some drivers who are driving vehicles not included in those three types).

6.85

Radiologist 1			
	Predicted Male	Predicted Female	Total
Baby Is Male	74	12	86
Baby Is Female	14	59	73
Total	88	71	159

a. $P(\text{prediction is male} \mid \text{baby is male}) = 74/86 = 0.860$.
b. $P(\text{prediction is female} \mid \text{baby is female}) = 59/73 = 0.808$.
c. Yes. Since the answer to (a) is greater than the answer to (b), the prediction is more likely to be correct when the baby is male. **d.** The radiologist was correct for $74 + 59 = 133$ of the 159 babies in the study. So $P(\text{correct}) = 133/159 = 0.836$.

6.87 **a.** i. 0.99 ii. 0.01 iii. 0.99 iv. 0.01 **b.** 0.02
c. 0.5, it is consistent with the quote.

6.89 **b.** 0.27 **c.** 0.43 **d.** Not independent

6.91 **a.** 0.622 **b.** 0.167 **c.** 0.117 **d.** 0.506

6.93 **a.** 0.201 **b.** 0.196 **c.** 0.335 **d.** 0.177

6.101 **b.** 0.025 **c.** 0.069

6.103 **a.** 0.49 **b.** 0.88 **c.** 0.7

6.105 The events *selected adult exercises at least 30 minutes 3 times per week* and *is a millennial* are dependent events. A millennial is more likely to exercise at least 30 minutes 3 times per week (57.1%) when compared with adults over age 35 (51.1%). The age of the adult (millennial or over age 35) has an impact on the length of time the selected adult exercises.

6.107 **a.** $P(D) = 0.148$, $P(T) = 0.374$, and $P(D \cap T) = 0.106$ **b.** $P(D \cup T) = 0.416$.

$P(\text{neither a change in diagnosis nor change in treatment}) = 1 - P(D \cup T) = 0.584$ **c.** 0.416

6.109 **a.** 0.018 **b.** $P(A_1|L) = 0.444$ **c.** $P(A_2|L) = 0.278$
d. $P(A_3|L) = 0.278$

6.111 **a.** For example, use a single-digit random number to represent the outcome of the game. The digits 0–7 will represent a win for seed 1, and digits 8–9 will represent a win for seed 4. **b.** For example, use a single-digit random number to represent the outcome of the game. The digits 0–5 will represent a win for seed 2, and digits 6–9 will represent a win for seed 3. **c.** For example, use a single-digit random number to represent the outcome of the game. If seed 1 won game 1 and seed 2 won game 2, the digits 0–5 will represent a win for seed 1, and digits 6–9 will represent a win for seed 2. If seed 1 won game 1 and seed 3 won game 2, the digits 0–6 will represent a win for seed 1, and digits 7–9 will represent a win for seed 3. If seed 4 won game 1 and seed 2 won game 2, the digits 0–6 will represent a win for seed 2, and digits 7–9 will represent a win for seed 4. If seed 4 won game 1 and seed 3 won game 2, the digits 0–5 will represent a win for seed 3, and digits 6–9 will represent a win for seed 4. **d.** Answers will vary.

6.113 **a.** 0.4 **b.** 0.75 **c.** 0.25

6.115 0.49

6.117 Since the total number of viewers was 2517 and the number of viewers of R-rated movies was 1140, $P(R_2|R_1) = 1139/2516$ and $P(R_2|R_1^C) = 1140/2516$. These probabilities are not equal, so the events R_1 and R_2 are not independent. However, as a result of the fact that the total number of viewers is large, the two probabilities are very close, and therefore from a practical point of view the events may be regarded as independent.

6.119 **a.** 0.383 **b.** 0.617

Chapter 7

7.1 **a.** Discrete **b.** Continuous **c.** Discrete **d.** Discrete **e.** Continuous

7.3 **a.** Positive integers **b.** For example, S with $y = 1$, RS, LS with $y = 2$, RLS, RRS, LES, LLS with $y = 2$, RLS, RRS, LES, LLS with $y = 3$.

7.5 a. The possible values are the real numbers between 0 and 100, inclusive; b. y is continuous.

7.7 a. 3, 4, 5, 6, 7 b. -3, -2, -1, 0, 1, 2, 3 c. 0, 1, 2 d. 0, 1

7.9 $P(3 \leq x \leq 6) = 0.09 + 0.25 + 0.4 + 0.16 = 0.9$.

$P(3 < x < 6) = 0.25 + 0.4 = 0.65$. The second probability is smaller (and therefore the two probabilities are not equal) because the second probability does not include the possibility of a student taking three or six courses, while the first does include this possibility.

7.11 a. 0.1 b. 0.95

7.13 a. 0.82 b. 0.18 c. 0.65, 0.27

7.15 Results of the simulation will vary.

7.17 a.

x	0	1	2	3	4
$p(x)$	0.4096	0.4096	0.1536	0.0256	0.0016

b. 0 and 1 c. 0.1808

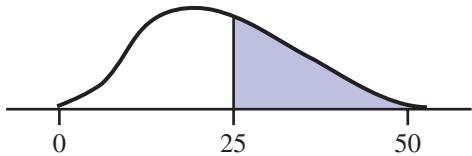
7.19 a. The smallest possible y value is 1, and the corresponding outcome is S. The second smallest y value is 2, and the corresponding outcome is FS. b. The set of positive integers.

c. $p(y) = (0.3)^{y-1}(0.7)$, for $y = 1, 2, 3, \dots$

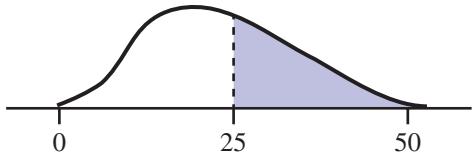
7.21

y	0	1	2	3
$p(y)$	0.16	0.33	0.32	0.19

7.23 a.



b.



7.25

$$P(2 < x < 3) = P(2 \leq x \leq 3) = (1/10)(3 - 2) = 0.1$$

$$P(x < 2) = (1/10)(2 - 0) = 0.2$$

$$P(x > 7) = (1/10)(10 - 7) = 0.3$$

Thus,

$$P(2 < x < 3) = P(2 \leq x \leq 3) < P(x < 2) < P(x > 7)$$

7.27 a. 0.25 b. 0.25 c. 0.25 d. These probabilities are the same because the density curve is uniform (always the same height) and the x -values of interest span the same range (the change in x is 100 in both cases). The rectangular regions under the density curve are the same size.

$$\text{7.29 a. } P\left(x > \frac{1}{2}\right) = \frac{1}{4}$$

$$\text{b. } P\left(x \geq \frac{1}{4}\right) = \frac{9}{16}$$

7.31 a. Area = $\frac{1}{2}(40)(0.05) = 1$, as required.

b. $P(w < 20) = 0.5$

c. $P(w < 10) = 0.125$

d. $P(w > 30) = 0.125$

7.33 3.306

7.35 a. 0.56 b. 65% c. The mean is not $(0 + 1 + 2 + 3 + 4)/5$ since, for example, far more cartons have 0 broken eggs than 4 broken eggs, and so 0 needs a much greater weighting than 4 in the calculation of the mean.

7.37 a. 0.14 b. $\mu - 2\sigma = 4.66 - 2(1.2) = 2.26$.

$\mu + 2\sigma = 4.66 + 2(1.2) = 7.06$. The values of x more than 2 standard deviations away from the mean are 1 and 2. c. The probability that x is more than 2 standard deviations away from its mean is 0.05.

7.39 a. 2.8, 1.288 b. 0.7, 0.781

7.41 a. 3.5, 2.27 b. 15.4, 75.94

7.43 a. 46.5 b. \$890

7.47 a. Whether y is positive or negative tells us whether or not the peg will fit into the hole. b. 0.003 c. 0.00632

d. Yes. Since there is no reason to believe that the pegs are being specially selected to match the holes (or vice versa), it is reasonable to think that x_1 and x_2 are independent.

e. Since 0 is less than half a standard deviation from the mean in the distribution of y , it is relatively likely that a value of y will be negative, and, therefore, that the peg will be too big to fit the hole.

7.49 a. 3.5, 2.9167, 1.7078 b. 3.5, 2.9167, 1.7078

7.51 a. 0.9938 b. 0.9996 c. 0.0004 d. 0.3347

7.53 a. 0.2458; In the long run, 24.58% of random samples of 6 passengers will contain exactly 4 passengers who travel with a smart phone. b. 0.2621 c. 0.9011

7.55 a. 0.264 b. 0.633 c. 0.367 d. 0.736

7.57 a. 0.755 b. $P(x > 15) = 0.051$ and $P(x < 5) = 0.0003$. Therefore, more than 15 having security solutions is more likely than fewer than 5 having security solutions.

7.59 a. The probability distribution of x is geometric, with success probability $p = 0.44$. The distribution is geometric because we are waiting until we find someone that washes sheets at least once a week. b. 0.138 c. 0.824 d. 0.176

7.61 a. 200 b. 13.416

7.63 a. Binomial distribution with $n = 100$ and $p = 0.2$

b. 20 c. 16, 4 d. A score of 50 is $(50 - 20)/4 = 7.5$ standard deviations from the mean in the distribution of x . So, a score of over 50 is very unlikely.

7.65 a. 0.017 b. 0.811, c. 0.425 d. The error probability when $p = 0.7$ is now 0.902. The error probability when $p = 0.6$ is now 0.586.

7.67 a. There is not a fixed number of trials. b. i. 0.062 ii. 0.284 iii. 0.716 iv. 0.779

7.69 a. 0.0975 b. 0.043 c. 0.815

7.71 a. 0.6887 b. 0.6826 c. 1.0000

7.73 a. 0.5 b. 1.0000 c. 0.939 d. 0.5910

7.75 a. 0.0706 b. 0.0228 c. 0.9996 d. approximately 1

7.77 a. 0.1469 b. 0.4577 c. 0.8944 d. 0.0730

7.79 a. $z^* = 1.88$ b. $z^* = 2.33$ c. $z^* = -1.75$

d. $z^* = -1.28$

7.81 a. $z^* = 1.34$ b. $z^* = 0.74$ c. $z^* = 0$ d. $z^* = -1.34$

e. The value of the 30th percentile is the negative of the value of the 70th percentile.

7.83 a. 0.2033 b. 0.5934 c. 0.0124

7.85 Those with emissions > 2.113 parts per billion

7.87 a. 0.2843 b. 0.0918 c. 0.4344 d. 29.928

7.89 a. 0.0035 b. 0.382

7.91 The second machine

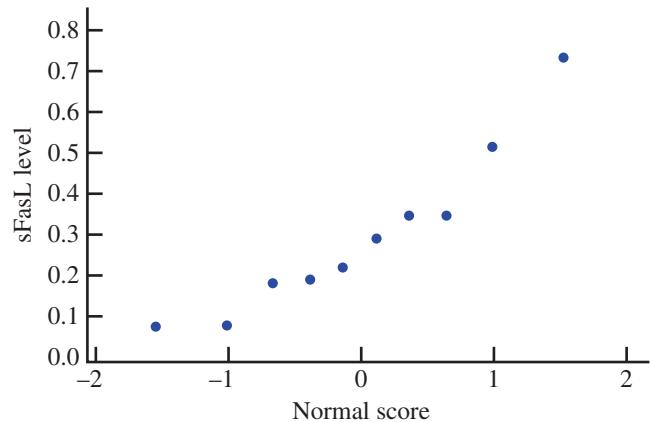
7.93 94.4 seconds or less

7.95 0.0013

7.97 b. The clear curve in the normal probability plot tells us that the distribution of fussing times is not normal.
d. The transformation results in a pattern that is much closer to linear than the pattern in Part (b).

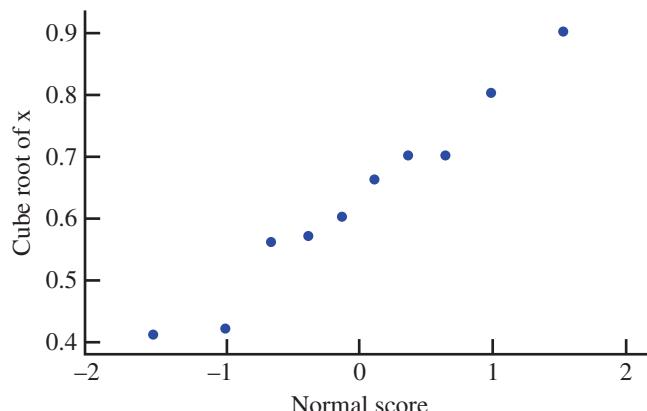
7.99 b. No. The distribution of x is positively skewed.
d. Yes. The histogram shows a distribution that is slightly closer to being symmetric than the distribution of the untransformed data. f. Both transformations produce histograms that are closer to being symmetric than the histogram of the untransformed data, but neither transformation produces a distribution that is truly close to being normal.

7.101 a.



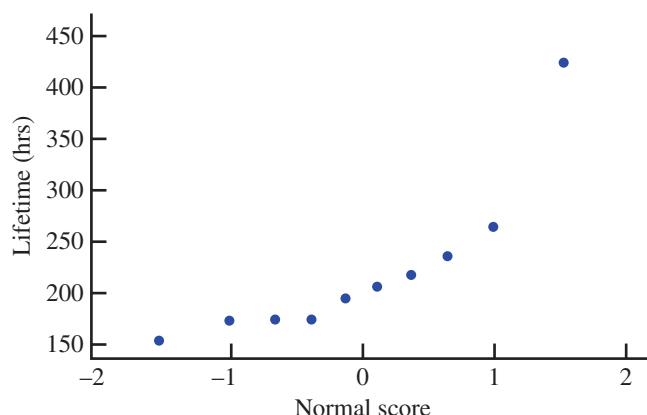
b. The normal probability plot appears curved.

c.



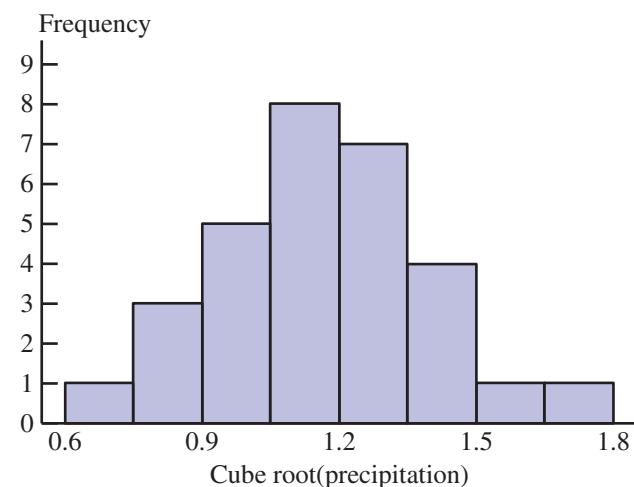
Yes. This normal probability plot appears more linear than the plot for the untransformed data.

7.103

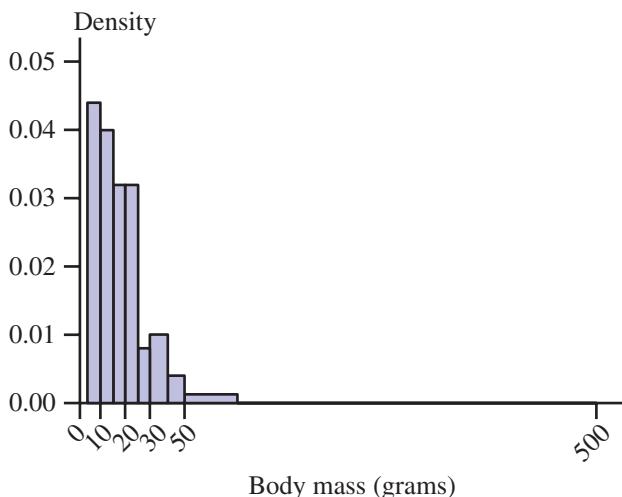


Since the normal probability plot shows a clear curve, the normal distribution is not a plausible model for power supply lifetime. However, it is worth noting that most of the apparent curved pattern is brought about by the single point with coordinates (1.539, 422.6).

