STATISTICS

for BUSINESS and ECONOMICS

14TH EDITION

James T. McClave • P. George Benson • Terry Sincich

APPLET CORRELATION

Applet	Concept Illustrated	Description	Applet Activity
Random numbers	Uses a random number generator to determine the experimental units to be included in a sample.	Generates random numbers from a range of integers specified by the user.	**1.1**, 44; **1.2**, 44; **3.6**, 198; **4.1**, 219; **4.2**, 219; **4.8**, 271
Sample from a population	Assesses how well a sample represents the population and the role that sample size plays in the process.	Produces random sample from population from specified sample size and population distribution shape. Reports mean, median, and standard deviation; applet creates plot of sample.	**4.4**, 233; **4.6**, 257; **4.7**, 271
Sampling distributions	Compares means and standard deviations of distributions; assesses effect of sample size; illustrates undbiasedness.	Simulates repeatedly choosing samples of a fixed size n from a population with specified sample size, number of samples, and shape of population distribution. Applet reports means, medians, and standard deviations; creates plots for both.	**5.1**, 310; **5.2**, 310
Long-run probability demonstrations illustrate the concept that theoretical probabilities are long-run experimental probabilities.			
Simulating probability of rolling a 6	Investigates relationship between theoretical and experimental probabilities of rolling 6 as number of die rolls increases.	Reports and creates frequency histogram for each outcome of each simulated roll of a fair die. Students specify number of rolls; applet calculates and plots proportion of 6s.	**3.1**, 162; **3.3**, 174; **3.4**, 175; **3.5**, 188
Simulating probability of rolling a 3 or 4	Investigates relationship between theoretical and experimental probabilities of rolling 3 or 4 as number of die rolls increases.	Reports outcome of each simulated roll of a fair die; creates frequency histogram for outcomes. Students specify number of rolls; applet calculates and plots proportion of 3s and 4s.	**3.3**, 174; **3.4**, 175
Simulating the probability of heads: fair coin	Investigates relationship between theoretical and experimental probabilities of getting heads as number of fair coin flips increases.	Reports outcome of each fair coin flip and creates a bar graph for outcomes. Students specify number of flips; applet calculates and plots proportion of heads.	**3.2**, 162; **4.2**, 219
Simulating probability of heads: unfair coin $(P(H) = .2)$	Investigates relationship between theoretical and experimental probabilities of getting heads as number of unfair coin flips increases.	Reports outcome of each flip for a coin where heads is less likely to occur than tails and creates a bar graph for outcomes. Students specify number of flips; applet calculates and plots the proportion of heads.	**4.3**, 233
Simulating probability of heads: unfair coin $(P(H) = .8)$	Investigates relationship between theoretical and experimental probabilities of getting heads as number of unfair coin flips increases.	Reports outcome of each flip for a coin where heads is more likely to occur than tails and creates a bar graph for outcomes. Students specify number of flips; applet calculates and plots the proportion of heads.	**4.3**, 233
Simulating the stock market	Theoretical probabilities are long run experimental probabilities.	Simulates stock market fluctuation. Students specify number of days; applet reports whether stock market goes up or down daily and creates a bar graph for outcomes. Calculates and plots proportion of simulated days stock market goes up.	**4.5**, 234
Mean versus median	Investigates how skewedness and outliers affect measures of central tendency.	Students visualize relationship between mean and median by adding and deleting data points; applet automatically updates mean and median.	**2.1**, 88; **2.2**, 88; **2.3**, 88

(*Continued*)

Applet	Concept Illustrated	Description	Applet Activity
Standard deviation	Investigates how distribution shape and spread affect standard deviation.	Students visualize relationship between mean and standard deviation by adding and deleting data points; applet updates mean and standard deviation.	**2.4**, 96; **2.5**, 96; **2.6**, 96; **2.7**, 118
Confidence intervals for a mean (the impact of confidence level)	Not all confidence intervals contain the population mean. Investigates the meaning of 95% and 99% confidence.	Simulates selecting 100 random samples from population; finds 95% and 99% confidence intervals for each. Students specify sample size, distribution shape, and population mean and standard deviation; applet plots confidence intervals and reports number and proportion containing true mean.	**6.1**, 336; **6.2**, 336
Confidence intervals for a mean (not knowing standard deviation)	Confidence intervals obtained using the sample standard deviation are different from those obtained using the population standard deviation. Investigates effect of not knowing the population standard deviation.	Simulates selecting 100 random samples from the population and finds the 95% z-interval and 95% t-interval for each. Students specify sample size, distribution shape, and population mean and standard deviation; applet plots confidence intervals and reports number and proportion containing true mean.	**6.3**, 346; **6.4**, 346
Confidence intervals for a proportion	Not all confidence intervals contain the population proportion. Investigates the meaning of 95% and 99% confidence.	Simulates selecting 100 random samples from the population and finds the 95% and 99% confidence intervals for each. Students specify population proportion and sample size; applet plots confidence intervals and reports number and proportion containing true proportion.	**6.5**, 354; **6.6**, 354
Hypothesis tests for a mean	Not all tests of hypotheses lead correctly to either rejecting or failing to reject the null hypothesis. Investigates the relationship between the level of confidence and the probabilities of making Type I and Type II errors.	Simulates selecting 100 random samples from population; calculates and plots t statistic and P-value for each. Students specify population distribution shape, mean, and standard deviation; sample size, and null and alternative hypotheses; applet reports number and proportion of times null hypothesis is rejected at both 0.05 and 0.01 levels.	**7.1**, 397; **7.2**, 408; **7.3**, 408; **7.4**, 408
Hypothesis tests for a proportion	Not all tests of hypotheses lead correctly to either rejecting or failing to reject the null hypothesis. Investigates the relationship between the level of confidence and the probabilities of making Type I and Type II errors.	Simulates selecting 100 random samples from population; calculates and plots z-statistic and P-value for each. Students specify population proportion, sample size, and null and alternative hypotheses; applet reports number and proportion of times null hypothesis is rejected at 0.05 and 0.01 levels.	**7.5**, 424; **7.6**, 425
Correlation by eye	Correlation coefficient measures strength of linear relationship between two variables. Teaches user how to assess strength of a linear relationship from a scattergram.	Computes correlation coefficient r for a set of bivariate data plotted on a scattergram. Students add or delete points and guess value of r; applet compares guess to calculated value.	**11.2**, 682
Regression by eye	The least squares regression line has a smaller SSE than any other line that might approximate a set of bivariate data. Teaches students how to approximate the location of a regression line on a scattergram.	Computes least squares regression line for a set of bivariate data plotted on a scattergram. Students add or delete points and guess location of regression line by manipulating a line provided on the scattergram; applet plots least squares line and displays the equations and the SSEs for both lines.	**11.1**, 657

Statistics

for Business and Economics

This page is intentionally left blank

14 EDITION

GLOBAL EDITION

Statistics

for Business and Economics

James T. **MCCLAVE**
Info Tech, Inc.
University of Florida

P. George **BENSON**
College of Charleston

Terry **SINCICH**
University of South Florida

Pearson

Harlow, England • London • New York • Boston • San Francisco • Toronto • Sydney • Dubai • Singapore • Hong Kong
Tokyo • Seoul • Taipei • New Delhi • Cape Town • São Paulo • Mexico City • Madrid • Amsterdam • Munich • Paris • Milan

Credits and acknowledgments borrowed from other sources and reproduced, with permission, in this textbook appear on the appropriate page of appearance or in the Credits on pages

Cover image by MikeDotta / Shutterstock

Pearson Education Limited
KAO Two
KAO Park
Hockham Way
Harlow
Essex
CM17 9SR
United Kingdom

and Associated Companies throughout the world

Visit us on the World Wide Web at: www.pearsonglobaleditions.com

ISBN 10: 1-292-41339-5 (print)
ISBN 13: 978-1-292-41339-6 (print)
eBook ISBN 13: 978-1-292-41352-5 (eBook/uPDF)

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

1 22

Typeset in Times NR MT Pro by B2R Technologies Pvt. Ltd.

Contents

Preface

This 14th edition of *Statistics for Business and Economics* is an introductory text emphasizing inference, with extensive coverage of data collection and analysis as needed to evaluate the reported results of statistical studies and make good decisions. As in earlier editions, the text stresses the development of statistical thinking, the assessment of credibility and value of the inferences made from data, both by those who consume and by those who produce them. It assumes a mathematical background of basic algebra.

The text incorporates the following features, developed from the American Statistical Association (ASA) sponsored conferences on *Making Statistics More Effective in Schools of Business* (MSMESB) and ASA's Guidelines for Assessment and Instruction in Statistics Education (GAISE) Project:

- Emphasize statistical literacy and develop statistical thinking
- Use real data in applications
- Use technology for developing conceptual understanding and analyzing data
- Foster active learning in the classroom
- Stress conceptual understanding rather than mere knowledge of procedures
- Emphasize intuitive concepts of probability

New in the 14th Edition

- **More than 1,200 exercises, with revisions and updates to 30%.** Many new and updated exercises, based on contemporary business-related studies and real data, have been added. Most of these exercises foster critical thinking skills.
- **Data Informed Development.** The authors analyzed aggregated student usage and performance data from MyLab Statistics for the previous edition of this text. The results of this analysis helped improve the quality and quantity of exercises that matter most to instructors and students.
- **Updated technology.** All printouts from statistical software (Excel 2019/XLSTAT, StatCrunch 3.0, Minitab 19, and the TI-84 Graphing Calculator) and corresponding instructions for use have been revised to reflect the latest versions of the software.
- **Statistics in Action Cases Updated.** Three of the 14 Statistics in Action cases have been updated. All cases are based on real data from a recent business study.
- **Continued Emphasis on Ethics.** Where appropriate, boxes have been added to emphasize the importance of ethical behavior when collecting, analyzing, and interpreting data with statistics.
- **Business Analytics.** The importance of statistical thinking to successful business analytics is established early in the text.
- **Short Video Tutorials.** New videos guide students through real-life applications of chapter topics to illustrate how these concepts translate to everyday life.

Hallmark Strengths

We have maintained the pedagogical features of *Statistics for Business and Economics* that we believe make it unique among introductory business statistics texts. These features, which assist the student in achieving an overview of statistics and an understanding of its relevance in both the business world and everyday life, are as follows:

- **Use of Examples as a Teaching Device** Almost all new ideas are introduced and illustrated by data-based applications and examples. We believe that students better understand definitions, generalizations, and theoretical concepts *after* seeing an application. All examples have three components: (1) "Problem," (2) "Solution," and (3) "Look Back" (or "Look Ahead"). This step-by-step process provides students with a defined structure by which to approach problems and enhances their problem-solving skills. The "Look Back/Look Ahead" feature often gives helpful hints to solving the problem and/or provides a further reflection or insight into the concept or procedure that is covered.
- **Now Work** A "Now Work" exercise suggestion follows each example. The Now Work exercise (marked with the NW icon in the exercise sets) is similar in style and concept to the text example. This provides students with an opportunity to immediately test and confirm their understanding.
- **Statistics in Action** Each chapter begins with a case study based on an actual contemporary, controversial or high-profile issue in business. Relevant research questions and data from the study are presented and the proper analysis is demonstrated in short "Statistics in Action Revisited" sections throughout the chapter. These motivate students to critically evaluate the findings and think through the statistical issues involved.
- **"Hands-On" Activities for Students** At the end of each chapter, students are provided with an opportunity to participate in hands-on classroom activities, ranging from real data collection to formal statistical analysis. These activities are designed to be performed by students individually or as a class.
- **Applet Exercises.** The text is accompanied by applets (short computer programs), available on the student resource site (www.pearsonglobaleditions.com) and in MyLab Statistics. These point-and-click applets allow students to easily run simulations that visually demonstrate some of the more difficult statistical concepts (e.g., sampling distributions and confidence intervals.) Each chapter contains several optional applet exercises in the exercise sets. They are denoted with the following Applet icon:  .
- **Real-World Business Cases** Seven extensive business problem-solving cases, with real data and assignments for the student, are provided. Each case serves as a good capstone and review of the material that has preceded it. Typically, these cases follow a group of two or three chapters and require the student to apply the methods presented in these chapters.
- **Real Data–Based Exercises** The text includes more than 1,200 exercises based on applications in a variety of business disciplines and research areas. All applied exercises use current real data extracted from current publications (e.g., newspapers, magazines, current journals, and the Internet). Some students have difficulty learning the mechanics of statistical techniques when all problems are couched in terms of realistic applications. For this reason, all exercise sections are divided into at least four parts:

  Learning the Mechanics. Designed as straightforward applications of new concepts, these exercises allow students to test their ability to comprehend a mathematical concept or a definition.

  Applying the Concepts—Basic. Based on applications taken from a wide variety of business journals, newspapers, and other sources, these short exercises help students to begin developing the skills necessary to diagnose and analyze real-world problems.

  Applying the Concepts—Intermediate. Based on more detailed real-world applications, these exercises require students to apply their knowledge of the technique presented in the section.

Applying the Concepts—Advanced. These more difficult real-data exercises require students to use their critical thinking skills.

Critical Thinking Challenges. Placed at the end of the "Supplementary Exercises" section only, this feature presents students with one or two challenging business problems.

- **Exploring Data with Statistical Computer Software and the Graphing Calculator** Each statistical analysis method presented is demonstrated using output from three leading Windows-based statistical software packages: Excel/XLSTAT, StatCrunch, and Minitab. Students are exposed early and often to computer printouts they will encounter in today's hi-tech business world.
- **"Using Technology" Tutorials** At the end of each chapter are statistical software tutorials with point-and-click instructions (with screen shots) for Minitab, StatCrunch, and Excel/XLSTAT. These tutorials are easily located and show students how to best use and maximize statistical software. In addition, output and keystroke instructions for the TI-84 Graphing Calculator are presented.
- **Profiles of Statisticians in History (Biography)** Brief descriptions of famous statisticians and their achievements are presented in side boxes. In reading these profiles, students will develop an appreciation for the statistician's efforts and the discipline of statistics as a whole.
- **Data and Applets** The text is accompanied by a website (www.pearsonglobaleditions.com) that contains files for all of the data sets marked with an icon  in the text. These include data sets for text examples, exercises, Statistics in Action, and Real-World cases. Data files are available in multiple formats: Excel and Minitab. This website also contains the applets that are used to illustrate statistical concepts.

Flexibility in Coverage

The text is written to allow the instructor flexibility in coverage of topics. Suggestions for two topics, probability and regression, are given below.

- **Probability and Counting Rules** One of the most troublesome aspects of an introductory statistics course is the study of probability. Probability poses a challenge for instructors because they must decide on the level of presentation, and students find it a difficult subject to comprehend. We believe that one cause for these problems is the mixture of probability and counting rules that occurs in most introductory texts. Consequently, we have included the counting rules (with examples) in an appendix (Appendix B) rather than in the body of Chapter 3. Thus, the instructor can control the level of coverage of probability.
- **Multiple Regression and Model Building** This topic represents one of the most useful statistical tools for the solution of applied problems. Although an entire text could be devoted to regression modeling, we feel that we have presented coverage that is understandable, usable, and much more comprehensive than the presentations in other introductory statistics texts. We devote two full chapters to discussing the major types of inferences that can be derived from a regression analysis, showing how these results appear in the output from statistical software, and, most important, selecting multiple regression models to be used in an analysis. Thus, the instructor has the choice of a one-chapter coverage of simple linear regression (Chapter 11), a two-chapter treatment of simple and multiple regression (excluding the sections on model building in Chapter 12), or complete coverage of regression analysis, including model building and regression diagnostics. This extensive coverage of such useful statistical tools will provide added evidence to the student of the relevance of statistics to real-world problems.
- **Role of Calculus in Footnotes** Although the text is designed for students with a non-calculus background, **footnotes** explain the role of calculus in various derivations. Footnotes are also used to inform the student about some of the theory underlying certain methods of analysis. These footnotes allow additional flexibility in the mathematical and theoretical level at which the material is presented.

Acknowledgments

This book reflects the efforts of a great many people over a number of years. First, we would like to thank the following professors, whose reviews and comments on this and prior editions have contributed to the 14th edition:

Reviewers of the 14th Edition of *Statistics for Business and Economics*

Anna Errore, University of Minnesota Twin Cities
Alka Gandhi, University of Maryland-College Park
Stacey Hachigian, Humber College
Seunghee Lee, Pellissippi State Community College
Amit Mitra, Auburn University
Yu Yue, Baruch College

Reviewers of Previous Editions

ALABAMA Volodymyr Melnykov, *University of Alabama–Tuscaloosa* **ARKANSAS** Julie Trivitt, *University of Arkansas* **CALIFORNIA** Joyce Curley-Daly, Jim Daly, Robert K. Smidt, *California Polytechnic State University* • Jim Davis, *Golden Gate University* • Carol Eger, *Stanford University* • Paul W. Guy, *California State University, Chico* • Judd Hammack, P. Kasliwal, *California State University, Los Angeles* • Mabel T. King, *California State University, Fullerton* • James Lackritz, *California State University, San Diego* • Beth Rose, *University of Southern California* • Daniel Sirvent, *Vanguard University* **COLORADO** Rick L. Edgeman, Charles F. Warnock, *Colorado State University* • Eric Huggins, *Fort Lewis College* • William J. Weida, *United States Air Force Academy* **CONNECTICUT** Alan E. Gelfand, Joseph Glaz, Timothy J. Killeen, *University of Connecticut* **DELAWARE** Christine Ebert, *University of Delaware* **DISTRICT OF COLUMBIA** Phil Cross, Jose Luis Guerrero-Cusumano, *Georgetown University* • Gaminie Meepagala, *Howard University* **FLORIDA** John M. Charnes, *University of Miami* • C. Brad Davis, *Clearwater Christian College* • Vivian Jones, *Bethune-Cookman University* • Fred Leysieffer, Pi-Erh Lin, Doug Zahn, *Florida State University* • P. V. Rao, *University of Florida* • Laura Reisert, *Florida International University* • Jeffrey W. Steagall, *University of North Florida* • Edna White, *Florida Atlantic University* **GEORGIA** Robert Elrod, *Georgia State University* • Karen Smith, *West Georgia University* **HAWAII** Steve Hora, *University of Hawaii, Hilo* **ILLINOIS** Arunas Dagys, *St. Xavier University* • Edward Minieka, *University of Illinois at Chicago* • Don Robinson, *Illinois State University* • Chipei Tseng, *Northern Illinois University* • Pankaj Vaish, *Arthur Andersen & Company* **IOWA** Dileep Dhavale, *University of Northern Iowa* • William Duckworth II, William Q. Meeker, *Iowa State University* • Tim E. McDaniel, *Buena Vista University* **KANSAS** Paul I. Nelson, *Kansas State University* • Lawrence A. Sherr, *University of Kansas* **KENTUCKY** Richard N. McGrath, *Bowling Green State University* **LOUISIANA** James Willis, *Louisiana State University* **MARYLAND** John F. Beyers, Michael Kulansky, *University of Maryland–University College* • Glenn J. Browne, Mary C. Christman, *University of Maryland* **MASSACHUSETTS** Warren M. Holt, *Southeastern Massachusetts University* • Remus Osan, *Boston University* **MICHIGAN** Atul Agarwal, Petros Ioannatos, *GMI Engineering and Management Institute* • Richard W. Andrews, Peter Lenk, Benjamin Lev, *University of Michigan* • Leszek Gawarecki, *Kettering University* • Toni M. Somers, *Wayne State University* • William Welch, *Saginaw Valley State University* • T. J. Wharton, *Oakland University* **MINNESOTA** Gordon J. Alexander, Donald W. Bartlett, David M. Bergman, Atul Bhatia, Steve Huchendorf, Benny Lo, Karen Lundquist, Vijay Pisharody, Donald N. Steinnes, Robert W. Van Cleave, Steve Wickstrom, *University of Minnesota* • Daniel G. Brick, Leigh Lawton, *University of St. Thomas* • Susan Flach, *General Mills, Inc.* • David D. Krueger, Ruth K. Meyer, Jan Saraph, Gary Yoshimoto, *St. Cloud State University* • Paula M. Oas, *General Office Products* • Fike Zahroom, *Moorhead State University* **MISSISSIPPI** Eddie M. Lewis, *University of Southern Mississippi* • Alireza Tahai, *Mississippi State University* **MISSOURI** James Holstein, Lawrence D. Ries,

University of Missouri, Columbia • Marius Janson, L. Douglas Smith, *University of Missouri, St. Louis* • Farroll Tim Wright, *University of Missouri* • Stephanie Schartel-Dunn *Missouri Western University* **NEBRASKA** James Wright, *Chadron State College* **NEW HAMPSHIRE** Ken Constantine, *University of New Hampshire* **NEW JERSEY** Lewis Coopersmith, Cengiz Haksever, *Rider University* • Lei Lei, Xuan Li, Zina Taran, *Rutgers University* • Philip Levine, Leonard Presby, *William Paterson University* **NEW MEXICO** S. Howard Kraye, *University of New Mexico* **NEW YORK** James Czachor, *Fordham-Lincoln Center, AT&T* • Bernard Dickman, *Hofstra University* • Joshua Fogel, *Brooklyn College of City University of New York* • Martin Labbe, *State University of New York, College at New Paltz* • Kenneth Leong, *College of New Rochelle* • Mark R. Marino, *Niagara University/Erie Community College* • G. E. Martin, *Clarkson University* • Thomas J. Pfaff, *Ithaca College* • Gary Simon, *New York University, Stern School of Business* • Rungrudee Suetorsak, *SUNY-Fredonia* **NORTH CAROLINA** Golam Azam, *North Carolina Agricultural & Technical University* • Edward Carlstein, Douglas A. Elvers, *University of North Carolina at Chapel Hill* • Barry P. Cuffe, *Wingate University* • Don Holbert, *East Carolina University* • J. Morgan Jones, *University of North Carolina* • Douglas S. Shafer, *University of North Carolina, Charlotte* **OHIO** William H. Beyer, *University of Akron* • Michael Broida, Tim Krehbiel, *Miami University of Ohio* • Chih-Hsu Cheng, Douglas A. Wolfe, *Ohio State University* • Ronald L. Coccari, *Cleveland State University* • Richard W. Culp, *Wright-Patterson AFB, Air Force Institute of Technology* **OKLAHOMA** Larry Claypool, Brenda Masters, Rebecca Moore, *Oklahoma State University* • Robert Curley, *University of Central Oklahoma* **PENNSYLVANIA** Mohammed Albohali, Douglas H. Frank, *Indiana University of Pennsylvania* • Sukhwinder Bagi, *Bloomsburg University* • Carl Bedell, *Philadelphia College of Textiles and Science* • Ann Hussein, *Philadelphia University* • Behnam Nakhai, *Millersville University* • Rose Prave, *University of Scranton* • Farhad Saboori, *Albright College* • Kathryn Szabet, *LaSalle University* • Pandu Tadikamalla *University of Pittsburgh* • Christopher J. Zappe, *Bucknell University* **SOUTH CAROLINA** Iris Fetta, Robert Ling, *Clemson University* • Kathleen M. Whitcomb, *University of South Carolina* **TENNESSEE** Francis J. Brewerton, *Middle Tennessee State University* **TEXAS** Larry M. Austin, *Texas Tech University* • Jim Branscome, Robert W. Brobst, Mark Eakin, Grace Esimai, Michael E. Hanna, Craig W. Slinkman, *University of Texas at Arlington* • Katarina Jegdic, *University of Houston–Downtown* • Virgil F. Stone, *Texas A & M University* **VIRGINIA** Edward R. Clayton, *Virginia Polytechnic Institute and State University* **WASHINGTON** June Morita, Kim Tamura, *University of Washington* **WISCONSIN** Ross H. Johnson, *Madison College* **WASHINGTON, D.C.** Keith Ord, *Georgetown University* • Balaji Srinivasan, *George Washington University* **CANADA** Clarence Bayne, *Concordia University* • Edith Gombay, *University of Alberta* **TURKEY** Dilek Onkal, *Bilkent University, Ankara* **OTHER** Michael P. Wegmann, *Keller Graduate School of Management*

Other Contributors

Special thanks are due to our supplements author, Mark Dummeldinger, who has worked with us for many years. Accuracy checker Engin Sungur helped ensure a highly accurate, clean text. Nati Jain and Shantel Vargas diligently researched new data and cases for the exercise sets. Finally, the Pearson Education staff of Amanda Brands, Suzanna Bainbridge, Karen Montgomery, Alicia Wilson, Demetrius Hall, Joe Vetere, Peggy McMahon, Jean Choe, and Bob Carroll, who helped greatly with all phases of the text development, production, and marketing effort. Our gratitude also goes to Heidi Aguiar of Spi Global for overseeing the production process.

Global Edition Acknowledgments

Pearson would like to thank Alicia Tan Yiing Fei, Taylor's University Malaysia, for developing content for this Global Edition. Pearson would also like to thank Simon Trimborn, City University of Hong Kong; and John Alexander Wright, The Chinese University of Hong Kong, for sharing suggestions that were valuable to us in developing the content for the Global Edition.

MyLab Statistics Resources for Success

MyLab Statistics is available to accompany Pearson's market-leading text options, including Statistics for Business and Economics, 14th Edition (access code required).

MyLab™ is the teaching and learning platform that empowers you to reach every student. MyLab Statistics combines trusted author content—including full eText and assessment with immediate feedback—with digital tools and a flexible platform to personalize the learning experience and improve results for each student. Integrated with StatCrunch®, a web-based statistical software program, students learn the skills they need to interact with data in the real world.

Student Resources

Excel Technology Manual, by Mark Dummeldinger (University of South Florida), provides tutorial instruction and worked-out text examples for Excel. The *Excel Technology Manual* is available for download at www.pearsonglobaleditions.com or within MyLab Statistics.

NEW! Videos by author Terry Sincich have been added to the Video and Resource Library. These videos illustrate the solutions to select exercises in the book, demonstrate key topics in detail, and include tutorials on real-world statistics in action.

Accessibility

Pearson works continuously to ensure our products are as accessible as possible to all students. Currently we work toward achieving WCAG 2.0 AA for our existing products (2.1 AA for future products) and Section 508 standards, as expressed in the Pearson Guidelines for Accessible Educational Web Media.

For the Instructor

Instructor's Solutions Manual, by Mark Dummeldinger (University of South Florida), provides detailed, worked-out solutions to all of the book's exercises. Careful attention has been paid to ensure that all methods of solution and notation are consistent with those used in the core text. Available for download at www.pearsonglobaleditions.com.

TestGen® (www.pearsoned.com/testgen) enables instructors to build, edit, print, and administer tests using a computerized bank of questions developed to cover all the objectives of the text. TestGen is algorithmically based, allowing instructors to create multiple but equivalent versions of the same question or test with the click of a button. Instructors can also modify test bank questions or add new questions. The software and testbank are available for download from Pearson's Instructor Resource Center.

Student's Solutions Manual, by Mark Dummeldinger (University of South Florida), provides detailed, worked-out solutions to all odd-numbered text exercises. This item is available within MyLab Statistics and can be shared by the instructor only.

Data Analytics

Instructors have a comprehensive gradebook with enhanced reporting functionality that makes it easier to understand which students are struggling, and which topics they struggle with most.

pearson.com/mylab/statistics

1

CONTENTS

WHERE WE'RE GOING

- Introduce the field of statistics (1.1)
- Demonstrate how statistics applies to business (1.2)
- Introduce the language of statistics and the key elements of any statistical problem (1.3)
- Differentiate between population and sample data (1.3)
- Differentiate between descriptive and inferential statistics (1.3)
- Introduce the key elements of a process (1.4)
- Identify the different types of data and data-collection methods (1.5–1.6)
- Discover how critical thinking through statistics can help improve our quantitative literacy (1.7)

Statistics, Data, and Statistical Thinking

STATISTICS IN ACTION

A 20/20 View of Surveys and Studies: Facts or Fake News?

Several years ago, the popular ABC television program 20/20 aired a story titled "Fact or Fiction?—Exposés of So-Called Surveys." The focus of the program segment was on what we now call "fake news" or "alternative facts," that is, false information that is often highly publicized in the media (newspapers, magazines, TV shows, Twitter, Instagram, etc.). Several misleading (and possibly unethical) surveys, conducted by businesses or special interest groups with specific objectives in mind, were presented on the ABC program. Several are listed in Table SIA1.1, as well as some recent misleading studies used in product advertisements.

For this *Statistics in Action*, we consider research sponsored by Mars Corp. and published in the journal *Nature Neuroscience* (Dec. 2014). Researchers from Columbia University wondered whether taking cocoa supplements would enhance a region of the brain that deteriorates with age and is associated with age-related memory loss. They concluded that cocoa supplements can indeed boost cognition in older adults. Other similar studies claim that chocolate will reduce cardiovascular disease risk and help with weight loss. These results were reported on by media outlets such as the *New York Times*, with headlines like, "*To improve a memory, consider chocolate!*", "*Good news for chocolate lovers: The more you eat, the lower your risk of heart disease*," or simply "*Chocolate is good for you.*" These reported "facts" have likely grown consumer demand for chocolate. At a time when candy sales overall has declined, chocolate retail sales in the United States have risen from $14.2 billion in 2007 to $18.9 billion in 2017.

Critical thinkers may question chocolate as a powerful health food. One such group, Vox Media (November 7, 2017), investigated this chocolate phenomenon.

STATISTICS IN ACTION

(*continued*)

First, Vox discovered that chocolate companies like Mars, "have spent millions of dollars for scientific studies and research grants that support cocoa science. And, of 100 [recent] Mars-sponsored studies on cocoa, chocolate, and health, 98 had conclusions that were favorable to the candy maker in some way." This unusually high percentage of favorable studies, according to Vox, "raises questions about the quality of the studies, given that Mars and other chocolate makers can use the positive findings to market their products." This finding motivated Vox Media to critically analyze the *Nature Neuroscience* chocolate study.

To conduct this study, the Vox Media researchers randomly assigned 37 people to one of four groups. Each subject in group 1 was given a high daily dose (900 mg) of cocoa flavanol supplements and assigned one hour of aerobic exercise four times per week. Subjects in group 2 received the same high dose of cocoa flavanol supplements but were not assigned to exercise. Group 3 subjects received a low dose of cocoa flavanols (10 mg) and were assigned one hour of aerobic exercise four times per week. Finally, the last group received a low dose of cocoa flavanols but was not assigned exercise (See Figure SIA1.1). After a 3-month period, the researchers tested whether cocoa flavanol supplements staved off cognitive decline in a region of the brain associated with age-related memory loss. They did this by measuring brain waves in an MRI machine and by using an object-recognition task to test memory and reaction time. The researchers also tested if exercise had any effect on memory.

Group 1: High Cocoa / Exercise (8 subjects)	*Group 2*: High Cocoa / No Exercise (11 subjects)
Group 3: Low Cocoa / Exercise (9 subjects)	*Group 4*: Low Cocoa / No Exercise (9 subjects)

Figure SIA1.1
Schematic of Chocolate Study

The researchers reported that exercise had no impact on brain function—but cocoa flavanols did. Subjects receiving a high cocoa flavanol dosage demonstrated a greater improvement in cognitive performance than those in the low dosage groups. However, as reported by Vox Media, the researchers drew conclusions that went beyond the scope of the study. For example, the researchers claimed that the effects they saw in the high-flavanol group demonstrated that cocoa could reverse age-related memory decline by 30 years. Vox also discovered problems with the study's small sample size and design.

Henry Drysdale, a doctor and fellow at Oxford University's Center for Evidence-Based Medicine elaborated on the design issue. First, he warned that eating cocoa supplements in order to improve memory in three months is not relevant to real-world age-related memory decline. Second, the doctor pointed out the need for a much larger (than 37) group of study participants, and to conduct the trial for several years. Finally, Drysdale commented on the study variables: "Instead of only tracking the study participants' brain waves in an MRI machine (which is not a measure of cognitive ability), or using an object recognition task to test memory, you'd also want to measure outcomes that matter in people's lives, like, whether those taking cocoa could remember what they did that morning or that they had a doctor's appointment next week better than the people who didn't take the cocoa."

Ultimately, Vox Media stated that, "this trial only demonstrated that supplements seem to enhance brain function over a period of weeks, and only according to a very specific (and not very widely used) test of cognitive function. That is far from valid proof that cocoa is a memory enhancer."

To conclude the introduction to this *Statistics in Action*, we return to the ABC *20/20* TV program segment. The segment ended with an interview of Cynthia Crossen, author of *Tainted Truth: The Manipulation of Fact in America*, an exposé of misleading and biased surveys.

Some 20 years before the term "fake news" was coined, Crossen warned, "If everybody is misusing numbers and scaring us with numbers to get us to do something,

however good [that something] is, we've lost the power of numbers. Now, we know certain things from research. For example, we know that smoking cigarettes is hard on your lungs and heart, and because we know that, many people's lives have been extended or saved. We don't want to lose the power of information to help us make decisions, and that's what I worry about."

Table SIA1.1: Examples of "Fake News"

Fake News (Source)	Actual Study Information/Flaw
1. Eating oat bran is a cheap and easy way to reduce your cholesterol. (*Quaker Oats*)	Diet must consist of nothing but oat bran to reduce your cholesterol count.
2. One in four American children under age 12 is hungry or at risk of hunger. (*Food Research and Action Center*)	Based on responses to questions: "Do you ever cut the size of meals?" "Do you ever eat less than you feel you should?" "Did you ever rely on limited numbers of foods to feed your children because you were running out of money to buy food for a meal?"
3. There is a strong correlation between a CEO's golf handicap and the company's stock performance: The lower the CEO's handicap (i.e., the better the golfer), the better the stock performs. (*New York Times*, May 31, 1998)	Survey sent to CEOs of 300 largest US companies; only 51 revealed their golf handicaps. Data for several top-ranking CEOs were excluded from the analysis.
4. Prior to the passing of the federal government's health reform act, 30% of employers are predicted to "definitely" or "probably" stop offering health coverage. (*McKinsey & Company Survey*, February 2011)	Online survey of 1,329 private-sector employers in the United States. Respondents were asked leading questions that made it logical to stop offering health insurance.
5. In an advertisement, "more than 80% of dentists surveyed recommend Colgate toothpaste to patients." (*Colgate-Palmolive Company*, January 2007)	The survey allowed each dentist to recommend more than one toothpaste.The Advertising Standards Authority cited and fined Colgate for a misleading ad (implying 80% of dentists recommend Colgate toothpaste in preference to all other brands) and banned the advertisement.
6. An advertisement for Kellogg's Frosted Mini-Wheats claimed that the cereal was "clinically shown to improve kids' attentiveness by nearly 20%." (*Kellogg Company*, 2009)	Only half of the kids in the study showed any improvement in attentiveness; only 1 in 7 improved by 18% or more, and only 1 in 9 improved by 20% or more; kids who ate Frosted Mini-Wheats were compared against kids who had only water for breakfast. (The Kellogg Company agreed to pay $4 million to settle suit over false ad claim.)
7. On the basis of a commissioned study, Walmart advertised that it "was responsible for an overall 3.1% decline in consumer prices" and it "saves customers over $700 per year." (*Global Insight*, 2005)	The Economic Policy Institute noted that the Global Insight study was based on the retailer's impact on the Consumer Price Index (CPI)—but 60% of the items in the CPI are services, not commodities that can be purchased at Walmart. (Walmart was forced to withdraw the misleading advertisement.)
8. In a survey commissioned by cable provider Comcast, respondents were asked to decide which cable provider, Comcast or DIRECTV, offered more HD channels. Respondents were shown channel lists for DIRECTV (List #387) and Comcast (List #429). (*NAD Case Report No. 5208*, August 25, 2010).	The National Advertising Division (NAD) of the Council of Better Business Bureaus rejected the survey after finding that the higher list number (#429) "served as a subtle, yet effective cue" that Comcast's list contained more channels.

STATISTICS IN ACTION
(*continued*)

Fake News (Source)	Actual Study Information/Flaw
9. NPR reported on a recent study that found that teens who spend five or more hours per day on their smartphones are 71 percent more likely to have one risk factor for suicide/depression (*Journal of Abnormal Psychology*, March 2019).	Data was based on teen's perceived use of their smartphones. Studies have shown that perceived use is poorly related to actual use measured with an app. Also, the measures of addiction or suicide risk were based on unreliable scales.

In the following *Statistics in Action Revisited* sections, we discuss several key statistical concepts covered in this chapter that are relevant to misleading surveys and studies.

STATISTICS IN ACTION REVISITED

- Identifying the population, sample, and inference **(p. 29)**
- Identifying the data-collection method and data type **(p. 39)**
- Critically assessing the ethics of a statistical study **(p. 42)**

1.1 The Science of Statistics

What does *statistics* mean to you? Does it bring to mind batting averages? Gallup polls, unemployment figures, or numerical distortions of facts (lying with statistics)? Or is it simply a college requirement you have to complete? We hope to persuade you that statistics is a meaningful, useful science with a broad scope of applications to business, government, and the physical and social sciences that is almost limitless. We also want to show that statistics can lie only when they are misapplied. Finally, we wish to demonstrate the key role statistics play in critical thinking—whether in the classroom, on the job, or in everyday life. Our objective is to leave you with the impression that the time you spend studying this subject will repay you in many ways.

Although the term can be defined in many ways, a broad definition of *statistics* is the science of collecting, classifying, analyzing, and interpreting information. Thus, a statistician isn't just someone who calculates batting averages at baseball games or tabulates the results of a Gallup poll. Professional statisticians are trained in *statistical science*—that is, they are trained in collecting information in the form of **data,** evaluating it, and drawing conclusions from it. Furthermore, statisticians determine what information is relevant in a given problem and whether the conclusions drawn from a study are to be trusted.

Statistics is the science of data. It involves collecting, classifying, summarizing, organizing, analyzing, and interpreting numerical and categorical information.

In the next section, you'll see several real-life examples of statistical applications in business and government that involve making decisions and drawing conclusions.

1.2 Types of Statistical Applications in Business

Statistics means "numerical descriptions" to most people. Monthly unemployment figures, the failure rate of startup companies, and the proportion of female executives in a particular industry all represent statistical descriptions of large sets of data collected on some phenomenon. Often the data are selected from some larger set of data that has characteristics we wish to estimate. We call this selection process *sampling*. For example,

BIOGRAPHY

FLORENCE NIGHTINGALE (1820–1910)
The Passionate Statistician

In Victorian England, the "Lady of the Lamp" had a mission to improve the squalid field hospital conditions of the British army during the Crimean War. Today, most historians consider Florence Nightingale to be the founder of the nursing profession. To convince members of the British Parliament of the need for supplying nursing and medical care to soldiers in the field, Nightingale compiled massive amounts of data from the army files. Through a remarkable series of graphs (which included the first "pie chart"), she demonstrated that most of the deaths in the war were due to illnesses contracted outside the battlefield or long after battle action from wounds that went untreated. Florence Nightingale's compassion and self-sacrificing nature, coupled with her ability to collect, arrange, and present large amounts of data, led some to call her the "Passionate Statistician."

you might collect the ages of a sample of customers of a video streaming services company to estimate the average age of *all* customers of the company. Then you could use your estimate to target the firm's advertisements to the appropriate age group. Notice that statistics involves two different processes: (1) describing sets of data and (2) drawing conclusions (making estimates, decisions, predictions, etc.) about the sets of data based on sampling. So, the applications of statistics can be divided into two broad areas: *descriptive statistics* and *inferential statistics*.

Descriptive statistics utilizes numerical and graphical methods to explore data, i.e., to look for patterns in a data set, to summarize the information revealed in a data set, and to present the information in a convenient form for the user.

Inferential statistics utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data.

Although we'll discuss both descriptive and inferential statistics in the following chapters, the primary theme of the text is **inference.**

Let's begin by examining some business studies that illustrate applications of statistics.

Study 1.1 "Best-Selling Girl Scout Cookies" (Source: www.girlscouts.org): Since 1917, the Girl Scouts of America have been selling boxes of cookies. In 2017, there were 12 varieties for sale: Thin Mints, Samoas, Lemonades, Tagalongs, Do-si-dos, Trefoils, Savannah Smiles, Thanks-A-Lot, Dulce de Leche, Cranberry Citrus Crisps, Chocolate Chip, and Thank U Berry Much. Each of the approximately 150 million boxes of Girl Scout cookies sold in 2017 was classified by variety. The results are summarized in Figure 1.2. From the graph, you can clearly see that the best-selling variety is Thin Mints (25%), followed by Samoas (19%) and Tagalongs (13%). Since Figure 1.1 *describes* the variety of categories of the boxes of Girl Scout cookies sold, the graphic is an example of *descriptive statistics.*

Study 1.2 "Executive Compensation vs. Typical Worker Pay" (*Source*: 24/7 Wall Street, *USA Today*, October 15, 2018): How big is the gap between what a firm pays its CEO and what it pays its typical worker? To answer this question, 24/7 Wall Street reviewed the ratio between annual CEO base pay and typical worker salary at 168 large US companies, using data from benefits and compensation information provided by the website Payscale. This information was used to compute the ratio of CEO pay to the typical worker salary

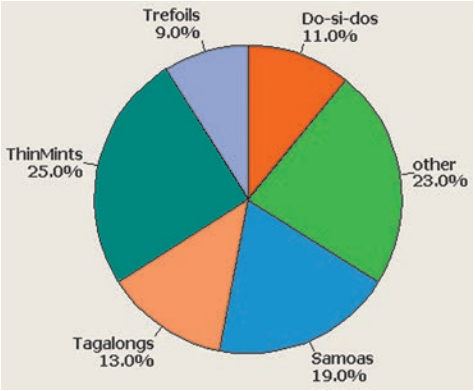


Figure 1.1
Best-selling Girl Scout cookies
Source: "Best-Selling Girl Scout Cookies," based on www.girlscouts.org.

at each company.* The data for the 10 companies with the highest ratio in the sample of 168 companies in the study are shown in Table 1.1. An analysis of the data for all 168 firms revealed that the "average" ratio of CEO pay to typical worker pay was 205.† In other words, on average, CEOs in the sample earn around 205 times what their firm's typical worker earns. Armed with this sample information, an economist might *infer* that the average ratio of CEO pay to typical worker pay for *all* US firms is 205. Thus, this study is an example of *inferential statistics*.

Table 1.1 Ratio of CEO Compensation to Typical Worker Pay

Company	CEO	CEO Base Pay	CEO Total Compensation	Typical Worker Pay	Ratio
1 CVS Health	Larry J. Merlo	$12,105,481	$22,855,374	$27,900	434
2 CBS Corp.	Leslie Moonves	$23,652,883	$56,352,801	$59,900	395
3 Walt Disney	Robert A. Iger	$26,208,003	$43,490,567	$71,400	367
4 TGX Comp.	Carol Meyrowitz	$7,330,584	$17,962,232	$22,400	327
5 21st Century Fox	K. Rupert Murdoch	$17,047,636	$22,192,923	$54,800	311
6 Comcast	Brian L. Roberts	$16,819,942	$27,520,744	$55,800	301
7 L Brands	Leslie H. Wexner	$9,665,925	$26,669,306	$33,900	285
8 Honeywell Int.	David M. Cote	$22,767,851	$33,105,851	$81,600	279
9 PepsiCo	Indra K. Nooyi	15,937,828	$22,189,307	$61,500	259
10 Wynn Resorts	Stephen A. Wynn	$11,930,391	$20,680,391	$50,100	238

Source: 24/7 Wall Street, *USA Today*, Oct. 15, 2018.

