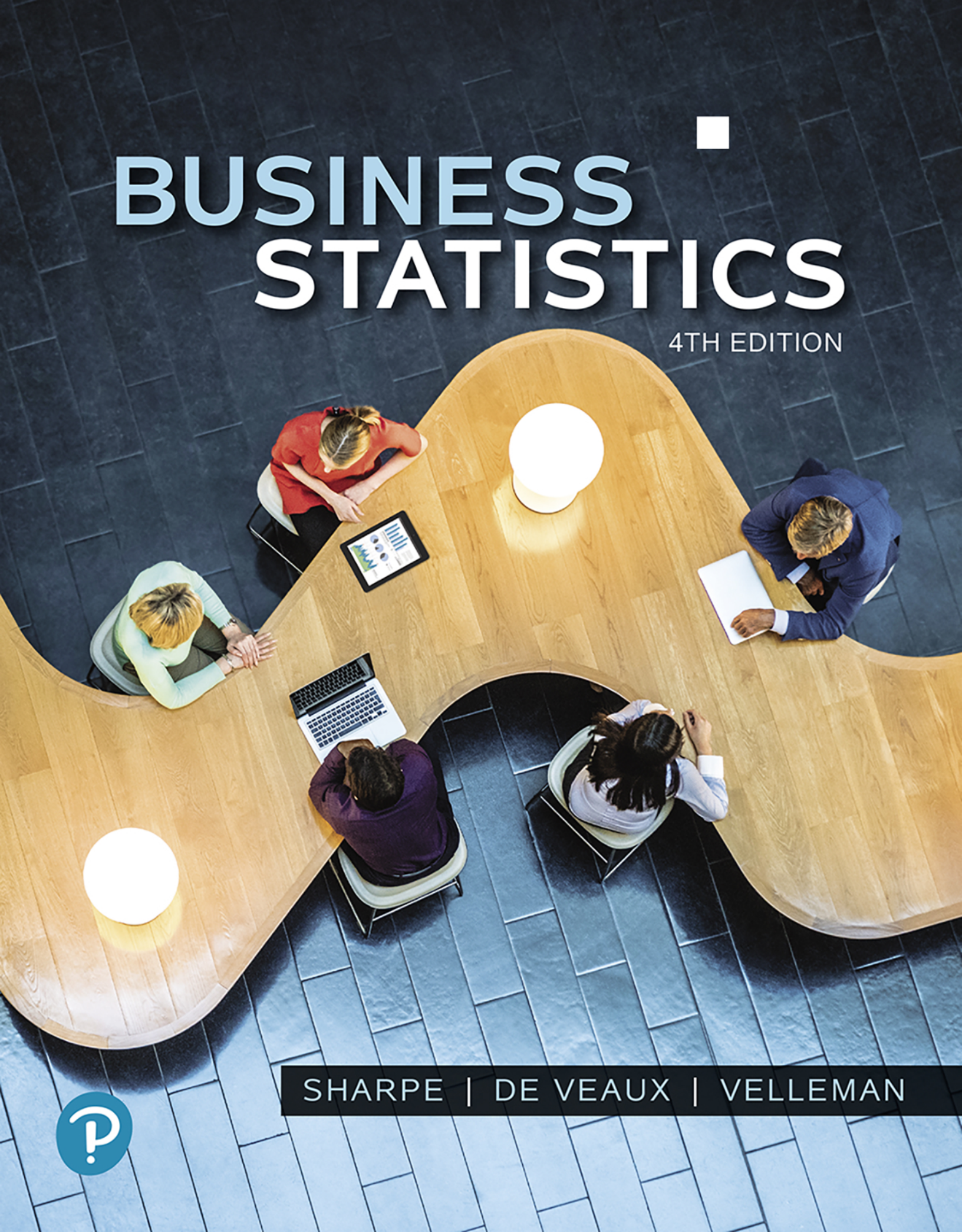
BUSINESS STATISTICS

4TH EDITION



SHARPE | DE VEAUX | VELLEMAN

4th Edition

Business Statistics

Norean R. Sharpe
St. John's University

Richard D. De Veaux
Williams College

Paul F. Velleman
Cornell University

With Contributions by David Bock
and Special Contributor Eric M. Eisenstein

Director, Portfolio Management: Deirdre Lynch
Senior Portfolio Management Analyst: Patrick Barbera
Editorial Assistant: Morgan Danna
Managing Producer: Scott Disanno
Content Producer: Peggy McMahon
Senior Publishing Services Analyst: Joe Vetere
Manager, Quality Control: Mary Durnwald
Manager, Content Development: Robert Carroll
Senior Producer: Aimee Thorne

Product Marketing Analyst: Kaylee Carlson
Marketing Support Assistant: Shannon McCormack
Manager, Rights/Permissions Gina Cheselka
Project Management, Rights & Permissions Editing: SPi Global
Text and Cover Design, Illustrations, Composition: Cenveo® Publisher Services
Cover Images: (Mobile Phone and Tablet Computer Screens) © Can Yesil/Shutterstock; (Business people meeting in modern office) © Johnny-Greig/E+/Getty Images

Library of Congress Cataloging-in-Publication Data

Names: Sharpe, Norean Radke, author. | De Veaux, Richard D., author. | Velleman, Paul F., 1949- author.
Title: Business statistics / Norean R. Sharpe, Georgetown University, Richard D. De Veaux, Williams College, Paul F. Velleman, Cornell University; with Contributions by David Bock and Eric Eisenstein.
Description: 4th Edition. | Boston, MA: Pearson, [2018] | Revised edition of the authors' Business statistics, [2015] | Includes index.
Identifiers: LCCN 2018019089 | ISBN 9780134705217 (student edition) | ISBN 0134705211
Subjects: LCSH: Commercial statistics.
Classification: LCC HF1017 .S467 2018 | DDC 650.01/5195—dc23 LC record available at https://lccn.loc.gov/2018019089

Pearson

ISBN-10: 0-134-70521-1
ISBN-13: 978-0-134-70521-7

To my loving family for their patience and support
—Norean

To my father, whose daily stories informed me how the world of business really worked, and to my family, for giving me the love and support that made this book possible
—Dick

To my father, who taught me about ethical business practice by his constant example as a small businessman and parent
—Paul

This page intentionally left blank

Meet the Authors

Norean R. Sharpe, Ph.D., is Dean and the Joseph H. and Maria C. Schwartz Distinguished Chair at The Peter J. Tobin College of Business at St. John's University. As the chief academic officer of the Tobin College of Business, she is responsible for the curriculum for 2500 undergraduate business majors and 600 graduate students in one of seven M.S./M.B.A. programs, all supported by more than 150 faculty and staff on the Manhattan, Queens, Staten Island, and Rome, Italy, campuses. Within the Tobin College is the Center for Enterprise Risk Management, the Applied Finance Institute, and the Global Business Stewardship Center, as well as the acclaimed School of Risk Management, Insurance, and Actuarial Science.
Dr. Sharpe is an accomplished scholar, with 30 years of teaching experience at Yale University, Bowdoin College, Babson College, and Georgetown University—and with more than 30 scholarly publications in analytics and statistics education. Her research interests include time series analysis, forecasting, analytics, and women's roles in entrepreneurship in the Middle East. Dr. Sharpe earned her B.A. from Mt. Holyoke College, her M.S. from the University of North Carolina, and her Ph.D. in Systems Engineering from the University of Virginia.

Richard D. De Veaux (Ph.D. Stanford University) is an internationally known educator, consultant, and lecturer. Dick has taught statistics at a business school (Wharton), an engineering school (Princeton), and a liberal arts college (Williams). While at Princeton, he won a Lifetime Award for Dedication and Excellence in Teaching. Since 1994, he has taught at Williams College, although he returned to Princeton for the academic year 2006–2007 as the William R. Kenan Jr. Visiting Professor of Distinguished Teaching. He is currently the C. Carlisle and Margaret Tippit Professor of Statistics at Williams College. Dick holds degrees from Princeton University in Civil Engineering and Mathematics and from Stanford University where he studied statistics with Persi Diaconis and dance with Inga Weiss. His research focuses on the analysis of large datasets and data mining in science and industry. Dick has won both the Wilcoxon and Shewell awards from the American Society for Quality. He is an elected member of the International Statistics Institute (ISI) and a Fellow of the American Statistical Association (ASA). Dick was elected Vice President of the ASA in 2018 and will serve from 2019 to 2021. Dick is also well known in industry, having consulted for such *Fortune* 500 companies as American Express, Hewlett-Packard, Alcoa, DuPont, Pillsbury, General Electric, and Chemical Bank. He was named the "Statistician of the Year" for 2008 by the Boston Chapter of the American Statistical Association. In his spare time he is an avid cyclist and swimmer, and is a frequent singer and soloist with various local choirs, including the Choeur Vittoria of Paris, France. Dick is the father of four children.

