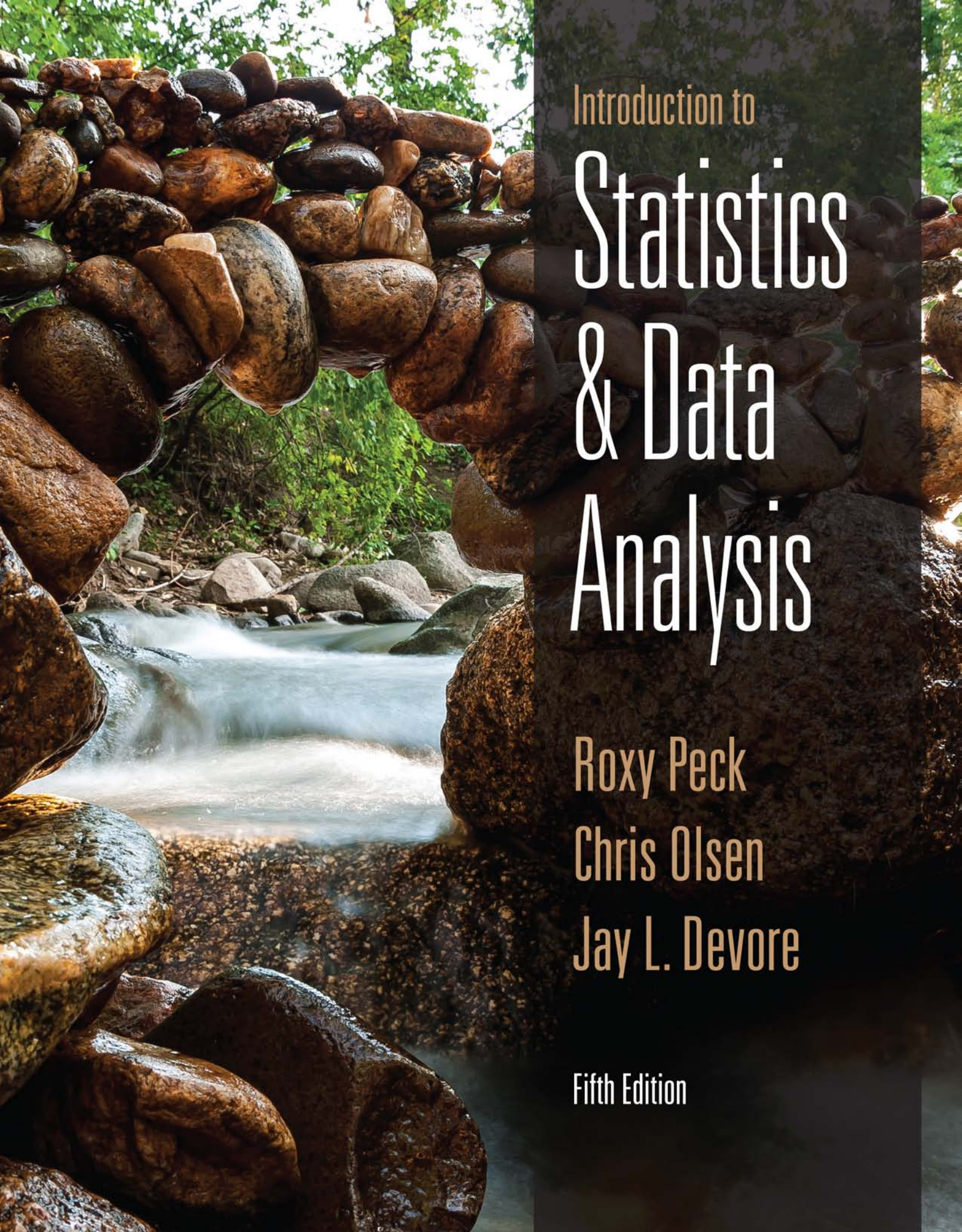


Introduction to Statistics & Data Analysis

Roxy Peck
Chris Olsen
Jay L. Devore

Fifth Edition

Introduction to Statistics and Data Analysis

EDITION 5

Introduction to Statistics and Data Analysis

Roxy Peck
California Polytechnic State University, San Luis Obispo

Chris Olsen
Grinnell College, Grinnell, Iowa

Jay L. Devore
California Polytechnic State University, San Luis Obispo

CENGAGE Learning®

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

CENGAGE Learning®

***Introduction to Statistics and Data Analysis*, Fifth Edition**
Roxy Peck, Chris Olsen, Jay L. Devore

Senior Product Team Manager: Richard Stratton
Senior Product Manager: Molly Taylor
Senior Content Developer: Jay Campbell
Associate Content Developer: Danielle Hallock
Product Assistant: Spencer Arritt
Media Developer: Andrew Coppola
Marketing Manager: Ryan Ahern
Content Project Manager: Jill Quinn
Senior Art Director: Linda May
Manufacturing Planner: Sandee Milewski
IP Analyst: Christina Ciaramella
IP Project Manager: Farah Fard
Production Service: MPS Limited
Compositor: MPS Limited
Text Designer: Dale Porter
Cover Designer: Travis Hoffman
Cover Image: © 2009-2014 Gravity Glue

Library of Congress Control Number: 2014946319

Student Edition:
ISBN: 978-1-305-11534-7

Cengage Learning
20 Channel Center Street
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at **www.cengage.com/global.**

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

To learn more about Cengage Learning Solutions, visit **www.cengage.com.**

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com.**

Printed in the United States of America
Print Number: 01 Print Year: 2014

To Lygia and Kyle

R. P.

To my wife, Sally, and my daughter, Anna

C. O.

To Carol, Allie, and Teri

J. D.

About the Authors

ROXY PECK is Emerita Associate Dean of the College of Science and Mathematics and Professor of Statistics Emerita at California Polytechnic State University, San Luis Obispo. A faculty member at Cal Poly from 1979 until 2009, Roxy served for 6 years as Chair of the Statistics Department before becoming Associate Dean, a position she held for 13 years. She received an M.S. in Mathematics and a Ph.D. in Applied Statistics from the University of California, Riverside. Roxy is nationally known in the area of statistics education, and she was presented with the Lifetime Achievement Award in Statistics Education at the U.S. Conference on Teaching Statistics in 2009. In 2003 she received the American Statistical Association's Founder's Award, recognizing her contributions to K–12 and undergraduate statistics education. She is a Fellow of the American Statistical Association and an elected member of the International Statistics Institute. Roxy served for 5 years as the Chief Reader for the Advanced Placement Statistics Exam and has chaired the American Statistical Association's Joint Committee with the National Council of Teachers of Mathematics on Curriculum in Statistics and Probability for Grades K–12 and the Section on Statistics Education. In addition to her texts in introductory statistics, Roxy is also co-editor of *Statistical Case Studies: A Collaboration Between Academe and Industry* and a member of the editorial board for *Statistics: A Guide to the Unknown,* 4th edition. Outside the classroom, Roxy likes to travel and spends her spare time reading mystery novels. She also collects Navajo rugs and heads to Arizona and New Mexico whenever she can find the time.