Study 1.3 "Does rudeness really matter in the workplace?" (*Academy of Management Journal*, October 2007): Many studies have established that rudeness in the workplace can lead to retaliatory and counterproductive behavior. However, there has been little research on how rude behaviors influence a victim's task performance. In one study, college students enrolled in a management course were randomly assigned to one of two experimental conditions: rudeness condition (45 students) and control group (53 students). Each student was asked to write down as many uses for a brick as possible in 5 minutes; this value (total number of uses) was used as a performance measure for each student. For those students in the rudeness condition, the facilitator displayed rudeness by berating the students in general for being irresponsible and unprofessional (due to a late-arriving associate of the researchers). No comments were made about the late-arriving associate of the researchers to students in the control group. As you might expect, the researchers discovered that the performance levels for students in the rudeness condition were generally lower than the performance levels for students in the control group; thus, they concluded that rudeness in the workplace negatively affects job performance. As in Study 1.2, this study is an example of the use of inferential statistics. The researchers used data collected on 98 college students in a simulated work environment to make an inference about the performance levels of all workers exposed to rudeness on the job.

These studies provide three real-life examples of the uses of statistics in business, economics, and management. Notice that each involves an analysis of data, either for the purpose of describing the data set (Study 1.1) or for making inferences about a data set (Studies 1.2 and 1.3).

*The ratio was calculated using the *median* worker salary at each firm. A formal definition of median is given in Chapter 2. For now, think of the median as the *typical* salary for a worker, i.e., one that falls in the middle of all worker salaries.

†A formal definition of *average* is also given in Chapter 2. Like the median, think of the average as another way to express the *middle* salary.

1.3 Fundamental Elements of Statistics

Statistical methods are particularly useful for studying, analyzing, and learning about *populations* of *experimental units*.

An **experimental (or observational) unit** is an object (e.g., person, thing, transaction, or event) upon which we collect data.

A **population** is a set of units (usually people, objects, transactions, or events) that we are interested in studying.

For example, populations may include (1) *all* employed workers in the United States; (2) *all* registered voters in California; (3) *everyone* who has purchased a particular brand of cell phone; (4) *all* the cars produced last year by a particular assembly line; (5) the *entire* stock of spare parts at United Airlines' maintenance facility; (6) *all* sales made at the drive-through window of a McDonald's restaurant during a given year; and (7) the set of *all* accidents occurring on a particular stretch of interstate during a holiday period. Notice that the first three population examples (1–3) are sets (groups) of people, the next two (4–5) are sets of objects, the next (6) is a set of transactions, and the last (7) is a set of events. Also notice that *each set includes all the experimental units in the population* of interest.

In studying a population, we focus on one or more characteristics or properties of the experimental units in the population. We call such characteristics *variables*. For example, we may be interested in the variables age, gender, income, and/or the number of years of education of the people currently unemployed in the United States.

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

The name *variable* is derived from the fact that any particular characteristic may vary among the experimental units in a population.

In studying a particular variable, it is helpful to be able to obtain a numerical representation for it. Often, however, numerical representations are not readily available, so the process of measurement plays an important supporting role in statistical studies. **Measurement** is the process we use to assign numbers to variables of individual population units. We might, for instance, measure the preference for a food product by asking a consumer to rate the product's taste on a scale from 1 to 10. Or we might measure workforce age by simply asking each worker, "How old are you?" In other cases, measurement involves the use of instruments such as timers, scales, and calipers.

If the population we wish to study is small, it is possible to measure a variable for every unit in the population. For example, if you are measuring the starting salary for all University of Michigan MBA graduates last year, it is at least feasible to obtain every salary. When we measure a variable for every experimental unit of a population, the result is called a **census** of the population. Typically, however, the populations of interest in most applications are much larger, involving perhaps many thousands or even an infinite number of units. Examples of large populations include the seven listed above, as well as all invoices produced in the last year by a *Fortune* 500 company, all potential buyers of a new iPad, and all stockholders of a firm listed on the New York

Stock Exchange. For such populations, conducting a census would be prohibitively time-consuming and/or costly. A reasonable alternative would be to select and study a *subset* (or portion) of the units in the population.

A **sample** is a subset of the units of a population.

For example, suppose a company is being audited for invoice errors. Instead of examining all 15,472 invoices produced by the company during a given year, an auditor may select and examine a sample of just 100 invoices (see Figure 1.2). If he is interested in the variable "invoice error status," he would record (measure) the status (error or no error) of each sampled invoice.

After the variable(s) of interest for every experimental unit in the sample (or population) is (are) measured, the data are analyzed, either by descriptive or by inferential statistical methods. The auditor, for example, may be interested only in *describing* the error rate in the sample of 100 invoices. More likely, however, he will want to use the information in the sample to make *inferences* about the population of all 15,472 invoices.

A **statistical inference** is an estimate or prediction or some other generalization about a population based on information contained in a sample.

That is, we use the information contained in the sample to learn about the larger population.* Thus, from the sample of 100 invoices, the auditor may estimate the total number of invoices containing errors in the population of 15,472 invoices. The auditor's inference about the quality of the firm's invoices can be used in deciding whether to modify the firm's billing operations.

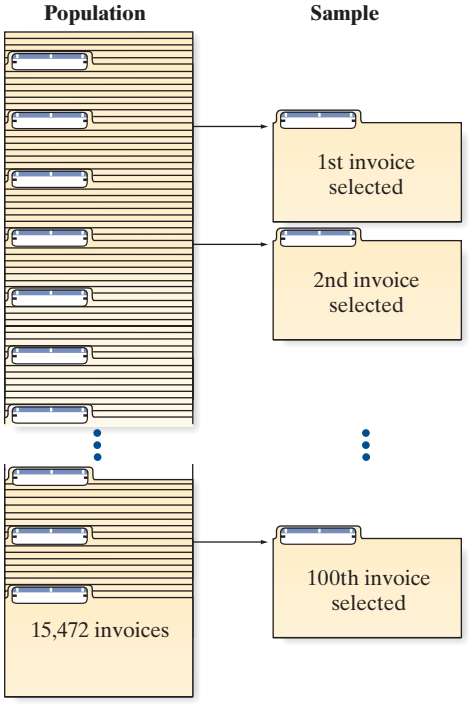


Figure 1.2
A sample of all company invoices

*The terms *population* and *sample* are often used to refer to the sets of measurements themselves, as well as to the units on which the measurements are made. When a single variable of interest is being measured, this usage causes little confusion. But when the terminology is ambiguous, we'll refer to the measurements as *population data sets* and *sample data sets*, respectively.

EXAMPLE 1.1

Key Elements of a Statistical Problem—Ages of Cable TV News Viewers

Problem According to the most recent Nielsen survey of cable TV news viewers, the average age of CNN viewers is 60 years. Suppose a rival network (e.g., FOX) executive hypothesizes that the average age of FOX viewers is greater than 60. To test her hypothesis, she samples 200 FOX viewers and determines the age of each.

a. Describe the population.
b. Describe the variable of interest.
c. Describe the sample.
d. Describe the inference.

Solution

a. The population is the set of units of interest to the TV executive, which is the set of all FOX viewers.
b. The age (in years) of each viewer is the variable of interest.
c. The sample must be a subset of the population. In this case, it is the 200 FOX viewers selected by the executive.
d. The inference of interest involves the *generalization* of the information contained in the sample of 200 viewers to the population of all FOX viewers. In particular, the executive wants to *estimate* the average age of the viewers in order to determine whether it exceeds 60 years. She might accomplish this by calculating the average age in the sample and using the sample average to estimate the population average.

Look Back A key to diagnosing a statistical problem is to identify the data set collected (in this example, the ages of the 200 FOX TV viewers) as a population or sample.

EXAMPLE 1.2

Key Elements of a Statistical Problem—Pepsi vs. Coca-Cola

Problem *Cola wars* is the popular term for the intense competition between Coca-Cola and Pepsi displayed in their marketing campaigns. Their campaigns have featured claims of consumer preference based on taste tests. For example, the *Huffington Post* (November 11, 2013) conducted a blind taste test of 9 cola brands that included Coca-Cola and Pepsi. (Pepsi finished 1st and Coke finished 5th.) Suppose, as part of a Pepsi marketing campaign, 1,000 cola consumers are given a blind taste test (i.e., a taste test in which the two brand names are disguised). Each consumer is asked to state a preference for brand A or brand B.

a. Describe the population.
b. Describe the variable of interest.
c. Describe the sample.
d. Describe the inference.

Solution

a. Because we are interested in the responses of cola consumers in a taste test, a cola consumer is the experimental unit. Thus, the population of interest is the collection or set of all cola consumers.
b. The characteristic that Pepsi wants to measure is the consumer's cola preference as revealed under the conditions of a blind taste test, so cola preference is the variable of interest.
c. The sample is the 1,000 cola consumers selected from the population of all cola consumers.
d. The inference of interest is the *generalization* of the cola preferences of the 1,000 sampled consumers to the population of all cola consumers. In particular, the

preferences of the consumers in the sample can be used to *estimate* the percentage of all cola consumers who prefer each brand.

Look Back In determining whether the statistical application is inferential or descriptive, we assess whether Pepsi is interested in the responses of only the 1,000 sampled customers (descriptive statistics) or in the responses for the entire population of consumers (inferential statistics).

Now Work Exercise 1.16

The preceding definitions and examples identify four of the five elements of an inferential statistical problem: a population, one or more variables of interest in a sample, and an inference. But making the inference is only part of the story. We also need to know its **reliability**–that is, how good the inference is. The only way we can be certain that an inference about a population is correct is to include the entire population in our sample. However, because of *resource constraints* (e.g., insufficient time and/or money), we usually can't work with whole populations, so we base our inferences on just a portion of the population (a sample). Consequently, whenever possible, it is important to determine and report the reliability of each inference made. Reliability, then, is the fifth element of inferential statistical problems.

The measure of reliability that accompanies an inference separates the science of statistics from the art of fortune-telling. A palm reader, like a statistician, may examine a sample (your hand) and make inferences about the population (your life). However, unlike statistical inferences, the palm reader's inferences include no measure of reliability.

Suppose, like the TV executive in Example 1.1, we are interested in the *error of estimation* (i.e., the difference between the average age of the population of TV viewers and the average age of a sample of TV viewers). Using statistical methods, we can determine a *bound on the estimation error*. This bound is simply a number that our estimation error (the difference between the average age of the sample and the average age of the population) is not likely to exceed. We'll see in later chapters that bound is a measure of the uncertainty of our inference. The reliability of statistical inferences is discussed throughout this text. For now, we simply want you to realize that an inference is incomplete without a measure of its reliability.

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

Let's conclude this section with a summary of the elements of both descriptive and inferential statistical problems and an example to illustrate a measure of reliability.

Four Elements of Descriptive Statistical Problems

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

Five Elements of Inferential Statistical Problems

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

EXAMPLE 1.3

Reliability of an Inference—Pepsi vs. Coca-Cola

Problem Refer to Example 1.2, in which the cola preferences of 1,000 consumers were indicated in a taste test. Describe how the reliability of an inference concerning the preferences of all cola consumers in the Pepsi bottler's marketing region could be measured.

Solution When the preferences of 1,000 consumers are used to estimate the preferences of all consumers in the region, the estimate will not exactly mirror the preferences of the population. For example, if the taste test shows that 56% of the 1,000 consumers chose Pepsi, it does not follow (nor is it likely) that exactly 56% of all cola drinkers in the region prefer Pepsi. Nevertheless, we can use sound statistical reasoning (which is presented later in the text) to ensure that our sampling procedure will generate estimates that are almost certainly within a specified limit of the true percentage of all consumers who prefer Pepsi. For example, such reasoning might assure us that the estimate of the preference for Pepsi from the sample is almost certainly within 5% of the actual population preference. The implication is that the actual preference for Pepsi is between 51% [i.e., (56 − 5)%] and 61% [i.e., (56 + 5)%]—that is, (56 ± 5)%. This interval represents a measure of reliability for the inference.

Look Back The interval 56 ± 5 is called a *confidence interval*, because we are "confident" that the true percentage of customers who prefer Pepsi in a taste test falls into the range (51, 61). In Chapter 6, we learn how to assess the degree of confidence (e.g., 90% or 95% confidence) in the interval.

STATISTICS IN ACTION REVISITED

Identifying the Population, Sample, and Inference

Consider the study on the link between a CEO's golf handicap and the company's stock performance, reported in the *New York Times*. The newspaper gathered information on golf handicaps of corporate executives obtained from a *Golf Digest* survey sent to CEOs of the 300 largest US companies. (A golf handicap is a numerical "index" that allows golfers to compare skills; the lower the handicap, the better the golfer.) For the 51 CEOs who reported their handicaps, the *New York Times* then determined each CEO's company stock market performance over a 3-year period (measured as a rate-of-return index, from a low value of 0 to a high value of 100). Thus, the experimental unit for the study is a corporate executive, and the two variables measured are golf handicap and stock performance index. Also, the data for the 51 CEOs represent a sample selected from the much larger population of all corporate executives in the United States. (These data are available in the **GLFCEO** file.)

The *New York Times* discovered a "statistical correlation" (a method discussed in Chapter 11) between golf handicap and stock performance. Thus, the newspaper inferred that the better the CEO is at golf, the better the company's stock performance.

Data Set: GLFCEO

1.4 Processes (Optional)

Sections 1.2 and 1.3 focused on the use of statistical methods to analyze and learn about populations, which are sets of *existing* units. Statistical methods are equally useful for analyzing and making inferences about *processes*.

A **process** is a series of actions or operations that transforms inputs to outputs. A process produces or generates output over time.

The most obvious processes of interest to businesses are those of production or manufacturing. A manufacturing process uses a series of operations performed by

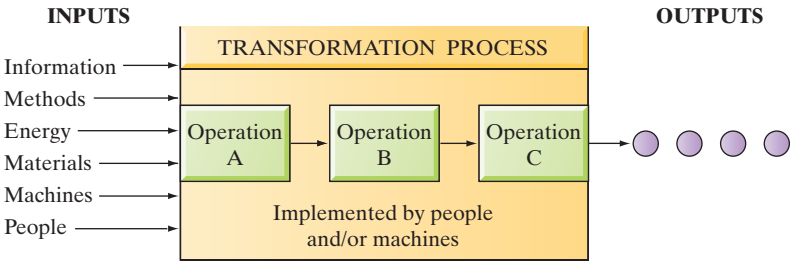


Figure 1.3
Graphical depiction of a manufacturing process

people and machines to convert inputs, such as raw materials and parts, to finished products (the outputs). Examples include the process used to produce the paper on which these words are printed, automobile assembly lines, and oil refineries.

Figure 1.3 presents a general description of a process and its inputs and outputs. In the context of manufacturing, the process in the figure (i.e., the transformation process) could be a depiction of the overall production process or it could be a depiction of one of the many processes (sometimes called *subprocesses*) that exist within an overall production process. Thus, the output shown could be finished goods that will be shipped to an external customer or merely the output of one of the steps or subprocesses of the overall process. In the latter case, the output becomes input for the next subprocess. For example, Figure 1.4 could represent the overall automobile assembly process, with its output being fully assembled cars ready for shipment to dealers. Or, it could depict the windshield assembly subprocess, with its output of partially assembled cars with windshields ready for "shipment" to the next subprocess in the assembly line.

Besides physical products and services, businesses and other organizations generate streams of numerical data over time that are used to evaluate the performance of the organization. Examples include weekly sales figures, quarterly earnings, and yearly profits. The US economy (a complex organization) can be thought of as generating streams of data that include the gross domestic product (GDP), stock prices, and the Consumer Price Index. Statisticians and other analysts conceptualize these data streams as being generated by processes. Typically, however, the series of operations or actions that cause particular data to be realized are either unknown or so complex (or both) that the processes are treated as *black boxes*.

A process whose operations or actions are unknown or unspecified is called a **black box.**

Frequently, when a process is treated as a black box, its inputs are not specified either. The entire focus is on the output of the process. A black box process is illustrated in Figure 1.4.

In studying a process, we generally focus on one or more characteristics, or properties, of the output. For example, we may be interested in the weight or the length of the units produced or even the time it takes to produce each unit. As with characteristics of population units, we call these characteristics *variables*. In studying processes whose output is already in numerical form (i.e., a stream of numbers), the characteristic, or property, represented by the numbers (e.g., sales, GDP, or stock prices) is typically the variable of interest. If the output is not numeric, we use *measurement processes* to assign

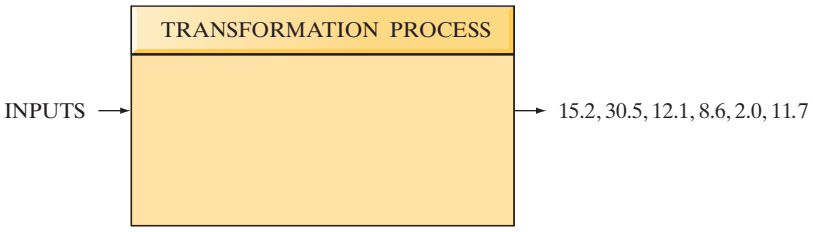


Figure 1.4
A black box process with numerical output

numerical values to variables.* For example, if in the automobile assembly process the weight of the fully assembled automobile is the variable of interest, a measurement process involving a large scale will be used to assign a numerical value to each automobile.

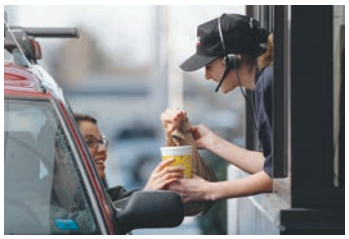
As with populations, we use sample data to analyze and make inferences (estimates, predictions, or other generalizations) about processes. But the concept of a sample is defined differently when dealing with processes. Recall that a population is a set of existing units and that a sample is a subset of those units. In the case of processes, however, the concept of a set of existing units is not relevant or appropriate. Processes generate or create their output *over time*—one unit after another. For example, a particular automobile assembly line produces a completed vehicle every 4 minutes. We define a sample from a process in the box.

Any set of output (object or numbers) produced by a process is also called a **sample.**

Thus, the next 10 cars turned out by the assembly line constitute a sample from the process, as do the next 100 cars or every fifth car produced today.

EXAMPLE 1.4

Key Elements of a Process—Waiting Time at a Fast-Food Window



Problem A particular fast-food restaurant chain has 6,289 outlets with drive-through windows. To attract more customers to its drive-through services, the company is considering offering a 50% discount to customers who wait more than a specified number of minutes to receive their order. To help determine what the time limit should be, the company decided to estimate the average waiting time at a particular drive-through window in Dallas, Texas. For 7 consecutive days, the worker taking customers' orders recorded the time that every order was placed. The worker who handed the order to the customer recorded the time of delivery. In both cases, workers used synchronized digital clocks that reported the time to the nearest second. At the end of the 7-day period, 2,109 orders had been timed.

a. Describe the process of interest at the Dallas restaurant.
b. Describe the variable of interest.
c. Describe the sample.
d. Describe the inference of interest.
e. Describe how the reliability of the inference could be measured.

Solution

a. The process of interest is the drive-through window at a particular fast-food restaurant in Dallas, Texas. It is a process because it "produces," or "generates," meals over time—that is, it services customers over time.
b. The variable the company monitored is customer waiting time, the length of time a customer waits to receive a meal after placing an order. Because the study is focusing only on the output of the process (the time to produce the output) and not the internal operations of the process (the tasks required to produce a meal for a customer), the process is being treated as a black box.
c. The sampling plan was to monitor every order over a particular 7-day period. The sample is the 2,109 orders that were processed during the 7-day period.
d. The company's immediate interest is in learning about the drive-through window in Dallas. They plan to do this by using the waiting times from the sample to make a statistical inference about the drive-through process. In particular, they might use the average waiting time for the sample to estimate the average waiting time at the Dallas facility.
e. As for inferences about populations, measures of reliability can be developed for inferences about processes. The reliability of the estimate of the average waiting

*A process with an output that is already in numerical form necessarily includes a measurement process as one of its subprocesses.

time for the Dallas restaurant could be measured by a bound on the error of estimation—that is, we might find that the average waiting time is 4.2 minutes, with a bound on the error of estimation of 0.5 minutes. The implication would be that we could be reasonably certain that the true average waiting time for the Dallas process is between 3.7 and 4.7 minutes.

Look Back Notice that there is also a population described in this example: the company's 6,289 existing outlets with drive-through facilities. In the final analysis, the company will use what it learns about the process in Dallas and, perhaps, similar studies at other locations to make an inference about the waiting times in its population of outlets.

Now Work Exercise 1.38

Note that output already generated by a process can be viewed as a population. Suppose a soft-drink canning process produced 2,000 twelve-packs yesterday, all of which were stored in a warehouse. If we were interested in learning something about those 2,000 twelve-packs—such as the percentage with defective cardboard packaging—we could treat the 2,000 twelve-packs as a population. We might draw a sample from the population in the warehouse, measure the variable of interest, and use the sample data to make a statistical inference about the 2,000 twelve-packs, as described in Sections 1.2 and 1.3.

In this optional section, we have presented a brief introduction to processes and the use of statistical methods to analyze and learn about processes. In Chapters 13 and 14 we present an in-depth treatment of these subjects.

1.5 Types of Data

You have learned that statistics is the science of data and that data are obtained by measuring the values of one or more variables on the units in the sample (or population). All data (and hence the variables we measure) can be classified as one of two general types: *quantitative* and *qualitative*.

Quantitative data are data that are measured on a naturally occurring numerical scale.* The following are examples of quantitative data:

1. The temperature (in degrees Celsius) at which each unit in a sample of 20 pieces of heat-resistant plastic begins to melt
2. The current unemployment rate (measured as a percentage) for each of the 50 states
3. The scores of a sample of 150 MBA applicants on the GMAT, a standardized business graduate school entrance exam administered nationwide
4. The number of female executives employed in each of a sample of 75 manufacturing companies

Quantitative data are measurements that are recorded on a naturally occurring numerical scale.

In contrast, qualitative data cannot be measured on a natural numerical scale; they can only be classified into categories.† Examples of qualitative data are as follows:

1. The political party affiliation (Democrat, Republican, or Independent) in a sample of 50 CEOs

*Quantitative data can be subclassified as either *interval* or *ratio*. For ratio data, the origin (i.e., the value 0) is a meaningful number. But the origin has no meaning with interval data. Consequently, we can add and subtract interval data, but we can't multiply and divide them. Of the four quantitative data sets listed, (1) and (3) are interval data, while (2) and (4) are ratio data.

†Qualitative data can be subclassified as either *nominal* or *ordinal*. The categories of an ordinal data set can be ranked or meaningfully ordered, but the categories of a nominal data set can't be ordered. Of the four qualitative data sets listed, (1) and (2) are nominal and (3) and (4) are ordinal.

2. The defective status (defective or not) of each of 100 computer chips manufactured by Intel
3. The size of a car (subcompact, compact, midsize, or full-size) rented by each of a sample of 30 business travelers
4. A taste tester's ranking (best, worst, etc.) of four brands of barbecue sauce for a panel of 10 testers

Often, we assign arbitrary numerical values to qualitative data for ease of computer entry and analysis. But these assigned numerical values are simply codes: They cannot be meaningfully added, subtracted, multiplied, or divided. For example, we might code Democrat = 1, Republican = 2, and Independent = 3. Similarly, a taste tester might rank the barbecue sauces from 1 (best) to 4 (worst). These are simply arbitrarily selected numerical codes for the categories and have no utility beyond that.

Qualitative data are measurements that cannot be measured on a natural numerical scale; they can only be classified into one of a group of categories.

EXAMPLE 1.5

Types of Data—Study of a River Contaminated by a Chemical Plant

Problem Chemical and manufacturing plants sometimes discharge toxic-waste materials such as DDT into nearby rivers and streams. These toxins can adversely affect the plants and animals inhabiting the river and the riverbank. The U.S. Army Corps of Engineers conducted a study of fish in the Tennessee River (in Alabama) and its three tributary creeks: Flint Creek, Limestone Creek, and Spring Creek. A total of 144 fish were captured, and the following variables were measured for each:

1. River/creek where each fish was captured
2. Species (channel catfish, largemouth bass, or smallmouth buffalo fish)
3. Length (centimeters)
4. Weight (grams)
5. DDT concentration (parts per million)

These data are saved in the **DDT** file. Classify each of the five variables measured as quantitative or qualitative.

Solution The variables length, weight, and DDT are quantitative because each is measured on a numerical scale: length in centimeters, weight in grams, and DDT in parts per million. In contrast, river/creek and species cannot be measured quantitatively: They can only be classified into categories (e.g., channel catfish, largemouth bass, and smallmouth buffalo fish for species). Consequently, data on river/creek and species are qualitative.

Look Back It is essential that you understand whether data are quantitative or qualitative in nature because the statistical method appropriate for describing, reporting, and analyzing the data depends on the data type (quantitative or qualitative).

Now Work Exercise 1.15

We demonstrate many useful methods for analyzing quantitative and qualitative data in the remaining chapters of the text. But first, we discuss some important ideas on data collection.

1.6 Collecting Data: Sampling and Related Issues

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

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

Sometimes, the data set of interest has already been collected for you and is available in a **published source,** such as a book, journal, newspaper, or Web site. For example, you may want to examine and summarize the unemployment rates (i.e., percentages of eligible workers who are unemployed) in the 50 states of the United States. You can find this data set (as well as numerous other data sets) at your library in the *Statistical Abstract of the United States*, published annually by the US government. Similarly, someone who is interested in monthly mortgage applications for new home construction would find this data set in the *Survey of Current Business*, another government publication. Other examples of published data sources include the *Wall Street Journal* (financial data) and the The *Elias Sports Bureau* (sports information).* The Internet (World Wide Web) provides a medium by which data from published sources are readily available.

A second method of collecting data involves conducting a **designed experiment,** in which the researcher exerts strict control over the units (people, objects, or events) in the study. For example, an often-cited medical study investigated the potential of aspirin in preventing heart attacks. Volunteer physicians were divided into two groups—the *treatment* group and the *control* group. In the treatment group, each physician took one aspirin tablet a day for 1 year, while each physician in the control group took an aspirin-free placebo (no drug) made to look like an aspirin tablet. The researchers, not the physicians under study, controlled who received the aspirin (the treatment) and who received the placebo. As you will learn in Chapter 9, properly designed experiment allows you to extract more information from the data than is possible with an uncontrolled study.

Finally, observational studies can be employed to collect data. In an **observational study,** the researcher observes the experimental units in their natural setting and records the variable(s) of interest. For example, a company psychologist might observe and record the level of "Type A" behavior of a sample of assembly line workers. Similarly, a finance researcher may observe and record the closing stock prices of companies that are acquired by other firms on the day prior to the buyout and compare them to the closing prices on the day the acquisition is announced. Unlike a designed experiment, an observational study is one in which the researcher makes no attempt to control any aspect of the experimental units.

The most common type of observational study is a **survey,** where the researcher samples a group of people, asks one or more questions, and records the responses. Probably the most familiar type of survey is the political poll, conducted by any one of a number of organizations (e.g., Harris, Gallup, Roper, and CNN) and designed to predict the outcome of a political election. Another familiar survey is the Nielsen survey, which provides the major television networks with information on the most watched TV programs. Surveys can be conducted through the mail, with telephone interviews, or with in-person interviews. Although in-person interviews are more expensive than mail or telephone surveys, they may be necessary when complex information must be collected.

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

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

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

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

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

For example, consider a political poll conducted during a presidential election year. Assume the pollster wants to estimate the percentage of all 145 million registered voters in the United States who favor the incumbent president. The pollster would be unwise to base the estimate on survey data collected for a sample of voters from the incumbent's own state. Such an estimate would almost certainly be *biased* high; consequently, it would not be very reliable.

The most common way to satisfy the representative sample requirement is to select a simple random sample. A **simple random sample** ensures that every subset of fixed size in the population has the same chance of being included in the sample. If the pollster samples 1,500 of the 145 million voters in the population so that every subset of 1,500 voters has an equal chance of being selected, she has devised a simple random sample.

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

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

EXAMPLE 1.6

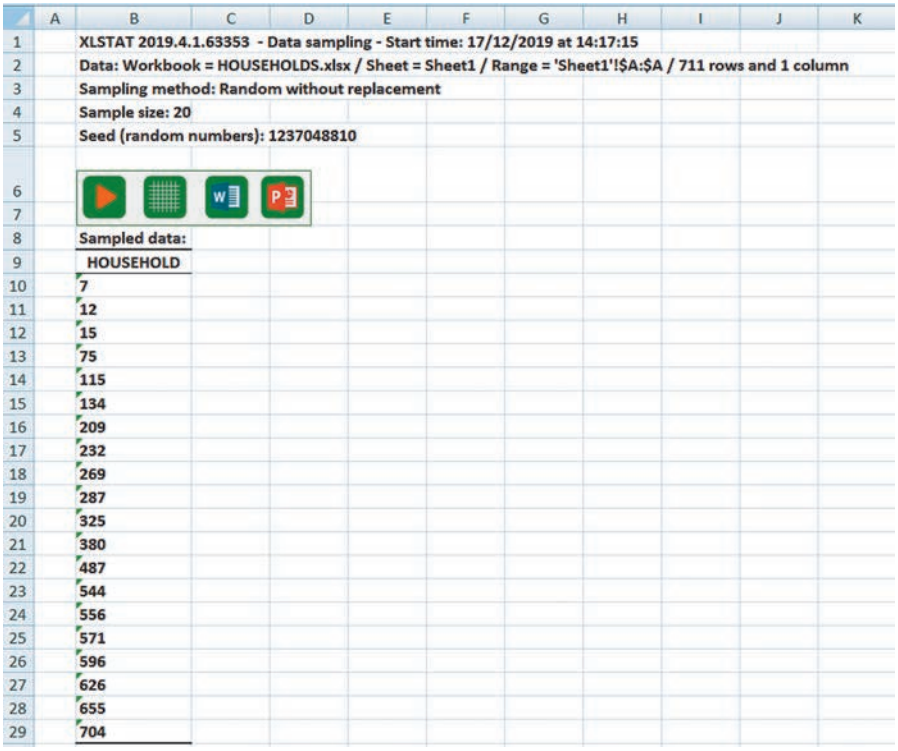
Generating a Simple Random Sample—Selecting Households for a Feasibility Study

Problem Suppose you wish to assess the feasibility of building a new high school. As part of your study, you would like to gauge the opinions of people living close to the proposed building site. The neighborhood adjacent to the site has 711 homes. Use a random number generator to select a simple random sample of 20 households from the neighborhood to participate in the study.

Solution In this study, your population of interest consists of the 711 households in the adjacent neighborhood. To ensure that every possible sample of 20 households selected from the 711 has an equal chance of selection (i.e., to ensure a simple random sample), first assign a number from 1 to 711 to each of the households in the population. These numbers were entered into an Excel worksheet. Now, apply the random number generator of Excel/ XLSTAT, requesting that 20 households be selected without replacement. Figure 1.5 shows one possible set of random numbers generated from XLSTAT. You can see that households numbered 7, 12, 15, . . . , 704 are the households to be included in your sample.

Look Back It can be shown (proof omitted) that there are over 3×10^{38} possible samples of size 20 that can be selected from the 711 households. Random number generators guarantee (to a certain degree of approximation) that each possible sample has an equal chance of being selected.

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



	A	B	C	D	E	F	G	H	I	J	K
1		XLSTAT 2019.4.1.63353 - Data sampling - Start time: 17/12/2019 at 14:17:15									
2		Data: Workbook = HOUSEHOLDS.xlsx / Sheet = Sheet1 / Range = 'Sheet1'!$A:$A / 711 rows and 1 column									
3		Sampling method: Random without replacement									
4		Sample size: 20									
5		Seed (random numbers): 1237048810									
6											
7											
8		Sampled data:									
9		HOUSEHOLD									
10		7									
11		12									
12		15									
13		75									
14		115									
15		134									
16		209									
17		232									
18		269									
19		287									
20		325									
21		380									
22		487									
23		544									
24		556									
25		571									
26		596									
27		626									
28		655									
29		704									

Figure 1.5
Random selection of 20 households using XLSTAT

Now Work Exercise 1.14

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

EXAMPLE 1.7

Randomization in a Designed Experiment—A Clinical Trial

Problem A designed experiment in the medical field involving human subjects is referred to as a *clinical trial*. One recent clinical trial was designed to determine the potential of using aspirin in preventing heart attacks. Volunteer physicians were randomly divided into two groups—the *treatment* group and the *control* group. Each physician in the treatment group took one aspirin tablet a day for one year, while the physicians in the control group took an aspirin-free placebo made to look identical to an aspirin tablet. Because the physicians did not know which group, treatment or control, they were assigned to, the clinical trial is called a *blind study*. Assume 20 physicians volunteered for the study. Use a random number generator to randomly assign half of the physicians to the treatment group and half to the control group.

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

The Minitab random sample procedure was employed, producing the printout shown in Figure 1.6. Numbering the physicians from 1 to 20, we see that physicians 1, 9, 20, 12, 3, 13, 4, 5, 14, and 15 are assigned to receive the aspirin (treatment). The remaining physicians are assigned the placebo (control).

Now Work Exercise 1.34e

	C1	C2
	Physician	Treatment
1	1	1
2	2	9
3	3	20
4	4	12
5	5	3
6	6	13
7	7	4
8	8	5
9	9	14
10	10	15
11	11	
12	12	
13	13	
14	14	
15	15	
16	16	
17	17	
18	18	
19	19	
20	20	
21		

Figure 1.6
Minitab Worksheet with Random Assignment of Physicians

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

Stratified random sampling is typically used when the experimental units associated with the population can be separated into two or more groups of units, called *strata*, where the characteristics of the experimental units are more similar within strata than across strata. Random samples of experimental units are obtained for each strata; then the units are combined to form the complete sample. For example, if you are gauging opinions of voters on a polarizing issue, like government-sponsored health care, you may want to stratify on political affiliation (Republicans and Democrats), making sure that representative samples of both Republicans and Democrats (in proportion to the number of Republicans and Democrats in the voting population) are included in your survey.

Sometimes it is more convenient and logical to sample natural groupings (*clusters*) of experimental units first, and then collect data from all experimental units within each cluster. This involves the use of *cluster sampling*. For example, suppose a marketer for a large upscale restaurant chain wants to find out whether customers like the new menu. Rather than collect a simple random sample of all customers (which would be very difficult and costly to do), the marketer will randomly sample 10 of the 150 restaurant locations (clusters), and then interview all customers eating at each of the 10 locations on a certain night.

Another popular sampling method is *systematic sampling*. This method involves systematically selecting every kth experimental unit from a list of all experimental units. For example, every fifth person who walks into a shopping mall could be asked his or her opinion on a business topic of interest. Or, a quality control engineer at a manufacturing plant may select every 10th item produced on an assembly line for inspection.

A fourth alternative to simple random sampling is *randomized response sampling*. This design is particularly useful when the questions of the pollsters are likely to elicit false answers. For example, suppose each person in a sample of wage earners is asked whether he or she ever cheated on an income tax return. A cheater might lie, thus biasing an estimate of the true likelihood of someone cheating on his or her tax return. To circumvent this problem, each person is presented with two questions, one being the object of the survey and the other an innocuous question, such as:

1. Did you ever cheat on your federal income tax return?
2. Did you drink coffee this morning?

One of the questions is chosen at random to be answered by the wage earner by flipping a coin; however, which particular question is answered is unknown to the interviewer. In this way, the random response method attempts to elicit an honest response to a sensitive question. Sophisticated statistical methods are then employed to derive an estimate of the percentage of "yes" responses to the sensitive question.

No matter what type of sampling design you employ to collect the data for your study, be careful to avoid *selection bias*. Selection bias occurs when some experimental units in the population have less chance of being included in the sample than others. This results in samples that are not representative of the population. Consider an opinion poll that employs either a telephone survey or a mail survey. After collecting a random sample of phone numbers or mailing addresses, each person in the sample is contacted via telephone or the mail and a survey conducted. Unfortunately, these types of surveys often suffer from selection bias due to *nonresponse*. Some individuals may not be home when the phone rings, or others may refuse to answer the questions or mail back the questionnaire. As a consequence, no data are obtained for the nonrespondents in the sample. If the nonrespondents and respondents differ greatly on an issue, then *nonresponse bias* exits. For example, those who choose to answer a question on a school board issue may have a vested interest in the outcome of the survey—say, parents with children of school age, schoolteachers whose jobs may be in jeopardy, or citizens whose taxes might be substantially affected. Others with no vested interest may have an opinion on the issue but might not take the time to respond.

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

Consider a sample of experimental units where some units produce data (i.e., responders) and no data is collected on the other units (i.e., nonresponders). **Nonresponse bias** is a type of selection bias that results when the response data differ from the potential data for the nonresponders.

Finally, even if your sample is representative of the population, the data collected may suffer from *measurement error*. That is, the values of the data (quantitative or qualitative) may be inaccurate. In sample surveys, opinion polls, etc., measurement error often results from *ambiguous* or *leading questions*. Consider the survey question: "How often did you change the oil in your car last year?" It is not clear whether the researcher wants to know how often you personally changed the oil in your car or how often you took your car into a service station to get an oil change. The ambiguous question may lead to inaccurate responses. On the other hand, consider the question: "Does the new health plan offer more comprehensive medical services at less cost than the old one?" The way the question is phrased *leads* the reader to believe that the new plan is better and to a "yes" response—a response that is more desirable to the researcher. A better, more neutral way to phrase the question is: "Which health plan offers more comprehensive medical services at less cost, the old one or the new one?"

Measurement error refers to inaccuracies in the values of the data collected. In surveys, the error may be due to ambiguous or leading questions and the interviewer's effect on the respondent.

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

EXAMPLE 1.8

Method of Data Collection—Survey of Online Shoppers

Problem What is the most popular device used by online shoppers? To find out, the mobile video ad network AdColony conducted a 2019 nationwide survey of 1,000 US online shoppers for Mobile Marketer. The most popular device was a smartphone, used by 56% of the online shoppers. Other results: 28% used a desktop or laptop computer, and 16% used a tablet.

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

Solution

a. The data-collection method is a survey: 1,000 online shoppers participated in the study.
b. Presumably, Mobile Marketer (who commissioned the survey) is interested in the devices used by all US online shoppers. Consequently, the target population is *all* consumers who use the Internet for online shopping.
c. Because the 1,000 respondents clearly make up a subset of the target population, they do form a sample. Whether or not the sample is representative is unclear because Mobile Marketer provided no detailed information on how the 1,000 shoppers were selected. If the respondents were obtained using, say, random-digit telephone dialing, then the sample is likely to be representative because it is a random sample. However, if the questionnaire was made available to anyone surfing the Internet, then the respondents are *self-selected* (i.e., each Internet user who saw the survey chose whether or not to respond to it). Such a survey often suffers from *nonresponse* bias. It is possible that many Internet users who chose not to respond (or who never

saw the questionnaire) would have answered the questions differently, leading to a lower (or higher) sample percentage.

Look Back Any inferences based on survey samples that employ self-selection are suspect due to potential nonresponse bias.

Now Work Exercise 1.27

EXAMPLE 1.9

Representative Data—Price Promotion Study

Problem Marketers use wording such as "was $100, now $80" to indicate a price promotion. The promotion is typically compared to the retailer's previous price or to a competitor's price. A study in the *Journal of Consumer Research* investigated whether between-store comparisons result in greater perceptions of value by consumers than within-store comparisons. Suppose 50 consumers were randomly selected from all consumers in a designated market area to participate in the study. The researchers randomly assigned 25 consumers to read a within-store price promotion advertisement ("was $100, now $80") and 25 consumers to read a between-store price promotion ("$100 there, $80 here"). The consumers then gave their opinion on the value of the discount offer on a 10-point scale (where 1 = lowest value and 10 = highest value). The value opinions of the two groups of consumers were compared.

a. Identify the data-collection method.

b. Are the sample data representative of the target population?

Solution

a. Here, the experimental units are the consumers. Because the researchers controlled which price promotion ad—"within-store" or "between-store"—the experimental units (consumers) were assigned to, a designed experiment was used to collect the data.

b. The sample of 50 consumers was randomly selected from all consumers in the designated market area. If the target population is all consumers in this market, it is likely that the sample is representative. However, the researchers warn that the sample data should not be used to make inferences about consumer behavior in other, dissimilar markets.

Look Back By using randomization in a designed experiment, the researcher is attempting to eliminate different types of bias, including self-selection bias.

Now Work Exercise 1.19

STATISTICS IN ACTION REVISITED

Identifying the Data-Collection Method and Data Type

First, refer to the *Nature Neuroscience* chocolate study. Recall that researchers randomly assigned 37 people to one of four groups: (1) High dose of cocoa flavanol and aerobic exercise; (2) High dose of cocoa flavanol but no exercise; (3) Low dose of cocoa flavanol and aerobic exercise; and (4) Low cocoa flavanol dose but without the exercise (See Figure SIA1.1). After a 3-month period, the researchers measured (among other variables) each subject's reaction time (in seconds) to an object-recognition task. Clearly, the data-collection method employed is a designed experiment—the experimental units (subjects) were randomly assigned to the groups. The groups are formed from two variables: dosage of cocoa flavanol (high or low) and aerobic exercise (yes or no). Consequently, these two categorical-type variables are qualitative in nature. The numerical reaction time variable, measured in seconds, is quantitative.

Now, let's consider the *New York Times* study on the link between a CEO's golf handicap and the company's stock performance. Recall that the newspaper gathered information on golf handicaps of corporate executives obtained from a *Golf Digest* survey

STATISTICS IN ACTION REVISTED (*continued*)

that was sent to 300 corporate executives. Thus, the data-collection method is a survey. In addition to golf handicap (a numerical "index" that allows golfers to compare skills), the *Times* measured the CEO's company stock market performance over a 3-year period on a scale of 0 to 100. Because both variables, golf handicap and stock performance, are numerical in nature, they are quantitative data.

Data Set: GLFCEO

1.7 Business Analytics: Critical Thinking with Statistics

BIOGRAPHY

H. G. WELLS (1866–1946)
Writer and Novelist

English-born Herbert George Wells published his first novel, *The Time Machine*, in 1895 as a parody of the English class division and as a satirical warning that human progress is inevitable. Although most famous as a science-fiction novelist, Wells was a prolific writer as a journalist, sociologist, historian, and philosopher. Wells's prediction about statistical thinking is just one of a plethora of observations he made about life in this world. Here are a few more of H. G. Wells's more famous quotes:

"Advertising is legalized lying."

"Crude classification and false generalizations are the curse of organized life."

"The crisis of today is the joke of tomorrow."

"Fools make researchers and wise men exploit them."

"The only true measure of success is the ratio between what we might have done and what we might have been on the one hand, and the thing we have made and the things we have made of ourselves on the other." (Quotes by Herbert George Wells.)

According to H. G. Wells, author of such science-fiction classics as *The War of the Worlds* and *The Time Machine, "Statistical thinking* will one day be as necessary for efficient citizenship as the ability to read and write." Written more than a hundred years ago, Wells's prediction is proving true today.

The growth in data collection associated with scientific phenomena, business operations, and government activities (e.g., marketing, quality control, statistical auditing, forecasting, etc.) has been remarkable over the past decade. This growth is due, in part, to technology now capable of capturing lots of data at a high rate, such as information-sensing mobile devices, cameras, radio-frequency identification (RFID) readers, and wireless sensor networks. In fact, the term "Big Data" is now commonly used by companies to describe this wealth of information.