Paul F. Velleman (Ph.D. Princeton University) has an international reputation for innovative statistics education. He designed the Data Desk® software package and is also the author and designer of the award-winning ActivStats® multimedia software, for which he received the EDUCOM Medal for innovative uses of computers in teaching statistics and the ICTCM Award for Innovation in Using Technology in College Mathematics. He is the founder and CEO of Data Description, Inc. (**www.datadesk.com**), which supports both of these programs. Data Description also developed and maintains the Internet site *Data and Story Library* (DASL; **dasl.datadescription.com**), which provides all of the datasets used in this text as well as many others useful for teaching statistics, and the statistics conceptual tools at **astools.datadesk.com**. Paul coauthored (with David Hoaglin) the book *ABCs of Exploratory Data Analysis*. Paul is Emeritus Professor of Statistical Sciences, at Cornell University where he was awarded the MacIntyre Prize for Exemplary Teaching. Paul earned his M.S. and Ph.D. from Princeton University, where he studied with John Tukey. His research often focuses on statistical graphics and data analysis methods. Paul is a Fellow of the American Statistical Association and of the American Association for the Advancement of Science. He was a member of the working group that developed the GAISE 2016 guidelines for teaching statistics. Paul's experience as a professor, entrepreneur, and business leader brings a unique perspective to the book.

Richard De Veaux and Paul Velleman have authored successful books in the introductory college and AP High School market with David Bock, including *Intro Stats*, Fifth Edition (Pearson, 2018); *Stats: Modeling the World*, Fifth Edition (Pearson, 2019); and *Stats: Data and Models*, Fourth Edition (Pearson, 2016).

Special Contributor

Eric M. Eisenstein (Ph.D. Wharton School of Business) is an internationally known educator, researcher, and consultant. Eric has taught at multiple business schools, including Wharton, Cornell's Johnson School, ESADE, and Temple University's Fox School of Business. At Fox, he serves as the Director of the MS in Business Analytics in the department of Statistical Science, Director of Graduate Programs in the department of Marketing and Supply Chain Management, and Chair of the Undergraduate Program (curriculum) Committee. Eric teaches data analytics, quantitative strategy, and marketing. His research focuses on the psychology of expertise, how to improve decision making, and strategic analytics. Prior to becoming an academic, Eric worked at Mercer Management Consulting (now Oliver Wyman) where he focused on management of technology and marketing research in the financial services and telecommunications industries. His teams won the outstanding team award three times consecutively; clients invested over $30 million based on the recommendations of his teams, and the teams' strategic recommendations affected more than $10 billion in revenue and $2 billion in profits. He continues to consult and serve on the board of numerous companies and charities. Eric earned his Ph.D. in Applied Economics and an M.A. in Statistics at the Wharton School of Business, University of Pennsylvania and graduated from the Management and Technology dual degree program at the University of Pennsylvania, where he concurrently earned a B.S. in Economics from Wharton and a B.S. in Computer Systems Engineering from the School of Engineering and Applied Science. He is the proud father to three children.

Contents

Preface

The question that should motivate a business student's study of statistics should be "Even without perfect information, how can I make better decisions?"[1] As entrepreneurs and consultants, we know that in today's data-rich environment, knowledge of statistics is essential to survive and thrive in the business world. But, as educators, we've seen a disconnect between the way business statistics is traditionally taught and the way it should be used in making business decisions. In *Business Statistics*, we try to narrow the gap between theory and practice by presenting relevant statistical methods that will empower business students to make effective, data-informed decisions.

Of course, students should come away from their statistics course knowing how to think statistically and how to apply statistics methods with modern technology. But they must also be able to communicate their analyses effectively to others. When asked about statistics education, a group of CEOs from *Fortune* 500 companies recently said that although they were satisfied with the technical competence of students who had studied statistics, they found the students' ability to communicate their findings to be woefully inadequate.

Our Plan, Do, Report rubric provides a structure for solving business problems that mimics the correct application of statistics to solving real business problems. Unlike many other authors, we emphasize the often neglected thinking (Plan) and communication (Report) steps in problem solving in addition to the methodology (Do). This approach requires up-to-date, real-world examples and data. So we constantly strive to illustrate our lessons with current business issues and examples.

What's New in This Edition?

We've been delighted with the reaction to previous editions of *Business Statistics.* We've made some changes to the organization of the fourth edition to help students focus on the essentials and think about the data-rich world they will find in the workplace. And, of course, we continue to update examples and exercises so that the story we tell is always tied to the ways statistics informs modern business practice.

- **Recent data.** We teach with real data whenever possible, so we've updated data throughout the book. New examples reflect current stories in the news and recent economic and business events. When a historical dataset is especially good at illuminating a pedagogical point, we have, from time to time, chosen pedagogy over recency.
- **Improved organization.** We have retained our "data first" presentation of topics because we find that it provides students with both motivation and a foundation in real business decisions on which to build an understanding.
    - Chapters 1–4 have been streamlined to cover collecting, displaying, summarizing, and understanding data in four chapters. We find that this provides students with a solid foundation to launch their study of probability and statistics.
    - Chapters 5–7 introduce students to randomness and probability models. We've moved the discussion of probability trees and Bayes' rule into these chapters.
    - Chapters 8 and 9 cover data collection by survey and by designed experiments. New discussions here address technology-enabled sampling, online data, and Big Data. We've moved the discussion of experiments up front because of the increased importance of online testing, but we've

[1] Unfortunately, not the question most students are asking themselves on the first day of the course.

moved the analysis of such designs (ANOVA), which many instructors find difficult to cover in a first course, to the online Chapter 25.
  - Chapters 10–15 cover inference for both proportions and means. We introduce inference by discussing proportions because most students are better acquainted with proportions reported in surveys and news stories. However, this edition ties in the discussion of means immediately so students can appreciate that the reasoning of inference is the same in a variety of contexts. We've added an optional discussion of bootstrapping. This may help students' intuition about inference as well as providing a relatively new modern method.
  - Chapters 16–19 cover regression-based models for decision making.
  - Chapter 20 discusses time series methods.
  - Chapter 21 is a newly expanded discussion of data mining and Big Data.
  - Chapters 22–24 discuss special topics that can be selected according to the needs of the course and the preferences of the instructor.
- **Streamlined design.** Our goal has always been a readable text. This edition sports a new design that clarifies the purpose of each text element. The major theme of each chapter is linear and easy to follow without distraction. Supporting material is clearly boxed and shaded, so students know where to focus their study efforts.
- **Enhanced Technology Help.** We've updated Technology Help (now called Tech Support) in almost every chapter.
- **Updated examples to reflect the changing world.** The time since our last revision has seen marked changes in the U.S. and world economies. This has required us to update many of our examples. Our selection of course content reflects the wisdom of the GAISE2016 report adopted by the American Statistical Association as a standard for introductory statistics teaching. Our "In Practice" elements have all been re-structured to reflect real-world business challenges. The result is a text that is realistic and useful.
- **Increased focus on core material.** Statistics in practice means making smart decisions based on data. Students need to know the methods, how to apply them, and the assumptions and conditions that make them work. We've tightened our discussions to get students there as quickly as possible, focusing increasingly on the central ideas and core material.

Our Approach

Statistical Thinking

For all of our improvements, examples, and updates in this edition of *Business Statistics* we haven't lost sight of our original mission—writing a modern business statistics text that addresses the importance of *statistical thinking* in making business decisions and that acknowledges how Statistics is actually used in business.

Statistics is practiced with technology, and this insight informs everything from our choice of forms for equations (favoring intuitive forms over calculation forms) to our extensive use of real data. But most important, understanding the value of technology allows us to focus on teaching statistical thinking rather than calculation. The questions that motivate each of our hundreds of examples are not "How do you find the answer?" but "How do you think about the answer?"; "How does it help you make a better decision?"; and "How can you best communicate your decision?" Our redesigned "In Practice" elements in each chapter have been recast as conversations between managers and analysts to emphasize the business relevance of each method and its importance in making good business decisions.

Our focus on statistical thinking ties the chapters of the book together. An introductory Business Statistics course covers an overwhelming number of new terms, concepts, and methods, and it is vital that students see their central core: how we can understand more about the world and make better decisions by understanding what the data tell us. From this perspective, it is easy to see that the patterns we look for in graphs are the same as those we think about when we prepare to make inferences. And it is easy to see that the many ways to draw inferences from data are several applications of the same core concepts. It follows naturally that when we extend these basic ideas into more complex (and even more realistic) situations, the same basic reasoning is still at the core of our analyses.