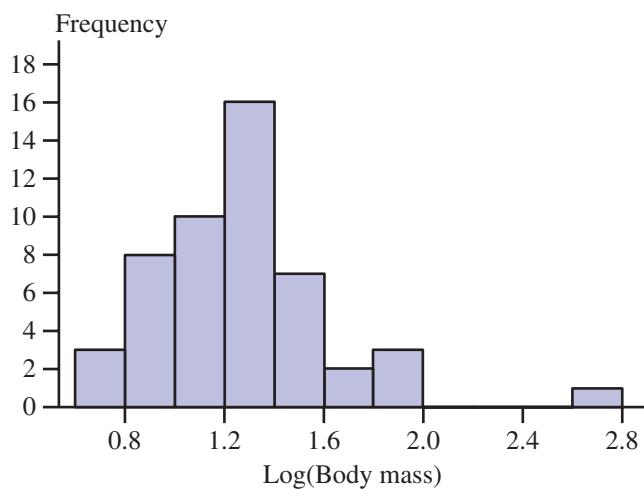
7.105



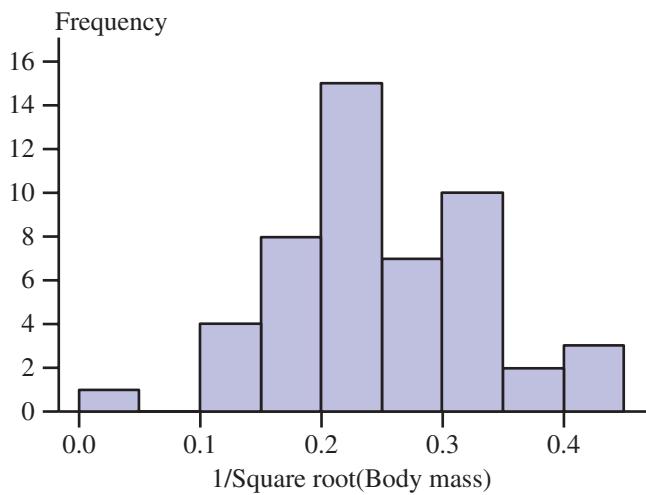
The cube-root transformation appears to result in the more symmetrical histogram.

7.107 a.

Yes. If a symmetrical distribution is required, a transformation is desirable.

b.

The distribution is closer to being symmetrical than the original distribution but is still positively skewed.

c.

The shape in the histogram roughly resembles that of a normal curve. Certainly, this transformation was more successful than the one in Part (b) in producing an approximately normal distribution.

7.109 a. 0.0240 b. 0.7580 c. 0.7357 d. 0.9108

7.111 a. 0.0233 b. 0.6536 c. 0.6585

7.113 a. 0.1114 b. 0.0409 c. 0.0968 d. 0.9429 e. 0.9001

7.115 a. When $p = 0.96$, $n(1 - p) = 60(0.04) = 2.4$, which is less than 10, and so the normal approximation to the binomial distribution should not be used. The binomial formula should be used. b. For a person who is not faking the test, a score of 42 or less is extremely unlikely (a probability of 0.000000000013). Therefore, if someone does get a score of 42 or less, we would be convinced that the person is faking the test.

7.117 a. 0.7970 b. 0.7016 c. 0.0143 d. 0.05

7.119 a. No, since $np = 50(0.05) = 2.5 < 10$. b. Now

$n = 500$ and $p = 0.05$, so $np = 500(0.05) = 25 \geq 10$.

So the techniques of this section can be used

$$P(\text{at least 20 are defective}) \approx 0.8708.$$

7.121 a. 0.8247 b. 0.0516 c. 0.68826 d. Yes, the probability of a pregnancy having a duration of at least 310 days is only 0.0030.

7.123 a. 0.7 b. 0.45 c. 0.55

7.125 a. $\mu_x = 2.64$ $\sigma_x^2 = 2.3704$. $\sigma_x = 1.540$

$$\mathbf{b. } \mu_x - 3\sigma_x = 2.64 - 3(1.540) = -1.98$$

$$\mu_x + 3\sigma_x = 2.64 + 3(1.540) = 7.26$$

Since all the possible values of x lie between these two values, the required probability is 0.

7.127 a. 0.1359 b. 0.0228 c. 0.595

7.129 a. If the coin is fair, then the distribution of x is binomial with $n = 25$ and $p = 0.5$. The probability that it is judged to be biased is $P(x \leq 7) + P(x \geq 18)$.

$$P(x \leq 7) = 0.014 + 0.005 + 0.002 = 0.021.$$

$$P(x \geq 18) = 0.014 + 0.005 + 0.002 = 0.021.$$

So the probability that the coin is judged to be biased is $0.021 + 0.021 = 0.042$. b. Now the distribution of x is binomial with $n = 25$ and $p = 0.9$. The probability that the coin is judged to be biased is $P(x \leq 7) + P(x \geq 18)$.

$$P(x \leq 7) = 0.000. P(x \geq 18) = 0.997.$$

So the probability that the coin is judged to be fair is $1 - 0.997 = 0.003$. When $P(H) = 0.1$, the distribution of x is binomial with $n = 25$ and $p = 0.1$. The probability that the coin is judged to be biased is $P(x \leq 7) + P(x \geq 18)$. $P(x \leq 7) = 0.998$.

$$P(x \geq 18) = 0.000.$$

So the probability that the coin is judged to be fair is $1 - 0.998 = 0.002$. c. When

$P(H) = 0.6$, the distribution of x is binomial with $n = 25$ and $p = 0.6$. The probability that the coin is judged to be biased is $P(x \leq 7) + P(x \geq 18)$. $P(x \leq 7) = 0.001$.

$$P(x \geq 18) = 0.153.$$

So the probability that the coin is judged to be fair is $1 - 0.153 = 0.847$. By symmetry,

when $P(H) = 0.4$, the probability that the coin is judged to be fair is 0.847. These probabilities are larger than

those in Part (b), since, when $P(H) = 0.6$ or $P(H) = 0.4$,

the probability of getting a head is closer to 0.5 than when $P(H) = 0.9$ or $P(H) = 0.1$. Thus it is more likely that the coin will be judged to be fair when $P(H) = 0.6$ or $P(H) = 0.4$. **d.** It is now more likely that the coin will be judged to be fair, and so the error probabilities are increased. This would seem to make the rule less good; however, this rule makes it more likely that the coin will be judged to be fair when in fact it is fair.

7.131 **a.** 0.9332 **b.** 72.8 minutes **c.** \$60

7.133 $P(x < 4.9) = 0.0228$ $P(x \geq 5.2) = 0.0000$

7.135 **a.**

y	0	1	2	3	4
$p(y)$	0.0625	0.4375	0.3125	0.1250	0.0625

$$\mu_y = 1.6875$$

b.

y	0	1	2	3	4
$p(y)$	0.0256	0.3264	0.3456	0.1728	0.1296

$$\mu_y = 2.0544$$

c.

z	1	2	3	4
$p(z)$	0.1250	0.5000	0.2500	0.1250

$$\mu_z = 2.375$$

7.137 **a.** 0.1552 **b.** 0.2688

c.

x	4	5	6	7
$p(x)$	0.1552	0.2688	0.29952	0.27648

$$\mu_x = 5.697$$

7.139 **a.** 0.7745 **b.** 0.1587 **c.** 0.3085 **d.** 6.1645

7.141 **a.** No. The proportion of the population that is being sampled is $5000/40000 = 0.125$, which is more than 5%.

b. $\mu = np = 100(0.275) = 27.5$ and

$\sigma = \sqrt{np(1-p)} = 4.465$. **c.** No. Since n is being doubled, the standard deviation, $\sqrt{np(1-p)}$, is multiplied by $\sqrt{2}$.

CUMULATIVE REVIEW EXERCISES 7

CR7.3 No. The percentages given in the graph are said to be, for each year, the “percent increase in the number of communities installing” red-light cameras. This presumably means the percentage increase in the number of communities with red-light cameras *installed*, in which case the positive results for all of the years 2003 to 2009 show that a great many more communities had red-light cameras installed in 2009 than in 2002.

CR7.5 0.1243

CR7.7 $P(\text{Service 1|Late}) = 0.349$ and

$P(\text{Service 2|Late}) = 0.651$. Service 2 is more likely to have been used.

CR7.9 1. $P(M) = 0.021$ 2. $P(M|B) = 0.043$

3. $P(M|W) = 0.07$

CR7.11 statement iv

CR7.13 **a.** 0.09 **b.** 0.045 **c.** 0.135 **d.** 0.667

CR7.15 **a.** $\mu_x = 2.3$. **b.** $\sigma_x^2 = 0.81$, $\sigma_x = 0.9$

CR7.17 Let x be the number of correct identifications. Assume that the graphologist was merely guessing—in other words, that the probability of success on each trial was 0.5. Then $P(x \geq 6) = 0.377$. Since this probability is not particularly small, it would not have been unlikely for him/her to get six or more correct identifications when guessing. No ability to distinguish the handwriting of psychotics is indicated.

CR7.19 **a.** 0.1587 **b.** 51.4 minutes **c.** 41.65 minutes

CR7.21 **b.** Positive **c.** Yes. If a symmetrical distribution were required, then a transformation would be advisable.

Chapter 8

8.1 A population characteristic is a quantity that summarizes the whole population. A statistic is a quantity calculated from the values in a sample.

8.3 **a.** Population characteristic **b.** Statistic **c.** Population characteristic **d.** Statistic **e.** Statistic

8.7 **a.** and **b.**

\bar{x}	1.5	2	2.5	3	3.5
$p(\bar{x})$	1/6	1/6	1/3	1/6	1/6

c.

\bar{x}	1	1.5	2	2.5	3	3.5	4
$p(\bar{x})$	1/16	1/8	3/16	1/4	3/16	1/8	1/16

d. Both distributions are symmetrical, and their means are equal (2.5). However, the “with replacement” version has a greater spread than the first distribution, with values ranging from 1 to 4 in the “with replacement” distribution and from 1.5 to 3.5 in the “without replacement” distribution. The stepped pattern of the “with replacement” distribution more closely resembles a normal distribution than does the shape of the “without replacement” distribution.

8.9 **a.** and **b.**

\bar{x}	$2\frac{2}{3}$	3	$3\frac{1}{3}$	$3\frac{2}{3}$
$p(\bar{x})$	0.1	0.4	0.3	0.2

Sample Median	3	4
$p(\text{Sample Median})$	0.7	0.3

$(\text{Max} + \text{Min})/2$	2.5	3	3.5
$p((\text{Max} + \text{Min})/2)$	0.1	0.5	0.4

c. Since $\mu = 3.2$ and $\mu_{\bar{x}} = 3.2$, \bar{x} is an unbiased estimator of μ , which is not the case for either of the two other statistics. Since the distribution of the sample mean has less variability than either of the other two sampling distributions, the sample mean will tend to produce values that are

closer to μ than the values produced by either of the other statistics.

8.11 The sampling distribution of \bar{x} will be approximately normal for the sample sizes in Parts (c)–(f), since those sample sizes are all greater than or equal to 30.

8.13 **a.** $\mu_{\bar{x}} = 40$, $\sigma_{\bar{x}} = 0.625$ approximately normal
b. 0.5762 **c.** 0.2628

8.15 **a.** $\mu_{\bar{x}} = 2$, $\sigma_{\bar{x}} = 0.267$ **b.** In each case $\mu_{\bar{x}} = 2$. When $n = 20$, $\sigma_{\bar{x}} = 0.179$, and when $n = 100$, $\sigma_{\bar{x}} = 0.008$. **c.** All three centers are the same, and the larger the sample size, the smaller the standard deviation of \bar{x} . Since the distribution of \bar{x} when $n = 100$ is the one with the smallest standard deviation of the three, this sample size is most likely to result in a value of \bar{x} close to μ .

8.17 **a.** 0.8185, 0.0013 **b.** 0.9772, 0.0000

8.19 $p(0.49 < \bar{x} < 0.51) = 0.9974$; the probability that the manufacturing line will be shut down unnecessarily is $1 - 0.9974 = 0.0026$.

8.21 Approximately 0.

8.23 **a.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.151$. **b.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.107$. **c.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.087$. **d.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.067$. **e.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.048$. **f.** $\mu_{\hat{p}} = 0.65$, $\sigma_{\hat{p}} = 0.034$.

8.25 **a.** Mean: $\mu_{\hat{p}} = 0.03$; Standard Deviation:

$$\sigma_{\hat{p}} = \sqrt{\frac{0.03(1 - 0.03)}{100}} = 0.017$$

b. The sampling distribution is not approximately normal because $np = (100)(0.03) = 3$ is less than the required 10. **c.** The change in sample size does not change the mean of the sampling distribution. However, the standard deviation will decrease to 0.009. The mean does not change when the sample size is increased because the sampling distribution is always centered at the population value (in this case, $\mu_{\hat{p}} = 0.03$) regardless of the sample size. The standard deviation of the sampling distribution will decrease as the sample size increases because the sample size (n) is in the denominator of the formula for standard deviation. As sample size increases, standard deviation of the sampling distribution of \hat{p} decreases. **d.** The sampling distribution of \hat{p} is approximately normal because $np = (400)(0.03) = 12$ and $n(1 - p) = (400)(1 - 0.03) = 388$. Both of these values are at least 10, so we have satisfied the rule of thumb.

8.27 **a.** $\mu_{\hat{p}} = 0.005$, $\sigma_{\hat{p}} = 0.007$ **b.** No, since $np = 0.5$, which is not greater than or equal to 10. **c.** We need both np and $n(1 - p)$ to be greater than or equal to 10; need $n \geq 2000$.

8.29 **a.** If $p = 0.8$, then $\mu_{\hat{p}} = 0.8$ and

$$\sigma_{\hat{p}} = \sqrt{\frac{0.8(1 - 0.8)}{225}} = 0.027.$$

b. If $p = 0.7$, then $\mu_{\hat{p}} = 0.7$ and

$$\sigma_{\hat{p}} = \sqrt{\frac{0.7(1 - 0.7)}{225}} = 0.031.$$

c. When $p = 0.8$, $np = 225(0.8) = 180$ and $n(1 - p) = 225(1 - 0.8) = 45$, which are both greater than 10. When $p = 0.7$, $np = 225(0.7) = 157.5$ and $n(1 - p) = 225(1 - 0.7) = 67.5$, which are also both greater than 10. In each case, np and $n(1 - p)$ are both greater than 10, so the sampling distribution of \hat{p} is approximately normal for $p = 0.8$ as well as $p = 0.7$.

8.31 **a.** 0.9744 **b.** Approximately 0

8.33 **a.** \bar{x} is approximately normally distributed with mean 50 and standard deviation 0.1. **b.** 0.9876 **c.** 0.5

8.35 **a.** 0.8185 **b.** **i.** 0.8357 **ii.** 0.9992

8.37 0.0793

Chapter 9

9.1 Statistics II and III are preferable to Statistic I since they are unbiased (their means are equal to the value of the population characteristic). However, Statistic II is preferable to Statistic III since its standard deviation is smaller. So Statistic II should be recommended.