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

JAY L. DEVORE is Professor Emeritus of Statistics at California Polytechnic State University. He earned his undergraduate degree in Engineering Science from the University of California, Berkeley; spent a year at the University of Sheffield in England; and finished his Ph.D. in statistics at Stanford University. Jay previously taught at the University of Florida and at Oberlin College and has had visiting appointments at Stanford, Harvard, the University of Washington, New York University, and Columbia. From 1998 to 2006, he served as Chair of the Statistics Department. In addition to this book, Jay has written several widely used engineering statistics texts and a book in applied mathematical statistics. He recently coauthored a text in probability and stochastic processes. He is the recipient of a distinguished teaching award from Cal Poly, is a Fellow of the American Statistical Association, and has served several terms as an Associate Editor of the *Journal of the American Statistical Association*. In his spare time, he enjoys reading, cooking and eating good food, playing tennis, and traveling to faraway places. He is especially proud of his wife, Carol, a retired elementary school teacher; his daughter Allison, who has held several high-level positions in nonprofit organizations in Boston and New York City; and his daughter Teresa, an ESL teacher in New York City.

Brief Contents

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

Contents

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

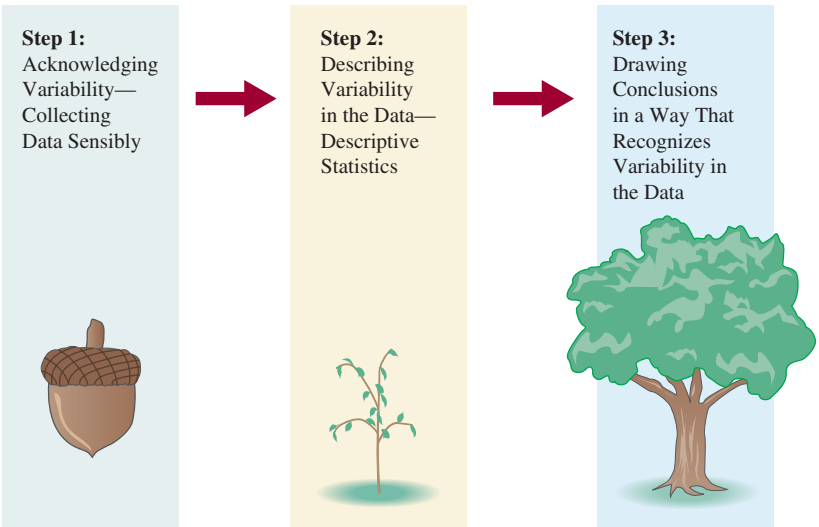
Preface

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

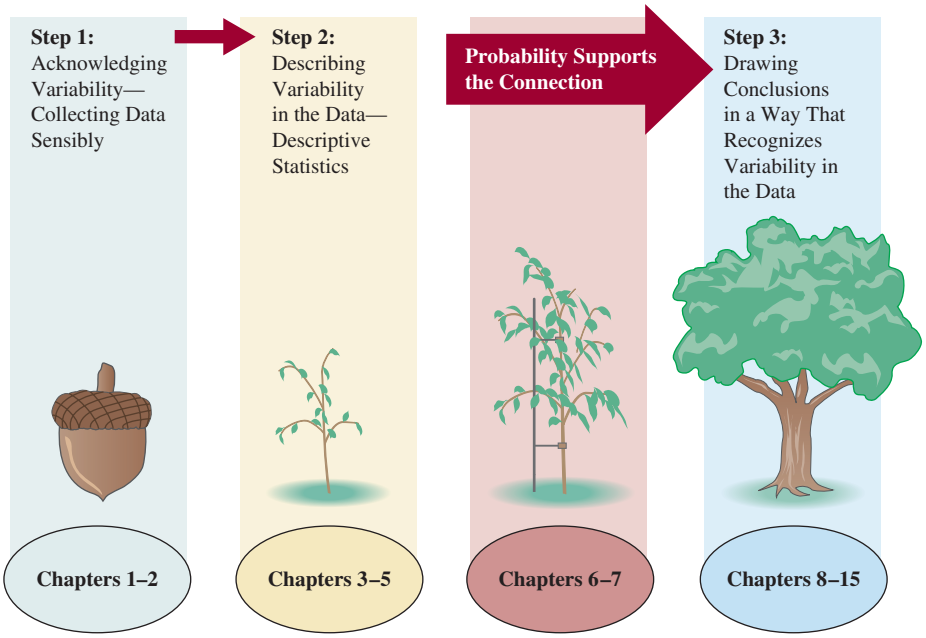
An Organization That Reflects the Data Analysis Process

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

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



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



The Importance of Context and Real Data

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

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

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

A Focus on Interpretation and Communication

Most chapters include a section titled "Interpreting and Communicating the Results of Statistical Analyses." These sections include advice on how to best communicate the results of a statistical analysis and also consider how to interpret statistical

summaries found in journals and other published sources. A subsection titled "A Word to the Wise" reminds readers of things that must be considered in order to ensure that statistical methods are employed in reasonable and appropriate ways.

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

In 2005, the American Statistical Association endorsed the report "College Guidelines in Assessment and Instruction for Statistics Education (GAISE Guidelines)," which included the following six recommendations for the introductory statistics course:

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

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

1. **Emphasize statistical literacy and develop statistical thinking.**
   Statistical literacy is promoted throughout the text in the many examples and exercises that are drawn from the popular press. In addition, a focus on the role of variability, consistent use of context, and an emphasis on interpreting and communicating results in context work together to help students develop skills in statistical thinking.
2. **Use real data.**
   The examples and exercises from *Introduction to Statistics and Data Analysis,* Fifth Edition, are context driven, and the reference sources include the popular press as well as journal articles.
3. **Stress conceptual understanding rather than mere knowledge of procedures.**
   Nearly all exercises in *Introduction to Statistics and Data Analysis,* Fifth Edition, are multipart and ask students to go beyond just computation. They focus on interpretation and communication, not just in the chapter sections specifically devoted to this topic, but throughout the text. The examples and explanations are designed to promote conceptual understanding. Hands-on activities in each chapter are also constructed to strengthen conceptual understanding. Which brings us to . . .
4. **Foster active learning in the classroom.**
   While this recommendation speaks more to pedagogy and classroom practice, *Introduction to Statistics and Data Analysis,* Fifth Edition, provides more than 30 hands-on activities in the text and additional activities in the accompanying instructor resources that can be used in class or assigned to be completed outside of class.
5. **Use technology for developing conceptual understanding and analyzing data.**
   The computer brings incredible statistical power to the desktop of every investigator. The wide availability of statistical computer packages such as Minitab,

JMP, and SPSS, and the graphical capabilities of the modern microcomputer have transformed both the teaching and learning of statistics. To highlight the role of the computer in contemporary statistics, we have included sample output throughout the book. In addition, numerous exercises contain data that can easily be analyzed using statistical software. However, access to a particular statistical package is not assumed. Technology manuals for specific software packages and for the graphing calculator are available in the online materials that accompany this text.

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

This text allows the instructor to balance the use of computers and calculators in a manner consistent with his or her philosophy and presents the power of the calculator in a series of Graphing Calculator Explorations. These are part of the online materials that accompany this text. As with computer packages, our exposition avoids assuming the use of a particular calculator and presents the calculator capabilities in a generic format. For those using a TI graphing calculator, there is a technology manual available in the online materials that accompany this text. As much as possible, the calculator explorations are independent of each other, allowing instructors to pick and choose calculator topics that are most relevant to their particular courses.

