

Robert Gould | Rebecca Wong | Colleen Ryan

Introductory **STATISTICS**

exploring the world through data



Third Edition

Introductory Statistics: Exploring the World Through Data

Third Edition

Robert Gould

University of California, Los Angeles

Rebecca Wong

West Valley College

Colleen Ryan

Moorpark Community College



Director, Portfolio Management: Deirdre Lynch
Senior Portfolio Manager: Suzanna Bainbridge
Portfolio Management Assistant: Morgan Danna
Content Producer: Tamela Ambush
Managing Producer: Karen Wernholm
Producer: Shana Siegmund
Associate Content Developer: Sneh Singh
Manager, Courseware QA: Mary Durnwald
Manager, Content Development: Robert Carroll

Product Marketing Manager: Emily Ockay
Field Marketing Manager: Andrew Noble
Marketing Assistant: Shannon McCormack
Senior Author Support/Technology Specialist: Joe Vetere
Manager, Rights and Permissions: Gina Cheselka
Manufacturing Buyer: Carol Melville, LSC Communications
Cover Design, Full Service Vendor, Composition: Pearson CSC
Full Service Project Management: Chere Bemelmans - Pearson CSC
Cover Image: Paul Scott/EyeEm/Getty Images

Copyright © 2020, 2016, 2013 by Pearson Education, Inc. 221 River Street, Hoboken, NJ 07030. All Rights Reserved. Printed in the United States of America. This publication is protected by copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise. For information regarding permissions, request forms and the appropriate contacts within the Pearson Education Global Rights & Permissions department, please visit www.pearsoned.com/permissions/.

Attributions of third party content appear on page C-1, which constitutes an extension of this copyright page.

PEARSON, ALWAYS LEARNING, and MYLAB are exclusive trademarks owned by Pearson Education, Inc. or its affiliates in the U.S. and/or other countries.

Unless otherwise indicated herein, any third-party trademarks that may appear in this work are the property of their respective owners and any references to third-party trademarks, logos or other trade dress are for demonstrative or descriptive purposes only. Such references are not intended to imply any sponsorship, endorsement, authorization, or promotion of Pearson's products by the owners of such marks, or any relationship between the owner and Pearson Education, Inc. or its affiliates, authors, licensees or distributors.

TI-84+C screenshots courtesy of Texas Instruments. Data and screenshots from StatCrunch used by permission of StatCrunch. Screenshots from Minitab courtesy of Minitab Corporation. XLSTAT screenshots by Addinsoft, Inc. All Rights Reserved. XLSTAT is a registered trademark of Addinsoft SARL.

Microsoft® and Windows® are registered trademarks of the Microsoft Corporation in the U.S.A. and other countries. Screen shots and icons reprinted with permission from the Microsoft Corporation. This book is not sponsored or endorsed by or affiliated with the Microsoft Corporation. Microsoft and/or its respective suppliers make no representations about the suitability of the information contained in the documents and related graphics published as part of the services for any purpose. All such documents and related graphics are provided "as is" without warranty of any kind. Microsoft and/or its respective suppliers hereby disclaim all warranties and conditions with regard to this information, including all warranties and conditions of merchantability, whether express, implied or statutory, fitness for a particular purpose, title and non-infringement. In no event shall Microsoft and/or its respective suppliers be liable for any special, indirect or consequential damages or any damages whatsoever resulting from loss of use, data or profits, whether in an action of contract, negligence or other tortious action, arising out of or in connection with the use or performance of information available from the services. The documents and related graphics contained herein could include technical inaccuracies or typographical errors. Changes are periodically added to the information herein. Microsoft and/or its respective suppliers may make improvements and/or changes in the product(s) and/or the program(s) described herein at any time. Partial screen shots may be viewed in full within the software version specified.

Library of Congress Cataloging-in-Publication Data

Names: Gould, Robert, 1965- author. | Ryan, Colleen N. (Colleen Nooter), 1939- author.

Title: Introductory statistics : exploring the world through data / Robert Gould (University of California, Los Angeles), Colleen Ryan (California Lutheran University).

Description: Third edition. | Boston : Pearson, [2020] | Includes index.

Identifiers: LCCN 2018045613 | ISBN 9780135188927 (third edition)

Subjects: LCSH: Statistics--Textbooks.

Classification: LCC QA276.12 .G687 2020 | DDC 519.5--dc23 LC record available at <https://lccn.loc.gov/2018045613>

Dedication

To my parents and family, my friends, and my colleagues who are also friends. Without their patience and support, this would not have been possible.

—Rob

To Nathaniel and Allison, to my students, colleagues, and friends. Thank you for helping me be a better teacher and a better person.

—Rebecca

To my teachers and students, and to my family who have helped me in many different ways.

—Colleen

About the Authors

Robert Gould



Robert L. Gould (Ph.D., University of California, Los Angeles) is a leader in the statistics education community. He has served as chair of the American Statistical Association's (ASA) Statistics Education Section, chair of the American Mathematical Association of Two-Year Colleges/ASA Joint Committee, and has served on the National Council of Teacher of Mathematics/ASA Joint Committee. He served on a panel of co-authors for the *2005 Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report* and is co-author on the revision for the *GAISE K-12 Report*. As lead principal investigator of the NSF-funded Mobilize Project, he led the development of the first high school level data science course, which is taught in the Los Angeles Unified School District and several other districts. Rob teaches in the Department of Statistics at UCLA, where he directs the undergraduate statistics program and is director of the UCLA Center for Teaching Statistics. In recognition for his activities in statistics education, in 2012 Rob was elected Fellow of the American Statistical Association.

In his free time, Rob plays the cello and enjoys attending concerts of all types and styles.

Rebecca Wong



Rebecca K. Wong has taught mathematics and statistics at West Valley College for more than twenty years. She enjoys designing activities to help students explore statistical concepts and encouraging students to apply those concepts to areas of personal interest.

Rebecca earned a B.A. in mathematics and psychology from the University of California, Santa Barbara, an M.S.T. in mathematics from Santa Clara University, and an Ed.D. in Educational Leadership from San Francisco State University. She has been recognized for outstanding teaching by the National Institute of Staff and Organizational Development and the California Mathematics Council of Community Colleges.

When not teaching, Rebecca is an avid reader and enjoys hiking trails with friends.

Colleen Ryan



Colleen N. Ryan has taught statistics, chemistry, and physics to diverse community college students for decades. She taught at Oxnard College from 1975 to 2006, where she earned the Teacher of the Year Award. Colleen currently teaches statistics part-time at Moorpark Community College. She often designs her own lab activities. Her passion is to discover new ways to make statistical theory practical, easy to understand, and sometimes even fun.

Colleen earned a B.A. in physics from Wellesley College, an M.A.T. in physics from Harvard University, and an M.A. in chemistry from Wellesley College. Her first exposure to statistics was with Frederick Mosteller at Harvard.

In her spare time, Colleen sings, has been an avid skier, and enjoys time with her family.

Contents

Preface ix
Index of Applications xix

CHAPTER 1	Introduction to Data 1
	CASE STUDY ► Dangerous Habit? 2
	1.1 What Are Data? 3
	1.2 Classifying and Storing Data 6
	1.3 Investigating Data 10
	1.4 Organizing Categorical Data 13
	1.5 Collecting Data to Understand Causality 18
	DATA PROJECT ► How Are Data Stored? 28
CHAPTER 2	Picturing Variation with Graphs 40
	CASE STUDY ► Student-to-Teacher Ratio at Colleges 41
	2.1 Visualizing Variation in Numerical Data 42
	2.2 Summarizing Important Features of a Numerical Distribution 47
	2.3 Visualizing Variation in Categorical Variables 57
	2.4 Summarizing Categorical Distributions 60
	2.5 Interpreting Graphs 64
	DATA PROJECT ► Asking Questions 67
CHAPTER 3	Numerical Summaries of Center and Variation 90
	CASE STUDY ► Living in a Risky World 91
	3.1 Summaries for Symmetric Distributions 92
	3.2 What's Unusual? The Empirical Rule and z-Scores 101
	3.3 Summaries for Skewed Distributions 107
	3.4 Comparing Measures of Center 114
	3.5 Using Boxplots for Displaying Summaries 119
	DATA PROJECT ► The Statistical Investigation Cycle 126
CHAPTER 4	Regression Analysis: Exploring Associations between Variables 149
	CASE STUDY ► Forecasting Home Prices 150
	4.1 Visualizing Variability with a Scatterplot 151
	4.2 Measuring Strength of Association with Correlation 156
	4.3 Modeling Linear Trends 164
	4.4 Evaluating the Linear Model 178
	DATA PROJECT ► Data Moves 186

CHAPTER 5 Modeling Variation with Probability 213**CASE STUDY** ▶ SIDS or Murder? 214

- 5.1 What Is Randomness? 215
- 5.2 Finding Theoretical Probabilities 218
- 5.3 Associations in Categorical Variables 228
- 5.4 Finding Empirical and Simulated Probabilities 240

DATA PROJECT ▶ Subsetting Data 248**CHAPTER 6 Modeling Random Events: The Normal and Binomial Models 266****CASE STUDY** ▶ You Sometimes Get More Than You Pay for 267

- 6.1 Probability Distributions Are Models of Random Experiments 267
- 6.2 The Normal Model 273
- 6.3 The Binomial Model 287

DATA PROJECT ▶ Generating Random Numbers 303**CHAPTER 7 Survey Sampling and Inference 323****CASE STUDY** ▶ Spring Break Fever: Just What the Doctors Ordered? 324

- 7.1 Learning about the World through Surveys 325
- 7.2 Measuring the Quality of a Survey 332
- 7.3 The Central Limit Theorem for Sample Proportions 341
- 7.4 Estimating the Population Proportion with Confidence Intervals 348
- 7.5 Comparing Two Population Proportions with Confidence 356

DATA PROJECT ▶ Population Proportions 364**CHAPTER 8 Hypothesis Testing for Population Proportions 382****CASE STUDY** ▶ Dodging the Question 383

- 8.1 The Essential Ingredients of Hypothesis Testing 384
- 8.2 Hypothesis Testing in Four Steps 392
- 8.3 Hypothesis Tests in Detail 401
- 8.4 Comparing Proportions from Two Populations 408

DATA PROJECT ▶ Dates as Data 416

- CHAPTER 9** **Inferring Population Means 435**
CASE STUDY ▶ You Look Sick! Are You Sick? 436
9.1 Sample Means of Random Samples 437
9.2 The Central Limit Theorem for Sample Means 440
9.3 Answering Questions about the Mean of a Population 448
9.4 Hypothesis Testing for Means 458
9.5 Comparing Two Population Means 464
9.6 Overview of Analyzing Means 479
DATA PROJECT ▶ Data Structures 484
- CHAPTER 10** **Associations between Categorical Variables 507**
CASE STUDY ▶ Popping Better Popcorn 508
10.1 The Basic Ingredients for Testing with Categorical Variables 509
10.2 The Chi-Square Test for Goodness of Fit 518
10.3 Chi-Square Tests for Associations between Categorical Variables 523
10.4 Hypothesis Tests When Sample Sizes Are Small 531
DATA PROJECT ▶ Stacking Data 538
- CHAPTER 11** **Multiple Comparisons and Analysis of Variance 558**
CASE STUDY ▶ Seeing Red 559
11.1 Multiple Comparisons 560
11.2 The Analysis of Variance 566
11.3 The ANOVA Test 573
11.4 Post Hoc Procedures 578
DATA PROJECT ▶ Where to Begin 586
- CHAPTER 12** **Experimental Design: Controlling Variation 604**
CASE STUDY ▶ Does Stretching Improve Athletic Performance? 605
12.1 Variation Out of Control 606
12.2 Controlling Variation in Surveys 614
12.3 Reading Research Papers 618
DATA PROJECT ▶ Keep It Real 628
- CHAPTER 13** **Inference without Normality 639**
CASE STUDY ▶ Contagious Yawns 640
13.1 Transforming Data 641
13.2 The Sign Test for Paired Data 649
13.3 Mann-Whitney Test for Two Independent Groups 653
13.4 Randomization Tests 658
DATA PROJECT ▶ Making Maps and Slicing Strings 666

CHAPTER 14 Inference for Regression 690**CASE STUDY** ► Another Reason to Stand at Your Desk? 691

14.1 The Linear Regression Model 692

14.2 Using the Linear Model 703

14.3 Predicting Values and Estimating Means 711

DATA PROJECT ► Think Small 719

Appendix A Tables A-1

Appendix B Answers to Odd-Numbered Exercises A-9

Appendix C Credits C-1

Index I-1

Preface

About This Book

We believe firmly that analyzing data to uncover insight and meaning is one of the most important skills to prepare students for both the workplace and civic life. This is not a book about “statistics,” but is a book about understanding our world and, in particular, understanding how statistical inference and data analysis can improve the world by helping us see more clearly.

Since the first edition, we’ve seen the rise of a new science of data and been amazed by the power of data to improve our health, predict our weather, connect long-lost friends, run our households, and organize our lives. But we’ve also been concerned by data breaches, by a loss of privacy that can threaten our social structures, and by attempts to manipulate opinion.

This is not a book meant merely to teach students to interpret the statistical findings of others. We *do* teach that; we all need to learn to critically evaluate arguments, particularly arguments based on data. But more importantly, we wish to inspire students to examine data and make their own discoveries. This is a book about *doing*. We are not interested in a course to teach students to memorize formulas or to ask them to mindlessly carry out procedures. Students must learn to think critically with and about data, to communicate their findings to others, and to carefully evaluate others’ arguments.

What’s New in the Third Edition

As educators and authors, we were strongly inspired by the spirit that created the Guidelines for Assessment and Instruction in Statistics Education (GAISE) (<http://amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx>), which recommends that we

- teach statistical thinking, which includes teaching statistics as an investigative process and providing opportunities for students to engage in multivariate thinking;
- focus on conceptual understandings;
- integrate real data with a context and purpose;
- foster active learning;
- use technology to explore concepts and to analyze data;
- use assessments to improve and evaluate student learning.

These have guided the first two editions of the book. But the rise of data science has led us to rethink how we engage students with data, and so, in the third edition, we offer some new features that we hope will prepare students for working with the complex data that surrounds us.

More precisely, you’ll find

- An emphasis on what we call the Data Cycle, a device to guide students through the statistical investigation process. The Data Cycle includes four phases: Ask Questions, Consider Data, Analyze Data, and Interpret Data. A new marginal icon indicates when the Data Cycle is particularly relevant.
- An increased emphasis on formulating “statistical investigative questions” as an important first step in the Data Cycle. Previous editions have emphasized the other three steps, but we feel students need practice in formulating questions that will help them interpret data. To formulate questions is to engage in mathematical and statistical modeling, and this edition spends more time teaching this important skill.

- The end-of-chapter activities have been replaced by a series of “Data Projects.” These are self-guided activities that teach students important “data moves” that will help them navigate through the large and complex data sets that are so often found in the real world.
- The addition of a “Data Moves” icon. Some examples are based on extracts of data from much larger data sets. The Data Moves icon points students to these data sets and also indicate the “data moves” used to extract the data. We are indebted to Tim Erickson for the phrase “data moves” and the ideas that motivate it.
- A smoother and more refined approach to simulations in Chapter 5.
- Updated technology guides to match current hardware and software.
- Hundreds of new exercises.
- New and updated examples in each chapter.
- New and updated data sets, with the inclusion of more large data.

Approach

Our text is concept-based, as opposed to method-based. We teach useful statistical methods, but we emphasize that applying the method is secondary to understanding the concept.

In the real world, computers do most of the heavy lifting for statisticians. We therefore adopt an approach that frees the instructor from having to teach tedious procedures and leaves more time for teaching deeper understanding of concepts. Accordingly, we present formulas as an aid to understanding the concepts, rather than as the focus of study.

We believe students need to learn how to

- Determine which statistical procedures are appropriate.
- Instruct the software to carry out the procedures.
- Interpret the output.

We understand that students will probably see only one type of statistical software in class. But we believe it is useful for students to compare output from several different sources, so in some examples we ask them to read output from two or more software packages.

Coverage

The first two-thirds of this book are concept-driven and cover exploratory data analysis and inferential statistics—fundamental concepts that every introductory statistics student should learn. The final third of the book builds on that strong conceptual foundation and is more methods-based. It presents several popular statistical methods and more fully explores methods presented earlier, such as regression and data collection.

Our ordering of topics is guided by the process through which students should analyze data. First, they explore and describe data, possibly deciding that graphics and numerical summaries provide sufficient insight. Then they make generalizations (inferences) about the larger world.

Chapters 1–4: Exploratory Data Analysis. The first four chapters cover data collection and summary. Chapter 1 introduces the important topic of data collection and compares and contrasts observational studies with controlled experiments. This chapter also teaches students how to handle raw data so that the data can be uploaded to their statistical software. Chapters 2 and 3 discuss graphical and numerical summaries of single

variables based on samples. We emphasize that the purpose is not just to produce a graph or a number but, instead, to explain what those graphs and numbers say about the world. Chapter 4 introduces simple linear regression and presents it as a technique for providing graphical and numerical summaries of relationships between two numerical variables.

We feel strongly that introducing regression early in the text is beneficial in building student understanding of the applicability of statistics to real-world scenarios. After completing the chapters covering data collection and summary, students have acquired the skills and sophistication they need to describe two-variable associations and to generate informal hypotheses. Two-variable associations provide a rich context for class discussion and allow the course to move from fabricated problems (because one-variable analyses are relatively rare in the real world) to real problems that appear frequently in everyday life. We return to regression in Chapter 14, when we discuss statistical inference in the context of regression, which requires quite a bit of machinery. We feel that it would be a shame to delay until the end of the course all the insights that regression without inference can provide.

Chapters 5–8: Inference. These chapters teach the fundamental concepts of statistical inference. The main idea is that our data mirror the real world, but imperfectly; although our estimates are uncertain, under the right conditions we can quantify our uncertainty. Verifying that these conditions exist and understanding what happens if they are not satisfied are important themes of these chapters.

Chapters 9–11: Methods. Here we return to the themes covered earlier in the text and present them in a new context by introducing additional statistical methods, such as estimating population means, analyzing categorical variables, and analyzing relations between a numerical and a categorical variable. We also introduce multiple comparisons and use them to motivate the need for the statistical method of ANOVA.