9.3 $\hat{p} = \frac{584}{1,668} = 0.350$

9.5 $\hat{p} = 0.7$

9.7 **a.** $\bar{x} = 421.429$ **b.** $s^2 = 10414.286$ **c.** $s = 102.050$. No, s is not an unbiased statistic for estimating σ .

9.9 **a.** $\bar{x} = 120.6$ therms **b.** The value of τ is estimated to be $10000(120.6) = 1,206,000$ therms. **c.** $\hat{p} = 0.8$ **d.** sample median = 120 therms

9.11 The sample of size $n = 100$ is likely to result in a wider confidence interval. The formula for the confidence interval is $\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$, and so a smaller value of n is likely to result in a wider confidence interval since n is in the denominator.

9.13 **a.** Yes, since $np = 50(0.3) = 15 \geq 10$ and $n(1 - p) = 50(0.7) = 35 \geq 10$. **b.** No, since $np = 50(0.05) = 2.5$, which is not greater than or equal to 10. **c.** No, since $np = 15(0.45) = 6.75$, which is not greater than or equal to 10. **d.** No, since $np = 100(0.01) = 1$, which is not greater than or equal to 10.

9.15 **a.** The larger the confidence level, the wider the interval. **b.** The larger the sample size, the narrower the interval. **c.** Values of \hat{p} farther from 0.5 give smaller values of $\hat{p}(1 - \hat{p})$. Therefore, the farther the value of \hat{p} from 0.5, the narrower the interval.

9.17 If a large number of random samples of size 1200 were to be taken, 90% of the resulting confidence intervals

would contain the true proportion of all Facebook users who would say it is not OK to “friend” someone who reports to you at work.

9.19 **a.** (0.675, 0.705). **b.** We are 98% confident that the proportion of all coastal residents who would evacuate is between 0.675 and 0.705. If we were to take a large number of random samples of size 5046, about 98% of the resulting confidence intervals would contain the true proportion of all coastal residents who would evacuate.

9.21 **a.** (0.625, 0.675). We are 90% confident that the proportion of dog owners who take more pictures of their dog than of their significant others or friends is between 0.625 and 0.675. **b.** (0.429, 0.491). We are 95% confident that the actual proportion of dog owners who are more likely to complain to their dog than to a friend is between 0.429 and 0.491. **c.** First, the confidence interval in Part (b) has a higher confidence level, which uses a larger z critical value in the computation of the confidence interval. Second, the standard error in Part (b) is larger than that in Part (a). Both of these reasons contribute to the wider confidence interval in Part (b).

9.23 (0.799, 0.861). We are 99% confident that the actual proportion of mothers of children under the age of 2 years who posted pictures of their new baby on social media from the delivery room is between 0.799 and 0.861.

9.25 (0.821, 0.859). We are 95% confident that the proportion of all adult Americans who prefer cheese on their burgers is between 0.821 and 0.859.

9.27 (0.362, 0.418). We are 90% confident that the proportion of all millennials who drive a car that is more than 5 years old who have named their cars is between 0.362 and 0.418.

9.29 It is unlikely that the estimated proportion of adult Americans who have pretended to know what the cloud is or how it works ($\hat{p} = 0.22$) will differ from the true proportion by more than 0.026.

9.31 **a.** (0.119, 0.286). We are 95% confident that the proportion of all patients under 50 years old who experience a failure within the first 2 years after receiving this type of defibrillator is between 0.119 and 0.286. **b.** (0.011, 0.061). We are 99% confident that the proportion of all patients age 50 or older who experience a failure within the first 2 years after receiving this type of defibrillator is between 0.011 and 0.061. **c.** Using the estimate of p from the study, 18/89, the required sample size is $n = 688.685$. A sample of size at least 689 is required.

9.33 A sample size of 2401 is required.

9.35 **a.** Assuming a 95% confidence level, and using the preliminary estimate of $p = 0.63$, the sample should contain 359 individuals. Using the conservative estimate of $p = 0.5$, the sample should contain 385 individuals.

b. The sample size computed using the conservative estimate is larger than the sample size computed using the preliminary estimate. The sample size 385 should be used for this study because it will guarantee a margin of error of

no greater than 0.05 because the margin of error is largest with $p = 0.5$.

9.37 **a.** 2.12 **b.** 1.80 **c.** 2.81 **d.** 1.71 **e.** 1.78 **f.** 2.26

9.39 The second interval is based on the larger sample size; the interval is narrower.

9.41 **a.** If the distribution of volunteer times is approximately normal, for the sample standard deviation of $s = 16.54$ hours and the sample mean of $\bar{x} = 14.76$ hours, approximately 18.6% of volunteer times would be negative, which is impossible. Therefore, it is not reasonable to think that the distribution of volunteer times is approximately normal. **b.** The two conditions required to use a one-sample t distribution are that the sample was either randomly selected or is representative of the population, and the population distribution is either normally distributed or that the sample size is at least 30. We are told that the sample is representative of the population, and that the sample size is 500. Both conditions are satisfied in the exercise. **c.** (13.307, 16.213). We are 95% confident that the mean number of hours spent in volunteer activities per year for South Korean middle school children is between 13.307 and 16.213 hours.

9.43 **a.** (35.424, 38.616). We are 95% confident that the mean procrastination scale for first-year students at this college is between 35.424 and 38.616. **b.** Forty (40) is not a plausible value for the mean population procrastination scale score because that value is not contained within the confidence interval. It seems plausible that, on average, students at this university do not have high levels of procrastination.

9.45 **a.** Narrower. **b.** The statement is not correct. The population mean, μ , is a constant, and therefore, we cannot talk about the probability that it falls within a certain interval. **c.** The statement is not correct. We can say that *on average* 95 out of every 100 samples will result in confidence intervals that will contain μ , but we cannot say that in 100 such samples, *exactly* 95 will result in confidence intervals that contain μ .

9.47 **a.** The samples from 12 to 23 month and 24 to 35 month are the ones with the greater variability. **b.** The less-than-12-month sample is the one with the greater sample size. **c.** The new interval has a 99% confidence level.

9.49 **a.** (179.02, 186.98). We are 95% confident that the mean summer weight is between 179.02 and 186.98 pounds. **b.** (185.423, 194.577). We are 95% confident that the mean winter weight is between 185.423 and 194.577 pounds. **c.** Based on the Frontier Airlines data, neither recommendation is likely to be an accurate estimate of the mean passenger weight, since 190 is not contained in the confidence interval for the mean summer weight and 195 is not contained in the confidence interval for the mean winter weight.

9.51 It is not reasonable to construct a confidence interval for the mean mileage rating of 2016 midsize hybrid

cars. There is no indication that the cars were randomly selected, which is the first required condition. The second required condition of approximate normality of the population of mileage ratings of the cars has been satisfied. A normal probability plot of the data shows some curvature.

9.53 A reasonable estimate of σ is given by (sample range)/4 = 162.5. A sample size of 1015 is needed.

9.55 First, we need to know that the information is based on a random sample of middle-income consumers age 65 and older. Second, it would be useful if some sort of margin of error was given for the estimated mean of \$10,235.

9.57 **a.** The paper states that queens flew for an *average* of 24.2 ± 9.21 minutes on their mating flights, and so this interval is a confidence interval for a population mean.

b. (3.301, 5.899)

9.59 **a.** No, it is not appropriate to use the large-sample confidence interval for a population proportion to estimate the proportion of the transportation services customers who have tried Uber or Lyft at least once. We don't have at least 10 successes in the sample, which is one of the necessary conditions. **b.** Yes, it is appropriate to use a bootstrap confidence interval for a population proportion to estimate the proportion of the transportation services customers who have tried Uber or Lyft because a random sample of regular customers who are 55 or older was taken. **c.** A bootstrap 95% confidence interval for the population proportion of customers 55 or older who have used Uber or Lyft at least once, p , is (0.000, 0.333). Based on this sample, you can be 95% confident that the actual proportion of customers 55 or older who have used Uber or Lyft at least once is somewhere between 0.000 and 0.333. **d.** Yes, the value obtained in the study (21%) is contained within the bootstrap confidence interval. Confidence intervals provide a range of plausible values for the true proportion. The interval contains 0.21, so we know that 0.21 is a plausible value for the true proportion of customers 55 or older who have used Uber or Lyft at least once.

9.61 **a.** It would not be appropriate to use a large-sample confidence interval for one proportion to estimate Kevin Love's success rate for three-point shots during the 2016 season because there are only 5 successes in the sample, which is fewer than the required 10 successes to use the large-sample interval. **b.** Different simulations will produce different results, so answers will vary. For one simulation, a bootstrap 90% confidence interval for the population proportion of Kevin Love's three-point shot success rate during the 2016 NBA season, p , was (0.105, 0.421). Based on this sample, you can be 90% confident that the actual proportion of Kevin Love's three-point shot success rate during the 2016 NBA season is somewhere between 0.105 and 0.421.

9.63 **a.** Yes, it would be appropriate to use the large-sample confidence interval for a population proportion to estimate the proportion of residential properties successfully sold at auction in the post-Brexit UK. The necessary

conditions (a representative sample of residential properties and at least 10 successes and failures in the sample) have both been satisfied. **b.** It is appropriate to use a bootstrap confidence interval because we have a representative sample of the population. **c.** A bootstrap 95% confidence interval for the population proportion of residential properties successfully sold at auction in the post-Brexit UK, p , is (0.346, 0.731). Based on this sample, you can be 95% confident that the actual proportion of residential properties successfully sold at auction in the post-Brexit UK is somewhere between 0.346 and 0.731. **d.** Yes, the success rate for properties sold at auction throughout the UK during one stretch a year earlier (0.72) is contained within the bootstrap confidence interval.

9.65 Different simulations will produce different results, so answers will vary. For one simulation, a bootstrap 95% confidence interval for the population proportion of all brand-name products that are active on Snapchat, p , is (0.617, 0.783). Based on this sample, you can be 95% confident that the actual proportion of all brand-name products that are active on Snapchat is somewhere between 0.617 and 0.783.

9.67 **a.** The sample size of $n = 21$ is smaller than 30, so the methods based on the t distribution may not be appropriate. **b.** The given simulation produced a confidence interval of (0.994, 1.082). We are 95% confident that the population mean time discrimination score for the population of male smokers who abstain from smoking for 24 hours lies somewhere between 0.994 and 1.082.

9.69 The given simulation produced a confidence interval of (4.579, 5.754). We are 95% confident that the population mean annual energy cost for the population of all small televisions lies somewhere between \$4.579 and \$5.754.

9.71 **a.** The sample size of $n = 15$ is smaller than 30, so the methods based on the t distribution may not be appropriate. **b.** Different simulations will produce different results, so answers will vary. One simulation produced a confidence interval of (38, 52). We are 95% confident that the mean interleague winning percentage for National League teams lies somewhere between 38% and 52%. **c.** In Part (b), the mean winning percentage for National League teams was estimated (with 95% confidence) to be between 38% and 52%, which indicates that 50% is a plausible value. Therefore, it is not reasonable to say that the National League or American League performs significantly better than the other in interleague play.

9.73 Intervals constructed when the sample proportions (\hat{p}) are closer to 0.5 are wider than those farther from 0.5 because the product $\hat{p}(1 - \hat{p})$ is largest when $\hat{p} = 0.5$. The sample proportion in this exercise is closer to 0.5 than that in the previous exercise. Therefore, the interval in this exercise will be wider than the interval in the previous exercise.

9.75 (0.477, 0.563). We are 90% confident that the true proportion of all homeowners in western United States who have considered installing solar panels is between 0.477 and 0.563.

9.77 Assuming a 95% confidence level, and using $p = 0.32$, a sample size of 335 is required. Using the conservative value, $p = 0.5$, a sample size of 385 is required. The conservative estimate of p gives the larger sample size. Since the relevant proportion could have changed significantly since 2011, it would be sensible to use a sample size of 385.

9.79 A sample size of 246 is required.

9.81 (8.571, 9429)

Chapter 10

10.1 \bar{x} is a *sample* statistic.

10.3 $H_a: \mu > 100$ will be used.

10.5 In this case, consider the population of youths with preexisting psychological conditions, and the following hypotheses: H_0 : Video games do not increase antisocial behavior *versus* H_a : Video games do increase antisocial behavior. The article states that the researchers found no evidence “that violent video games increase antisocial behavior in youths with pre-existing psychological conditions.” Finding no evidence simply means that the data did not provide strong support for the alternative, not that the null hypothesis is true. Therefore, the title is misleading in that it is implying that the null hypothesis is true.

10.7 $H_0: p = 0.5, H_a: p > 0.5$

10.9 $H_0: p = 0.5, H_a: p > 0.5$

10.11 $H_0: p = 0.7, H_a: p \neq 0.7$

10.13 **a.** This is a Type II error because the statement describes the result of failing to reject the null hypothesis when the null hypothesis is actually false (concluding that the woman does not have breast cancer when, in actuality, she does). This probability is approximately

$$P(\text{Type II error}) \approx \frac{1}{13} = 0.077. \quad \text{b.}$$

The other error that is possible is a Type I error, in which a true null hypothesis is rejected. In this scenario, a Type I error is concluding that a woman has cancer when she really does not have cancer. This probability is approximately

$$P(\text{Type I error}) \approx \frac{90}{637} = 0.141.$$

10.15 **a.** A Type I error is when it is concluded that a particular man is the father when, in fact, he is not the father. A Type II error is when it is concluded that a particular man is not the father when, in fact, he is the father.

b. $\alpha = 0$ and $\beta = 0.0001$

10.17 **a.** A Type I error concluding that there is evidence that more than 1% of a shipment is defective when in fact

(at least) 1% of the shipment is defective. A Type II error is not being convinced that more than 1% of a shipment is defective when in fact more than 1% of the shipment is defective. **b.** Type II **c.** Type I

10.19 The probability of a Type I error is equal to the significance level. Here the aim is to reduce the probability of a Type I error, so a small significance level (such as 0.01) should be used.

10.21 **a.** The phrase “No evidence of increased risk of thyroid cancer in areas that were near a nuclear power plant” indicates that the researchers failed to reject the null hypothesis. **b.** If the researchers are incorrect in their conclusion, then they would have failed to reject H_0 when H_0 was actually false. This is a Type II error. **c.** No. The study did not provide sufficient evidence to reject the null hypothesis. In other words, there was not convincing evidence that the proportion of the population in areas near nuclear power plants who are diagnosed with thyroid cancer during a given year is greater than the proportion of the population in areas where no nuclear power plants are present. This does not mean that there was no effect, only that the data did not provide evidence to reject the null hypothesis.

10.23 **a.** A *P*-value of 0.0003 means that it is very unlikely (probability = 0.0003), assuming that H_0 is true, that you would get a sample result at least as inconsistent with H_0 as the one obtained in the study. Thus, H_0 is rejected.

b. A *P*-value of 0.350 means that it is not particularly unlikely (probability = 0.350), assuming that H_0 is true, that you would get a sample result at least as inconsistent with H_0 as the one obtained in the study. Thus, there is no reason to reject H_0 .

10.25 **a.** H_0 is not rejected. **b.** H_0 is not rejected. **c.** H_0 is not rejected.

10.27 **a.** 0.081 **b.** 0.176 **c.** 0.025 **d.** 0.007 **e.** 0.567

10.29 $H_0: p = 0.1, H_a: p > 0.1, z = 1.647, P\text{-value} = 0.0498$, fail to reject H_0 .