However, with big data, comes the need for methods of analysis—**business analytics**—that ultimately lead to good business decisions. A key to successful applications of business analytics is *quantitative literacy* (i.e., the ability to evaluate data intelligently). Whether the data of interest is "big" or not, each of us has to develop the ability to use rational thought to interpret and understand the meaning of the data. Business analytics and quantitative literacy can help you make intelligent decisions, inferences, and generalizations from data; that is, it helps you *think critically* using statistics. We term this skill **statistical thinking.**

Business analytics refers to methodologies (e.g., statistical methods) that extract useful information from data in order to make better business decisions.

Statistical thinking involves applying rational thought and the science of statistics to critically assess data and inferences. Fundamental to the thought process is that variation exists in populations and process data.

To gain some insight into the role statistics plays in business analytics, we present two examples of some misleading or faulty surveys.

EXAMPLE 1.10

Biased Sample—Motorcycle Helmet Law

Problem An article in the *New York Times* considered the question of whether motorcyclists should be required by law to wear helmets. In supporting his argument for no helmets, the editor of a magazine for Harley-Davidson bikers presented the results of one study that claimed "nine states without helmet laws had a lower fatality rate (3.05 deaths per 10,000 motorcycles) than those that mandated helmets (3.38)" and a survey that found "of 2,500 bikers at a rally, 98% of the respondents opposed such laws." Based on this information, do you think it is safer to ride a motorcycle without a helmet? What further statistical information would you like?

Solution You can use statistical thinking to help you critically evaluate the study. For example, before you can evaluate the validity of the 98% estimate, you would want to know how the data were collected. If a survey was, in fact, conducted, it's possible that

the 2,500 bikers in the sample were not selected at random from the target population of all bikers, but rather were "self-selected." (Remember, they were all attending a rally—a rally likely for bikers who oppose the law.) If the respondents were likely to have strong opinions regarding the helmet law (e.g., strongly oppose the law), the resulting estimate is probably biased high. Also, if the selection bias in the sample was intentional, with the sole purpose to mislead the public, the researchers would be guilty of **unethical statistical practice.**

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

You would also want more information about the study comparing the motorcycle fatality rate of the nine states without a helmet law to those states that mandate helmets. Were the data obtained from a published source? Were all 50 states included in the study, or were only certain states selected? That is, are you seeing sample data or population data? Furthermore, do the helmet laws vary among states? If so, can you really compare the fatality rates?

Look Back Questions such as these led a group of mathematics and statistics teachers attending an American Statistical Association course to discover a scientific and statistically sound study on helmets. The study reported a dramatic *decline* in motorcycle crash deaths after California passed its helmet law.

EXAMPLE 1.11

Manipulative or Ambiguous Survey Questions—Satellite Radio Survey

Problem When talk-show host Howard Stern moved his controversial radio program from free, over-the-air (AM/FM) radio to Sirius XM satellite radio, the move was perceived in the industry to boost satellite radio subscriptions. This led American Media Services, a developer of AM/FM radio properties, to solicit a nationwide random-digit dialing phone survey of 1,008 people. The purpose of the survey was to determine how much interest Americans really have in buying satellite radio service. After providing some background on Howard Stern's controversial radio program, one of the questions asked, "How likely are you to purchase a subscription to satellite radio after Howard Stern's move to Sirius?" The result: 86% of the respondents stated that they aren't likely to buy satellite radio because of Stern's move. Consequently, American Media Services concluded that "the Howard Stern Factor is overrated" and that "few Americans expect to purchase satellite radio"—claims that made the headlines of news reports and Web blogs. Do you agree?

Solution First, we need to recognize that American Media Services had a vested interest in the outcome of the survey—the company makes its money from over-the-air broadcast radio stations. Second, although the phone survey was conducted using random-digit dialing, there is no information provided on the response rate. It's possible that nonrespondents (people who were not home or refused to answer the survey questions) tend to be people who use cell phones more than their landline phone and, consequently, are more likely to use the latest in electronic technology, including satellite radio. Finally, the survey question itself is ambiguous. Do the respondents have negative feelings about satellite radio, Howard Stern, or both? If not for Howard Stern's program, would the respondents be more likely to buy satellite radio? To the critical thinker, it's unclear what the results of the survey imply.

Look Back Examining the survey results from the perspective of satellite radio providers, 14% of the respondents indicated that they would be likely to purchase satellite radio. Projecting the 14% back to the population of all American adults, this figure represents about 50 million people; what is interpreted as "few Americans" by American Media Services could be music to the ears of satellite radio providers.

ETHICS in STATISTICS
Intentionally selecting a nonrandom sample in an effort to support a particular viewpoint is considered *unethical statistical practice.*

As with many statistical studies, both the motorcycle helmet study and the satellite radio study are based on survey data. Most of the problems with these surveys result from the use of *nonrandom* samples. These samples are subject to potential errors, such as *selection bias, nonresponse bias*, and *measurement error*. Researchers who are aware of these problems and continue to use the sample data to make inferences are practicing *unethical statistics*.

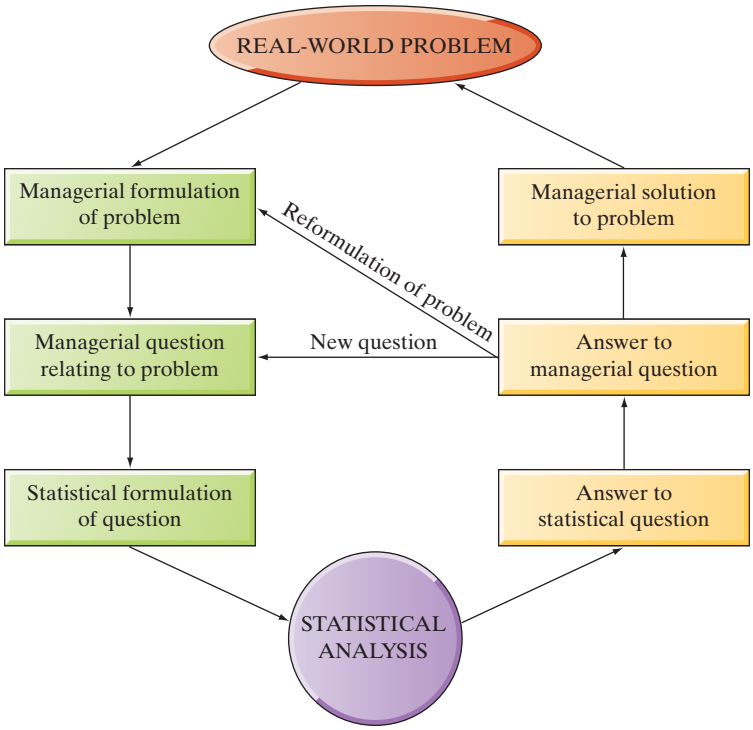


Figure 1.7
Flow diagram showing the role of statistics in business analytics
Source: From *The American Statistician* by George Benson. Copyright © by George Benson. Used by permission of George Benson.

As stated earlier, business analytics relies heavily on statistical thinking to help firms make better business decisions. The role statistics can play in a manager's use of business analytics is displayed in Figure 1.7. Every managerial decision-making problem begins with a real-world problem. This problem is then formulated in managerial terms and framed as a managerial question. The next sequence of steps (proceeding counterclockwise around the flow diagram) identifies the role that statistics can play in this process. The managerial question is translated into a statistical question, the sample data are collected and analyzed, and the statistical question is answered. The next step in the process is using the answer to the statistical question to reach an answer to the managerial question. The answer to the managerial question may suggest a reformulation of the original managerial problem, suggest a new managerial question, or lead to the solution of the managerial problem.

One of the most difficult steps in the decision-making process—one that requires a cooperative effort among managers and statisticians—is the translation of the managerial question into statistical terms (for example, into a question about a population). This statistical question must be formulated so that, when answered, it will provide the key to the answer to the managerial question. Thus, as in the game of chess, you must formulate the statistical question with the end result, the solution to the managerial question, in mind.

In the remaining chapters of the text, you'll become familiar with the business analytic tools essential for building a firm foundation in statistics and statistical thinking.

STATISTICS IN ACTION REVISITED

Critically Assessing the Ethics of a Statistical Study

The *New York Times* reported a strong link between a corporate executive's golf handicap and his/her company's stock performance. Thus, the newspaper inferred that the better the CEO is at golf, the better the company's stock performance will be. To critically assess this study, consider the following facts:

1. *Golf Digest* sent surveys to the CEOs at the 300 largest US firms. Only 74 executives agreed to reveal their golf handicaps. Of these 74 CEOs, the *Times* collected data on stock performance for only 51 of the companies. (The other 23 CEOs were not in the stock performance database used by the newspaper.)

STATISTICS IN ACTION REVISTED (*continued*)

2. The *New York Times* researcher who performed the analysis of the data stated that "for all the different factors I've tested as possible links to predicting which CEOs are going to perform well or poorly, [golf handicap] is certainly one of the . . . strongest."
3. According to the *Times*, the researcher "scientifically sifted out a handful of CEOs because of their statistical extremes," in effect "removing seven CEOs from the final analysis because [their data] destroyed the trend lines."

These observations lead a critical thinker to doubt the validity of the inference made by the *New York Times* researcher. Consider first that the sample of CEOs analyzed was not randomly selected from all CEOs in the United States. In fact, it was self-selected—only those CEOs who chose to report their golf handicap were included in the study. (Not even all these "self-reporters" were included; some were eliminated because the newspaper did not have information on their company's stock performance in the database.) Thus, the potential for selection and/or nonresponse bias is extremely high.

Second, based on fact #2, it is likely that the researcher tested a multitude of factors and found only one (golf handicap) that had a link to stock performance. We will learn in subsequent chapters that a plethora of irrelevant variables are tested statistically, by chance one or more of the variables will be found "statistically significant."

Finally, the researcher removed the data for seven CEOs based on their "statistical extremes." In the next chapter, we learn about statistical "outliers"—how to detect them and how to treat them when discovered. However, it can be shown (using the methods outlined in the text) that these seven data points are not outliers. If the data points are included in the analysis, the link between golf handicap and stock performance is found to be weak, at best.

Data Set: GLFCEO

CHAPTER NOTES

Key Terms

Note: Starred (*) terms are from the optional section in this chapter.

*Black box 30
Big data 40
Business analytics 40
Census 25
Cluster sampling 37
Data 21
Descriptive statistics 23
Designed experiment 34
Experimental (or observational) unit 25
Inference 23
Inferential statistics 23
Measurement 25
Measurement error 37
Measure of reliability 28
Nonresponse bias 38
Observational study 34
Population 25
*Process 29
Published source 34
Qualitative data 33
Quantitative data 32
Quantitative literacy 39
Randomized response sampling 37
Random number generator 35
Reliability 28
Representative sample 35
Sample 26
Selection bias 38
Simple random sample 35
Statistical inference 26
Statistical thinking 40
Statistics 22
Stratified random sampling 37
Survey 34
Systematic sampling 37
Unethical statistical practice 41
Variable 25

Key Ideas

Types of Statistical Applications

Descriptive

1. Identify **population** or **sample** (collection of **experimental units**)
2. Identify **variable(s)**
3. Collect **data**
4. **Describe** data

Inferential

1. Identify **population** (collection of *all* **experimental units**)
2. Identify **variable(s)**
3. Collect **sample** data (*subset* of population)
4. **Inference** about population based on sample
5. **Measure of reliability** for inference

Types of Data

1. **Quantitative** (numerical in nature)
2. **Qualitative** (categorical in nature)

Data-Collection Methods

1. **Observational** (e.g., survey)
2. **Published source**
3. **Designed experiment**

Types of Random Samples

1. **Simple random sample**
2. **Stratified random sample**
3. **Cluster sample**
4. **Systematic sample**
5. **Random response sample**

Problems with Nonrandom Samples

1. **Selection bias**
2. **Nonresponse bias**
3. **Measurement error**

EXERCISES 1.1–1.40

Note: Starred (*) exercises are from the optional section in this chapter.

Learning the Mechanics

1.1 What is statistics?

1.2 Explain the difference between descriptive and inferential statistics.

1.3 List and define the four elements of a descriptive statistics problem.

1.4 List and define the five elements of an inferential statistical analysis.

1.5 List the three major methods of collecting data and explain their differences.

1.6 Explain the difference between quantitative and qualitative data.

1.7 Explain how populations and variables differ.

1.8 Explain how populations and samples differ.

1.9 What is a representative sample? What is its value?

1.10 Why would a statistician consider an inference incomplete without an accompanying measure of its reliability?

1.11 Give an example of unethical statistical practice.

1.12 Define *statistical thinking*.

1.13 Suppose you have to sort sample units into categories according to their region of origin. The regions are "Africa," "Americas," "Asia," "Europe," and "Oceania." For further analysis with a statistical software, you replace each region name with a numerical code: 1 for Africa, 2 for Americas, and so on. Are the data consisting of the region names qualitative or quantitative? Are the numerical codes qualitative or quantitative? Explain your answer.

1.14 Suppose that a production batch contains 1,000 units and you have to select 10 units for quality assurance. Use a random number generator to select a simple random sample of $n = 10$ from the production batch.

Applet Exercise 1.1

The *Random Numbers* applet generates a list of n random numbers from 1 to N, where n is the size of the sample and N is the size of the population. The list generated often contains repetitions of one or more numbers.

a. Using the applet *Random Numbers*, enter 1 for the minimum value, 10 for the maximum value, and 10 for the sample size. Then click on *Sample*. Look at the results and list any numbers that are repeated and the number of times each of these numbers occurs.

b. Repeat part **a,** changing the maximum value to 20 and keeping the size of the sample fixed at 10. If you still have repetitions, repeat the process, increasing the maximum value by 10 each time but keeping the size of the sample fixed. What is the smallest maximum value for which you had no repetitions?

c. Describe the relationship between the population size (maximum value) and the number of repetitions in the list of random numbers as the population size increases and the sample size remains the same. What can you conclude about using a random number generator to choose a relatively small sample from a large population?

Applet Exercise 1.2

The *Random Numbers* applet can be used to select a random sample from a population, but can it be used to simulate data? In parts **a** and **b**, you will use the applet to create data sets. Then you will explore whether those data sets are realistic.

a. In the activity *Keep the Change* on page 48, a data set called *Amounts Transferred* is described. Use the *Random Numbers* applet to simulate this data set by setting the minimum value equal to 0, the maximum value equal to 99, and the sample size equal to 30. Explain what the numbers in the list produced by the applet represent in the context of the activity. (You may need to read the activity.) Do the numbers produced by the applet seem reasonable? Explain.

b. Use the *Random Numbers* applet to simulate grades on a statistics test by setting the minimum value equal to 0, the maximum value equal to 100, and the sample size equal to 30. Explain what the numbers in the list produced by the applet represent in this context. Do the numbers produced by the applet seem reasonable? Explain.

c. Referring to parts **a** and **b,** why do the randomly generated data seem more reasonable in one situation than in the other? Comment on the usefulness of using a random number generator to produce data.

Applying the Concepts—Basic

1.15 NW **Wind turbine database**. A wind turbine turns wind energy into electricity using the aerodynamic force from the rotor blades. In 2019, wind power accounted for 5 percent of all electricity generated in the United States. The US Geological Survey compiled a database of over 53,700 wind turbines in the country. Several variables measured for each wind turbine in the database are listed below. Determine the type of data (quantitative or qualitative) recorded for each variable.

a. Electrical generation capacity (measured in kilowatts)
b. Hub height (measured in meters)
c. Rotor diameter (measured in meters)
d. Location (state/county)
e. Number of turbines in the project

[*Note*: A project consists of multiple wind turbines that form a single system.]

1.16 NW CAPRATE **Valuation of single-tenant properties**. In court cases involving real estate, experts are hired to estimate the value of properties. The estimate will vary depending on the method of valuation applied. Research on valuation methods for single-tenant properties was published in *The Appraisal Journal* (Summer 2019). One method examines the ratio of net operating income to property asset value – called the capitalization rate. The research focused on 80 tenants of properties owned by the Boulder Group consulting firm. Data on S&P credit ratings and capitalization rates for the 13 Boulder Group tenants of retail properties are shown in the table (next page). Capitalization rates are provided for tenants with 5, 10, 15, and 20 years remaining on the lease.

a. What is the experimental unit for this study?
b. Describe the variables measured in the study. Do these variables produce quantitative or qualitative data?
c. Do the 13 tenants in the table represent a population or a sample?
d. If the 13 tenants represent a representative sample from a population, describe the population.
e. One inference derived from the data is that "capitalization rates increase . . . as the remaining lease term decreases." Do you agree?

	S&P	Capitalization Rate (%)			
Tenant	Credit Rating	5-Year	10-Year	15-Year	20-Year
Best Buy	BBB	8.25	7.25	N/A	N/A
BJ's Warehouse	B	7.40	6.85	6.40	6.00
Dollar General	BBB−	8.45	7.75	6.90	N/A
Dollar Tree	BBB−	7.50	6.85	N/A	N/A
Family Dollar	BBB−	8.50	7.50	N/A	N/A
Kohl's	BBB−	8.50	7.25	6.50	6.35
Kroger	BBB	7.25	6.75	6.25	6.00
Lowe's	BBB+	6.70	6.10	5.45	5.00
Sherwin-Williams	BBB	6.75	5.90	N/A	N/A
The Home Depot	A	6.75	6.00	5.40	5.00
United Rentals	BB	8.15	7.25	N/A	N/A
Walmart	AA	6.75	5.90	5.50	5.15
Whole Foods Market	AA−	5.50	5.15	4.60	4.25

Source: Sellers, L.P., et al. "Valuation Methods and Dark Big-Box Theories", *The Appraisal Journal*, Vol. 87, No. 3, Summer 2019 (Exhibit 1).

1.17 Parking at a university. Parking at a large university has become a big problem. The university's administrators want to determine the average parking time of its students. An administrator inconspicuously followed 250 students and carefully recorded the time it took them to find a parking spot.
a. What is the population of interest to the university administration?
b. Identify the sample of interest to the university administration.
c. What is the experimental unit of interest to the university administration?
d. What is the variable of interest to the university administration?

1.18 Car comparison. *Carbase.my* is a Malaysian website that helps car buyers choose the perfect car. Its database is updated daily, and it provides potential buyers with detailed information on car types, brands, prices, and user reviews to help them make an informed decision. Jim would like to buy a new sedan car, so he visits *Carbase.my* to get information on the models available in different price ranges.
a. Do the prices selected from *Carbase.my* for different brands represent a population or a sample? Explain.
b. Describe what the price population is.
c. Identify the variable "car prices" as qualitative or quantitative.
d. Identify the variable "car types" as qualitative or quantitative.
e. Explain how to measure car types quantitatively.

1.19 NW **Opinion polls.** Pollsters regularly conduct opinion polls to determine the popularity rating of the current president. Suppose a poll is to be conducted tomorrow in which 2,000 individuals will be asked whether the president is doing a good or bad job. The 2,000 individuals will be selected by random-digit telephone dialing and asked the question over the phone.
a. What is the relevant population?
b. What is the variable of interest? Is it quantitative or qualitative?
c. What is the sample?
d. What is the inference of interest to the pollster?
e. What method of data collection is employed?
f. How likely is the sample to be representative?

1.20 Cybersecurity survey. The information systems organization, ISACA, conducts an annual survey of cybersecurity at firms from around the world. ISACA sends survey questionnaires via email (SurveyMonkey) to all professionals that hold ISACA's Certified Information Security Manager designation. Over 1,500 professionals participated in the 2019 cybersecurity survey. Each was asked whether or not they expect to experience a cyberattack (e.g., a Malware, hacking, or phishing attack) against their firm in the coming year. About 80% of the respondents expect to experience a cyberattack during the year (*State of Cybersecurity: 2019: Part 2*, ISACA and RSA Conference Survey).
a. Identify the population of interest to ISACA.
b. Identify the data-collection method used by ISACA. Are there any potential biases in the method used?
c. Describe the variable measured in the ISACA survey. Is it quantitative or qualitative?
d. What inference can be made from the study result?

1.21 COVID-19. The *New York Times* and the *JHU CSSE* COVID-19 report the latest worldwide coronavirus statistics by location. On March 16, 2021, it was reported that the United States, Brazil, and India had the highest total number of cases (29.5 million, 11.5 million, and 11.4 million respectively). Concurrently, *Our World in Data* provides the number of vaccine doses by location and has reported that 109 million, 11.9 million, and 32.9 million doses have been given to these three countries respectively. Adrian uses this data to conduct research into the number of COVID-19 cases compared to the number of individuals vaccinated in these three countries. He will not be drawing any estimates or making any predictions from the data collected at this stage. What type of statistical application does Adrian's research represent right now? Explain.

1.22 Entertainment market. A researcher used the ticket sales data provided by *The Ticketing Business* on Pollstar's 2020 annual mid-year analysis, in order to project the impact of COVID-19 on the entertainment market in the second half of 2020. According to the data, *Live Nation* is at the top with an average gross per ticket of $89, which is roughly the same as what it was in the previous year ($87.5 per ticket), and *AEG Presents* is in second place with 1.9 million tickets sold for $181m grossed, which is about half of what it was in mid-year 2019. *Feld Entertainment* is third with 1.57 million tickets sold, which is only a bit less than 2019. Does this study represent a descriptive or inferential statistical study? Explain.

Applying the Concepts—Intermediate

1.23 **Delivery times for online orders**. The *Journal of Marketing Research* (October 2019) published a study of delivery times for online orders. During a recent year, a major apparel retailer fulfilled all its online orders from a single distribution center (DC) located in the eastern United States. Later that year, the retailer opened a second DC located in the western United States, with the goal of reducing delivery times to western US customers. The researchers collected data on delivery times for a sample of online orders fulfilled by both the eastern and western US distribution centers for several western states. For the state of Washington, the typical delivery time was 7 business days from the eastern DC and 5 business days from the western DC. For the state of Montana, the typical delivery time was 7 business days from either of the distribution centers.

a. What is the experimental unit for this study?
b. Identify the variables measured and their type (quantitative or qualitative).
c. Explain why this is an example of inferential statistics. What inference can you make?

1.24 **Who is better at multi-tasking?** In business, employees are often asked to perform a complex task when their attention is divided (i.e., *multi-tasking). Human Factors* (May 2014) published a study designed to determine whether video game players are better than non–video game players at multi-tasking. Each in a sample of 60 college students was classified as a video game player or a non-player. Participants entered a street crossing simulator and were asked to cross a busy street at an unsigned intersection. The simulator was designed to have cars traveling at various high rates of speed in both directions. During the crossing, the students also performed a memory task as a distraction. Two variables were measured for each student: (1) a street crossing performance score (measured out of 100 points) and (2) a memory task score (measured out of 20 points). The researchers found no differences in either the street crossing performance or memory task score of video game players and non-gamers. "These results," say the researchers, "suggest that action video game players [and non-gamers] are equally susceptible to the costs of dividing attention in a complex task."

a. Identify the experimental unit for this study.
b. Identify the variables measured as quantitative or qualitative.
c. Is this an application of descriptive statistics or inferential statistics? Explain.

1.25 **Dissatisfaction among auditors and specialists.** *Behavioral Research in Accounting* (BRIA), published by the American Accounting Association, provides data on accounting and its relation to individuals and organizations based on empirical research. BRIA (Vol. 32(2), 2020) published a study on auditors' and specialists' views about the use of specialists during an audit. This study was conducted through interviews on 12 auditors (partners and managers) and 22 specialists (tax, IT, valuation, forensic) from six accounting firms. The details of interviewees' demographic characteristics, such as the identifier (ordered alphabetically), years at firm, years in position, audit experience (in years), specialist experience (in years), education, designations, gender, and age, are listed in a table.

a. Identify the data-collection method used.
b. What is the experimental unit for this study?
c. Identify the type (quantitative or qualitative) of the variables listed in the interviewees' demographics table.
d. Suppose 55% of the auditors and 63% of the specialists are dissatisfied with the current standard of specialists used in audits. Make an inference about the population of interest.

1.26 "Does the adoption of descriptive analytics impact online retailer performance? If so, how?" This is the research question of interest in a working paper published by the Harvard Business School (November 2020). The paper found that there was a 13% to 20% increase in the average weekly revenues after more than 1,000 e-commerce websites adopted a retail analytics dashboard. The paper argues that this increase in revenue did not result from product price changes or advertising (cost and platform) optimization but rather from the addition of prospecting and personalization technologies to retailer websites. The paper concludes that without using the descriptive dashboard for monitoring the diversity of the products sold, the number of transactions, the numbers of website visitors, and the revenue from customers (new or repeating), retailers are unable to gain the benefits that these technologies offer.

a. What is the experimental unit for this study?
b. Identify the type (quantitative or qualitative) of the variables measured.
c. Assume you own an online retail and plan to use this study to project your retail's performance and revenue. Is this an application of descriptive or inferential statistics? Explain.

1.27 NW **The economic return to earning an MBA.** What are the economic rewards (e.g., higher salary) to obtaining an MBA degree? This was the question of interest in an article published in the *International Economic Review* (August 2008). The researchers made inferences based on wage data collected for a sample of 3,244 individuals who sat for the Graduate Management Admissions Test (GMAT). (The GMAT exam is required for entrance into most MBA programs.) The following sampling scheme was employed. All those who took the GMAT exam in any of four selected time periods were mailed a questionnaire. Those who responded to the questionnaire were then sent three follow-up surveys (one survey every 3 months). The final sample of 3,244 represents only those individuals who responded to all four surveys. (For example, about 5,600 took the GMAT in one time period; of these, only about 800 responded to all four surveys.)

a. For this study, describe the population of interest.
b. What method was used to collect the sample data?
c. Do you think the final sample is representative of the population? Why or why not? Comment on potential biases in the sample.

1.28 **Corporate sustainability and firm characteristics.** *Corporate sustainability* refers to business practices designed around social and environmental considerations (e.g., "going green"). *Business and Society* (March 2011) published a paper on how firm size and firm type impact sustainability behaviors. The researchers added questions on sustainability to a quarterly survey of Certified Public Accountants (CPAs). The survey was sent to approximately 23,500 senior managers at CPA firms, of which 1,293 senior managers responded. (*Note:* It is not clear how the 23,500 senior managers were selected.) Due to missing

data (incomplete survey answers), only 992 surveys were analyzed. These data were used to infer whether larger firms are more likely to report sustainability policies than smaller firms and whether public firms are more likely to report sustainability policies than private firms.

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

1.29 **Property value.** Property values in Singapore are influenced by logical factors, such as economic theories and population density, and other factors, including the feel of a neighborhood and future growth expectations. The following are some of the key variables that affect property values. Classify each variable as quantitative or qualitative.

a. The number of properties supplied and demanded
b. Location (good, fair or, poor)
c. Interest rates (%)
d. Economic growth rate (%)
e. Designated parking space (Yes or No)
f. Migration level
g. Home improvements (renovated or not renovated)

1.30 **Bridge tour.** *Highestbridges.com*, created by bridge architecture design and engineering expert Eric Sakowski, will be conducting the eighth annual High Over China Bridge Tour in September 2021. During this tour, participants will visit some major bridges in China, including the Beipanjiang Bridge, which is 565 m high and, as of 2016, the highest bridge in the world. To make its selection of bridges for this tour, *Highestbridges.com* has referred to information from various repositories including, China's BRIDGE website (en.chinabridge.org.cn).

a. What is the variable of interest to the tour organizers?
b. Is the variable of interest quantitative or qualitative?
c. Are the bridges that will be visited a sample or a population in this context? Explain.
d. What data-collection method is used by the organizers?

***1.31** **Monitoring product quality.** The Wallace Company of Houston is a distributor of pipes, valves, and fittings to the refining, chemical, and petrochemical industries. The company was a recent winner of the Malcolm Baldrige National Quality Award. One of the steps the company takes to monitor the quality of its distribution process is to send out a survey twice a year to a subset of its current customers, asking the customers to rate the speed of deliveries, the accuracy of invoices, and the quality of the packaging of the products they have received from Wallace.

a. Describe the process studied.
b. Describe the variables of interest.
c. Describe the sample.
d. Describe the inferences of interest.
e. What are some of the factors that are likely to affect the reliability of the inferences?

1.32 **Guilt in decision making.** The effect of guilt emotion on how a decision maker focuses on the problem was investigated in the *Journal of Behavioral Decision Making* (January 2007). A total of 171 volunteer students participated in the experiment, where each was randomly assigned to one of three emotional states (guilt, anger, or neutral) through a reading/writing task. Immediately after the task, the students were presented with a decision problem (e.g., whether or not to spend money on repairing a very old car). The researchers found that a higher proportion of students in the guilty-state group chose to repair the car than those in the neutral-state and anger-state groups.

a. Identify the population, sample, and variables measured for this study.
b. Identify the data-collection method used.
c. What inference was made by the researcher?
d. In later chapters you will learn that the reliability of an inference is related to the size of the sample used. In addition to sample size, what factors might affect the reliability of the inference drawn in this study?

1.33 A study published in *Behavioral Research in Accounting* (Vol. 32, 2020) investigated whether client status is a significant factor in auditor–client negotiations. The researchers reached out to thousands of finance and accounting executives via a subscription-based database and attendees of an education program. In all, 85 took part in the study. The following variables pertaining to accounting executives were measured: the overall work experience of the participants (in years), their accounting-related work experience (in years), their age (in years), and their job title (CFO, Controller/Chief Accounting Officer, CPA, etc.). Later, the participants took part in an experiment to investigate if, among other factors, the status of their clients (with or without a CPA license) affected their financial reporting aggressiveness (warranty cost estimate as a % of sales). The study provided evidence that client status is a significant factor in auditor–client negotiations and can influence the reporting aggressiveness.

a. What is the population of interest to the researcher?
b. What type of data (quantitative or qualitative) is produced by each of the variables measured?
c. Identify the sample.
d. Identify the data-collection method used.
e. What inference was made by the researcher?
f. How might the selection bias impact the inference?

1.34 **Can money spent on gifts buy love?** Is the gift you purchased for that special someone really appreciated? This was the question of interest to business professors at Stanford University. Their research was published in the *Journal of Experimental Social Psychology* (Vol. 45, 2009). In one study, the researchers investigated the link between engagement ring price (dollars) and level of appreciation of the recipient (measured on a 7-point scale where 1 = "not at all" and 7 = "to a great extent"). Participants for the study were those who used a popular Web site for engaged couples. The Web site's directory was searched for those with "average" American names (e.g., "John Smith," "Sara Jones"). These individuals were then invited to participate in an online survey in exchange for a $10 gift certificate. Of the respondents, those who paid really high or really low prices for the ring were excluded, leaving a sample size of 33 respondents.

a. Identify the experimental units for this study.
b. What are the variables of interest? Are they quantitative or qualitative in nature?

c. Describe the population of interest.
d. Do you believe the sample of 33 respondents is representative of the population? Explain.
e. In a second designed study, the researchers investigated NW whether the link between gift price and level of appreciation is stronger for birthday gift-givers than for birthday gift-receivers. The participants were randomly assigned to play the role of gift-giver or gift-receiver. Assume that the sample consists of 50 individuals. Use a random number generator to randomly assign 25 individuals to play the gift-receiver role and 25 to play the gift-giver role.

Applying the Concepts—Advanced

1.35 Transit. The *American Public Transportation Association* (APTA) published a 2020 update of the Economic Impact of Public Transportation Investment. This report mentions that due to shifts in consumer patterns and constrained budgets, the economic impacts and benefits of public transportation services must be consistently documented by transit agencies. This will not only help such agencies in their decision-making processes but also enable them to get increased investments to improve urban transportation systems.
a. Construct a brief questionnaire (with two or three questions) that could be used to help the transit agencies know why they need to assess the economic impacts and benefits of transit.
b. Describe the population about which inferences could be made from the results of the survey.
c. Discuss the pros and cons of sending the questionnaire to all transit agencies versus to a sample of 200.

1.36 Random-digit dialing. To ascertain the effectiveness of their advertising campaigns, firms frequently conduct telephone interviews with consumers using *random-digit dialing*. With this approach, a random number generator mechanically creates the sample of phone numbers to be called. Each digit in the phone number is randomly selected from the possible digits 0, 1, 2, . . . , 9. Use the procedure to generate five seven-digit telephone numbers whose first three digits (area code) are 373.

1.37 Lazada Group. Lazada Group is a leading e-commerce platform in Southeast Asia. It was founded in 2012 and is available in Indonesia, Malaysia, the Philippines, Singapore, Thailand, and Vietnam. It connects this huge and diverse region through its advanced technology, logistics, and payments capabilities. In order to grow the business and serve its customers better, survey questionnaires have been sent out to identify the platform's service attributes and customer satisfaction. In a recent survey, more than 85% of Lazada Group's customers provided positive feedback and expressed satisfaction with the platform's services. However, studying the customers' negative feedback can help the platform make improvements.
a. Define the population of interest to Lazada Group for their customer satisfaction survey.
b. What variables are being measured? Are they quantitative or qualitative?
c. Is the problem of interest to Lazada Group descriptive or inferential?

***1.38 Monitoring the production of soft-drink cans.** NW The Wakefield plant of Coca-Cola and Schweppes Beverages Limited (CCSB) can produce 4,000 cans of soft drink per minute. The automated process consists of measuring and dispensing the raw ingredients into storage vessels to create the syrup, and then injecting the syrup, along with carbon dioxide, into the beverage cans. In order to monitor the subprocess that adds carbon dioxide to the cans, five filled cans are pulled off the line every 15 minutes, and the amount of carbon dioxide in each of these five cans is measured to determine whether the amounts are within prescribed limits.
a. Describe the process studied.
b. Describe the variable of interest.
c. Describe the sample.
d. Describe the inference of interest.
e. *Brix* is a unit for measuring sugar concentration. If a technician is assigned the task of estimating the average brix level of all 240,000 cans of beverage stored in a warehouse near Wakefield, will the technician be examining a process or a population? Explain.

1.39 Sampling TV markets for a court case. A recent court case involved a claim of satellite television subscribers obtaining illegal access to local TV stations. The defendant (the satellite TV company) wanted to sample TV markets nationwide and determine the percentage of its subscribers in each sampled market who have illegal access to local TV stations. To do this, the defendant's expert witness drew a rectangular grid over the continental United States, with horizontal and vertical grid lines every .02 degrees of latitude and longitude, respectively. This created a total of 500 rows and 1,000 columns, or (500)(1,000) = 500,000 intersections. The plan was to randomly sample 900 intersection points and include the TV market at each intersection in the sample. Explain how you could use a random number generator to obtain a random sample of 900 intersections. Develop at least two plans: one that numbers the intersections from 1 to 500,000 prior to selection and another that selects the row and column of each sampled intersection (from the total of 500 rows and 1,000 columns).

Critical Thinking Challenge

1.40 *20/20* survey exposé. Refer to the "Statistics in Action" box of this chapter (p. 19). Recall that the popular prime-time ABC television program *20/20* presented several misleading (and possibly unethical) surveys in a segment titled "Fact or Fiction?—Exposés of So-Called Surveys." The information reported from two of these surveys and several others is listed here (actual survey facts are provided in parentheses).
- *Quaker Oats study:* Eating oat bran is a cheap and easy way to reduce your cholesterol count. (Fact: Diet must consist of nothing but oat bran to achieve a slightly lower cholesterol count.)
- *March of Dimes report:* Domestic violence causes more birth defects than all medical issues combined. (Fact: No study—false report.)
- *American Association of University Women (AAUW) study:* Only 29% of high school girls are happy

with themselves, compared to 66% of elementary school girls. (Fact: Of 3,000 high school girls, 29% responded, "Always true" to the statement "I am happy the way I am." Most answered, "Sort of true" and "Sometimes true.")

- *Food Research and Action Center study:* One in four American children under age 12 is hungry or at risk of hunger. (Fact: Based on responses to questions: "Do you ever cut the size of meals?" and "Do you ever eat less than you feel you should?" and "Did you ever rely on limited numbers of foods to feed your children because you were running out of money to buy food for a meal?")
- *McKinsey survey on the health reform act:* Thirty percent of employers would "definitely" or "probably" stop offering health coverage to their employees if the government-sponsored act is passed. (Fact: Employers were asked leading questions that made it seem logical for them to stop offering insurance. For example, respondents were told that the new health insurance exchanges would become "an easy, affordable way for individuals to obtain health insurance" outside the company. Then they were given examples of how little their workers would pay for this insurance. Only then were they asked how likely they would be to stop offering health insurance.)

a. Refer to the Quaker Oats study relating oat bran to cholesterol levels. Discuss why it is unethical to report the results as stated.

b. Consider the false March of Dimes report on domestic violence and birth defects. Discuss the type of data required to investigate the impact of domestic violence on birth defects. What data-collection method would you recommend?

c. Refer to the AAUW study of self-esteem of high school girls. Explain why the results of the study are likely to be misleading. What data might be appropriate for assessing the self-esteem of high school girls?

d. Refer to the Food Research and Action Center study of hunger in America. Explain why the results of the study are likely to be misleading. What data would provide insight into the proportion of hungry American children?

e. Refer to the McKinsey survey on the health reform act. Explain what a "leading question" is and why it might produce responses that bias the results.

ACTIVITY 1.1 *Keep the Change:* Collecting Data

Bank of America has a savings program called *Keep the Change.* Each time a customer enrolled in the program uses his or her debit card to make a purchase, the difference between the purchase total and the next higher dollar amount is transferred from the customer's checking account to a savings account. For example, if you were enrolled in the program and used your debit card to purchase a latte for $3.75, then $0.25 would be transferred from your checking to your savings account. For the first 90 days that a customer is enrolled in the program, Bank of America matches the amounts transferred up to $250. In this and subsequent activities, we will investigate the potential benefit to the customer and cost to the bank.

1. Simulate the program by keeping track of all purchases that you make during one week that could be made with a debit card, even if you use a different form of payment. For each purchase, record both the purchase total and the amount that would be transferred from checking to savings with the *Keep the Change* program.

2. You now have two sets of data: *Purchase Totals* and *Amounts Transferred*. Both sets contain quantitative data. For each data set, identify the corresponding naturally occurring numerical scale. Explain why each set has an obvious lower bound but only one set has a definite upper bound.

3. Find the total of the amounts transferred for the one-week period. Because 90 days is approximately 13 weeks, multiply the total by 13 to estimate how much the bank would have to match during the first 90 days. Form a third data set, *Bank Matching*, by collecting the 90-day estimates of all the students in your class. Identify the naturally occurring scale, including bounds, for this set of data.

Keep the data sets from this activity for use in other activities. We suggest you save the data using statistical software (e.g., Minitab) or a graphing calculator.

ACTIVITY 1.2 Identifying Misleading Statistics

In the *Statistics in Action* feature at the beginning of this chapter, several examples of false or misleading statistics were discussed. Claims such as *One in four American children under age 12 is hungry or at risk of hunger* are often used to persuade the public or the government to donate or allocate more money to charitable groups that feed the poor. Researchers sometimes claim a relationship exists between two seemingly unrelated quantities such as a CEO's golf handicap and the company's stock performance; such relationships are often weak at best and of little practical importance. Read the *Statistics in Action* and *Statistics in Action Revisited* features in this chapter before completing this activity.

1. Look for an article in a newspaper or on the Internet in which a large proportion or percentage of a population is purported to be "at risk" of some calamity, as in the childhood hunger example. Does the article cite a source or provide any information to support the proportion or

(*continued*)

percentage reported? Is the goal of the article to persuade some individual or group to take some action? If so, what action is being requested? Do you believe that the writer of the article may have some motive for exaggerating the problem? If so, give some possible motives.

2. Look for another article in which a relationship between two seemingly unrelated quantities is purported to exist, as in the CEO golf handicap and stock performance study. Select an article that contains some information on how the data were collected. Identify the target population and the data-collection method. Based on what is presented in the article, do you believe that the data are representative of the population? Explain. Is the purported relationship of any practical interest? Explain.

References

Careers in Statistics, American Statistical Association, 2011 (www.amstat.org).

Cochran, W. G. *Sampling Techniques*, 3rd ed. New York: Wiley, 1977.

Deming, W. E. *Sample Design in Business Research.* New York: Wiley, 1963.

Dillman, D. A., Smyth, J. D., and Christian, L. M. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. New York: Wiley, 2008.

Ethical Guidelines for Statistical Practice, American Statistical Association, 1999 (www.amstat.org).

Hahn, G. J. and Doganaksoy, N. *The Role of Statistics in Business and Industry*. New York: Wiley, 2008.

Huff, D. *How to Lie with Statistics*. New York: Norton, 1982 (paperback 1993).

Hoerl, R. and Snee, R. *Statistical Thinking: Improving Business Performance*. Boston: Duxbury, 2002.

Kish, L. *Survey Sampling.* New York: Wiley, 1965 (paperback, 1995).

Peck, R., Casella, G., Cobb, G., Hoerl, R., Nolan, D., Starbuck, R., and Stern, H. *Statistics: A Guide to the Unknown*, 4th ed. Cengage Learning, 2005.

Scheaffer, R., Mendenhall, W., and Ott, R. L. *Elementary Survey Sampling*, 6th ed. Boston: Duxbury, 2005.

What Is a Survey? American Statistical Association (F. Scheuren, editor), 2nd ed., 2005 (www.amstat.org).

USING TECHNOLOGY

Technology images shown here are taken from Minitab 19, XLSTAT 2019 and StatCrunch 3.0.

Minitab: Accessing and Listing Data

When you start a Minitab session, you will see a screen similar to Figure 1.M.1. The bottom portion of the screen is an empty spreadsheet—called a Minitab worksheet—with columns representing variables and rows representing observations (or cases). The very top of the screen is the Minitab main menu bar, with buttons for the different functions and procedures available in Minitab. Once you have entered data into the spreadsheet, you can analyze the data by clicking the appropriate menu buttons. The results will appear in the window above the worksheet.

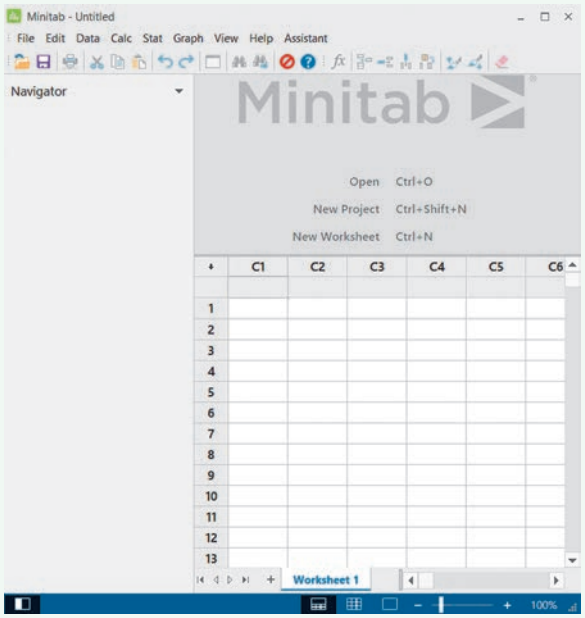


Figure 1.M.1 Initial screen viewed by the Minitab user

Entering Data

Create a Minitab data file by entering data directly into the worksheet. Figure 1.M.2 shows data entered for a variable called "GPA." Name the variables (columns) by typing in the name of each variable in the box below the column number.