Our Goal: Read This Book!

The best textbook in the world is of little value if it isn't read. Here are some of the ways we made *Business Statistics* more approachable:

- ***Readability.*** We strive for a conversational, approachable style, and we introduce anecdotes to maintain interest. Instructors report (to their amazement) that their students read ahead of their assignments voluntarily. Students tell us (to *their* amazement) that they actually enjoy the book. In this edition, we've focused our discussions even more clearly on the central ideas we want to convey.
- ***Focus on assumptions and conditions.*** More than any other textbook, *Business Statistics* emphasizes the need to verify assumptions when using statistical procedures. We reiterate this focus throughout the examples and exercises. We make every effort to provide templates that reinforce the practice of checking these assumptions and conditions, rather than rushing through the computations. Business decisions have consequences. Blind calculations open the door to errors that could easily be avoided by taking the time to graph the data, check assumptions and conditions, and then check again that the results and residuals make sense.
- ***Emphasis on graphing and exploring data.*** Our consistent emphasis on the importance of displaying data is evident from the first chapters on understanding data to the sophisticated model-building chapters at the end. Examples often illustrate the value of examining data graphically, and the exercises reinforce this. Good graphics reveal structures, patterns, and occasional anomalies that could otherwise go unnoticed. These patterns often raise new questions and inform both the path of a resulting statistical analysis and the business decisions. Hundreds of new graphics found throughout the book demonstrate that the simple structures that underlie even the most sophisticated statistical inferences are the same ones we look for in the simplest examples. This helps tie the concepts of the book together to tell a coherent story.
- ***Consistency.*** We work hard to avoid the "do what we say, not what we do" trap. Having taught the importance of plotting data and checking assumptions and conditions, we are careful to model that behavior throughout the book. (Check the exercises in the chapters on multiple regression or time series and you'll find us still requiring and demonstrating the plots and checks that were introduced in the early chapters.) This consistency helps reinforce these fundamental principles and provides a familiar foundation for the more sophisticated topics.
- ***The need to read.*** In this book, important concepts, definitions, and sample solutions are not always set aside in boxes. The book needs to be read, so we've tried to make the reading experience enjoyable. The common approach of skimming for definitions or starting with the exercises and looking up examples just won't work here. (It never did work as a way to learn about and understand statistics.)

Coverage

The topics covered in a Business Statistics course are generally mandated by our students' needs in their studies and in their future professions. But the *order* of these topics and the relative emphasis given to each is not well established. *Business Statistics* presents some topics sooner or later than other texts. Although many chapters can be taught in a different order, we urge you to consider the order we have chosen.

We've been guided in the order of topics by the fundamental goal of designing a coherent course in which concepts and methods fit together to provide a new understanding of how reasoning with data can uncover new and important truths. Each new topic should fit into the growing structure of understanding that students develop throughout the course. For example, we teach inference concepts with proportions first and then with means. Most people have a wider experience with proportions, seeing them in polls and advertising. And by starting with proportions, we can teach inference with the Normal model and then introduce inference for means with the Student's *t*-distribution.

We introduce the concepts of association, correlation, and regression early in *Business Statistics*. Our experience in the classroom shows that introducing these fundamental ideas early makes statistics useful and relevant even at the beginning of the course. By Chapter 4, students can discuss relationships among variables in a meaningful way. Later in the semester, when we discuss inference, it is natural and relatively easy to build on the fundamental concepts learned earlier and enhance them with inferential methods.

GAISE Report

We've been guided in our choice of what to emphasize by the GAISE 2016 (Guidelines for Assessment and Instruction in Statistics Education) Report, which emerged from extensive studies of how students best learn Statistics (**www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf**). The GAISE Report was extensively revised in 2016 to reflect the evolution of technology and new wisdom about teaching statistics. The new recommendations have been officially adopted and recommended by the American Statistical Association and urge (among other detailed suggestions) that statistics education should:

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

In this sense, this book is thoroughly modern.

Syllabus Flexibility

To be effective, a course must fit comfortably with the instructor's preferences. The early chapters—Chapters 1–15—cover core material that will be part of most introductory courses. Chapters 16–20—multiple regression, model building, and time series. Analysis of Variance—may be included in an introductory course, but our organization provides flexibility in the order and choice of specific topics. Chapters 21–25 may be viewed as "special topics" and selected and sequenced to suit the instructor or the course requirements.

Here are some specific notes:

- Chapter 4, Correlation and Linear Regression, may be postponed until just before covering regression inference in Chapter 16. (But we urge you to teach it where it appears.) Chapter 4 now includes an early glimpse of multiple regression (as advised by GAISE 2016). We urge you not to skip that discussion.
- Chapter 19, Building Multiple Regression Models, must follow the introductory material on multiple regression in Chapter 18.
- Chapters 20 and 25, Time Series Analysis and ANOVA, require material on multiple regression from Chapter 18.

The following topics can be introduced in any order (or omitted) after basic inference has been covered:

- Chapter 15, Inference for Counts: Chi-Square Tests
- Chapter 21, Introduction to Big Data and Data Mining
- Chapter 22, Quality Control
- Chapter 23, Nonparametric Methods
- Chapter 24, Decision Making and Risk

Continuing Features

A textbook isn't just words on a page. A textbook is many elements that come together to form a big picture. The features in *Business Statistics* provide a real-world context for concepts, help students apply these concepts, promote problem solving, and integrate technology—all of which help students understand and see the big picture of Business Statistics.

Providing Real-World Context

Motivating Vignettes. Each chapter opens with a motivating vignette, often taken from the authors' consulting experiences. Companies featured include Amazon.com, Zillow.com, Keen Inc., and Whole Foods Market. We analyze data from or about the companies in the motivating vignettes throughout the chapter.

Brief Cases. Each chapter includes one or more Brief Cases that use real data and ask students to investigate a question or make a decision. Students define the objective, plan the process, complete the analysis, and report a conclusion. Data for the Brief Cases are available on the website, formatted for various technologies.

Case Studies. Throughout the book we present Case Studies. Students are given realistically large datasets and challenged to respond to open-ended business questions using the data. Students can bring together methods they have learned throughout the book to address the issues raised. Students will have to use a computer to work with the large datasets that accompany these Case Studies.

What Can Go Wrong? In each chapter, What Can Go Wrong? highlights the most common statistical errors and the misconceptions about statistics. The most common mistakes for the new user of statistics often involve misusing a method—not miscalculating a statistic. One of our goals is to arm students with the tools to detect statistical errors and to offer practice in debunking misuses of Statistics, whether intentional or not.

Applying Concepts

In Practice. Almost every section of every chapter includes focused examples that illustrate and apply the concepts or methods of that section to a real-world business context. Each one now ends with a specific written report. They are now structured as conversations between a manager and an analyst or employee with the requirement that a report be made to the manager. This format helps to frame the issues in a practical way.

Step-by-Step Guided Examples. The answer to a statistical question is almost never just a number. Statistics is about understanding the world and making better decisions with data. Guided Examples model a thorough solution in the right column with commentary in the left column. The overall analysis follows our innovative **Plan, Do, Report** template. Each analysis begins with a clear question about a business decision and an examination of the data (**Plan**), moves to calculating the selected statistics (**Do**), and finally concludes with a **Report** that specifically addresses the question. To emphasize that our goal is to address the motivating question, we present the **Report** step as a business memo that summarizes the results in the context of the example and states a recommendation if the data are able to support one. To preserve the realism of the example, whenever it is appropriate, we include limitations of the analysis or models in the concluding memo, as one should in making such a report.

By Hand. Even though we encourage the use of technology to calculate statistical quantities, we recognize the pedagogical benefits of occasionally doing a calculation by hand. The By Hand boxes break apart the calculation of some of the simpler formulas and help the student through the calculation of a worked example.

Reality Check. We regularly offer reminders that statistics is about understanding the world and making decisions with data. Results that make no sense are probably wrong, no matter how carefully we think we did the calculations. Mistakes are often easy to spot with a little thought, so we ask students to stop for a reality check before interpreting results.

Notation Alert. Throughout this book, we emphasize the importance of clear communication. Proper notation is part of the vocabulary of statistics, but it can be daunting. We've found that it helps students when we are clear about the letters and symbols statisticians use to mean very specific things, so we've included Notation Alerts whenever we introduce a special notation that students will see again.

Math Boxes. When we present the mathematical underpinnings of the statistical methods and concepts, we set proofs, derivations, and justifications apart from the narrative. In this way, the underlying mathematics is there for those who want greater depth, but the text itself presents the logical development of the topic at hand without distractions.