10.31 **a.** $H_0: p = 0.4, H_a: p < 0.4, z = -2.969, P\text{-value} = 0.001$, reject H_0 . **b.** $H_0: p = 1/3, H_a: p > 1/3, z = 2.996, P\text{-value} = 0.001$, reject H_0 .

10.33 $H_0: p = 0.5, H_a: p < 0.5, z = -2.536, P\text{-value} = 0.006$, reject H_0 .

10.35 $H_0: p = 0.69$ versus $H_a: p \neq 0.69$, where p = proportion of teens at the high school who access social media from a mobile phone

10.37 $H_0: p = 0.87, H_a: p \neq 0.87$

10.39 **a.** Here $np = 728(0.25) = 182 \geq 10$ and $n(1-p) = 728(0.75) = 546 \geq 10$, so the sampling distribution of \hat{p} is approximately normal. The sampling distribution of \hat{p} has mean $p = 0.25$ and standard deviation $\sqrt{p(1-p)/n} = \sqrt{(0.25)(0.75)/728} = 0.016$. **b.** 0.106; this probability is not particularly small, so it would not be particularly surprising to observe a sample proportion as large as $\hat{p} = 0.27$ if the null hypothesis were true.

c. 0.0001; this probability is small, so it would be surprising to observe a sample proportion as large as $\hat{p} = 0.31$ if the null hypothesis were true.

10.41 **a.** $H_0: p = 0.6, H_a: p > 0.6, z = 2.178, P\text{-value} = 0.015$, fail to reject H_0 .

b. $H_0: p = 0.5, H_a: p > 0.5, z = 5.621, P\text{-value} \approx 0$, reject H_0

10.43 **a.** 0.040 **b.** 0.003 **c.** 0.019 **d.** 0.000

10.45 **a.** 0.484 **b.** 0.686 **c.** 0.025 **d.** 0.000 **e.** 0.097

10.47 **a.** H_0 is rejected. **b.** H_0 is not rejected. **c.** H_0 is not rejected.

10.49 **a.** $t = 0.748, P\text{-value} = 0.468$, fail to reject H_0

b. $t = 8.731, P\text{-value} \approx 0$, reject H_0

10.51 **a.** No. The study was conducted in New York City only, and therefore the results cannot be generalized to the lunchtime fast-food purchases of all adult Americans.

b. If you ask the customers what they purchased, some customers might misremember or might give false answers. By looking at the receipt, you know that you are receiving an accurate response. **c.** Yes. It is possible that knowing that a record of their lunch order was going to be seen might have influenced the amount of food customers ordered. For example, this knowledge might cause customers to order less, out of the fear of embarrassment at people seeing the sizes of their orders.

10.53 $H_0: \mu = 5.04, H_a: \mu < 5.04, t = -6.36, P\text{-value} \approx 0$, reject H_0 .

10.55 **a.** The fact that the sample standard deviation is greater than the sample mean indicates that there is a quite a bit of variability in the distribution of purchase amounts. In addition, since purchase amounts cannot be negative, the larger standard deviation indicates that the distribution is positively skewed. **b.** $H_0: \mu = 2.80, H_a: \mu > 2.80, t = 0.857, P\text{-value} = 0.196$, fail to reject H_0 .

10.57 Since the sample was large, it was possible for the hypothesis test to provide convincing evidence that the mean score for the population of children who spent long hours in child care was greater than the mean score for third graders in general, even though the obtained sample mean didn't differ greatly from the known mean for third graders in general.

10.59 **a.** $H_0: \mu = 120, H_a: \mu > 120, t = 0.649, P\text{-value} = 0.258$, fail to reject H_0 . **b.** $H_0: \mu = 120, H_a: \mu > 120, t = 2.049, P\text{-value} = 0.020$, reject H_0 . **c.** The sample standard deviation of 117.1 minutes in Part (a) indicates that there is likely to be more variability in the number of minutes of video or computer game playing than when the sample standard deviation was 37.1 minutes, as in Part (b). If the mean of the population of average time spent playing video or computer games for male Canadian high school students was indeed 120 minutes, it is less likely that a sample mean of 123.4 minutes would be obtained when the standard deviation was 37.1 minutes than when the standard deviation was 117.1 minutes. The larger standard deviation

in Part (a) makes it more difficult to detect a difference, resulting in a failure to reject the null hypothesis.

10.61 **a.** Yes. Since the pattern in the normal probability plot is roughly linear, and since the sample was a random sample from the population, the t test is appropriate.

b. The boxplot shows a median of around 245, and since the distribution is a roughly symmetrical distribution, this tells us that the sample mean is also around 245. This might initially suggest that the population mean differs from 240. But, the sample is relatively small, and the sample values range all the way from 225 to 265; such a sample mean would still be feasible if the population mean were 240. **c.** $t = 1.212, P\text{-value} = 0.251$, fail to reject H_0

10.63 **a.** Increasing the sample size increases the power.

b. Increasing the significance level increases the power.

10.65 **a.** A Type II error occurs when H_0 is not rejected, and, as shown in the sketch, H_0 will not be rejected for values of \bar{x} less than 152.546. **b.** $\beta \approx 0$ **c.** A Type II error could have occurred.

10.67 **a.** $P\text{-value} = 0.044$. H_0 is rejected, and we have convincing evidence that more than 75% of apartments exclude children. **b.** 0.351

10.69 **a. i.** $\beta \approx 0.84; \beta \approx 0.84$ **ii.** $\beta \approx 0.57$ **iii.** $\beta \approx 0.10$ **iv.** $\beta \approx 0$ **b.** When σ increases, each d value decreases, and, looking at the graphs in Appendix Table 5, we see that each value of β will be greater than in Part (a).

10.71 **a.** $\beta \approx 0$ **b.** $\beta \approx 0.04$ **c.** $\beta \approx 0.01$

10.73 **a.** $z = 3.370, P\text{-value} = 0.0004$, reject H_0

10.75 **a.** $H_0: p = 0.20, H_a: p < 0.20$ **b.** It is not stated that the sample was randomly selected or that the sample is representative of the population of all hospital patients who had been treated for pneumonia using a respiratory therapist protocol. As such, we must use caution and assume the sample was selected in a reasonable way. The large sample size condition has been satisfied because there are at least 10 successes and 10 failures in the sample. **c.** The output for the exact binomial test indicates that the P -value is 0.000. Since the P -value of 0.000 is less than any reasonable significance level, we reject the null hypothesis. We have sufficient evidence to conclude that the proportion of subjects who will be readmitted to a hospital within 30 days after following a respiratory therapist protocol for treatment of pneumonia is less than 0.20. **d.** $P\text{-value} = 0.003$, the same conclusion will be drawn.

10.77 **a.** These data should not be analyzed using a large-sample hypothesis test for one population proportion because the number of successes and failures are not both at least 10. In this case, the number of successes is 17, and the number of failures is 8, which is less than 10.

b. Because the P -value of 0.000 is less than any reasonable significance level, we reject the null hypothesis. There is sufficient evidence to conclude that the proportion of people who quit smoking prior to surgery who would smoke

again after 1 year is greater than 0.10. **c.** The patients who were smokers were required to quit for their surgery, but maybe they quit against their will. These patients perhaps never wanted to quit and so began smoking again once they were allowed to do so.

10.79 **a.** $H_0: p = 0.50$, $H_a: p > 0.50$, P -value of 0.007. Because the P -value of 0.007 is less than any reasonable significance level, we reject the null hypothesis. We have sufficient evidence to conclude that the proportion of coin flip calls that the Patriots win is greater than 0.50. **b.** In order to carry out the exact binomial version of the hypothesis test, there are four necessary conditions: (1) There is a fixed number of trials; (2) each trial can result in one of only two possible outcomes; (3) outcomes of different trials are independent; and (4) the success probability is the same for each trial. In this case, all the conditions have been satisfied. Rejecting the null hypothesis does not mean that the alternative hypothesis is true. Rather, it means we have evidence that supports the alternative hypothesis based on the data we collected. It could be that we incorrectly rejected the null hypothesis based on this sample (which is a Type I error).

10.81 **a.** The sample size of $n = 21$ is smaller than 30, so the methods based on the t distribution may not be appropriate. **b.** The given randomization test resulted in a P -value of 0.029. Because this P -value of 0.029 is less than a significance level of 0.05, we reject the null hypothesis. We have convincing evidence that the population mean time discrimination score for male smokers who abstain from smoking for 24 hours is significantly greater than 1.

10.83 **a.** The given randomization test resulted in a P -value of 0.000. Because this P -value of 0.000 is less than a significance level of 0.05, we reject the null hypothesis. We have convincing evidence that the population mean 2000-meter ergometer time for U.S. junior male sculls rowers differs from the 2007 international standard of 387 seconds. **b.** Based on the result of the hypothesis test, we can say that there is evidence to conclude that the mean 2000-meter ergometer time for U.S. junior male sculls rowers differs from the 2007 international standard of 387 seconds. The sample mean time of 394.25 seconds, given in the Shiny App output, is greater than 387 seconds. Therefore, since the null hypothesis was rejected and the sample mean is greater than the hypothesized mean, we can conclude that the U.S. junior male sculls rowers seem to have “caught up,” on average, with the international championship rowers from 2007.

10.85 **a.** The fact that the sample size ($n = 15$) is smaller than 30 indicates that the methods based on the t distribution might not be appropriate. **b.** Different simulations will produce different results, so answers will vary. One randomization test resulted in a P -value of 0.170. Because this P -value of 0.170 is greater than a 5% significance level, we fail to reject the null hypothesis. We do not have convincing evidence that the population mean interleague winning percentage for National League teams differs from 50%.

c. In Part (b), the null hypothesis was not rejected, which indicates that there is insufficient evidence to conclude that the mean winning percentage for National League teams differs from 50%. Therefore, it is not reasonable to say that the National League or American League performs significantly better than the other in interleague play.

10.87 $z = -11.671$, P -value ≈ 0 , reject H_0

10.89 $H_0: \mu = 7.85$, $H_a: \mu > 7.85$, $t = 3.28$, P -value = 0.0013, reject H_0 .

10.91 **a.** We would need to assume that the sample was randomly selected from the population of white-collar workers in the United States or that the sample is representative of the population. **b.** $H_0: \mu = 4$, $H_a: \mu > 4$, $t = 2.437$, P -value = 0.0075, reject H_0 .

10.93 $H_0: p = 0.25$, $H_a: p > 0.25$, $z = 1.03$, P -value = 0.15, fail to reject H_0 .

10.95 $z = 1.069$, P -value = 0.143, fail to reject H_0

10.97 $z = 4.186$, P -value ≈ 0 , reject H_0

10.99 $t = -5.324$, P -value ≈ 0 , reject H_0

CUMULATIVE REVIEW EXERCISES 10

CR10.3 **a.** Three airlines stand out from the rest and have large numbers of delayed flights. These airlines are ExpressJet, Delta, and Continental, with 93, 81, and 72 delayed flights, respectively. **b.** A typical number of flights delayed per 100,000 flights is around 1.1, with most rates lying between 0 and 1.6. Four airlines stand out from the rest and have high rates, with two of those four having *particularly* high rates. **c.** The rate per 100,000 flights data should be used, since this measures the likelihood of any given flight being late. An airline could stand out in the number of flights delayed data purely as a result of having a large number of flights.

CR10.5 **a.** 0.134 **b.** 0.041 **c.** $\mu_x = 25$, $\sigma_x = 4.330$ **d.** 0.102

CR10.7 **a.** 0.4 **b.** 0.18 **c.** 0.26 **d.** 0.45 **e.** Since anyone who accepts a job offer must have received at least one job offer, $P(O|A) = 1$. **f.** 0.18

CR10.9 **a.** (0.244, 0.416). We are 95% confident that the proportion of all U.S. medical residents who work moonlighting jobs is between 0.244 and 0.416. **b.** (0.131, 0.252). We are 90% confident that the proportion of all U.S. medical residents who have credit card debt of more than \$3000 is between 0.131 and 0.252.

c. The interval in Part (a) is wider than the interval in Part (b) because the confidence level in Part (a) (95%) is greater than the confidence level in Part (b) (90%) and because the sample proportion in Part (a) (38/115) is closer to 0.5 than the sample proportion in Part (b) (22/115).

CR10.11 A reasonable estimate of σ is given by $(\text{sample range})/4 = 0.1$. A sample size of 385 is needed.

CR10.13 $z = 7.557$, P -value ≈ 0 , reject H_0

CR10.15 **a.** With a sample mean of 14.6, the sample standard deviation of 11.6 places zero just over one standard deviation below the mean. Since no teenager can spend a negative time online, to get a typical deviation from the mean of just over 1, there must be values that are

substantially more than one standard deviation above the mean. This suggests that the distribution of online times in the sample is positively skewed. **b.** $t = 9.164$, $P\text{-value} \approx 0$, reject H_0

Chapter 11

11.1 The distribution of $\bar{x}_1 - \bar{x}_2$ is approximately normal with mean 5 and standard deviation 0.529.

11.3 $H_0: \mu_1 - \mu_2 = 0$, $H_a: \mu_1 - \mu_2 \neq 0$, $t = -4.536$, $df = 71$, $P\text{-value} \approx 0$, reject H_0 . There is convincing evidence that the mean reading level for health-related pages differs for Wikipedia and WebMD.

11.5 **a.** Since boxplots are roughly symmetrical and since there is no outlier in either sample, the assumption of normality is justified, and it is reasonable to carry out a two-sample t test. **b.** $t = 3.332$, $P\text{-value} = 0.001$, reject H_0 **c.** (0.423, 2.910). We are 98% confident that the difference between the mean number of hours per day spent using electronic media in 2009 and 1999 is between 0.423 and 2.910.

11.7 **a.** (7.206, 20.794). We can be 95% confident that the difference in the mean amount of sleep on a work night for adults in Canada and adults in England is between 7.206 and 20.794 minutes. **b.** The confidence interval allows us to conclude that there is evidence of a difference in the mean amount of sleep on a work night for the two countries because zero is not contained in the confidence interval.

11.9 **a.** $H_0: \mu_1 - \mu_2 = 0$, $H_a: \mu_1 - \mu_2 \neq 0$, $t = 4.024$, $df = 112$, $P\text{-value} \approx 0.0001$, reject H_0 . There is convincing evidence that the mean time spent on Facebook is not the same for males and for females. **b.** It is not reasonable to generalize the conclusion from the hypothesis test in Part (a) to all male and all female college students in the United States because the sample was taken from one large university in Southern California and is therefore not representative of all college students in the United States.

11.11 $H_0: \mu_1 - \mu_2 = 0$, $H_a: \mu_1 - \mu_2 > 0$, $t = 0.486$, $df = 81$, $P\text{-value} \approx 0.314$, fail to reject H_0 . We do not have convincing evidence that the mean time per day male students at this university spend using a computer is greater than the mean time for female students.

11.13 $(-1.671, -0.704)$. We can be 95% confident that the difference in the mean sit-stand-sit time for the population of stroke patients who receive biofeedback weight training for 8 weeks and the population of stroke patients who did not receive biofeedback weight training for 8 weeks is between -1.671 and -0.704 seconds. Both endpoints of this interval are negative, so you can estimate that the mean sit-stand-sit time for the population of stroke patients who receive biofeedback weight training for 8 weeks is less than the mean sit-stand-sit time for the population of stroke patients who did not receive biofeedback

weight training for 8 weeks by somewhere between 0.7035 and 1.6705 seconds.