Chapters 12–14: Special Topics. Students who have covered all topics up to this point will have a solid foundation in statistics. These final chapters build on that foundation and offer more details, as we explore the topics of designing controlled experiments, survey sampling, additional contexts for hypothesis testing, and using regression to make inferences about a population.

In Chapter 12 we provide guidance for reading scientific literature. Even if your schedule does not allow you to cover Chapter 12, we recommend using Section 12.3 to offer students the experience of critically examining real scientific papers.

Organization

Our preferred order of progressing through the text is reflected in the Contents, but there are some alternative pathways as well.

10-week Quarter. The first eight chapters provide a full, one-quarter course in introductory statistics. If time remains, cover Sections 9.1 and 9.2 as well, so that students can solidify their understanding of confidence intervals and hypothesis tests by revisiting the topic with a new parameter.

Proportions First. Ask two statisticians, and you will get three opinions on whether it is best to teach means or proportions first. We have come down on the side of proportions for a variety of reasons. Proportions are much easier to find in popular news media (particularly around election time), so they can more readily be tied to students' everyday lives. Also, the mathematics and statistical theory are simpler; because there's no need to provide a separate estimate for the population standard deviation, inference is based on the Normal distribution, and no further approximations (that is, the t -distribution) are required. Hence, we can quickly get to the heart of the matter with fewer technical diversions.

The basic problem here is how to quantify the uncertainty involved in estimating a parameter and how to quantify the probability of making incorrect decisions when posing hypotheses. We cover these ideas in detail in the context of proportions. Students can then more easily learn how these same concepts are applied in the new context of means (and any other parameter they may need to estimate).

Means First. Conversely, many people feel that there is time for only one parameter and that this parameter should be the mean. For this alternative presentation, cover Chapters 6, 7, and 9, in that order. On this path, students learn about survey sampling and the terminology of inference (population vs. sample, parameter vs. statistic) and then tackle inference for the mean, including hypothesis testing.

To minimize the coverage of proportions, you might choose to cover Chapter 6, Section 7.1 (which treats the language and framework of statistical inference in detail), and then Chapter 9. Chapters 7 and 8 develop the concepts of statistical inference more slowly than Chapter 9, but essentially, Chapter 9 develops the same ideas in the context of the mean.

If you present Chapter 9 before Chapters 7 and 8, we recommend that you devote roughly twice as much time to Chapter 9 as you have devoted to previous chapters, because many challenging ideas are explored in this chapter. If you have already covered Chapters 7 and 8 thoroughly, Chapter 9 can be covered more quickly.


Features

We've incorporated into this text a variety of features to aid student learning and to facilitate its use in any classroom.

Integrating Technology







Modern statistics is inseparable from computers. We have worked to make this textbook accessible for any classroom, regardless of the level of in-class exposure to technology, while still remaining true to the demands of the analysis. We know that students sometimes do not have access to technology when doing homework, so many exercises provide output from software and ask students to interpret and critically evaluate that given output.

Using technology is important because it enables students to handle real data, and real data sets are often large and messy. The following features are designed to guide students.

- **TechTips** outline steps for performing calculations using TI-84[®] (including TI-84 + C[®]) graphing calculators, Excel[®], Minitab[®], and StatCrunch[®]. We do not want students to get stuck because they don't know how to reproduce the results we show in the book, so whenever a new method or procedure is introduced, an icon, , refers students to the TechTips section at the end of the chapter. Each set of TechTips contains at least one mini-example, so that students are not only learning to use the technology but also practicing data analysis and reinforcing ideas discussed in the text. Most of the provided TI-84 steps apply to all TI-84 calculators, but some are unique to the TI-84 + C calculator. Throughout the text, screenshots of TI calculators are labeled "TI-84" but are, in fact, from a TI-84 Plus C Silver Edition.
- All **data sets** used in the exposition and exercises are available at <http://www.pearsonhighered.com/mathstatsresources/>.

Guiding Students

- Each chapter opens with a **Theme**. Beginners have difficulty seeing the forest for the trees, so we use a theme to give an overview of the chapter content.

- Each chapter begins by posing a real-world **Case Study**. At the end of the chapter, we show how techniques covered in the chapter helped solve the problem presented in the Case Study.
- **Margin Notes** draw attention to details that enhance student learning and reading comprehension.
 -  **Caution** notes provide warnings about common mistakes or misconceptions.
 -  **Looking Back** reminders refer students to earlier coverage of a topic.
 -  **Details** clarify or expand on a concept.
-  **Key Points** highlight essential concepts to draw special attention to them. Understanding these concepts is essential for progress.
-  **Snapshots** break down key statistical concepts introduced in the chapter, quickly summarizing each concept or procedure and indicating when and how it should be used.
-  **Data Moves** point students toward more complete source data.
- An abundance of worked-out **examples** model solutions to real-world problems relevant to students' lives. Each example is tied to an end-of-chapter exercise so that students can practice solving a similar problem and test their understanding. Within the exercise sets, the icon **TRY** indicates which problems are tied to worked-out examples in that chapter, and the numbers of those examples are indicated.
- The **Chapter Review** that concludes each chapter provides a list of important new terms, student learning objectives, a summary of the concepts and methods discussed, and sources for data, articles, and graphics referred to in the chapter.

Active Learning

- Each chapter ends in a **Data Project**. These are activities designed for students to work alone or in pairs. Data analysis requires practice, and these sections, which grow increasingly more complex, are intended to guide students through basic “data moves” to help them find insight in complex data.
- All exercises are located at the end of the chapter. **Section Exercises** are designed to begin with a few basic problems that strengthen recall and assess basic knowledge, followed by mid-level exercises that ask more complex, open-ended questions. **Chapter Review Exercises** provide a comprehensive review of material covered throughout the chapter.

The exercises emphasize good statistical practice by requiring students to verify conditions, make suitable use of graphics, find numerical values, and interpret their findings in writing. All exercises are paired so that students can check their work on the odd-numbered exercise and then tackle the corresponding even-numbered exercise. The answers to all odd-numbered exercises appear in the back of the student edition of the text.

Challenging exercises, identified with an asterisk (*), ask open-ended questions and sometimes require students to perform a complete statistical analysis.

- Most chapters include select exercises, marked with a **g** within the exercise set, to indicate that problem-solving help is available in the **Guided Exercises** section. If students need support while doing homework, they can turn to the Guided Exercises to see a step-by-step approach to solving the problem.

Acknowledgments

We are grateful for the attention and energy that a large number of people devoted to making this a better book. We extend our gratitude to Chere Bemelmans, who handled production, and to Tamela Ambush, content producer. Many thanks to John Norbutas for his technical advice and help with the TechTips. We thank Deirdre Lynch, editor-in-chief, for signing us up and sticking with us, and we are grateful to Emily Ockay for her market development efforts.

We extend our sincere thanks for the suggestions and contributions made by the following reviewers of this edition:

Beth Burns, *Bowling Green State University*

Rod Elmore, *Mid Michigan Community College*

Carl Fetteroll, *Western New England University*

Elizabeth Flynn, *College of the Canyons*

David French, *Tidewater Community College*

Terry Fuller, *California State University, Northridge*

Kimberly Gardner, *Kennesaw State University*

Ryan Girard, *Kauai Community College*

Carrie Grant, *Flagler College*

Deborah Hanus, *Brookhaven College*

Kristin Harvey, *The University of Texas at Austin*

Abbas Jaffary, *Moraine Valley Community College*

Tony Jenkins, *Northwestern Michigan College*

Jonathan Kalk, *Kauai Community College*

Joseph Kudrle, *University of Vermont*

Matt Lathrop, *Heartland Community College*

Raymond E. Lee, *The University of North Carolina at Pembroke*

Karen McNeal, *Moraine Valley Community College*

Tejal Naik, *West Valley College*

Hadley Pridgen, *Gulf Coast State College*

John M. Russell, *Old Dominion University*

Amy Salvati, *Adirondack Community College*

Marcia Siderow, *California State University, Northridge*

Kenneth Strazzeri, *George Mason University*

Amy Vu, *West Valley College*

Rebecca Walker, *Guttman Community College*

We would also like to extend our sincere thanks for the suggestions and contributions made by the following reviewers, class testers, and focus group attendees of the previous edition.

Arun Agarwal, *Grambling State University*

Anne Albert, *University of Findlay*

Michael Allen, *Glendale Community College*

Eugene Allevato, *Woodbury University*

Dr. Jerry Allison, *Trident Technical College*

Polly Amstutz, *University of Nebraska*

Patricia Anderson, *Southern Adventist University*

MaryAnne Anthony-Smith, *Santa Ana College*

David C. Ashley, *Florida State College at Jacksonville*

Diana Asmus, *Greenville Technical College*

Kathy Autrey, *Northwestern State University of Louisiana*

Wayne Barber, *Chemeketa Community College*

Roxane Barrows, *Hocking College*

Jennifer Beineke, *Western New England College*

Diane Benner, *Harrisburg Area Community College*

Norma Biscula, *University of Maine, Augusta*

K.B. Boomer, *Bucknell University*

Mario Borha, *Loyola University of Chicago*

David Bosworth, *Hutchinson Community College*

Diana Boyette, *Seminole Community College*

Elizabeth Paulus Brown, *Waukesha County Technical College*

Leslie Buck, *Suffolk Community College*

R.B. Campbell, *University of Northern Iowa*

Stephanie Campbell, *Mineral Area College*

Ann Cannon, *Cornell College*

Rao Chaganty, *Old Dominion University*

Carolyn Chapel, *Western Technical College*

Christine Cole, *Moorpark College*

Linda Brant Collins, *University of Chicago*

James A. Condor, *Manatee Community College*

Carolyn Cuff, *Westminster College*

Phyllis Curtiss, *Grand Valley State University*

Monica Dabos, *University of California, Santa Barbara*

Greg Davis, *University of Wisconsin, Green Bay*

Bob Denton, *Orange Coast College*

Julie DePree, *University of New Mexico—Valencia*

Jill DeWitt, *Baker Community College of Muskegon*

Paul Drelles, *West Shore Community College*

Keith Driscoll, *Clayton State University*

Rob Eby, *Blinn College*

Nancy Eschen, *Florida Community College at Jacksonville*

Karen Estes, *St. Petersburg College*

Mariah Evans, *University of Nevada, Reno*

Harshini Fernando, *Purdue University North Central*

Stephanie Fitchett, *University of Northern Colorado*

Elaine B. Fitt, *Bucks County Community College*

Michael Flesch, *Metropolitan Community College*

Melinda Fox, *Ivy Tech Community College, Fairbanks*

Joshua Francis, *Defiance College*

Michael Frankel, *Kennesaw State University*

Heather Gamber, *Lone Star College*

Debbie Garrison, *Valencia Community College, East Campus*

Kim Gilbert, *University of Georgia*

Stephen Gold, *Cypress College*

Nick Gomersall, *Luther College*

Mary Elizabeth Gore, *Community College of Baltimore County—Essex*

- Ken Grace, *Anoka Ramsey Community College*
- Larry Green, *Lake Tahoe Community College*
- Jeffrey Grell, *Baltimore City Community College*
- Albert Groccia, *Valencia Community College, Osceola Campus*
- David Gurney, *Southeastern Louisiana University*
- Chris Hakenkamp, *University of Maryland, College Park*
- Melodie Hallet, *San Diego State University*
- Donnie Hallstone, *Green River Community College*
- Cecil Hallum, *Sam Houston State University*
- Josephine Hamer, *Western Connecticut State University*
- Mark Harbison, *Sacramento City College*
- Beverly J. Hartter, *Oklahoma Wesleyan University*
- Laura Heath, *Palm Beach State College*
- Greg Henderson, *Hillsborough Community College*
- Susan Herring, *Sonoma State University*
- Carla Hill, *Marist College*
- Michael Huber, *Muhlenberg College*
- Kelly Jackson, *Camden County College*
- Bridgette Jacob, *Onondaga Community College*
- Robert Jernigan, *American University*
- Chun Jin, *Central Connecticut State University*
- Jim Johnston, *Concord University*
- Maryann Justinger, Ed.D., *Erie Community College*
- Joseph Karnowski, *Norwalk Community College*
- Susitha Karunaratne, *Purdue University North Central*
- Mohammed Kazemi, *University of North Carolina–Charlotte*
- Robert Keller, *Loras College*
- Omar Keshk, *Ohio State University*
- Raja Khoury, *Collin County Community College*
- Brianna Killian, *Daytona State College*
- Yoon G. Kim, *Humboldt State University*
- Greg Knofczynski, *Armstrong Atlantic University*
- Jeffrey Kollath, *Oregon State University*
- Erica Kwiatkowski-Egizio, *Joliet Junior College*
- Sister Jean A. Lanahan, *OP, Molloy College*
- Katie Larkin, *Lake Tahoe Community College*
- Michael LaValle, *Rochester Community College*
- Deann Leoni, *Edmonds Community College*
- Lenore Lerer, *Bergen Community College*
- Quan Li, *Texas A&M University*
- Doug Mace, *Kirtland Community College*
- Walter H. Mackey, *Owens Community College*
- Keith McCoy, *Wilbur Wright College*
- Elaine McDonald-Newman, *Sonoma State University*
- William McGregor, *Rockland Community College*
- Bill Meisel, *Florida State College at Jacksonville*
- Bruno Mendes, *University of California, Santa Cruz*
- Wendy Miao, *El Camino College*
- Robert Mignone, *College of Charleston*
- Ashod Minasian, *El Camino College*
- Megan Mocko, *University of Florida*
- Sumona Mondal, *Clarkson University*
- Kathy Mowers, *Owensboro Community and Technical College*
- Mary Moyinhan, *Cape Cod Community College*
- Junalyn Navarra-Madsen, *Texas Woman’s University*
- Azarnia Nazanin, *Santa Fe College*
- Stacey O. Nicholls, *Anne Arundel Community College*
- Helen Noble, *San Diego State University*
- Lyn Noble, *Florida State College at Jacksonville*
- Keith Oberlander, *Pasadena City College*
- Pamela Omer, *Western New England College*
- Ralph Padgett Jr., *University of California – Riverside*
- Nabendu Pal, *University of Louisiana at Lafayette*
- Irene Palacios, *Grossmont College*
- Ron Palcic, *Johnson County Community College*
- Adam Pennell, *Greensboro College*
- Patrick Perry, *Hawaii Pacific University*
- Joseph Pick, *Palm Beach State College*
- Philip Pickering, *Genesee Community College*
- Victor I. Piercey, *Ferris State University*
- Robin Powell, *Greenville Technical College*
- Nicholas Pritchard, *Coastal Carolina University*
- Linda Quinn, *Cleveland State University*
- William Radulovich, *Florida State College at Jacksonville*
- Mumunur Rashid, *Indiana University of Pennsylvania*
- Fred J. Rispoli, *Dowling College*
- Danielle Rivard, *Post University*
- Nancy Rivers, *Wake Technical Community College*
- Corlis Robe, *East Tennessee State University*
- Thomas Roe, *South Dakota State University*
- Alex Rolon, *North Hampton Community College*
- Dan Rowe, *Heartland Community College*
- Ali Saadat, *University of California – Riverside*
- Kelly Sakkinen, *Lake Land College*
- Carol Saltsgaver, *University of Illinois–Springfield*
- Radha Sankaran, *Passaic County Community College*
- Delray Schultz, *Millersville University*
- Jenny Shook, *Pennsylvania State University*
- Danya Smithers, *Northeast State Technical Community College*
- Larry Southard, *Florida Gulf Coast University*
- Dianna J. Spence, *North Georgia College & State University*
- René Sporer, *Diablo Valley College*
- Jeganathan Sriskandarajah, *Madison Area Technical College–Traux*
- David Stewart, *Community College of Baltimore County–Cantonsville*
- Linda Strauss, *Penn State University*
- John Stroyls, *Georgia Southwestern State University*
- Joseph Sukta, *Moraine Valley Community College*
- Sharon I. Sullivan, *Catawba College*
- Lori Thomas, *Midland College*
- Malissa Trent, *Northeast State Technical Community College*
- Ruth Trygstad, *Salt Lake Community College*
- Gail Tudor, *Husson University*
- Manuel T. Uy, *College of Alameda*
- Lewis Van Brackle, *Kennesaw State University*
- Mahbobeh Vezvaei, *Kent State University*
- Joseph Villalobos, *El Camino College*
- Barbara Wainwright, *Salisbury University*
- Henry Wakhungu, *Indiana University*
- Jerimi Ann Walker, *Moraine Valley Community College*
- Dottie Walton, *Cuyahoga Community College*
- Jen-ting Wang, *SUNY, Oneonta*
- Jane West, *Trident Technical College*
- Michelle White, *Terra Community College*
- Bonnie-Lou Wicklund, *Mount Wachusett Community College*
- Sandra Williams, *Front Range Community College*
- Rebecca Wong, *West Valley College*
- Alan Worley, *South Plains College*
- Jane-Marie Wright, *Suffolk Community College*
- Haishen Yao, *CUNY, Queensborough Community College*
- Lynda Zenati, *Robert Morris Community College*
- Yan Zheng-Araujo, *Springfield Community Technical College*
- Cathleen Zucco-Teveloff, *Rider University*
- Mark A. Zuiker, *Minnesota State University, Mankato*

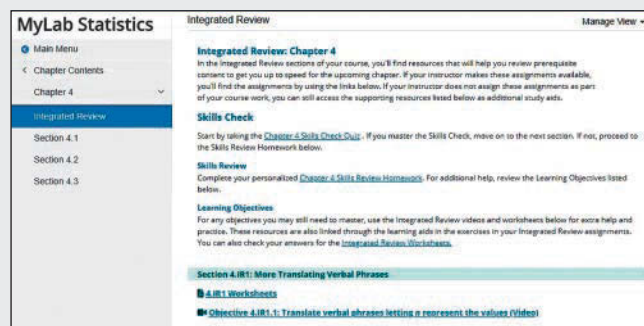
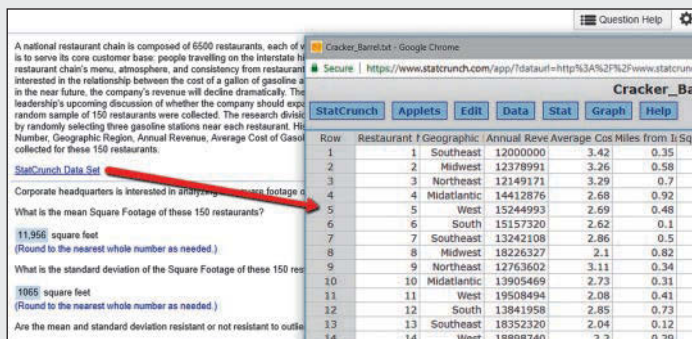
MyLab Statistics Online Course for *Introductory Statistics: Exploring the World Through Data, 3e*

(Access Code Required)

MyLab™ Statistics is available to accompany Pearson’s market-leading text offerings. To give students a consistent tone, voice, and teaching method, each text’s flavor and approach is tightly integrated throughout the accompanying MyLab Statistics course, making learning the material as seamless as possible.