6. **Use assessments to improve and evaluate student learning.** Assessment materials in the form of a test bank, quizzes, and chapter exams are available in the instructor resources that accompany this text. The items in the test bank reflect the data-in-context philosophy of the text's exercises and examples.

Advanced Placement Statistics

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

Topic Coverage

Our book can be used in courses as short as one quarter or as long as one year in duration. Particularly in shorter courses, an instructor will need to be selective in deciding which topics to include and which to set aside. The book divides naturally into four major sections: collecting data and descriptive methods (Chapters 1–5), probability

(Chapters 6–8), the basic one- and two-sample inferential techniques (Chapters 9–12), and more advanced inferential methodology (Chapters 13–16).

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

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

A Note on Probability

The content of the probability chapters is consistent with the Advanced Placement Statistics course description. It includes both a traditional treatment of probability and probability distributions at an introductory level, as well as a section on the use of simulation as a tool for estimating probabilities. For those who prefer a briefer and more informal treatment of probability, two chapters previously from the book *Statistics: The Exploration and Analysis of Data,* by Roxy Peck and Jay Devore are available as a custom option. Please contact your Cengage learning consultant for more information about this alternative and other alternative customized options available to you.

In This Edition

Look for the following in the Fifth Edition:

- **NEW A more informal writing style.** In this revision, an effort was made to accommodate a broader range of student reading levels.
- **NEW Helpful hints in exercises.** To help students who might be having trouble getting started, hints have been added to many of the exercises directing students to relevant examples in the text.
- **NEW Changes in exercise layout.** Many of the multistep exercises carried over from the previous edition have now been broken into parts to make them more manageable and to help students organize their thinking about the solution.
- **NEW Margin notes to highlight the importance of context and the process of data analysis.** Margin notations have been added in appropriate places in the examples. These include *Understanding the context*, *Consider the data, Formulate a plan*, *Do the work*, and *Interpret the results*. These notes are designed to increase student awareness of the steps in the data analysis process.
- **Updated examples and exercises that use data from current newspapers and journals are included.** In addition, more of the exercises specifically ask students to write (for example, by requiring students to explain their reasoning, interpret results, and comment on important features of an analysis).
- **Activities at the end of each chapter.** These activities can be used as a chapter capstone or can be integrated at appropriate places as the chapter material is covered in class.

- **Students can now go online with Aplia and CourseMate to further their understanding** of the material covered in each chapter.
- **Advanced topics** that are often omitted in a one-quarter or one-semester course, such as survey design (Section 2.6), logistic regression (Section 5.6), inference based on the estimated regression line (Sections 13.4 and 13.5), inference and variable selection methods in multiple regression (Sections 14.3 and 14.4), analysis of variance for randomized block and two-factor designs (Sections 15.3 and 15.4), and distribution-free procedures (Chapter 16) **are available in the online materials that accompany this text.**
- **Updated materials for instructors** are included. In addition to the usual instructor supplements such as a complete solutions manual and a test bank, the following are also available to instructors:
  - **An Instructor's Resource Binder,** which contains additional examples that can be incorporated into classroom presentations and cross-references to resources such as Fathom, Workshop Statistics, and Against All Odds. Of particular interest to those teaching Advanced Placement Statistics, the binder also includes additional data analysis questions of the type encountered on the free response portion of the Advanced Placement exam, as well as a collection of model responses.
  - For those who use student-response systems in class, **a set of "clicker" questions** (see JoinIn™ on TurningPoint® under Instructor Resources—Media) for assessing student understanding is available.

Instructor and Student Resources

MindTap™

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

Go to http://www.cengage.com/mindtap for more information.

Aplia™

Content

Aplia™ is an online interactive learning solution that improves comprehension and outcomes by increasing student effort and engagement. Founded by a professor to enhance his own courses, Aplia provides automatically graded assignments with detailed, immediate explanations for every question, along with innovative teaching materials. Our easy-to-use system has been used by more than 1,000,000 students at over 1,800 institutions. Exercises are taken directly from text.

Aplia homework engages students in critical thinking, requiring them to synthesize and apply knowledge, not simply recall it. The diverse types of questions reflect the types of exercises that help students learn. All homework is written by subject matter experts in the field who have taught the course before.

Aplia contains a robust course management system with powerful analytics, enabling professors to track student performance easily.

Service

Your adoption of Aplia® includes CourseCare, Cengage Learning's industry leading service and training program designed to ensure that you have everything that you need to make the most of your use of Aplia. CourseCare provides one-on-one service, from finding the right solutions for your course to training and support. A team of Cengage representatives, including Digital Solutions Managers and Coordinators as well as Service and Training Consultants assist you every step of the way. For additional information about CourseCare, please visit http://www.cengage.com/coursecare.

Our Aplia training program provides a comprehensive curriculum of beginner, intermediate, and advanced sessions, designed to get you started and effectively integrate Aplia into your course. We offer a flexible online and recorded training program designed to accommodate your busy schedule. Whether you are using Aplia for the first time or are an experienced user, there is a training option to meet your needs.

JMP Statistical Software

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

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

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

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

Student Resources

Digital

CENGAGE brain .com

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

- Complete step-by-step instructions for JMP, TI-84 Graphing Calculators, Excel, Minitab, and SPSS indicated by the  icon throughout the text.

- Data sets in JMP, TI-84, Excel, Minitab, SPSS, SAS, and ASCII file formats indicated by the ● icon throughout the text.
- Applets used in the Activities found in the text.

Print

Student Solutions Manual ***(ISBN: 978-1-3052-6582-0):*** The Student Solutions Manual, prepared by Michael Allwood, contains fully worked-out solutions to all of the odd-numbered exercises in the text, giving students a way to check their answers and ensure that they took the correct steps to arrive at an answer.

Instructor Resources

Print

Annotated Instructor's Edition ***(ISBN: 978-1-3052-5252-3):*** The Annotated Instructor's Edition contains answers for all exercises, including those not found in the answer section of the student edition. There also are suggested assignments and teaching tips for each section in the book written by Kathy Fritz, an experienced AP Statistics teacher, along with an annotated table of contents with comments written by Roxy Peck.

Teacher's Resource Binder ***(ISBN: 978-1-3052-6605-6):*** The Teacher's Resource Binder, prepared by Chris Olsen, is full of wonderful resources for both college professors and AP Statistics teachers. These include

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

Online

- **Instructor Companion Site:** Everything you need for your course in one place! This collection of book-specific lecture and class tools is available online via www.cengage.com/login. Access and download PowerPoint presentations, images, instructor's manual, and more.
  - **JoinIn™ on TurningPoint®:** The easiest student classroom response system to use, JoinIn features instant classroom assessment and learning.

- **Cengage Learning Testing Powered by Cognero (ISBN: 978-1-3052-6589-9)** is a flexible, online system that allows you to author, edit, and manage test bank content, create multiple test versions in an instant, and deliver tests from your LMS, your classroom or wherever you want. This is available online via www.cengage.com/login.
- **Complete Solutions Manual** This manual contains solutions to all exercises from the text, including Chapter Review Exercises and Cumulative Review Exercises. This manual can be found on the Instructors Companion Site.