From Learning to Earning. Each chapter ends with a From Learning to Earning summary that includes learning objectives and definitions of terms introduced in the chapter. Students should use these as study guides. We encourage them to take this opportunity to see the "big picture" of the chapter and see how it applies to making business decisions.

Promoting Problem Solving

Just Checking. Throughout each chapter we pose short questions to help students check their understanding. The answers are at the end of the exercise sets in each chapter to make them easy to check. The questions can also be used to motivate class discussion.

Ethics in Action. Statistics is not just plugging numbers into formulas; most statistical analyses require a fair amount of judgment. Ethics in Action vignettes—updated for this edition—in each chapter provide a context for some of the judgments needed in statistical analyses. Possible errors, a link to the American Statistical Association's Ethical Guidelines, and ethically and statistically sound alternative approaches are presented in the Instructor's Solutions Manual.

Section Exercises. The exercises for each chapter begin with straightforward exercises targeted at the topics in each section. These are designed to check understanding of specific topics. Because they are labeled by section, it is easy to turn back to the chapter to clarify a concept or review a method.

Chapter Exercises. These exercises are designed to be more realistic than section exercises and to lead to conclusions about the real world. They may combine concepts and methods from different sections, and they contain relevant, modern, and real-world questions. Many come from news stories; some come from recent research articles. The exercises marked with a **T** indicate that the data are available on the book's companion website, in a variety of formats. We pair the exercises so that each odd-numbered exercise (with answer in the back of the book) is followed by an even-numbered exercise on the same statistics topic. Exercises are roughly ordered within each chapter by both topic and level of difficulty.

Integrating Technology

Data and Sources. Most of the data used in examples and exercises are from real-world sources and whenever we can, we include URLs for Internet data sources. The data we use, are usually available at the online Data and Story Library (DASL) at **dasl.datadescription.com** and on the companion website, **www.pearsonhighered.com/sharpe**.

Videos with Optional Captioning. Videos, featuring the *Business Statistics* authors, review the high points of each chapter. The presentations feature the same student-friendly style and emphasis on critical thinking as the textbook. In addition, 10 *Business Insight Videos* feature Deckers, Southwest Airlines, Starwood, and other companies and focus on statistical concepts as they pertain to the real world. Videos are available with captioning. They can also be viewed from within the online MyLab Statistics course.

Tech Support. In business, statistics is practiced with computers using a variety of statistics packages. In Business-school statistics classes, however, Excel is the software most often used. In the Tech Support sections at the end of each chapter, we summarize what students can find in the most common software, often with annotated output. In updating for this edition, we offer extended guidance for Excel 2016, and start-up pointers for Minitab, SPSS, JMP, StatCrunch, R, and XLStat, formatted in easy-to-read bulleted lists. This advice is not intended to replace the documentation for any of the software, but rather to point the way and provide start-up assistance.

Get the Most Out of MyLab Statistics

Pearson MyLab

MyLab™ Statistics is the leading online homework, tutorial, and assessment program for teaching and learning statistics, built around Pearson's best-selling content. MyLab Stats helps students and instructors improve results; it provides engaging experiences and personalized learning for each student so learning can happen in any environment. Plus, it offers flexible and time-saving course management features to allow instructors to easily manage their classes while remaining in complete control, regardless of course format.

Preparedness

One of the biggest challenges in many mathematics and statistics courses is making sure students are adequately prepared with the prerequisite skills needed to successfully complete their course work. Pearson offers a variety of content and course options to support students with just-in-time remediation and key-concept review.

- Build homework assignments, quizzes, and tests to support your course learning outcomes. From Getting Ready (GR) questions to the Conceptual Question Library (CQL), we have your assessment needs covered from the mechanics to the critical understanding of Statistics. The exercise libraries include technology-led instruction, including new Excel-based exercises, and learning aids to reinforce your students' success.
- Using proven, field-tested technology, auto-graded Excel Projects allow instructors to seamlessly integrate Microsoft® Excel® content into their course without having to manually grade spreadsheets. Students have the opportunity to practice important statistical skills in Excel, helping them to master key concepts and gain proficiency with the program.

---

pearson.com/mylab/statistics

Resources for Success

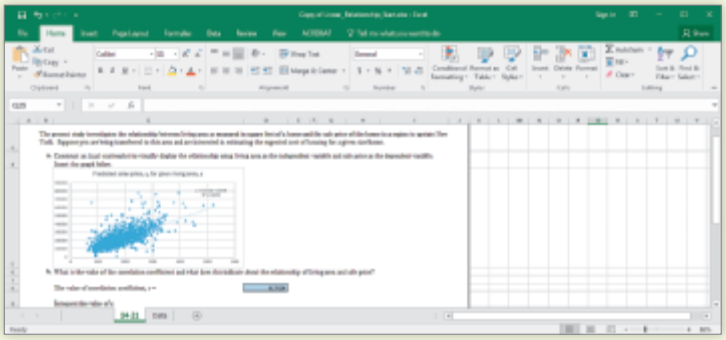
MyLab Statistics Online Course for Business Statistics, Fourth Edition, by Sharpe/De Veaux/Velleman

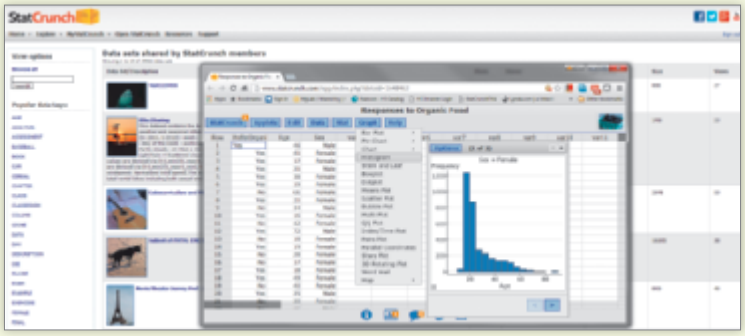
(access code required)

MyLab™ Stats 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! Auto-Graded Excel Grader Projects

Using proven, field-tested technology, auto-graded Excel Projects allow instructors to seamlessly integrate Microsoft® Excel® content into their course without having to manually grade spreadsheets.



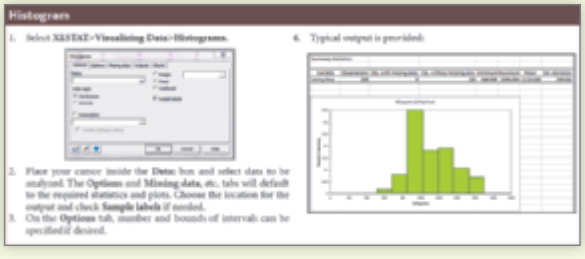


StatCrunch

StatCrunch, a powerful, web-based statistical software, is integrated into MyLab, so students can quickly and easily analyze datasets from their text and exercises. In addition, MyLab includes access to **www.StatCrunch.com**, the full web-based program where users can access tens of thousands of shared datasets, create and conduct online surveys, interact with a full library of applets, and perform complex analyses using the powerful statistical software.

Technology Tutorials and Study Cards

Excel® tutorials provide brief video walk-throughs and step-by-step instructional study cards on common statistical procedures such as Confidence Intervals, ANOVA, Simple & Multiple Regression, and Hypothesis Testing. Tutorials will capture methods in Microsoft Windows Excel® 2010, 2013, and 2016 versions.



pearson.com/mylab/statistics

Pearson MyLab

Resources for Success

Instructor Supplements

Instructor's Edition contains answers to all exercises. (ISBN-13: 978-0-13-468758-2; ISBN-10: 0-13-468758-2)

Instructor's Resource Guide (download only), written by the authors, contains chapter-by-chapter comments on the major concepts, tips on presenting topics (and what to avoid), teaching examples, suggested assignments, basic exercises, and web links and lists of other resources. Available to qualified instructors through Pearson's online catalog at **www.pearson.com/us/higher-education** or within MyLab Statistics.

Online Test Bank (download only), by Dirk Tempelaar, Maastricht University, includes chapter quizzes and part-level tests. Available to qualified instructors through Pearson's online catalog at **www.pearson.com/us/higher-education** or within MyLab Statistics.

Instructor's Solutions Manual (download only), by Linda Dawson, University of Washington, contains detailed solutions to all of the exercises. The Instructor's Solutions Manual is available to qualified instructors through Pearson's online catalog at **www.pearson.com/us/higher-education** or within MyLab Statistics.

TestGen® Computerized Test Bank (**www.pearsoned.com/testgen**) enables instructors to build, edit, print, and administer tests using a computerized bank of questions developed to cover all the objectives of the text. TestGen is algorithmically based, allowing instructors to create multiple but equivalent versions of the same question or test with the click of a button. Instructors can also modify test bank questions or add new questions. The software and test bank are available for download from Pearson's online catalog at **www.pearson.com/us/higher-education**. Test Forms (download only) are also available from the online catalog.