NEW! Integrated Review

This MyLab includes a full suite of supporting Integrated Review resources for the Gould, *Introductory Statistics* course, including pre-made, assignable (and editable) quizzes to assess the prerequisite skills needed for each chapter, and personalized remediation for any gaps in skills that are identified. Each student, therefore, receives just the help that he or she needs—no more, no less.

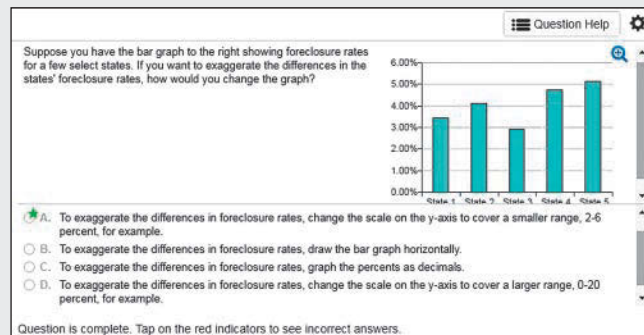
Row	Restaurant	Geographic	Annual Revenue	Average Cost	Miles from Interstate
1	1	Southeast	12000000	3.42	0.35
2	2	Midwest	12378991	3.26	0.58
3	3	Northeast	12149171	3.29	0.7
4	4	Midatlantic	14412876	2.68	0.92
5	5	West	15244993	2.69	0.48
6	6	South	15157320	2.62	0.1
7	7	Southeast	13242108	2.86	0.5
8	8	Midwest	18226327	2.1	0.82
9	9	Northeast	12763602	3.11	0.34
10	10	Midatlantic	13905469	2.73	0.31
11	11	West	19508494	2.08	0.41
12	12	South	13841958	2.85	0.73
13	13	Southeast	18352320	2.04	0.12
14	14	West	18898740	2.2	0.29

NEW! Data Projects

Data Projects from the text are assignable in MyLab Statistics and provide opportunities for students to practice statistical thinking beyond the classroom. **StatCrunch Projects** that either span the entire curriculum or focus on certain key concepts are also assignable in MyLab Statistics and encourage students to apply concepts to real situations and make data-informed decisions.

UPDATED! Conceptual Questions

The Conceptual Question Library in MyLab Statistics includes 1,000 assignable questions that assess conceptual understanding. These questions are now correlated by chapter to make it easier than ever to navigate and assign these types of questions.



Suppose you have the bar graph to the right showing foreclosure rates for a few select states. If you want to exaggerate the differences in the states' foreclosure rates, how would you change the graph?

State	Foreclosure Rate
State 1	3.5%
State 2	4.5%
State 3	3.0%
State 4	4.8%
State 5	5.2%

- A. To exaggerate the differences in foreclosure rates, change the scale on the y-axis to cover a smaller range, 2-6 percent, for example.
- B. To exaggerate the differences in foreclosure rates, draw the bar graph horizontally.
- C. To exaggerate the differences in foreclosure rates, graph the percents as decimals.
- D. To exaggerate the differences in foreclosure rates, change the scale on the y-axis to cover a larger range, 0-20 percent, for example.

Resources for Success

Student Resources

StatCrunch

StatCrunch® is powerful web-based statistical software that allows users to collect, crunch, and communicate with data. The vibrant online community offers tens of thousands of shared data sets for students and instructors to analyze, in addition to all of the data sets in the text or online homework. StatCrunch is integrated directly into MyLab Statistics or it can be purchased separately. Learn more at www.statcrunch.com.

Video Resources

Chapter Review videos walk students through solving some of the more complex problems and review key concepts from each chapter. Data Cycle of Everyday Things videos demonstrate for students that data collection and data analysis can be applied to answer questions about everyday life. StatTalk Videos, hosted by fun-loving statistician Andrew Vickers, demonstrate important statistical concepts through interesting stories and real-life events. Assessment questions for each video are also available.

Data Sets

All data sets from the textbook are available in MyLab Statistics. They can be analyzed in StatCrunch or downloaded for use in other statistical software programs.

Statistical Software Support

Instructors and students can copy data sets from the text and MyLab Statistics exercises directly into software such as StatCrunch or Excel®. Students can also access instructional support tools including tutorial videos, Study Cards, and manuals for a variety of statistical software programs including, StatCrunch, Excel, Minitab®, JMP®, R, SPSS, and TI 83/84 calculators.

Student Solutions Manual

Written by James Lapp, this manual provides detailed, worked-out solutions to all odd-numbered text exercises. It is available in print and can be downloaded from MyLab Statistics. (ISBN-13: 978-0-13-518923-8; ISBN-10: 0-13-518923-3)

Instructor Resources

Instructor's Edition

Includes answers to all text exercises, as well as a set of Instructor Notes at the front of the text that offer chapter-by-chapter teaching suggestions and commentary. (ISBN-13: 978-0-13-516300-9; ISBN-10: 0-13-516300-5)

Instructor Solutions Manual

Written by James Lapp, the Instructor Solutions Manual contains worked-out solutions to all text exercises. It can be downloaded from MyLab Statistics or from www.pearson.com.

PowerPoint Slides

PowerPoint slides provide an overview of each chapter, stressing important definitions and offering additional examples. They can be downloaded from MyLab Statistics or from www.pearson.com.

TestGen

TestGen® (www.pearson.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, and modify test bank questions or add new questions. It is available for download from Pearson's online catalog, www.pearson.com. The questions are also assignable in MyLab Statistics.

Learning Catalytics

Now included in all MyLab Statistics courses, this student response tool uses students' smartphones, tablets, or laptops to engage them in more interactive tasks and thinking during lecture. Learning Catalytics™ fosters student engagement and peer-to-peer learning with real-time analytics. Access pre-built exercises created specifically for statistics.

Question Libraries

In addition to StatCrunch Projects and the Conceptual Question Library, MyLab Statistics also includes a Getting Ready for Statistics library that contains more than 450 exercises on prerequisite topics.

Statistical Software Bundle Options

Minitab and Minitab Express™

Bundling Minitab software with educational materials ensures students have access to the software they need in the classroom, around campus, and at home. And having 12-month access to Minitab and Minitab Express ensures students can use the software for the duration of their course. ISBN 13: 978-0-13-445640-9
ISBN 10: 0-13-445640-8

JMP Student Edition

An easy-to-use, streamlined version of JMP desktop statistical discovery software from SAS Institute, Inc. is available for bundling with the text. (ISBN-13: 978-0-13-467979-2 ISBN-10: 0-13-467979-2)

XLSTAT™

An Excel add-in that enhances the analytical capabilities of Excel. XLSTAT is used by leading businesses and universities around the world. It is available to bundle with this text. For more information go to www.pearsonhighered.com/xlstatupdate. (ISBN-13: 978-0-321-75940-5; ISBN-10: 0-321-75940-0)

Index of Applications

BIOLOGY

age and weight, 197, 206
animal gestation periods, 75
animal longevity, 75
arm spans, 71, 310–311, 707–709, 710–711
baby seal length, 278, 279–280, 282, 283
birthdays, 252, 259
birth lengths, 130, 310, 314, 316
birth weights, 310, 315, 488
blood types, 371
body temperature, 314, 496
boys' foot length, 310, 316
boys' heights, 312
brain size, 498
caloric restriction of monkeys, 526–527
cats' birth weights, 312
children's ages and heights, 208
color blindness, 371
cousins, 194
elephants' birth weights, 312
eye color, 261
finger length, 32–33, 542
gender of children, 252, 256, 306, 312
grandchildren, 258
hand and foot length, 196–197
handedness, 33, 258
height and age, 721
height and arm spans, 195–196, 197, 707–709, 710–711
height prediction, 726
heights and weights, 160–162, 191, 204, 727
heights of adults, 139
heights of children, 112–113
heights of college women, 132
heights of females, 492
heights of men, 132–133, 311, 312, 492, 497
heights of sons and dads, 135
heights of students and their parents, 499–500
heights of 12th graders, 490–491
heights of women, 133, 307–308, 311, 312, 497
heights of youths, 134
hippopotamus gestation periods, 311
human body temperatures, 492, 500–501, 679
life expectancy and gestation periods for animals, 729
life on Mars, 237–238
mother and daughter heights, 195
newborn hippo weights, 311
night shifts, 547
pregnancy length, 307
sea lion health, 563
siblings, 72
St. Bernard dogs' weights, 309
stem cell research, 81
student and parent heights, 725
tomato plants and colored light, 593–594, 595
tomato plants and fertilizer, 633
whales' gestation periods, 309
women's foot length, 310

BUSINESS AND ECONOMICS

apartment rents, 588, 592, 599–600
baseball salaries, 495
baseball strike, 134
CEO salaries, 82
college costs, 55–56, 71, 130, 191, 449–451, 455–456, 461–463, 477–478, 490, 613–614
consumer price index, 135, 143
earnings and gender, 128, 195

eBay auctions, 705–706
economic class, 60–61, 62
fast food employee wages, 116
food security, 545
gas prices, 95, 100–101, 109
gas taxes, 138–138
grocery delivery, 499
home prices, 130–131, 134, 150, 184–185, 190, 194, 195, 726
houses with swimming pools, 128
income in Kansas, 488
industrial energy consumption, 133
Internet advertising, 386–387, 388
law school tuition, 75–76
logging, 726
movie budgets, 207, 724, 725
post office customers, 70
poverty, 20–21, 136, 207–208
rents in San Francisco, 74
shrinking middle class, 62
wedding costs, 129

CRIME AND CORRECTIONS

capital punishment, 134–135, 259, 368
FBI, 373
gender and type of crime, 541–542
incarceration rates, 33
jury duty, 258
juvenile delinquency, 672
marijuana legalization, 260, 315, 372, 546–547
parental training and criminal behavior of children, 550
race of defendant, 551
recidivism rates, 261, 292–293, 636
“Scared Straight” programs, 37
stolen bicycles, 291–292
stolen cars, 17

EDUCATION

ACT scores, 141, 191
age and credits, 189
age and gender of psychology majors, 78
age and GPA, 189
alumni donations, 636
bar exam pass rates, 45–46, 53, 206, 260–261, 721–722
college costs, 55–56, 71, 130, 191, 449–451, 455–456, 461–463, 477–478, 490, 613–614
college enrollment, 371, 375, 457
college graduation rates, 136–137, 359, 360–362, 369, 370
college majors, 79
college tours, 635
community college applicants, 77
course enrollment rates, 34
credits and GPA, 190
educational attainment, 140, 428
embedded tutors, 419, 420, 421
employment after law school, 80–81, 206, 420
entry-level education, 77
exam scores, 105, 141, 203, 206, 209, 253, 487, 499
exercise and language learning, 35
faculty-to-student ratio, 634
final exam grades, 139
4th-grade reading and math scores, 202–203
GPA, 168–170, 189, 190, 194, 311, 490, 491, 594, 595, 597, 678, 726–727
grades and student employment, 202

guessing on tests, 252
heights and test scores, 206
high school graduation rates, 207–208, 313, 374, 377, 545–546
law school selectivity and employment, 206
law school tuition, 75–76
life expectancy and education, 194
LSAT scores, 191, 206
marital status and education, 223, 224–225, 226, 231, 235
math scores, 93–94
MCAT scores, 311
medical licensing, 314
medical school acceptance, 192, 490
medical school GPAs, 311, 490
multiple-choice exams, 253, 254, 256, 262, 421
music practice, 79
note taking, 582–583
opinion about college, 260
of parents, 724–725, 728
party affiliation and education, 547
passing bar exam, 138
Perry Preschool, 374, 375, 377, 545–546, 550
piano practice, 681
postsecondary graduation rates, 176–177
poverty and high school graduation rates, 207–208
professor evaluation, 194, 634
reading comprehension, 594
reading scores and teaching method, 595
relevance, 544
salary and education, 190, 207
SAT scores, 74, 141, 168–170, 175–176, 194, 203, 307, 308–309, 310, 312, 315, 316–317, 632, 726–727
school accountability, 729
school bonds, 376
semesters and credits, 722
spring break, 324, 363
state and federal spending on, 724
student ages, 141, 487, 489, 493
student gender, 427
student loans, 420, 428, 542
student records, 634
student-to-teacher ratio at colleges, 41–42, 66
study hours, 208, 591, 592, 595
success rates and retention, 725
teacher effectiveness, 251
teacher pay, 201–202
textbook prices, 673
travel time to school, 491, 594–595
true/false tests, 256, 425, 429, 430
tutoring and math grades, 34
vacations and education, 252
years of formal education, 72

EMPLOYMENT

age discrimination, 421
CEO salaries, 82
commuting, 253
duration of employment, 488
earnings and gender, 195
employment after law school, 80–81, 206, 420
gender discrimination in tech industry, 253, 263
grades and student employment, 202
harassment in workplace, 550
law school selectivity and employment, 206
personal care aides, 33

retirement age, 139
 salaries, 189, 190, 194–195, 200, 207,
 593, 679, 723
 self-employment, 428
 sleep and economic group, 569–570
 teacher pay, 201–202
 technology and, 375
 textbook prices, 74
 turkey costs, 201
 unemployment rates, 140
 wages of twins, 723–724
 work and sleep, 190
 work and TV, 190

ENTERTAINMENT

Broadway ticket prices, 141, 493
 cable TV subscriptions, 34
 commercial radio formats, 78
 concerts, 634
 DC movies, 133
 Marvel movies, 133
 movie budgets, 207, 724, 725
 movie ratings, 12–13, 199, 656
 movies with dinner, 32
 movie ticket prices, 493
 MP3 song lengths, 114–115
 music listening, 591–592, 611–612
 streaming TV, 259
 streaming video, 139
 work and TV, 190

ENVIRONMENT

Chicago weather, 311
 city temperatures, 130, 315
 climate change, 261, 525–526
 concern over nuclear power, 595–596
 daily temperatures, 106
 environmental quality, 426
 environment vs. energy development, 371
 global warming, 80, 259, 423
 New York City weather, 311
 opinions on nuclear energy, 81
 pollution index, 132, 142
 pollution reduction, 498
 rainfall, 677–678
 river lengths, 129
 satisfaction with, 375, 429
 smog levels, 96, 99, 102–103
 snow depth, 306

FINANCE

age and value of cars, 177–178
 alumni donations, 636
 Broadway ticket prices, 141, 493
 car insurance and age, 198
 cell phone bills, 674–675
 charitable donations, 597–598
 credit card balances, 670, 673–674,
 676–677
 financial incentives, 425
 gas prices, 95, 100–101, 109, 588–589, 592
 health insurance, 33–34
 home prices, 130–131, 134, 150, 184–185, 190,
 194, 195, 726
 investing, 200
 life insurance and age, 198
 millionaires, 199
 movie ticket prices, 493
 online grocery prices, 679
 professional sport ticket prices, 136, 140–141
 retirement income, 488
 tax rates, 75
 textbook prices, 495, 496, 673
 ticket prices, 679
 train ticket prices, 194
 vacation rentals, 699–701

FOOD AND DRINK

alcoholic drinks, 76, 131, 201, 496–497, 499,
 547, 698
 beer, 76, 496–497, 499
 bottled vs. tap water, 419
 breakfast habits, 542
 butter taste test, 427
 butter vs. margarine, 424
 caloric restriction of monkeys, 526–527
 carrots, 490
 cereals, 70–71, 142
 chain restaurant calories, 139–140
 coffee, 2, 27, 36–37, 630
 Coke vs. Pepsi, 420, 422
 cola taste test, 427
 dairy products and muscle, 630
 diet and depression, 36, 633
 dieting, 475–476, 493, 545
 drink size, 491
 eating out, 497
 energy drinks, 681
 fast food calories, carbs, and sugar, 70, 71,
 204–205
 fast food employee wages, 116
 fast food habits, 542–543
 fat in sliced turkey, 109–110
 fish consumption and arthritis, 635
 fish oil and asthma risk, 35
 food security, 545
 French fries, 498
 frozen dinners, 713
 granola bars, 205
 grocery delivery, 499
 ice cream cones, 267, 302, 498, 679, 681–682
 ice cream preference, 77
 meat-eating behavior, 673
 mercury in freshwater fish, 424, 694
 milk and cartilage, 35
 mixed nuts, 420
 no-carb diet, 425
 nutrition labels, 373
 online grocery prices, 679
 orange juice prices, 130
 oranges, 490
 organic products, 372, 375
 peanut allergies, 549
 picky eaters, 373
 pizza size, 454–455
 popcorn, 508, 536–537, 563–564, 576–577
 potatoes, 491, 492
 salad and stroke, 36
 skipping breakfast and weight gain, 25–26
 snack food calories, 206
 soda, 262–263, 356, 420, 675, 677
 sugary beverages, 36, 374, 631
 tea and divergent creativity, 636
 tomatoes, 492
 turkey costs, 201
 vegetarians, 419, 420, 421
 vitamin C and cancer, 34
 water taste test, 427
 wine, 201, 698

GAMES

blackjack, 209
 brain games, 23–24, 622–623, 632
 cards, 234, 251–252
 coin flips, 236–237, 241–242, 252, 255, 257,
 258, 260, 312, 370, 424, 427, 543
 coin spinning, 395, 402, 543
 dice, 220–221, 227, 243–244, 253, 255, 256,
 257, 258, 262, 270–271, 306, 312, 544
 drawing cubes, 260
 dreidel spinning, 544
 gambling, 257–258
 roller coaster endurance, 50