11.15 **a.** If the distributions were both approximately normal, then 22.8% of the male Internet Addiction scores would be negative, and 25.6% of the female Internet Addiction scores would be negative. Assuming that the scores must be positive, these percentages are too large to make it reasonable to believe that the distributions are approximately normal. **b.** It would be appropriate to use the two-sample t test to test the null hypothesis that there is no difference in the mean Internet addiction scores for male and female Chinese sixth grade students. The sample sizes are both much larger than 30. **c.** $H_0: \mu_F - \mu_M = 0$, $H_a: \mu_F - \mu_M < 0$, $t = -4.934$, $df = 1600$, $P\text{-value} \approx 0$, reject H_0 . There is convincing evidence that the mean Internet Addiction score is greater for male Chinese sixth grade students than for female Chinese sixth grade students.

11.17 **a.** $t = 10.359$, $P\text{-value} \approx 0$, reject H_0 **b.** $t = -16.316$, $P\text{-value} \approx 0$, reject H_0 **c.** $t = 4.690$, $P\text{-value} \approx 0$, reject H_0 **d.** The results do seem to provide convincing evidence of a gender bias in the monkeys' choices of how much time to spend playing with each toy, with the male monkeys spending significantly more time with the "masculine toy" than the female monkeys, and with the female monkeys spending significantly more time with the "feminine toy" than the male monkeys. However, the data also provide convincing evidence of a difference between male and female monkeys in the time they choose to spend playing with a "neutral toy." It is possible that it was some attribute other than masculinity/femininity in the toys that was attracting the different genders of monkey in different ways. **e.** The given mean time playing with the police car and mean time playing with the doll for female monkeys are sample means for the same sample of female monkeys. The two-sample t test can only be performed when there are two independent random samples.

11.19 **a.** Since the samples are small it is necessary to know—or to assume—that the distributions from which the random samples were taken are normal. However, in this case, since both standard deviations are large compared to the means, it seems unlikely that these distributions would have been normal. **b.** Since the samples are large, it is appropriate to carry out the two-sample t test. **c.** $t = -2.207$, $P\text{-value} = 0.030$, fail to reject H_0

11.21 **a.** $t = -9.863$, $P\text{-value} \approx 0$, reject H_0 **b.** For the two-sample t test, $t = -9.979$, $df = 22.566$, and $P\text{-value} \approx 0$. Thus, the conclusion is the same.

11.23 $H_0: \mu_d = 0$, $H_a: \mu_d > 0$. $t = -0.247$, $P\text{-value} = 0.595$, fail to reject H_0 . There is not convincing evidence that the mean time to exhaustion for experienced triathletes is greater when they run while listening to motivational music.

11.25 $(-58.198, 46.598)$. We can be 95% confident that the difference in mean time to exhaustion for experienced

triathletes when running to motivational music and the mean time when running with no music is somewhere between -58.198 and 46.598 seconds. Because the endpoints of the confidence interval have opposite signs, zero is included in the interval, and there may be no difference in the mean time to exhaustion for experienced triathletes when running to motivational music and the mean time when running with no music.

11.27 **a.** $H_0: \mu_d = 0$, $H_a: \mu_d \neq 0$. $t = 5.29$, $P\text{-value} \approx 0$, reject H_0 . There is convincing evidence of a difference in the mean reported weight and the mean actual weight for male college students. **b.** $H_0: \mu_d = 0$, $H_a: \mu_d \neq 0$. $t = 18.89$, $P\text{-value} \approx 0$, reject H_0 . There is evidence of a significance difference in the mean reported height and the mean actual height for male college students. **c.** Both of the hypothesis tests in Parts (a) and (b) resulted in rejecting the null hypothesis. Additionally, the t test statistics are both positive, which tells us that the sample mean differences are greater than zero for reported value minus actual value. There is convincing evidence that male college students tend to over-report both height and weight.

11.29 $H_0: \mu_d = 0$, $H_a: \mu_d > 0$. $t = 2.724$, $P\text{-value} = 0.004$, reject H_0 . There is convincing evidence to support the claim that the mean following distance for Greek taxi drivers is greater when there are no distractions than when the driver is talking on a mobile phone. This conclusion is consistent with the claim made by the authors.

11.31 $(0.862, 1.738)$. We can be 95% confident that the mean following distance for Greek taxi drivers while driving with no distractions and while driving and texting is between 0.862 and 1.738 meters. Both endpoints of this interval are positive, so you can estimate that the mean following distance for Greek taxi drivers while driving with no distractions and while driving and texting is somewhere between 0.862 and 1.738 meters.

11.33 $t = -0.515$, $P\text{-value} = 0.612$, fail to reject H_0

11.35 $t = 0.639$, $P\text{-value} = 0.267$, fail to reject H_0

11.37 **a.** $t = 4.451$, $P\text{-value} \approx 0$, reject H_0 **b.** $(-0.210, 0.270)$. **c.** $t = 3.094$, $P\text{-value} = 0.001$, reject H_0 **d.** In Part (a), the male profile heights and the male actual heights are paired (according to which individual has the actual height and the height stated in the profile), and with paired samples, we use the paired t test. In Part (c), we were dealing with two independent samples (the sample of males and the sample of females), and therefore, the two-sample t test was appropriate.

11.39 **a.** $t = 4.321$, $P\text{-value} \approx 0$, reject H_0 **b.** $t = 1.662$, $P\text{-value} = 0.055$, fail to reject H_0 **c.** A smaller standard deviation in the sample of differences means that we have a lower estimate of the standard deviation of the population of differences. Assuming that the mean wrist extensions for the two mouse types are the same (in other words, that the mean of the population of differences is zero), a sample mean difference of as much as 8.82 is much less likely when the standard deviation of the population of

differences is around 10 than when the standard deviation of the population of differences is around 26 .

11.41 $z = -3.94$, $P\text{-value} \approx 0$, reject H_0

11.43 **a.** $z = 1.172$, $P\text{-value} = 0.121$, fail to reject H_0

b. $(-0.036, 0.096)$. We are 99% confident that the difference between the proportion of Gen Y and the proportion of Gen X who made a donation via text message is between -0.036 and 0.096 . In repeated sampling with random samples of size 400 , approximately 99% of the resulting confidence intervals would contain the true difference in proportions who donated via text message.

11.45 No, I would not use the large-sample test for a difference in population proportions because this question is phrased as a one-proportion hypothesis test, not a difference in proportions.

11.47 $(-0.053, -0.005)$. We can be 90% confident that the difference in the proportion living in poverty for men and women is between -0.053 and -0.005 .

11.49 **a.** $H_0: p_1 - p_2 = 0$, $H_a: p_1 - p_2 \neq 0$, $z = -0.574$, $P\text{-value} = 0.566$, fail to reject H_0 . We do not have convincing evidence of a difference between the proportion of young adults who think that their parents would provide financial support for marriage and the proportion of parents who say they would provide financial support for marriage. **b.** $H_0: p_1 - p_2 = 0$, $H_a: p_1 - p_2 > 0$, $z = 2.993$, $P\text{-value} = 0.001$, reject H_0 . We have convincing evidence that the proportion of parents who say they would help with buying a house or apartment is less than the proportion of young adults who think that their parents would help.

11.51 $(-0.016, 0.030)$. We are 95% confident that the actual difference in the proportion of college graduates who were unemployed in October 2013 and the proportion of college graduates who were unemployed in October 2014 is somewhere between -0.016 and 0.030 . Because the endpoints of the confidence interval have opposite signs, zero is included in the interval, and there may be no difference in the proportion of college graduates who were unemployed in October 2013 and the proportion who were unemployed in October 2014.

11.53 **a.** Yes, sample size conditions are met. **b.** $(0.027, 0.173)$. We are 90% confident that the actual difference in the proportion of adult Americans age 18 to 29 who believe the foods made with genetically modified ingredients are bad for their health and the corresponding proportion for adult Americans age 50 to 64 is somewhere between 0.027 and 0.173 . **c.** Zero is not included in the confidence interval. Because both endpoints of the confidence interval are positive, we believe that the percent of adult Americans age 18 to 29 who believe the foods made with genetically modified ingredients are bad for their health is greater than the corresponding proportion for adult Americans age 50 to 64 by somewhere between 2.7 and 17.3 percentage points.

11.55 **a.** It is quite possible that a patient's *knowledge* of being given one of the treatments will itself contribute

positively to the health of the patient, and that the effect of the knowledge of being given an injection might be greater than the corresponding effect of the knowledge of being given a nasal spray (or vice versa). **b.** (0.033, 0.061). We are 99% confident that the proportion of children who get sick with the flu after being vaccinated with an injection minus the proportion of children who get sick with the flu after being vaccinated with the nasal spray is between 0.033 and 0.061. **c.** Yes. Since zero is not included in the interval, there is evidence that the proportions of children who get the flu are different for the two vaccination methods.

11.57 (0.032, 0.108). We are 95% confident that the actual difference in the proportion of adults and teens age 13–17 who believe in reincarnation is somewhere between 0.032 and 0.108.

11.59 $H_0: p_1 - p_2 = 0$, $H_a: p_1 - p_2 > 0$, $z = 3.76$, $P\text{-value} \approx 0$, reject H_0 . There is convincing evidence that the proportion of Gen Xers who do not pay off their credit cards each month is greater than this proportion for millennials.

11.61 **b.** If we want to know whether the e-mail intervention *reduces* (as opposed to *changes*) adolescents' display of risk behavior in their profiles, then we use one-sided alternative hypotheses and the P -values are halved. If that is the case, using a 0.05 significance level, we are convinced that the intervention is effective with regard to reduction of references to sex and that the proportion showing any of the three protective changes is greater for those receiving the e-mail intervention. Each of the other two apparently reduced proportions could have occurred by chance.

11.63 Different simulations will produce different results, so answers will vary. For one randomization test, results indicated a P -value of 0.001. Because this P -value of 0.001 is less than the significance level of 0.05, the null hypothesis is rejected. There is convincing evidence that the difference in the population mean heart rate percentage for drivers in a two-firefighter team is greater than the mean heart rate for drivers in a team with more than two firefighters.

11.65 Different simulations will produce different results, so answers will vary. For one randomization test, results indicated a P -value of 0.001. Because this P -value of 0.001 is less than a significance level of 0.05 or 0.01, the null hypothesis is rejected at either of these significance levels. There is convincing evidence that the mean Personal Meaning Score for patients taking a high dose is greater than the mean score for patients taking a low dose.

11.67 **a.** Because the sample sizes for each treatment group are smaller than 30, we must be wary of using the two-sample t methods. Boxplots for ADHD Severity Scores show that the data are skewed. **b.** Different simulations will produce different results, so answers will vary. For one randomization test, results indicated a P -value of 0.000. Because this P -value of 0.000 is less than a

significance level of 0.05, the null hypothesis is rejected. There is convincing evidence that the mean 1-year improvement in ADHD Severity Score for the ONTRAC treatment is different than the mean 1-year improvement in ADHD Severity Score for the control treatment.

11.69 **a.** The sample size is not greater than 30 and a dotplot of the difference data suggests that it may not be reasonable to assume that the difference distribution is normal. **b.** Different simulations will produce different results, so answers will vary. For one randomization test, results indicated a P -value of 0.003. Because this P -value of 0.003 is less than any reasonable significance level, the null hypothesis is rejected. There is convincing evidence that the mean difference in movement is greater than 0.

11.71 **a.** In order to use the large-sample hypothesis test for the difference in two population proportions, the number of successes and failures in each sample must be at least 10. In this case, the number of successes in one of the groups is $n_1\hat{p}_1 = (21)(0.05) = 1.05$, which is less than 10. **b.** For the given simulation, results indicate a P -value of 0.002. Because this P -value of 0.002 is less than any reasonable significance level, we reject the null hypothesis. We have convincing evidence that the proportion of Tasmanian devils with the genetic marker was greater after DFTD than before DFTD. **c.** The given simulation produced a confidence interval of $(-0.548, -0.214)$. We are 95% confident that the true difference in the rates of occurrence of the specific genetic marker in the genes of Tasmanian devils, before and after DFTD, lies somewhere between -0.548 and -0.214 .

11.73 **a.** Data from this study should not be analyzed using a large-sample hypothesis test for a difference in two population proportions because the number of successes in each group (5 in the high-intensity group and 0 in the regular-intensity group) are both less than 10.

b. Different simulations will produce different results, so answers will vary. For one randomization test, results indicated a P -value of 0.017. Because this P -value of 0.017 is less than any reasonable significance level, we reject the null hypothesis. We have convincing evidence of a difference in the population proportions of patients in the two exercise groups (high-intensity and regular-intensity) who die within 1.5 years. **c.** Different simulations will produce different results, so answers will vary. One simulation resulted in a confidence interval of $(0.123, 0.133)$. We are 95% confident that the actual difference in the population proportions of patients who die within 1.5 years for the two exercise groups is somewhere between 0.123 and 0.133.

11.75 **a.** The given randomization test resulted in a P -value of 0.000. Because this P -value of 0.000 is less than any reasonable significance level, we reject the null hypothesis. We have convincing evidence that the proportion who think it is OK to use a cell phone at a restaurant is higher for the 18 to 29 age group than for the 30 to 49 age group. **b.** The given simulation resulted in a confidence interval of

(0.058, 0.139). We are 90% confident that the actual difference in the proportions of people age 18 to 29 and those age 30 to 49 who would say that it is acceptable to use a cell phone in a restaurant is somewhere between 0.058 and 0.139. **c.** No, the interpretation would not change. The endpoints of the confidence interval change slightly—from Example 11.1, the confidence interval is (0.061, 0.139)—but the conclusion and interpretation remain the same.

11.77 **a.** (0.091, 0.265). We are 90% confident that the difference in the population proportions of cell phone users age 20 to 39 and those age 40 to 49 who say that they sleep with their cell phones is between 0.091 and 0.265. **b.** The given simulation resulted in a bootstrap confidence interval of (0.093, 0.267). Therefore, we are 90% confident that the difference in the population proportions of cell phone users age 20 to 39 and those age 40 to 49 who say that they sleep with their cell phones is between 0.093 and 0.267. **c.** The confidence intervals in Parts (a) and (b) are very similar. The lower endpoints are 0.091 and 0.093, respectively, and the upper endpoints are 0.265 and 0.267, respectively. The interpretations are essentially the same, with the only differences due to the slight differences in the endpoints.

11.79 **a.** $t = -17.382$, $P\text{-value} \approx 0$, reject H_0 **b.** $t = 2.440$, $P\text{-value} = 0.030$, reject H_0 **c.** No, the paired t test would not be appropriate since the treatment and control groups were not paired samples.

11.81 **a.** (−0.056, 0.070) **b.** We are 90% confident that the proportion of people receiving insulin who develop diabetes minus the proportion of people not receiving insulin who develop diabetes is between −0.056 and 0.070. Interpretation of Confidence Level: Consider the process of randomly selecting 339 people at random from the population of people thought to be at risk of developing type I diabetes, randomly assigning them to groups of the given sizes, performing the experiment as described, and then calculating the confidence interval by the method shown in Part (a). Ninety percent of the time, this process will result in a confidence interval that contains $p_1 - p_2$, where p_1 is the proportion of all such people who would develop diabetes having been given the insulin treatment, and p_2 is the proportion of all such people who would develop diabetes having not been given insulin. **c.** Since zero is included in the interval, we do not have convincing evidence of a difference between the proportions of people developing diabetes for the two treatments.

11.83 **a.** Yes. The P -value for this test is given to be less than 0.001. **b.** Yes. The P -value for this test is given to be less than 0.05. **c.** No. We are only given information concerning the difference in the means for male trial and nontrial lawyers and the difference in the means for female trial and nontrial lawyers. We have no information concerning a comparison of the two sexes. In order to conduct the test we would need the values of the sample means for male and female trial lawyers, along with the associated sample standard deviations.