	C1	C2
	GPA	
1	2.66	
2	3.50	
3	3.91	
4	2.85	
5	3.04	

Figure 1.M.2 Data entered into the Minitab worksheet

Opening a Minitab Data File

If the data have been previously saved as a Minitab (.mtw) file, access the data as follows.

Step 1 Click the "File" button on the menu bar, and then click "Open" as shown in Figure 1.M.3. A dialog box similar to Figure 1.M.4 will appear.

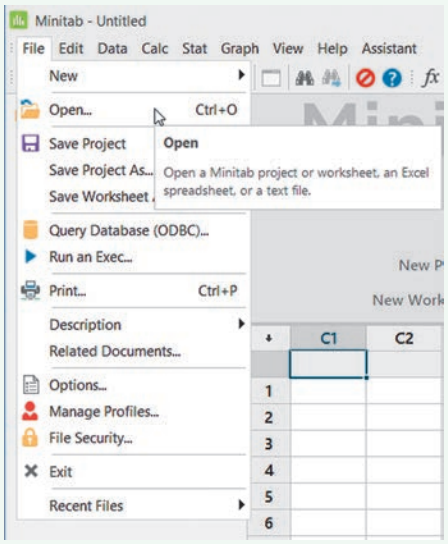


Figure 1.M.3 Options for opening a Minitab data file

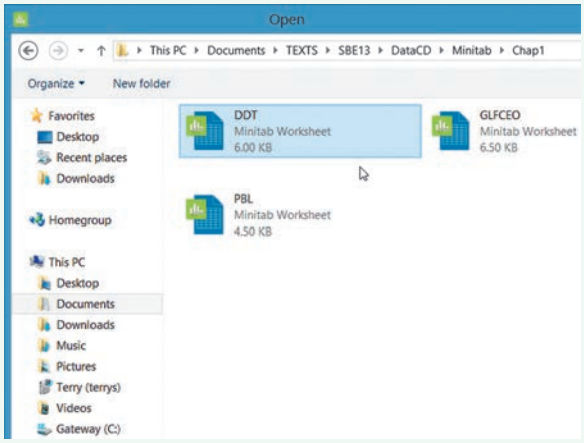


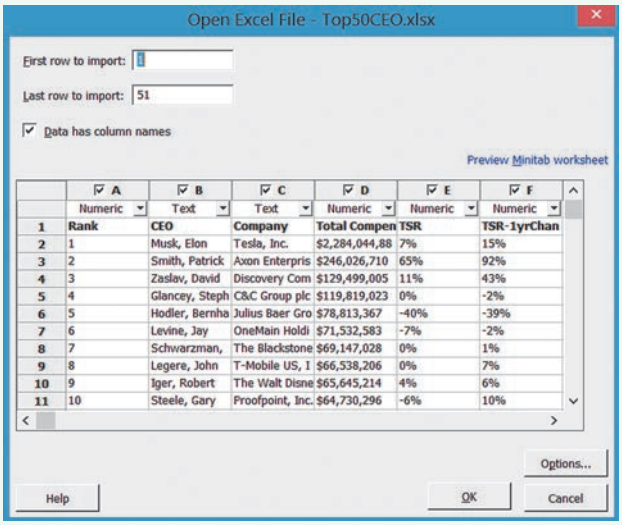
Figure 1.M.4 Selecting the Minitab data file to open

Step 2 Specify the location (folder) that contains the data, click on the Minitab data file, and then click "Open" (see Figure 1.M.4). The data will appear in the Minitab worksheet, as shown in Figure 1.M.5 at bottom of the page.

Accessing Data from an Excel File

Step 1 Click the "File" button on the menu bar, and then click "Open" as shown in Figure 1.M.3.

Step 2 Specify the location (folder) that contains the Excel data file and the file type (e.g., .xls), and then click on the file name, and "Open." This reveals the dialog box shown in Figure 1.M.6.

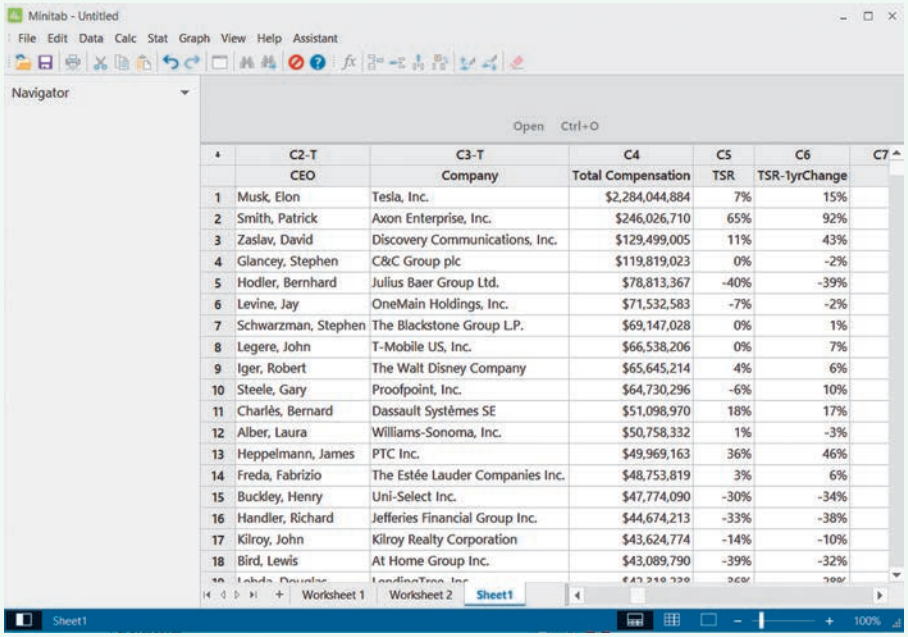


	A	B	C	D	E	F
	Numeric	Text	Text	Numeric	Numeric	Numeric
1	Rank	CEO	Company	Total Compen	TSR	TSR-1yrChan
2	1	Musk, Elon	Tesla, Inc.	$2,284,044,88	7%	15%
3	2	Smith, Patrick	Axon Enterpris	$246,026,710	65%	92%
4	3	Zaslav, David	Discovery Com	$129,499,005	11%	43%
5	4	Glancey, Steph	C&C Group plc	$119,819,023	0%	-2%
6	5	Hodler, Bernha	Julius Baer Gro	$78,813,367	-40%	-39%
7	6	Levine, Jay	OneMain Holdi	$71,532,583	-7%	-2%
8	7	Schwarzman,	The Blackstone	$69,147,028	0%	1%
9	8	Legere, John	T-Mobile US, I	$66,538,206	0%	7%
10	9	Iger, Robert	The Walt Disne	$65,645,214	4%	6%
11	10	Steele, Gary	Proofpoint, Inc.	$64,730,296	-6%	10%

Figure 1.M.6 Open Excel file dialog box

Step 3 If the Excel spreadsheet has column names, check the appropriate box on the screen. Then click "OK." The data will appear in the MINITAB worksheet.

Reminder: If the Excel spreadsheet does not have column names, the variables (columns) can be named by typing in the name of each variable in the box under the column number.



	C2-T	C3-T	C4	C5	C6	C7
	CEO	Company	Total Compensation	TSR	TSR-1yrChange	
1	Musk, Elon	Tesla, Inc.	$2,284,044,884	7%	15%	
2	Smith, Patrick	Axon Enterprise, Inc.	$246,026,710	65%	92%	
3	Zaslav, David	Discovery Communications, Inc.	$129,499,005	11%	43%	
4	Glancey, Stephen	C&C Group plc	$119,819,023	0%	-2%	
5	Hodler, Bernhard	Julius Baer Group Ltd.	$78,813,367	-40%	-39%	
6	Levine, Jay	OneMain Holdings, Inc.	$71,532,583	-7%	-2%	
7	Schwarzman, Stephen	The Blackstone Group L.P.	$69,147,028	0%	1%	
8	Legere, John	T-Mobile US, Inc.	$66,538,206	0%	7%	
9	Iger, Robert	The Walt Disney Company	$65,645,214	4%	6%	
10	Steele, Gary	Proofpoint, Inc.	$64,730,296	-6%	10%	
11	Charlès, Bernard	Dassault Systèmes SE	$51,098,970	18%	17%	
12	Alber, Laura	Williams-Sonoma, Inc.	$50,758,332	1%	-3%	
13	Heppelmann, James	PTC Inc.	$49,969,163	36%	46%	
14	Freda, Fabrizio	The Estée Lauder Companies Inc.	$48,753,819	3%	6%	
15	Buckley, Henry	Uni-Select Inc.	$47,774,090	-30%	-34%	
16	Handler, Richard	Jefferies Financial Group Inc.	$44,674,213	-33%	-38%	
17	Kilroy, John	Kilroy Realty Corporation	$43,624,774	-14%	-10%	
18	Bird, Lewis	At Home Group Inc.	$43,089,790	-39%	-32%	

Figure 1.M.5 The Minitab worksheet showing the opened Minitab data file

Listing (Printing) Data

Step 1 Click on the "Data" button on the Minitab main menu bar, and then click on "Display Data." (See Figure 1.M.7.) The resulting menu, or dialog box, appears as in Figure 1.M.8.

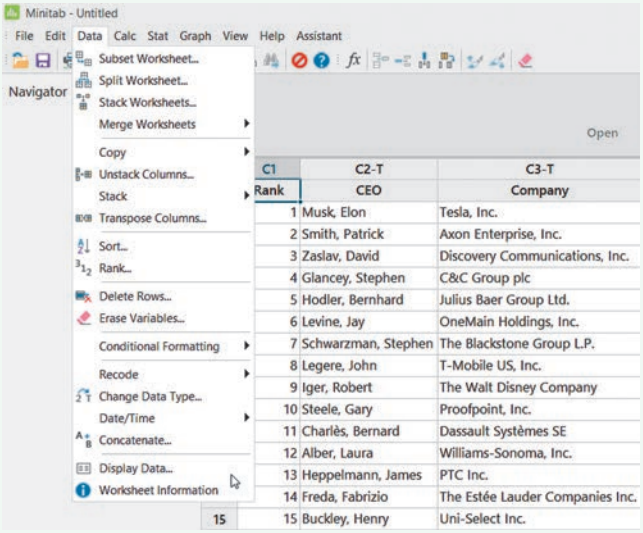


Figure 1.M.7 Minitab options for displaying data

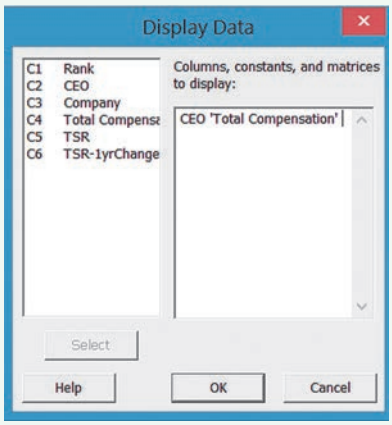


Figure 1.M.8 Minitab Display Data dialog box

Step 2 Enter the names of the variables you want to print in the "Columns, constants, and matrices to display" box (you can do this by simply double clicking on the variables), and then click "OK." The printout will show up on your Minitab session screen.

Minitab: Generating a Random Sample

Step 1 Click on the "Calc" button on the Minitab menu bar and then click on "Random Data," and finally, click on "Sample From Columns," as shown in Figure 1.M.9. The resulting dialog box appears as shown in Figure 1.M.10.

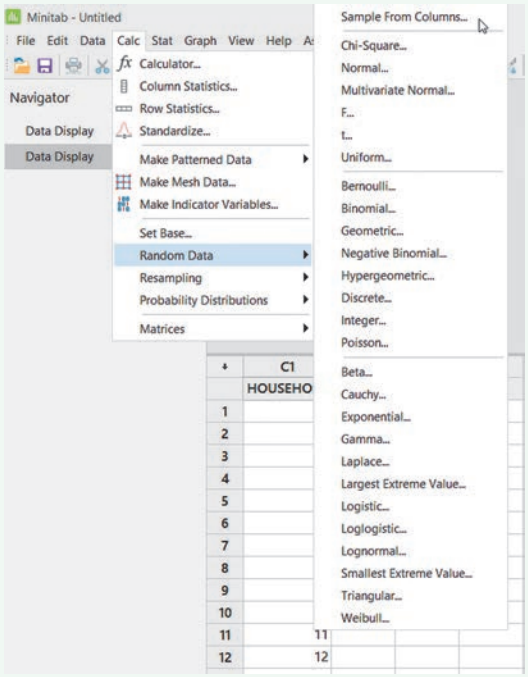


Figure 1.M.9 Minitab menu options for sampling from a data set

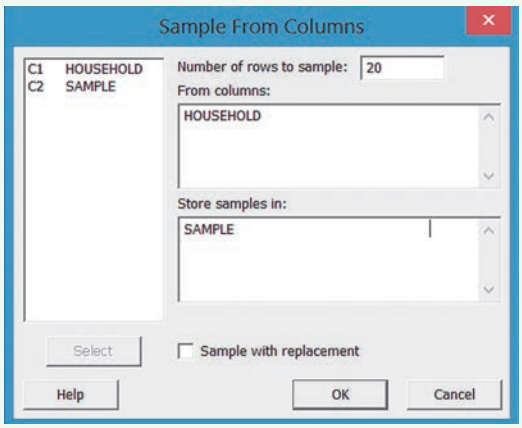


Figure 1.M.10 Minitab options for selecting a random sample from worksheet columns

Step 2 Specify the sample size (i.e., number of rows), the variable(s) to be sampled, and the column(s) where you want to save the sample.

Step 3 Click "OK" and the Minitab worksheet will reappear with the values of the variable for the selected (sampled) cases in the column specified.

In Minitab, you can also generate a sample of case numbers (e.g., a sample of cases from a population of cases numbered from 1 to 500).

Step 1 From the Minitab menu, click on the "Calc" button and then click on "Random Data," and finally, click on the "Integer" option (see Figure 1.M.9).

Step 2 In the resulting dialog box (shown in Figure 1.M.11), specify the number of cases (rows, i.e., the sample size), and the column where the case numbers selected will be stored. Also, specify the total number of observations in the population in the "Maximum value" box.

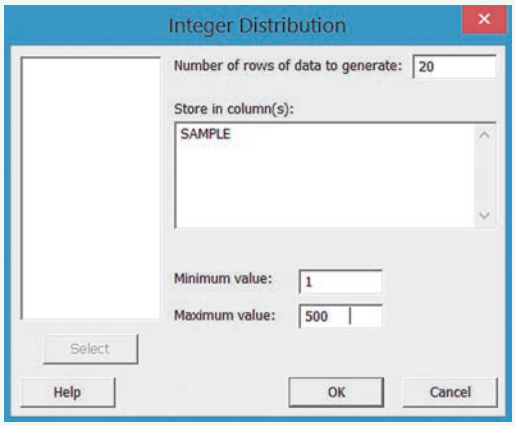


Figure 1.M.11 Minitab options for selecting a random sample of cases

Step 3 Click "OK" and the Minitab worksheet will reappear with the case numbers for the selected (sampled) cases in the column specified.

[*Note:* If you want the option of generating the same (identical) sample multiple times from the data set, then first click on the "Set Base" option shown in Figure 1.M.9. Specify an integer in the resulting dialog box. If you always select the same integer. Minitab will select the same sample when you choose the random sampling options.]

XLSTAT: Accessing and Listing Data

When you open XLSTAT, you will see a screen similar to Figure 1.E.1. The majority of the screen window is a spreadsheet—called an Excel workbook—with columns (labeled A, B, C, etc.) representing variables, and rows representing observations (or cases). The very top of the screen is the Excel main menu bar, with buttons for the different functions and procedures available in Excel. XLSTAT will be one of the menu options. Once you have entered data into the spreadsheet, you can analyze the data by clicking the appropriate XLSTAT menu buttons. The results will appear in a new workbook.

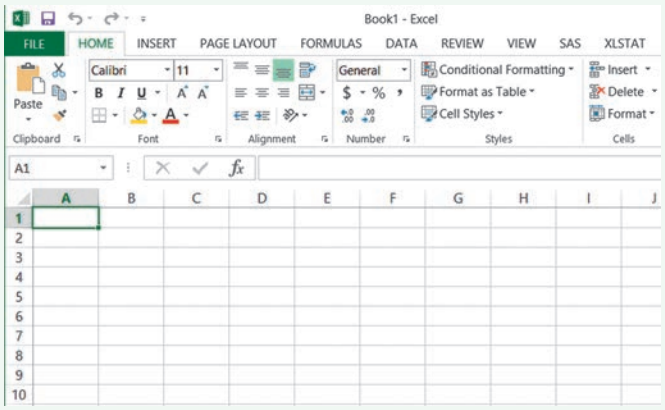


Figure 1.E.1 Initial screen viewed by XLSTAT user

Entering Data

Enter data directly into the appropriate row and column of the spreadsheet. Figure 1.E.2 shows data entered in the first (A) column. Optionally, you can add names for the variables (columns) in the first row of the workbook.

	A	B
1	GPA	
2	2.66	
3	3.91	
4	3.08	
5	2.74	
6	3.22	
7		

Figure 1.E.2 Data entered into the Excel workbook

Opening an Excel File

If the data have been previously saved as an Excel (.xls) file, access the data as follows.

Step 1 Click "File" at the far left of the menu bar, and then click "Open," as shown in Figure 1.E.3.

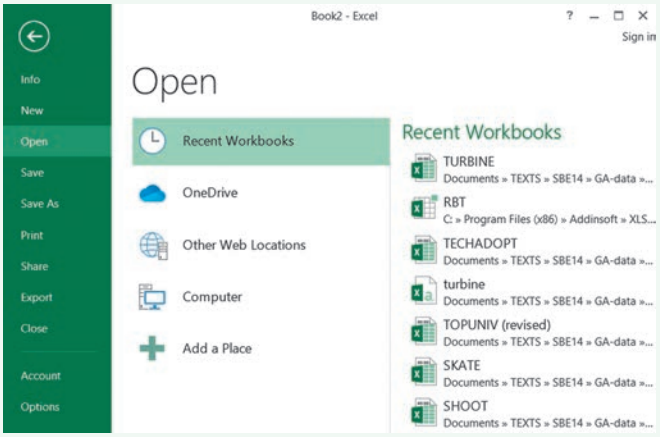


Figure 1.E.3 Options for opening an Excel file

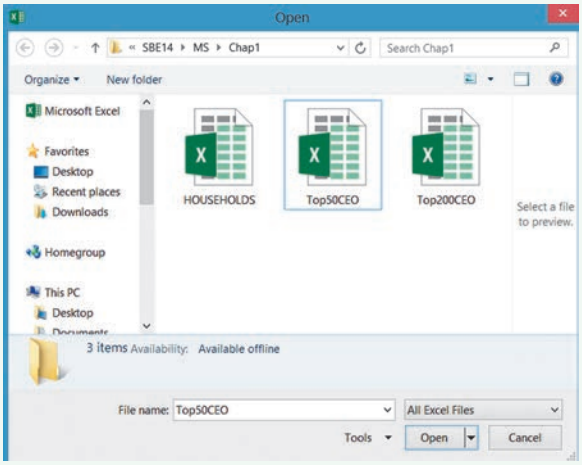
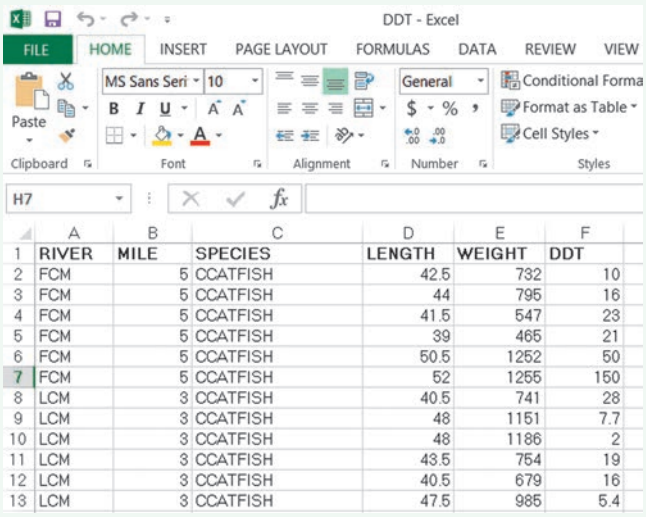


Figure 1.E.4 Selecting the Excel file to open

Step 2 Specify the location (folder) that contains the data, click on the Excel file, and then click "Open" (see Figure 1.E.4). The data will appear in the Excel spreadsheet as shown in Figure 1.E.5.



	A	B	C	D	E	F
1	RIVER	MILE	SPECIES	LENGTH	WEIGHT	DDT
2	FCM	5	CCATFISH	42.5	732	10
3	FCM	5	CCATFISH	44	795	16
4	FCM	5	CCATFISH	41.5	547	23
5	FCM	5	CCATFISH	39	465	21
6	FCM	5	CCATFISH	50.5	1252	50
7	FCM	5	CCATFISH	52	1255	150
8	LCM	3	CCATFISH	40.5	741	28
9	LCM	3	CCATFISH	48	1151	7.7
10	LCM	3	CCATFISH	48	1186	2
11	LCM	3	CCATFISH	43.5	754	19
12	LCM	3	CCATFISH	40.5	679	16
13	LCM	3	CCATFISH	47.5	985	5.4

Figure 1.E.5 The Excel spreadsheet showing the opened Excel file

Accessing Data from a .TXT or .DAT File

Step 1 Click "File", and then click "Open,"

Step 2 Specify the location (folder) that contains the data file and the file type, and then click on the file name and click on "Open," (similar to Figure 1.E.4). The Excel Text Import Wizard opens (Figure 1.E.6).

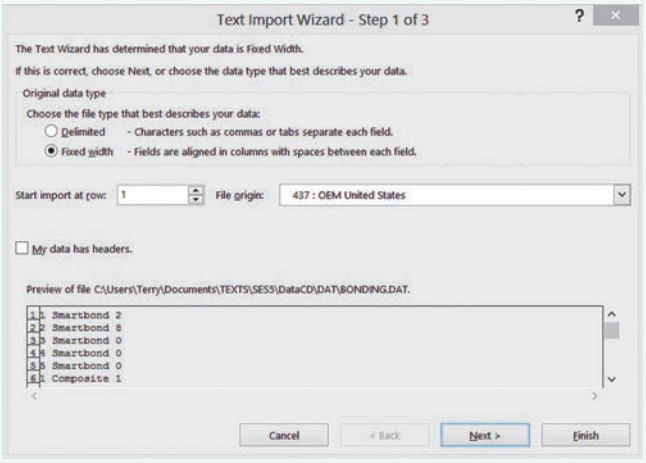


Figure 1.E.6 Excel Text Import Wizard, Screen 1

Step 3 Make the appropriate selections on the screen, and click "Next" to go to the next screen; then click "Next" again.

Step 4 When finished, click "Finish." The Excel workbook will reappear with the data from the .txt or .dat file.

Naming Variables

Step 1 Select "Insert" from the Excel main menu, and then select "Rows." A blank (empty) row will be added in the first row of the spreadsheet.

Step 2 Type the name of each variable in the first row under the appropriate column.

Listing (Printing) Data

Step 1 Click "File" on the main menu bar.

Step 2 Click on "Print."

XLSTAT: Generating a Random Sample

To obtain a random sample of data from an Excel file, perform the following:

Step 1 Highlight the column of data that you want to sample from on the Excel workbook.

Step 2 Select "XLSTAT" from the main menu bar, then "Preparing Data", then "Data Sampling", as shown in Figure 1.E.7.

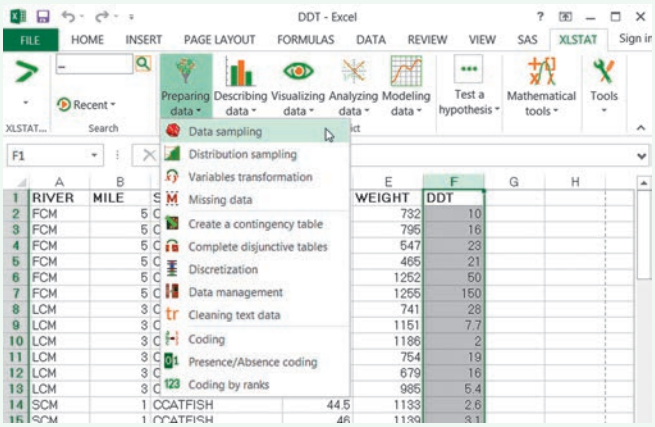


Figure 1.E.7 XLSTAT menu options for generating a random sample of numbers

Step 3 On the "Data sampling" dialog box (see Figure 1.E.8), the column of data should appear in the "Data" box at the top. Now specify "Random without replacement" in the "Sampling" box and select the "Sample size." Then click "OK" and "Continue" to generate the random sample. (The sampled data values will appear on a new Excel workbook.)

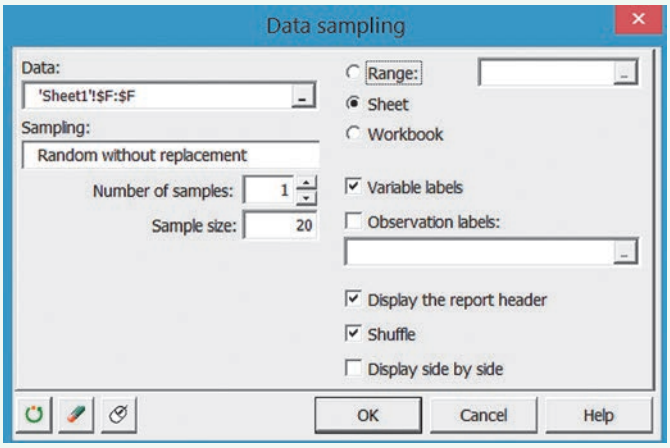


Figure 1.E.8 XLSTAT Data Sampling Dialog Box

StatCrunch: Accessing and Loading Data

When you start a StatCrunch session, you will see a screen similar to the one shown in Figure 1.S.1. The main portion of the screen is an empty spreadsheet with columns representing variables and rows representing observations (or cases). The very top of the screen is the StatCrunch main menu bar, with buttons for the different functions and procedures available in StatCrunch. Once you have entered data into the spreadsheet, you can analyze the data by clicking the appropriate menu buttons. The results will appear in another window.

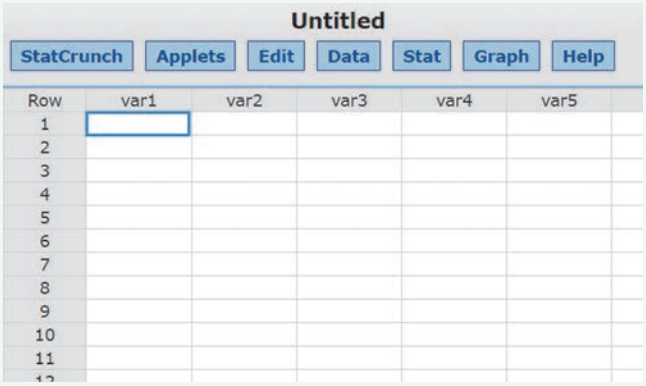


Figure 1.S.1 Initial screen viewed by the StatCrunch user

Entering Data

You can enter data directly into the spreadsheet. Figure 1.S.2 shows data entered for a variable called "GPA." Name the variables (columns) by clicking on the column number cell (e.g., "var1") and typing in the new name.

Row	GPA	var2
1	3.95	
2	2.77	
3	3.02	
4	2.99	
5	3.56	
6		

Figure 1.S.2 Data entered into the StatCrunch spreadsheet

Loading Data from a File

If the data have been previously saved as an Excel (.xls) or text (.txt) or raw data (.dat) file, access the data as follows.

Step 1 Click "Data" on the main menu bar. Then click "Load", "From file", and "on my computer", as shown in Figure 1.S.3.

Step 2 On the resulting dialog box (see Figure 1.S.4), click "Choose file." Specify the folder where the file resides, and then click on the file name. Then click "Open", as shown in Figure 1.S.5.

Step 3 When you return to the Load data dialog box, the data file name will appear. Click "Load" at the bottom of the screen. The data will appear in the spreadsheet (see Figure 1.S.6).

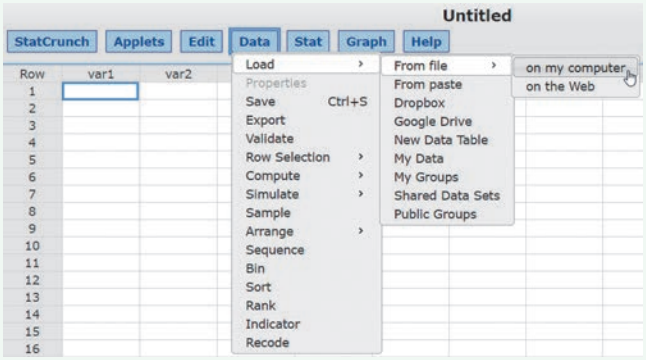


Figure 1.S.3 StatCrunch options for loading data

Figure 1.S.4 StatCrunch Load Data dialog box

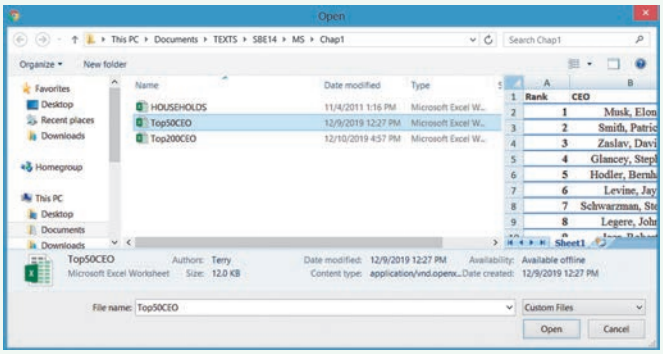


Figure 1.S.5 Open the selected data file

Row	CEO	Company	Total Compensation	TSR	TSR-1yrChange
1	Musk, Elon	Tesla, Inc.	2.2840449e9	7	15
2	Smith, Patrick	Axon Enterprise, Inc.	2.4602671e8	65	92
3	Zaslav, David	Discovery Communica	1.2949901e8	11	43
4	Glancey, Stephen	C&C Group plc	1.1981902e8	0	-2
5	Hodler, Bernhard	Julius Baer Group Ltd	78813367	-40	-39
6	Levine, Jay	OneMain Holdings, In	71532583	-7	-2
7	Schwarzman, Step	The Blackstone Group	69147028	0	1
8	Legere, John	T-Mobile US, Inc.	66538206	0	7
9	Iger, Robert	The Walt Disney Com	65645214	4	6
10	Steele, Gary	Proofpoint, Inc.	64730296	-6	10
11	Charlès, Bernard	Dassault Systèmes SE	51098970	18	17

Figure 1.S.6 XLSTAT spreadsheet showing the selected data file

StatCrunch: Generating a Random Sample

Step 1 Click on the "Data" button on the StatCrunch menu bar and then click on "Sample", as shown in Figure 1.S.7. The resulting dialog box appears as shown in Figure 1.S.8.

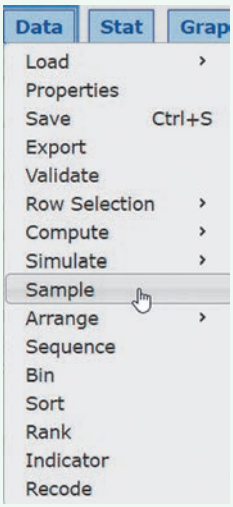


Figure 1.S.7 XLSTAT menu options for obtaining a random sample

Step 2 Specify the column (or variable) that you want to sample data from in the "Select columns" box, and enter a value for the sample size, as shown in Figure 1.S.8.

Step 3 Click "Compute" and the StatCrunch worksheet will reappear with the values of the variable for the selected (sampled) cases in a new column.

[*Note:* If you want the option of generating the same (identical) sample multiple times from the data set, then select "Use fixed seed" and specify an integer for the seed at the bottom of the dialog box shown in Figure 1.S.8. If you always select the same integer, StatCrunch will select the same sample when you choose the random sampling options.]

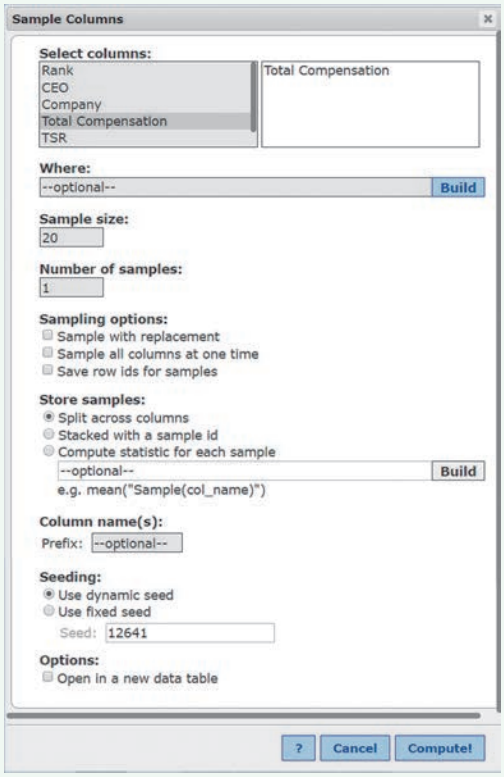


Figure 1.S.8 XLSTAT Sample Columns dialog box

2

CONTENTS

WHERE WE'VE BEEN

- Examined the difference between inferential and descriptive statistics
- Described the key elements of a statistical problem
- Learned about the two types of data—quantitative and qualitative
- Discussed the role of statistical thinking in managerial decision making

WHERE WE'RE GOING

- Describe qualitative data using graphs (2.1)
- Describe quantitative data using graphs (2.2)
- Describe quantitative data using numerical measures (2.3–2.7)
- Describe the relationship between two quantitative variables using graphs (2.8–2.9)
- Detecting descriptive methods that distort the truth (2.10)

Methods for Describing Sets of Data

STATISTICS IN ACTION **Can Money Buy Love?**

*Every day, millions of shoppers hit the stores in full force—both online and on foot—searching frantically for the perfect gift. . . . Americans [spend] over $30 billion at retail stores in the month of December alone. [Yet] many dread the thought of buying gifts; they worry that their purchases will disappoint rather than delight the intended recipients.**

With this paragraph, Stanford University Graduate School of Business researchers Francis J. Flynn and Gabrielle S. Adams introduce their study "Money Can't Buy Love: Asymmetric Beliefs About Gift Price and Feelings of Appreciation," published in the *Journal of Experimental Social Psychology* (Vol. 45, 2009). The researchers investigated the relationship between the price paid for a gift and the level of appreciation felt by the recipient. Gift-givers who spend more money on a gift often do so to send a strong signal of their love to the recipient. The researchers theorized that these gift-givers would expect the recipient to express a high level of appreciation for the gift. However, the researchers did not expect gift-recipients to associate a greater level of appreciation with a higher gift price. That is, "The link between gift price and feelings of appreciation will be stronger for gift-givers than for gift-recipients."

To test this theory, the researchers conducted an experimental study involving a representative sample of 237 adults from across the nation. Each subject completed an online survey in exchange for a $5 gift certificate to a major online retailer. The survey asked questions about a birthday gift that the subject either received or gave. The participants were randomly assigned to the role of either gift-giver or

**Source:* Republished with permission of Elsevier, from Money can't buy love: Asymmetric beliefs about gift price and feelings of appreciation in *Journal of Experimental Social Psychology*, Francis J. Flynn and Gabrielle S. Adams, Volume 45, no. 02, pp. 404–409, 2009; permission conveyed through Copyright Clearance Center, Inc.

STATISTICS IN ACTION

(*continued*)

gift-receiver. (In other words, gift-givers were asked about a birthday gift they recently gave, while gift-recipients were asked about a birthday gift they recently received.) Gifts of cash, gift cards, or gift certificates were excluded from the study. Data were collected on the following variables measured for each participant:

1. *Role* (gift-giver or gift-recipient)
2. *Gender* (male or female)
3. *Gift price* (measured in dollars)
4. *Feeling of appreciation* (measured on a 7-point scale in response to the question: "To what extent do you or does the recipient appreciate this gift?," where 1 = "Not at all," 2 = "A little," 3 = "More than a little," 4 = "Somewhat," 5 = "Moderately so," 6 = "Very much," and 7 = "To a great extent")
5. *Feeling of gratefulness* (measured on a 7-point scale in response to the question: "To what extent do you or does the recipient feel grateful for this gift?," where 1 = "Not at all," 2 = "A little," 3 = "More than a little," 4 = "Somewhat," 5 = "Moderately so," 6 = "Very much," and 7 = "To a great extent")
6. *Overall level of appreciation* (measured as the sum of the two 7-point scales—possible values are 2, 3, 4, . . . , 13, and 14)

These data are saved in the **BUYLOV** file.

The Stanford University researchers' analysis of the data led them to conclude that "gift-givers and gift-receivers disagree about the link between gift price and gift-recipients' feelings of appreciation. Givers anticipated that recipients would appreciate more expensive gifts, but gift-recipients did not base their feelings of appreciation on how much the gift cost."

In the following *Statistics in Action Revisited* sections, we apply the graphical and numerical descriptive techniques of this chapter to the **BUYLOV** data to demonstrate the conclusions reached by the Stanford University researchers.

STATISTICS IN ACTION REVISITED

Interpreting pie charts and bar graphs **(p. 64)**

Interpreting histograms **(p. 75)**

Interpreting numerical descriptive measures **(p. 102)**

Detecting outliers **(p. 117)**

Interpreting scatterplots **(p. 123)**

Data Set: BUYLOVE

Suppose you wish to evaluate the managerial capabilities of a class of 400 MBA students based on their Graduate Management Aptitude Test (GMAT) scores. How would you describe these 400 measurements? Characteristics of the data set include the typical or most frequent GMAT score, the variability in the scores, the highest and lowest scores, the "shape" of the data, and whether or not the data set contains any unusual scores. Extracting this information by "eye-balling" the data isn't easy. The 400 scores may provide too many bits of information for our minds to comprehend. Clearly, we need some formal methods for summarizing and characterizing the information in such a data set. Methods for describing data sets are also essential for statistical inference. Most populations make for large data sets. Consequently, we need methods for describing a sample data set that let us make statements (inferences) about the population from which the sample was drawn.

Two methods for describing data are presented in this chapter, one *graphical* and the other *numerical*. Both play an important role in statistics. Section 2.1 presents both graphical and numerical methods for describing qualitative data. Graphical methods for describing quantitative data are presented in Sections 2.2 and 2.7, and optional Sections 2.8 and 2.9; numerical descriptive methods for quantitative data are presented in Sections 2.3–2.6. We end this chapter with a section on the *misuse* of descriptive techniques.

2.1 Describing Qualitative Data

Recall the study of executive compensation by 24/7 Wall Street (see Study 1.2 in Section 1.2). A similar study by Equilar, Inc., was reported in the *New York Times* (May 24, 2019.) In addition to salary information, personal data on the CEOs was collected, including level of education. Do most CEOs have advanced degrees, such as master's degrees or doctorates? To answer this question, Table 2.1 gives the highest college degree obtained (bachelor's, MBA, master's, law, PhD, or no terminal degree) for each of the 50 highest paid CEOs in 2018.

Table 2.1 Data on 50 Highest Paid CEOs

Rank	CEO (Company)	Total Compensation ($ millions)	Degree	Age
1	**Elon Musk** Tesla	2,284.04	Bachelor's	48
2	**David M. Zaslav** Discovery	129.50	Law	60
3	**Nikesh Arora** Palo Alto Networks	125.07	Bachelor's	51
4	**Mark V. Hurd** Oracle	108.30	Bachelor's	62
5	**Safra A. Catz** Oracle	108.28	Law	50
6	**John J. Legere** T-Mobile US	66.54	MBA	61
7	**Robert A. Iger** Walt Disney	65.65	Bachelor's	68
8	**James Heppelmann** PTC	49.97	Bachelor's	54
9	**Fabrizio Freda** Estée Lauder	48.16	Bachelor's	61
10	**Vivek Shah** j2 Global	45.06	Bachelor's	45
11	**Richard B. Handler** Jefferies Financial	44.67	MBA	58
12	**James R. Murdoch** 21st Century Fox	44.42	None	47
13	**Stephen P. MacMillan** Hologic	42.04	Bachelor's	54
14	**Joseph M. Hogan** Align Technology	41.76	MBA	61
15	**Paul C. Saville** NVR	39.13	MBA	63
16	**Daniel H. Schulman** PayPal	37.76	MBA	61
17	**Reed Hastings** Netflix	36.08	Master's	59
18	**Jeff K. Storey** CenturyLink	35.66	Master's	59
19	**Brian R. Niccol** Chipotle Mex Grill	33.52	MBA	44
20	**Robert A. Kotick** Activision Blizzard	30.84	Bachelor's	55
21	**James Dimon** JPMorgan Chase	30.02	MBA	63
22	**Howard W. Lutnick** BGC Partners	29.69	Bachelor's	58
23	**Brian L. Roberts** Comcast	29.33	Bachelor's	60
24	**Shantanu Narayen** Adobe	28.40	MBA	56
25	**Marc Benioff** salesforce.com	28.39	Bachelor's	55
26	**Hamid R. Moghadam** Prologis	28.20	MBA	63
27	**Richard J. Tobin** Dover	27.93	MBA	55
28	**Robert Greenberg** Skechers U.S.A.	27.36	Bachelor's	78
29	**Joseph R. Ianniello** CBS	27.36	MBA	53
30	**Laura Alber** Williams-Sonoma	27.25	Bachelor's	51
31	**James P. Gorman** Morgan Stanley	26.55	MBA	61
32	**Laurence D. Fink** BlackRock	26.54	MBA	67
33	**Leonard S. Schleifer** Regeneron Pharm	26.52	PhD	67
34	**Michael F. Neidorff** Centene	26.12	Master's	76
35	**Satya Nadella** Microsoft	25.84	MBA	52
36	**Randall L. Stephenson** AT&T	25.60	Master's	59
37	**Chris E. Kubasik** L3 Technologies	25.51	Bachelor's	48
38	**Jay Bray** Mr. Cooper Group	25.12	Bachelor's	52
39	**James M. Cracchiolo** Ameriprise Financial	24.82	MBA	61
40	**Michael L. Corbat** Citigroup	24.18	Bachelor's	59
41	**Sheldon G. Adelson** Las Vegas Sands	24.01	None	86
42	**John D. Wren** Omnicom Group	23.95	MBA	67
43	**Alan B. Miller** Universal Health	23.55	MBA	82
44	**Ronald N. Tutor** Tutor Perini	23.49	Bachelor's	78
45	**Kevin Stein** TransDigm Group	23.47	PhD	52

(*continued*)

Rank	CEO (Company)	Total Compensation ($ millions)	Degree	Age
46	**John Visentin** Xerox	23.46	Bachelor's	54
47	**Dennis A. Muilenburg** Boeing	23.39	Master's	54
48	**Miles D. White** Abbott Laboratories	22.87	MBA	63
49	**Robert D. Lawler** Chesapeake Energy	22.75	MBA	51
50	**Barry L. Cottle** Scientific Games	22.61	MBA	57

Data Set: CEO50

For this study, the variable of interest, highest college degree obtained, is qualitative in nature. Qualitative data are nonnumerical; thus, the value of a qualitative variable can be classified only into categories called *classes*. The possible degree types—bachelor's, MBA, master's, law, PhD, or none—represent the classes for this qualitative variable. We can summarize such data numerically in two ways: (1) by computing the *class frequency*—the number of observations in the data set that fall into each class or (2) by computing the *class relative frequency*—the proportion of the total number of observations falling into each class.