Acknowledgments

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

Reviewers for the Fifth Edition

Kathleen Dale
Eastern Connecticut State University

Kendra Mhoon
University of Houston-Downtown

Mamunur Rashid
Indiana University-Purdue University Indianapolis

Leah Rathbun
Colorado State University

Cathleen ZuccoTeveloff
Rider University

Reviewers for the Fourth, Third, and Second Editions

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

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

- Molly Taylor, our editor, for her support and guidance.
- Jay Campbell, our developmental editor, for his kindness and humor as he kept us on track and moving forward.

- Danielle Hallock for leading the development of all of the supporting ancillaries.
- Jill Quinn, our Cengage production editor.
- Edward Dionne, our project editor at MPS Limited.
- Christine Sabooni for her careful manuscript review and copyediting.
- Michael Allwood for his work in creating new student and instructor solutions manuals to accompany the text.
- Kathy Fritz for creating the new interactive PowerPoint presentations that accompany the text and also for sharing her insight in writing the Teaching Tips that accompany each chapter in the annotated instructor editions.
- Stephen Miller for checking the accuracy of examples and solutions.
- Josh Tabor and Peter-Flannagan Hyde for their contributions to the test bank that accompanies the book.
- Beth Chance and Francisco Garcia for producing the applet used in the confidence interval activities.
- Gary McClelland for producing the applets from Seeing Statistics used in the regression activities.
- Carolyn Crockett, our former editor at Cengage, for her support on the previous editions of this book.

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

Roxy Peck
Chris Olsen
Jay Devore

Peck, Olsen, Devore's

Introduction to Statistics and Data Analysis, Fifth Edition . . .

. . . Emphasizes Statistical Literacy and Statistical Thinking

Context-Driven Applications

Real data examples and exercises throughout the text are drawn from the popular press, as well as journal articles. Data sources are in a colored font for easy identification.

EXAMPLE 2.2 Think Before You Order That Burger!

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

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

Page 37

EXERCISES 2.1 - 2.12

2.1 The article **"How Dangerous Is a Day in the Hospital?" (*Medical Care* [2011]: 1068–1075)** describes a study to determine if the risk of an infection is related to the length of a hospital stay. The researchers looked at a large number of hospitalized patients and compared the proportion who got an infection for two groups of patients—those who were hospitalized overnight and those who were hospitalized for more than one night. Indicate whether the study is an observational study or an experiment. Give a brief explanation for your choice.

2.2 The authors of the paper **"Fudging the Numbers: Distributing Chocolate Influences Student Evaluations of an Undergraduate Course" (*Teaching in Psychology* [2007]: 245–247)** carried out a study to see if events unrelated to an undergraduate course could affect chocolate prior to having them fill out course evaluations. Students in the other three sections were not offered chocolate.

The researchers concluded that "Overall, students offered chocolate gave more positive evaluations than students not offered chocolate." Indicate whether the study is an observational study or an experiment. Give a brief explanation for your choice.

2.3 The article **"Why We Fall for This" (*AARP Magazine*, May/June 2011)** described a study in which a business professor divided his class into two groups. He showed students a mug and then asked students in one of the groups how much they would pay for the mug. Students in the other group were asked how much they would sell the mug for if it belonged to them. Surprisingly, the average value assigned to the

Page 33

Focus on Interpreting and Communicating

Chapter sections on interpreting and communicating results are designed to emphasize the importance of being able to interpret statistical output and communicate its meaning to non-statisticians. A subsection titled "A Word to the Wise" reminds students of things that must be considered in order to ensure that statistical methods are used in reasonable and appropriate ways.

3.5 Interpreting and Communicating the Results of Statistical Analyses

A graphical display, when used appropriately, can be a powerful tool for organizing and summarizing data. By sacrificing some of the detail of a complete listing of a data set, important features of the data distribution are more easily seen and more easily communicated to others.

Page 127

A Word to the Wise: Cautions and Limitations

When constructing and interpreting graphical displays, you need to keep in mind these things:

1. **Areas should be proportional to frequency, relative frequency, or magnitude of the number being represented.** The eye is naturally drawn to large areas in graphical displays, and it is natural for the observer to make informal comparisons based on area. Correctly constructed graphical displays, such as pie charts,

Page 131

What to Look for in Published Data

Here are some questions you might ask yourself when attempting to extract information from a graphical data display:

- Is the chosen display appropriate for the type of data collected?
- For graphical displays of univariate numerical data, how would you describe the shape of the distribution, and what does this say about the variable being summarized?
- Are there any outliers (noticeably unusual values) in the data set? Is there any plausible explanation for why these values differ from the rest of the data? (The presence of outliers often leads to further avenues of investigation.)
- Where do most of the data values fall? What is a typical value for the data set? What does this say about the variable being summarized?
- Is there much variability in the data values? What does this say about the variable being summarized?

Page 130

Interpreting the Results of Statistical Analyses

When someone uses a web search engine, do they rely on the ranking of the search results returned or do they first scan the results looking for the most relevant? The authors of the paper **"Learning User Interaction Models for Predicting Web Search Result Preferences" (*Proceedings of the 29th Annual ACM Conference on Research and Development in Information Retrieval*, 2006)** attempted to answer this question by observing user behavior when they varied the position of the most relevant result in the list of resources returned in response to a web search.

They concluded that people clicked more often on results near the top of the list, even when they were not relevant. They supported this conclusion with the comparative bar graph in Figure 3.37.

Page 128

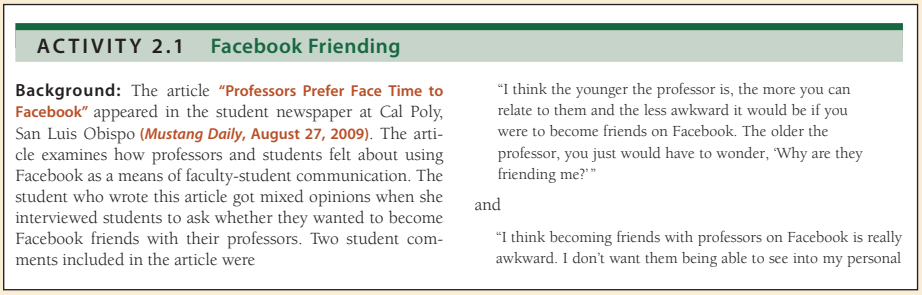
Peck, Olsen, Devore's

Introduction to Statistics and Data Analysis, Fifth Edition . . .

. . . Encourages Conceptual Understanding and Active Learning

Hands-on Activities in Every Chapter

More than 30 hands-on activities in the text, and additional activities in the accompanying instructor resources, can be used to encourage active learning inside or outside the classroom.