GENERAL INTEREST

apartment rents, 588, 592
 book width, 167, 701–702
 boys' heights, 141
 caregiving responsibilities, 426
 children of first ladies, 129
 children's heights, 141
 dandruff shampoo effectiveness, 611
 DMV wait times, 679
 draft lottery, 593
 drought-resistant plants in front yard, 634
 energy consumption, 133
 ethnicity of active military, 550
 exercise hours, 128
 frequency of *e* in English language, 346–347,
 354–355
 gun availability, 77
 hand folding, 255, 263
 hand washing, 376, 429
 home ownership, 389–390, 391, 393–394
 home sizes, 646–647
 houses with garages, 78
 houses with swimming pools, 128
 improving tips, 620
 libraries, 137, 313, 723, 725
 logging, 726
 marijuana, 252–253, 372–373, 546–547
 morning routine, 670–671
 numbers of siblings, 190
 passports, 313
 pet ownership, 75, 312, 340, 344, 635
 population density, 136
 proportion of *a*'s in English language, 424
 proportion of *r*'s in English language, 424
 reading colored paper, 633, 672–673
 reading habits, 259, 260, 370, 426
 renting vs. buying a home, 354
 research abstracts, 608–609
 residential energy consumption, 132
 roller coaster heights, 128, 138
 seesaw heights, 197
 shoe sizes, 82, 205, 727
 shower duration, 487, 498
 sibling ages, 130
 sitting and brain, 691, 716–718
 superpowers, 254, 424
 tall buildings, 120–121, 128, 142, 207, 721
 thumbtacks, 254, 306
 trash weight and household size, 202, 725
 U.S. population, 722
 vacations, 256, 422, 423
 violins, 543, 544
 waist size and height, 727, 728
 weight of coins, 71

HEALTH

acetaminophen and asthma, 635
 ages of women who give birth, 191
 aloe vera juice, 34
 anesthesia care and postoperative
 outcomes, 637
 antibiotics vs. placebo, 544
 arthritis, 425
 autism and MMR vaccine, 35–36, 637
 beliefs about, 662–663
 blood pressure, 200, 495, 590
 BMI, 71, 492
 calcium levels in the elderly, 480–481
 cardiovascular disease and gout, 429
 causes of death, 63
 cervical cancer, 256
 cholesterol, 209, 492, 590, 598
 civic engagement and health, 630–631
 coffee and cancer, 2, 27
 coronary artery bypass grafting, 638
 CPR in Sweden, 550

Crohn's disease, 26, 362
 diabetes, 372, 421, 544, 638
 ear infections, 37
 embryonic stem cell use, 357
 exercise and weight loss, 632
 fast eating and obesity, 634
 fish consumption and arthritis, 635, 636
 fish oil, 35, 374
 fitness among adults, 78
 flu vaccine, 419, 420, 632
 glucose readings, 70
 glycemic load and acne, 35
 hand washing, 429
 health insurance, 33–34
 heart attack prevention, 632–633
 heart rate, 497, 498–499
 HIV treatment, 425–426
 hormone replacement therapy, 79–80
 hospital readmission, 421
 hospital rooms, 635
 hours of sleep, 71
 HPV vaccination, 545
 ideal weight, 81–82
 identifying sick people, 436, 482–483
 infant formula and diabetes risk, 638
 intravenous fluids, 635
 lead exposure, 671–672
 life expectancy and education, 194
 light exposure effects, 37
 low-birth-weight babies, 132
 marijuana use and bone density, 630, 631
 medical group, 251
 melanoma, 638
 men's health, 399, 400–401
 mercury in freshwater fish, 424
 milk and cartilage, 35
 multiple sclerosis treatment, 547
 mummies with heart disease, 544
 music and pain, 611–612
 no-carb diet, 425
 obesity, 77, 634
 ondansetron for nausea during pregnancy, 375
 opioid crisis, 429
 peanut allergies, 549
 personal data collection, 9
 pet ownership and cardiovascular disease, 635, 637
 pneumonia vaccine for young children, 35
 pregnancy lengths, 132
 preventable deaths, 76–77
 pulse rates, 69, 70, 71, 444–445, 493–494, 495, 501, 591, 673, 679
 racket sports and health, 631
 red blood cells, 310
 rheumatoid arthritis treatment, 631
 salad and stroke, 36
 scorpion antivenom, 548
 sea lion health, 563
 SIDS, 214–215, 247
 skipping breakfast and weight gain, 25–26
 sleep hours, 78, 79, 140, 190–191, 206, 520–522, 569–570, 594, 674, 678
 smell sense, 651–652
 smoking, 82, 131, 197–198, 371, 618
 smoking cessation, 426, 631, 634
 sodium levels, 128
 stroke, 36, 37, 631, 633
 sugary beverages and brain health, 36, 631
 swine flu hospitalizations, 531–533
 systolic blood pressures, 314
 treating depression, 34–35
 triglycerides, 71, 494–495, 598
 vitamin C and cancer, 34
 vitamin D and osteoporosis, 37
 weight gain during pregnancy, 130
 white blood cells, 310
 wine consumption and mortality, 698
 yoga and cellular aging, 635, 636

LAW

capital punishment, 259, 368
 gun laws, 259, 371, 428
 jury duty, 258, 315
 jury pool, 420
 marijuana legalization, 252–253, 260, 315, 372, 546–547
 three-strikes law, 428
 trust in judiciary, 373
 trust in legislative branch, 376

POLITICS

climate change, 261
 common ground between political parties, 410–411
 concern over nuclear power, 595–596
 education and party affiliation, 547
 equal rights for women, 254–255
 free press, 372
 gender and party affiliation, 548
 party and opinion about right direction, 544
 political debates, 383, 415
 political parties, 254, 410–411, 548
 presidential elections, 345–346, 373, 376, 428
 presidents' ages, 493
 unpopular views in a democracy, 372
 voters' polls, 376
 votes for independents, 429
 voting, 370, 371–372
 young voters, 424

PSYCHOLOGY

adult abusiveness and viewing TV violence as a child, 515
 age and gender of psychology majors, 78
 body image, 78
 brain games, 23–24, 632
 coffee and depression, 630
 confederates and compliance, 36
 diet and depression, 36, 633
 dreaming, 376, 421
 epilepsy treatment, 632
 expression of feelings, 631
 extrasensory perception, 288–289, 293–296, 369, 427, 429
 extraversion and sports, 675–676, 679
 financial incentives, 425
 happiness, 373, 548, 594, 597, 675, 676
 ketamine and social anxiety disorder, 636–637, 638
 music and divergent thinking, 636
 neurofeedback and ADHD, 37
 parental training and criminal behavior of children, 550
 poverty and IQ, 20–21
 psychological distance and executive function, 637
 psychometric scores, 172
 rat experiment, 549
 reaction distance, 492, 499, 595
 reaction times, 593
 risk perception, 91, 124–125
 robot cockroaches, 550–551
 schizophrenia treatment, 631
 smiling, 545, 632
 stress, 315, 370
 Stroop effect, 633–634, 672
 sugary drinks and brain health, 631
 tea and divergent creativity, 636
 treating depression, 34–35
 unusual IQs, 132
 yawning, 640, 664–665

SOCIAL ISSUES

age and marriage, 32, 673
 body piercings, 56–57
 contact with mother, 597, 679, 680
 gay marriage, 428
 happiness of marriage and gender, 545
 marital status and blood pressure, 590
 marital status and cholesterol, 590
 marital status and education, 223, 224–225, 226, 231, 235
 marriage and divorce rates, 34, 386
 online dating, 259
 opioid crisis, 429
 percentage of elderly, 34
 phubbing and relationship satisfaction, 635
 population density, 33
 right of way, 412–414
 same-sex marriage, 513, 546
 secondhand smoke exposure and young children, 36
 sexual harassment, 332
 spring break, 324, 363
 yoga and high-risk adolescents, 37

SPORTS

athletes' weights, 130
 baseball, 259–260
 baseball players' ages and weights, 728
 baseball position and hits, 589
 baseball runs scored, 131–132, 674, 680–681
 baseball salaries, 495
 baseball strike, 134
 basketball free-throw shots, 300–301, 312, 313–314
 batting averages, 373
 car race finishing times, 174–175
 college athletes' weights, 495
 college athletics, 371
 deflated footballs, 492–493
 energy drinks, 681
 exercise and study hours, 208, 671
 exercise and weight loss, 632
 extraversion and sports, 675–676
 fitness, 544–545
 GPA and gym use, 194
 heights of basketball players, 500
 marathon finishing times, 51, 118, 139
 MLB pitchers, 199–200
 MLB player ages, 134
 Olympics, 129, 130, 372
 Olympic viewing, 423
 predicting home runs, 202
 predicting 3-point baskets, 202
 professional basketball player weights, 142
 professional sport ticket prices, 136, 140–141
 race times, 141–142
 racket sports and health, 631
 RBIs, 491
 reaction times, 593
 speed skating suits, 632
 steps walked and calories burned, 713–714
 stretching and athletic performance, 605, 626–627
 Super Bowl, 371
 surfing, 129–130, 497
 team uniform color, 559, 584–585
 tennis winning percentage, 197
 200-meter run, 129
 weights of athletes, 74, 499, 675, 728
 working out, 422–423

SURVEYS AND OPINION POLLS

age and Internet, 326
 alien life, 372
 artificial intelligence, 373
 baseball, 259–260

cell phone security, 260
 college graduation rates, 359, 360–362
 common ground between political parties, 410–411
 concern over nuclear power, 595–596
 data security and age, 230
 diabetes, 372
 embryonic stem cell use, 357
 environmental satisfaction, 375
 environment vs. energy
 development, 371
 equal rights for women, 254–255, 259
 FBI, 373
 freedom of religion, 426
 freedom of the press, 372, 426
 happiness, 373
 marijuana legalization, 252–253, 315, 372
 marijuana use, 372–373
 news sources, 255, 301, 375
 nutrition labels, 373
 online presence, 254
 opinion about college, 260
 opinion on same-sex marriage, 513
 opinions on nuclear energy, 81
 organic products, 372, 375
 picky eaters, 373
 presidential elections, 345–346, 428
 reading habits, 259, 260, 315, 370
 renting vs. buying a home, 354
 satisfaction with environment, 429
 sexual harassment, 332
 social media use, 427
 streaming TV, 259
 stress, 255, 315, 370
 sugary sodas, 356
 teachers and digital devices, 352
 technology anxiety, 375
 television viewing, 428
 travel by Americans, 313
 trust in executive branch, 376
 trust in judiciary, 373
 trust in legislative branch, 376

unpopular views in a democracy, 372
 vacations, 256
 voters' polls, 376
 watching winter Olympics, 372

TECHNOLOGY

audio books, 426, 634
 cell phones, 79, 256, 260, 313, 488
 diet apps, 545
 drones, 313
 employment and, 375
 Facebook, 32, 227–228, 427
 fitness apps, 545, 552
 gender discrimination in tech
 industry, 253, 263
 Instagram, 371
 Internet advertising, 386–387, 388
 Internet browsers, 78
 Internet usage, 326, 543
 iPad batteries, 452
 iTunes music, 440
 landlines, 313
 Netflix cheating, 370
 news sources, 423, 429
 online dating, 259, 260
 online presence, 254
 online shopping, 259
 phubbing and relationship satisfaction, 635
 reading electronics, 472–474
 robots, 580–581
 social media, 32, 83, 227–228, 371, 424, 427, 596
 streaming TV, 370
 teachers and digital devices, 352
 texting/text messages, 76, 200, 313, 429, 543, 680
 TV ownership, 493, 494, 501
 TV viewing, 499, 591, 598, 671
 Twitter, 424
 vehicle sales, 550
 virtual reality and fall risk, 37
 voice-controlled assistants, 315

TRANSPORTATION

age and value of cars, 177–178, 489
 airline arrival times, 312
 airline ticket prices, 193–194, 198–199
 airport screeners, 239–240
 car insurance and age, 198
 car MPG, 72, 201
 commuting times, 588, 592, 594–595
 crash-test results, 7
 DMV wait times, 679
 driver's licenses, 369, 370, 377
 driving exam, 254, 259, 261, 314
 electric car charging stations, 634
 flight times/distances, 209
 fuel economy, 572
 fuel-efficient cars, 207
 gas prices, 95, 100–101, 588–589, 592
 gas taxes, 138
 hybrid car sales, 420
 miles driven, 488
 monthly car costs, 72
 MPH, 82
 parking tickets, 71
 pedestrian fatalities, 34
 plane crashes, 424
 red cars and stop signs, 542
 right of way, 412–414
 seat belt use, 14–16, 422
 self-driving cars, 421, 497
 speeding, 37, 139
 stolen cars, 17
 teen drivers, 419
 texting while driving, 313, 429, 543
 traffic cameras, 81
 traffic lights, 262
 train ticket prices, 194
 travel time to school, 491
 turn signal use, 374–375
 used cars, 155, 720, 726
 waiting for a bus, 272–273

1

Introduction to Data



THEME

Statistics is the science of data, so we must learn the types of data we will encounter and the methods for collecting data. The method used to collect data is very important because it determines what types of conclusions we can reach and, as you'll learn in later chapters, what types of analyses we can do. By organizing the data we've collected, we can often spot patterns that are not otherwise obvious.

This text will teach you to examine data to better understand the world around you. If you know how to sift data to find patterns, can communicate the results clearly, and understand whether you can generalize your results to other groups and contexts, you will be able to make better decisions, offer more convincing arguments, and learn things you did not know before. Data are everywhere, and making effective use of them is such a crucial task that one prominent economist has proclaimed statistics one of the most important professions of the decade (*McKinsley Quarterly* 2009).

The use of statistics to make decisions and convince others to take action is not new. Some statisticians date the current practice of statistics back to the mid-nineteenth century. One famous example occurred in 1854, when the British were fighting the Russians in the brutal Crimean War. A British newspaper had criticized the military medical facilities, and a young but well-connected nurse, Florence Nightingale, was appointed to study the situation and, if possible, to improve it.

Nightingale carefully recorded the numbers of deaths, the causes of the deaths, and the times and dates of the deaths. She organized these data graphically,

and these graphs enabled her to see a very important pattern: A large percentage of deaths were due to contagious disease, and many deaths could be prevented by improving sanitary conditions. Within six months, Nightingale had reduced the death rate by half. Eventually she convinced Parliament and military authorities to completely reorganize the medical care they provided. Accordingly, she is credited with inventing modern hospital management.

In modern times, we have equally important questions to answer. Do cell phones cause brain tumors? Are alcoholic drinks healthful in moderation? Which diet works best for losing weight? What percentage of the public is concerned about job security? **Statistics**—the science (and art!) of collecting and analyzing observations to learn about ourselves, our surroundings, and our universe—helps answer questions such as these.

Data are the building blocks of statistics. This chapter introduces some of the basic types of data and explains how we collect them, store them, and organize them. You'll be introduced to the Data Cycle, a guide for how to interact with data in a productive way. These ideas and skills will provide a basic foundation for your study of the rest of the text.

CASE STUDY

Dangerous Habit?

Will your coffee habit give you cancer? A court in California considered whether Californians' morning cup of coffee should include a health warning. In 1986, Californians voted for the Safe Drinking Water and Toxic Enforcement Act, which requires that products that contain harmful chemicals be labeled as hazardous. Coffee contains a chemical, acrylamide, that, in the official jargon "is known to the State of California to cause cancer." In 2010, a lawyer sued the coffee industry to force companies to either label their product as hazardous or to remove the chemical from their product. As of the date of publication of this book, the lawsuit continues. Complicating this lawyer's efforts is the fact that recent research suggests that drinking coffee is possibly beneficial to our health and maybe even reduces the risk of cancer.

Does coffee cause cancer? Does it prevent cancer? Why are there conflicting opinions? In this chapter we explore questions such as these, and consider the different types of evidence needed to make *causal* claims, such as the claim that drinking coffee will give you cancer.

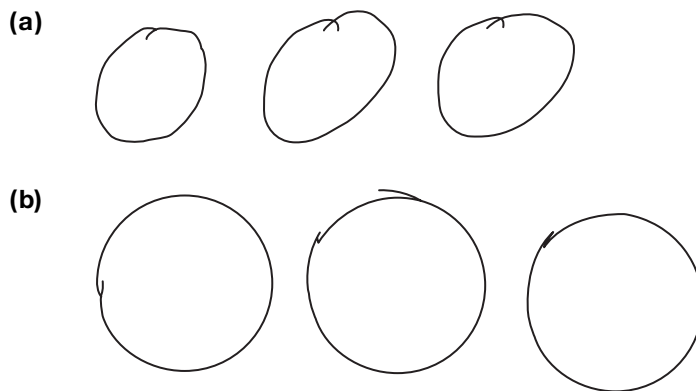


SECTION 1.1

What Are Data?

The study of statistics rests on two major concepts: variation and data. **Variation** is the more fundamental of these concepts. To illustrate this idea, draw a circle on a piece of paper. Now draw another one, and try to make it look just the same. Now another. Are all three exactly the same? We bet they're not. They might be slightly different sizes, for instance, or slightly different versions of round. This is an example of variation. How can you reduce this variation? Maybe you can get a penny and outline the penny. Try this three times. Does variation still appear? Probably it does, even if you need a magnifying glass to see, say, slight variations in the thickness of the penciled line.

Data are observations that you or someone else (or *something* else) records. The drawings in Figure 1.1 are data that record our attempts to draw three circles that look the same. Analyzing pictorial data such as these is not easy, so we often try to quantify such observations—that is, to turn them into numbers. How would you measure whether these three circles are the same? Perhaps you would compare diameters or circumferences, or somehow try to measure how and where these circles depart from being perfect circles. Whatever technique you chose, these measurements could also be considered data.



 **Details**

Data Are What Data Is

If you want to be “old school” grammatically correct, then the word *data* is plural. So we say “data *are*” and not “data *is*.” The singular form is *datum*. However, this usage is changing over time, and some dictionaries now say that *data* can be used as both a singular and a plural noun.