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

The **class frequency** is the number of observations in the data set falling into a particular class.

The **class relative frequency** is the class frequency divided by the total number of observations in the data set; that is,

$$\text{class relative frequency} = \frac{\text{class frequency}}{n}$$

Examining Table 2.1, we observe that 2 of the 50 highest-paid CEOs did not obtain a college degree, 19 obtained bachelor's degrees, 20 MBAs, 5 master's degrees, 2 PhDs, and 2 law degrees. These numbers—2, 19, 20, 5, 2, and 2—represent the class frequencies for the six classes and are shown in the summary table, Figure 2.1, produced using StatCrunch.

The **class percentage** is the class relative frequency multiplied by 100; that is,

$$\text{class percentage} = (\text{class relative frequency}) \times 100$$

Figure 2.1 also gives the relative frequency of each of the five degree classes. We know that we calculate the relative frequency by dividing the class frequency by the total number of observations in the data set. Thus, the relative frequencies for the five degree types are

$$\text{Bachelor's: } \frac{19}{50} = .38$$

$$\text{Law: } \frac{2}{50} = .04$$

$$\text{MBA: } \frac{20}{50} = .40$$

$$\text{Master's: } \frac{5}{50} = .10$$

$$\text{None: } \frac{2}{50} = .04$$

$$\text{Ph.D.: } \frac{2}{50} = .04$$

Options

Frequency table results for Degree:
Count = 50

Degree	Frequency	Relative Frequency
Bachelor's	19	0.38
Law	2	0.04
MBA	20	0.4
Master's	5	0.1
None	2	0.04
PhD	2	0.04

Figure 2.1
StatCrunch Summary Table for Degrees of 50 Highest Paid CEOs

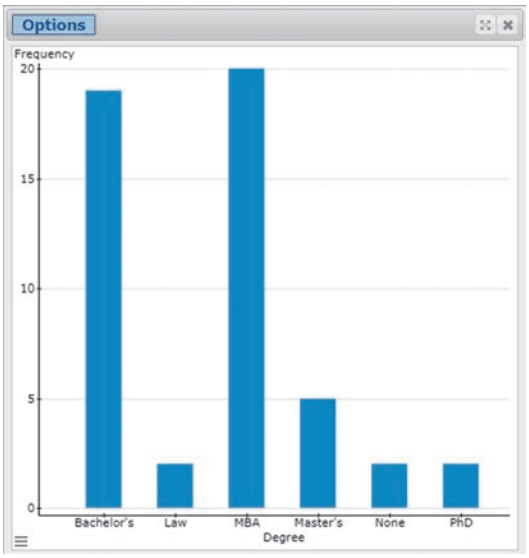


Figure 2.2
StatCrunch Bar Graph for Degrees of 50 Highest Paid CEOs

BIOGRAPHY

VILFREDO PARETO (1848–1923)
The Pareto Principle

Born in Paris to an Italian aristocratic family, Vilfredo Pareto was educated at the University of Turin, where he studied engineering and mathematics. After the death of his parents, Pareto quit his job as an engineer and began writing and lecturing on the evils of the economic policies of the Italian government. While at the University of Lausanne in Switzerland in 1896, he published his first paper, "Cours d'economie politique." In the paper, Pareto derived a complicated mathematical formula to prove that the distribution of income and wealth in society is not random but that a consistent pattern appears throughout history in all societies. Essentially, Pareto showed that approximately 80% of the total wealth in a society lies with only 20% of the families. This famous law about the "vital few and the trivial many" is widely known as the Pareto principle in economics.

These values are shown in the far right column in the StatCrunch summary table, Figure 2.1. If we sum the relative frequencies for Law, MBA, master's, and PhD, we obtain .04 + .40 + .10 + .04 = .58. Therefore, 58% of the 50 highest-paid CEOs obtained at least a master's degree (MBA, master's, law, or PhD).

Although the summary table in Figure 2.1 adequately describes the data in Table 2.1, we often want a graphical presentation as well. Figures 2.2 and 2.3 show two of the most widely used graphical methods for describing qualitative data—**bar graphs** and **pie charts.** Figure 2.2 is a bar graph for "highest degree obtained" produced with StatCrunch. Note that the height of the rectangle, or "bar," over each class is equal to the class frequency. (Optionally, the bar heights can be proportional to class relative frequencies.) In contrast, Figure 2.3 (created using Minitab) shows the relative frequencies (expressed as a percentage) of the six degree types in a *pie chart*. Note that the pie is a circle (spanning 360°), and the size (angle) of the "pie slice" assigned to each class is proportional to the class relative frequency. For example, the slice assigned to the MBA degree is 40% of 360°, or $(.40)(360°) = 144°$.

Before leaving the data set in Table 2.1, consider the bar graph shown in Figure 2.4, produced using Minitab. Note that the bars for the CEO degree categories are arranged in descending order of height from left to right across the horizontal axis—that is, the tallest bar (MBA) is positioned at the far left and the shortest bar is at the far right. This rearrangement of the bars in a bar graph is called a Pareto diagram. One goal of a Pareto

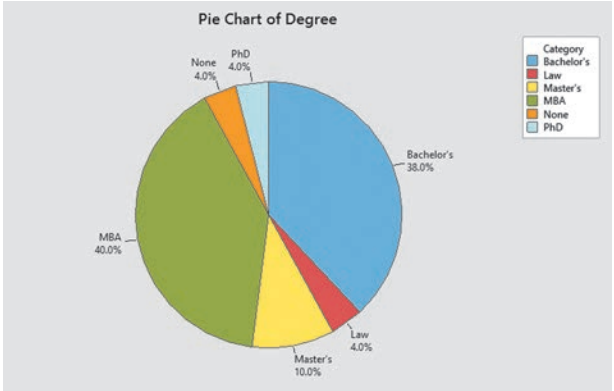


Figure 2.3
Minitab Pie Chart for Degrees of 50 Highest Paid CEOs

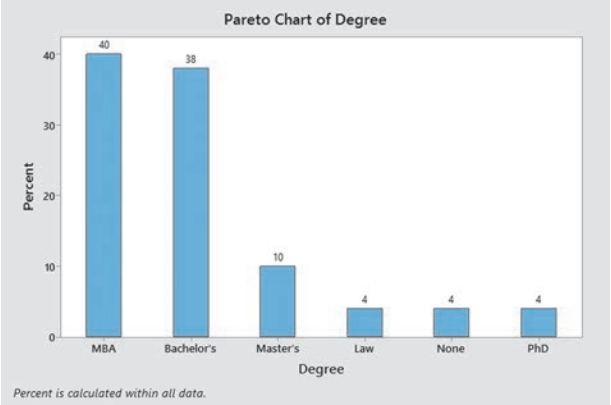


Figure 2.4
Minitab Pareto Diagram for Degrees of 50 Highest Paid CEOs

diagram (named for the Italian economist Vilfredo Pareto) is to make it easy to locate the "most important" categories—those with the largest frequencies. For the 50 highest-paid CEOs in 2018, an MBA degree was most common (40%), and law, PhD, or no degree the least common (4%) highest degree obtained.

Summary of Graphical Descriptive Methods for Qualitative Data

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

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

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

Let's look at a practical example that requires interpretation of the graphical results.

EXAMPLE 2.1

Graphing and Summarizing Qualitative Data—Blood Loss Study

Problem A group of cardiac physicians in southwest Florida have been studying a new drug designed to reduce blood loss in coronary artery bypass operations. Blood loss data for 114 coronary artery bypass patients (some who received a dosage of the drug and others who did not) are saved in the **BLOOD** file. Although the drug shows promise in reducing blood loss, the physicians are concerned about possible side effects and complications. So their data set includes not only the qualitative variable, DRUG, which indicates whether or not the patient received the drug, but also the qualitative variable, COMP, which specifies the type (if any) of complication experienced by the patient. The four values of COMP recorded by the physicians are (1) redo surgery; (2) post-op infection; (3) both; or (4) none.

a. Figure 2.5, generated using XLSTAT, shows summary tables for the two qualitative variables, DRUG and COMP. Interpret the results.

b. Interpret the Minitab output shown in Figure 2.6 and the XLSTAT output shown in Figure 2.7.

Solution

a. The top table in Figure 2.5 is a summary frequency table for DRUG. Note that exactly half (57) of the 114 coronary artery bypass patients received the drug and half did not. The bottom table in Figure 2.5 is a summary frequency table for COMP. We see that about 69% of the 114 patients had no complications, leaving about 31% who experienced either a redo surgery, a post-op infection, or both.

Descriptive statistics (Qualitative data):

Variable\ Statistic	Nbr. of observations	Categories	Frequency per category	Rel. frequency per category (%)
DRUG	114	NO	57.0000	50.0000
		YES	57.0000	50.0000

Descriptive statistics (Qualitative data):

Variable\ Statistic	Nbr. of observations	Categories	Frequency per category	Rel. frequency per category (%)
COMP	114	BOTH	6.0000	5.2632
		INFECT	15.0000	13.1579
		NONE	79.0000	69.2982
		REDO	14.0000	12.2807

Figure 2.5
XLSTAT summary tables for DRUG and COMP

b. Figure 2.6 is a Minitab side-by-side bar graph for the data. The four bars on the left represent the frequencies of COMP for the 57 patients who did not receive the drug; the four bars on the right represent the frequencies of COMP for the 57 patients who did receive a dosage of the drug. The graph clearly shows that patients who did not get the drug suffered fewer complications. The exact percentages are displayed in the XLSTAT summary tables of Figure 2.7. About 56% of the patients who got the drug had no complications, compared with about 83% for the patients who did not get the drug.

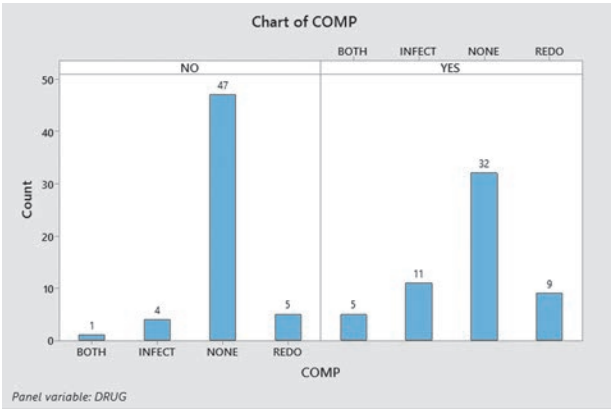


Figure 2.6
Minitab side-by-side bar graphs for COMP, by value of DRUG

Descriptive statistics (Qualitative data):

Variable\Statistic	Nbr. of observations	Categories	Frequency per category	Rel. frequency per category (%)
COMP | NO	57	BOTH	1.0000	1.7544
		INFECT	4.0000	7.0175
		NONE	47.0000	82.4561
		REDO	5.0000	8.7719
COMP | YES	57	BOTH	5.0000	8.7719
		INFECT	11.0000	19.2982
		NONE	32.0000	56.1404
		REDO	9.0000	15.7895

Figure 2.7
XLSTAT summary tables for COMP by value of DRUG

Look Back Although these results show that the drug may be effective in reducing blood loss, Figures 2.6 and 2.7 imply that patients on the drug may have a higher risk of complications. But before using this information to make a decision about the drug, the physicians will need to provide a measure of reliability for the inference—that is, the physicians will want to know whether the difference between the percentages of patients with complications observed in this sample of 114 patients is generalizable to the population of all coronary artery bypass patients.

Now Work Exercise 2.12

STATISTIC IN ACTION REVISITED

Interpreting Pie Charts and Bar Graphs

In the *Journal of Experimental Social Psychology* (Vol. 45, 2009) study on whether money can buy love, Stanford University researchers measured several qualitative (categorical) variables for each of 237 adults: *Gender* (male or female), *Role* (gift-giver or gift-recipient), *Feeling of appreciation for the gift* (measured on an ordinal 7-point scale), and *Feeling of gratefulness for the gift* (measured on an ordinal 7-point scale). We classify the last two variables listed as qualitative, since the numerical values represent distinct response categories (e.g., 1 = "Not at all," 2 = "A little," 3 = "More than a little," 4 = "Somewhat," 5 = "Moderately so," 6 = "Very much," and 7 = "To a great extent") that portray an opinion on how one feels about giving or receiving a gift. Pie charts and bar graphs can be used to summarize and describe the responses for these variables. Recall that the data are saved in the **BUYLOV** file. We used XLSTAT and Minitab to create pie charts for two of these variables: Role (Figure SIA2.1) and Feeling of Gratefulness (Figure SIA2.2).

First, notice in Figure SIA2.1 that of the 237 adults, 56.5% were gift-givers and 43.5% were gift-recipients. Next, from Figure SIA2.2 you can see that 23.2% of adults responded, "Not at all" to the Feeling of gratefulness question, compared with 3% who responded, "To a great extent."

Of interest in the study is whether gift-givers and gift-recipients would respond differently to the Feeling of gratefulness question. We can gain insight into this question by forming bar graphs of the Feeling of gratefulness responses, one graph for gift-givers and one for gift-recipients. These bar graphs are shown in Figure SIA2.3. You can see that about 32% of the gift-givers responded, "Not at all" (top graph) as compared with about 12% of the gift-recipients (bottom graph). Similarly, 6.7% of the gift-givers

Descriptive statistics (Qualitative data):

Variable\ Statistic	Nbr. of observations	Categories	Frequency per category	Rel. frequency per category (%)
Role	237	Giver	134.0000	56.5401
		Receiver	103.0000	43.4599

Pie charts:

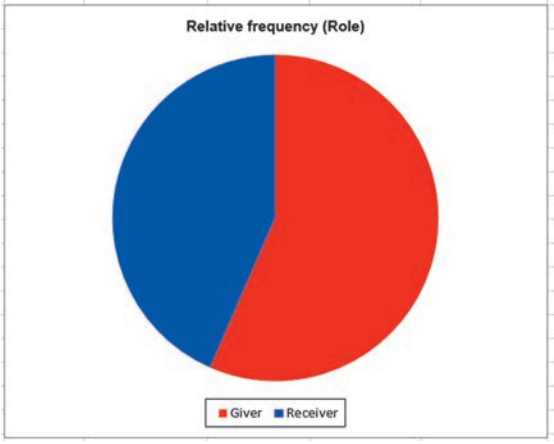


Figure SIA2.1
XLSTAT pie chart for role

responded, "Somewhat" as compared with about 23% of the gift-recipients, and 1.5% of the gift-givers responded, "To a great extent" as compared with 5% of the gift-recipients. Thus, it does appear that gift-givers and gift-recipients respond differently, with gift-recipients more likely to express a greater level of gratefulness for the gift than what gift-givers perceive.

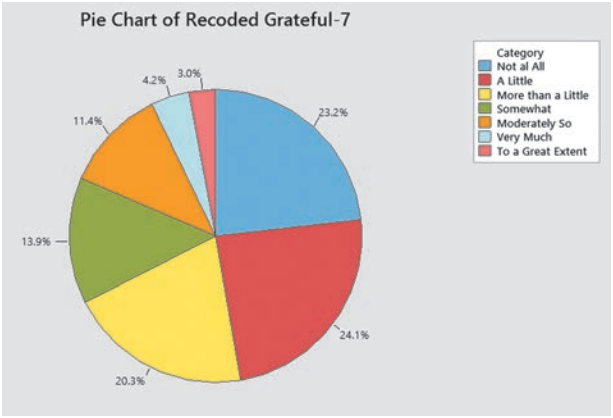
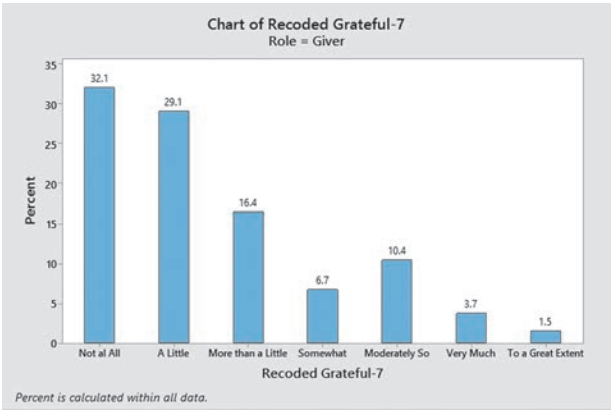


Figure SIA2.2
Minitab pie chart for feeling of gratefulness



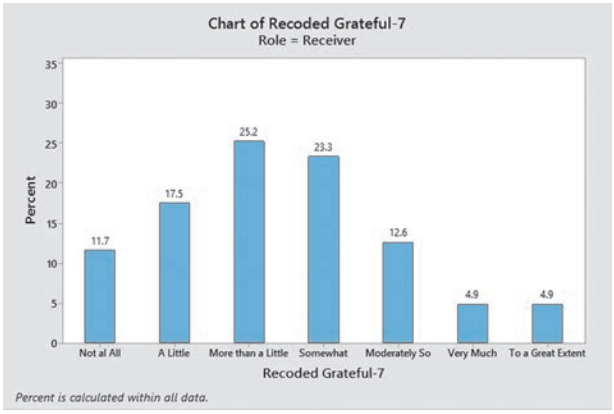


Figure SIA2.3
Minitab bar graphs for feeling of gratefulness by role

Caution: The information produced in these graphs should be limited to describing the sample of 237 adults who participated in the study. If one is interested in making inferences about the population of all gift-givers and gift-recipients (as were the Stanford University researchers), inferential statistical methods need to be applied to the data. These methods are the topics of later chapters.

Data Set: BUYLOVE

Exercises 2.1–2.17

Learning the Mechanics

2.1 Complete the following table on customer statistics.

Age of Customer	Frequency	Relative Frequency
15 or younger	36	—
16 to 25	96	—
25 to 35	48	—
36 to 50	—	0.2
older than 50	12	—
Total	240	1.00

2.2 A qualitative variable is measured for 20 company companies randomly sampled and the data are classified into three classes, small (S), medium (M), and large (L), based on the number of employees in each company. The data (observed class for each company) are listed below.

S	S	L	M	S	M	M	S	M	S
L	M	S	S	S	S	M	L	S	L

a. Compute the frequency for each of the three classes.
b. Compute the relative frequency for each of the three classes.
c. Display the results, part **a,** in a frequency bar graph.
d. Display the results, part **b,** in a pie chart.

Applying the Concepts—Basic

2.3 **STEM programs and jobs**. Do high school programs that emphasize science, technology, engineering and mathematics (STEM) lead to jobs? This was a question of interest in the second *Longitudinal Study of Young People in England* (February 2019). The study included approximately 10,000 students of high school age. Each student was asked which one of seven different areas of study would most likely lead to a job in the future. The percentages (rounded) of males and females who selected each area of study are shown in the accompanying table.

STEM1

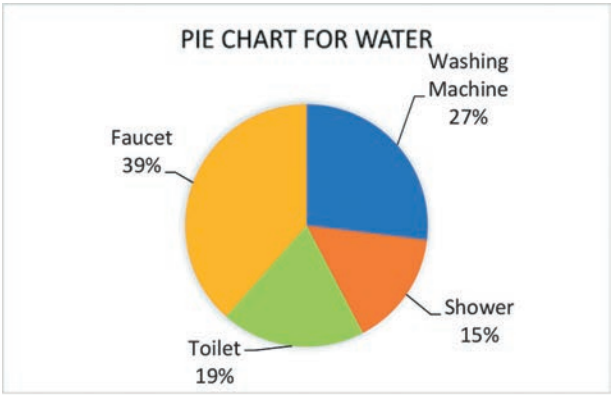
Area of Study	Males	Females
Science	16%	21%
Mathematics	36	23
Technology	17	7
English	20	32
Foreign Languages	2	3
Arts	4	9
Humanities	5	5
	100%	100%

Adapted from: "Attitudes toward STEM subjects by Gender at KS4", *Longitudinal Study of Young People in England 2*, Research Brief, February 2019 (Figure 3).

a. Construct a bar graph that describes the male students' opinions.
b. Construct a bar graph that describes the female students' opinions.
c. Identify the subject areas on which the male and female students have similar opinions.
d. Identify the subject areas on which the male and female students have the greatest differences in opinions.

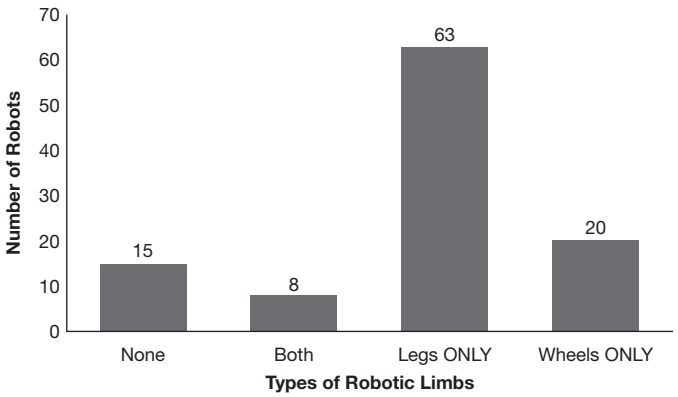
2.4 **Water consumption.** A recent report, Saving Water in Your Home, published by the City Government of Portland established the daily water usage of an average household. Assume their data suggests that in a day a household uses about 25 gallons of water for flushing the toilet, 20 gallons for showering, 50 gallons through faucets, and 35 gallons for running the washing machine. The results are shown in the following Excel pie chart.

WATER



a. According to the pie chart, what is the proportion of water that is used per day for showering? Verify the accuracy of this proportion using the information provided in the question text.
b. If the City Government of Portland is no longer interested in analyzing the water used for the washing machine, the new pie chart should only have three slices. Reconstruct the pie chart to reflect this and interpret the results.

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



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

2.6 **Global credit cards.** There are currently six global credit card companies that allow worldwide usage–Visa, Mastercard, American Express, UnionPay, JCB, and Discover. The table gives a breakdown on the number of purchase transactions for each company in 2018.

GLOBALCC

Credit Card	Number of Worldwide Transactions (billions)
Visa	147.9
Mastercard	75.8
American Express	7.5
UnionPay	58.6
JCB	3.4
Discover	2.5

Source: The Nilson Report, January 2, 2020

a. One of the worldwide credit card transactions in 2018 is selected at random, and the credit card company is determined. What type of data (quantitative or qualitative) is measured?
b. For each credit card company in the table, calculate the percentage of the 2018 worldwide transactions.
c. Use the percentages from part **b** to construct a relative frequency bar graph for the data summarized in the table.
d. Based on the bar graph, make a statement about the most used credit cards worldwide.

2.7 **Microsoft program security issues.** To help its users combat malicious attacks (e.g., worms, viruses) on its computer software, Microsoft periodically issues a security bulletin that reports the software affected by the vulnerability. In *Computers & Security* (July 2013), researchers focused on reported security issues with three Microsoft products: Office, Windows, and Explorer

SECURITY

a. In a sample of 50 security bulletins issued in a recent year, 32 reported a security issue with Windows, 6 with Explorer, and 12 with Office. Construct a pie chart to describe the Microsoft products with security issues. Which product had the lowest proportion of security issues?
b. The researchers also categorized the security bulletins according to the expected repercussion of the vulnerability. Categories were Denial of service, Information disclosure, Remote code execution, Spoofing, and Privilege elevation. Suppose that of the 50 bulletins sampled, the following numbers of bulletins were classified into each respective category: 6, 8, 22, 3, 11. Construct a Pareto diagram to describe the expected repercussions from security issues. Based on the graph, what repercussion would you advise Microsoft to focus on?

Applying the Concepts—Intermediate

2.8 **Car comparison.** Refer to Exercise 1.18 (p. 45). Jim has a budget of RM80,000 to RM100,00 for purchasing multiple cars. (Note: The Malaysian Ringgit, RM, is the official currency of Malaysia.) From *Carbase.my*, he has retrieved information about the car brands and number of cars that he can afford and listed them in the following table.
a. Construct a Pareto diagram for the data. Interpret the results.
b. Construct a pie chart for the top three brands that Jim can afford. Interpret the results.

Brand with Car Comparison and Production Year	Number of Cars That Jim Can Afford
Chevrolet Sonic Sedan LTZ 1.4 (2014)	1
Nissan Almera Turbo 1.0 (2020)	3
Toyota Vios 1.5 (2020)	4
Volkswagen Vento 1.6 Comfortline (2020)	1
Kia Cerato 1.6 (2017)	1
Honda City 1.5 (2020)	3

2.9 **The *Apprentice* contestants' performance ratings.** *The Apprentice* was a TV show that gave the winning contestant the opportunity to work with a famous successful business leader (e.g., before-he-was- President Donald Trump in the United States and Lord Alan Sugar in the United Kingdom). A study was conducted to investigate what separates the successful candidates from the losers (*Significance*, April 2015). Data were collected for 159 contestants on the United Kingdom's version of *The Apprentice* over a 10-year period. (For the first 6 years in the study, the prize was a 100,000 pounds per year job with Lord Sugar; in the next 4 years, the prize was a business partnership with Sugar.) Each contestant was rated (on a 20-point scale) based on their performance. In addition, the higher education degree earned by each contestant—no degree, first (bachelor's) degree, or postgraduate degree—was recorded. These data (simulated, based on statistics reported in the article) are saved in the **APPREN** file. Use statistical software to construct a graph that describes the highest degree obtained by the 159 contestants. Interpret the results.

APPREN

2.10 **The economic return to earning an MBA.** Refer to the *International Economic Review* (August 2008) study on the economic rewards to obtaining an MBA degree, Exercise 1.27 (p. 46). Job status information was collected for a sample of 3,244 individuals who sat for the GMAT in each of four time periods (waves). Summary information (number of individuals) for Wave 1 (at the time of taking the GMAT) and Wave 4 (7 years later) is provided in the accompanying table. Use a graph to compare and contrast the job status distributions of GMAT takers in Wave 1 and Wave 4.

MBAJOB

Job Status	Wave 1	Wave 4
Working, No MBA	2,657	1,787
Working, Have MBA	0	1,372
Not Working, Business School	0	7
Not Working, Other Graduate School	36	78
Not Working, 4-Year Institution	551	0
Total	3,244	3,244

Source: Data from P. Arcidiancono, P. Cooley, and A. Hussey, "The Economic Returns to an MBA," *International Economic Review*, Vol. 49, No. 3, August 2008, Table 1.

2.11 **Profiling UK rental malls.** An analysis of the retail rental levels of tenants of United Kingdom regional shopping malls was published in *Urban Studies* (June 2011). One aspect of the study involved describing the type of tenant typically found at a UK shopping mall. Data were collected for 148 Shopping malls, which housed 1,821 stores. Tenants were categorized into five different-size groups based on amount of floor space: *anchor tenants* (more than 30,000 square feet), *major space users* (between 10,000 and 30,000 sq. ft.), *large standard tenants* (between 4,000 and 10,000 sq. ft.), *small standard tenants* (between 1,500 and 4,000 sq. ft.), and *small tenants* (less than 1,500 sq. ft.). The number of stores in each tenant category was reported as 14, 61, 216, 711, and 819, respectively. Use this information to construct a Pareto diagram for the distribution of tenant groups at UK shopping malls. Interpret the graph.

MALL

2.12 **Ambient scents effect food choices.** Managers of retail stores, hotels, and supermarkets often use ambient scents to increase sales and influence buyers' choices. The *Journal of Marketing Research* (February 2019) published a study that investigated whether food-related ambient scents had an effect on consumers' product choices. At a supermarket, the researchers compared two ambient scents–chocolate chip cookie (the indulgent scent) and strawberry (the non-indulgent scent). Each scent was run for one hour. At the end of the hour, all purchases at the store were classified as either healthy (e.g., fish and fruit), unhealthy (e.g., candy and ice cream), or neutral/nonfood. The number of products purchased in each category for both the indulgent and non-indulgent scents are summarized in the next table. Construct side-by-side bar charts to compare the percentages of food choices for the indulgent and non-indulgent scents. According to the researchers, "exposure to an indulgent (vs. non-indulgent) ambient scent leads to lower (higher) degree of unhealthy (healthy) food purchases." Do you agree? Explain.

AMBIENT

Scent	Total Number Purchased	Unhealthy	Healthy	Neutral
Indulgent (Cookie)	292	86 (29.5%)	114 (39.1%)	92 (31.3%)
Nonindulgent (Strawberry)	527	240 (45.4%)	135 (25.7%)	152 (28.9%)

Source: Biswas, D. & Szocs, C. "The Smell of Healthy Choices: Cross-Modal Sensory Compensation Effects of Ambient Scent on Food Purchases", *Journal of Marketing Research,* Vol. 56, No. 1, February 2019 (adapted from Table 3.B).

2.13 **Educational.** On September 17, 2020, the Department of Statistics Malaysia published the *Salaries & Wages Survey Report, Malaysia, 2019*. The survey's main objective is to collect information on monthly salaries and wages from employees (Malaysian citizens only). The following table shows the average salaries and wages by educational attainment.

EDULVL

Education level	2018	2019
No formal education	RM1,481	RM1,608
Primary level	RM1,821	RM1,929
Secondary level	RM2,215	RM2,372
Tertiary level	RM4,553	RM4,643

a. Use a pie chart to describe the average monthly salaries and wages by educational attainment for the year 2018.
b. Use a pie chart to describe the average monthly salaries and wages by educational attainment for the year 2019.
c. Compare the two pie charts. What conclusion can you draw?

Applying the Concepts—Advanced

2.14 **Museum management.** What criteria do museums use to evaluate their performance? In a worldwide survey reported in *Museum Management and Curatorship* (June 2010), managers of 30 leading museums of contemporary art were asked to provide the performance measure used most often. A summary of the results is provided in the table. The researcher concluded that "there is a large amount of variation within the museum community with regard to . . . performance measurement and evaluation," Do you agree? Use a graph to support your conclusion.

MMC

Performance Measure	Number of Museums
Total visitors	8
Paying visitors	5
Big shows	6
Funds raised	7
Members	4

2.15 **Projected undernourished individuals.** In 2020, The FAO, IFAD, UNICEF, WFP, and WHO published a brief report on the state of food security and nutrition in the world. There were 687.8 million undernourished people

NOURISH

in the world in 2019. One of the sections in the report shows the distribution of these people as shown in the following table:

Region	Percentage of undernourished people (%)
Africa	36.4
Asia	55.4
Northern America and Europe	0.9
Oceania	0.4
Latin America and the Caribbean	6.9

a. Describe the variable measured in the table.
b. These results were displayed in the form of a pie chart. Reconstruct the pie chart from the information given in the table.
c. How many of the 687.8 million undernourished people in the world in 2019 were in Africa?
d. As Africa has many third world countries, people might expect Africa to have the highest number of undernourished people in the world. Does the information in the pie chart, part **b,** support this contention? Explain why or why not.

2.16 Motivation and right-oriented bias. Evolutionary theory suggests that motivated decision makers tend to exhibit a right-oriented bias. (For example, if presented with two equally valued brands of detergent on a supermarket shelf, consumers are more likely to choose the brand on the right.) In *Psychological Science* (November 2011), researchers tested this theory using data on all penalty shots attempted in World Cup soccer matches (a total of 204 penalty shots). The researchers believed that goalkeepers, motivated to make a penalty-shot save but with little time to make a decision, would tend to dive to the right. The results of the study (percentages of dives to the left, middle, or right) are provided in the table. Note that the percentages in each row, corresponding to a certain match situation, add to 100%. Use graphs to illustrate the distribution of dives for the three match situations. What inferences can you draw from the graphs?

Match Situation	Dive Left	Stay Middle	Dive Right
Team behind	29%	0%	71%
Tied	48%	3%	49%
Team ahead	51%	1%	48%

Source: Based on M. Roskes et al., "The Right Side? Under Time Pressure, Approach Motivation Leads to Right-Oriented Bias," *Psychological Science*, Vol. 22, No. 11, November 2011 (adapted from Figure 2).

2.17 Expat teacher's income. A study in the journal *Data in Brief* (Vol. 31, 2020) investigated the decisions of expat teachers in Southeast Asia regarding whether to leave due to the COVID-19 pandemic. The study's dataset includes 307 responses of expat teachers who were working in the Philippines, Indonesia, Singapore, Thailand, and Vietnam during the pandemic. The dataset also included the participants' income before, during, and after COVID-19 and the number of participants in different income ranges. Partial data for income ranges from $3500 to $3999 and more than $4000 for the three different periods are shown in the following table.

COVIDINC

$3500 to $3999			More than $4000		
Before	During	After	Before	During	After
Before	During	After	Before	During	After
Before	During	After	Before	During	After
Before	During	After	Before	During	
Before	During	After	Before	During	
Before	During	After	Before	After	
Before	During	After	Before	After	
Before	During	After	Before	After	
Before	During	After	Before	After	
Before	During	After	Before	After	
Before	During	After	Before	After	
Before	During	After	Before	After	
Before	During	After	Before	After	
Before	During	After	During	After	
Before	After	After	During	After	
Before	After	After	During	After	
Before	After	After	During	After	
Before	After		During	After	

Source: Based on Hoanga A. D., et al., "Dataset of the intention of ex-pat teachers in Sotheast Asia to leave due to the COVID-19 pandemic," *Journal Data in Brief*, Vol. 39, 105913.

a. Use graphical methods to describe each of the three different periods for the two different income ranges.
b. Use side-by-side bar charts to compare the proportions of the two different income ranges for the three different periods.
c. Use side-by-side bar charts to compare the proportions of the three different periods for two different income ranges.
d. What inferences can be made from the bar charts, parts **a–c?**

2.2 Graphical Methods for Describing Quantitative Data

Recall from Section 1.5 that quantitative data sets consist of data that are recorded on a meaningful numerical scale. For describing, summarizing, and detecting patterns in such data, we can use three graphical methods: **dot plots**, **stem-and-leaf displays**, and **histograms**. Because almost all statistical software packages can produce these graphs, we'll focus here on their interpretations rather than their construction.

Table 2.2 Percentage of Revenues Spent on Research and Development

Company	Percentage	Company	Percentage	Company	Percentage	Company	Percentage
1	13.5	14	9.5	27	8.2	39	6.5
2	8.4	15	8.1	28	6.9	40	7.5
3	10.5	16	13.5	29	7.2	41	7.1
4	9.0	17	9.9	30	8.2	42	13.2
5	9.2	18	6.9	31	9.6	43	7.7
6	9.7	19	7.5	32	7.2	44	5.9
7	6.6	20	11.1	33	8.8	45	5.2
8	10.6	21	8.2	34	11.3	46	5.6
9	10.1	22	8.0	35	8.5	47	11.7
10	7.1	23	7.7	36	9.4	48	6.0
11	8.0	24	7.4	37	10.5	49	7.8
12	7.9	25	6.5	38	6.9	50	6.5
13	6.8	26	9.5				

Data Set: R&D

For example, suppose a financial analyst is interested in the amount of resources spent by computer hardware and software companies on research and development (R&D). She samples 50 of these high-technology firms and calculates the amount each spent last year on R&D as a percentage of their total revenue. The results are given in Table 2.2. As numerical measurements made on the sample of 50 units (the firms), these percentages represent quantitative data. The analyst's initial objective is to summarize and describe these data in order to extract relevant information.

A visual inspection of the data indicates some obvious facts. For example, the smallest R&D percentage is 5.2% (company 45) and the largest is 13.5% (companies 1 and 16). But it is difficult to provide much additional information on the 50 R&D percentages without resorting to some method of summarizing the data. One such method is a dot plot.

Dot Plots

A **dot plot** for the 50 R&D percentages, produced using Minitab, is shown in Figure 2.8. The horizontal axis of Figure 2.8 is a scale for the quantitative variable, percent. The numerical value of each measurement in the data set is located on the horizontal scale by a dot. When data values repeat, the dots are placed above one another, forming a pile at that particular numerical location. As you can see, this dot plot shows that 45 of the 50 R&D percentages (90%) are between 6% and 12%, with most falling between 7% and 9%.

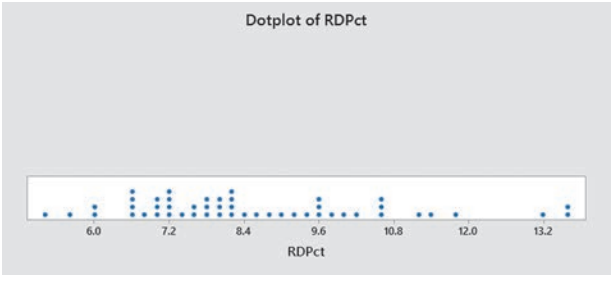


Figure 2.8
Minitab dot plot for 50 R&D percentages

Stem-and-Leaf Display

We used XLSTAT to generate another graphical representation of these same data, a **stem-and-leaf display,** in Figure 2.9. In this display the *stem* is the portion of the measurement (percentage) to the left of the decimal point, while the remaining portion to the right of the decimal point is the *leaf.*

The stems for the data set are listed in the first column of Figure 2.9 from the smallest (5) to the largest (13). Then the leaf for each observation is recorded in the row of the display corresponding to the observation's stem. For example, the leaf 5 of the first observation (13.5) in Table 2.2 is placed in the row corresponding to the stem 13. Similarly, the leaf 4 for the second observation (8.4) in Table 2.2 is recorded in the row corresponding to the stem 8, while the leaf 5 for the third observation (10.5) is recorded

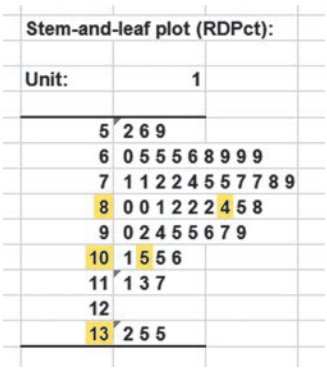


Figure 2.9
XLSTAT stem-and-leaf display for 50 R&D percentages

in the row corresponding to the stem 10. (The leaves for these first three observations are shaded in Figure 2.9.) Typically, the leaves in each row are ordered as shown in Figure 2.9.

The stem-and-leaf display presents another compact picture of the data set. You can see at a glance that most of the sampled computer companies (37 of 50) spent between 6.0% and 9.9% of their revenues on R&D, and 11 of them spent between 7.0% and 7.9%. Relative to the rest of the sampled companies, three spent a high percentage of revenues on R&D—in excess of 13%.

The definitions of the stem and leaf can be modified to alter the graphical display. For example, suppose we had defined the stem as the tens digit for the R&D percentage data, rather than the ones and tens digits. With this definition, the stems and leaves corresponding to the measurements 13.5 and 8.4 would be as follows:

Stem	*Leaf*
1	3

Stem	*Leaf*
0	8

Note that the decimal portion of the numbers has been dropped. Only one digit is displayed in the leaf.

If you look at the data, you'll see why we didn't define the stem this way. All the R&D measurements fall below 13.5, so all the leaves would fall into just two stem rows—1 and 0—in this display. The picture resulting from using only a few stems would not be nearly as informative as Figure 2.9.

Histograms

A Minitab **histogram** for these 50 R&D measurements is displayed in Figure 2.10. The horizontal axis for Figure 2.10, which gives the percentage amounts spent on R&D for each company, is divided into **class intervals** commencing with the interval (5.0–6.0) and proceeding in intervals of equal size to (13.0–14.0). The vertical axis gives the number (or *frequency*) of the 50 measurements that fall in each class interval. You can see that the class interval (7.0–8.0) (i.e., the class with the highest bar) contains the largest frequency of 11 R&D percentage measurements; the remaining class intervals tend to contain a smaller number of measurements as R&D percentage gets smaller or larger.

Histograms can be used to display either the *frequency* or *relative frequency* of the measurements falling into the class intervals. The class intervals, frequencies, and relative frequencies for the 50 R&D measurements are shown in Table 2.3.* By summing

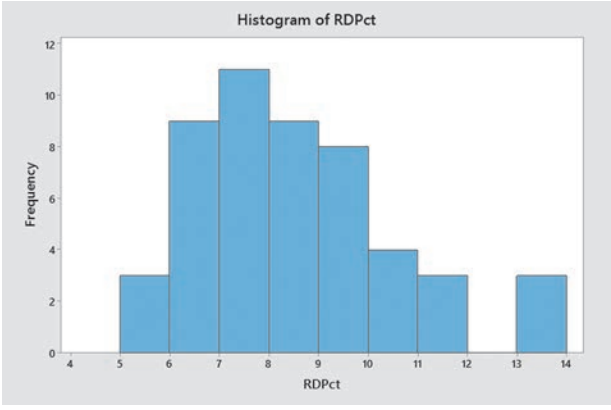


Figure 2.10
Minitab histogram for 50 R&D percentages

*Minitab, like many statistical software packages, will classify an observation that falls on the borderline of a class interval into the next highest class interval. For example, the R&D measurement of 8.0, which falls on the border between the intervals (7.0–8.0) and (8.0–9.0), is classified into the (8.0–9.0) interval. The frequencies in Table 2.3 reflect this convention.

BIOGRAPHY

JOHN TUKEY (1915–2000)
The Picasso of Statistics

Like the legendary artist Pablo Picasso, who mastered and revolutionized a variety of art forms during his lifetime, John Tukey is recognized for his contributions to many subfields of statistics. Born in Massachusetts, Tukey was home-schooled, graduated with his bachelor's and master's degrees in chemistry from Brown University, and received his PhD in mathematics from Princeton University. While at Bell Telephone Laboratories in the 1960s and early 1970s, Tukey developed "exploratory data analysis," a set of graphical descriptive methods for summarizing and presenting huge amounts of data. Many of these tools, including the stem-and-leaf display and the box plot (see Section 2.7), are now standard features of modern statistical software packages. (In fact, it was Tukey himself who coined the term *software* for computer programs.)

Table 2.3 Class Intervals, Frequencies, and Relative Frequencies for the 50 R&D Measurements

Class	Class Interval	Class Frequency	Class Relative Frequency
1	5.0–6.0	3	3/50 = .06
2	6.0–7.0	9	9/50 = .18
3	7.0–8.0	11	11/50 = .22
4	8.0–9.0	9	9/50 = .18
5	9.0–10.0	8	8/50 = .16
6	10.0–11.0	4	4/50 = .08
7	11.0–12.0	3	3/50 = .06
8	12.0–13.0	0	0/50 = .00
9	13.0–14.0	3	3/50 = .06
Totals		50	1.00

the relative frequencies in the intervals (6.0–7.0), (7.0–8.0), (8.0–9.0), (9.0–10.0), and (10.0–11.0), we find that .18 + .22 + .18 + .16 + .08 = .82, or 82%, of the R&D measurements are between 6.0 and 11.0. Similarly, summing the relative frequencies in the last two intervals, (12.0–13.0) and (13.0–14.0), we find that 6% of the companies spent over 12.0% of their revenues on R&D. Many other summary statements can be made by further study of the histogram. Note that the sum of all class frequencies will always equal the sample size n.

When interpreting a histogram, consider two important facts. First, the proportion of the total area under the histogram that falls above a particular interval of the horizontal axis is equal to the relative frequency of measurements falling in the interval. For example, the relative frequency for the class interval 7.0–8.0 in Figure 2.10 is .22. Consequently, the rectangle above the interval contains 22% of the total area under the histogram.