ACTIVITY 2.1 Facebook Friending

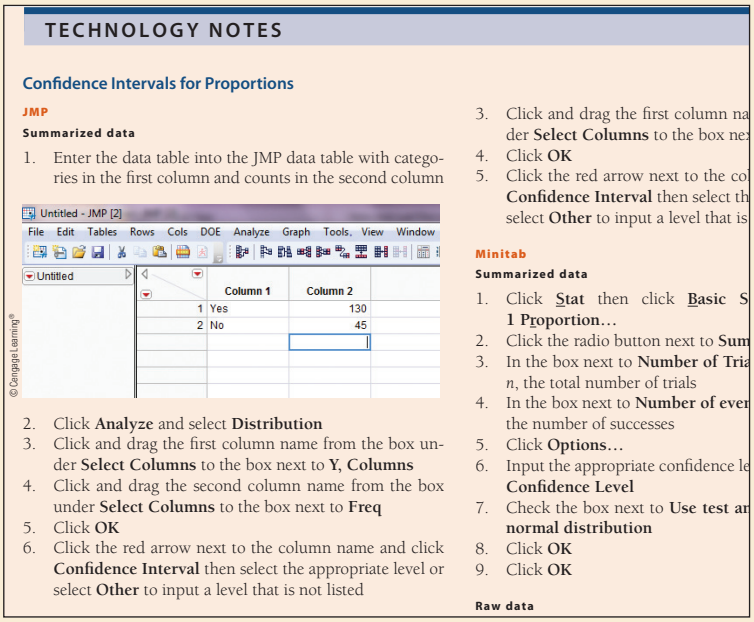
Background: The article **"Professors Prefer Face Time to Facebook"** appeared in the student newspaper at Cal Poly, San Luis Obispo (***Mustang Daily*, August 27, 2009**). The article examines how professors and students felt about using Facebook as a means of faculty-student communication. The student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

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

and

"I think becoming friends with professors on Facebook is really awkward. I don't want them being able to see into my personal

Page 69



TECHNOLOGY NOTES

Confidence Intervals for Proportions

JMP

Summarized data

1. Enter the data table into the JMP data table with categories in the first column and counts in the second column

2. Click **Analyze** and select **Distribution**
3. Click and drag the first column name from the box under **Select Columns** to the box next to **Y, Columns**
4. Click and drag the second column name from the box under **Select Columns** to the box next to **Freq**
5. Click **OK**
6. Click the red arrow next to the column name and click **Confidence Interval** then select the appropriate level or select **Other** to input a level that is not listed
3. Click and drag the first column na
der **Select Columns** to the box ne
4. Click **OK**
5. Click the red arrow next to the co
Confidence Interval then select th
select **Other** to input a level that is

Minitab

Summarized data

1. Click **Stat** then click **Basic S**
1 Proportion…
2. Click the radio button next to **Sum**
3. In the box next to **Number of Tria**
n, the total number of trials
4. In the box next to **Number of ever**
the number of successes
5. Click **Options…**
6. Input the appropriate confidence le
Confidence Level
7. Check the box next to **Use test ar**
normal distribution
8. Click **OK**
9. Click **OK**

Raw data

Page 502

Technology Notes

Technology Notes appear at the end of most chapters and give students helpful hints and guidance on completing tasks associated with a particular chapter. The following technologies are included in the notes: JMP, Minitab, SPSS, Microsoft Excel, and TI-83/84.

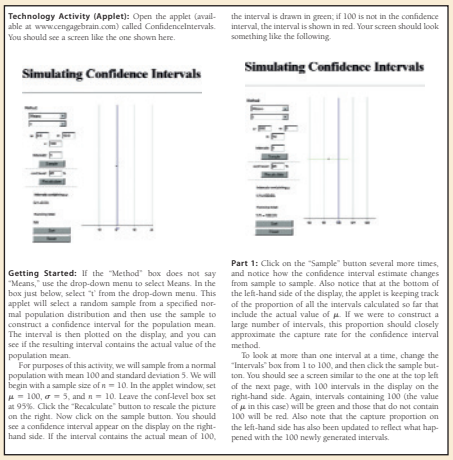
Peck, Olsen, Devore's

Introduction to Statistics and Data Analysis, Fifth Edition . . .

. . . Uses Technology to Develop Conceptual Understanding

Applets Allow Students to See the Concepts

Within the Activities, applets are used to illustrate and promote a deeper understanding of the key statistical concepts.



Technology Activity (Applet): Open the applet (available at www.cengagebrain.com) called ConfidenceIntervals. You should see a screen like the one shown here.

Simulating Confidence Intervals

Getting Started: If the "Method" box does not say "Means," use the drop-down menu to select Means. In the box just below, select 't' from the drop-down menu. This applet will select a random sample from a specified normal population distribution and then use the sample to construct a confidence interval for the population mean. The interval is then plotted on the display, and you can see if the resulting interval contains the actual value of the population mean.

For purposes of this activity, we will sample from a normal population with mean 100 and standard deviation 5. We will begin with a sample size of $n = 10$. In the applet window, set $\mu = 100$, $\sigma = 5$, and $n = 10$. Leave the conf-level box set at 95%. Click the "Recalculate" button to rescale the picture on the right. Now click on the sample button. You should see a confidence interval appear on the display on the right-hand side. If the interval contains the actual mean of 100, the interval is drawn in green; if 100 is not in the confidence interval, the interval is shown in red. Your screen should look something like the following.

Simulating Confidence Intervals

Part 1: Click on the "Sample" button several more times, and notice how the confidence interval estimate changes from sample to sample. Also notice that at the bottom of the left-hand side of the display, the applet is keeping track of the proportion of all the intervals calculated so far that include the actual value of μ. If we were to construct a large number of intervals, this proportion should closely approximate the capture rate for the confidence interval method.

To look at more than one interval at a time, change the "Intervals" box from 1 to 100, and then click the sample button. You should see a screen similar to the one at the top left of the next page, with 100 intervals in the display on the right-hand side. Again, intervals containing 100 (the value of μ in this case) will be green and those that do not contain 100 will be red. Also note that the capture proportion on the left-hand side has also been updated to reflect what happened with the 100 newly generated intervals.

Page 497

. . . And Analyze Data

Real Data Sets

Real data sets promote statistical analysis, as well as technology use. They are formatted for JMP, Minitab, SPSS, Microsoft Excel, TI-83/84, and ASCII and are indicated by the ● icon throughout the text.

EXAMPLE 3.11 Progress for Children

● The report **"Progress for Children" (UNICEF, April 2005)** included the accompanying data on the percentage of primary-school-age children who were enrolled in school for 19 countries in Northern Africa and for 23 countries in Central Africa.

Northern Africa

54.6 34.3 48.9 77.8 59.6 88.5 97.4 92.5 83.9 96.9 88.9
98.8 91.6 97.8 96.1 92.2 94.9 98.6 86.6

Central Africa

58.3 34.6 35.5 45.4 38.6 63.8 53.9 61.9 69.9 43.0 85.0
63.4 58.4 61.9 40.9 73.9 34.8 74.4 97.4 61.0 66.7 79.6
98.9

We will construct a comparative stem-and-leaf display using the first digit of each observation as the stem and the remaining two digits as the leaf. To keep the display simple the leaves will be truncated to one digit. For example, the observation 54.6 would be processed as

Page 95

3.21 ● The percentage of teens not in school or working in 2010 for the 50 states were given in the **2012 Kids Count Data Book (www.aecf.org)** and are shown in the following table:

State	Rate	State	Rate
Alabama	11%	Kansas	6%
Alaska	11%	Kentucky	11%
Arizona	12%	Louisiana	14%
Arkansas	12%	Maine	7%
California	8%	Maryland	8%
Colorado	7%	Massachusetts	5%
Connecticut	5%	Michigan	9%
Delaware	9%	Minnesota	5%
Florida	10%	Mississippi	13%
Georgia	12%	Missouri	9%
Hawaii	12%	Montana	9%

Page 98

Step-by-Step Technology Instructions

Complete online step-by-step instructions for JMP, Minitab, SPSS, Microsoft Excel, and TI-83/84 are indicated by the icon throughout the text.