PowerPoint Lecture Slides: Free to qualified adopters, this classroom lecture presentation software is geared specifically to the sequence and philosophy of *Business Statistics*. Key graphics from the book are included to help bring the statistical concepts alive in the classroom. These files are available to qualified instructors through Pearson's online catalog at **www.pearson.com/us/higher-education** or within MyLab Statistics.

Learning Catalytics™ is a web-based engagement and assessment tool. As a "bring-your-own-device" direct response system, Learning Catalytics offers a diverse library of dynamic question types that allow students to interact with and think critically about statistical concepts. As a real-time resource, instructors can take advantage of critical teaching moments both in the classroom and through assignable and gradable homework.

Student Resources

Business Statistics, for-sale student edition. (ISBN-13: 978-0-13-470521-7; ISBN-10: 0-13-470521-1)

Student's Solutions Manual, by Linda Dawson, University of Washington, provides detailed, worked-out solutions to odd-numbered exercises. (ISBN-13: 978-0-13-470548-4; ISBN-10: 0-13-470548-3)

Study Cards for Business Statistics Software: This series of study cards, available for Excel 2016 with DAT: 0-13-457679-9; Excel 2016 with XLSTAT: 0-13-457683-7; StatCrunch: 0-13-397513-4, R: 0-13-522870-0; and R Studio: 0-13-522869-7 provides students with easy step-by-step guides to the most common business statistics software.

Resources for Success

Technology Resources

MyLab Statistics Online Course (access code required) MyLab™ Statistics is the teaching and learning platform that empowers you to reach *every* student. By combining trusted author content with digital tools and a flexible platform, MyLab Statistics personalizes the learning experience and improves results for each student. With MyLab Statistics and StatCrunch®, an integrated web-based statistical software program, students learn the skills they need to interact with data in the real world. Learn more about MyLab Statistics at **pearson.com/mylab/statistics**.

Used by nearly one million students a year, MyLab Statistics is the world's leading online program for teaching and learning statistics. MyLab Statistics delivers assessment, tutorials, and multimedia resources that provide engaging and personalized experiences for each student, so learning can happen in any environment. Each course is developed to accompany Pearson's best-selling content, authored by thought leaders across the statistics curriculum, and can be easily customized to fit any course format.

Methods for teaching statistics are continuously evolving to provide today's students with the skills they need to interact with data in the real world. In addition, statistics students are coming to the classroom with a wide range of backgrounds and learner styles. The flexibility to build a course that fits instructors' individual course formats and every student's needs—with a variety of content options and multimedia resources all in one place—has made MyLab Statistics the market-leading solution for teaching and learning statistics since its inception.

Thanks to feedback from instructors and students from more than 10,000 institutions, MyLab Statistics continues to transform—delivering new content, innovative learning resources, and platform updates to support students and instructors, today and in the future.

Deliver Trusted Content

You deserve teaching materials that meet your own high standards for your course. That's why Pearson partners with highly respected authors to develop interactive content and course-specific resources that you can trust—and that keep your students engaged.

Tutorial Exercises with Multimedia Learning Aids: The homework and practice exercises in MyLab Statistics align with the exercises in the textbook, and they regenerate algorithmically to give students unlimited opportunity for practice and mastery. Exercises offer immediate helpful feedback, guided solutions, sample problems, animations, videos, and eText clips for extra help at point-of-use.

Auto-Graded Excel Projects: Using proven, field-tested technology, auto-graded Excel Projects let you seamlessly integrate Microsoft® Excel® content into your course without having to manually grade spreadsheets. Students can practice important statistical skills in Excel, helping them master key concepts and gain proficiency with the program. They simply download a spreadsheet, work live on a statistics problem in Excel, and then upload that file back into the MyLab. Within minutes, they receive a report that provides personalized, detailed feedback to pinpoint where they went wrong in the problem.

StatCrunch: MyLab Statistics integrates the web-based statistical software, StatCrunch, within the online assessment platform so that students can easily analyze datasets from exercises and the text. In addition, MyLab Statistics includes access to **www.StatCrunch.com**, a website where users can access tens of thousands of shared datasets, conduct online surveys, perform complex analyses using the powerful statistical software, and generate compelling reports.

Business Insight Videos: Ten engaging videos show managers at top companies using statistics in their everyday work. Assignable questions encourage debate and discussion.

StatTalk Videos: Fun-loving statistician Andrew Vickers takes to the streets of Brooklyn, New York, to demonstrate important statistical concepts through interesting stories and real-life events. This series of 24 videos includes available assessment questions and an instructor's guide.

Empower Each Learner

Each student learns at a different pace. Personalized learning pinpoints the precise areas where each student needs practice, giving all students the support they need—when and where they need it—to be successful.

Pearson MyLab

- **Study Plan:** Acts as a tutor, providing personalized recommendations for each of your students based on his or her ability to master the learning objectives in your course. This allows students to focus their study time by pinpointing the precise areas they need to review, and allowing them to use customized practice and learning aids—such as videos, eText, tutorials, and more—to get them back on track. Using the report available in the Gradebook, you can tailor course lectures to prioritize the content where students need the most support, offering you better insight into classroom and individual performance.
- With the **Companion Study Plan Assignments** you can now assign the Study Plan as a prerequisite to a test or quiz, guiding students through the concepts they need to master.
- **Getting Ready for Statistics:** A library of questions now appears within each MyLab Statistics course to offer the developmental math topics students need for the course. These can be assigned as a prerequisite to other assignments, if desired.

Conceptual Question Library: In addition to algorithmically regenerated questions that are aligned with your textbook, there is a library of 1,000 Conceptual Questions available in the assessment manager that require students to apply their statistical understanding.

Teach the Course Your Way

Your course is unique. So whether you'd like to build your own assignments, teach multiple sections, or set prerequisites, MyLab gives you the flexibility to easily create *your* course to fit *your* needs.

- **Learning Catalytics:** Generate class discussion, guide your lecture, and promote peer-to-peer learning with real-time analytics. MyLab Statistics now provides Learning Catalytics™—an interactive student response tool that uses students' smartphones, tablets, or laptops to engage them in more sophisticated tasks and thinking.
- **LMS Integration:** You can now link Blackboard Learn™, Brightspace® by D2L®, Canvas™, or Moodle® to the MyLabs. Access assignments, rosters, and resources, and synchronize grades with your LMS gradebook. For students, single sign-on provides access to all the personalized learning resources that make studying more efficient and effective.

Improve Student Results

When you teach with MyLab, student performance improves. That's why instructors have chosen MyLab for over 15 years, touching the lives of more than 50 million students.

MathXL Online Course (access code required)

Part of the world's leading collection of online homework, tutorial, and assessment products, MathXL® delivers assessment and tutorial resources that provide engaging and personalized experiences for each student. Each course is developed to accompany Pearson's best-selling content, authored by thought leaders across the math curriculum, and can be easily customized to fit any course format.

With MathXL, instructors can:

- Create, edit, and assign online homework and tests using algorithmically generated exercises correlated at the objective level to the textbook.
- Create and assign their own online exercises and import TestGen tests for added flexibility.
- Maintain records of all student work tracked in MathXL's online gradebook.

With MathXL, students can:

- Take chapter tests in MathXL and receive personalized study plans and/or personalized homework assignments based on their test results.
- Use the study plan and/or the homework to link directly to tutorial exercises for the objectives they need to study.
- Access supplemental animations and video clips directly from selected exercises.

MathXL is available to qualified adopters. For more information, visit our web site at **www.mathxl.com** or contact your Pearson representative.

StatCrunch

Integrated directly into MyLab Statistics, StatCrunch® is powerful web-based statistical software that allows users to perform complex analyses, share datasets, and generate compelling reports of their data.

pearson.com/mylab/statistics

The vibrant online community offers tens of thousands of shared datasets for students to analyze.

- **Collect.** Users can upload their own data to StatCrunch or search a large library of publicly shared datasets, spanning almost any topic of interest. Datasets from the text and from online homework exercises can also be accessed and analyzed in StatCrunch. An online survey tool allows users to quickly collect data via web-based surveys.
- **Crunch.** A full range of numerical and graphical methods allows users to analyze and gain insights from any dataset. Interactive graphics help users understand statistical concepts, and are available for export to enrich reports with visual representations of data.
- **Communicate.** Reporting options help users create a wide variety of visually appealing representations of their data.

StatCrunch is also available by itself to qualified adopters. It can be accessed on your laptop, smartphone, or tablet when you visit the StatCrunch website from your device's browser. For more information, visit the StatCrunch website at **www.StatCrunch.com** or contact your Pearson representative.