Second, imagine the appearance of the relative frequency histogram for a very large set of data (say, a population). As the number of measurements in a data set is increased, you can obtain a better description of the data by decreasing the width of the class intervals. When the class intervals become small enough, a relative frequency histogram will (for all practical purposes) appear as a smooth curve (see Figure 2.11). Some recommendations for selecting the number of intervals in a histogram for smaller data sets are given in the box below Figure 2.11.

While histograms provide good visual descriptions of data sets—particularly very large ones—they do not let us identify individual measurements. In contrast, each of the original measurements is visible to some extent in a dot plot and clearly visible in a stem-and-leaf display. The stem-and-leaf display arranges the data in ascending order, so it's easy to locate the individual measurements. For example, in Figure 2.9, we can easily see that three of the R&D measurements are equal to 8.2, but we can't see that fact by inspecting the histogram in Figure 2.10. However, stem-and-leaf displays can become unwieldy for very large data sets. A very large number of stems and leaves causes the vertical and horizontal dimensions of the display to become cumbersome, diminishing the usefulness of the visual display.

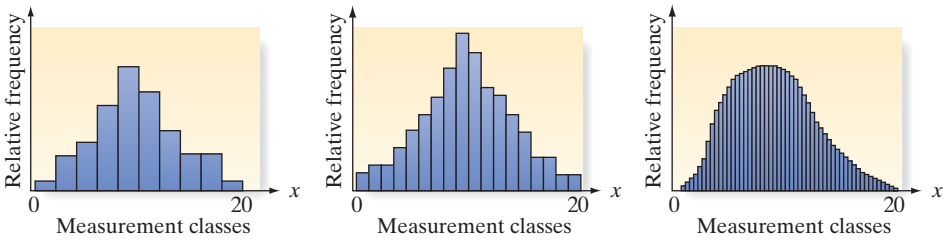


Figure 2.11
Effect of the size of a data set on the outline of a histogram

Determining the Number of Classes in a Histogram

Number of Observations in Data Set	Number of Classes
Less than 25	5–6
25–50	7–14
More than 50	15–20

EXAMPLE 2.2

Graphs for a Quantitative Variable—Lost Price Quotes

Problem A manufacturer of industrial wheels suspects that profitable orders are being lost because of the long time the firm takes to develop price quotes for potential customers. To investigate this possibility, 50 requests for price quotes were randomly selected from the set of all quotes made last year, and the processing time was determined for each quote. The processing times are displayed in Table 2.4, and each quote was classified according to whether the order was "lost" or not (i.e., whether or not the customer placed an order after receiving a price quote).

a. Use a statistical software package to create a frequency histogram for these data. Then shade the area under the histogram that corresponds to lost orders. Interpret the result.

b. Use a statistical software package to create a stem-and-leaf display for these data. Then shade each leaf of the display that corresponds to a lost order. Interpret the result.

Solution

a. We used Minitab to generate the frequency histogram in Figure 2.12. Note that 20 classes were formed. The class intervals are (1.0–2.0), (2.0–3.0), . . . , (20.0–21.0).

Table 2.4 Price Quote Processing Time (Days)

Request Number	Processing Time	Lost?	Request Number	Processing Time	Lost?
1	2.36	No	26	3.34	No
2	5.73	No	27	6.00	No
3	6.60	No	28	5.92	No
4	10.05	Yes	29	7.28	Yes
5	5.13	No	30	1.25	No
6	1.88	No	31	4.01	No
7	2.52	No	32	7.59	No
8	2.00	No	33	13.42	Yes
9	4.69	No	34	3.24	No
10	1.91	No	35	3.37	No
11	6.75	Yes	36	14.06	Yes
12	3.92	No	37	5.10	No
13	3.46	No	38	6.44	No
14	2.64	No	39	7.76	No
15	3.63	No	40	4.40	No
16	3.44	No	41	5.48	No
17	9.49	Yes	42	7.51	No
18	4.90	No	43	6.18	No
19	7.45	No	44	8.22	Yes
20	20.23	Yes	45	4.37	No
21	3.91	No	46	2.93	No
22	1.70	No	47	9.95	Yes
23	16.29	Yes	48	4.46	No
24	5.52	No	49	14.32	Yes
25	1.44	No	50	9.01	No

Data Set: QUOTES

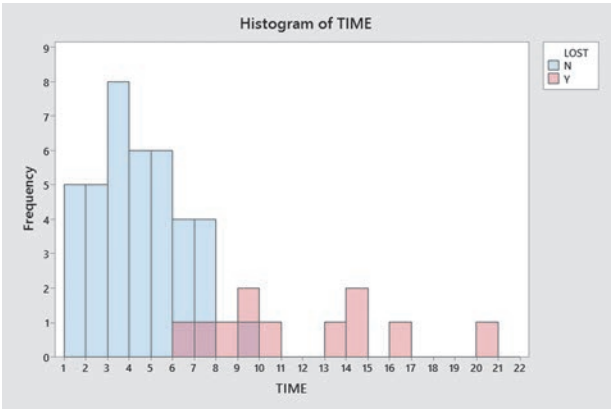


Figure 2.12
Minitab frequency histogram for price quote data

This histogram clearly shows the clustering of the measurements in the lower end of the distribution (between approximately 1 and 8 days), and the relatively few measurements in the upper end of the distribution (greater than 12 days). The shading of the area of the frequency histogram corresponding to lost orders (red bars) clearly indicates that they lie in the upper tail of the distribution.

b. We used StatCrunch to generate the stem-and-leaf display in Figure 2.13. Note that the stem (the first column of the printout) consists of the number of whole days (digits to the left of the decimal). The leaf (the second column of the printout) is the tenths digit (rounded first digit after the decimal) of each measurement. Thus, the leaf 2 in the stem 20 (the last row of the printout) represents the time of 20.23 days. Like the histogram, the stem-and-leaf display shows the shaded "lost" orders in the upper tail of the distribution.

Look Back As is usually the case for data sets that are not too large (say, fewer than 100 measurements), the stem-and-leaf display provides more detail than the histogram without being unwieldy. For instance, the stem-and-leaf display in Figure 2.13 clearly indicates that the lost orders are associated with high processing times (as does the histogram in Figure 2.12), and exactly which of the times correspond to lost orders. Histograms are most useful for displaying very large data sets, when the overall shape of the distribution of measurements is more important than the identification of individual measurements. Nevertheless, the message of both graphical displays is clear: Establishing processing time limits may well result in fewer lost orders.

Now Work Exercise 2.22

Variable: TIME

Decimal point is at the colon.
Leaf unit = 0.1

```
 1 : 34799
 2 : 04569
 3 : 23445699
 4 : 044579
 5 : 115579
 6 : 02468
 7 : 35568
 8 : 2
 9 : 05
10 : 01
11 :
12 :
13 : 4
14 : 13
15 :
16 : 3
17 :
18 :
19 :
20 : 2
```

Figure 2.13
StatCrunch stem-and-leaf display for price quote data

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

Summary of Graphical Descriptive Methods for Quantitative Data

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

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

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

STATISTICS IN ACTION REVISITED

Interpreting Histograms

In the *Journal of Experimental Social Psychology* (Vol. 45, 2009) study on whether money can buy love (p. 39), the researchers randomly assigned participants to the role of either gift-giver or gift-receiver. (Gift-givers, recall, were asked about a birthday gift they recently gave, while gift-recipients were asked about a birthday gift they recently received.) Two quantitative variables were measured for each of the 237 participants: *gift price* (measured in dollars) and *overall level of appreciation for the gift* (measured as the sum of the two 7-point appreciation scales, with higher values indicating a higher level of appreciation). One of the objectives of the research was to investigate whether givers and receivers differ on the price of the gift reported and on the level of appreciation reported. We can explore this phenomenon graphically by forming side-by-side histograms for the quantitative variables, one histogram for gift-givers and one for gift-recipients. These histograms, produced from a Minitab analysis of the data in the **BUYLOV** file, are shown in Figures SIA2.4a and Figures SIA2.4b.

First, examine the histograms for birthday gift price (Figure SIA2.4a). The prices reported by gift-recipients tended to be higher than the prices reported by gift-givers. For example, receivers reported more birthday gift prices of at least $300 than givers, while givers reported more prices of $100 or less than receivers.

Next, examine the histograms for overall level of appreciation (Figure SIA2.4b). For gift-givers, the histogram of appreciation scores is centered around 5 points, while

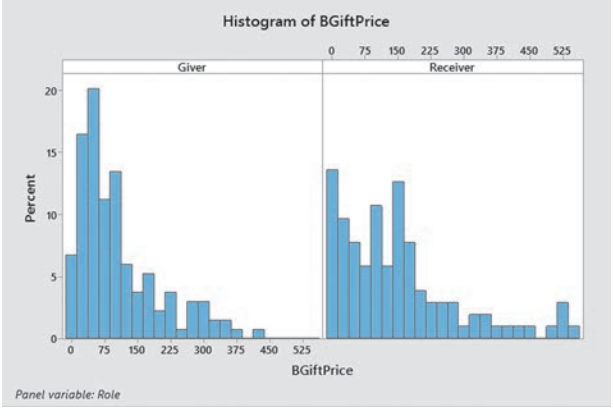


Figure SIA2.4a
Side-by-side histograms for birthday gift price

STATISTICS IN ACTION REVISTED (*continued*)

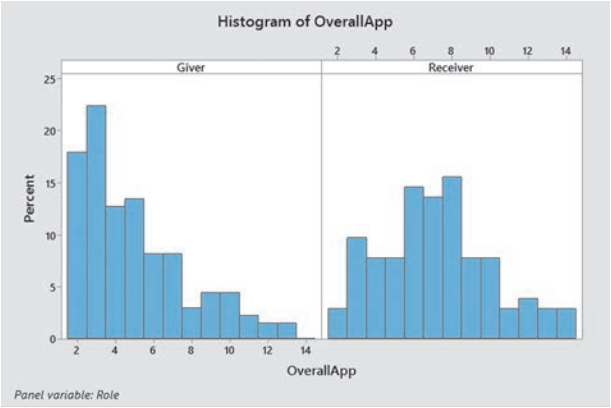


Figure SIA2.4b
Side-by-side histograms for overall level of appreciation

for gift-recipients the histogram is centered higher, at about 8 points. Also from the histograms you can see that about 65% of the givers reported an appreciation level of less than 6 compared with about 28% for gift-recipients. As with the bar graphs in the previous *Statistics in Action Revisited*, it appears that gift-givers and gift-recipients respond differently, with gift-recipients more likely to express a greater level of appreciation for the gift than what gift-givers perceive. In later chapters, we'll learn how to attach a measure of reliability to such an inference.

Data Set: BUYLOV

Exercises 2.18–2.34

Learning the Mechanics

2.18 A company is analyzing the prices at which its items are sold. Graph the relative frequency histogram for the 600 items summarized in the accompanying relative frequency table.

Items Class	Relative Frequency
.50 but less than 1.50	.06
1.50 but less than 2.50	.30
2.50 but less than 3.50	.20
3.50 but less than 4.50	.15
4.50 but less than 5.50	.14
5.50 but less than 6.50	.10
6.50 but less than 7.50	.05

2.19 Refer to Exercise 2.18. Calculate the number of the 600 items falling into each of the item classes. Then graph a frequency histogram for these data.

2.20 Consider the following stem-and-leaf display shown here.

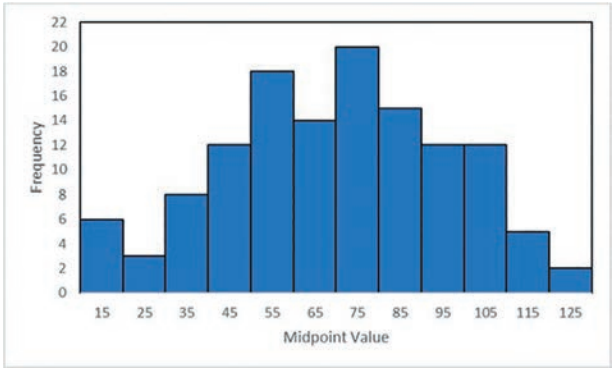
Stem	Leaf
2	9
3	133
4	122
5	
6	00115

a. Identify the numbers in the original data set represented by the stem and its leaves. Assume that the data was not rounded.

b. Construct a dot plot of the original data set.

c. Now assume that the original data might have been rounded off to the nearest whole number. Identify the interval of the number represented in the first row of the stem-and-leaf display.

2.21 Answer the following questions based on the following histogram:



a. Is this a frequency histogram or a relative frequency histogram? Explain.

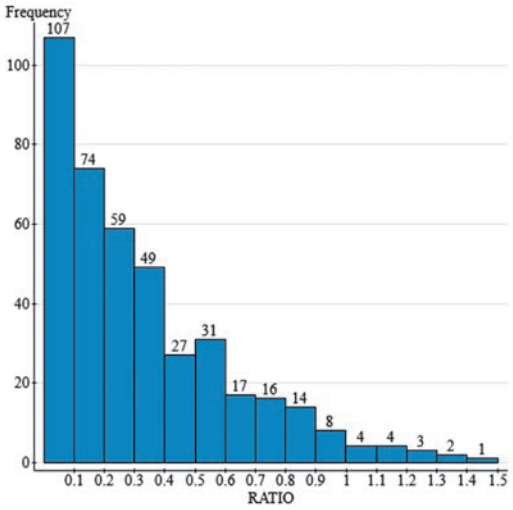
b. How many measurement classes were used in the construction of this histogram?

c. How many measurements are in the data set described by this histogram?

Applying the Concepts—Basic

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

NW FUP



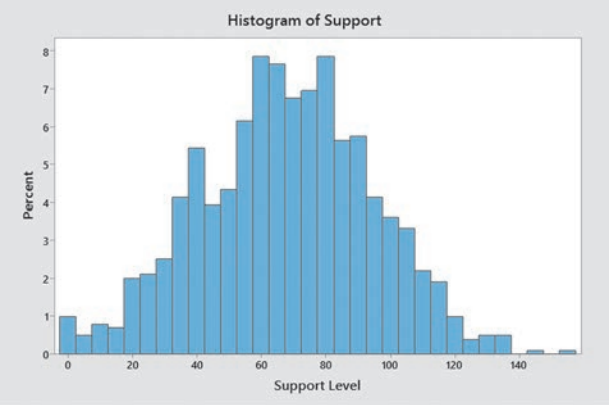
a. What type of graph is displayed?
b. What is the quantitative variable summarized in the graph?
c. Determine the proportion of fup/fumic ratios that fall above 1.
d. Determine the proportion of fup/fumic ratios that fall below .4.

2.23 **Corporate sustainability of CPA firms.** Refer to the *Business and Society* (March 2011) study on the sustainability behaviors of CPA corporations, Exercise 1.28 (p. 46). *Corporate sustainability*, recall, refers to business practices designed around social and environmental considerations. Data on the level of support for corporate sustainability were obtained for 992 senior managers. Level of support was measured quantitatively. Simulation was used to convert the data from the study to a scale ranging from 0 to 160 points, where higher point values indicate a higher level of support for sustainability.

CORSUS

a. A histogram for level of support for sustainability is shown next. What type of histogram is produced, frequency or relative frequency?
b. Use the graph to estimate the percentage of the 992 senior managers who reported a high (100 points or greater) level of support for corporate sustainability.

Minitab histogram for Exercise 2.23



2.24 **Cruise ships sanitation scores.** The Vessel Sanitation Program (VSP) at the Centers for Disease Control and Prevention (CDC) assists the cruise ship industry with preventing and controlling the introduction, transmission, and spread of gastrointestinal (GI) illnesses on cruise ships. The cruise ships are rated on a 100-point scale by the CDC and scores of 85 or lower are considered "not satisfactory." The sanitation scores for 129 cruise ships between January 2019 and October 2020 are saved in the accompanying file. The first four and last five observations in the data set are listed in the following table.

CRUISE

Cruise Ship	Date	Score
Adventure of the Seas	07/12/2019	95
AIDAdiva	11/10/2019	99
AIDAluna	06/10/2019	91
AIDAVITA	16/08/2019	94
.		
.		
.		
Vision of the Seas	01/04/2019	99
Volendam	31/05/2019	98
Westerdam	03/07/2019	98
Zaandam	17/08/2019	99
Zuiderdam	20/12/2019	98

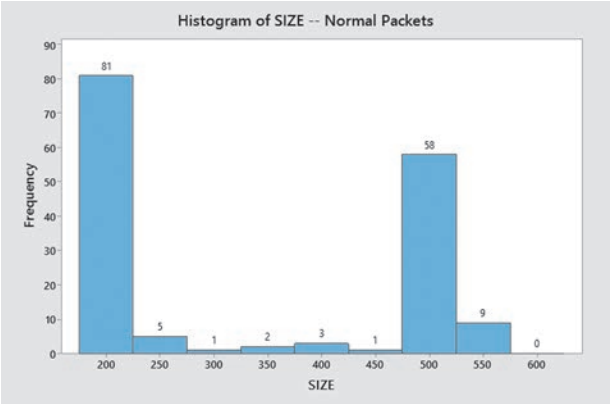
Source: Data from the Centers for Disease Control and Prevention (CDC), 2021. https://www.cdc.gov/nceh/vsp/desc/aboutvsp.htm

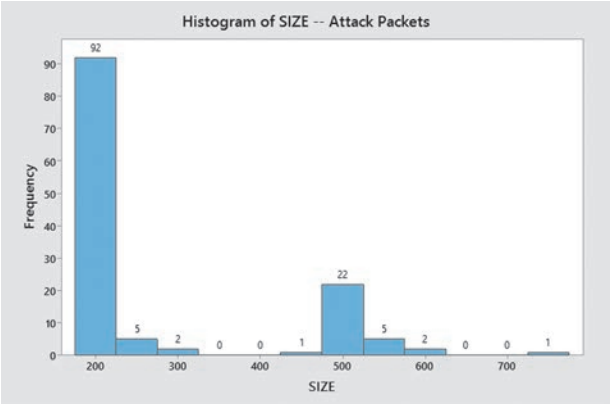
a. Generate both a stem-and-leaf display and a histogram of the data.
b. Use the graphs to estimate the percentage of cruise ships that are "not satisfactory". Which graph did you use?
c. Locate the sanitation score of 92 (Carnival Magic) on the graph. Which graph did you use?

2.25 **Malicious attacks on a cryptocurrency network.** Ethereum is an open-sourced software system initially developed for cryptocurrency. In *Computers & Security* (January 2020), researchers developed a detector of malicious attacks on data traveling through the Ethereum network. Initially, the researchers used code to create a large number of normal (i.e., honest) data packets on the network. The data packets varied in size (number of bytes). Then they created code intended to maliciously attack the data packets. Samples of data packets from both the normal (160 packets) and attack (130 packets) states were collected and the variable,

ATTACK

packet size, was determined for each packet. Minitab histograms for the two states are shown below.





a. Convert each of the frequency histograms into relative frequency histograms using the numbers shown above the bars.

b. Do you detect differences in the pattern of packet sizes for normal and attacked data packets? Explain.

Applying the Concepts—Intermediate

2.26 **Wastewater and Electricity.** A study has been done on the daily electricity consumed and volume of wastewater generated at the Dillman Road wastewater treatment plant for each of the first five days of each month in 2020. The data in the study consists of a total of 60 days' billed electricity consumption in kilowatt-hours (kWh), solar supply in kilowatt-hours (kWh), total supply in kilowatt-hours (kWh), daily peak electricity demand in kilowatts (kW), and volume of wastewater in million gallons (MG).

CONSUMP

a. Use a graph to describe the distribution of billed electricity consumption for the 60 days.

b. Use a graph to describe the distribution of the solar supply for the 60 days.

c. Use a graph to describe the distribution of the total supply for the 60 days.

d. Use a graph to describe the distribution of the daily peak electricity demand.

e. Use a graph to describe the distribution of volume of wastewater for the 60 days.

f. Compare and contrast the graphs, parts **a**–**e**.

RANK	TEAM	Current Value ($mil)	1-Year Value Change (%)	Debt/Val (%)	Revenue ($mil)	Oper. Income ($mil)
1	Dallas Cowboys	5500	10	7	950	420
2	New England Patriots	4100	8	5	600	240
3	New York Giants	3900	18	13	519	142
4	Los Angeles Rams	3800	19	86	401	30
5	San Francisco 49ers	3500	15	10	492	93
6	Chicago Bears	3450	19	3	453	62
7	Washington Redskins	3400	10	7	493	120
8	New York Jets	3200	12	17	635	115
9	Houston Texans	3100	11	1	497	176
10	Philadelphia Eagles	3050	11	6	482	150
11	Denver Broncos	3000	13	10	446	94
12	Oakland Raiders	2900	20	52	357	28
13	Green Bay Packers	2850	9	4	456	39
14	Pittsburgh Steelers	2800	8	7	439	102
15	Seattle Seahawks	2775	8	5	439	106
16	Miami Dolphins	2760	7	16	443	67
17	Atlanta Falcons	2755	6	30	458	97
18	Baltimore Ravens	2750	6	10	438	131
19	Minnesota Vikings	2700	12	21	427	65
20	Indianapolis Colts	2650	11	4	393	104
21	Los Angeles Chargers	2500	10	36	375	72
22	Carolina Panthers	2400	4	8	424	78
23	Jacksonville Jaguars	2325	12	3	424	77
24	Kansas City Chiefs	2300	10	5	410	83
25	New Orleans Saints	2275	10	9	441	126
26	Arizona Cardinals	2250	5	7	400	87
27	Tampa Bay Buccaneers	2200	10	8	400	66
28	Cleveland Browns	2175	12	10	399	32
29	Tennessee Titans	2150	5	7	394	53
30	Cincinnati Bengals	2000	11	5	380	58
31	Detroit Lions	1950	15	13	385	73
32	Buffalo Bills	1900	19	11	386	82

Source: "The Most Valuable Teams in the NFL," *Forbes*, copyright © September 4, 2019.

2.27 SATSCORE **SAT scores.** Educators are constantly evaluating the efficacy of public schools in the education and training of American students. One quantitative assessment of change over time is the difference in scores on the SAT, which has been used for decades by colleges and universities as one criterion for admission. The following table shows the reading and mathematics scores in the first five and the last two states for the years 2019 and 2020. This comprises the 50 states and District of Columbia provided by the SAT.

STATE	Read 2019	Math 2019	Read 2020	Math 2020
Alabama	583	560	576	551
Alaska	556	541	555	543
Arizona	569	565	571	568
Arkansas	582	559	590	567
California	534	531	527	522
⋮				
Wisconsin	635	648	615	628
Wyoming	623	615	614	606

Source: The College Board, 2020.

a. Use graphs to display the SAT mathematics score distributions in 2019 and 2020. How did the distributions of the states' scores change over the one-year period?

b. As another method of comparing the 2019 and 2020 SAT mathematics scores, compute the paired difference by subtracting the 2019 score from the 2020 score for each state. Summarize these differences in a graph. Compare the results with those of part **a**.

c. Based on the graph of part **b,** what is the largest improvement in SAT mathematics score? Identify the state associated with this improvement.

2.28 CAPRATE **Valuation of single-tenant properties**. Refer to *The Appraisal Journal* (Summer, 2019) study of the valuation of single-tenant properties, Exercise 1.16 (p. 44). Recall that the ratio of net operating income to property asset value–called the capitalization rate–was determined for a sample of 13 retail property tenants. Data on S&P credit ratings and 5-year capitalization rates for the 13 tenants are reproduced in the table. Use a graphical method to describe the 5-year capitalization rates. Select a graph that will allow you to determine where the tenants with a low (BBB-) S&P rating fall within the capitalization rate distribution. What do you observe?

Tenant	S&P Credit Rating	5-Year Capitalization Rate (%)
Best Buy	BBB	8.25
BJ's Warehouse	B	7.40
Dollar General	BBB−	8.45
Dollar Tree	BBB−	7.50
Family Dollar	BBB−	8.50
Kohl's	BBB−	8.50
Kroger	BBB	7.25
Lowe's	BBB+	6.70
Sherwin-Williams	BBB	6.75
The Home Depot	A	6.75
United Rentals	BB	8.15
Walmart	AA	6.75
Whole Foods Market	AA−	5.50

Source: Sellers, L.P., et al. "Valuation Methods and Dark Big-Box Theories", *The Appraisal Journal*, Vol. 87, No. 3, Summer 2019 (Exhibit 1).

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

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

Source: Based on A. Permanyer, J. L. R. Gallego, M. A. Caja, and D. Dessort, "Crude Oil Biodegradation and Environmental Factors at the Riutort Oil Shale Mine, SE Pyrenees," *Journal of Petroleum Geology*, Vol. 33, No. 2, April 2010, (Table 1).

2.30 **Gross domestic product.** *The Economy* is an annual publication by the Government of Newfoundland and Labrador (https://www.gov.nl.ca/) on the region's economic performance. From data provided by Statistics Canada Department of Finance, The Economy 2020 reported the estimated Gross Domestic Product (GDP) (in $ millions) in 2018 for the following 12 industries: the services producing sector, wholesale trade, retail trade, transportation and warehousing, finance, insurance, real estate & business support services, professional, scientific, & technical services, educational services, health care & social assistance, information, culture & recreation, accommodation & food services, public administration, and other services. The data is listed in the accompanying table. In this data, GDP is expressed at basic prices, measuring payments made to the owners of factor inputs in production. This differs from GDP at market prices. The difference is attributable to taxes less subsidies on products and imports. Industry components may not sum to total due to independent rounding.

17,167.5	668.4	1,553.5	973.6	4,413.2	1,007.2
1,797.6	2,633.9	702.0	570.7	2,350.0	497.5

Source: https://www.economics.gov.nl.ca/E2018/TheEconomy2018.pdf

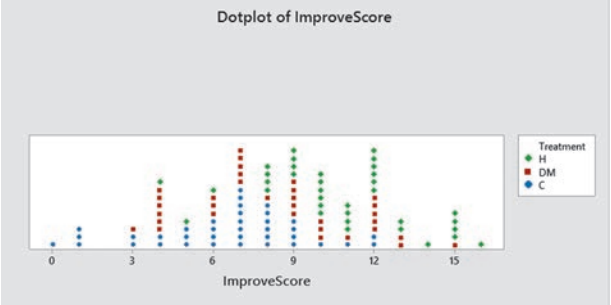
a. Use a relative frequency histogram to show the distribution of the estimated GDP for the 12 industries in 2018.
b. Identify the class interval that includes the largest and the lowest proportions of the estimated GDP.

Applying the Concepts—Advanced

2.31 HCOUGH **Is honey a cough remedy?** Does a teaspoon of honey before bed really calm a child's cough? To test the folk remedy, pediatric researchers carried out a designed study conducted over two nights (*Archives of Pediatrics and Adolescent Medicine*, December 2007). A sample of 105 children who were ill with an upper respiratory tract infection and their parents participated in the study. On the first night, the parents rated their children's cough symptoms on a scale from 0 (no problems at all) to 6 (extremely severe) in five different areas. The total symptoms score (ranging from 0 to 30 points) was the variable of interest for the 105 patients. On the second night, the parents were instructed to give their sick child a dosage of liquid "medicine" prior to bedtime. Unknown to the parents, some were given a dosage of dextromethorphan (DM)—an over-the-counter cough medicine—while others were given a similar dose of honey. Also, a third group of parents (the control group) gave their sick children no dosage at all. Again, the parents rated their children's cough symptoms, and the improvement in total cough symptoms score was determined for each child. The data (improvement scores) for the study are shown in the table below, followed (in the next column) by a Minitab dot plot of the data. Notice that the green dots represent the children who received a dose of honey, the red dots represent those who got the DM dosage, and the black dots represent the children in the control group. What conclusions can pediatric researchers draw from the graph? Do you agree with the statement (extracted from the article), "Honey may be a preferable treatment for the cough and sleep difficulty associated with childhood upper respiratory tract infection"?

Honey Dosage:	12	11	15	11	10	13	10	4	15	16	9	14	
	10	6	10	8	11	12	12	8	9	5	12		
	12	9	11	15	10	15	9	13	8	12	10	8	
DM Dosage:	4	6	9	4	7	7	7	9	12	10	11	6	3
	4	9	12	7	6	8	12	12	4	12	10	15	9
	13	7	10	13	9	4	4						
No Dosage (Control):	5	8	6	1	0	8	12	8	7	7	1	6	7
	7	12	7	9	7	9	5	11	9	5	1	4	3
	6	8	8	6	7	10	9	4	8	7	3		

Source: Based on I. M. Paul et al. "Effect of Honey, Dextromethorphan, and No Treatment on Nocturnal Cough and Sleep Quality for Coughing Children and Their Parents," *Archives of Pediatrics and Adolescent Medicine,* Vol. 161, No. 12, December 2007 (data simulated).



2.32 VESSEL **Shipments.** *The Review of Maritime Transport* is a publication prepared by the United Nations Conference on Trade and Development (UNCTAD) secretariat. It provides an overview of certain characteristics of vessels such as age, size, cargo carrying capacity, and container carrying capacity, as well as the time vessels spend in a country's ports over a certain period. The age of 32 vessels from three different regions in Asia (East, Southeast, and West) for 2020 are listed in the accompanying table.

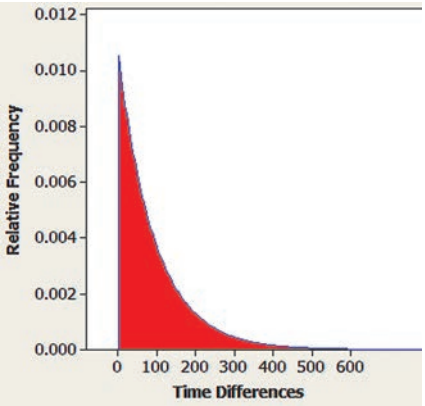
Asia Region	Age of Vessels					
East Asia	13	14	14	30	15	
Southeast Asia	14	14	15	14	15	20
	11	17	16	14		
West Asia	32	16	30	19	23	16
	19	15	15	23	13	12
	14	21	25	15	21	

Source: UNCTAD, based on data provided by MarineTraffic (www.marinetraffic.com).

a. Construct a stem-and-leaf display for the age of all 32 vessels.
b. Summarize the information reflected in the stem-and-leaf display from part **a.** Make a general statement about the age of vessels for West Asia.
c. Select a graphical method that will permit a comparison of the age of vessels for the three different regions in Asia.
d. Mark the vessels from East Asia, which has the smallest number of vessels, on the stem-and-leaf display in part a, by circling their ages. Do you observe any pattern in the graph? Explain.

2.33 PHISH **Phishing attacks to email accounts.** *Phishing* is the term used to describe an attempt to extract personal/financial information (e.g., PIN numbers, credit card information, bank account numbers) from unsuspecting people through fraudulent email. An article in *Chance* (Summer 2007) demonstrates how statistics can help identify phishing attempts and make e-commerce safer. Data from an actual phishing attack against an organization were used to determine whether the attack may have been an "inside job" that originated within the company. The company set up a publicized email account—called a "fraud box"—that enabled employees to notify them if they suspected an email phishing attack. The interarrival times, i.e., the time differences (in seconds), for 267 fraud box email notifications were

recorded and saved in the file. *Chance* showed that if there is minimal or no collaboration or collusion from within the company, the interarrival times would have a frequency distribution similar to the one shown in the accompanying figure. Construct a frequency histogram for the interarrival times. Give your opinion on whether the phishing attack against the organization was an "inside job."



2.34 **Made-to-order delivery times.** Production processes may be classified as *make-to-stock processes* or *make-to-order processes*. Make-to-stock processes are designed to produce a standardized product that can be sold to customers from the firm's inventory. Make-to-order processes are designed to produce products according to customer specifications (Schroeder, *Operations Management,* 2008). In general, performance of make-to-order processes is measured by delivery time—the time from receipt of an order until the product is delivered to the customer. The accompanying data set is a sample of delivery times (in days) for a particular make-to-order firm last year. The delivery times marked by an asterisk are associated with customers who subsequently placed additional orders with the firm.

MTO

50*	64*	56*	43*	64*	82*	65*	49*	32*	63*	44*	71
54*	51*	102	49*	73*	50*	39*	86	33*	95	59*	51*
68											

Concerned that they are losing potential repeat customers because of long delivery times, the management would like to establish a guideline for the maximum tolerable delivery time. Use a graphical method to help suggest a guideline. Explain your reasoning.

2.3 Numerical Measures of Central Tendency

When we speak of a data set, we refer to either a sample or a population. If statistical inference is our goal, we'll wish ultimately to use sample **numerical descriptive measures** to make inferences about the corresponding measures for the population.

As you'll see, a large number of numerical methods are available to describe quantitative data sets. Most of these methods measure one of two data characteristics:

1. The **central tendency** of the set of measurements—that is, the tendency of the data to cluster, or center, about certain numerical values (see Figure 2.14a).
2. The **variability** of the set of measurements—that is, the spread of the data (see Figure 2.14b).

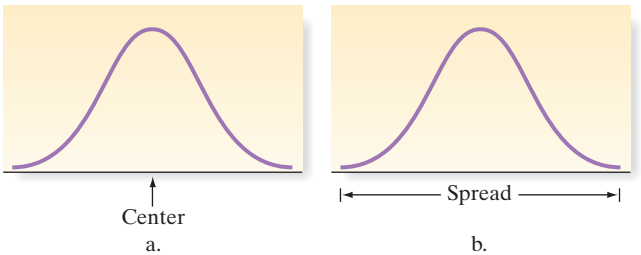


Figure 2.14
Numerical descriptive measures

In this section we concentrate on **measures of central tendency.** In the next section, we discuss measures of variability. The most popular and best-understood measure of central tendency for quantitative data is the arithmetic mean (or simply the mean) of a data set.

The **mean** of a set of quantitative data is the sum of the measurements divided by the number of measurements contained in the data set.

In everyday terms, the mean is the average value of the data set and is often used to represent a "typical" value. We denote the **mean of a sample** of measurements by $\bar{x}$ (read "x-bar") and represent the formula for its calculation as shown in the box below.

Formula for a Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

[*Note:* $\sum_{i=1}^{n} x_i = (x_1 + x_2 + \cdots + x_n)$. For more details on this summation notation, see Appendix A.]

EXAMPLE 2.3

Calculating the Sample Mean

Problem Calculate the mean of the following five sample measurements: 5, 3, 8, 5, 6.

Solution Using the definition of sample mean and the summation notation, we find

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{5 + 3 + 8 + 5 + 6}{5} = \frac{27}{5} = 5.4$$

Thus, the mean of this sample is 5.4.

Look Back There is no specific rule for rounding when calculating $\bar{x}$ because $\bar{x}$ is specifically defined to be the sum of all measurements divided by n—that is, it is a specific fraction. When $\bar{x}$ is used for descriptive purposes, it is often convenient to round the calculated value of $\bar{x}$ to the number of significant figures used for the original measurements. When $\bar{x}$ is to be used in other calculations, however, it may be necessary to retain more significant figures.

Now Work Exercise 2.36

EXAMPLE 2.4

Finding the Mean on a Printout—R&D Expenditures

Problem Calculate the sample mean for the R&D expenditure percentages of the 50 companies given in Table 2.2 on page 52.

Solution The mean R&D percentage for the 50 companies is denoted

$$\bar{x} = \frac{\sum_{i=1}^{50} x_i}{50}$$

We employed Excel and XLSTAT to compute the mean. The XLSTAT printout is shown in Figure 2.15. The sample mean, highlighted on the printout, is $\bar{x} = 8.492$.

Descriptive statistics (Quantitative data):

Statistic	RDPct
Nbr. of observations	50
Minimum	5.2000
Maximum	13.5000
Range	8.3000
1st Quartile	7.1000
Median	8.0500
3rd Quartile	9.5750
Mean	8.4920
Variance (n-1)	3.9228
Standard deviation (n-1)	1.9806

Figure 2.15
XLSTAT numerical descriptive measures for 50 R&D percentages

Look Back Given this information, you can visualize a distribution of R&D percentages centered in the vicinity $\bar{x} = 8.492$. An examination of the relative frequency histogram (Figure 2.10) confirms that $\bar{x}$ does, in fact, fall near the center of the distribution.

The sample mean $\bar{x}$ will play an important role in accomplishing our objective of making inferences about populations based on sample information. For this reason, we need to use a different symbol for the *mean of a population*–the mean of the set of measurements on every unit in the population. We use the Greek letter μ (mu) for the population mean.

Symbols for the Sample and Population Mean

In this text, we adopt a general policy of using Greek letters to represent population numerical descriptive measures and Roman letters to represent corresponding descriptive measures for the sample. The symbols for the mean are

$\bar{x}$ = Sample mean

μ = Population mean*

We'll often use the sample mean $\bar{x}$ to estimate (make an inference about) the population mean, μ. For example, the percentages of revenues spent on R&D by the population consisting of *all* US companies has a mean equal to some value, μ. Our sample of 50 companies yielded percentages with a mean of $\bar{x} = 8.492$. If, as is usually the case, we don't have access to the measurements for the entire population, we could use $\bar{x}$ as an estimator or approximator for μ. Then we'd need to know something about the reliability of our inference—that is, we'd need to know how accurately we might expect $\bar{x}$ to estimate μ. In Chapter 6, we'll find that accuracy depends on two factors:

1. The *size of the sample*. The larger the sample, the more accurate the estimate will tend to be.
2. The *variability,* or *spread, of the data*. All other factors remaining constant, the more variable the data, the less accurate the estimate.

Another important measure of central tendency is the *median*.

The **median** of a quantitative data set is the middle number when the measurements are arranged in ascending (or descending) order.

The median is of most value in describing large data sets. If the data set is characterized by a relative frequency histogram (Figure 2.16), the median is the point on the x-axis such that half the area under the histogram lies above the median and half lies below. [*Note:* In Section 2.2, we observed that the relative frequency associated with a particular interval on the horizontal axis is proportional to the amount of area under the histogram that lies above the interval.] We denote the *median of a sample* by m. Like with the population mean, we use a Greek letter (η) to represent the population median.

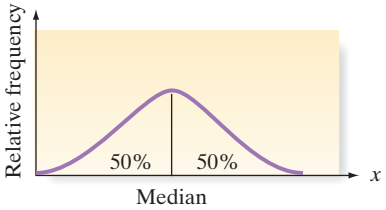


Figure 2.16
Location of the median

*The population mean μ is calculated as $\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$, where N is the population size.

Calculating a Sample Median, m

Arrange the n measurements from smallest to largest.

1. If n is odd, m is the middle number.
2. If n is even, m is the mean of the middle two numbers.

Symbols for the Sample and Population Median

m = sample median

η = population median*

EXAMPLE 2.5

Computing the Median

Problem Consider the following sample of $n = 7$ measurements: 5, 7, 4, 5, 20, 6, 2.

a. Calculate the median m of this sample.

b. Eliminate the last measurement (the 2) and calculate the median of the remaining $n = 6$ measurements.

Solution

a. The seven measurements in the sample are ranked in ascending order: 2, 4, 5, 5, 6, 7, 20. Because the number of measurements is odd, the median is the middle measurement. Thus, the median of this sample is $m = 5$ (the second 5 listed in the sequence).

b. After removing the 2 from the set of measurements, we rank the sample measurements in ascending order as follows: 4, 5, 5, 6, 7, 20. Now the number of measurements is even, so we average the middle two measurements. The median is $m = (5 + 6)/2 = 5.5$.

Look Back When the sample size n is even and the two middle numbers are different (as in part **b**), exactly half of the measurements will fall below the calculated median m. However, when n is odd (as in part **a**), the percentage of measurements that fall below m is approximately 50%. This approximation improves as n increases.

Now Work Exercise 2.35

In certain situations, the median may be a better measure of central tendency than the mean. In particular, the median is less sensitive than the mean to extremely large or small measurements. Note, for instance, that all but one of the measurements in part **a** of Example 2.5 center about $x = 5$. The single relatively large measurement, $x = 20$, does not affect the value of the median, 5, but it causes the mean, $\bar{x} = 7$, to lie to the right of most of the measurements.

As another example of data for which the central tendency is better described by the median than the mean, consider the salaries of professional athletes (e.g., National Basketball Association players). The presence of just a few athletes (e.g., LeBron James) with extremely high salaries will affect the mean more than the median. Thus, the median will provide a more accurate picture of the typical salary for the professional league. The mean could exceed the vast majority of the sample measurements (salaries), making it a misleading measure of central tendency.

*The population median η is calculated like the sample median, but with all N observations in the population arranged from smallest to largest.

EXAMPLE 2.6

Finding the Median on a Printout—R&D Expenditures

Problem Calculate the median for the 50 R&D percentages given in Table 2.2 on page 52. Compare the median to the mean found in Example 2.4.

Solution For this large data set, we again resort to a computer analysis. The median is highlighted on the XLSTAT printout, Figure 2.15. You can see that the median is 8.05. This value implies that half of the 50 R&D percentages in the data set fall below 8.05 and half lie above 8.05.

Note that the mean (8.492) for these data is larger than the median. This fact indicates that the data are **skewed** to the right–that is, there are more extreme measurements in the right tail of the distribution than in the left tail (recall the histogram in Figure 2.10).

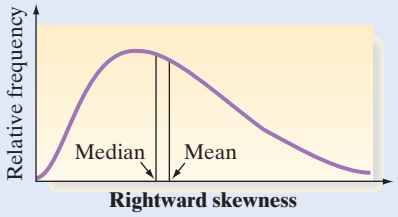
Look Back In general, extreme values (large or small) affect the mean more than the median because these values are used explicitly in the calculation of the mean. On the other hand, the median is not affected directly by extreme measurements because only the middle measurement (or two middle measurements) is explicitly used to calculate the median. Consequently, if measurements are pulled toward one end of the distribution (as with the R&D percentages), the mean will shift toward that tail more than the median.

A data set is said to be **skewed** if one tail of the distribution has more extreme observations than the other tail.

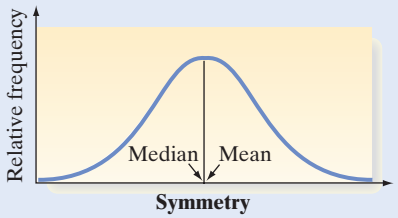
A comparison of the mean and the median gives us a general method for detecting skewness in data sets, as shown in the next box. With *rightward skewed* data, the right tail (high end) of the distribution has more extreme observations. These few, but large, measurements tend to pull the mean away from the median toward the right; that is, rightward skewness typically indicates that the mean is greater than the median. Conversely, with *leftward skewed* data, the left tail (low end) of the distribution has more extreme observations. These few, but small, measurements also tend to pull the mean away from the median but toward the left; consequently, leftward skewness typically implies that the mean is smaller than the median.

Detecting Skewness by Comparing the Mean and the Median

If the data set is skewed to the right, then typically the median is less than the mean.

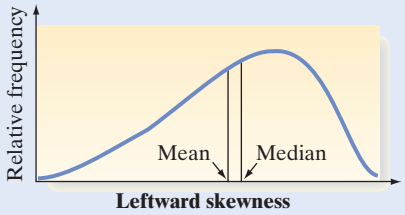


If the data set is symmetric, then the mean equals the median.



Continued

If the data set is skewed to the left, then typically the mean is less than (to the left of) the median.



A third measure of central tendency is the *mode* of a set of measurements.

The **mode** is the measurement that occurs most frequently in the data set.

EXAMPLE 2.7

Finding the Mode

Problem Each of 10 taste testers rated a new brand of barbecue sauce on a 10-point scale, where 1 = awful and 10 = excellent. Find the mode for the 10 ratings shown below.