Step-by-step technology instructions available online

Understand the data

EXAMPLE 3.16 TV Viewing Habits of Children

The article **"Early Television Exposure and Subsequent Attention Problems in Children" (*Pediatrics*, April 2004)** investigated the television viewing habits of children in the United States. Table 3.5 gives approximate relative frequencies (read from graphs that appeared in the article) for the number of hours spent watching TV per day for a sample of children at age 1 year and a sample of children at age 3 years. The data summarized in the article were obtained as part of a large-scale national survey.

Page 105

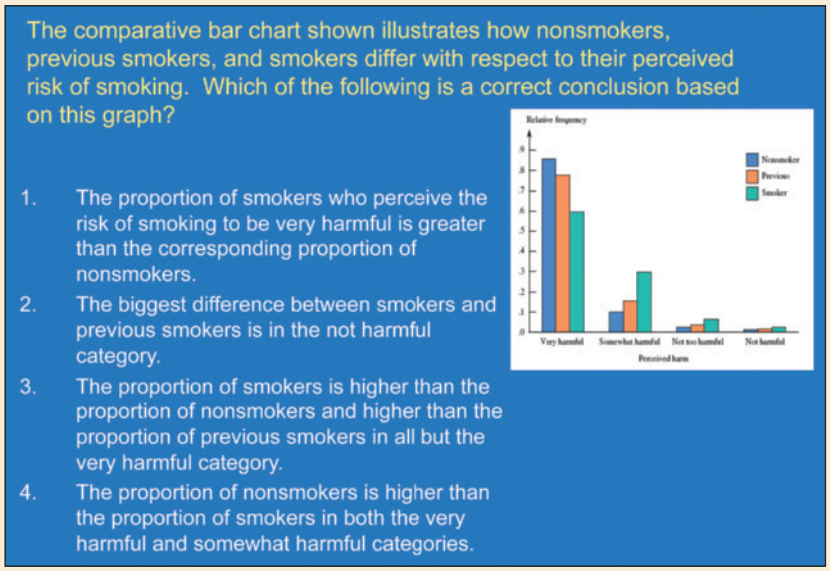
Peck, Olsen, Devore's

Introduction to Statistics and Data Analysis, Fifth Edition . . .

. . . Evaluates Students' Understanding

Evaluate as You Teach Using Clickers

Using clicker content authored by Roxy Peck, evaluate your students' understanding immediately—in class—after teaching a concept. Whether it's a quick quiz, a poll to be used for in-class data, or just checking in to see if it is time to move on, our quality, tested content creates truly interactive classrooms with students' responses shaping the lecture as you teach.



3.11 ▼ Poor fitness in adolescents and adults increases the risk of cardiovascular disease. In a study of 3110 adolescents and 2205 adults (*Journal of the American Medical Association*, December 21, 2005), researchers found 33.6% of adolescents and 13.9% of adults were unfit; the percentage was similar in adolescent males (32.9%) and females (34.4%), but was higher in adult females (16.2%) than in adult males (11.8%).

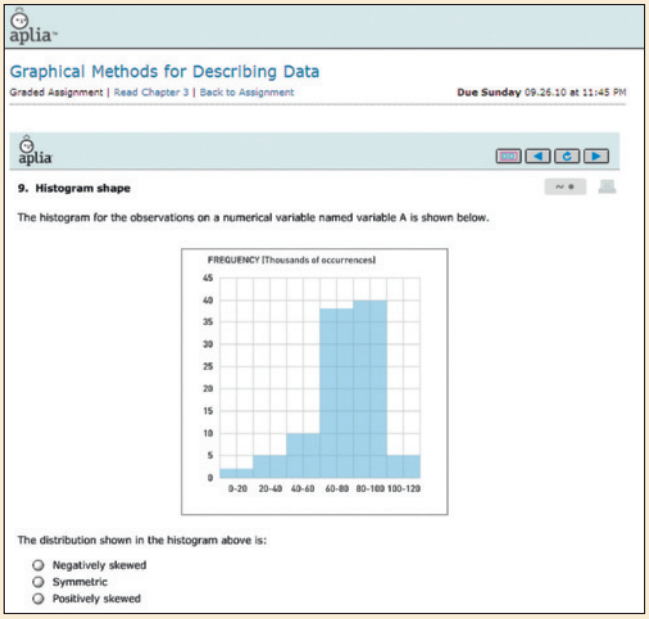
a. Summarize this information using a comparative bar graph that shows differences between males and females within the two different age groups.

b. Comment on the interesting features of your graphical display.

Page 90

Video Solutions Motivate Student Understanding

More than 90 exercises have video solutions, presented by Brian Kotz of Montgomery College, which can be viewed online or downloaded for later viewing. These exercises are designated by the ▼ in the text.



APLIA™ Online Interactive Learning Solution

Aplia provides automatically graded assignments with detailed, immediate explanations for every question, along with innovative teaching materials.

Resources for Students*

Digital

MindTap MindTap™

CENGAGE brain.com Cengagebrain.com

- Step-by-step instructions for JMP, TI-84 Graphing Calculators, Excel, Minitab, and SPSS
- Data sets formatted for JMP, TI-84, Excel, Minitab, SPSS, SAS, and ASCII
- Video solutions
- Applets used in the Activities

aplia Aplia™

ENHANCED WebAssign Enhanced WebAssign

JMP statistical software

Print

Student Solutions Manual
(978-1-3052-6582-0)

Resources for Instructors*

Digital

MindTap MindTap™

aplia Aplia™

Companion website: cengage.com/login

- PowerPoint presentations
- Images
- Complete solutions manual
- Cengage Learning Testing by Cognero
- JoinIn on TurningPoint

JMP statistical software

Print

Annotated Instructor's Edition
(978-1-3052-5252-3)

Teacher's Resource Binder
(978-1-3052-6605-6)

* See the full preface for complete descriptions.

Options That **SAVE** Your Students Money

ONLINE LEARNING WITH INTERACTIVE EBOOK

aplia™

Teach more, grade less with Aplia.

Aplia™ is an online interactive learning solution that improves comprehension and outcomes by increasing student effort and engagement.

Founded by a professor to enhance his own courses, **Aplia** provides automatically graded assignments with detailed, immediate explanations on every question, and innovative teaching materials. Our easy to use system has been used by more than 1,000,000 students at over 1800 institutions.

2-Semester Slim Pack
ISBN: 0-538-73475-2

Aplia can be bundled with the text. Contact your Cengage Learning representative to find out more about bundling options.

BOOK RENTALS UP TO 60% OFF

CENGAGEbrain.com

Students have the CHOICE to purchase the eBook or rent the text at CengageBrain.com

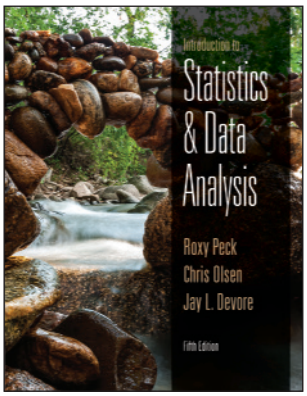
OR

an eTextbook in PDF format is also available for instant access for your students at **www.coursesmart.com**.