TestGen

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

PowerPoint Lecture Slides

PowerPoint® Lecture Slides provide an outline to use in a lecture setting, presenting definitions, key concepts, and figures from the text. These slides are available within MyLab Statistics and in the Instructor Resource Center at **www.pearsonhighered.com/irc**.

Foster student engagement and peer-to-peer learning

Generate class discussion, guide your lecture, and promote peer-to-peer learning with real-time analytics.

MyLab™ Math and MyLab Statistics now provide Learning Catalytics™—an interactive student response tool that uses students' smartphones, tablets, or laptops to engage them in more sophisticated tasks and thinking.

Instructors, you can:

- Pose a variety of open-ended questions that help your students develop critical thinking skills.
- Monitor responses to find out where students are struggling.
- Use real-time data to adjust your instructional strategy and try other ways of engaging your students during class.
- Manage student interactions by automatically grouping students for discussion, teamwork, and peer-to-peer learning.

XLSTAT™ for Pearson

Used by leading businesses and universities, XLSTAT is an Excel® add-in that offers a wide variety of functions to enhance the analytical capabilities of Microsoft Excel, making it the ideal tool for your everyday data analysis and statistics requirements. XLSTAT is compatible with all Excel versions. Available for bundling with the text as ISBN-13: 978-0-321-75940-5; ISBN-10: 0-321-75940-0. Standalone: ISBN-13: 978-0-321-75932-0; ISBN-10: 0-321-75932-X

Minitab and Minitab Express™

Minitab and Minitab Express™ make learning statistics easy and provide students with a skill set that's in demand in today's data-driven workforce. 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 the latest version of Minitab ensures that students can use the software for the duration of their course. Access Card only; not sold as standalone: ISBN 13: 978-0-13-445640-9; ISBN 10: 0-13-445640-8.

JMP Student Edition

JMP Student Edition is an easy-to-use, streamlined version of JMP desktop statistical discovery software from SAS Institute, Inc., and is available for bundling with the text. Access Card only; not sold as standalone: ISBN-13: 978-0-13-467979-2; ISBN-10: 0-13-467979-2.

Acknowledgments

This book would not have been possible without many contributions from David Bock, our coauthor on several other texts. Many of the explanations and exercises in this book benefit from Dave's pedagogical flair and expertise. We are honored to have him as a colleague and friend.

Many people have contributed to this book from the first day of its conception to its publication. *Business Statistics* would have never seen the light of day without the assistance of the incredible team at Pearson. The Director of Portfolio Management, Deirdre Lynch, was central to the support, development, and realization of the book from day one. Patrick Barbera, Senior Portfolio Management Analyst; Morgan Danna, Editorial Assistant; Kaylee Karlson, Product Marketing Manager; and Shannon McCormack, Marketing Support Assistant, were essential in managing all of the behind-the-scenes work that needed to be done. Peggy McMahon, Content Producer, and Chere Bemelmans, Project Manager at SPi Global, worked miracles to get the book out the door. We are indebted to them. Aimee Thorne, Senior Producer, put together a top-notch media package for this book. Designer Jerilyn Bokorick and Cenveo® Publisher Services are responsible for the wonderful way the book looks.

We'd also like to thank our accuracy checker, whose monumental task was to make sure we said what we thought we were saying: Dirk Tempelaar, Maastricht University.

We also thank those who provided feedback through focus groups, class tests, and reviews:

Hope M. Baker, Kennesaw State University
John F. Beyers, University of Maryland—University College
Scott Callan, Bentley College
Laurel Chiappetta, University of Pittsburgh
Anne Davey, Northeastern State University
Joan Donohue, The University of South Carolina
Robert Emrich, Pepperdine University
Michael Ernst, St. Cloud State
Mark Gebert, University of Kentucky
Kim Gilbert, University of Georgia
Nicholas Gorgievski, Nichols College
Clifford Hawley, West Virginia University
Kathleen Iacocca, University of Scranton
Chun Jin, Central Connecticut State University
Austin Lampros, Colorado State University
Roger Lee, Salt Lake Community College
Monnie McGee, Southern Methodist University
Richard McGowan, Boston College
Mihail Motzev, Walla Walla University
Robert Potter, University of Central Florida
Eugene Round, Embry-Riddle Aeronautical University
Sunil Sapra, California State University—Los Angeles
Dmitry Shishkin, Georgia Gwinnett College
Courtenay Stone, Ball State University
Gordon Stringer, University of Colorado—Colorado Springs

Arnold J. Stromberg, University of Kentucky
Joe H. Sullivan, Mississippi State University
Timothy Sullivan, Towson University
Minghe Sun, University of Texas—San Antonio
Patrick Thompson, University of Florida
Jackie Wroughton, Northern Kentucky University
Ye Zhang, Indiana University—Purdue Indianapolis

Finally, we want to thank our families. This has been a long project, and it has required many nights and weekends. Our families have sacrificed so that we could write the book we envisioned.

Norean Sharpe
Richard De Veaux
Paul Velleman
Eric Eisenstein

Index of Applications

Note: Page numbers followed by n indicate footnotes.

Consumers

Demographics

Distribution and Operations Management

E-Commerce

Economics

Education

Food/Drink

Games

Government, Labor, and Law

Human Resource Management/ Personnel

Insurance

Management

Manufacturing

Marketing

Media and Entertainment

Pharmaceuticals, Medicine, and Health

Politics and Popular Culture

Quality Control

Real Estate

Salary and Benefits

Sales and Retail

Science

Service Industries and Social Issues

Sports

Surveys and Opinion Polls

Technology

Transportation

CHAPTER

1

Data and Decisions

H&M

Even if you haven't bought something from H&M recently, chances are good that you've passed by one of their stores. With over 4000 stores in 64 markets worldwide, they are one of the largest and fastest-growing clothing retailers in the world. Over the past decade, H&M has built new stores at an astounding rate of over 10% a year. Thanks to this growth, the CEO, Karl-Johan Persson, grandson of the founder, is now the richest person in Sweden.

Like most companies, H&M's online presence has been increasing as well. Of their 64 worldwide markets, 35 offer e-commerce where customers can shop 24 hours a day, 7 days a week, with just the click of a mouse. H&M now reaches their customers in ways no one could even imagine just a generation ago. But what of the future? Will the company be better off continuing to grow brick and mortar stores at the same pace, or should they devote more resources into the digital space?[1]

[1]We developed this hypothetical example in late 2017 based on our business and consulting experience. As we were going to press, the news caught up with us. It turns out that indeed H&M had been struggling with their balance of online sales vs. brick and mortar inventory. Perhaps if this book had been published a year earlier, they could have solved the problem: www.nytimes.com/2018/03/27/business/hm-clothes-stock-sales.html

A few generations ago, many store owners knew their customers and their business well. With that knowledge, they could forecast growth, see trends, and even personalize their suggestions to customers, guessing which items that particular customers might like. Businesses today rely on similar information to make decisions, but most never meet their customers. With 4000 different stores and thousands of online customers, H&M has to obtain and analyze their data in other ways.

The key to turning data into information and knowledge is Statistics—the collection of tools that extract information from data. These tools that you will learn also provide the foundation for more advanced methods like data mining and analytics. According to CEO Karl-Johan Persson, "advanced analytics provide an important support for our operations. The algorithms we have started to use will contribute to improvements within everything from assortment planning and logistics to sales."[2] Using statistical methods to turn data into information, information into knowledge, and knowledge into smart business decisions is the key to all successful modern business enterprises.

And it all starts . . . with data.

Thomasine has just landed her first job out of school as a marketing and strategy analyst working for H&M. Her team's first assignment is to decide whether to build more brick and mortar stores or invest more in online operations. To help make the decision, they investigate store sales data over the past ten years and display them in the following graph:

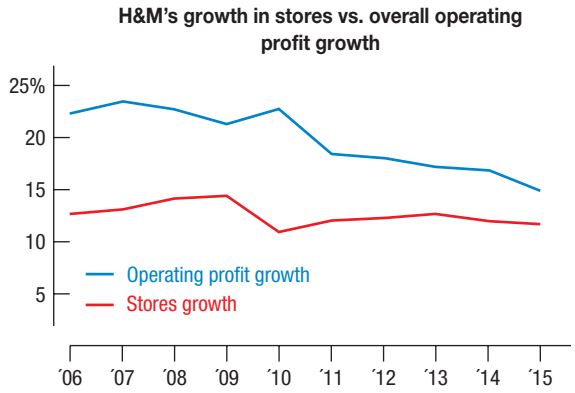


FIGURE 1.1 H&M's store growth has remained steady at just over 10% a year, but operating profit growth seems to be coming down.