◀ **FIGURE 1.1** (a) Three circles drawn by hand. (b) Three circles drawn using a coin. It is clear that the circles drawn by hand show more variability than the circles drawn with the aid of a coin.

Data are more than just numbers, though. David Moore, a well-known statistician, defined data as “numbers in context.” By this he meant that data consist not only of the numbers we record but also of the story behind the numbers. For example,

10.00, 9.88, 9.81, 9.81, 9.75, 9.69, 9.5, 9.44, 9.31

are just numbers. But in fact these numbers represent “Weight in pounds of the ten heaviest babies in a sample of babies born in North Carolina in 2004.” Now these numbers have a context and have been elevated into data. See how much more interesting data are than numbers?

These data were collected by the state of North Carolina in part to help researchers understand the factors that contribute to low-weight and premature births. If doctors understand the causes of premature birth, they can work to prevent it—perhaps by helping expectant mothers change their behavior, perhaps by medical intervention, and perhaps by a combination of both.

KEY POINT

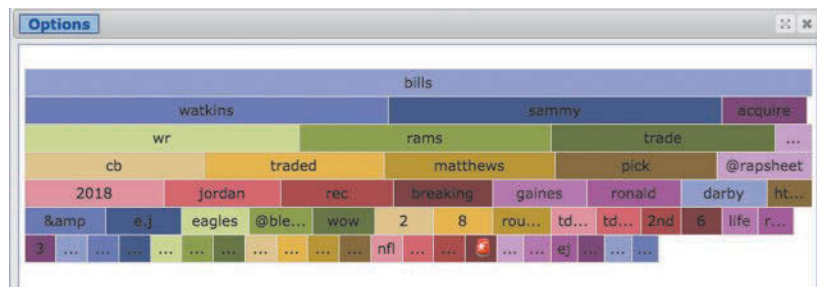
Data are “numbers in context.”

Data play a pivotal role in our economy, culture, and everyday lives. Much of this textbook is concerned with data collected for what we might call “professional” purposes, such as answering scientific questions or making business decisions. But in fact, data are everywhere. Google, for example, saves every search you make and combines this with data on which links you click in order to improve the way it presents information (and, of course, to determine which advertisements will appear on your search results).

In this book you’ll see data from a variety of sources. For example, thanks to small, portable sensors, you can now join the “Personal Data Movement.” Members of this movement record data about their daily lives and analyze it in order to improve their health, to run faster, or just to make keepsakes—a modern-day scrapbook of personal data visualizations. Maybe you or a friend uses a smart watch to keep track of your runs. One of the authors of this text carries a FitBit to record his daily activity. From this he learned that he typically takes 2500 more steps on days that he lectures than on days that he does not.

Speaking of Twitter, did you know every tweet in the “twitterverse” is saved and can be accessed? Twitter, like many other websites, provides what’s called an API for accessing data. API stands for Application Program Interface, and it’s basically a language that allows programmers to communicate with websites in order to access data that the website wishes to make public. For example, the statistical analysis software package StatCrunch makes use of an API provided by Twitter to create a “Word Wall” on whatever keywords you choose to type or show you currently trending tags. See Figures 1.2 and 1.3.

▼ FIGURE 1.2 Trending Tags on a day in August 2017. Can you guess what day of the week this was?



▲ FIGURE 1.3 A StatCrunch-generated “word wall” showing most common words appear in tweets that include the word *Bills*.











Another common way to store data on the Internet is using HTML tables. HTML (Hyper Text Markup Language) is another example of a software language that tells your browser how to display a web page. HTML tells the browser which words are “headers,” which are paragraphs, and which should be displayed in a table. For example, while reading a Wikipedia article about coffee, the authors came across a table showing coffee production by country. These data are stored in an HTML file. This table is relatively small, and so it is simple to enter it into statistical analysis software. But other tables are quite large, and software packages must be employed to “scrape” the data (Figure 1.4).

Sometimes you can find data by turning to your government. The website data.gov has over 197,000 data sets available. The city of Miami, Florida, is one of many cities around the United States that provides data on a variety of topics. Figure 1.5 shows the first few rows of a data set that provides salaries for roughly 28,000 employees of the city of Miami.

Data provided through an open data portal, such as these data, can be stored in a variety of formats. These data can be downloaded as “CSV,” “CSV for Excel,” “JSON,” “XML,” and some other formats as well. You don’t need to worry about these, but

Production [edit]

Top ten green coffee producers in 2014

Rank	Country	Teragrams ^[8]
1	 Brazil	2.8
2	 Vietnam	1.4
3	 Colombia	0.7
4	 Indonesia	0.6
5	 Ethiopia	0.4
6	 India	0.3
7	 Honduras	0.3
8	 Guatemala	0.2
9	 Peru	0.2
10	 Uganda	0.2
	World	8.8

StatCrunchThis Output

URL:

<https://en.wikipedia.org/wiki/Coffee>

Table 2								
Row	Rank	Country	Teragrams[8]	var4	var5	var6	var7	var8
1	1	Brazil	2.8					
2	2	Vietnam	1.4					
3	3	Colombia	0.7					
4	4	Indonesia	0.6					
5	5	Ethiopia	0.4					
6	6	India	0.3					
7	7	Honduras	0.3					
8	8	Guatemala	0.2					
9	9	Peru	0.2					
10	10	Uganda	0.2					
11		World	8.8					
12								
13								
14								

▲ FIGURE 1.4 Coffee production HTML table from en.wikipedia.org/wiki/Coffee (viewed on August 11, 2017; left) and the same table “scraped” by StatCrunch (right).

Title	Department	Annual Salary	Gross Pay Last Year	Gross Year 1	Gross Year 2
W&S SEWER WATER AND		42044.6	3411.87	45630.98	44282.92
AIRPORT OPERATIONS	AVIATION	50068.72	2086.67	36306	35244.41
ASST W&S WATER AND		96744.96	3764.42	60230.72	58444.64
BUS OPERATIONS	TRANSPORT	46332	2183.58	38911.15	37764.66
CORRECTIONS	CORRECTIONS	102386.44	4079.85	65808.99	63897.41
LANDSCAPE MAINTENANCE	PARKS, RECREATION	43264	1707.46	29852.11	28977.41
TREATMENT PLANTS	WATER AND SEWER	66755.52	4175.56	60779.63	58977.12
CORRECTIONS	CORRECTIONS	77477.92	3023.38	49120.06	47671.78
ENVIRONMENTAL SERVICES	REGULATORY	63657.36	2486.82	39789.12	38613.92
BUS OPERATIONS	TRANSPORT	50616.8	2701.54	43895.25	42599.25
SR SYSTEMS	INFORMATION TECHNOLOGY	113020.96	4410.42	70266.72	68180.16
ADMINISTRATIVE SERVICES	PARKS, RECREATION	83850.26	3563.23	50379.58	48889.04
BUS MAINTENANCE	TRANSPORT	49576.8	2501.1	46157.56	44803.71
POLICE OFFICERS	POLICE	85223.06	5221.32	82099.06	79708.32
SEAPORT ENGINEERS	SEAPORT	63283.74	2979.46	60668.57	58903.53
POLICE RESERVE	POLICE	49352.42	1941.63	33237.68	32284.86

◀ FIGURE 1.5 The first few rows of public employee salary data sets provided by the city of Miami, Florida.

for most applications, “CSV,” which stands for “comma-separated values” will be understandable by most data analysis packages. (For example, Excel, Minitab, and StatCrunch can all import CSV files.)

You won’t have to scrape and download data on your own in order to do the examples and exercises in this textbook (unless you want to, of course). The data you need are provided for you, ready to upload into one of several common statistical analysis packages (Excel, Minitab, StatCrunch, and the TI-84 graphing calculator). However, this book includes projects that might lead you to uncharted waters, and so you should be aware that different data storage types exist.

In the next section, you’ll see that you can store data in different structures, and some structures are particularly helpful in some circumstances.

What Is Data Analysis?

In this text you will study the science of data. Most important, you will learn to analyze data. What does this mean? You are analyzing data when you examine data of some sort and explain what they tell us about the real world. In order to do this, you must first learn about the different types of data, how data are stored and structured, and how they are summarized. The process of summarizing data takes up a big part of this text;

indeed, we could argue that the entire text is about summarizing data, either through creating a visualization of the data or distilling them down to a few numbers that we hope capture their essence.

KEY POINT

Data analysis involves creating summaries of data and explaining what these summaries tell us about the real world.

SECTION 1.2

Classifying and Storing Data



▲ **FIGURE 1.6** A photo of Carhenge, Nebraska.



▲ **FIGURE 1.7** Satellites in NASA's Earth Observing Mission record ultraviolet reflections and transmit these data back to Earth. Such data are used to construct images of our planet. Earth Observer (<http://eos.gsfc.nasa.gov/>).

The first step in understanding data is to understand the different types of data you will encounter. As you've seen, data are numbers in context. But that's only part of the story; data are also recorded observations. Your photo from your vacation to Carhenge in Nebraska is data (Figure 1.6). The ultraviolet images streaming from the Earth Observer Satellite system are data (Figure 1.7). These are just two examples of data that are not numbers. Statisticians work hard to help us analyze complex data, such as images and sound files, just as easily as we study numbers. Most of the methods involve recoding the data into numbers. For example, your photos can be digitized in a scanner, converted into a very large set of numbers, and then analyzed. You might have a digital camera that gives you feedback about the quality of a photo you've taken. If so, your camera is not only collecting data but also analyzing it!

Almost always, our data sets will consist of characteristics of people or things (such as gender and weight). These characteristics are called **variables**. Variables are not “unknowns” like those you studied in algebra. We call these characteristics variables because they have variability: The values of the variable can be different from person to person.

KEY POINT

Variables in statistics are different from variables in algebra. In statistics, variables record characteristics of people or things.

Details

More Grammar

We're using the word *sample* as a noun—it is an object, a collection of data that we study. Later we'll also use the word *sample* as a verb—that is, to describe an action. For example, we'll sample ice cream cones to measure their weight.

When we work with data, they are grouped into a collection, which we call either a **data set** or a **sample**. The word *sample* is important, because it implies that the data we see are just one part of a bigger picture. This “bigger picture” is called a **population**. Think of a population as the Data Set of Everything—it is the data set that contains all of the information about everyone or everything with respect to whatever variable we are studying. Quite often, the population is really what we want to learn about, and we learn about it by studying the data in our sample. However, many times it is enough just to understand and describe the sample. For example, you might collect data from students in your class simply because you want to know about the students in your class, not because you wish to use this information to learn about all students at your school. Sometimes, data sets are so large that they effectively *are* the population, as you'll soon see in the data reflecting births in North Carolina.

Two Types of Variables

While variables can be of many different types, there are two basic types that are very important to this book. These basic types can be broken into small subcategories, which we'll discuss later.

Numerical variables describe quantities of the objects of interest. The values will be numbers. The weight of an infant is an example of a numerical variable.

Categorical variables describe qualities of the objects of interest. These values will be categories. The sex of an infant is an example of a categorical variable. The possible values are the categories “male” and “female.” Eye color of an infant is another example; the categories might be brown, blue, black, and so on. You can often identify categorical variables because their values are *usually* words, phrases, or letters. (We say “usually” because we sometimes use numbers to represent a word or phrase. Stay tuned.)

EXAMPLE 1 Crash-Test Results

The data in Table 1.1 are an excerpt from crash-test dummy studies in which cars are crashed into a wall at 35 miles per hour. Each row of the data set represents the observed characteristics of a single car. This is a small sample of the database, which is available from the National Transportation Safety Administration. The *head injury* variable reflects the risk to the passengers’ heads. The higher the number, the greater the risk.

Make	Model	Doors	Weight	Head Injury
Acura	Integra	2	2350	599
Chevrolet	Camaro	2	3070	733
Chevrolet	S-10 Blazer 4X4	2	3518	834
Ford	Escort	2	2280	551
Ford	Taurus	4	2390	480
Hyundai	Excel	4	2200	757
Mazda	626	4	2590	846
Volkswagen	Passat	4	2990	1182
Toyota	Tercel	4	2120	1138

QUESTION For each variable, state whether it is numerical or categorical.

SOLUTION The variables *make* and *model* are categorical. Their values are descriptive names. The units of *doors* are, quite simply, the number of doors. The units of *weight* are pounds. The variables *doors* and *weight* are numerical because their values are measured quantities. The units for *head injury* are unclear; head injury is measured using some scale that the researchers developed.

TRY THIS! Exercise 1.3



Details

Quantitative and Qualitative Data

Some statisticians use the word *quantitative* to refer to numerical variables (think “quantity”) and *qualitative* to refer to categorical variables (think “quality”). We prefer *numerical* and *categorical*. Both sets of terms are commonly used, and you should be prepared to hear and see both.

◀ **TABLE 1.1** Crash-test results for cars.

Details

Categorical Doors

Some people might consider *doors* a categorical variable, because nearly all cars have either two doors or four doors, and for many people, the number of doors designates a certain type of car (small or larger). There’s nothing wrong with that.

Coding Categorical Data with Numbers

Sometimes categorical variables are “disguised” as numerical. The *smoke* variable in the North Carolina data set (Table 1.2) has numbers for its values (0 and 1), but in fact those numbers simply indicate whether or not the mother smoked. Mothers were asked, “Did you smoke?” and if they answered “Yes,” the researchers coded this categorical response with a 1. If they answered “No,” the response was coded with a 0. These particular numbers represent categories, not quantities. *Smoke* is a categorical variable.

Coding is used to help both humans and computers understand what the values of a variable represent. For example, a human would understand that a “yes” under the

Weight	Female	Smoke
7.69	1	0
0.88	0	1
6.00	1	0
7.19	1	0
8.06	1	0
7.94	1	0

▲ **TABLE 1.2** Data for newborns with coded categorical variables.

! Caution**Don't Just Look for Numbers!**

You can't always tell whether a variable is categorical simply by looking at the data table. You must also consider what the variable represents. Sometimes, researchers code categorical variables with numerical values.

Details**Numerical Categories**

Categories might be numbers. Sometimes, numerical variables are coded as categories, even though we wish to use them as numbers. For example, *number of siblings* might be coded as "none," "one," "two," "three," and so on. Although words are used, this is really a numerical variable since it is counting something.

"Smoke" column would mean that the person was a smoker, but to the computer, "yes" is just a string of symbols. If instead we follow a convention where a 1 means "yes" and a 0 means "no," then a human understands that the 1s represent smokers, and a computer can easily add the values together to determine, for example, how many smokers are in the sample.

This approach for coding categorical variables is quite common and useful. If a categorical variable has only two categories, as do *gender* and *smoke*, then it is almost always helpful to code the values with 0 and 1. To help readers know what a "1" means, rename the variable with either one of its category names. A "1" then means the person belongs to that category, and a 0 means the person belongs to the other category. For example, instead of calling a variable *gender*, we rename it *female*. And then if the baby is a boy we enter the code 0, and if it's a girl we enter the code 1.

Sometimes your computer does the coding for you without your needing to know anything about it. So even if you see the words *female* and *male* on your computer, the computer has probably coded these with values of 0 and 1 (or vice versa).

Storing Your Data

The format in which you record and store your data is very important. Computer programs will require particular formats, and by following a consistent convention, you can be confident that you'll remember the qualities of your own data set if you need to revisit it months or even years later. Data are often stored in a spreadsheet-like format in which each row represents the object (or person) of interest. Each column represents a variable. In Table 1.3, each row represents a movie. The column heads are variables: *Title*, *Rating*, *Runtime*, *Critics Rating*. (The "rating" is the Motion Pictures Association of America rating to indicate the movie's intended audience. The Critics Rating is a score from 0 to 100 from the website Rotten Tomatoes. High scores are good.) This format is sometimes referred to as the **stacked data** format.

Title	Rating	Runtime	Critics Rating
<i>Cars 2</i>	G	106	39
<i>Alvin and the Chipmunks: Chipwrecked</i>	G	87	12
<i>Monsters University</i>	G	104	78
<i>Alice Through the Looking Glass</i>	PG	113	30
<i>Chasing Mavericks</i>	PG	116	31
<i>Despicable Me 2</i>	PG	98	73
<i>Cloudy with a Chance of Meatballs 2</i>	PG	95	70
<i>Hotel Transylvania</i>	PG	91	45

▲ TABLE 1.3 Data for a few movies rated G or PG.

When you collect your own data, the stacked format is almost always the best way to record and store your data. One reason is that it allows you to easily record several different variables for each subject. Another reason is that it is the format that most software packages will assume you are using for most analyses. (The exceptions are the TI-84 and Excel.)

Some technologies, such as the TI calculators, require, or at least accommodate, data stored in a different format, called **unstacked data**. Unstacked data tables are also common in some books and media publications. In this format, each column represents a variable from a different group. For example, one column could represent the length in minutes of movies rated G and another column could represent the length of movies rated PG. The data set, then, is a single variable (*Runtime*) broken into distinct groups. The groups are determined by a categorical variable (in this case, *Rating*.) Table 1.4

Rated G	Rated PG
112	113
90	116
95	95
	98
	91

▲ TABLE 1.4 Movie runtime (in minutes) by rating group (unstacked).

shows an example of the *Runtime* variable in Table 1.3. Figure 1.8 shows the same data in TI-84 input format.

The great disadvantage of the unstacked format is that it can store only two variables at a time: the variable of interest (for example, Runtime) and a categorical variable that tells us which group the observation belongs in (for example, Rating). However, most of the time we record many variables for each observation. For example, recording a movie's title, rating, running time and critics' rating in the stacked format enables us to display as many variables as we wish.

L1	L2	L3	L4	L5	2
112	113	---	---	---	
90	116				
95	95				
---	98				
	91				

▲ FIGURE 1.8 TI-84 data input screen (unstacked data).

EXAMPLE 2 Personal Data Collection

Using a sensor worn around her wrist, Safaa recorded the amount of sleep she got on several nights. She also recorded whether it was a weekend or a weeknight. For the weekends, she recorded (in hours): 8.1, 8.3. For the weeknights she recorded 7.9, 6.5, 8.2, 7.0, 7.3.

QUESTION Write these data in both the stacked format and the unstacked format.