8 7 9 6 8 10 9 9 5 7

Solution Because 9 occurs most often, the mode of the 10 taste-ratings is 9.

Look Back Note that the data are actually qualitative in nature (e.g., "awful," "excellent"). The mode is particularly useful for describing qualitative data. The modal category is simply the category (or class) that occurs most often.

Now Work Exercise 2.39

Because it emphasizes data concentration, the mode is used with quantitative data sets to locate the region in which much of the data are concentrated. A retailer of men's clothing would be interested in the modal neck size and sleeve length of potential customers. The modal income class of the laborers in the United States is of interest to the Labor Department.

For some quantitative data sets, the mode may not be very meaningful. For example, consider the percentages of revenues spent on R&D by 50 companies, Table 2.2. A reexamination of the data reveals that three of the measurements are repeated three times: 6.5%, 6.9%, and 8.2%. Thus, there are three modes in the sample, and none is particularly useful as a measure of central tendency.

A more meaningful measure can be obtained from a relative frequency histogram for quantitative data. The class interval containing the largest relative frequency is called the **modal class.** Several definitions exist for locating the position of the mode within a modal class, but the simplest is to define the mode as the midpoint of the modal class. For example, examine the relative frequency histogram for the price quote processing times in Figure 2.12. You can see that the modal class is the interval (3.0–4.0). The mode (the midpoint) is 3.5. This modal class (and the mode itself) identifies the area in which the data are most concentrated, and in that sense it is a measure of central tendency. However, for most applications involving quantitative data, the mean and median provide more descriptive information than the mode.

EXAMPLE 2.8

Comparing the Mean, Median, and Mode—CEO Salaries

Problem Each year the Equilar Institute and *New York Times* publish a list of salaries for the 200 highest paid CEOs in the United States. Data for the top 50 CEOs in 2019, saved in the **CEO50** file, includes the quantitative variables CEO total compensation (in billions of dollars) and pay (in thousands of dollars) of typical workers at the CEO's firm. Examining the file, you'll see that Elon Musk (of Tesla) was the highest paid CEO at over $2,284 billion. The next highest – well below Musk -- was David Zaslav (of Discovery) at $129.5 billion. Find the mean, median, and mode for both of these variables. Which measure of central tendency is better for describing the distribution of CEO annual salary? Typical worker pay?

Solution Measures of central tendency for the two variables were obtained using StatCrunch. The means, medians, and modes are displayed at the top of the StatCrunch printout, Figure 2.17a. For CEO salary, the mean and median are $83.8 billion and $28.3 billion, respectively. (No mode is reported since there is no single salary value that is repeated in the data set.) Note that the mean is much greater than the median, indicating that the data are highly skewed right. This rightward skewness, graphically shown on the Minitab histogram for CEO salary in Figure 2.17b (top), is mostly due to Elon Musk's exceptionally high salary of $2,284 billion in 2019. In fact, Musk's salary is so large that it is not shown on the histogram. Consequently, we would probably want to use the median, $28.3 billion, as the "typical" value for the top 50 CEO salaries.

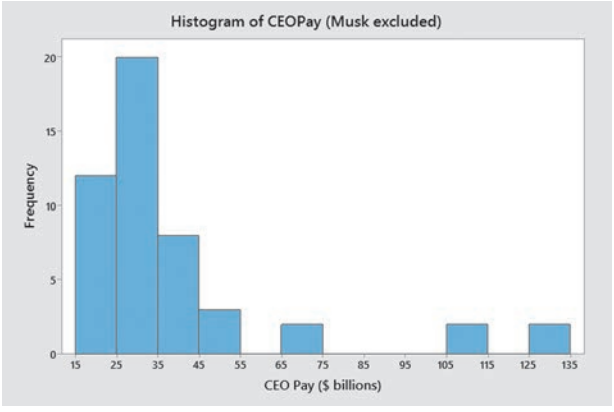
For typical worker pay, the mean, median, and mode shown in Figure 2.17a are 84.3, 79.4, and 89.9 thousand dollars, respectively. The mean and median are nearly the same, which is a property of symmetric distributions. The mode results from the fact that Oracle placed two CEOs (Mark Hurd and Safra Catz) on the list, hence their worker pay value was the same ($89.9 thousand). Thus, for this data set the mode is not very meaningful. From the histogram (bottom) in Figure 2.17b, you can see that the worker pay distribution is nearly symmetric, and that the modal class centered at $75 thousand includes both the mean and the median. Consequently, either the mean or the median could be used to describe the "middle" of the typical worker pay distribution.

Look Back The choice of which measure of central tendency to use will depend on the properties of the data set analyzed and on the application. Consequently, it is vital that you understand how the mean, median, and mode are computed.

Summary statistics:

Column	n	Mean	Median	Mode
CEO Pay ($ billions)	50	83.814913	28.296623	No mode
Worker Pay ($ thousands)	50	84.3027	79.3735	89.887

Figure 2.17a
StatCrunch Descriptive Statistics for CEO Compensation and Worker Pay



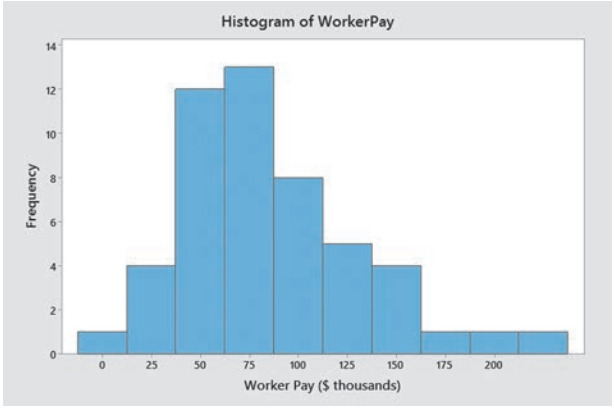


Figure 2.17b
Minitab Histograms for CEO Compensation and Worker Pay

Now Work Exercise 2.46a, b

Exercises 2.35–2.55

Learning the Mechanics

2.35 Calculate the mean and median of the following selected 5 monthly salaries (SGD):

5,020 3,200 1,200 3,700 2,800

2.36 Calculate the mean for samples where
a. $n = 56, \sum x = 784$
b. $n = 105, \sum x = 5250$
c. $n = 12, \sum x = 1158$
d. $n = 23, \sum x = 437$

2.37 Explain how the relationship between the mean and median provides information about the symmetry or skewness of the data's distribution.

2.38 **Parking issues at a university.** A university's administrators are interested in determining the average time it takes a student to find a parking spot. An administrator inconspicuously followed 190 students and recorded how long it took each of them to find a parking spot. The durations had a distribution that was skewed to the left. Based on this information, discuss the relationship between the mean and the median for the 190 times collected.

2.39 Calculate the mode, mean, and median of the following data:

22 9 18 10 19 20 13 11 14 10

2.40 Calculate the mean, the median, and the mode for the following samples.
a. −2, 0, 5, 5, 5, 10, 11
b. 37, 98, 15, 38, 63, 19
c. Five food tasters rated a new product either "'awful' or "'excellent'. Their ratings were awful, awful, excellent, excellent, and awful, respectively.

2.41 Describe how the mean compares to the median for a distribution as follows:
a. Skewed to the left
b. Skewed to the right
c. Symmetric

Applet Exercise 2.1

Use the applet entitled *Mean versus Median* to find the mean and median of each of the three data sets in Exercise 2.40. For each data set, set the lower limit to a number less than all of the data, set the upper limit to a number greater than all of the data, and then click on *Update*. Click on the approximate location of each data item on the number line. You can get rid of a point by dragging it to the trash can. To clear the graph between data sets, simply click on the trash can.

a. Compare the means and medians generated by the applet to those you calculated by hand in Exercise 2.40. If there are differences, explain why the applet might give values slightly different from the hand calculations.
b. Despite providing only approximate values of the mean and median of a data set, describe some advantages of using the applet to find these values.

Applet Exercise 2.2

Use the applet *Mean versus Median* to illustrate your descriptions in Exercise 2.41. For each part **a, b,** and **c,** create a data set with 10 items that has the given property. Using the applet, verify that the mean and median have the relationship you described in Exercise 2.41.

Applet Exercise 2.3

Use the applet *Mean versus Median* to study the effect that an extreme value has on the difference between the mean and median. Begin by setting appropriate limits and plotting the given data on the number line provided in the applet.

0 6 7 7 8 8 8 9 9 10

a. Describe the shape of the distribution and record the value of the mean and median. Based on the shape of the distribution, do the mean and median have the relationship that you would expect?
b. Replace the extreme value of 0 with 2, then 4, and then 6. Record the mean and median each time. Describe what is happening to the mean as 0 is replaced by higher numbers. What is happening to the median? How is the difference between the mean and the median changing?
c. Now replace 0 with 8. What values does the applet give you for the mean and the median? Explain why the mean and the median should be the same.

Applying the Concepts—Basic

2.42 **Hotels' use of ecolabels.** Ecolabels such as *Energy Star*, *Green Key*, and *Audubon International* are used by hotels to advertise their energy-saving and conservation policies. The *Journal of Vacation Marketing* (January 2016) published a study to investigate how familiar travelers are with these ecolabels and whether travelers believe they are credible. A sample of 392 adult travelers were administered a questionnaire. One question showed a list of 6 different ecolabels, and asked, "How familiar are you with this ecolabel, on a scale of 1 (*not familiar at all*) to 5 (*very familiar*)." Summarized results for the numerical responses are given in the table.
a. Give a practical interpretation of the mean response for Energy Star.
b. Give a practical interpretation of the median response for Energy Star.
c. Give a practical interpretation of the response mode for Energy Star.
d. Based on these summary statistics, which ecolabel appears to be most familiar to travelers?

Ecolabel	Mean	Median	Mode
Energy Star	4.44	5	5
TripAdvisor Greenleaders	3.57	4	4
Audubon International	2.41	2	1
US Green Building Council	2.28	2	1
Green Business Bureau	2.25	2	1
Green Key	2.01	1	1

Source: S. Park and M. Millar, "The US Traveler's Familiarity with and Perceived Credibility of Lodging Ecolabels," *Journal of Vacation Marketing*, Vol. 22, No. 1, January 2016 (Table 3).

Rank	University	Public/ Private	Academic Reputation Score (100-pt. scale)	Average Financial Aid Awarded	Average Net Cost to Attend	Median Salary During Early Career	% High Meaning*	% STEM† Degrees
1	Yale	Private	98	$42,792	$16,528	$70,300	53	22
2	Harvard	Private	99	$41,505	$16,445	$74,800	54	19
22	Virginia	Public	76	$18,236	$13,463	$64,500	46	23
23	Emory	Private	77	$30,768	$27,412	$62,000	44	20
24	UCLA	Public	76	$17,703	$13,686	$62,000	48	31
49	Purdue	Public	56	$10,902	$13,541	$60,900	48	38
50	Texas A&M	Public	55	$10,034	$11,323	$60,600	53	30

*% high meaning represents the percentage of alumni who say their work makes the world a better place;

†% STEM represents the percentage of degrees awarded in science, technology, engineering, or mathematics.

2.43 ENERGY **Household energy expenditure.** The annual total household expenditure on energy in the United Kingdom from 2011 to 2020 was selected from the Office of National Statistics' Consumer Trends Time Series data. The data consists of the annual total household expenditure (in £ millions) on gas, electricity, liquid fuels, and the total of individual household's expenditure on energy which denoted as the total expenditure (in £ millions). The data are saved in the **ENERGY** file. The first two and last two rows are listed in the table above.

Year	Gas	Electricity	Liquid Fuels	Total Expenditure
2011	12,119	13,389	1,250	983,041
2012	12,961	13,924	1,254	1,000,382
⋮	⋮	⋮	⋮	⋮
2019	11,225	12,259	1,258	1,166,026
2020	10,603	12,341	1,327	1,042,290

Source: Consumer Trends Time Series, Office for National Statistics, https://www.ons.gov.uk/.

a. Calculate the average of the total expenditure on energy for the ten years (2011 to 2020) and determine whether this statistic represents a population or sample mean. Interpret this value.

b. Find the median of the total expenditure on energy for the ten years. Does this statistic represent a population or sample median? Interpret this value.

2.44 SCREEN **Performance of stock screeners.** Investment companies provide their clients with automated tools—called *stock screeners*—to help them select a portfolio of stocks to invest in. The American Association of Individual Investors (AAII) provides statistics on stock screeners at its Web site, www.aaii.com. The next table lists the annualized percentage return on investment (as compared to the Standard & Poor's 500 Index) for 10 randomly selected stock screeners. (*Note:* A negative annualized return reflects a stock portfolio that performed worse than the S&P 500.)

9.0	−.1	−1.6	14.6	16.0	7.7	19.9	9.8	3.2	24.8

a. Compute the mean for the data set. Interpret its value.

b. Compute the median for the data set. Interpret its value.

2.45 CAPRATE **Valuation of single-tenant properties**. Refer to *The Appraisal Journal* (Summer, 2019) study of the valuation of single-tenant properties, Exercise 2.28 (p. 79). Consider, again, the 5-year capitalization rates for the 13 tenants in the sample. Find the mean, median, and mode for the sample data. Interpret, practically, the values of each.

2.46 ATTACK **Malicious attacks on a cryptocurrency network.** Refer to the *Computers & Security* (January 2020) study of malicious attacks on a cryptocurrency network called Ethereum, Exercise 2.25 (p. 77). Recall that a sample of 160 normal data packets traveling through the network was compared to a sample of 130 data packets that were maliciously attacked. The quantitative variable of interest was the size (number of bytes) of each data packet. The mean and median size for both the normal and attacked packets are shown in the accompanying XLSTAT printout.

a. Locate and interpret the mean size for each sample.

b. Locate and interpret the median size for each sample.

c. Based on this information, what statement can you make about the sizes of normal and attacked data packets?

Descriptive statistics (Quantitative data):

Statistic	Size | Attack	Size | Normal
Nbr. of observation	130	160
Minimum	176.0000	175.0000
Maximum	772.0000	564.0000
Median	207.5000	223.0000
Mean	276.8538	337.4500

Applying the Concepts—Intermediate

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

saved in the **STONE** file. Measures of central tendency for the permeability measurements of each sandstone group are displayed in the accompanying Minitab printout.

Statistics

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

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

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

2.48 CORSUS **Corporate sustainability of CPA firms.** Refer to the *Business and Society* (March 2011) study on the sustainability behaviors of CPA corporations, Exercise 2.23 (p. 77). Recall that level of support for corporate sustainability (measured on a quantitative scale ranging from 0 to 160 points) was obtained for each of 992 senior managers at CPA firms. Numerical measures of central tendency for level of support are shown in the accompanying StatCrunch printout.

Summary statistics:

Column	n	Mean	Median	Min	Max	Mode
Support	992	67.75504	68	0	155	64

a. Locate the mean on the printout. Comment on the accuracy of the statement: "On average, the level of support for corporate sustainability for the 992 senior managers was 67.76 points."
b. Locate the median on the printout. Comment on the accuracy of the statement: "Half of the 992 senior managers reported a level of support for corporate sustainability below 68 points."
c. Locate the mode on the printout. Comment on the accuracy of the statement: "Most of the 992 senior managers reported a level of support for corporate sustainability below 64 points."
d. Based on the values of the measures of central tendency, make a statement about the type of skewness (if any) that exists in the distribution of 992 support levels. Check your answer by examining the histogram shown in Exercise 2.23.

2.49 HCOUGH **Is honey a cough remedy?** Refer to the *Archives of Pediatrics and Adolescent Medicine* (December 2007) study of honey as a remedy for coughing, Exercise 2.31 (p. 80). Recall that the 105 ill children in the sample were randomly divided into three groups: those who received a dosage of an over-the-counter cough medicine (DM), those who received a dosage of honey (H), and those who received no dosage (control group). The coughing improvement scores (as determined by the children's parents) for the patients are reproduced in the table.

Honey Dosage:	12	11	15	11	10	13	10	4	15	16	9	14			
	10	6	10	8	11	12	12	8	12	9	11	15			
	10	15	9	13	8	12	10	8	9	5	12				
DM Dosage:	4	6	9	4	7	7	7	9	12	10	11	6	3	4	
	9	12	7	6	8	12	12	4	12	13	7	10			
	13	9	4	4	10	15	9								
No Dosage (Control):	5	8	6	1	0	8	12	8	7	7	1	6	7	7	12
	7	9	7	9	5	11	9	5	6	8					
	8	6	7	10	9	4	8	7	3	1	4	3			

Source: Based on I. M. Paul et al., "Effect of Honey, Dextromethorphan, and No Treatment on Nocturnal Cough and Sleep Quality for Coughing Children and Their Parents," *Archives of Pediatrics and Adolescent Medicine,* Vol. 161, No. 12, December 2007 (data simulated).

a. Find the median improvement score for the honey dosage group.
b. Find the median improvement score for the DM dosage group.
c. Find the median improvement score for the control group.
d. Based on the results, parts **a–c,** what conclusions can pediatric researchers draw? (We show how to support these conclusions with a measure of reliability in subsequent chapters.)

2.50 BIODEG **Crude oil biodegradation.** Refer to the *Journal of Petroleum Geology* (April 2010) study of the environmental factors associated with biodegradation in crude oil reservoirs, Exercise 2.29 (p. 79). Recall that amount of dioxide (milligrams/liter) and presence/absence of crude oil was determined for each of 16 water specimens collected from a mine reservoir. The data are repeated in the accompanying table.

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

Source: Based on A. Permanyer, J. L. R. Gallego, M. A. Caja, and D. Dessort, "Crude Oil Biodegradation and Environmental Factors at the Riutort Oil Shale Mine, SE Pyrenees," *Journal of Petroleum Geology*, Vol. 33, No. 2, April 2010, (Table 1).

a. Find the mean dioxide level of the 16 water specimens. Interpret this value.
b. Find the median dioxide level of the 16 water specimens. Interpret this value.

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

2.51 **Symmetric or Skewed?** Would you expect the data sets described below to possess relative frequency distributions that are symmetric, skewed to the right, or skewed to the left? Explain.
a. The daily stock market returns
b. The income in an underdeveloped country
c. The average age of death in a particular country
d. The average grade point for college students
e. The amount of time spent on reading by a working adult per month
f. The age of retirement

2.52 **Ranking driving performance of professional golfers.** A group of researchers developed a new method for ranking the total driving performance of golfers on the Professional Golf Association (PGA) tour (*The Sport Journal,* Winter 2007). The method requires knowing a golfer's average driving distance (yards) and driving accuracy (percent of drives that land in the fairway). The values of these two variables are used to compute a driving performance index. Data for the top 25 PGA golfers (as ranked by the new method for the 2019 tour year) are saved in the accompanying file. The first five and last five observations are listed in the table below.

PGA

Rank	Player	Driving Distance (yards)	Driving Accuracy (%)	Driving Performance Index
1	Woodland	322.3	72.56	9.43
2	Champ	327.1	57.14	6.50
3	Trahan	308.5	68.25	5.49
4	Kang	316.2	60.48	5.10
5	Furyk	287.5	80.95	4.39
21	Riley	300.9	58.93	1.55
22	Gligic	301.6	58.21	1.51
23	Herron	282	71.43	0.83
24	Bae	302.3	54.59	0.74
25	Jacobson	293.9	60.71	0.56

Source: Based on Frederick Wiseman, Ph.D., Mohamed Habibullah, Ph.D., and Mustafa Yilmaz, Ph.D, "Ranking Driving Performance on the PGA Tour," *Sports Journal*, Vol. 10, No. 1, Winter 2007 (Table 2). 2019 data from www.pgatour.com/stats.

a. Find the mean, median, and mode for the 25 driving performance index values.
b. Interpret each of the measures of central tendency, part **a.**
c. Use the results, part **a,** to make a statement about the type of skewness in the distribution of driving performance indexes. Support your statement with a graph.

Applying the Concepts—Advanced

2.53 **Shipments.** Refer to Exercise 2.32 (p. 80) on the average age of 32 vessels for three different regions in Asia (East, Southeast, and West) for the year 2020. Is it reasonable to use a single number (e.g., mean or median) to describe the center of the average age of vessels distributions? Or should three "centers" be calculated, one for each of the three regions of Asia? Explain.

VESSEL

2.54 **Traffic.** A survey was conducted on the traffic counts and accidents that occurred in Bloomington during 2019. The data set contains road names listed in alphabetical order from A to B and considers the accident on that street which involved the highest number of vehicles. The dataset mentions whether the accident was a hit and run or not (Y or N) and lists the number of vehicles involved.

TRAFFIC

Roadway Name	Hit and Run?	Vehicles Involved
Acadia	Y	2
Adams	Y	4
Addisyn	N	2
Airport	N	2
Alder	Y	2
Alexander	Y	1
Allen	Y	1
Amy	Y	1
Anderson	N	1
Andy	N	1
Apache	N	2
Arch Haven	N	2
Arlington	N	2
Association	N	2
Atwater	N	3
Auto Mall	N	2
Baby Creek	N	1
Ballantine	N	2
Basswood	N	1
Beechwood	N	1
Bell	N	3
Bell Trace	Y	2
Bennington	N	1
Bethel	N	1
Black Foot	N	2
Bloomfield	N	3
Boardwalk	N	2
Bolin	N	1
Boltinghouse	N	1
Bottom	N	1
Braeside	N	2
Brandon	Y	2
Breeden	N	2
Brownstone	N	2
Brummetts Creek	N	1
Buick Cadillac	N	2
Burks	N	1
Burma	N	2
Bushmill	Y	1
Business 37	N	1

Source: data.world

a. Find the mean, median, and mode of the number of the vehicles involved in each accident considered per street. Interpret these values.
b. Repeat part **a** for the accidents that involved a hit and run only.
c. Repeat part **a** for the accidents that did not involve a hit and run.

d. Compare the results, parts **b** and **c** What inference can you make about the impact of the hit and run status of the accident on the number of vehicles involved?

e. Eliminate the road name with the highest number of vehicles involved during its accident from the data set and repeat part a. Does dropping this measurement have any effect on the measures of central tendency found in part **a**?

f. Rearrange the 40 values in the table from the lowest to the highest. Next, eliminate the four lowest values and the four highest values from the data set and find the mean of the remaining data values. The result is called a 10% trimmed mean, because it is calculated after removing the highest 10% and the lowest 10% of the data values. What advantages does a trimmed mean have over the regular arithmetic mean?

2.55 Professional athletes' salaries. The salaries of superstar professional athletes receive much attention in the media. The multimillion-dollar long-term contract is now commonplace among this elite group. Nevertheless, rarely does a season pass without negotiations between one or more of the players' associations and team owners for additional salary and fringe benefits for *all* players in their particular sports.

a. If a players' association wanted to support its argument for higher "average" salaries, which measure of central tendency do you think it should use? Why?

b. To refute the argument, which measure of central tendency should the owners apply to the players' salaries? Why?

2.4 Numerical Measures of Variability

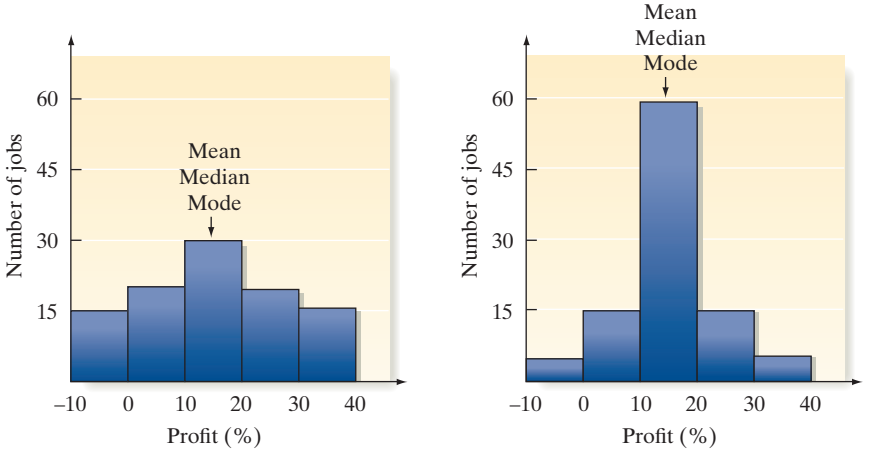
Measures of central tendency provide only a partial description of a quantitative data set. The description is incomplete without a **measure of the variability,** or **spread,** of the data set. Knowledge of the data's variability along with its center can help us visualize the shape of a data set as well as its extreme values.

For example, suppose we are comparing the profit margin per construction job (as a percentage of the total bid price) for 100 construction jobs for each of two cost estimators working for a large construction company. The histograms for the two sets of 100 profit margin measurements are shown in Figure 2.18. If you examine the two histograms, you will notice that both data sets are symmetric with equal modes, medians, and means. However, cost estimator A (Figure 2.18a) has profit margins spread with almost equal relative frequency over the measurement classes, while cost estimator B (Figure 2.18b) has profit margins clustered about the center of the distribution. Thus, estimator B's profit margins are *less variable* than estimator A's. Consequently, you can see that we need a measure of variability as well as a measure of central tendency to describe a data set.

Perhaps the simplest measure of the variability of a quantitative data set is its *range*.

The **range** of a quantitative data set is equal to the largest measurement minus the smallest measurement.

The range is easy to compute and easy to understand, but it is a rather insensitive measure of data variation when the data sets are large. This is because two data sets can have the same range and be vastly different with respect to data variation. This



a. Cost estimator A

b. Cost estimator B

Figure 2.18
Profit margin histograms for two cost estimators

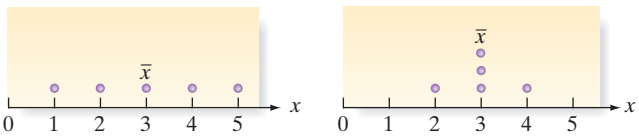
phenomenon is demonstrated in Figure 2.18. Although the ranges are equal and all central tendency measures are the same for these two symmetric data sets, there is an obvious difference between the two sets of measurements. The difference is that estimator B's profit margins tend to be more stable—that is, to pile up or to cluster about the center of the data set. In contrast, estimator A's profit margins are more spread out over the range, indicating a higher incidence of some high profit margins but also a greater risk of losses. Thus, even though the ranges are equal, the profit margin record of estimator A is more variable than that of estimator B, indicating a distinct difference in their cost-estimating characteristics.

Let's see if we can find a measure of data variation that is more sensitive than the range. Consider the two samples in Table 2.5: Each has five measurements. (We have ordered the numbers for convenience.)

Table 2.5 Two Hypothetical Data Sets

	Sample 1	Sample 2
Measurements	1, 2, 3, 4, 5	2, 3, 3, 3, 4
Mean	$\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$	$\bar{x} = \frac{2+3+3+3+4}{5} = \frac{15}{5} = 3$
Deviations of measurement values from $\bar{x}$	$(1-3), (2-3), (3-3), (4-3), (5-3)$, or $-2, -1, 0, 1, 2$	$(2-3), (3-3), (3-3), (3-3), (4-3)$, or $-1, 0, 0, 0, 1$

Note that both samples have a mean of 3 and that we have also calculated the distance and direction, or *deviation*, between each measurement and the mean. What information do these deviations contain? If they tend to be large in magnitude, as in sample 1, the data are spread out, or highly variable, as shown in Figure 2.19a. If the deviations are mostly small, as in sample 2, the data are clustered around the mean, $\bar{x}$, and therefore do not exhibit much variability, as shown in Figure 2.19b. You can see that these deviations provide information about the variability of the sample measurements.



a. Sample 1 b. Sample 2

Figure 2.19
Dot plots for two data sets

The next step is to condense the information in these deviations into a single numerical measure of variability. Averaging the deviations from $\bar{x}$ won't help because the negative and positive deviations cancel; that is, the sum of the deviations (and thus the average deviation) is always equal to zero.

Two methods come to mind for dealing with the fact that positive and negative deviations from the mean cancel. The first is to treat all the deviations as though they were positive, ignoring the sign of the negative deviations. We won't pursue this line of thought because the resulting measure of variability (the mean of the absolute values of the deviations) presents analytical difficulties beyond the scope of this text. A second method of eliminating the minus signs associated with the deviations is to square them. The quantity we can calculate from the squared deviations will provide a meaningful description of the variability of a data set and present fewer analytical difficulties in inference making.

To use the squared deviations calculated from a data set, we first calculate the *sample variance*.

The **sample variance** for a sample of n measurements is equal to the sum of the squared deviations from the mean divided by $(n-1)$. The symbol s^2 is used to represent the sample variance.

Formula for the Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

Note: A shortcut formula for calculating s^2 is

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

Referring to the two samples in Table 2.5, you can calculate the variance for sample 1 as follows:

$$s^2 = \frac{(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5-1}$$

$$= \frac{4 + 1 + 0 + 1 + 4}{4} = 2.5$$

The second step in finding a meaningful measure of data variability is to calculate the *standard deviation* of the data set.

The **sample standard deviation,** s, is defined as the positive square root of the sample variance, s^2. Thus, $s = \sqrt{s^2}$.

The population variance, denoted by the symbol σ^2 (sigma squared), is the average of the squared distances of the measurements on *all* units in the population from the mean, μ, and σ (sigma) is the square root of this quantity. Because we rarely, if ever, have access to the population data, we do not compute σ^2 or σ. We simply denote these two quantities by their respective symbols.*

Symbols for Variance and Standard Deviation

s^2 = Sample variance
s = Sample standard deviation
σ^2 = Population variance
σ = Population standard deviation

Notice that, unlike the variance, the standard deviation is expressed in the original units of measurement. For example, if the original measurements are in dollars, the variance is expressed in the peculiar units "dollars squared," but the standard deviation is expressed in dollars. Consequently, you can think of s as a "typical" distance of an observation x from its mean, $\bar{x}$.

You may wonder why we use the divisor $(n - 1)$ instead of n when calculating the sample variance. Wouldn't using n be more logical so that the sample variance would be the average squared deviation from the mean? The trouble is that using n tends to produce an underestimate of the population variance, σ^2, so we use $(n - 1)$ in the denominator to provide the appropriate correction for this tendency.† Because sample statistics such as s^2 are primarily used to estimate population parameters such as σ^2, $(n - 1)$ is preferred to n when defining the sample variance.

*The population variance, σ^2, is calculated as $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$, where N is the size of the population.

†*Appropriate* here means that s^2 with the divisor $(n - 1)$ is an *unbiased estimator* of σ^2. We define and discuss *unbiasedness* of estimators in Chapter 5.

EXAMPLE 2.9

Computing Measures of Variation

Problem Calculate the variance and standard deviation of the following sample: 2, 3, 3, 3, 4. (These data are entered into an Excel spreadsheet, shown in Figure 2.20.)

Solution If you calculate the values of s and s^2 using the formulas in the boxes on pages 76–77, you first need to compute $\bar{x}$. From Figure 2.20 below, we see that $\Sigma x = 15$. Thus, $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{15}{5} = 3$. Now, for each measurement, find $(x - \bar{x})$ and $(x - \bar{x})^2$, as shown in the Excel spreadsheet.

	A	B	C	D	E
1		X	(X-XBAR)	$(X-XBAR)^2$	
2		2	-1	1	
3		3	0	0	
4		3	0	0	
5		3	0	0	
6		4	1	1	
7					
8	Sum	15	0	2	
9					
10			Variance	0.5	
11					
12			Std. Dev.	0.71	
13					

Figure 2.20
Excel Spreadsheet Showing Variance Calculations

Then we use*

$$s^2 = \frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n - 1} = \frac{2}{5 - 1} = \frac{2}{4} = .5$$

$$s = \sqrt{.5} = .71$$

Look Ahead As the sample size n increases, these calculations can become very tedious. As the next example shows, we can use the computer to find s^2 and s.

Now Work Exercise 2.57

EXAMPLE 2.10

Finding Measures of Variation on a Printout—R&D Expenditures

Problem Use the computer to find the sample variance s^2 and the sample standard deviation s for the 50 companies' percentages of revenues spent on R&D, given in Table 2.2 (p. 70).

Solution The XLSTAT printout describing the R&D percentage data is reproduced in Figure 2.21. The variance and standard deviation, highlighted on the printout, are $s^2 = 3.9228$ and $s = 1.9806$. The value $s = 1.98$ represents a typical deviation of an R&D percentage from the sample mean, $\bar{x} = 8.49\%$. We have more to say about the interpretation of s in the next section.

Descriptive statistics (Quantitative data):

Statistic	RDPct
Minimum	5.2000
Maximum	13.5000
Range	8.3000
1st Quartile	7.1000
Median	8.0500
3rd Quartile	9.5750
Mean	8.4920
Variance (n-1)	3.9228
Standard deviation (n-1)	1.9806

Figure 2.21
XLSTAT numerical descriptive measures for 50 R&D percentages

Now Work Exercise 2.67

You now know that the standard deviation measures the variability of a set of data. The larger the standard deviation, the more variable the data. The smaller the standard deviation, the less variable the data. But how can we practically interpret the standard deviation and use it to make inferences? This is the topic of Section 2.5.

*When calculating s^2, how many decimal places should you carry? Although there are no rules for the rounding procedure, it's reasonable to retain twice as many decimal places in s^2 as you ultimately wish to have in s. If you wish to calculate s to the nearest hundredth (two decimal places), for example, you should calculate s^2 to the nearest ten-thousandth (four decimal places).

Exercises 2.56–2.70

Learning the Mechanics

2.56 Answer the following questions about variability of data sets:
- **a.** What is the primary disadvantage of using the range to compare the variability of data sets?
- **b.** Describe the sample variance using words rather than a formula. Do the same with the population variance.
- **c.** Can the variance of a data set ever be negative? Explain. Can the variance ever be smaller than the standard deviation? Explain.

2.57 Calculate the range, variance, and standard deviation for the following samples:
- **a.** $-1, 1, 2, 0, 8$
- **b.** $3, 1, 5, 1, 4, 0, 2$
- **c.** $4, -5, 0, 1, 1, 3, 3, 2, -2, -3$
- **d.** $1, 4, 1, 3, 2, 2, 5, 1, 3, 3, 3, 4, 5, 3, 2, 1$

2.58 Calculate the variance and standard deviation for samples where
- **a.** $n = 15, \Sigma x^2 = 556, \Sigma x = 54$
- **b.** $n = 24, \Sigma x^2 = 84, \Sigma x = 13$
- **c.** $n = 57, \Sigma x^2 = 237, \Sigma x = 106$

2.59 Compute $\bar{x}$, s^2, and s for each of the following data sets. If appropriate, specify the units in which of your answer is expressed.
- **a.** \$2, \$11, \$17, \$20, \$9
- **b.** 25 minutes, 8 minutes, 17 minutes, 15 minutes
- **c.** $-6\%, -1\%, 0\%, 5\%, 4\%, 2\%$
- **d.** 4 cm, 4 cm, 4 cm, 4 cm, 4 cm, 4 cm, 4 cm

2.60 Calculate the range, variance, and standard deviation for the following samples:
- **a.** $79, 12, 25, 33, 68$
- **b.** $105, 2, 8, 16, 90, 53, 81, 47, 34$
- **c.** $-2, 5, -11, 4, 9, 7, 2, 1$

2.61 Using only integers between 0 and 10, construct two data sets with at least 10 observations each that have the same range but different means. Construct a dot plot for each of your data sets, and mark the mean of each data set on its dot diagram.

2.62 Using only integers between 0 and 10, construct two data sets with at least 10 observations each so that the two sets have the same mean but different variances. Construct dot plots for each of your data sets and mark the mean of each data set on its dot diagram.

2.63 Consider the following sample of six measurements: 10, 11, 13, 15, 18.
- **a.** Calculate the range, s^2, and s.
- **b.** Add 2 to each measurement and repeat part **a.**
- **c.** Subtract 10 from each measurement and repeat part **a.**
- **d.** Considering your answers to parts **a–c,** what seems to be the effect on the variability of a data set by adding the same number to or subtracting the same number from each measurement?

Applet Exercise 2.4

Use the applet entitled *Standard Deviation* to find the standard deviation of each of the four data sets in Exercise 2.57. For each data set, set the lower limit to a number less than all of the data, set the upper limit to a number greater than all of the data, and then click on *Update*. Click on the approximate location of each data item on the number line. You can get rid of a point by dragging it to the trash can. To clear the graph between data sets, simply click on the trash can.
- **a.** Compare the standard deviations generated by the applet to those you calculated by hand in Exercise 2.57. If there are differences, explain why the applet might give values slightly different from the hand calculations.
- **b.** Despite providing a slightly different value of the standard deviation of a data set, describe some advantages of using the applet.

Applet Exercise 2.5

Use the applet *Standard Deviation* to study the effect that multiplying or dividing each number in a data set by the same number has on the standard deviation. Begin by setting appropriate limits and plotting the given data on the number line provided in the applet.

0 1 1 1 2 2 3 4

- **a.** Record the standard deviation. Then multiply each data item by 2, plot the new data items, and record the standard deviation. Repeat the process, first multiplying each of the original data items by 3 and then by 4. Describe what is happening to the standard deviation as the data items are multiplied by higher numbers. Divide each standard deviation by the standard deviation of the original data set. Do you see a pattern? Explain.
- **b.** Divide each of the original data items by 2, plot the new data, and record the standard deviation. Repeat the process, first dividing each of the original data items by 3 and then by 4. Describe what is happening to the standard deviation as the data items are divided by higher numbers. Divide each standard deviation by the standard deviation of the original data set. Do you see a pattern? Explain.
- **c.** Using your results from parts **a** and **b,** describe what happens to the standard deviation of a data set when each of the data items in the set is multiplied or divided by a fixed number n. Experiment by repeating parts **a** and **b** for other data sets if you need to.

Applet Exercise 2.6

Use the applet *Standard Deviation* to study the effect that an extreme value has on the standard deviation. Begin by setting appropriate limits and plotting the given data on the number line provided in the applet.

0 6 7 7 8 8 8 9 9 10

- **a.** Record the standard deviation. Replace the extreme value of 0 with 2, then 4, and then 6. Record the standard deviation each time. Describe what is happening to the standard deviation as 0 is replaced by higher numbers.
- **b.** How would the standard deviation of the data set compare to the original standard deviation if the 0 were replaced by 16? Explain.

Applying the Concepts—Basic

2.64 Hotels' use of ecolabels. Refer to the *Journal of Vacation Marketing* (January 2016) study of travelers' familiarity with ecolabels used by hotels, Exercise 2.42 (p. 88). Recall that a sample of 392 adult travelers were shown a list of 6 different ecolabels, and asked, "How familiar are you with this ecolabel, on a scale of 1 (*not familiar at all*) to 5 (*very familiar*)." The mean and standard deviation of the responses for each ecolabel are provided in the table. Which of the ecolabels had the most variation in numerical responses? Explain.

Ecolabel	Mean	Std. Dev.
Energy Star	4.44	0.82
TripAdvisor Greenleaders	3.57	1.38
Audubon International	2.41	1.44
US Green Building Council	2.28	1.39
Green Business Bureau	2.25	1.39
Green Key	2.01	1.30

Source: S. Park and M. Millar, "The US Traveler's Familiarity with and Perceived Credibility of Lodging Ecolabels," *Journal of Vacation Marketing*, Vol. 22, No. 1, January 2016 (Table 3).

2.65 Permeability of sandstone during weathering. Refer to the *Geographical Analysis* (Vol. 42, 2010) study of the decay properties of sandstone when exposed to the weather, Exercise 2.47 (p. 89). Recall that slices of sandstone blocks were tested for permeability under three conditions: no exposure to any type of weathering (A), repeatedly sprayed with a 10% salt solution (B), and soaked in a 10% salt solution and dried (C). Measures of variation for the permeability measurements (mV) of each sandstone group are displayed in the accompanying Minitab printout.

STONE

Statistics

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

a. Find the range of the permeability measurements for Group A sandstone slices. Verify its value using the minimum and maximum values shown on the printout.
b. Find the standard deviation of the permeability measurements for Group A sandstone slices. Verify its value using the variance shown on the printout.
c. Which condition (A, B, or C) has the more variable permeability data?

2.66 Performance of stock screeners. Refer to the American Association of Individual Investors (AAII) statistics on stock screeners, Exercise 2.44 (p. 89). Annualized percentage return on investment (as compared to the Standard & Poor's 500 Index) for 10 randomly selected stock screeners are reproduced in the table.

SCREEN

9.0	−.1	−1.6	14.6	16.0	7.7	19.9	9.8	3.2	24.8

a. Find the range of the data for the 10 stock screeners. Give the units of measurement for the range.
b. Find the variance of the data for the 10 stock screeners. If possible, give the units of measurement for the variance.
c. Find the standard deviation of the data for the 10 stock screeners. Give the units of measurement for the standard deviation.

Applying the Concepts—Intermediate

2.67 Corporate sustainability of CPA firms. Refer to the *Business and Society* (March 2011) study on the sustainability behaviors of CPA corporations, Exercise 2.48 (p. 90). Numerical measures of variation for level of support for the 992 senior managers are shown in the accompanying StatCrunch printout.

CORSUS
NW

Summary statistics:

Column	n	Mean	Std. dev.	Variance	Range	Min	Max
Support	992	67.75504	26.870724	722.0358	155	0	155

a. Locate the range on the printout. Comment on the accuracy of the statement: "The difference between the largest and smallest values of level of support for the 992 senior managers is 155 points."
b. Locate the variance on the printout. Comment on the accuracy of the statement: "On average, the level of support for corporate sustainability for the 992 senior managers is 722 points."
c. Locate the standard deviation on the printout. Does the distribution of support levels for the 992 senior managers have more or less variation than another distribution with a standard deviation of 50? Explain.
d. Which measure of variation best describes the distribution of 992 support levels? Explain.

2.68 Is honey a cough remedy? Refer to the *Archives of Pediatrics and Adolescent Medicine* (December 2007) study of honey as a remedy for coughing, Exercise 2.31 (p. 80). The coughing improvement scores (as determined by the children's parents) for the patients in the over-the-counter cough medicine dosage (DM) group, honey dosage group, and control group are reproduced in the accompanying table.

HCOUGH

Honey Dosage:	12	11	15	11	10	13	10	4	15	16	9			
	14	10	6	10	8	11	12	12	8					
	12	9	11	15	10	15	9	13	8	12	10			
	8	9	5	12										
DM Dosage:	4	6	9	4	7	7	7	9	12	10	11	6		
	3	4	9	12	7	6	8	12	12	4	12			
	13	7	10	13	9	4	4	10	15	9				
No Dosage (Control):	5	8	6	1	0	8	12	8	7	7	1	6		
	7	7	12	7	9	7	9	5	11	9	5			
	6	8	8	6	7	10	9	4	8	7	3	1	4	3

Source: Based on I. M. Paul et al., "Effect of Honey, Dextromethorphan, and No Treatment on Nocturnal Cough and Sleep Quality for Coughing Children and Their Parents," *Archives of Pediatrics and Adolescent Medicine*, Vol. 161, No. 12, December 2007 (data simulated).

a. Find the standard deviation of the improvement scores for the honey dosage group.
b. Find the standard deviation of the improvement scores for the DM dosage group.

c. Find the standard deviation of the improvement scores for the control group.
d. Based on the results, parts **a–c,** which group appears to have the most variability in coughing improvement scores? The least variability?