11.85 **a.** In each case, the standard deviation of the differences is large compared to the magnitude of the mean. This can be the case, since differences can be negative. (Also, the large standard deviations reflect the fact that shoppers' consumption can vary considerably from week to week.) **b.** (−0.738, 1.498). We are 95% confident that the mean drink consumption for credit card shoppers in 1994 minus the mean drink consumption for credit card shoppers in 1995 is between −0.738 and 1.498. Since zero is included in this interval, we do not have convincing evidence that the mean number of drinks decreased. **c.** $H_0: \mu_d = 0$, $H_a: \mu_d \neq 0$, $t = 0.764$, $P\text{-value} = 0.445$, fail to reject H_0 . We do not have convincing evidence of a change in the mean number of drinks between 1994 and 1995 for non-credit card shoppers. The P -value tells us that if the mean number of drinks consumed by non-credit card customers were the same for 1994 as for 1995, then the probability of obtaining a sample mean difference as far from zero (or farther) as the one obtained in this study would be 0.445.

11.87 **a.** $H_0: p_1 - p_2 = 0$, $H_a: p_1 - p_2 > 0$, $z = 6.826$, $P\text{-value} \approx 0$, reject H_0 . We have convincing evidence that the use of breakaway bases reduces the proportion of games with a player suffering a sliding injury. **b.** Since the conclusion states that use of breakaway bases *reduces* the proportion of games with a player suffering a sliding injury (that is, *causes* a lower proportion of games with sliding injuries), we need to assume that this was an experiment in which the 2500 games were randomly assigned to the two treatments (stationary bases and breakaway bases). It seems more likely that the treatments were assigned by league (or by region of the country), in which case it would be difficult to argue that each treatment group was representative of all games in terms of sliding injuries.

11.89 **a.** (−4.738, 22.738). **b.** $t = 0.140$, $P\text{-value} = 0.890$, fail to reject H_0 **c.** $t = -0.446$, $P\text{-value} = 0.330$, fail to reject H_0

11.91 **a.** (−0.186, 0.111). We are 90% confident that the mean difference between surface and subsoil pH is between −0.186 and 0.111. **b.** We must assume that the distribution of differences across all locations is normal.

Chapter 12

12.1 **a.** Fail to reject H_0 **b.** Reject H_0 **c.** Reject H_0

12.3 $X^2 = 251.981$, $P\text{-value} \approx 0$, reject H_0 . There is convincing evidence that one or more of the age groups buys a disproportionate share of Texas Lottery tickets.

12.5 $X^2 = 6.017$, $P\text{-value} = 0.872$, fail to reject H_0 . There is not convincing evidence that the proportion of births is not uniform. This is consistent with the researcher's claim that the frequency of births is not related to lunar cycle.

12.7 $X^2 = 20.686$, $P\text{-value} = 0.0003$, reject H_0

12.9 $X^2 = 19.599$, P -value ≈ 0 , reject H_0

12.11 Since the P -value is small, there is convincing evidence that the population proportions of people who respond “Monday” and “Friday” are not equal.

12.13 **a.** $X^2 = 4.63$, P -value = 0.099, fail to reject H_0 . We do not have convincing evidence that the theory is incorrect. **b.** The analysis and conclusion would be the same.

12.15 **a.** P -value = 0.844, fail to reject H_0 **b.** P -value = 0.106, fail to reject H_0

12.17 $X^2 = 48.00$, P -value ≈ 0 , reject H_0 . There is convincing evidence that the proportions in each of the two credit card categories are not the same for all three types of colleges.

12.19 $X^2 = 2.557$, P -value = 0.634, fail to reject H_0 . There is not convincing evidence that the proportions falling into each of the response categories were not the same for high school students in 2004 and 2014.

12.21 $X^2 = 148.250$, P -value ≈ 0 , reject H_0 . There is convincing evidence that there is an association between income category and education level.

12.23 $X^2 = 29.507$, P -value = 0.001, reject H_0

12.25 **a.** $X^2 = 2.314$, P -value = 0.128, fail to reject H_0 **b.** Yes **c.** Yes. Since P -value = 0.127 > 0.05, we do not reject H_0 . **d.** The two P -values are almost equal; in fact, the difference between them is only due to rounding errors in the Minitab program.

12.27 $X^2 = 46.515$, P -value ≈ 0 , reject H_0

12.29 **a.** H_0 : Gender and number of servings of water consumed per day are independent

H_a : Gender and number of servings of water consumed per day are not independent $df = (5 - 1)(2 - 1) = 4$

b. The P -value for the test was 0.086, which is greater than the new significance level of 0.05. So, for a significance level of 0.05, we do not have convincing evidence of a difference between males and females with regard to water consumption. **c.** $X^2 = 15.153$, P -value = 0.015, reject H_0 . There is convincing evidence of an association between gender and consumption of fried potatoes. This agrees with the authors’ conclusion that there was a significant association between gender and consumption of fried potatoes.

12.31 It is not possible to decide which, if either, of the two conclusions is correct. Since the results were obtained from an observational study, no conclusion regarding causality can be reached.

12.33 $X^2 = 11.242$, P -value = 0.105, fail to reject H_0 . We do not have convincing evidence of a color preference.

12.35 $X^2 = 17.346$, P -value = 0.008, reject H_0 . There is convincing evidence that there is an association between age and how students perceive their money management skills.

Chapter 13

13.1 **a.** $y = -5.0 + 0.017x$

13.3 **a.** For $x = 10$, $y = -0.12 + 0.095(10) = 0.83$

For $x = 15$, $y = -0.12 + 0.095(15) = 1.305$

b. Since the slope of the population regression line is 0.095, the average increase in flow rate associated with a 1-inch increase in pressure drop is 0.095.

13.5 **a.** For each one-unit increase in horsepower the predicted fuel efficiency decreases by 0.150 mpg.

b. The number $\hat{y} = 29.0$ is both an estimate of the mean fuel efficiency when the horsepower is 100 and a prediction for the horsepower of a car whose horsepower is 100.

c. When $x = 300$, $\hat{y} = -1.0$. This result cannot be valid, since it is not possible to have a car whose fuel efficiency is negative. This has probably occurred because 300 is outside the range of horsepower ratings for the small cars used in the sample; the estimated regression equation is not valid for values outside this range.

13.7 **a.** 47, 4700 **b.** 0.3156, 0.0643

13.9 **a.** $r^2 = 23.48\%$ **b.** The point estimate of σ_e is $s_e = 5.36559$. The typical amount by which the depression score change value deviates from the value in the sample and what was predicted using the least-squares regression line is 5.366. **c.** 5.08 **d.** 12.966

13.11 **a.** $r^2 = 0.830$. Eighty-three percent of the variability in breast cancer incidence can be explained by the approximate linear relationship between breast cancer incidence and percent of women using HRT. **b.** $s_e = 4.154$. This is a typical deviation of breast cancer incidence value in this data set from the value predicted by the estimated regression line.

13.13 **a.** The plot shows a linear pattern, and the vertical spread of points does not appear to be changing over the range of x values in the sample. If we assume that the distribution of errors at any given x value is approximately normal, then the simple linear regression model seems appropriate. **b.** $\hat{y} = -0.00227 + 1.247x$; when $x = 0.09$, $\hat{y} = 0.110$. **c.** $r^2 = 0.436$, 43.6% of the variation in market share can be explained by the linear regression model relating market share and advertising share. **d.** $s_e = 0.0263$, $df = 8$

13.15 **a.** Yes. P -value = 0.023. Since this value is less than 0.05, we have convincing evidence at the 0.05 significance level of a useful linear relationship between shrimp catch per tow and oxygen concentration density. **b.** Yes. $r = \sqrt{0.496} = 0.704$. So there is a moderate to strong linear relationship for the values in the sample. **c.** (17.363, 177.077) We are 95% confident that the slope of the population regression line relating mean catch per tow to O₂ saturation is between 17.363 and 177.077.

13.17 **a.** 0.1537 **b.** (2.17, 2.83) **c.** Yes, the interval is relatively narrow.

13.19 **a.** $a = 592.1$, $b = 97.26$ **b.** When $x = 2$, $\hat{y} = 786.62$, $y - \hat{y} = -29.62$. **c.** (87.76, 106.76)

13.21 $t = -3.66$, $P\text{-value} \approx 0$, reject H_0

13.23 **a.** (0.081, 0.199) We are 95% confident that the mean change in pleasantness rating associated with an increase of 1 impulse per second in firing frequency is between 0.081 and 0.199. **b.** $t = 5.451$, $P\text{-value} = 0.001$, reject H_0

13.25 **a.** This is the slope of the estimated regression line, or $b = 0.359 \mu\text{g/kg}$. **b.** When $x = 250$ s, the predicted acrylamide concentration is $176.75 \mu\text{g/kg}$. **c.** $t = 0.82$, $P\text{-value} = 0.459$, fail to reject H_0

13.27 **a.** $t = 6.493$, $P\text{-value} \approx 0$, reject H_0 **b.** $t = 1.56$, $P\text{-value} = 0.079$, fail to reject H_0

13.29 **a.** 0.253 **b.** 0.179; no **c.** 4

13.31 There is a random scatter of points in the residual plot, implying that a linear model relating squirrel population density to percentage of logging is appropriate. The residual plot shows no tendency for the size (magnitude) of the residuals to either increase or decrease as percentage of logging increases. So it is justifiable to assume that the vertical deviations from the population regression line have equal standard deviations. The last condition is that the vertical deviations be normally distributed. The fact that the boxplot of the residuals is roughly symmetrical and shows no outliers suggests that this condition is satisfied.

13.33 There is a curved pattern in the residual plot, which suggest that the simple linear regression model is not appropriate.

13.35 No. The pattern in the scatterplot looks linear, but the spread around the line does not appear to be constant. The histogram of the residuals also appears to be positively skewed.

CUMULATIVE REVIEW EXERCISES 13

CR13.1 Randomly assign the 400 students to two groups of equal size, Group A and Group B. Have the 400 students take the same course, attending the same lectures and being given the same homework assignments. The only difference between the two groups should be that the students in Group A should be given daily quizzes and the students in Group B should not. After the final exam the exam scores for the students in Group A should be compared to the exam scores for the students in Group B.

CR13.3 **b.** The two airlines with the highest numbers of fines assessed may not be the worst in terms of maintenance violations since these airlines might have more flights than the other airlines.

CR13.5 **a.** (0.651, 0.709) We are 95% confident that the proportion of all adult Americans who view a landline phone as a necessity is between 0.651 and 0.709. **b.** $z = 1.267$, $P\text{-value} = 0.103$, fail to reject H_0 **c.** $z = 9.513$, $P\text{-value} \approx 0$, reject H_0

CR13.7 **a.** 0.62 **b.** 0.1216 **c.** 0.19 **d.** 0.0684

CR13.9 **b.** $\hat{y} = -12.887 + 21.126x$ **d.** $t = 21.263$, $P\text{-value} \approx 0$, reject H_0

CR13.11 $X^2 = 26.175$, $P\text{-value} \approx 0$, reject H_0

CR13.13 $X^2 = 15.106$, $P\text{-value} = 0.002$, reject H_0

CR13.15 $t = -113.17$, $df = 45$, $P\text{-value} \approx 0$, reject H_0

CR13.17 $X^2 = 4.8$, $P\text{-value} = 0.684$, fail to reject H_0

Chapter 14

14.1 **a.** A deterministic model does not have the random deviation component e , while a probabilistic model does contain such a component.

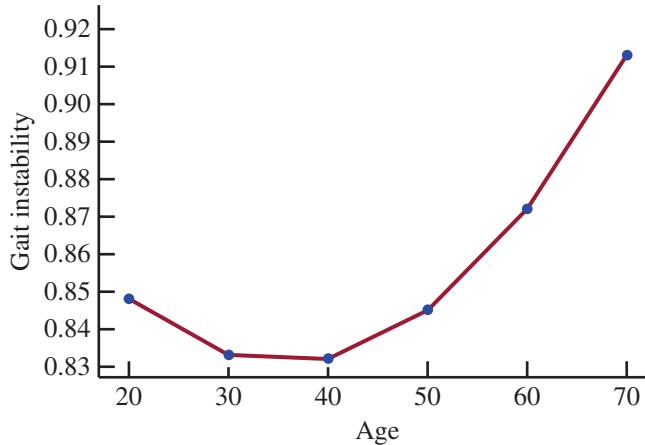
14.3 **a.** (mean y value for fixed values of x_1, x_2, x_3) = $30 + 0.90x_1 + 0.08x_2 - 4.5x_3$

b. $\beta_0 = 30$, $\beta_1 = 0.9$, $\beta_2 = 0.08$, $\beta_3 = -4.50$ **c.** The average change in acceptable load associated with a 1-cm increase in left lateral bending, when grip endurance and trunk extension ratio are held fixed, is 0.90 kg. **d.** The average change in weight associated with a 1 N/kg increase in trunk extension ratio, when grip endurance and left lateral bending are held fixed, is -4.5 kg.

14.5 **a.** 13.552 g. **b.** When length is fixed, the mean increase in weight associated with a 1-mm increase in width is 0.828 g. When width is fixed, the mean increase in weight associated with a 1-mm increase in length is 0.373 g.

14.7 **a.** 35,893 **b.** $\beta_1 = 0.974$ is the change in power consumption associated with an increase of 1 million in population when gross domestic product is held constant.

14.9 **a.**



b. From 20 to 30, decreases by 0.15. From 50 to 60 increases by 0.027.

14.11 **c.** The parallel lines in each graph are attributable to the lack of interaction between the two independent variables.

14.13 **a.** $y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + e$

b. $y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1^2 + \beta_5x_2^2 + \beta_6x_3^2 + e$

c. $y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_2 + e$

$y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_3 + e$

$y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_2x_3 + e$

d. $y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1^2 + \beta_5x_2^2 + \beta_6x_3^2 + \beta_7x_1x_2 + \beta_8x_1x_3 + \beta_9x_2x_3 + e$

14.15 **a.** Three dummy variables would be needed to incorporate a nonnumerical variable with four categories. For example, you could define $x_3 = 1$ if the car is a subcompact and 0 otherwise, $x_4 = 1$ if the car is a compact and 0 otherwise, and $x_5 = 1$ if the car is a midsize and 0 otherwise. The model equation is then $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + e$. **b.** For the variables defined in Part (a), $x_6 = x_1 x_3$, $x_7 = x_1 x_4$, and $x_8 = x_1 x_5$ are the additional predictors needed to incorporate interaction between age and type of car.

14.17 **a.** $0.01 < P\text{-value} < 0.05$ **b.** $P\text{-value} > 0.10$
c. $P\text{-value} = 0.01$ **d.** $0.001 < P\text{-value} < 0.01$

14.19 **a.** $F = 12118$, $P\text{-value} \approx 0$, reject H_0 **b.** Since the $P\text{-value}$ is small and r^2 is close to 1, there is strong evidence that the model is useful. **c.** The model in Part (b) should be recommended, since adding the variables x_1 and x_2 to the model [to obtain the model in Part (a)] only increases the value of R^2 a small amount (from 0.994 to 0.996).

14.21 $F = 24.41$, $P\text{-value} < 0.001$, reject H_0 and conclude that the model is useful.

14.23 $F = 3196.02$, $P\text{-value} \approx 0$, reject H_0 and conclude that the model is useful.