SOLUTION In the stacked format, each row represents a unit of observation, and each column measures a characteristic of that observation. For Safaa, the unit of observation was a night of sleep, and she measured two characteristics: time and whether it was a weekend. In stacked format, her data would look like this:

Time	Weekend
8.1	Yes
8.3	Yes
7.9	No
6.5	No
8.2	No
7.0	No
7.3	No

(Note that you might have coded the “Weekend” variable differently. For example, instead of entering “Yes” or “No,” you might have written either “Weekend” or “Weeknight” in each row.)

In the unstacked format, the numerical observations appear in separate columns, depending on the value of the categorical variable:

Weekend	Weeknight
8.1	7.9
8.3	6.5
	8.2
	7.0
	7.3

See the Tech Tips to review how to enter data like these using your technology.



TRY THIS! Exercise 1.11

! Caution

Look at the Data Set!

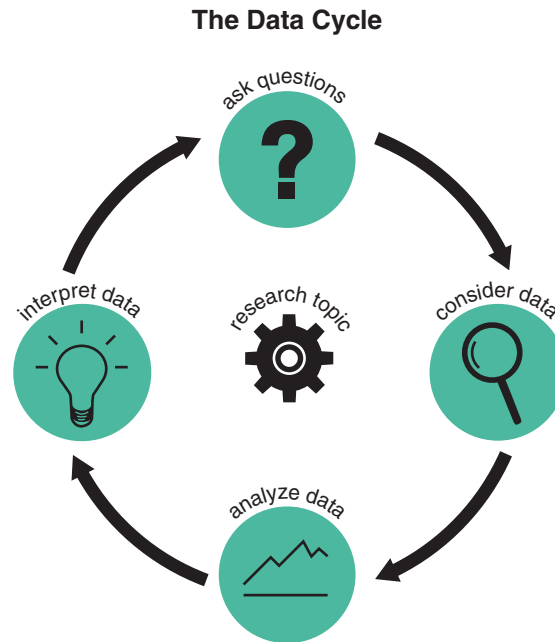
The fact that different people use different formats to store data means that the first step in any data investigation is to look at the data set. In most real-life situations, stacked data are the more useful format because this format permits you to work with several variables at the same time.

SECTION 1.3

Investigating Data

Now that you've seen some examples of data, it's time to learn what to do when you interact with data, as this book will ask you to do. To help guide you, consider the Data Cycle (see Figure 1.9).

► **FIGURE 1.9** The Data Cycle reminds us of the four stages of a statistical investigation. (Designed by Alyssa Brode for the Mobilize Project at UCLA.)



The Data Cycle is a representation of a statistical investigation cycle: the stages we go through when analyzing data. In real life, you might not necessarily go in this order of the stages listed here. In fact, you are likely to alternate between various stages. But it is useful to plan your analysis in this order.

As we move through the book, we'll deepen our understanding of what each stage entails. For now, the goal is to show you how the Data Cycle is used. Don't worry if you don't understand all of the details in what follows. We'll go into detail in subsequent chapters.

The cycle revolves around a research topic. Research topics can be very broad and serious; for example, "What makes a runner faster or slower?" or "What effect do taxes have on the economy?" Research topics might be geared toward answering some pointed, important questions: "Does human activity contribute to global warming?" or "Do cell phones cause cancer?" The first step is to break this big topic into smaller steps. These smaller steps should be phrased as questions.

Ask Questions. The trick here is to ask *good* questions, and you'll get better at that as you read on. Good questions are questions that can be answered with data. Better questions are questions that address the research topic and increase your understanding of the issues. Often in this book, we will ask you to consider a data set and to pose questions that you want answered.

For example, Table 1.5 shows a few randomly selected rows from a data set containing all 19,212 runners in the 2017 Los Angeles Marathon. Most of the variables are

Bib.Number	Age	Place	Gender.Place	5k.Split	15k.Split	Clock.Time	Net.Time	Hometown	Gender
8752	36	3,718	2,874	1,607	4,825	15,695	15,539	Pasadena, CA	M
11478	31	14,785	5,585	1,881	6,814	23,487	22,816	Victorville, CA	F
3372	47	2,246	1,839	1,530	4,763	14,368	14,330	Danville, CA	M

▲ **TABLE 1.5** Three randomly selected runners from the 2017 Los Angeles Marathon.

concerned with time given in seconds, for example, the time from when the race begins to when the runner crosses the finish line (*Net.Time*). The variable *15k.Split* gives the time it took this runner to run the first 15 kilometers. You might have some ideas about what qualities lead to faster and slower runners. This will be our research topic: Can we better understand what qualities are associated with running a fast Los Angeles Marathon?

Consider the variables provided and write down two questions you would like to know about this marathon; focus your attention on questions that could be answered with these variables (assuming you can see the full data set).

Perhaps your questions might have concerned the variables themselves. What does *Bib.Number* mean? What's the difference between *Clock.Time* and *Net.Time*? A natural question to ask is, "Who won?" but, since we don't have the runners' names, this can't be answered. So instead you might ask, "What was the fastest time?" (8938 minutes, or about 2 hours 29 minutes). You might also have wondered how different the speeds were for men and women. Or perhaps you wondered if older people ran this race slower than younger people?

These are all examples of important questions, although only the last two address the research topic. In Chapter 2 you'll learn about Statistical Questions, which are a particular type of question that can be very productive in a statistical investigation.

Consider Data. In this stage, we consider which data are available to answer the question. In fact, many statistical investigations begin at this stage: you are given data and need to generate useful questions to help you understand what the data are about. This was the case with the L.A. Marathon data.

When considering if your data are helpful for answering the questions you've posed, it's very important to understand the context of the data. Who are what was observed? What variables were measured, and how were they measured? What were the units of measurement? Who collected the data? How did they collect the data? Why did they collect the data? When did they collect the data? Where did they collect the data.

Sometimes these questions cannot be answered—we simply don't have the information at hand. If so, this could be a reason to distrust the data and proceed with extreme caution. If you collected the data yourself, be sure to record this information so that future analysts can understand and make use of your efforts.

In this case, the data came from the website <http://www.lamarathon.com/race-weekend/results>. (The data were modified somewhat for pedagogical purposes.) The data were collected on every runner in the race, as is typical for these large events. We aren't provided with information about how the data were collected, but it seems reasonable to assume that the data came from entry forms and from the race officials themselves.

- Who or what was observed? All participants in the 2017 Los Angeles Marathon.
- What variables were measured, and how? We may have to do a little research into the context in order to understand what the variables here mean. For example, with a little bit of googling, we can learn that in a large running event such as a marathon, it often takes a runner awhile after the start of the race before she or he gets to the actual start line. For this reason, the "Clock Time," the time from when the race began and the runner crossed the finish line, is often longer than the "Net Time" the time it took the runner to go from the start line to the finish line. "Bib Number" is the runner's assigned ID number. The units for the times are in seconds.
- Who collected the data? The data are official results, and we can assume they were assembled and collected by the race officials.
- When and where did they collect the data? The running times were collected on the day of the race, March 19, 2017. These times were collected at the start and finish line of the race.



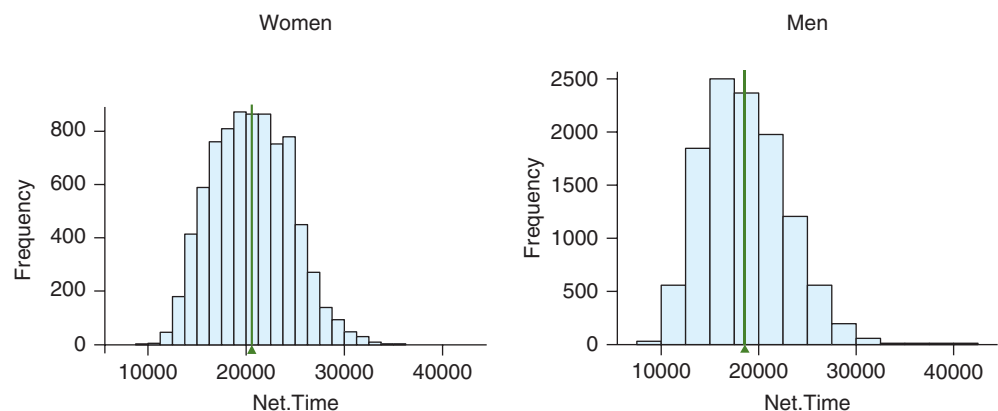
Data Moves ▶ These data were scraped from a web page that provides results

of marathons from across the country. A "script" was written in the statistical programming language R to convert the many pages of HTML files on this website into a file ready to be analyzed.

File ▶ [lamarathon.csv](#)

Analyze Data. This is the primary topic of this book. You’ll learn in Chapter 2 that the first step in an analysis is to visualize the data and that, sometimes, this visualization is enough to answer the question of interest. Let’s consider the question, “How different were the speeds for men and women?” This is a tricky question, because the running times varied considerably for men and women. Some women were faster than some men; some men were faster than some women. To help us answer, we refine this question to “What was the typical difference in speed between men and women in this race?” You’ll see in Chapter 2 that a visualization such as the one in Figure 1.10 is a helpful first step toward answering questions like ours. Figure 1.10 displays the *distributions* of the *Net.Time* variable for both men and women. (You’ll learn about how to read such visualizations in Chapter 2.) We’ve added a vertical bar to indicate the location of the *mean* time for women (top) and men (bottom). You’ll learn about how, when, and why to use the mean in Chapter 3. From the graph, it looks like the mean running time for women is a bit over 20,000 seconds (a bit more than 5.5 hours) and for men is less than 20,000 seconds.

► **FIGURE 1.10** Visualization of running times for women (left) and men (right). The vertical line indicates the mean running time for each group.



Interpret Data. The final step is to interpret your analysis. *Interpret* is a fancy word for “answer your question!” Our question was, “What was the typical difference in speed between men and women in this race?” In Chapter 3 you’ll see that the mean value gives us one way of measuring this notion of “typical,” and from Figure 1.10 we can roughly judge that the means differ by about 2000 seconds. (This is actually a tricky judgment call from this graphic, but you’ll soon learn some tools that help you judge these distances.) Our answer to the question is: Typically, the men were 2000 seconds faster than the women in this race.

Examples 3 and 4 will give you practice with the “Ask Questions” part of the Data Cycle.

EXAMPLE 3 At the Movies



Data Moves ► James Molyneux merged data from several sources to compile this data

set. Merging related datasets can often lead to greater insights.

File ► movieratings.csv

The data file *movieratings.csv*, compiled by statistician James Molyneux, contains data on almost 5000 movies. The following variables are provided:

Title, year, runtime, mpaa_rating, studio, color, director, language, country, aspect_ratio, n_post_face, n_critics, n_audience, reviews_num, audience_rating, critics_rating, budget, gross, imdbi_id

The Data Cycle begins with questioning. You might not know precisely what these variables mean or how the data were measured, but use your general knowledge about movies.

QUESTION Which variables would you consider in order to answer this question: Do the critics tend to rate movies lower than the audience rates them?

SOLUTION We see that there are two variables, *audience_rating* and *critics_rating*, that would likely help us determine if critics tend to rate movies differently than “regular” people.

TRY THIS! Exercise 1.15



EXAMPLE 4 Movie Ratings

The variables provided in the datafile *movieratings.csv* include the following:

Title, year, runtime, mpaa_rating, studio, color, director, language, country, aspect_ratio, n_post_face, n_critics, n_audience, reviews_num, audience_rating, critics_rating, budget, gross, imdb_id

Often we are provided with a “data dictionary” that gives us some guidelines as to what the variables mean. Here’s a simplified data dictionary: *runtime* is the length of the movie; *color* is whether the movie is in color or black-and-white; *mpaa_rating* is the Motion Pictures Association of America guide to the age level of the movie—G, PG, PG-13, and so on; *critics_rating* and *audience_rating* are average scores of critics from the Rotten Tomatoes website.

QUESTION Which of these questions cannot be answered with the data in *movieratings.csv*?

- Do critics rate R-rated movies more highly than G-rated movies?
- Are comedies shorter than dramas?
- Do audiences prefer shorter movies over longer movies?
- Do movies that have a large budget receive higher audience ratings?

SOLUTION Question (b) cannot be answered with these data because it requires knowing the *genre* (comedy, drama, documentary, etc.) of the movie, and we do not have a variable that provides this information.

TRY THIS! Exercise 1.19

The Data Cycle is not meant to be a hard-and-fast rule that, if followed rigorously, is promised to bring you success and happiness. Think of it, instead, as a guiding principle, a device to help you if you get stuck. For example, if you’ve just done an analysis and are thinking, “Am I done?” take a look at the Data Cycle. After the analysis comes the interpretation, and so, unless you’ve done an interpretation, you’re not yet done!

SECTION 1.4

Organizing Categorical Data

Once we have a data set, we next need to organize and display the data in a way that helps us see patterns. This task of organization and display is not easy, and we discuss it throughout the entire text. In this section we introduce the topic for the first time, in the context of categorical variables.

With categorical variables, we are usually concerned with knowing how often a particular category occurs in our sample. We then (usually) want to compare how often a category occurs for one group with how often it occurs for another (liberal/conservative, man/woman). To do these comparisons, you need to understand how to calculate percentages and other rates.

	Male	Female
Not Always	2	3
Always	3	7

▲ **TABLE 1.6** This two-way table shows counts for 15 youths who responded to a survey about wearing seat belts.

Male	Not Always
1	1
1	1
1	0
1	0
1	0
0	1
0	1
0	1
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0

▲ **TABLE 1.7** This data set is equivalent to the two-way summary shown in Table 1.6. We highlighted in red those who did not always wear a seat belt (the risk takers).

A common method for summarizing two potentially related categorical variables is to use a two-way table. **Two-way tables** show how many times each combination of categories occurs. For example, Table 1.6 is a two-way table from the Youth Behavior Risk Survey that shows gender and whether or not the respondent always (or almost always) wears a seat belt when riding in or driving a car. The actual Youth Behavior Risk Survey has over 10000 respondents, but we are practicing on a small sample from this much larger data set.

The table tells us that two people were male and did not always wear a seat belt. Three people were female and did not always wear a seat belt. These counts are also called frequencies. A **frequency** is simply the number of times a value is observed in a data set.

Some books and publications discuss two-way tables as if they displayed the original data collected by the investigators. However, two-way tables do not consist of “raw” data but, rather, are summaries of data sets. For example, the data set that produced Table 1.6 is shown in Table 1.7.

To summarize this table, we simply count how many of the males (a 1 in the Male column) also do not always wear seat belts (a 1 in the Not Always column). We then count how many both are male and always wear seat belts (a 1 in the Male column, a 0 in the Not Always column), how many both are female and don’t always wear seat belts (a 0 in the Male column, a 1 in the Not Always column), and finally, how many both are female and always wear a seat belt (a 0 in the Male column, a 0 in the Not Always column).

Example 5 illustrates that summarizing the data in a two-way table can make it easy to compare groups.

EXAMPLE 5 Percentages of Seat Belt Wearers

The 2011 Youth Behavior Risk Survey is a national study that asks American youths about potentially risky behaviors. We show the two-way summary again. All of the people in the table were between 14 and 17 years old. The participants were asked whether they wear a seat belt while driving or riding in a car. The people who said “always” or “almost always” were put in the Always group. The people who said “sometimes” or “rarely” were put in the Not Always group.

	Male	Female
Not Always	2	3
Always	3	7

! Caution

Two-Way Tables Summarize Categorical Variables

It is tempting to look at a two-way table like Table 1.6 and think that you are looking at numerical variables, because you see numbers. But the values of the variables are actually categories (gender and whether or not the subject always wears a seat belt). The numbers you see are summaries of the data.

QUESTIONS

- How many men are in this sample? How many women? How many people do not always wear seat belts? How many always wear seat belts?
- What percentage of the sample are men? What percentage are women? What percentage don’t always wear seat belts? What percentage always wear seat belts?
- Are the men in the sample more likely than the women in the sample to take the risk of not wearing a seat belt?

SOLUTIONS

- We can count the men by adding the first column: $2 + 3 = 5$ men. Adding the second column gives us the number of women: $3 + 7 = 10$.

We get the number who do not always wear seat belts by adding the first row: $2 + 3 = 5$ people don’t always wear seat belts. Adding the second row gives us the number who always wear seat belts: $3 + 7 = 10$.

- b. This question asks us to convert the numbers we found in part (a) to percentages. To do this, we divide the numbers by 15, because there were 15 people in the sample. To convert to percentages, we multiply this proportion by 100%.

The proportion of men is $5/15 = 0.333$. The percentage is $0.333 \times 100\% = 33.3\%$. The proportion of women must be $100\% - 33.3\% = 66.7\%$ ($10/15 \times 100\% = 66.7\%$).

The proportion who do not always wear seat belts is $5/15 = 0.333$, or 33.3%. The proportion who always wear seat belts is $100\% - 33.3\% = 66.7\%$.

- c. You might be tempted to answer this question by counting the number of males who don't always wear seat belts (2 people) and comparing that to the number of females who don't always wear seat belts (3 people). However, this is not a fair comparison because there are more females than males in the sample. Instead, we should look at the percentage of those who don't always wear seat belts in each group. This question should be reworded as follows:

Is the percentage of males who don't always wear seat belts greater than the percentage of females who don't always wear seat belts?

Because 2 out of 5 males don't always wear seat belts, the percentage of males who don't always wear seat belts is $(2/5) \times 100\% = 40\%$.

Because 3 out of 10 females don't always wear seat belts, the percentage of females who don't always wear seat belts is $(3/10) \times 100\% = 30\%$.

In fact, females in this sample engage in this risky behavior less often than males. Among all U.S. youth, it is estimated that about 28% of males do not always wear their seat belt, compared to 23% of females.



TRY THIS! Exercise 1.21

The calculations in Example 5 took us from frequencies to percentages. Sometimes, we want to go in the other direction. If you know the total number of people in a group and are given the percentage that meets some qualification, you can figure out *how many* people in the group meet that qualification.

EXAMPLE 6 Numbers of Seat Belt Wearers

A statistics class has 300 students, and they are asked whether they always ride or drive with a seat belt.

QUESTIONS

- Suppose that 30% of the students do not always wear a seat belt. How many students is this?
- Suppose we know that in another class, 20% of the students do not always wear seat belts, and this is equal to 43 students. How many students are in the class?