Thomasine wonders if the decline she sees in the stores' profit growth (the blue line in Figure 1.1) means she should recommend putting more resources into online sales instead of just building more stores.

Displays like this, called *data visualizations*, can summarize large amounts of data in a concise way that helps make good business decisions, and can often reveal things that weren't expected.

IN PRACTICE 1.1 Business insights from visualizations

One of the authors was consulting for a large multinational firm and was given access to their sales data. Management wondered if there might be sales opportunities around the world and where they might be. Because the company sold many consumer items to individuals, the consultant decided that rather than focus on the total sales

[2] 2016 H&M Group annual report, about.hm.com/en/media/news/financial-reports/2017/1/2441626.html

(in dollars) in each country he should divide the total sales in each country by the population size, creating the new variable *Sales per Capita*. When he displayed this variable on a map, management was shocked:[3]

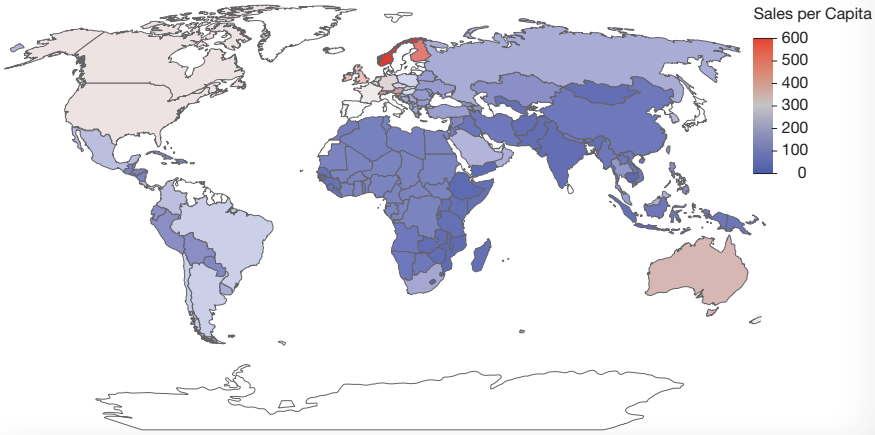
MANAGER We know that we sell more in the United States than anywhere else in the world, but why are some countries redder than the U.S.?

CONSULTANT In this color scheme, low *Sales per Capita* ($ spent per customer) is indicated by dark blue, average by white (grey) and higher than average by red. The countries in the brightest red are the ones with the highest sales per person.

MANAGER You mean we sell more per person in Norway, Finland, and even Australia than in the U.S.?

CONSULTANT Exactly. Norway has the highest sales at more than $600 per person, compared with the U.S. at $364.

MANAGER Wow! I had no idea. I never would have guessed that. Thank you for the insight!



We will be using visualizations, summaries, and models of data to understand, explain, and predict throughout the course. Along the way we will encounter many types of data and corresponding ways to visualize, model, and analyze the data we collect. And because it all starts with data, we'll spend the rest of this chapter getting to know more about the nature of data.

1.1 Data

> "Data is king at Amazon. Clickstream and purchase data are the crown jewels at Amazon. They help us build features to personalize the website experience."
>
> —Ronny Kohavi, former director of data mining and personalization, amazon.com

Every time you make an online purchase, more information is captured than just the details of the purchase itself. What pages did you search to get to your purchase? How much time did you spend looking at each? These recorded values, whether numbers or labels, together with their context are called **data**. They are recorded and stored electronically, in vast digital repositories called **data warehouses**. Businesses have always relied on data to make good decisions, but today, more than ever before, companies use data to make decisions about virtually all aspects of their business, from inventory to advertising to website design.

[3]This is based on a true story. We can't reveal the name of the company due to a non-disclosure agreement.

Every swipe of your credit card and every click of your mouse has helped these data warehouses grow. The challenges of collecting, managing, storing, and curating all of this information collectively fall under the term **Big Data**.

But data alone can't make good decisions. To start the process of turning data into useful information, you first need to know what decisions you want to make. Without a question, you have no idea what might be interesting about the data. Should you look at the time of transactions, their location, their price, which products were bought, or something else? Your knowledge of the business issues and the questions you want to answer will help guide your search for insights from the data, and help you harness data to make better decisions.

Why are you taking this course?
The typical answer is "because it's required." But why is it required? Because these are the tools that will help you leverage your business domain knowledge with data.

Once you have data and a clear vision of the problem, the statistics techniques in this book can empower your decision making. They will help you in two ways: You'll learn how to estimate the likely values needed for your decisions and—possibly more important—you'll learn how to quantify the *uncertainty* of those estimates.

Albert Einstein is credited with saying "If I had one hour to save the world, I'd spend 55 minutes defining the problem and 5 minutes solving it."[4] The wisdom of using your business acumen to define your question will be clear throughout this book.

Before H&M introduces a new product they usually test market it to a small sample of customers and collect data on the product's performance before committing to it worldwide. Statistics helps them make the leap from a sample to an understanding of the world at large. We hope this text will empower you to draw conclusions from data and make valid business decisions in response to such questions as:

- Will the new design of our website increase click-through rates and result in more sales?
- What is the effect of advertising on sales?
- Do aggressive, "high-growth" mutual funds really have higher returns than more conservative funds?
- Is there a seasonal cycle in your firm's profits?
- What is the relationship between shelf location and cereal sales?
- Do students around the world perceive issues in business ethics differently?
- Are there common characteristics about your customers and why they choose your products?—and, more importantly, are those characteristics the same among those who aren't your customers?

Plan (1–2)
1. **Define** the problem.
2. **Collect** and/or find data and **identify** the variables.

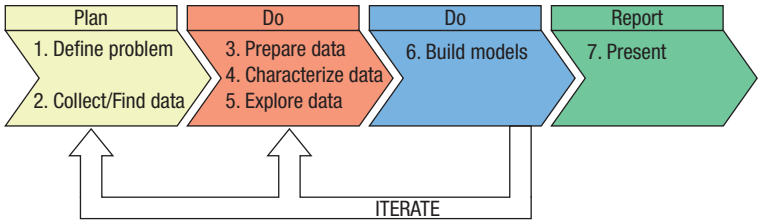
Do (3–6)
3. **Prepare** and wrangle data.
4. **Characterize** the data.
5. **Explore** the data.
   Summarize
   Visualize
6. **Model** (if appropriate).
   Check conditions and assumptions for modeling.
   Fit the model and make the necessary calculations.

Report (7)
7. Communicate and **present**.

Your ability to answer questions such as these and make sound business decisions with data depends largely on your ability to take a business problem, *translate* it into a question that data can answer, and *communicate* that answer to others. The steps to follow are shown in the box in the margin. The **Plan**, **Do**, and **Report** strategy is found throughout the book. The main headings will stay the same although the specific subparts will vary slightly depending on the topic we're learning.

Rarely does the journey from problem definition to solution proceed straight from Step 1 to Step 7. As you learn more about your data you'll probably want to rethink earlier steps, possibly even modifying the original question itself. Or you may decide to collect different data after you see the limitations of your current model. But bearing this process in mind will help you to strategize your data analytics process and keep you on the road toward the goal of delivering good decisions.

[4]According to quoteinvestigator.com there is "no substantive evidence that Einstein ever made a remark of this type." It appeared in a paper by William H. Markle, who credited an unnamed Yale professor. But many people, including those at goodreads.com, still give the credit to Einstein.



1.2 The Role of Data in Decision Making

Q: What is analytics?

A: Analytics is the term for extracting information from data.

Q: Is there really a difference between statistics and analytics?

A: Essentially no. We'll use the terms interchangeably. Some use the term "advanced analytics" to include modern machine learning methods not traditionally found in statistics. (See Chapter 21)

When companies try to obtain actionable information from data that may have been collected in the course of doing business (such as records of transactions or a customer database) it is usually called **data mining**. Sometimes the analysis is called **predictive analytics** if it focuses on future performance. The more general term, **business analytics** (or sometimes simply analytics), refers to any use of data and statistical analysis to inform business decisions. Leading companies are embracing analytics to extract value from their data. As Clive Humby, author of *The Loyalty Myth*, said, "data is the new oil." For example, Zillow recently offered $1,200,000 to improve the prediction of home sale prices from publicly available data. As of June 2017, 791 teams were competing for the prize.

Companies use data to make decisions about nearly every aspect of their business. By studying the past behavior of customers and predicting their responses, they hope to better serve their customers and to compete more effectively.

eBay collected data and used analytics to examine its own use of computer resources. Although not obvious to its own technical people, once they crunched the data they found huge inefficiencies. According to Forbes, eBay was able to "save millions in capital expenditures within the first year."