2.69 TRAFFIC **Traffic.** Refer to Exercise 2.54 and the data on the traffic counts and accidents that happened in Bloomington during 2019.
a. Find the range, variance, and standard deviation of this data set.
b. Eliminate the largest value from the data set and repeat part **a.** What effect does dropping this measurement have on the measures of variation found in part **a?**
c. Eliminate the smallest and largest values from the data set and repeat part **a.** What effect does dropping both of these measurements have on the measures of variation found in part **a?**

Applying the Concepts—Advanced

2.70 TIMES **Estimating production time.** A widely used technique for estimating the length of time it takes workers to produce a product is the **time study.** In a time study, the task to be studied is divided into measurable parts, and each is timed with a stopwatch or filmed for later analysis. For each worker, this process is repeated many times for each subtask. Then the average and standard deviation of the time required to complete each subtask are computed for each worker. A worker's overall time to complete the task under study is then determined by adding his or her subtask-time averages (Gaither and Frazier, *Operations Management*, 2001). The data (in minutes) given in the table are the result of a time study of a production operation involving two subtasks.

	Worker A		Worker B	
Repetition	Subtask 1	Subtask 2	Subtask 1	Subtask 2
1	30	2	31	7
2	28	4	30	2
3	31	3	32	6
4	38	3	30	5
5	25	2	29	4
6	29	4	30	1
7	30	3	31	4

a. Find the overall time it took each worker to complete the manufacturing operation under study.
b. For each worker, find the standard deviation of the seven times for subtask 1.
c. In the context of this problem, what are the standard deviations you computed in part **b** measuring?
d. Repeat part **b** for subtask 2.
e. If you could choose workers similar to A or workers similar to B to perform subtasks 1 and 2, which type would you assign to each subtask? Explain your decisions on the basis of your answers to parts **a–d.**

2.5 Using the Mean and Standard Deviation to Describe Data

We've seen that if we are comparing the variability of two samples selected from a population, the sample with the larger standard deviation is the more variable of the two. Thus, we know how to interpret the standard deviation on a relative or comparative basis, but we haven't explained how it provides a measure of variability for a single sample.

To understand how the standard deviation provides a measure of variability of a data set, consider a specific data set and answer the following questions: How many measurements are within 1 standard deviation of the mean? How many measurements are within 2 standard deviations? For a specific data set, we can answer these questions by counting the number of measurements in each of the intervals. However, if we are interested in obtaining a general answer to these questions, the problem is more difficult.

Rules 2.1 and 2.2 give two sets of answers to the questions of how many measurements fall within 1, 2, and 3 standard deviations of the mean. The first, which applies to *any* set of data, is derived from a theorem proved by the Russian mathematician P. L. Chebyshev. The second, which applies to **mound-shaped, symmetric distributions** of data (where the mean, median, and mode are all about the same), is based upon empirical evidence that has accumulated over the years. However, the percentages given for the intervals in Rule 2.2 provide remarkably good approximations even when the distribution of the data is slightly skewed or asymmetric. Note that both rules apply to either population data sets or sample data sets.

Rule 2.1: Using the Mean and Standard Deviation to Describe Data: Chebyshev's Rule

Chebyshev's Rule applies to *any data set*, regardless of the shape of the frequency distribution of the data.

a. No useful information is provided on the fraction of measurements that fall within 1 standard deviation of the mean [i.e., within the interval $(\bar{x} - s, \bar{x} + s)$ for samples and $(\mu - \sigma, \mu + \sigma)$ for populations].

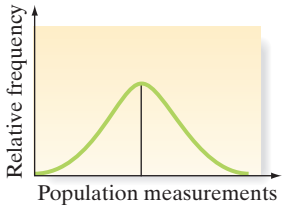
b. At least $\frac{3}{4}$ will fall within 2 standard deviations of the mean [i.e., within the interval $(\bar{x} - 2s, \bar{x} + 2s)$ for samples and $(\mu - 2\sigma, \mu + 2\sigma)$ for populations].

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

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

Rule 2.2: Using the Mean and Standard Deviation to Describe Data: The Empirical Rule

The **Empirical Rule** is a rule of thumb that applies to data sets with frequency distributions that are *mound-shaped and symmetric*, as shown below.



a. Approximately 68% of the measurements will fall within 1 standard deviation of the mean [i.e., within the interval $(\bar{x} - s, \bar{x} + s)$ for samples and $(\mu - \sigma, \mu + \sigma)$ for populations].

b. Approximately 95% of the measurements will fall within 2 standard deviations of the mean [i.e., within the interval $(\bar{x} - 2s, \bar{x} + 2s)$ for samples and $(\mu - 2\sigma, \mu + 2\sigma)$ for populations].

c. Approximately 99.7% (essentially all) of the measurements will fall within 3 standard deviations of the mean [i.e., within the interval $(\bar{x} - 3s, \bar{x} + 3s)$ for samples and $(\mu - 3\sigma, \mu + 3\sigma)$ for populations].

EXAMPLE 2.11

Interpreting the Standard Deviation—R&D Expenditures

Problem The 50 companies' percentages of revenues spent on R&D are repeated in Table 2.6. We have previously shown (see Figure 2.21, p. 95) that the mean and standard deviation of these data (rounded) are 8.49 and 1.98, respectively. Calculate the fraction of these measurements that lie within the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, and $\bar{x} \pm 3s$ and compare the results with those predicted by Rules 2.1 and 2.2.

Table 2.6 R&D Percentages for 50 Companies

13.5	9.5	8.2	6.5	8.4	8.1	6.9	7.5	10.5	13.5
7.2	7.1	9.0	9.9	8.2	13.2	9.2	6.9	9.6	7.7
9.7	7.5	7.2	5.9	6.6	11.1	8.8	5.2	10.6	8.2
11.3	5.6	10.1	8.0	8.5	11.7	7.1	7.7	9.4	6.0
8.0	7.4	10.5	7.8	7.9	6.5	6.9	6.5	6.8	9.5

Data Set: R&D

BIOGRAPHY

PAFNUTY L. CHEBYSHEV (1821–1894)
The Splendid Russian Mathematician

P. L. Chebyshev was educated in mathematical science at Moscow University, eventually earning his master's degree. Following his graduation, Chebyshev joined St. Petersburg (Russia) University as a professor, becoming part of the well-known "Petersburg mathematical school." It was here that Chebyshev proved his famous theorem about the probability of a measurement being within k standard deviations of the mean (Rule 2.1). His fluency in French allowed him to gain international recognition in probability theory. In fact, Chebyshev once objected to being described as a "splendid Russian mathematician," saying he surely was a "worldwide mathematician." One student remembered Chebyshev as "a wonderful lecturer" who "was always prompt for class," and "as soon as the bell sounded, he immediately dropped the chalk, and, limping, left the auditorium."

Solution We first form the interval

$$(\bar{x} - s, \bar{x} + s) = (8.49 - 1.98, 8.49 + 1.98) = (6.51, 10.47)$$

A check of the measurements reveals that 34 of the 50 measurements, or 68%, are within 1 standard deviation of the mean.

The next interval of interest,

$$(\bar{x} - 2s, \bar{x} + 2s) = (8.49 - 3.96, 8.49 + 3.96) = (4.53, 12.45),$$

contains 47 of the 50 measurements, or 94%.

Finally, the 3-standard-deviation interval around $\bar{x}$,

$$(\bar{x} - 3s, \bar{x} + 3s) = (8.49 - 5.94, 8.49 + 5.94) = (2.55, 14.43),$$

contains all, or 100%, of the measurements.

In spite of the fact that the distribution of these data is skewed to the right (see Figure 2.10, p. 71), the percentages within 1, 2, and 3 standard deviations (68%, 94%, and 100%) agree very well with the approximations of 68%, 95%, and 99.7% given by the Empirical Rule (Rule 2.2).

Look Back You will find that unless the distribution is extremely skewed, the mound-shaped approximations will be reasonably accurate. Of course, no matter what the shape of the distribution, Chebyshev's Rule (Rule 2.1) ensures that at least 75% and at least 89% of the measurements will lie within 2 and 3 standard deviations of the mean, respectively.

Now Work Exercise 2.74

EXAMPLE 2.12

Check on the Calculation of s—R&D Expenditures

Problem Chebyshev's Rule and the Empirical Rule are useful as a check on the calculation of the standard deviation. For example, suppose we calculated the standard deviation for the R&D percentages (Table 2.6) to be 3.92. Are there any "clues" in the data that enable us to judge whether this number is reasonable?

Solution The range of the R&D percentages in Table 2.6 is $13.5 - 5.2 = 8.3$. From Chebyshev's Rule and the Empirical Rule we know that most of the measurements (approximately 95% if the distribution is mound-shaped) will be within 2 standard deviations of the mean. And, regardless of the shape of the distribution and the number of measurements, almost all of them will fall within 3 standard deviations of the mean. Consequently, we would expect the range of the measurements to be between 4 (i.e., $\pm 2s$) and 6 (i.e., $\pm 3s$) standard deviations in length (see Figure 2.22).

For the R&D data, this means that s should fall between

$$\frac{\text{Range}}{6} = \frac{8.3}{6} = 1.38 \quad \text{and} \quad \frac{\text{Range}}{4} = \frac{8.3}{4} = 2.08$$

In particular, the standard deviation should not be much larger than $\frac{1}{4}$ of the range, particularly for the data set with 50 measurements. Thus, we have reason to believe that the calculation of 3.92 is too large. A check of our work reveals that 3.92 is the variance s^2, not the standard deviation s (see Example 2.10). We "forgot" to take the square root (a common error); the correct value is $s = 1.98$. Note that this value is between $\frac{1}{6}$ and $\frac{1}{4}$ of the range.

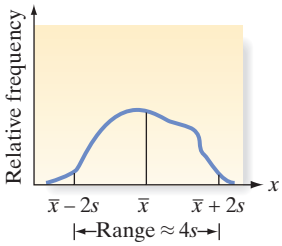


Figure 2.22
The relation between the range and the standard deviation

Look Ahead In examples and exercises we'll sometimes use $s \approx$ range/4 to obtain a crude, and usually conservatively large, approximation for s. However, we stress that this is no substitute for calculating the exact value of s when possible.

Now Work Exercise 2.75

In the next example, we use the concepts in Chebyshev's Rule and the Empirical Rule to build the foundation for statistical inference making.

EXAMPLE 2.13

Making a Statistical Inference—Car Battery Guarantee

Problem A manufacturer of automobile batteries claims that the average length of life for its grade A battery is 60 months. However, the guarantee on this brand is for just 36 months. Suppose the standard deviation of the life length is known to be 10 months, and the frequency distribution of the life-length data is known to be mound-shaped.

a. Approximately what percentage of the manufacturer's grade A batteries will last more than 50 months, assuming the manufacturer's claim is true?

b. Approximately what percentage of the manufacturer's batteries will last less than 40 months, assuming the manufacturer's claim is true?

c. Suppose your battery lasts 37 months. What could you infer about the manufacturer's claim?

Solution If the distribution of life length is assumed to be mound-shaped with a mean of 60 months and a standard deviation of 10 months, it would appear as shown in Figure 2.23. Note that we can take advantage of the fact that mound-shaped distributions are (approximately) symmetric about the mean, so that the percentages given by the Empirical Rule can be split equally between the halves of the distribution on each side of the mean.

For example, because approximately 68% of the measurements will fall within 1 standard deviation of the mean, the distribution's symmetry implies that approximately $\frac{1}{2}(68\%) = 34\%$ of the measurements will fall between the mean and 1 standard deviation on each side. This concept is illustrated in Figure 2.23. The figure also shows that 2.5% of the measurements lie beyond 2 standard deviations in each direction from the mean. This result follows from the fact that if approximately 95% of the measurements fall within 2 standard deviations of the mean, then about 5% fall outside 2 standard deviations; if the distribution is approximately symmetric, then about 2.5% of the measurements fall beyond 2 standard deviations on each side of the mean.

a. It is easy to see in Figure 2.23 that the percentage of batteries lasting more than 50 months is approximately 34% (between 50 and 60 months) plus 50% (greater than 60 months). Thus, approximately 84% of the batteries should have life length exceeding 50 months.

b. The percentage of batteries that last less than 40 months can also be easily determined from Figure 2.23. Approximately 2.5% of the batteries should fail prior to 40 months, assuming the manufacturer's claim is true.

c. If you are so unfortunate that your grade A battery fails at 37 months, you can make one of two inferences: either your battery was one of the approximately 2.5% that fail prior to 40 months, or something about the manufacturer's claim is not true. Because the chances are so small that a battery fails before 40 months, you would have good reason to have serious doubts about the manufacturer's claim. A mean smaller than 60 months and/or a standard deviation longer than 10 months would both increase the likelihood of failure prior to 40 months.*

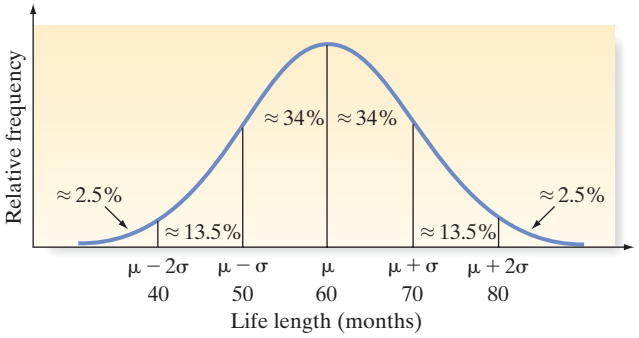


Figure 2.23
Battery life-length distribution: Manufacturer's claim assumed true

*The assumption that the distribution is mound-shaped and symmetric may also be incorrect. However, if the distribution were skewed to the right, as life-length distributions often tend to be, the percentage of measurements more than 2 standard deviations *below* the mean would be even less than 2.5%.

Look Back The approximations given in Figure 2.23 are more dependent on the assumption of a mound-shaped distribution than those given by the Empirical Rule (Rule 2.2) because the approximations in Figure 2.23 depend on the (approximate) symmetry of the mound-shaped distribution. We saw in Example 2.11 that the Empirical Rule can yield good approximations even for skewed distributions. This will *not* be true of the approximations in Figure 2.23; the distribution *must* be mound-shaped and approximately symmetric.

Example 2.13 is our initial demonstration of the statistical inference-making process. At this point you should realize that we'll use sample information (in Example 2.13, your battery's failure at 37 months) to make inferences about the population (in Example 2.13, the manufacturer's claim about the life length for the population of all batteries). We'll build on this foundation as we proceed.

STATISTICS IN ACTION REVISITED

Interpreting Numerical Descriptive Measures

We return to the *Journal of Experimental Social Psychology* (Vol. 45, 2009) study on whether money can buy love for two groups of participants—those assigned the role of gift-giver and those assigned the role of gift-recipient. Recall that the researchers investigated whether givers and receivers differ on the price of the birthday gift reported and on the overall level of appreciation reported. The Minitab descriptive statistics printout for the **BUYLOV** data is displayed in Figure SIA2.5, with means and standard deviations highlighted.

First, we focus on the quantitative variable, *birthday gift price*. The sample mean gift price for givers is \$105.84 compared with \$149.00 for receivers. Our interpretation is that receivers report a higher average gift price than givers—a difference of about \$43.

To interpret the gift price standard deviations (93.47 for givers and 134.5 for receivers), we substitute into the formula, $\bar{x} \pm 2s$, to obtain the following intervals:

Gift-Giver: $\bar{x} \pm 2s = 105.84 \pm 2(93.47) = 105.84 \pm 186.94 = (-81.1, 292.78)$
Gift-Receiver: $\bar{x} + 2s = 149.00 \pm 2(134.50) = 149 \pm 269 = (-120, 418)$

Because gift price cannot have a negative value, the two intervals for givers and receivers are more practically given as (0, 293) and (0, 418), respectively. Since the distributions of gift price are not mound-shaped and symmetric (see Figure SIA2.4a), we apply Chebyshev's Rule (Rule 2.1). Thus, we know that at least 75% of the gift-givers in the study reported a gift price between \$0 and \$293, and at least 75% of the gift-recipients reported a gift price between \$0 and \$418. You can see that the upper endpoint of the interval for givers lies below that for receivers. Consequently, we can infer that prices reported by gift-recipients tend to be higher than the prices reported by gift-givers. Also, if a gift price of \$400 is observed, it is much more likely to be reported by a gift-receiver than a gift-giver.

A similar analysis performed for the variable *overall level of appreciation* yielded the following intervals:

Gift-Giver: $\bar{x} \pm 2s = 4.985 \pm 2(2.775) = 4.985 \pm 5.55 = (-.565, 10.535)$
Gift-Receiver: $\bar{x} \pm 2s = 7.165 \pm 2(2.928) = 7.165 \pm 5.856 = (1.309, 13.021)$

Since overall level of appreciation cannot have a value less than 2 and is a whole number, the two intervals for givers and receivers are more practically given as (2, 10) and

Statistics

Variable	Role	N	Mean	StDev	Variance	Minimum	Median	Maximum
BGiftPrice	Giver	134	105.84	93.47	8736.78	2.00	75.50	431.00
	Receiver	103	149.0	134.5	18083.8	1.0	133.0	548.0
OverallApp	Giver	134	4.985	2.775	7.699	2.000	4.000	13.000
	Receiver	103	7.165	2.928	8.571	2.000	7.000	14.000

Figure SIA2.5
Minitab descriptive statistics for gift price and appreciation level, by role

STATISTICS IN ACTION REVISTED (*continued*)

(2, 13), respectively. Applying Chebyshev's Rule, we know that at least 75% of the gift-givers in the study reported an appreciation level between 2 and 10 points, and at least 75% of the gift-recipients reported an appreciation level between 2 and 13 points. Again, the upper endpoint of the interval for givers lies below that for receivers; thus, we infer that overall levels of appreciation reported by gift-recipients tend to be higher than those reported by gift-givers.

Now, how does this information help the researchers determine whether there are "significant" differences in the means for gift-givers and gift-recipients? In Chapters 6 and 7 we present inferential methods that will answer such a question and provide a measure of reliability for the inference.

Data Set: BUYLOV

Exercises 2.71–2.89

Learning the Mechanics

2.71 The output from a statistical software package indicates that the mean and standard deviation of a data set consisting of 200 measurements are \$1,500 and \$300, respectively.

a. What are the units of measurement of the variable of interest? Based on the units, what type of data is this: quantitative or qualitative?

b. What can be said about the number of measurements between \$900 and \$2,100? Between \$600 and \$2,400? Between \$1,200 and \$1,800? Between \$1,500 and \$2,100?

2.72 For any set of data, what can be said about the percentage of the measurements contained in each of the following intervals?

a. $\bar{x} - s$ to $\bar{x} + s$

b. $\bar{x} - 2s$ to $\bar{x} + 2s$

c. $\bar{x} - 3s$ to $\bar{x} + 3s$

2.73 For a set of data with a mound-shaped relative frequency distribution, what can be said about the percentage of the measurements contained in each of the intervals specified in Exercise 2.72?

2.74 NW L02074 The following is a sample of 25 measurements:

7	6	6	11	8	9	11	9	10	8	7	7	
5	9	10	7	7	7	7	9	12	10	10	8	6

a. Compute $\bar{x}$, s^2,, and s for this sample.

b. Count the number of measurements in the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$. Express each count as a percentage of the total number of measurements.

c. Compare the percentages found in part **b** to the percentages given by the Empirical Rule and Chebyshev's Rule.

d. Calculate the range and use it to obtain a rough approximation for s. Does the result compare favorably with the actual value for s found in part **a?**

2.75 NW Given a data set with a largest value of 760 and a smallest value of 135, what would you estimate the standard deviation to be? Explain the logic behind the procedure you used to estimate the standard deviation. Suppose the standard deviation is reported to be 25. Is this feasible? Explain.

Applying the Concepts—Basic

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

a. For a sag distribution with any shape, what proportion of transformers will have between 263 and 443 sags per week? Which rule did you apply and why?

b. For a sag distribution that is mound-shaped and symmetric, what proportion of transformers will have between 263 and 443 sags per week? Which rule did you apply and why?

c. For a swell distribution with any shape, what proportion of transformers will have between 109 and 259 swells per week? Which rule did you apply and why?

d. For a swell distribution that is mound-shaped and symmetric, what proportion of transformers will have between 109 and 259 swells per week? Which rule did you apply and why?

2.77 STONE **Permeability of sandstone during weathering.** Refer to the *Geographical Analysis* (Vol. 42, 2010) study of the decay properties of sandstone when exposed to the weather, Exercises 2.47 and 2.65 (pp. 89 and 97). Recall that slices of sandstone blocks were measured for permeability under three conditions: no exposure to any type of weathering (A), repeatedly sprayed with a 10% salt solution (B), and soaked in a 10% salt solution and dried (C).

a. Combine the mean (from Exercise 2.47) and standard deviation (from Exercise 2.65) to make a statement about where most of the permeability measurements for Group A sandstone slices will fall. Which rule did you use to make this inference and why?

b. Repeat part **a** for Group B sandstone slices.

c. Repeat part **a** for Group C sandstone slices.
d. Based on all your analyses, which type of weathering (type A, B, or C) appears to result in faster decay (i.e., higher permeability measurements)?

2.78 **Do social robots walk or roll?** Refer to the *International Conference on Social Robotics* (Vol. 6414, 2010) study on the current trend in the design of social robots, Exercise 2.5 (p. 66). Recall that in a random sample of social robots obtained through a Web search, 28 were built with wheels. The number of wheels on each of the 28 robots is listed in the accompanying table.

ROBOTS

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

Source: Based on S. Chew et al., "Do Social Robots Walk or Roll?" *International Conference on Social Robotics*, Vol. 6414, 2010 (adapted from Figure 2).

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

2.79 **Parents Against Watching Television.** A society called Parents Against Watching Television (PAWT) is primarily concerned with the amount of television viewed by today's youth. It asked 300 parents of elementary school aged children to estimate the number of hours their child spent watching television in any given week. The mean and the standard deviation for their responses were 17 and 3, respectively. PAWT then constructed a stem-and-leaf display for the data, which showed that the distribution of the number of hours was a symmetric, mound-shaped distribution. Identify the interval where you believe approximately 95% of the television viewing times fell in the distribution.

2.80 **House prices.** An article in the *Journal of Business, Finance, and Accounting* (Vol. 33, pp 1535–55, January 2020) predicts downside risks to future real house price growth in 22 advanced economies and 10 emerging market economies. The data was collected from South Africa, North America (Canada, Mexico, and U.S.), South America (Brazil, Chile, and Colombia), Europe (16 countries including Austria, Belgium, France, Germany, Ireland, Italy, Netherlands, four Nordic countries, Spain, Switzerland, U.K., as well as Russia and Turkey), and Asia-Pacific (Australia, China, Hong Kong SAR, Japan, Malaysia, New Zealand, and Singapore) from the early 1990s to early 2018. The advanced economies have a sample size of 2384 quarterly observations on the change rate in real house prices, and the mean and standard deviations are 0.48% and 2.36% respectively. Concurrently, the emerging market economies have a sample size of 960 quarterly observations on the change rate in real house prices, and the mean and standard deviations are 0.63% and 3.02% respectively. Assume the distributions for both economics of real house prices quarterly change rate are bell-shaped and symmetric.
a. Give a range of real house prices quarterly change rates that will contain about 95% of quarterly change rates in advanced economies.
b. Give a range of real house prices quarterly change rates that will contain about 95% of quarterly change rates in emerging market economies.
c. What proportion of the real house prices quarterly have change rate in emerging market economies below −5.41%?

Applying the Concepts—Intermediate

2.81 **Cruise ships sanitation scores.** Refer to the *Centers for Disease Control and Prevention* (CDC) listing of the January 2019 to October 2020 sanitation scores for 129 cruise ships, Exercise 2.24 (p. 77).

CRUISE

a. Find the mean and standard deviation of the sanitation scores.
b. Calculate the intervals $\bar{x} \pm s$, $\bar{x} \pm 2s$, $\bar{x} \pm 3s$.
c. Find the percentage of measurements in the data set that fall within each of the intervals in part **b.** Do these percentages agree with Chebyshev's Rule? The Empirical Rule?

2.82 **Gross domestic product.** Refer to the *Economy 2020* data on the estimated Gross Domestic Product (GDP) (in $ millions) in 2018 for the 12 industries, Exercise 2.30 (p. 79). Use the data provided to work out an interval that will most likely represent the actual GDP over a year.

17,167.5	668.4	1,553.5	973.6	4,413.2	1,007.2
1,797.6	2,633.9	702.0	570.7	2,350.0	497.5

2.83 **Auditing water resources in Australia.** Australia has developed a General Purpose Water Accounting (GPWA) reporting system in an effort to provide a financial accounting of water use in the country. The perceptions of potential users of the reports (e.g., water resource managers) were investigated in *Accounting, Auditing & Accountability Journal* (Vol. 29, 2016). Each in a sample of 36 potential users of GPWA reports was asked to complete a survey. Two key survey questions (with possible responses) were as follows:

Q1: Should there be national standards for water reporting? (Yes, No, or Undecided)

Q2: How useful will the GPWA reports be for water users? (1- to 5-point scale, where 1 = not useful at all and 5 = very useful)

The data (simulated from results reported in the journal article) for the 36 potential users are listed in the next table. For those users who believe there should be national standards, give a range that is likely to contain the user's answer to Q2.

GPWA

Data for Exercise 2.83

User	Q1	Q2
1	Yes	5
2	Yes	4
3	Yes	4
4	Yes	3
5	Undecided	2
6	Yes	4
7	Yes	4
8	Yes	4
9	Yes	4
10	Yes	5
11	Yes	4
12	Yes	4
13	Yes	3
14	Yes	3
15	Yes	3
16	Yes	4
17	Yes	3
18	Yes	3
19	Yes	5
20	Yes	4
21	Undecided	5
22	Yes	4
23	No	2
24	Yes	5
25	Yes	5
26	Yes	5
27	Yes	3
28	Yes	4
29	Yes	5
30	Undecided	5
31	Undecided	5
32	Yes	4
33	Yes	5
34	Undecided	5
35	Yes	5
36	Yes	3

2.84 ***The Apprentice* contestants' performance ratings.** Refer to the *Significance* (April 2015) study of contestants' performance on the TV show *The Apprentice*, Exercise 2.9 (p. 67). Recall that each of 159 contestants was rated (on a 20-point scale) based on their performance. The accompanying Minitab printout gives the mean and standard deviation of the contestant ratings, categorized by highest degree obtained (no degree, first degree, or postgraduate degree) and prize (job or partnership with Lord Sugar).

APPRENTICE

Results for Prize = Job

Statistics

Variable	Degree	N	Mean	StDev	Minimum	Maximum
Rating	First	54	7.796	4.231	1.000	17.000
	None	35	7.457	4.388	1.000	20.000
	Post	10	9.80	4.54	2.00	17.00

Results for Prize = Partner

Statistics

Variable	Degree	N	Mean	StDev	Minimum	Maximum
Rating	First	33	8.212	4.775	1.000	20.000
	None	21	10.62	4.83	3.00	20.00
	Post	6	6.50	3.33	2.00	12.00

a. Give a practical interpretation of the mean rating for contestants with a first (bachelor's) degree who competed for a job with Lord Sugar.

b. Find an interval that captures about 95% of the ratings for contestants with a first (bachelor's) degree who competed for a job with Lord Sugar.

c. An analysis of the data led the researchers to conclude that "when the reward for winning . . . was a job, more academically qualified contestants tended to perform less well; however, this pattern is reversed when the prize changed to a business partnership." Do you agree? Explain.

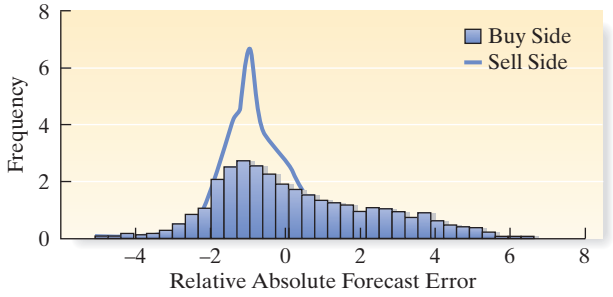
2.85 **Shopping vehicle and judgment.** While shopping at the grocery store, are you more likely to buy a vice product (e.g., a candy bar) when pushing a shopping cart or when carrying a shopping basket? This was the question of interest in a study published in the *Journal of Marketing Research* (December 2011). The researchers believe that when your arm is flexed (as when carrying a basket), you are more likely to choose a vice product than when your arm is extended (as when pushing a cart). To test this theory in a laboratory setting, the researchers recruited 22 consumers and asked each to push his or her hand against a table while being asked a series of shopping questions. Half of the consumers were told to put their arms in a flex position (similar to a shopping basket) and the other half were told to put their arms in an extended position (similar to a shopping cart). Participants were offered several choices between a vice and a virtue (e.g., a movie ticket vs. a shopping coupon, paying later with a larger amount vs. paying now), and a choice score (on a scale of 0 to 100) was determined for each. (Higher scores indicate a greater preference for vice options.) The average choice score for consumers with a flexed arm was 59, while the average for consumers with an extended arm was 43.

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

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

2.86 **Buy-side vs. sell-side analysts' earnings forecasts.** Financial analysts who make forecasts of stock prices and recommendations about whether to buy, sell, or hold specific securities can be categorized as either "buy-side" analysts or "sell-side" analysts. A group of Harvard Business School professors compared earnings forecasts of buy-side and sell-side analysts (*Financial Analysts Journal*, July/August 2008). Data were collected on 3,526 forecasts made by buy-side analysts and 58,562 forecasts made by sell-side analysts, and the relative absolute forecast error was determined for each.

a. Frequency distributions for buy-side and sell-side analysts forecast errors (with the sell-side distribution superimposed over the buy-side distribution) are shown in the accompanying figure. Based on the figure, the researchers concluded "that absolute forecast errors for buy-side analysts have a higher mean and variance than those for the sell-side analysts." Do you agree? Explain.



b. The mean and standard deviation of forecast errors for both buy-side and sell-side analysts are given in the following table. For each type of analyst, provide an interval that will contain approximately 95% of the forecast errors. Compare these intervals. Which type of analyst is likely to have a relative forecast error of +2.00 or higher?

	Buy-Side Analysts	Sell-Side Analysts
Mean	0.85	−0.05
Standard deviation	1.93	0.85

Source: Based on B. Groysberg, P. Healy, and C. Chapman, *Financial Analysis Journal,* Vol. 64, No. 4. July/August 2008 (Table 2).

Applying the Concepts—Advanced

2.87 **Land purchase decision.** A buyer for a lumber company must decide whether to buy a piece of land containing 5,000 pine trees. If 1,000 of the trees are at least 40 feet tall, the buyer will purchase the land; otherwise, he won't. The owner of the land reports that the height of the trees has a mean of 30 feet and a standard deviation of 3 feet. Based on this information, what is the buyer's decision?

2.88 **Delivery times for online orders.** Refer to the *Journal of Marketing Research* (Oct., 2019) study of delivery times for online orders, Exercise 1.23 (p. 46). Recall that a major apparel retailer originally fulfilled all its online orders from a single distribution center (DC) located in the Eastern U.S. Later that year, the retailer opened a second DC located in the Western U.S. The researchers collected data on delivery times (in business days) for a sample of online orders made by customers near the Eastern distribution center. Summary statistics for those orders filled by the Eastern DC and the Western DC are shown in the next table. Suppose that one of online orders is randomly selected and found to have a delivery time of 5 days. Is this order more likely to have been filled by the Eastern DC or the Western DC? Explain.

	Mean Number of Days	Standard Deviation
From Eastern DC	5.22	.77
From Western DC	6.95	.55

Source: Fisher, M.L., et al. "The Value of Rapid Delivery in Omnichannel Retailing", *Journal of Marketing Research,* Vol. 56, No. 5, October 2019 (Table 3).

2.89 HRVST **Monitoring harvest.** Corn can take from 60 to 100 days to reach harvest depending on the variety and the temperature during the growing season. On average, well-grown corn will produce one bushel of corn that weighs 56 pounds with a standard deviation of .25 pounds. The corn farmer monitors his harvest every two months. If one bushel of corn weighs less than two standard deviations from the mean (using the mean and standard deviation given above), the farmer needs to add extra fertilizers to improve the soil fertility and maintain the amount of nutrient-rich organic matter. The data given in the following table are the weights of one bushel of corn measured by the farmer over the past two and half years. Assume that the farmer has never added extra fertilizers to improve the soil quality before February 2019. By using the provided data, justify when the farmer adds extra fertilizers to improve the soil quality.

Month	Weight (pounds)
Feb 2019	55.99
Apr	56.68
Jun	55.97
Aug	56.15
Oct	57.45
Dec	54.85
Feb 2020	56.06
Apr	54.10
Jun	56.67
Aug	54.57
Oct	55.70
Dec	55.59
Feb 2021	55.15
Apr	56.51
Jun	54.71

2.6 Numerical Measures of Relative Standing

We've seen that numerical measures of central tendency and variability describe the general nature of a quantitative data set (either a sample or a population). In addition, we may be interested in describing the *relative* quantitative location of a particular measurement within a data set. Descriptive measures of the relationship of a measurement to the rest of the data are called **measures of relative standing.**

One measure of the relative standing of a measurement is its **percentile ranking,** or **percentile score**. For example, if oil company A reports that its yearly sales are in the 90th percentile of all companies in the industry, the implication is that 90% of all oil companies have yearly sales less than company A's, and only 10% have yearly sales

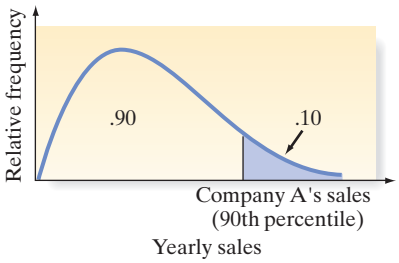


Figure 2.24
Location of 90th percentile for yearly sales of oil companies

exceeding company A's. This is demonstrated in Figure 2.24. Similarly, if the oil company's yearly sales are in the 50th percentile (the median of the data set), 50% of all oil companies would have lower yearly sales and 50% would have higher yearly sales.

Percentile rankings are of practical value only for large data sets. Finding them involves a process similar to the one used in finding a median. The measurements are ranked in order, and a rule is selected to define the location of each percentile. Because we are primarily interested in interpreting the percentile rankings of measurements (rather than finding particular percentiles for a data set), we define the *pth percentile* of a data set.

For any set of n measurements (arranged in ascending or descending order), the ***pth* percentile** is a number such that p% of the measurements fall below the *pth* percentile and $(100 - p)$% fall above it.

EXAMPLE 2.14

Finding and Interpreting Percentiles—R&D Expenditures

Problem Refer to the percentages spent on R&D by the 50 high-technology firms listed in Table 2.6 (p. 99). A portion of the XLSTAT descriptive statistics printout is shown in Figure 2.25. Locate the 10th percentile and 95th percentile on the printout and interpret these values.

Solution Both the 10th percentile and 95th percentile are highlighted on the XLSTAT printout, Figure 2.25. These values are 6.5 and 13.2, respectively. Our interpretations are as follows: 10% of the 50 R&D percentages fall below 6.5 and 95% of the R&D percentages fall below 13.2.

Look Back The method for computing percentiles with small data sets varies according to the software used. As the sample size increases, these percentile values from the different software packages will converge to a single number.

Summary statistics:

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
RDPct	50	5.2000	13.5000	8.4920	1.9806

Percentile table (Empirical distribution function):

Percentile	Value
Maximum 100%	13.5000
99%	13.5000
95%	13.2000
90%	11.1000
3rd Quartile 75%	9.6000
Median 50%	8.0000
1st Quartile 25%	7.1000
10%	6.5000
5%	5.9000
1%	5.6000
Minimum 0%	5.2000

Figure 2.25
XLSTAT percentiles for 50 R&D percentages

Now Work Exercise 2.91

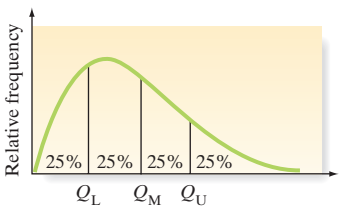


Figure 2.26
The quartiles for a data set

Percentiles that partition a data set into four categories, each category containing exactly 25% of the measurements, are called **quartiles.** The *lower quartile* (Q_L) is the 25th percentile, the *middle quartile* (Q_M) is the median or 50th percentile, and the *upper quartile* (Q_U) is the 75th percentile, as shown in Figure 2.26. Therefore, in Example 2.14, we have (from the XLSTAT printout, Figure 2.25), $Q_L = 7.1$, $Q_M = 8$, and $Q_U = 9.6$. Quartiles will prove useful in finding unusual observations in a data set (Section 2.7).

The **lower quartile (Q_L)** is the 25th percentile of a data set. The **middle quartile (Q_M)** is the median or 50th percentile. The **upper quartile (Q_U)** is the 75th percentile.

Another measure of relative standing in popular use is the **z-score.** As you can see in the definition of z-score below, the z-score makes use of the mean and standard deviation of the data set in order to specify the relative location of a measurement. Note that the z-score is calculated by subtracting $\bar{x}$ (or μ) from the measurement x and then dividing the result by s (or σ). The final result, the z-score, represents the distance between a given measurement x and the mean, expressed in standard deviations.

The **sample z-score** for a measurement x is

$$z = \frac{x - \bar{x}}{s}$$

The **population z-score** for a measurement x is

$$z = \frac{x - \mu}{\sigma}$$

EXAMPLE 2.15

Finding a z-Score—GMAT Results

Problem A random sample of 2,000 students who sat for the Graduate Management Admission Test (GMAT) is selected. For this sample, the mean GMAT score is $\bar{x} = 540$ points and the standard deviation is $s = 100$ points. One student from the sample, Kara Smith, had a GMAT score of $x = 440$ points. What is Kara's sample z-score?

Solution First, note that Kara's GMAT score lies below the mean score for the 2,000 students (see Figure 2.27). Now we compute

$$z = (x - \bar{x})/s = (440 - 540)/100 = -100/100 = -1.0$$

This z-score implies that Kara Smith's GMAT score is 1.0 standard deviations below the sample mean GMAT score, or, in short, her sample z-score is -1.0.

Look Back The numerical value of the z-score reflects the relative standing of the measurement. A large positive z-score implies that the measurement is larger than almost all other measurements, whereas a large (in magnitude) negative z-score indicates that the measurement is smaller than almost every other measurement. If a z-score is 0 or near 0, the measurement is located at or near the mean of the sample or population.

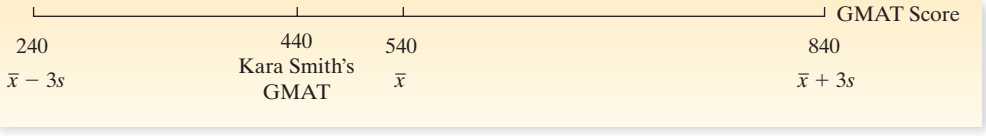


Figure 2.27
GMAT scores of a sample of test takers

Now Work Exercise 2.90

If we know that the frequency distribution of the measurements is mound-shaped, the following interpretation of the z-score can be given.

Interpretation of z-Scores for Mound-Shaped Distributions of Data

1. Approximately 68% of the measurements will have a z-score between -1 and 1.
2. Approximately 95% of the measurements will have a z-score between -2 and 2.
3. Approximately 99.7% (almost all) of the measurements will have a z-score between -3 and 3.

Note that this interpretation of z-scores is identical to that given by the Empirical Rule for mound-shaped distributions (Rule 2.2). The statement that a measurement falls in the interval $(\mu - \sigma)$ to $(\mu + \sigma)$ is equivalent to the statement that a measurement has a population z-score between -1 and 1 because all measurements between $(\mu - \sigma)$ and $(\mu + \sigma)$ are within 1 standard deviation of μ. These z-scores are displayed in Figure 2.28.

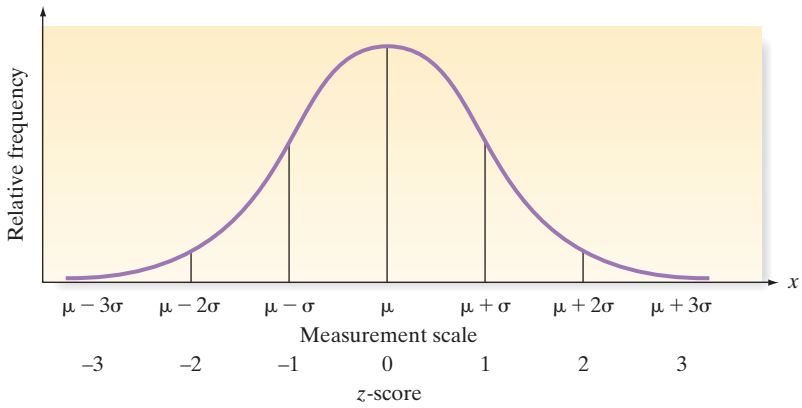


Figure 2.28
Population z-scores for a mound-shaped distribution

Exercises 2.90–2.105

Learning the Mechanics

2.90 NW Compute the z-score corresponding to each of the following values of x:
- **a.** $x = 40, s = 5, \bar{x} = 30$
- **b.** $x = 90, \mu = 89, \sigma = 2$
- **c.** $\mu = 50, \sigma = 5, x = 50$
- **d.** $s = 4, x = 20, \bar{x} = 30$
- **e.** In parts **a–d,** state whether the z-score locates x within a sample or a population.
- **f.** In parts **a–d,** state whether each value of x lies above or below the mean and by how many standard deviations.

2.91 NW Give the percentage of measurements in a data set that are above and below each of the following percentiles:
- **a.** 75th percentile
- **b.** 50th percentile
- **c.** 20th percentile
- **d.** 84th percentile

2.92 In terms of percentiles, define Q_L, Q_M, and Q_U.

2.93 Compare the z-scores to decide which of the following x values lie the greatest distance above the mean and the greatest distance below the mean.
- **a.** $x = 3, \mu = 6, \sigma = 2$
- **b.** $x = 450, \mu = 305, \sigma = 50$
- **c.** $x = 78, \mu = 36, \sigma = 35$
- **d.** $x = 33, \mu = 51, \sigma = 20$

2.94 Suppose that 40 and 90 are two elements of a population data set and that their z-scores are -2 and 3, respectively. Using only this information, is it possible to determine the population's mean and standard deviation? If so, find them. If not, explain why it's not possible.

Applying the Concepts—Basic

2.95 **Number of enrollment.** A university reports that in the last semester, the mean number of students enrolled in its business program was 354, with a 25th percentile of 280, 75th percentile of 408, and 90th percentile of 483. Interpret each of these numerical descriptive measures.

2.96 **Voltage sags and swells.** Refer to the *Electrical Engineering* (Vol. 95, 2013) study of transformer voltage sags and swells, Exercise 2.76 (p. 103). Recall that for a sample of 103 transformers built for heavy industry, the mean number of sags per week was 353 and the mean number of swells per week was 184. Assume the standard deviation of the sag distribution is 30 sags per week and the standard deviation of the swell distribution is 25 swells per week. Suppose one of the transformers is randomly selected and found to have 400 sags and 100 swells in a week.
- **a.** Find the z-score for the number of sags for this transformer. Interpret this value.
- **b.** Find the z-score for the number of swells for this transformer. Interpret this value.

2.97 **Average commute times.** A study analyzes the average commute times (in minutes) of workers aged 16 and above in 96 cities. The study reports that the mean for the average commute time is 25.41 minutes, the median is 24.35 minutes, and the 90th percentile for the average commute time is 31.2 minutes. Describe the distribution of the average commute time by interpreting each of these summary statistics.