SOLUTIONS

- We need to find 30% of 300. When working with percentages, first convert the percentage to its decimal equivalent:

$$30\% \text{ of } 300 = 0.30 \times 300 = 90$$

Therefore, 90 students don't always wear seat belts.

- b. The question tells us that 20% of some unknown larger number (call it y) must be equal to 43.

$$0.20y = 43$$

Divide both sides by 0.20 and you get

$$y = 215$$

There are 215 total students in the class, and 43 of them don't always wear seat belts.

TRY THIS! Exercise 1.29

Sometimes, you may come across data summaries that are missing crucial information. Pay close attention when tables give counts of things. Suppose we wanted to know which city was most at risk for a crime such as burglary? Table 1.8 gives the number of burglaries in 2009 reported in several major as reported by the Federal Bureau of Investigation. We show only the ten cities with the greatest number of burglaries.

► **TABLE 1.8** The ten U.S. cities with the greatest numbers of burglaries reported.

State	City	Number of Burglaries
Texas	Houston	19,858
California	Los Angeles	16,160
Nevada	Las Vegas	14,876
New York	New York City	14,100
Illinois	Chicago	13,152
Arizona	Phoenix	12,798
Texas	San Antonio	11,633
Texas	Dallas	11,121
Indiana	Indianapolis	11,085
Tennessee	Memphis	10,272

It looks like Houston, Texas, is the most dangerous city in terms of burglaries and Memphis, Tennessee, one of the least dangerous (among these top ten). But this table is missing a crucial piece of information: the number of people who live in the city. A city with 1 million people is probably going to have more burglaries than a town of 40,000.

How do we control for the difference in populations? Table 1.9 includes the information missing from Table 1.8: the population counts.

State	City	Population	Number of Burglaries
Texas	Houston	2,275,221	19,858
California	Los Angeles	3,962,726	16,160
Nevada	Las Vegas	1,562,134	14,876
New York	New York City	8,550,861	14,100
Illinois	Chicago	2,728,695	13,152
Arizona	Phoenix	1,559,744	12,798
Texas	San Antonio	1,463,586	11,633
Texas	Dallas	1,301,977	11,121
Indiana	Indianapolis	863,675	11,085
Tennessee	Memphis	657,936	10,272

▲ **TABLE 1.9** The same cities with population sizes included.

With this extra information we can figure out which city is most risky based on its size. For example, what percentage of residents reported burglaries in Houston? There were 2,275,221 residents and 19,858 burglaries. And so the percentage burgled is $(19858/2275221) \times 100\% = 0.87\%$, after rounding.

Sometimes, with percentages as small as this, we understand the numbers more easily if instead of reporting the “per cent” we report “per 1000” or even “per 10,000. We call such numbers **rates**. To get the burglary rate per 1000 residents, instead of multiplying $(19858/2275221)$ by 100 we multiply by 1000: $(19858/2275221) \times 1000 = 8.7$ burglaries per 1000 people. These results are shown in Table 1.10 (rounded to the hundredths place for easier reading).

State	City	Population	Number of Burglaries	Burglaries per 1000 residents
Texas	Houston	2,275,221	19,858	8.73
California	Los Angeles	3,962,726	16,160	4.08
Nevada	Las Vegas	1,562,134	14,876	9.52
New York	New York City	8,550,861	14,100	1.65
Illinois	Chicago	2,728,695	13,152	4.82
Arizona	Phoenix	1,559,744	12,798	8.21
Texas	San Antonio	1,463,586	11,633	7.95
Texas	Dallas	1,301,977	11,121	8.54
Indiana	Indianapolis	863,675	11,085	12.83
Tennessee	Memphis	657,936	10,272	15.61

▲ **TABLE 1.10** The ten cities with the most burglaries with burglaries per 1000 residents.

We now see that Memphis has the highest burglary rate among these cities and New York City the lowest!

EXAMPLE 7 Comparing Rates of Stolen Cars

Which model of car has the greatest risk of being stolen? The Highway Loss Data Institute reports that the Ford F-250 pickup truck is the most stolen car; 7 F-250s are reported stolen out of every 1000 that are insured. By way of contrast, the Jeep Compass is the least stolen; only 0.5 Jeep Compass is reported stolen for every 1000 insured (Insurance Institute for Highway Safety 2013).

QUESTION Why does the Highway Loss Data Institute report theft rates rather than the number of each type of car stolen?

SOLUTION We need to take into account the fact that some cars are more popular than others. Suppose there were many more Jeep Compasses than Ford F-250s. In that case, we might see a greater number of stolen Jeeps, simply because there are more of them to steal. By looking at the *theft rate*, we adjust for the total number of cars of that particular kind on the road.



TRY THIS! Exercise 1.31

KEY POINT

In order for us to compare groups, the groups need to be similar. When the data consist of counts, then percentages or rates are often better for comparisons because they take into account possible differences among the sizes of the groups.

SECTION 1.5

Collecting Data to Understand Causality

Often, the most important questions in science, business, and everyday life are questions about **causality**. These are usually phrased in the form of “what if” questions. What if I take this medicine; will I get better? What if I change my Facebook profile; will my profile get more hits?

Questions about causality are frequently in the news. The *Los Angeles Times* reported that many people believe a drink called peanut milk can cure gum disease and slow the onslaught of baldness. The BBC News (2010) reported that “[h]appiness wards off heart disease.” Statements such as these are everywhere we turn these days. How do we know whether to believe these claims?

The methods we use to collect data determine what types of conclusions we can make. Only one method of data collection is suitable for making conclusions about causal relationships, but as you’ll see, that doesn’t stop people from making such conclusions anyway. In this section we talk about three methods commonly used to collect data in an effort to answer questions about causality: anecdotes, observational studies, and controlled experiments.

Most questions about causality can be understood in terms of two variables: the **treatment variable** and the **outcome variable**. (The outcome variable is also sometimes called the **response variable** because it responds to changes in the treatment.) We are essentially asking whether the treatment variable causes changes in the outcome variable. For example, the treatment variable might record whether or not a person drinks peanut milk, and the outcome variable might record whether or not that person’s gum disease improved. Or the treatment variable might record whether or not a person is generally happy, and the outcome variable might record whether or not that person suffered from heart disease in a ten-year period.

People who receive the treatment of interest (or have the characteristic of interest) are said to be in the **treatment group**. Those who do not receive that treatment (or do not have that characteristic) are in the **comparison group**, which is also called the **control group**.

Anecdotes

Peanut milk is a drink invented by Jack Chang, an entrepreneur in San Francisco, California. He noticed that after he drank peanut milk for a few months, he stopped losing hair and his gum disease went away. According to the *Los Angeles Times* (Glionna 2006), another regular drinker of peanut milk says that the beverage caused his cancer to go into remission. Others have reported that drinking the beverage has reduced the severity of their colds, has helped them sleep, and has helped them wake up.

This is exciting stuff! Peanut milk could very well be something we should all be drinking. But can peanut milk really solve such a wide variety of problems? On the face of it, it seems that there’s evidence that peanut milk has cured people of illness. The *Los Angeles Times* reports the names of people who claim that it has. However, the truth is that this is simply not enough evidence to justify any conclusion about whether the beverage is helpful, harmful, or without any effect at all.

These testimonials are examples of anecdotes. An **anecdote** is essentially a story that someone tells about her or his own (or a friend’s or relative’s) experience. Anecdotes are an important type of evidence in criminal justice because eyewitness testimony can carry a great deal of weight in a criminal investigation. However, for answering questions about groups of people with great variability or diversity, anecdotes are essentially worthless.

The primary reason why anecdotes are not useful for reaching conclusions about cause-and-effect relationships is that the most interesting things that we study have so

much variety that a single report can't capture the variety of experience. For example, have you ever bought something because a friend recommended it, only to find that after a few weeks it fell apart? If the object was expensive, such as a car, you might have been angry at your friend for recommending such a bad product. But how do you know whose experience was more typical, yours or your friend's? Perhaps the car is in fact a very reliable model, and you just got a lemon.

A very important question to ask when someone claims that a product brings about some kind of change is to ask, "Compared to what?" Here the claim is that drinking peanut milk will make you healthier. The question to ask is, "Healthier compared to what?" Compared to people who don't drink peanut milk? Compared to people who take medicine for their particular ailment? To answer these questions, we need to examine the health of these other groups of people who do not drink peanut milk.

Anecdotes do not give us a comparison group. We might know that a group of people believe that peanut milk made them feel better, but we don't know how the milk drinkers' experiences compare to those of people who did not drink peanut milk.

KEY POINT

When someone makes a claim about causality, ask, "Compared to what?"

Another reason for not trusting anecdotal evidence is a psychological phenomenon called the placebo effect. People often react to the idea of a treatment, rather than to the treatment itself. A **placebo** is a harmless pill (or sham procedure) that a patient believes is actually an effective treatment. Often, the patient taking the pill feels better, even though the pill actually has no effect whatsoever. In fact, a survey of U.S. physicians published in the *British Medical Journal* (Britt 2008) found that up to half of physicians prescribe sugar pills—placebos—to manage chronic pain. This psychological wish fulfillment—we feel better because we think we *should* be feeling better—is called the **placebo effect**.

Observational Studies

The identifying mark of an **observational study** is that the subjects in the study are put into the treatment group or the control group either by their own actions or by the decision of someone else who is not involved in the research study. For example, if we wished to study the effects on health of smoking cigarettes (as many researchers have), then our treatment group would consist of people who had chosen to smoke, and the control group would consist of those who had chosen not to smoke.

Observational studies compare the outcome variable in the treatment group with the outcome variable in the control group. Thus, if many more people are cured of gum disease in the group that drinks peanut milk (treatment) than in the group that does not (control), then we would say that drinking peanut milk is associated with improvement in gum disease; that is, there is an **association** between the two variables. If fewer people in the happy group tend to have heart disease than those in the not-happy group, we would say that happiness is associated with improved heart health.

Note that we do not conclude that peanut milk *caused* the improvement in gum disease. In order for us to draw this conclusion, the treatment group and the control group must be very similar in every way except that one group gets the treatment and the other doesn't. For example, if we knew that the group of people who started drinking peanut milk and the group that did not drink peanut milk were alike in every way—both groups had the same overall health; were roughly the same ages; included the same mix of genders and races, education levels; and so on—then if the peanut milk group members are healthier after a year, we would be fairly confident in concluding that peanut milk is the reason for their better health.

Unfortunately, in observational studies this goal of having very similar groups is *extremely* difficult to achieve. *Some* characteristic is nearly always different in one group

than in the other. This means that the groups may experience different outcomes because of this different characteristic, not because of the treatment. A difference between the two groups that could explain why the outcomes were different is called a **confounding variable** or **confounding factor**.

For example, early observational studies on the effects of smoking found that a greater percentage of smokers than of nonsmokers had lung cancer. However, some scientists argued that genetics was a confounding variable (Fisher 1959). They maintained that the smokers differed genetically from the nonsmokers. This genetic difference made some people more likely to smoke and more susceptible to lung cancer.

This was a convincing argument for many years. It not only proposed a specific difference between the groups (genetics) but also explained how that difference might come about (genetics makes some people smoke more, perhaps because it tastes better to them or because they have addictive personalities). And the argument also explained why this difference might affect the outcome (the same genetics cause lung cancer). Therefore, the skeptics said, genetics—and not smoking—might be the cause of lung cancer.

Later studies established that the skeptics were wrong about genetics. Some studies compared pairs of identical twins in which one twin smoked and the other did not. These pairs had the same genetic makeup, and still a higher percentage of the smoking twins had cancer than of the nonsmoking twins. Because the treatment and control groups had the same genetics, genetics could not explain why the groups had different cancer rates. When we compare groups in which we force one of the variables to be the same, we say that we are *controlling for* that variable. In these twin studies, the researchers controlled for genetics by comparing people with the same genetic makeup (Kaprio and Koskenvuo 1989).

A drawback of observational studies is that we can never know whether there exists a confounding variable. We can search very hard for it, but the mere fact that we don't find a confounding variable does not mean it isn't there. For this reason, we can never make cause-and-effect conclusions from observational studies.

KEY POINT

We can never draw cause-and-effect conclusions from observational studies because of potential confounding variables. A single observational study can show only that there is an *association* between the treatment variable and the outcome variable.

EXAMPLE 8 Does Poverty Lower IQ?

“Chronic Poverty Can Lower Your IQ, Study Shows” is a headline from the online magazine *Daily Finance*. The article (Nisen 2013) reported on a study published in the journal *Science* (Mani et al. 2013) that examined the effects of poverty on problem-solving skills from several different angles. In one part of the study, researchers observed sugar cane farmers in rural India both before and after harvest. Before the harvest, the farmers typically have very little money and are often quite poor. Researchers gave the farmers IQ exams before the harvest and then after the harvest, when they had more money, and found that the farmers scored much higher after the harvest.

QUESTION Based on this evidence alone, can we conclude that poverty lowers people's IQ scores? If yes, explain why. If no, suggest a possible confounding factor.

SOLUTION No, we cannot. This is an observational study. The participants are in or out of the treatment group (“poverty”) because of a situation beyond the researchers' control. A possible confounding variable is nutrition; before harvest, without much money, the farmers are perhaps not eating well, and this could lower their IQ scores.

(In fact, the researchers considered this confounding variable. They determined that nutrition was relatively constant both before and after the harvest, so it was ruled out as a confounding variable. But other confounding variables may still exist.)

TRY THIS! Exercise 1.47



Controlled Experiments

In order to answer cause-and-effect questions, we need to create a treatment group and a control group that are alike in every way possible, except that one group gets a treatment and the other does not. As you've seen, this cannot be done with observational studies because of confounding variables. In a **controlled experiment**, researchers take control by assigning subjects to the control or treatment group. If this assignment is done correctly, it ensures that the two groups can be nearly alike in every relevant way except whether or not they receive the treatment under investigation.

Well-designed and well-executed controlled experiments are the only means we have for definitively answering questions about causality. However, controlled experiments are difficult to carry out (this is one reason why observational studies are often done instead). Let's look at some of the attributes of a well-designed controlled experiment.

A well-designed controlled experiment has four key features:

- The sample size must be large so that we have opportunities to observe the full range of variability in the humans (or animals or objects) we are studying.
- The subjects of the study must be assigned to the treatment and control groups at random.
- Ideally, the study should be “double-blind,” as explained later.
- The study should use a placebo if possible.

These features are all essential in order to ensure that the treatment group and the control group are as similar as possible.

To understand these key design features, imagine that a friend has taken up a new exercise routine in order to lose weight. By exercising more, he hopes to burn more calories and so lose weight. But he notices a strange thing: the more he exercises, the hungrier he gets, and so the more he eats. If he is eating more, can he lose weight? This is a complex issue, because different people respond differently both to exercise and to food. How can we know whether exercise actually leads to weight loss? (See Rosenkilde et al. (2012) for a study related to the hypothetical one presented here.) To think about how you might answer this question, suppose you select a group of slightly overweight young men to participate in your study. For a comparison group, you might ask some of them not to exercise at all during the study. The men in the treatment group, however, you will ask to exercise for about 30 minutes each day at a moderate level.

Sample Size

A good controlled experiment designed to determine whether exercise leads to weight loss should have a large number of people participate in the study. People react to changes in their activity level in a variety of ways, so the effects of exercise can vary from person to person. To observe the full range of variability, you therefore need a large number of people. How many people? This is a hard question to answer, but in general, the more the better. You should be critical of studies with very few participants.

Random Assignment

The next step is to assign people to the treatment group and the comparison group such that the two groups are similar in every way possible. As we saw when we discussed observational studies, letting the participants choose their own group doesn't work, because people who like to exercise might differ in important ways (such as level of motivation) that would affect the outcome.

Instead, a good controlled experiment uses **random assignment** to assign participants to groups. One way of doing this is to flip a coin. Heads means the participant goes into the treatment group, and tails means she or he goes into the comparison group (or the other way around—as long as you're consistent). In practice, the randomizing might instead be done with a computer or even with the random-number generator on a calculator, but the idea is always the same: No human determines group assignment. Rather, assignment is left to chance.

If both groups have enough members, random assignment will “balance” the groups. The variation in weights, the mix of metabolisms and daily calorie intake, and the mix of most variables will be similar in both groups. Note that by “similar” we don't mean exactly the same. We don't expect both groups to have exactly the same percentage of people who like to exercise, for example. Except in rare cases, random variation results in slight differences in the mixes of the groups. But these differences should be small.

Whenever you read about a controlled experiment that does not use random assignment, remember that there is a very real possibility that the results of the study are invalid. The technical term for what happens with nonrandomized assignment is **bias**. We say that a study exhibits bias when the results are influenced in one particular direction. A researcher who puts the heaviest people in the exercise group, for example, is biasing the outcome. It's not always easy, or even possible, to predict what the effects of the bias will be, but the important point is that the bias creates a confounding variable and makes it difficult, or impossible, to determine whether the treatment we are investigating really affects the outcome we're observing.

KEY POINT

Random assignment (assignment to treatment groups by a randomization procedure) helps balance the groups to minimize bias. This helps make the groups comparable.

Blinding

So far, we've recruited a large number of men and randomly assigned half to exercise and half to remain sedentary. In principle, these two groups will be very similar. However, there are still two potential differences.

First, we might know who is in which group. This means that when we interact with a participant, we might consciously or unconsciously treat that person differently, depending on which group he or she belongs to. For example, if we believe strongly that exercise helps with weight loss, we might give special encouragement or nutrition advice to people who are in the exercise group that we don't give to those in the comparison group. If we do so, then we have biased the study.

To prevent this from happening, researchers should be **blind** to assignment. This means that an independent party—someone who does not regularly see the participants and who does not participate in determining the results of the study—handles the assignment to groups. The researchers who measure the participants' weight loss do not know who is in which group until the study has ended; this ensures that their measurements will not be influenced by their prejudices about the treatment.

Second, we must consider the participants themselves. If they know they are in the treatment group, they may behave differently than they would if they knew nothing about their group assignment. Perhaps they will work harder at losing weight.

Or perhaps they will eat much more because they believe that the extra exercise allows them to eat anything they want.