14.25 $F = 7.986$, $P\text{-value} < 0.001$, reject H_0 and conclude that the model is useful.

14.27 **a.** $\hat{y} = 1.44 - 0.0523\text{length} + 0.00397\text{speed}$
b. 1.3245 **c.** $F = 24.02$, $P\text{-value} \approx 0$, reject H_0 and conclude that the model is useful. **d.** $\hat{y} = 1.59 - 1.40\left(\frac{\text{length}}{\text{speed}}\right)$
e. The model in Part (a) has $R^2 = 0.75$ and $R^2 \text{ adjusted} = 0.719$, whereas the model in Part (d) has $R^2 = 0.543$ and $R^2 \text{ adjusted} = 0.516$.

14.29 **a.** $\text{SSResid} = 390.4347$, $\text{SSTo} = 1618.2093$, $\text{SSRegr} = 1227.7746$ **b.** $R^2 = 0.759$; this means that 75.9% of the variation in the observed shear strength values has been explained by the fitted model. **c.** $F = 5.039$, $0.01 < P\text{-value} < 0.05$, reject H_0 , and conclude that the model is useful.

14.31 $F = 96.64$, $P\text{-value} < 0.001$, reject H_0 , and conclude that the model is useful.

14.35 $\hat{y} = 35.8 - 0.68x_1 + 1.28x_2$, $F = 18.95$, $P\text{-value} < 0.001$, reject H_0 , and conclude that the model is useful.

Chapter 15

15.1 **a.** $0.001 < P\text{-value} < 0.01$ **b.** $P\text{-value} > 0.10$
c. $P\text{-value} = 0.01$ **d.** $P\text{-value} < 0.001$ **e.** $0.05 < P\text{-value} < 0.10$ **f.** $0.01 < P\text{-value} < 0.05$ (using $df_1 = 4$ and $df_2 = 60$)
15.3 **a.** $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$, H_a : At least two of the four μ_i 's are different. **b.** $P\text{-value} = 0.012$, fail to reject H_0
c. $P\text{-value} = 0.012$, fail to reject H_0

15.5 $F = 6.687$, $P\text{-value} = 0.001$, reject H_0

15.7 **a.** $F = 3.86$, $P\text{-value} = 0.035$, reject H_0 . There is sufficient evidence to conclude that the mean hunger rating is not the same for all three treatments (“healthy” snack, “tasty” snack, no snack). **b.** It is not reasonable to conclude that the mean hunger rating is greater for people who do not get a snack because the sample mean hunger rating is highest for the healthy snack group.

15.9 $F = 5.273$, $P\text{-value} = 0.002$, reject H_0

15.11 **a.** Random assignment ensures that the experiment does not systematically favor one experimental condition over any other and attempts to create experimental groups that are as much alike as possible. **b.** $F = 2.44$, $P\text{-value} = 0.094$, fail to reject H_0 . There is not sufficient evidence to support the claim that the mean average speed is not the same for all three treatments (Stretching Method 1, Stretching Method 2, and No Stretching). **c.** The authors were likely surprised by the results of this study because they did not find evidence that supports a difference in mean average speed for the three treatments (the null hypothesis was not rejected). This means that there is not convincing evidence that stretching improves performance. These results contradict the results of other studies, which concluded that stretching can improve performance.

15.13

Source of Variation	df	Sum of Squares	Mean Square	F
Treatments	3	75,081.72	25,027.24	1.70
Error	16	235,419.04	14,713.69	
Total	19	310,500.76		

$F = 1.70$, $P\text{-value} > 0.10$, fail to reject H_0

15.15 Since there is a significant difference in all three of the pairs we need a set of intervals that does not include zero. Set 3 is therefore the required set.

15.17 **a.** In decreasing order of the resulting mean numbers of pretzels eaten the treatments were slides with related text, slides with no text, slides with unrelated text, and no slides. There were no significant differences between the results for slides with no text and slides with unrelated text and for slides with unrelated text and no slides. However, there was a significant difference between the results for slides with related text and each one of the other treatments and between the results for no slides and for slides with no text (and for slides with related text).

b. The results for the women and men are almost exactly the reverse of one another, with, for example, slides with related text (treatment 2) resulting in the smallest mean number of pretzels eaten for the women and the largest mean number of pretzels eaten for the men. For the men, treatment 2 was significantly different from all the other treatments; however, for women treatment 2 was not significantly different from treatment 1. For both women and men there was a significant difference between treatments

1 and 4 and no significant difference between treatments 3 and 4. However, between treatments 1 and 3 there was a significant difference for the women but no significant difference for the men.

15.19 **a.** $F = 3.34$, $P\text{-value} = 0.041$, reject H_0 . There is sufficient evidence to conclude that the mean gender stereotyping scores for the three different treatment groups differ.

b.

Control	X-Men	Spider Man
2.668	3.020	3.120

c. The mean measure of gender stereotyping is significantly different for the Control group and the Spider-Man group. The mean for X-Men was not significantly different from either of the other two groups. Because lower values indicate attitudes more accepting of equality, you can conclude that those who did not watch a video were more accepting of equality than those who watched Spider-Man.

15.21 **a.** $F = 45.64$, $P\text{-value} \approx 0$, reject H_0 **b.** Yes; T-K interval is $(0.388, 0.912)$

Index

1 in k systematic sample, 41

A

A and B, 280

A or B, 280

Addition rule

for mutually exclusive events, 293
general, 315–317

Additive multiple regression model, 731

Additive probabilistic model, 690

Adjusted R^2 , 747

Alternative hypothesis, 508–509

Analysis of variance (ANOVA). *See also* Single-factor analysis of variance (ANOVA)

ANOVA table, 768

B

Bar chart

for categorical data, 12–13
comparative, 78–79
uses of, 84–85

Bayes, Reverend Thomas, 321

Bayes' Rule, 321–323

Bell-shaped curve, 107. *See also* Normal distribution

Bias in sampling, 34–36

Biased statistic, 456

Bimodal histogram, 106–107

Binomial distribution, 372–379

defined, 372
formula for, 373–374
mean of, 378
normal approximation to, 412–413
standard deviation of, 378

Binomial experiment, 372

Binomial random variable, 372

Bivariate data

analysis, 260–261
cautions and limitations, 261–263
scatterplot of, 116

Bivariate data set

defined, 10

Blocking, 49

defined, 46

extraneous variables and, 46

random assignment and, 55

Bootstrap confidence interval

defined, 490, 496

for difference between two population means

using independent samples, 628–629

using paired samples, 626–628

for difference between two population proportions or two treatment proportions, 633–634

for difference between two treatment means, 625–626

for one proportion, 491–492

for a population mean, 496–497

Bootstrap distribution, 490, 496

Boxplots, 168–170

comparative, 169

modified, 170

outlier and, 169

skeletal, 169

C

Categorical data, 9, 155

bar chart for, 12–13

chi-square test for, 655–659

comparative bar chart for, 78–79

frequency distribution for, 11–12

multivariate, 675

pie chart for, 79–81

Causation

correlation and, 205

Cell count. *See also* Expected cell count; Observed cell count

Census, 34

Center of data set, 149–156

Central Limit Theorem, 436–437

Chance experiment, 278–279

Chebyshev's rule, 175–177

Chi-square distribution, 658

Chi-square tests

goodness-of-fit, 659–662

homogeneity, 665–675

independence, 671–673

for univariate data, 655–662

Class interval, 101, 104

Cluster sampling, 40–41

Clusters, 40

- Coefficient of determination, r^2 , 228–231
 Coefficient of multiple determination, R^2 , 746
 Common population proportion, 610
 Comparative bar chart, 78–79
 Comparative boxplots, 169
 Comparative stem-and-leaf display, 92–93
 Complement of an event, 280
 Complete second order model, 736
 Completely randomized design, 56
 Conditional probability, 297–303
 compared to unconditional probability, 301, 320
 Confidence interval
 for comparing two population or treatment means using
 independent samples, 587–588
 paired samples, 601–603
 for comparing two population or treatment proportions,
 614–615
 general form of, 466–467
 for population mean, 476
 for population proportion, 462–465
 for the slope of the population regression line, 703
 Confidence level, 460
 Confounded variables, 46
 Confounding variables, 30
 Contingency table, 665–666
 Continuity correction, 411
 Continuous numerical data
 defined, 10
 frequency distribution, 101–102
 histogram, 102–104
 Continuous random variable, 344, 353
 mean of, 364
 probability distribution of, 353–356
 standard deviation of, 364
 Control group, 53, 60
 Convenience sampling, 41
 Correlation, 198–205
 causation and, 205
 Correlation coefficient, 198–199
 cautions and limitations, 261–263
 coefficient of determination and, 230
 formula for, 199
 Cumulative area, 356
 Cumulative relative frequency, 110–111
 Cumulative relative frequency plot, 111–112
 Curvature in residual plot, 224
- D**
- Danger of extrapolation, 213–214, 262, 697
 Data
 types of, 9–10
 Data analysis process, 6–7
 Degrees of freedom
 ANOVA and, 763
 chi-square distributions and, 658
 simple linear regression and, 697
 t distribution and, 474–475
 two-sample t test and, 586
 Degrees of freedom (df), 163
 Density, 104
 Density curve, 354
 Density function, 354
 Density scale, 104
 Dependent events, 308
 Dependent variable, 209
 Descriptive statistics, 7
 Deterministic relationship, 690
 Deviations from the mean, 159–161
 Diagram of experimental design, 53
 Direct control, 46, 49–50
 Discrete data, 10
 frequency distribution, 97–101
 histogram, 97–101
 Discrete probability distribution, 355–356
 binomial distribution, 372–379
 geometric distribution, 379–380
 properties of, 349
 Discrete random variable
 defined, 344
 mean value of, 360–362
 probability distribution of, 347–351
 standard deviation of, 362–364
 variance of, 363
 Disjoint, 281
 Disjoint events. *See* Mutually exclusive events
 Dotplot, 13–16
 Double-blind experiment, 61–62
- E**
- Empirical Rule, 177–178
 standard normal distribution and,
 385, 388
 Equally likely outcomes, 285
 Error sum of squares (SSE), 763
 Errors in hypothesis testing, 512–515
 Estimated standard deviation, 697
 Estimation
 bias in, 456
 choosing a statistic, 455–458
 interval estimate (*See* Confidence interval)
 point estimate, 454
 Event, 279–283
 foming new events, 280–282
 simple, 279
 Exact binomial test
 for population proportion, 556–559
 Expected cell count, 668
 Expected value. *See* Mean
 Experiment, 29, 30, 45
 completely randomized, 56
 double-blind, 61–62
 replicating, 48–49
 single-blind, 61–62

using a control group, 53, 60

using volunteers, 41, 63

Experimental conditions, 45

Experimental design

evaluation of, 51

principles of, 49

Experimental unit, 62

Explanatory variables, 45, 49, 209

Extraneous variables, 46–47, 50

Extrapolation, danger of, 213–214, 262, 697

Extreme outlier, 169

F

F distribution, 747–748

F test

ANOVA, 760–769

for model utility, 748–749

Factor, 29, 30, 45, 760

Finite population correction factor, 487

Fitted value. *See* Predicted value

Five-number summary, 168

Frequency, 11

Frequency distribution, 11

for categorical data, 11–12

for continuous numerical data, 101–102

for discrete numerical data, 97–101

Full quadratic model, 736

Fundamental identity for single-factor ANOVA, 768

G

General addition rule, 315–317

General additive multiple regression model, 731

General multiplication rule, 317–319

Geometric probability distribution, 379–380

Geometric random variable, 379

Goodness-of-fit test, 659–662

Goodness-of-fit statistic, 658

Grand total, 666

Graphical data display

cautions and limitations, 129–132

H

Heavy-tailed curve, 107

Histograms

continuous numerical data, 102–104

with equal width intervals, 102

with unequal width intervals, 104

discrete numerical data, 97–101

probability, 349, 353, 377

shapes, 106–107

using density, 104–105

Homogeneity, chi-square test of, 665–675

Hypergeometric distribution, 376

Hypotheses, 508

alternative hypothesis, 508

null hypothesis, 508

Hypothesis test

comparing two population or treatment means using

independent samples, 577–590

paired samples, 595–603

errors in, 512–515

for comparing two population or treatment proportions, 608–615

for goodness-of-fit, 659

for homogeneity, 669

for independence, 672–673

for population mean, 530–537

for population proportion, 517–527

for the slope of the population regression line, 705–709

power of (*See* Power)

randomization test

for one mean, 562–563

for one proportion, 552–556

for two means, 623–631

for two proportions, 633–636

significance level of, 513

steps for, 525

I

Implicit null hypothesis, 509

Independence, 307–312

testing for, 672–673

Independent events

defined, 308

multiplication rule for more than two, 310

multiplication rule for two, 308–309

Independent samples, 577–578, 595

comparing two population or treatment means using, 577–590

simulation-based inference for difference in two population means using, 626–628

Independent variable, 209

Indicator variable, 738

Inferential statistics, 7

Influential observation, 227–228, 717

Interaction, 734–738

predictor in multiple regression, 738–739

Intercept, 210

Interquartile range (iqr), 163–164

Intersection of two events, 280

Interval estimate. *See* Confidence interval

J

Joint probability table, 307

L

Large sample confidence interval

for the difference in two population or treatment proportions, 614–615

for population proportion, 462–465

- Large-sample hypothesis test
 for the difference in two population or treatment proportions, 609–614
 for a population proportion, 517–527
- Law of large numbers, 288, 328
- Law of total probability, 320–321
- Leaf, 88–89
- Least squares regression line
 assessing fit of, 221–235
 defined, 211
 deviations from, 210–211
 influential observations and, 227
 making predictions with, 211
 outliers and, 227
 population regression line and, 694
 slope of, 211, 212
 standard deviation about, 231–235
 y intercept of, 211
- Least-squares estimates, 744
- Left skewed histogram, 107
- Light-tailed curve, 107
- Line
 assessing fit of, 221–235
 equation of, 210
- Linear combination of random variables, 366, 368–369
 mean and standard deviation, 365–368
- Linear function of random variable, 365, 368–369
 mean and standard deviation of, 364–365
- Linear regression, 234
 model, 690–694
 model utility test for simple linear regression, 707
- Linear relationship, strength of, 198–199, 203
- Lower quartile, 163–164
- Lower tail of histogram, 107
- Lower-tailed test, 523, 532
- M**
- Margin of error
 defined, 467
 sample size choice and, 467–468, 480
- Marginal total, 666
- Mean
 of binomial distribution, 378
 of a continuous random variable, 364
 of a data set, 149–153
 defined, 149
 deviations from, 159–161
 of discrete random variable, 360–362
 of a linear combination, 366
 of a linear function, 364–365
 of normal distribution, 383
 outliers and, 152
 population mean, 151
 of random variable, 358
 sample mean, 150
- of sampling distribution, 432–439
 vs. median, 154–155
- Mean square, 763
 for error (MSE), 763–765
 for treatments (MSTr), 763–765
- Measure of center, 149–156
- Measure of relative standing, 178–179
- Measurement bias, 34–35
- Median
 of a data set, 153–154
 defined, 153
 outliers and, 153–154
 sample median, 153
 vs. mean, 154–155
- Mild outlier, 169
- Mode, 106
- Model equation, 690
- Model utility test
 in multiple regression, 748–749
 for simple linear regression, 707
- Modified boxplots, 170
- Multimodal histogram, 106
- Multiple comparison procedure, 772–778
- Multiple regression
 general additive model, 731
 interaction variables in, 734–738
 model utility F test, 748–749
 qualitative predictors in, 738–739
- Multiplication rule
 general, 317–319
 for independent events, 308–310
- Multivariate data, 10
- Mutually exclusive events, 281–282
 addition rule for, 293
- N**
- Negative skew, 107
- Nonlinear regression, 241–256
 choosing models, 253–256
 common nonlinear functions, 241–243
- Nonresponse bias, 35
- Normal approximation to the binomial distribution, 412–413
- Normal curve, 107. *See also* Normal distribution
 discrete variables and, 410–412
- Normal distribution, 383–397
 extreme values in, 395–397
 mean value of, 383
 standard deviation of, 383
 standard normal distribution, 384–389
- Normal probability plot, 400–401
- Normal score, 400
- Null hypothesis, 508, 509–510
- Numerical data, 9
 continuous, 10–11
 discrete, 10–11

displaying bivariate, 116–122
dotplots for, 13–16
frequency distribution for
continuous numerical data, 101–102
discrete numerical data, 97–101
histograms for
continuous numerical data, 102–104
discrete numerical data, 97–101
stem-and-leaf display for, 88–93
types, 10–11
Numerical descriptive measures
cautions and limitations, 185–186