CourseSmart

Students can rent Peck/Olsen/Devore's *Introduction to Statistics and Data Analysis*, Fifth Edition, for up to 60% off list price.

CUSTOM SOLUTIONS



Custom Solutions to Fit Every Need

Contact your Cengage Learning representative to learn more about what custom solutions are available to meet your course needs.

- Adapt existing Cengage Learning content by adding or removing chapters
- Incorporate your own materials

CHAPTER 1

The Role of Statistics and the Data Analysis Process

© Andresr/Shutterstock.com

We encounter data and make conclusions based on data every day. **Statistics** is the scientific discipline that provides methods to help us make sense of data. Statistical methods, used intelligently, offer a set of powerful tools for gaining insight into the world around us. The widespread use of statistical analyses in diverse fields such as business, medicine, agriculture, social sciences, natural sciences, and engineering has led to increased recognition that statistical literacy—a familiarity with the goals and methods of statistics—should be a basic component of a well-rounded educational program.

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

Chapter 1: Learning Objectives

STUDENTS WILL UNDERSTAND:

- the steps in the data analysis process.

STUDENTS WILL BE ABLE TO:

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

1.1 Why Study Statistics?

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

It is difficult to function in today's world without a basic understanding of statistics. For example, here are just a few headlines from articles that draw conclusions based on data that all appeared over two days in ***USA Today*** **(December 19 and 20, 2013):**

- The article **"American Attitudes Toward Global Warming"** summarized data from a nationwide survey of adults. A variety of graphs and charts provide information on opinions regarding the existence of global warming, the impact of global warming, and what actions should be taken in response to global warming.
- **"Standardized Testing Fails College Students"** is the title of an article describing the use of standardized tests to place college students into appropriate college mathematics courses. The article concludes that many students are not aware of the importance of these exams and do not prepare for them. This results in many students being placed in developmental mathematics courses that slow their progress toward getting their degree.
- The article **"Shoppers Say Ho-Hum to Discounts"** describes conclusions drawn from a study of how consumers respond to e-mail advertising campaigns that offer discounts. The article concludes that this practice has become so widespread that shoppers largely ignore these e-mails and delete them without even reading them. This is information that retailers should consider when planning future advertising campaigns.
- **"Older Americans Could Opt Out of Blood Pressure Meds"** is the title of an article describing a study of the effect of high blood pressure on health for those over age 60. Data from this study lead to the conclusion that for older Americans, there is no further benefit to reducing blood pressure below 150/90. This is of interest to doctors because the previous recommendation was that blood pressure should be 140/90 or lower.
- The article **"College Coaching Gender Gap Persists"** reported data from a study of colleges in six large NCAA sports conferences. The study found that only 39.6% of the coaches of women's sports teams in 2013 were women, which is even lower than in previous years. The article concluded that although Title IX increased opportunities for participation of women in collegiate sports, it has not yet resulted in increased opportunities for women as coaches.

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

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

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

Studying statistics will also enable you to collect data in a sensible way and then use the data to answer questions of interest. In addition, studying statistics will allow you to

critically evaluate the work of others by providing you with the tools you need to make informed judgments.

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

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

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

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

1.2 The Nature and Role of Variability

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

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

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

EXAMPLE 1.1 If the Shoe Fits

Understand the context ⟩

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

Consider the data ⟩

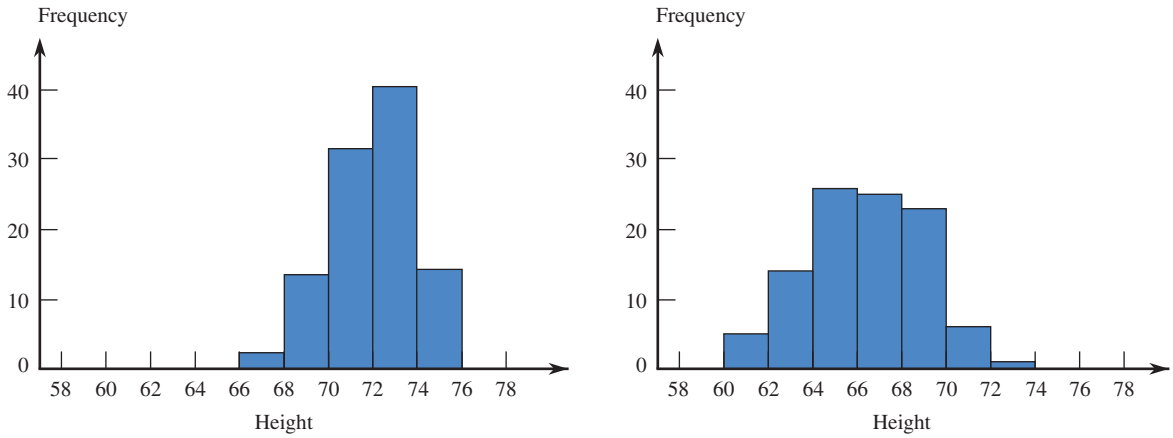


FIGURE 1.1
Histograms of heights (in inches) of female athletes:
(a) 100 basketball players;
(b) 100 gymnasts.

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

Interpret the results ⟩

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

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

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

EXAMPLE 1.2 Monitoring Water Quality


David Chasey/Photodisc/Getty Images

Understand the context ⟩

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

Now suppose that a chemical spill has occurred at a manufacturing plant 1 mile from the well. It is not known whether a spill of this nature would contaminate groundwater in the area

of the spill and, if so, whether a spill this distance from the well would affect the quality of well water.

Consider the data ⟩

One month after the spill, five water specimens are collected from the well, and the average contamination is 15.5 ppm. Considering the variation before the spill, would you interpret this as convincing evidence that the well water was affected by the spill? What if the calculated average was 17.4 ppm? 22.0 ppm? How is your reasoning related to the histogram in Figure 1.2?

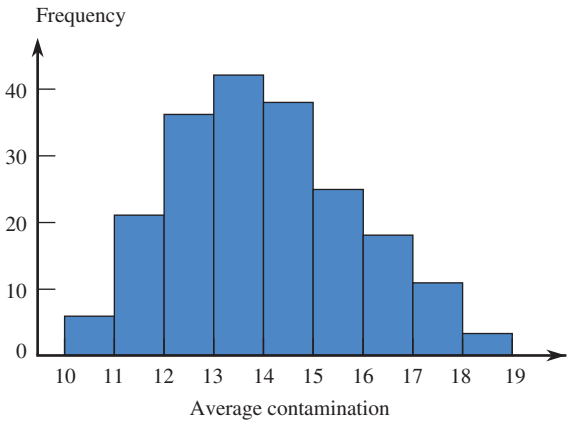


FIGURE 1.2
Average contamination concentration (in parts per million) measured each day for 200 days.

Interpret the results ⟩

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

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

1.3 Statistics and the Data Analysis Process

Statistics involves collecting, summarizing, and analyzing data. All three tasks are critical. Without summarization and analysis, raw data are of little value. Even sophisticated analyses can't produce meaningful information from data that were not collected in a sensible way.

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

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

The Data Analysis Process

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

To illustrate these steps, consider the following example. The admissions director at a large university might be interested in learning why some applicants who were accepted for the fall 2014 term failed to enroll at the university. The **population** of interest to the director consists of all accepted applicants who did not enroll in the fall 2014 term. Because this population is large and it may be difficult to contact all the individuals, the director might decide to collect data from only 300 selected students. These 300 students constitute a **sample**.