To prevent this from happening, the participants should also not know whether they are in the treatment group or the comparison group. In some cases, this can be accomplished by not even telling the participants the intent of the study; for example, the participants might not know whether the goal is to examine the effect of a sedentary lifestyle on weight or whether it is about the effects of exercise. (However, ethical considerations often forbid the researchers from engaging in deception.)

When neither the researchers nor the participants know whether the participants are in the treatment or the comparison group, we say that the study is **double-blind**. The double-blind format helps prevent the bias that can result if one group acts differently from the other because they know they are being treated differently or because the researchers treat the groups differently or evaluate them differently because of what the researchers hope or expect.

Placebos

The treatment and comparison groups might differ in still another way. People often react not just to a particular medical treatment, but also to the very idea that they are getting medical treatment. This means that patients who receive a pill, a vaccine, or some other form of treatment often feel better even when the treatment actually does absolutely nothing. Interestingly, this placebo effect also works in the other direction: if they are told that a certain pill might cause side effects (for example, a rash), some patients experience the side effects even though the pill they were given is just a sugar pill.

To neutralize the placebo effect, it is important that the comparison group receive attention similar to what the treatment group receives, so that both groups feel they are being treated the same by the researchers. In our exercise study, the groups behave very differently. However, the sedentary group might receive weekly counseling about lifestyle change or might be weighed and measured just as frequently as the treatment group. If, for instance, we were studying whether peanut milk improves baldness, we would require the comparison group to take a placebo drink so that we could rule out any placebo effect and thus perform a valid comparison between the treatment and control groups.

KEY POINT

The following qualities are the “gold standard” for experiments:

Large sample size. This ensures that the study captures the full range of variation among the population and allows small differences to be noticed.

Controlled and randomized. Random assignment of subjects to treatment or comparison groups helps to minimize bias.

Double-blind. Neither subjects nor researchers know who is in which group.

Placebo (if appropriate). This format controls for possible differences between groups that occur simply because some subjects are more likely than others to expect their treatment to be effective.

Details

The Real Deal

A study similar to the one described here was carried out in Denmark in 2012. The researchers found that young, overweight men who exercised 30 minutes per day lost more body fat than those who exercised 60 minutes per day. Both groups of exercisers lost more body fat than a group that did not exercise at all. One conclusion is that more intense exercise for overweight people leads to an increase in appetite, and so moderate levels of exercise are best for weight loss. (Rosenkilde, et al. 2012).

EXAMPLE 9 Brain Games

Brain-training video games, such as Nintendo’s Brain Age, claim to improve basic intelligence skills, such as memory. A study published in the journal *Nature* investigated whether playing such games can actually boost intelligence (Owen et al. 2010). The researchers explain that 11,430 people logged onto a web page and were randomly assigned to one of three groups. Group 1 completed six training tasks that emphasized “reasoning, planning and problem-solving.” Group 2 completed games that emphasized a broader range of cognitive skills. Group 3 was a control group and didn’t play any of

these games; instead, members were prompted to answer “obscure” questions. At the end of six weeks, the participants were compared on several different measures of thinking skills. The results? The control group did just as well as the treatment groups.

QUESTION Which features of a well-designed controlled experiment does this study have? Which features are missing?

SOLUTION **Sample size:** The sample size of 11,430 is quite large. Each of the three groups will have about 3800 people.

Randomization: The authors state that patients were randomly assigned to one of the three groups.

Double-blind format: Judging on the basis of this description, there was no double-blind format. It’s possible (indeed, it is likely) that the researchers did not know, while analyzing the outcome, to which treatment group individuals had been assigned. But we do not know whether *participants* were aware of the existence of the three different groups and how they differed.

Placebo: The control group participated in a “null” game, in which they simply answered questions. This activity is a type of placebo because the participants could have thought that this null game was a brain game.

TRY THIS! Exercise 1.49



! Caution

At Random

The concept of randomness is used in two different ways in this section. *Random assignment* is used in a controlled experiment. Subjects are randomly assigned to treatment and control groups in order to achieve a balance between groups. This ensures that the groups are comparable to each other and that the only difference between the groups is whether or not they receive the treatment under investigation. **Random selection** occurs when researchers select subjects from some larger group via a random method. We must employ random selection if we wish to extend our results to a larger group.

Extending the Results

In both observational studies and controlled experiments, researchers are often interested in knowing whether their findings, which are based on a single collection of people or objects, will extend to the world at large.

The researchers in Example 9 concluded that brain games are not effective, but might it just be that the games weren’t effective for those people who decided to participate? Maybe if the researchers tested people in another country, for example, the findings would be different.

It is usually not possible to make generalizations to a larger group of people unless the subjects for the study are representative of the larger group. The only way to collect a sample that is representative is to collect the objects we study at random. We will discuss how to collect a random sample, and why we can then make generalizations about people or objects who were not in the sample, in Chapter 7.

Selecting subjects using a random method is quite common in polls and surveys (which you’ll also study in Chapter 7), but it is much less common in other types of studies. Most medical studies, for example, are not conducted on people selected randomly, so even when a cause-and-effect relationship emerges between the treatment and the response, it is impossible to say whether this relationship will hold for a larger (or different) group of people. For this reason, medical researchers often work hard to replicate their findings in diverse groups of people.

Statistics in the News

When reading in a newspaper or blog about a research study that relies on statistical analysis, you should ask yourself several questions to evaluate how much faith you can put in the conclusions reached in the study:

1. *Is this an observational study or a controlled experiment?*

If it’s an observational study, then you can’t conclude that the treatment caused the observed outcome.

2. *If the study is a controlled experiment, was there a large sample size? Was randomization used to assign participants to treatment groups? Was the study double-blind? Was there a placebo?*

See the relevant section of this chapter for a review of the importance of these attributes.

3. *Was the paper published in a peer-reviewed journal? What is the journal's reputation?*

“Peer-reviewed” means that each paper published in the journal is rigorously evaluated by at least two anonymous researchers familiar with the field. The best journals are very careful about the quality of the research they report on. They have many checkpoints to make sure that the science is as good as it can be. (But remember, this doesn't mean the science is perfect. If you read a medical journal regularly, you'll see much debate from issue to issue about certain results.) Other journals, by contrast, sometimes allow sloppy research results, and you should be very wary of these journals.

4. *Did the study follow people for a long enough time?*

Some treatments take a long time to work, and some illnesses take a long time to show themselves. For example, many cost-conscious people like to refill water bottles again and again with tap water. Some fear that drinking from the same plastic bottle again and again might lead to cancer. If this is true, it might take a very long time for a person to get cancer from drinking out of the same bottle day after day. So researchers who wish to determine whether drinking water from the same bottle causes cancer should watch people for a very long time.

Often it is hard to get answers to all these questions from a newspaper article. Fortunately, the Internet has made it much easier to find the original papers, and your college library or local public library will probably have access to many of the most popular journals.

Even when a controlled experiment is well designed, things can still go wrong. One common way in which medical studies go astray is that people don't always do what their doctor tells them to do. Thus, people randomized to the treatment group might not actually take their treatments. Or people randomized to the Atkins diet might switch to Weight Watchers because they don't like the food on the Atkins diet. A good research paper will report on these difficulties and will be honest about the effect on the researchers' conclusions.

EXAMPLE 10 Does Skipping Breakfast Make You Gain Weight?

The New York Times reported on a seven-year study about the eating habits of a sample of 50,000 adults in the United States (Rabin 2017). The participants were all members of the Seventh Day Adventist religion. According to the report, “breakfast eaters” were more likely to keep their weight down after seven years than were “breakfast skippers.”

QUESTION Is this most likely an observational study or a controlled experiment? Why? Does this mean that if you skip breakfast you will gain weight?

SOLUTION This is most likely an observational study. The treatment variable is whether or not someone eats breakfast. Although you could possibly assign people to eat or not eat breakfast for a short time, it is very unlikely this could be sustained for seven years and with such a large number of participants. Therefore, most likely researchers simply observed the habits that the participants voluntarily adopted. (This is, in fact, how the study was conducted.)

Because this is an observational study, we cannot conclude that the *treatment variable* (skipping breakfast) affects the *outcome variable* (weight gain). Possibly, people who regularly skip breakfast have other lifestyle characteristics that might cause them to gain weight. A potential confounding variable is that people who skip breakfast regularly might be under considerable time pressure, and this stress results in overeating at other times of day.



TRY THIS! Exercise 1.53

EXAMPLE 11 Crohn's Disease

Crohn's disease is a bowel disease that causes cramping, abdominal pain, fever, and fatigue. A study reported in the *New England Journal of Medicine* (Columbel et al. 2010) tested two medicines for the disease: injections of infliximab (Inflix) and oral azathioprine (Azath). The participants were randomized into three groups. All groups received an injection and a pill (some were placebos but still a pill and an injection). One group received Inflix injections alone (with placebo pills), one received Azath pills alone (with placebo injections), and one group received both injections and pills. A good outcome was defined as the disease being in remission after 26 weeks. The accompanying table shows the summary of the data that this study yielded.

	Combination	Inflix Alone	Azath Alone
Remission	96	75	51
Not in Remission	73	94	119

QUESTIONS

- Compare the percentages in remission for the three treatments. Which treatment was the most effective and which was the least effective for this sample?
- Can we conclude that the combination treatment causes a better outcome? Why or why not?

SOLUTIONS

- For the combination: $96/169$, or 56.8%, success
 For the Inflix alone: $75/169$, or 44.4%, success
 For the Azath alone: $51/170$, or 30%, success

The combination treatment was the most effective for this sample, and Azath alone was the least effective.

- Yes, we can conclude that the combination of drugs causes a better outcome than the single drugs. The study was placebo-controlled and randomized. The sample size was reasonably large. Blinding was not mentioned, but at least, thanks to the placebos, the patients did not know what treatment they were getting.



TRY THIS! Exercise 1.55

**CASE STUDY
REVISITED****Dangerous Habit?**

Does drinking coffee cause cancer? This question turns out to be difficult to answer. It is true that coffee contains acrylamide, but it is difficult to know if it contains it in large enough quantities to cause harm in humans. Acrylamide is produced by some plant-based foods when they are heated up or cooked. (Not just coffee, but also french fries and potato chips.) Controlled studies carried out on rats and mice *did* show that ingesting acrylamide through water increased their cancer risk. But controlled studies are not possible or ethical with humans, and according to the American Cancer Society, attempts to find significant differences in cancer rates between those who eat acrylamide-rich diets and those who do not have either failed to find an increased risk or have found mixed results. In 2016, the International Agency for Research on Cancer, an agency run by the World Health Organization (WHO), downgraded the risks of coffee from “probably carcinogenic for humans” to “not classifiable as to its carcinogenicity to humans.” In slightly plainer language, this means that the current evidence does not allow one to conclude that coffee causes cancer in humans.

This does not mean that scientists have proved that drinking coffee is safe. As time goes on, we will no doubt learn more about the risks, or lack of risks, of drinking coffee as potential confounding variables are ruled out.

Sources:

<https://www.cancer.org/cancer/cancer-causes/acrylamide.html>

<https://www.iarc.fr/en/media-centre/iarcnews/2016/DebunkMyth.php>

DATA PROJECT ▶ How Are Data Stored?



1

OVERVIEW

To begin your data exploration, you'll learn about how data are stored and how some files can be downloaded onto your computer and then uploaded into statistical software.

2

GOAL

Download an interesting data file and understand how it is structured.

3

YOUR CITY, YOUR DATA

Many governments throughout the world are participating in “Open Data” initiatives. Open Data are data collected, usually at taxpayer expense, to help inform policy and decision making. Because taxpayers pay for the collection, advocates of open data argue that the public should be able to view and use the data.

To see if your local government has open data, perform an internet search using the words “open data” and the name of your city or county, or a city near you. Most medium-to-large cities participate, as do many smaller towns and cities, although some have much more data than others.

Most cities that have open data will have many, many different opportunities. Not all these will be interesting to you, and not all of them, frankly, contain data that are useful or, indeed, any data at all.

You should approach this project in the mindset of an explorer. The structure and quality of these websites varies drastically. Some make it really easy (such as Santa Monica, CA), a few make it really hard. Some will claim they have data, but what they really provide are tools to analyze data that they won't allow you to download or view. Others will provide photos or screenshots of data but not anything you can analyze. So be brave and open-minded and expect to encounter some dead-ends!

Project

Once you've identified a government with open data, find an area or theme that interests you. For example, you might be interested in response times of the fire department, crimes, how much government employees are paid, or the health ratings of restaurants. Some cities organize data by theme, such as Finance, Public Safety, and so on. You might have to do some poking around to find a data set that is interesting.

Your goal is to download a data file. But not just any data file; we need one we can that your statistical

software will understand. To do this, you'll want to look for some sort of option to “Export” or “Download” a file. In some cases, these words might not be used and you're expected to simply click on a link.

There are three things you need to pay attention to: the format, the structure, and the existence of *metadata*.

Format refers to the types of files provided and the software that can open them. For example, files that end in *.txt*—text files—can usually be open by many types of software, whereas files that end in *.docx* can be opened only by Microsoft Word. StatCrunch wants files that end in *.xlsx*, *.xls* (Excel spreadsheets), *.ods* (Open Office spreadsheet), or *.csv*, *.txt*, *.tsc* (text files).

Datafiles often use a *delimiter* to determine where one entry ends and another begins. For example, if a data file contains this:

```
1345
```

you don't know whether it is the number one thousand three hundred forty-five or the four digits 1, 3, 4, 5 or the two numbers 13 and 45 or anything else. A delimiter tells you when one value ends and the next begins. A comma-delimited files, which usually ends in the extension *.csv*, uses commas to do this:

```
1, 34, 5
```

for the values one, thirty-four, and five.

A tab-delimited file (*.tsv* or *.txt*) uses tabs

```
1      34      5
```

Metadata is documentation that helps you understand the data. Sometimes it is called a “codebook” or “data dictionary.” It should answer the who, what, when, where, why, and how of the data. But most important, it should tell you what the variable names mean and what the values mean. With any luck, it will also tell you the data type (for example, is it a number or a character?).

Warning: you might have to look at a great many data files before you find one that is useful.

Assignment

Download a data set and write a report answering these questions

1. What is the URL of the page from which you downloaded your data?
2. What file types are available for your data? What file type did you download?
3. How many variables are there? How many observations?
4. What does a row in the data set represent? In other words, what is the basic unit of observation?
5. Why were these data collected? What is their purpose? How were they collected?
6. In terms of megabytes (or gigabytes), how large is the file? (This information might not be available on the website; you might need to determine using your computer's operating system).
7. Why were *you* interested in this data set? Give some examples of what you want to know from these data.
8. Try to upload the data to StatCrunch. Describe what happens. Did you get what you expected? Did you get an error? Describe the outcome.

CHAPTER REVIEW

KEY TERMS

statistics, 2	stacked data, 8	treatment group, 18	confounding variable
variation, 3	unstacked data, 8	comparison group (or control	(or confounding
data, 3	two-way table, 14	group), 18	factor), 20
variables, 6	frequency, 14	anecdotes, 18	controlled experiment, 21
data set, 6	rate, 17	placebo, 19	random assignment, 22
sample, 6	causality, 18	placebo effect, 19	bias, 22
population, 6	treatment variable, 18	observational	blind, 22
numerical variable, 7	outcome variable (or response	study, 19	double-blind, 23
categorical variable, 7	variable), 18	association, 19	random selection, 24

LEARNING OBJECTIVES

After reading this chapter and doing the assigned homework problems, you should

- Be able to distinguish between numerical and categorical variables and understand methods for coding categorical variables.
- Know how to find and use rates (including percentages) and understand when and why they are more useful than counts for describing and comparing groups.
- Understand when it is possible to infer a cause-and-effect relationship from a research study and when it is not.
- Be able to explain how confounding variables prevent us from inferring causation and suggest confounding variables that are likely to occur in some situations.
- Be able to distinguish between observational studies and controlled experiments.

SUMMARY

Statistics is the science (and art) of collecting and analyzing observations (called data) and communicating your discoveries to others. Often, we are interested in learning about a population on the basis of a sample taken from that population.

Statistical investigations often progress through four stages (which we call the Data Cycle): Asking questions, considering the data available to answer the questions, analyzing the data, and interpreting the analysis to answer the questions.

With categorical variables, we are often concerned with comparing rates or frequencies between groups. A two-way table is sometimes a useful summary. Always be sure that you are making valid comparisons by comparing proportions or percentages of groups or that you are comparing the appropriate rates.

Many studies are focused on questions of causality: If we make a change to one variable, will we see a change in the other?

Anecdotes are not useful for answering such questions. Observational studies can be used to determine whether associations exist between treatment and outcome variables, but because of the possibility of confounding variables, observational studies cannot support conclusions about causality. Controlled experiments, if they are well designed, do allow us to draw conclusions about causality.

A well-designed controlled experiment should have the following attributes:

- A large sample size
- Random assignment of subjects to a treatment group and to a control group
- A double-blind format
- A placebo

SOURCES

- BBC News. 2010. Happiness wards off heart disease, study suggests. February 18, 2010. <http://news.bbc.co.uk/2/hi/health/8520549.stm>.
- Britt, C. 2008. U.S. doctors prescribe drugs for placebo effects, survey shows. *British Medical Journal*. October 23, 2008. Bloomberg.com.
- Colombel, S., et al. 2010. Infliximab, azathioprine, or combination therapy for Crohn's disease. *New England Journal of Medicine*, vol. 362 (April 15): 1383–1395.
- Fisher, R. 1959. *Smoking: The cancer controversy*. Edinburgh, UK: Oliver and Boyd.
- Glionna, J. 2006. Word of mouth spreading about peanut milk. *Los Angeles Times*, May 17.
- Insurance Institute for Highway Safety. 2013. <http://www.iihs.org/news/>
- Kahleova, H., Lloren, J. I., Maschak, A., Hill, M., Fraser, G. (2017) Meal frequency and timing are associated with changes in body index in Adventist Health Study 2. *Journal of Nutrition* doi: 10.3945/jn.116.244749.
- Kaprio, J., and M. Koskenvuo. 1989. Twins, smoking and mortality: A 12-year prospective study of smoking-discordant twin pairs. *Social Science and Medicine*, vol. 29, no. 9: 1083–1089.