O

Observational study, 29–31
Observed cell count, 666, 668
Observed significance level, 520. *See also P-value*
One sample z confidence interval, 472
One-sample t confidence interval, 476
One-sample t test, 532
One-way frequency table, 655
Outlier, 89, 152, 153–154, 169, 227, 717
Overcoverage, 65

P

Paired data, 15
Paired samples, 578, 596
comparing two population or treatment means using, 595–603
paired t test and, 597–601
simulation-based inference about two population means using, 628–631
Paired t confidence interval, 601–603
Paired t test, 597–601
Pearson's sample correlation coefficient. *See Correlation coefficient*
Percentile, 179–180
Pie chart
for categorical data, 79–81
uses of, 84–85
Placebo, 61
Plotting residuals, 224–228, 716–721
Point estimation, 454–458
Polynomial regression, 732–734
Pooled t test, 586
Population, 6
Population correlation coefficient, 205
Population mean, 151
bootstrap confidence intervals for, 496–497
comparing two using independent samples, 577–590
comparing two using paired samples, 595–603
confidence interval for, 476
hypothesis test for, 530–537
one-sample t test for, 532
randomization test for, 562–563
simulation-based inference for two means, 623–631

Population proportion
bootstrap confidence interval for, 489–494
comparing two using large samples, 609–614
exact binomial test for, 556–559
hypothesis test for, 517–527
large-sample confidence interval for, 462–465
large-sample hypothesis test for, 517–527
randomization test for, 552–556
simulation-based inference for two proportions, 633–636
Population proportion of S's, 156
Population regression coefficients, 731
Population regression function, 732, 733
Population regression line, 691–694
Population standard deviation, 163
Population variance, 163
Positive skew, 107
Potentially influential observation, 717
Power
calculating, 543–544
defined, 541
effect of sample size and significance level on, 541–543
type II error probabilities and, 541–548
Practical significance, 537, 551
Predicted value, 221–223, 745
Predictor variable, 209
Principle of least-squares, 210–215
Probabilistic model, 690
Probability
of an event, 285, 288, 290–291
basic properties of, 290–294
classical approach for equally likely outcomes, 285–286
of the complement of an event, 310
conditional, 297–303
estimating empirically, 328–329
estimating using simulation, 329–334
histogram, 349, 353, 377
of the intersection of two events, 308, 317
relative frequency approach to, 286–289
subjective approach to, 289–290
of the union of two events, 293, 316
Probability distribution
for continuous random variable, 353–356
for discrete random variable, 347–351
Probability histogram, 349, 353, 377
P-value
alternative hypotheses and, 524
defined, 520
determining for a chi-square test, 659
determining for a t test, 532
determining for a z test, 522–527

Q

Quadratic regression, 244–246, 733, 736
Qualitative data, 9
Qualitative predictor variables, 738–739
Quantitative data, 9

Quartile

- lower, 163–164
- upper, 163–164

R

r. See Correlation coefficient

Random assignment, 46–51

Random deviation, 691, 699

Random number, 37–38

Random sampling, 36–37

Random selection, 39

Random variable, 344–346

- linear combination of, 366

- linear function of, 365

- mean value of, 358

- standard deviation of, 358

Randomization distribution, 553

Randomization test

- for difference between two means, 623–631

- for difference between two proportions, 634–636

- for population mean, 562–563

- for population proportion, 552–556

Randomized block design, 56

Range, 159

Re-expression. *See* Transformations

Regression. *See* Least squares regression line; Linear regression

Regression analysis, 215–216

- linear regression, 234

- nonlinear regression, 241–256

Regression sum of squares (SSRegr), 746, 748–749

Relative frequency, 11

Relative frequency distribution, 11

Replacement, sampling with and without, 38, 310–312, 376–377

Replication, 48, 49, 62–63

Resampling, 490

Residual

- in linear regression, 221–223

- in multiple regression, 745

- plotting, 224–228, 716–721

Residual analysis, 714–716

Residual plot, 224–228, 716–721

- defined, 716

- standardized residual plot, 716–718

Residual sum of squares (SSResid), 228–230, 746

Resistant methods, 721

Response bias, 34–35

Response variable, 45, 49, 209

Right skewed histogram, 107

Robust methods, 721

*r*th percentile, 179

S

Sample, 6

- independent samples, 577–578, 595

- paired samples, 595–603

Sample correlation coefficient, 199–200

- properties, 202–203

Sample mean, 150

Sample median, 153

Sample proportion of successes, 155

Sample regression line, 211, 215–216. *See also* Least squares regression line

Sample size, 39

- margin of error and, 467–468, 480

- power and, 541–543

Sample space, 278–279

Sample standard deviation, 161

Sample variance, 161

Sampling, 34–44

- bias in, 34–36

- cluster, 40–41

- convenience, 41

- random, 36–37

- stratified, 40

- systematic, 41

- with and without replacement, 38, 310–312, 376–377

Sampling distribution, 427, 430

- of the difference in sample means (independent samples), 578

- of the difference in sample proportions, 609

- of the sample mean, 432–439

- of the sample proportion, 441–445

- of the slope of the least squares regression line, 702

Sampling frame, 37

Sampling variability, 108, 428–430

Scatterplot, 116

Segmented bar charts, 82–84

Selection bias, 34, 35

Significance level, 513, 521

- power and, 541–543

Simple comparative experiment, 45–60

- design strategies for, 46

Simple event, 279

Simple linear regression. *See* Least squares regression line;

- Linear regression

Simple random sample, 36

Simulation to estimate probabilities, 329–334

Simulation-based inference

- for two means, 623–631

- for two proportions, 633–636

Simultaneous confidence level, 775

Single-blind experiment, 61–62

Single-factor analysis of variance (ANOVA), 759

- F* test for, 760–769

- fundamental identity for, 768

Skeletal boxplots, 169

Skewed histogram, 107

Slope

- defined, 210

- estimate for population regression line, 702–709

- of least-squares regression line, 211, 212, 262

Smoothed histogram, 106

Stacked bar chart. *See* Segmented bar charts

- Standard deviation
 about the least-squares line, 231–235
 of binomial distribution, 378
 of a continuous random variable, 364
 of a data set, 161–163
 of difference in means, 578
 of discrete random variable, 362–364
 of a linear combination, 366
 of a linear function, 364–365
 of normal distribution, 383
 of random variable, 358
 of the slope of the least-squares line, 702
- Standard error, 467, 485
- Standard normal distribution, 384–389
- Standardized residual, 714
- Standardized residual plot, 716–718
- Standardizing, endpoints, 391
- Statistical significance, 537
- Statistics
 defined, 1, 428
 descriptive, 7
 inferential, 7
- Stem-and-leaf display, 88–93
 comparative, 92–93
 repeated, 91–92
- Stems, 88–89
- Strata, 40
- Stratified random sampling, 40
- Studentized range distribution, 773
- Sum of the squared deviations, 211
- Sum of squares
 error (SSE), 763
 regression (SSReg), 746, 748
 total (SSTo), 228–230, 746, 768
 treatment (SSTr), 763
- Symmetric histogram, 106
- Systematic sampling, 41
- T**
- t* confidence interval
 for a difference in population or treatment means, 587–588
 for a population mean, 476
- t* critical value, 475
- t* distribution, 474–475
 degrees of freedom and, 474
 properties of, 474
- t* test
 finding *P*-values for, 532
 one-sample *t* test for a population mean, 530–537
 paired *t* test, 597–601
 pooled *t* test, 586
 power of and type II error probabilities, 546–548
 for the slope of the population regression line, 706
 two-sample *t* test for difference in population or treatment means, 577–590
- Test of hypotheses. *See* Hypothesis test
- Test procedure, 508, 512
- Test statistic, 520
- Time series plot, 119–120
- Total sum of squares (SSTo), 228–230, 746, 768
- Transformations, 246–248
 defined, 248
 models transforming x , 248–251
 models transforming y , 251–253
 nonlinear regression models, 241–256
- Transforming data to normal distribution, 402–406
- Treatment, 45, 49
- Treatment means
 comparing using independent samples, 584–586
 comparing using paired samples, 595–603
 confidence interval for comparing using independent samples, 587–588
 confidence interval for comparing using paired samples, 601–603
 simulation-based inference about the difference in, 623–626
- Treatment proportions
 bootstrap confidence intervals for the difference in, 633–634
 confidence interval for difference in, 614–615
 large sample test for difference in, 609–614
 randomization tests for the difference in, 634–636
- Treatment sum of squares (SSTr), 763
- Tree diagram, 279
- Tukey-Kramer (T-K) multiple comparison procedure, 772–776
 simultaneous confidence level and, 775
- Two-sample *t* confidence interval, 587–588
- Two-sample *t* test
 for comparing two population means, 580
 for comparing two treatment means, 584
- Two-tailed test, 523, 532
- Two-way frequency table, 665–666
- Type I error, 512–515
 probability of, 513
- Type II error, 512–515
 power and, 541–548
 probability of, 513, 541–548
- U**
- Unbiased statistic, 456–457
- Unconditional probability
 compared to conditional probability, 301, 320
- Undercoverage, 34, 35, 65
- Unexplained variability, 229
- Uniform distribution, 355
- Unimodal histogram, 106–107
- Union of two events, 280
- Univariate data set, 9
- Upper quartile, 163–164
- Upper tail of histogram, 107
- Upper-tailed test, 523, 532

V

Variability

- in data set, 159–165
- interpreting, 175–180
- role of, 3–5
- sampling, 428–430
- unexplained, 229

Variables, 9

- blocking, 46
 - confounding, 30, 46
 - continuous random, 344
 - dependent, 209
 - discrete random, 344
 - explanatory, 45, 209
 - extraneous, 46–47, 50
 - independent, 209
 - indicator, 738
 - interaction between, 734–738
 - predictor, 209
 - random, 45, 344–346
 - response, 45, 49, 209
- Variance, 161–163. *See also* Standard deviation
of discrete random variable, 363
- Venn diagram, 281–282
- Vertical intercept, 210

Voluntary response sampling, 41
Volunteers in experiments, 41, 63

W

Whiskers in a boxplot, 168–170

Y

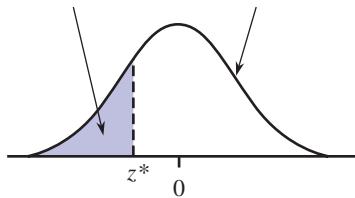
y-intercept, 210
of the least-squares regression line, 211, 262
of the population regression line, 692

Z

z confidence interval
for a difference in population or treatment proportions, 614–615
for a population mean, 472
for a population proportion, 466
z critical value, 464, 465
z curve, 384. *See also* Standard normal distribution
z test
for a difference in population proportions, 610
for a population proportion, 520–521
z-score, 179, 391
and the correlation coefficient, 199

**Standard normal
probabilities
(cumulative z curve
areas)**

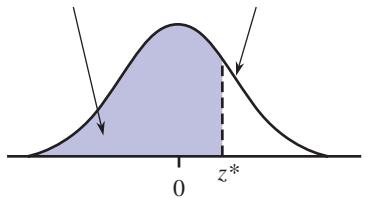
Tabulated area = probability Standard normal (z) curve

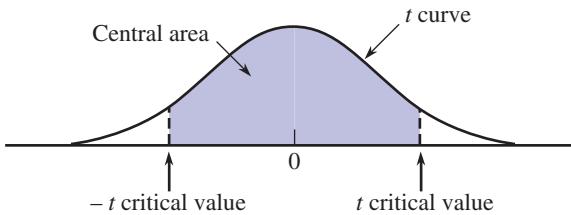


z^*	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.8	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
-3.7	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-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

Standard normal probabilities (*continued*)

Tabulated area = probability Standard normal (z) curve



***t* critical values**

Central area captured: Confidence level:	.80 80%	.90 90%	.95 95%	.98 98%	.99 99%	.998 99.8%	.999 99.9%	
Degrees of freedom	1	3.08	6.31	12.71	31.82	63.66	318.31	636.62
	2	1.89	2.92	4.30	6.97	9.93	23.33	31.60
	3	1.64	2.35	3.18	4.54	5.84	10.21	12.92
	4	1.53	2.13	2.78	3.75	4.60	7.17	8.61
	5	1.48	2.02	2.57	3.37	4.03	5.89	6.86
	6	1.44	1.94	2.45	3.14	3.71	5.21	5.96
	7	1.42	1.90	2.37	3.00	3.50	4.79	5.41
	8	1.40	1.86	2.31	2.90	3.36	4.50	5.04
	9	1.38	1.83	2.26	2.82	3.25	4.30	4.78
	10	1.37	1.81	2.23	2.76	3.17	4.14	4.59
	11	1.36	1.80	2.20	2.72	3.11	4.03	4.44
	12	1.36	1.78	2.18	2.68	3.06	3.93	4.32
	13	1.35	1.77	2.16	2.65	3.01	3.85	4.22
	14	1.35	1.76	2.15	2.62	2.98	3.79	4.14
	15	1.34	1.75	2.13	2.60	2.95	3.73	4.07
	16	1.34	1.75	2.12	2.58	2.92	3.69	4.02
	17	1.33	1.74	2.11	2.57	2.90	3.65	3.97
	18	1.33	1.73	2.10	2.55	2.88	3.61	3.92
	19	1.33	1.73	2.09	2.54	2.86	3.58	3.88
	20	1.33	1.73	2.09	2.53	2.85	3.55	3.85
	21	1.32	1.72	2.08	2.52	2.83	3.53	3.82
	22	1.32	1.72	2.07	2.51	2.82	3.51	3.79
	23	1.32	1.71	2.07	2.50	2.81	3.49	3.77
	24	1.32	1.71	2.06	2.49	2.80	3.47	3.75
	25	1.32	1.71	2.06	2.49	2.79	3.45	3.73
	26	1.32	1.71	2.06	2.48	2.78	3.44	3.71
	27	1.31	1.70	2.05	2.47	2.77	3.42	3.69
	28	1.31	1.70	2.05	2.47	2.76	3.41	3.67
	29	1.31	1.70	2.05	2.46	2.76	3.40	3.66
	30	1.31	1.70	2.04	2.46	2.75	3.39	3.65
	40	1.30	1.68	2.02	2.42	2.70	3.31	3.55
	60	1.30	1.67	2.00	2.39	2.66	3.23	3.46
	120	1.29	1.66	1.98	2.36	2.62	3.16	3.37
<i>z</i> critical values	∞	1.28	1.645	1.96	2.33	2.58	3.09	3.29