DEFINITION

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

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

Deciding how to select the 300 students and what data should be collected from each student are steps 2 and 3 in the data analysis process. Step 4 in the process involves organizing and summarizing data. Methods for organizing and summarizing data, such as the use of tables, graphs, or numerical summaries, make up the branch of statistics called **descriptive statistics**. The second major branch of statistics, **inferential statistics**, involves generalizing from a sample to the population from which it was selected. When we generalize in this way, we run the risk of an incorrect conclusion, because a conclusion about the population is based on incomplete information. An important aspect in the development of inferential techniques involves quantifying the chance of an incorrect conclusion.

DEFINITION

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

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

Example 1.3 illustrates the steps in the data analysis process.

EXAMPLE 1.3 The Benefits of Acting Out

Understand the context ⟩

A number of studies have reached the conclusion that stimulating mental activities can lead to improved memory and psychological wellness in older adults. The article **"A Short-Term Intervention to Enhance Cognitive and Affective Functioning in Older Adults" (*Journal of Aging and Health* [2004]: 562–585)** describes a study to investigate whether training in acting has similar benefits. Acting requires a person to consider the goals of the characters in the story, to remember lines of dialogue, to move on stage as scripted, and to do all of this at the same time. The researchers conducting the study wanted to see if participation in this type of complex multitasking would lead to an improvement in the ability to function independently.

Participants in the study were assigned to one of three groups. One group took part in an acting class for 4 weeks. One group spent a similar amount of time in a class on visual arts. The third group was a comparison group (called the "no-treatment group") that did not take either class. A total of 124 adults age 60 to 86 participated in the study.

Interpret the results ⟩

At the beginning of the 4-week study period and again at the end of the 4-week study period, each participant took several tests designed to measure problem-solving ability, memory span, self-esteem, and psychological well-being. After analyzing the data from this study, the researchers concluded that those in the acting group showed greater gains than both the visual arts group and the no-treatment group in both problem solving and psychological well-being.

Several new areas of research were suggested in the discussion that followed the analysis. The researchers wondered whether the effect of studying writing or music would be similar to what was observed for acting and described plans to investigate this further. They also noted that the participants in this study were generally well educated and recommended study of a more diverse group before generalizing conclusions about the benefits of studying acting to the larger population of all older adults.

This study illustrates the nature of the data analysis process. A clearly defined research question and an appropriate choice of how to measure the variables of interest (the tests used to measure problem solving, memory span, self-esteem, and psychological well-being) preceded the data collection. Assuming that a reasonable method was used to collect the data (we will see how this can be evaluated in Chapter 2) and that appropriate methods of analysis were employed, the investigators reached the conclusion that the study of acting showed promise. However, they recognized the limitations of the study, which in turn led to plans for further research. As is often the case, the data analysis cycle led to new research questions, and the process began again. ■

EXERCISES 1.1 - 1.11

1.1 Give a brief definition of the terms *descriptive statistics* and *inferential statistics*.

1.2 Give a brief definition of the terms *population* and *sample*.

1.3 The following conclusion from a study appeared in the article **"Smartphone Nation" (*AARP Bulletin*, September 2009)**: "If you love your smart phone, you are not alone. Half of all boomers sleep with their cell phone within arm's length. Two of three people age 50 to 64 use a cell phone to take photos, according to a 2010 Pew Research Center report." Are the given proportions (half and two of three) population values, or were they calculated from a sample?

1.4 Based on a study of 2121 children between the ages of 1 and 4, researchers at the Medical College of Wisconsin concluded that there was an association between iron deficiency and the length of time that a child is bottle-fed **(*Milwaukee Journal Sentinel*, November 26, 2005)**. Describe the sample and the population of interest for this study.

1.5 The student senate at a university with 15,000 students is interested in the proportion of students who favor a change in the grading system to allow for plus and minus grades (e.g., B+, B, B−, rather than just B). Two hundred students are interviewed to determine their attitude toward this proposed change.

a. What is the population of interest?

b. What group of students constitutes the sample in this problem?

1.6 The increasing popularity of online shopping has many consumers using Internet access at work to browse and shop online. In fact, the Monday after Thanksgiving has been nicknamed "Cyber Monday" because of the large increase in online purchases that occurs on that day. Data from a large-scale survey by a market research firm **(*Detroit Free Press*, November 26, 2005)** was used to compute estimates of the percent of men and women who shop online while at work. The resulting estimates probably won't make most employers happy—42% of the men and 32% of the women in the sample were shopping online at work!

Are the estimates given computed using data from a sample or for the entire population?

1.7 The supervisors of a rural county are interested in the proportion of property owners who support the construction of a sewer system. Because it is too costly to contact all 7000 property owners, a survey of 500 owners is undertaken. Describe the population and sample for this problem.

1.8 A consumer group conducts crash tests of new model cars. To determine the severity of damage to 2014 Toyota Camrys resulting from a 10-mph crash into a concrete wall, the research group tests six cars of this type and assesses the amount of damage. Describe the population and sample for this problem.

1.9 A building contractor has a chance to buy an odd lot of 5000 used bricks at an auction. She is interested in determining the proportion of bricks in the lot that are cracked and therefore unusable for her current project, but she does not have enough time to inspect all 5000 bricks. Instead, she checks 100 bricks to determine whether each is cracked. Describe the population and sample for this problem.

1.10 The article **"Brain Shunt Tested to Treat Alzheimer's" (*San Francisco Chronicle*, October 23, 2002)** summarizes the findings of a study that appeared in the journal *Neurology*. Doctors at Stanford Medical Center were interested in determining whether a new surgical approach to treating Alzheimer's disease results in improved memory functioning. The surgical procedure involves implanting a thin tube, called a shunt, which is designed to drain toxins from the fluid-filled space that cushions the brain. Eleven patients had shunts implanted and were followed for a year, receiving quarterly tests of memory function. Another sample of Alzheimer's patients was used as a comparison group. Those in the comparison group received the standard care for Alzheimer's disease. After analyzing the data from this study, the investigators concluded that the "results suggested the treated patients essentially held their own in the cognitive tests while the patients in the control group steadily declined. However, the study was too small to produce conclusive statistical evidence."

a. What were the researchers trying to learn? What questions motivated their research?

b. Do you think that the study was conducted in a reasonable way? What additional information would you want in order to evaluate this study? (Hint: See Example 1.3.)

1.11 In a study of whether taking a garlic supplement reduces the risk of getting a cold, participants were assigned to either a garlic supplement group or to a group that did not take a garlic supplement

("Garlic for the Common Cold," Cochrane Database of Systematic Reviews, 2009). Based on the study, it was concluded that the proportion of people taking a garlic supplement who get a cold is lower than the proportion of those not taking a garlic supplement who get a cold.

a. What were the researchers trying to learn? What questions motivated their research?

b. Do you think that the study was conducted in a reasonable way? What additional information would you want in order to evaluate this study?

Bold exercises answered in back ● Data set available online ▼ Video Solution available

1.4 Types of Data and Some Simple Graphical Displays

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

Types of Data

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

DEFINITION

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

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

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

DEFINITION

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

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

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