THE W'S:
WHO
WHAT
WHEN
WHERE
WHY

Data come in many forms. Some are numerical (consisting only of numbers), others are alphabetic (consisting only of letters), and yet others are alphanumerical (mixed numbers and letters). But data are useless unless we know their **context**. Newspaper journalists know that the lead paragraph of a good story should establish the "Five W's": *who*, *what*, *when*, *where*, and (if possible) *why*. Often, we add *how* to the list as well. Answering these questions connects the data to the business problem at hand. The answers to the first two questions are essential. If we don't know *who* and *what*, we don't have any useful information.

We can make the meaning clear if we add the context of *who* the data are about and *what* was measured and organize the values into a **data table**. Table 1.1 shows part of a data table of purchase records from an online music retailer. Each row represents a purchase of a music album. The most general term for a row of a data table is **case** or **record**. Each column of the table records some characteristic

Order Number	Name	State/Country	Price	Area Code	Album Download	Gift?	Stock ID	Artist
105-2686834-3759466	Katherine H.	Ohio	5.99	440	Identity	N	B00000I5Y6	James Fortune & Flya
105-9318443-4200264	Samuel P.	Illinois	9.99	312	Port of Morrow	Y	B000002BK9	The Shins
105-1872500-0198646	Chris G.	Massachusetts	9.99	413	Up All Night	N	B000068ZVQ	Syco Music UK
103-2628345-9238664	Monique D.	Canada	10.99	902	Fallen Empires	N	B000001OAA	Snow Patrol
002-1663369-6638649	Katherine H.	Ohio	11.99	440	Sees the Light	N	B002MXA7Q0	La Sera

TABLE 1.1 Example of a data table. The variable names are in the top row. Typically, the *Who* of the table are found in the leftmost column.

of the cases. The columns are called **variables**. You'll usually find the name of the variable at the top of the column as in Table 1.1.

We call cases by different names, depending on the situation. Individuals who answer a survey are referred to as **respondents**. People on whom we experiment are **subjects** or (in an attempt to acknowledge the importance of their role in the experiment) **participants**, but animals, plants, websites, and other inanimate subjects are often called **experimental units**. Often we call cases just what they are: for example, *customers*, *economic quarters*, or *companies*. When referring to a transaction, rows are often called *records*. In Table 1.1, the rows are the individual orders, or purchase records. A common place to find the *who* of the table is the leftmost column. It's often an identifying variable for the cases, in this example, the order number.

JUST CHECKING

1 What is the "*who*" of Table 1.1? That is, does each row refer to a) a person or b) an order? How can you tell?

Metadata

Metadata became a common term when the National Security Agency (NSA) claimed that they weren't collecting Americans' phone calls but only the information about the phone calls, the phone numbers of the caller and recipient, the time and duration of the call and any bank information used to make the call—in other words—the metadata.

If you collect the data yourself, you'll know what the cases are and how the variables are defined. But, often, you'll be looking at data that someone else collected. The information about the data, called the metadata, might have to come from the company's database administrator or from the information technology department of a company. **Metadata** typically contains information about *how*, *when*, and *where* (and possibly *why*) the data were collected; *who* each case represents; and the definitions of all the variables.

A general term for a data table like the one shown in Table 1.1 is a **spreadsheet**, a name that comes from bookkeeping ledgers of financial information. The data were typically spread across facing pages of a bound ledger, the book used by an accountant for keeping records of expenditures and sources of income. For the accountant, the columns were the types of expenses and income, and the rows were transactions, typically invoices or receipts. These days, it is common to keep modest-size datasets in a spreadsheet even if no accounting is involved. It is usually easy to move a data table from a spreadsheet program to a program designed for statistical graphics and analysis, either directly or by copying the data table and pasting it into the statistics program.

Although data tables and spreadsheets are great for relatively small data sets, they are cumbersome for the complex data sets that companies must maintain on a day-to-day basis. Try to imagine a spreadsheet from a company the size of Amazon with customers in the rows and products in the columns. Amazon has hundreds of millions of customers and millions of products. But very few customers have purchased more than a few dozen items, so almost all the entries in the spreadsheet would be blank—not a very efficient way to store information. For that reason, various other database architectures are used to store data. The most common is a relational database.

In a **relational database**, two or more separate data tables are linked together so that information can be merged across them. Each data table is a *relation* because it is about a specific set of cases with information about each of these cases for all (or at least most) of the variables ("fields" in database terminology). For example, a table of H&M customers, along with demographic information on each, is such a relation. A data table of all the items sold by the company, including information on price, inventory, and past history, is another relation. Transactions may be held in a third "relation" that references each of the other two relations. Table 1.2 shows a small example.

In statistics, analyses are typically performed on a single relation because all variables must refer to the same cases. But often the data must be retrieved from a

relational database. Retrieving data from these databases may require specific expertise with that software. In the rest of the book, we'll assume that the data have been retrieved and placed in a data table or spreadsheet with variables listed as columns and cases as the rows.

Customers

Customer Number	Name	City	State	ZIP Code	Customer since	Gold Member?
473859	R. De Veaux	Williamstown	MA	01267	2007	No
127389	N. Sharpe	New York City	NY	10021	2000	Yes
335682	P. Velleman	Ithaca	NY	14580	2003	No
…						

Items

Product ID	Name	Price	Currently in Stock?
42-8719	Resort Shirt	24.99	Yes
73-2671	Lace Dress	69.99	No
35-0518	Cashmere Sweater	129.00	Yes
72-9665	Leather Derby Shoes	69.00	Yes

Transactions

Transaction Number	Date	Customer Number	Product ID	Quantity	Shipping Method	Free Ship?
T23478923	9/15/17	473859	42-8719	1	UPS 2nd Day	N
T23478924	9/15/17	473859	35-0518	1	UPS 2nd Day	N
T63928934	10/20/17	335682	73-2671	3	UPS Ground	N
T72348299	12/22/17	127389	72-9665	1	Fed Ex Ovnt	Y

TABLE 1.2 A relational database shows all the relevant information for three separate relations linked together by customer and product numbers.

IN PRACTICE 1.2 Gaining insight from data by identifying variables and the W's

Carly is an analyst at a credit card issuer. Her manager wants to know if an offer mailed 3 months ago has affected customers' use of their cards. To answer, Carly asks the IT department to assemble some data on recent customer spending. The IT department sends her a spreadsheet. The first six rows look like this:

Account ID	Pre Spending	Spending	Age	Segment	Enroll?	Offer	Segment Spend
393371	$2,698.12	$6,261.40	25–34	Travel/Ent	NO	None	$887.36
462715	$2,707.92	$3,397.22	45–54	Retail	NO	Gift Card	$5,062.55
433469	$800.51	$4,196.77	65+	Retail	NO	None	$673.80
462716	$3,459.52	$3,335.00	25–34	Services	YES	Double Miles	$800.75
420605	$2,106.48	$5,576.83	35–44	Leisure	YES	Double Miles	$3,064.81
473703	$2,603.92	$7,397.50	<25	Travel/Ent	YES	Double Miles	$491.29

(*continued*)

MANAGER Thanks for the information. I'm not quite sure about the structure. Can you tell me what each row represents and what was measured on each?

ANALYST (CARLY) The cases are individual customers. The data are from our internal records for the past 6 months (3 months before and 3 months after an offer was sent to the customers). The variables include the account ID of the customer (*Account ID*), and the amounts charged on the card before (*Pre Spending*) and after (*Post Spending*) the offer was sent out. We also have the customer's *Age*, marketing *Segment*, whether they enrolled on the website (*Enroll?*), what offer they were sent (*Offer*), and how much they charged on the card in their marketing segment (*Segment Spend*). (The marketing *Segment* classifies cardholders based on their spending patterns.)

1.3 Variable Types

When the values of a variable are simply the names of categories we call it a **categorical**, or **qualitative**, **variable**. When the values of a variable are measured numerical quantities, we call it a **quantitative variable**.

Descriptive responses to questions are often categories. For example, the responses to the questions "What type of mutual fund do you invest in?" or "What kind of advertising does your firm use?" yield categorical values. An important special case of categorical variables is one that has only two possible responses (usually "yes" or "no"), which arise naturally from questions like "Do you invest in the stock market?" or "Do you make online purchases from this website?"

Question	Categories or Responses
Do you invest in the stock market?	__ Yes __ No
What kind of advertising do you use?	__ Newspapers __ Internet __ Direct mailings
What is your class at school?	__ Freshman __ Sophomore __ Junior __ Senior
I would recommend this course to another student.	__ Strongly Disagree __ Slightly Disagree __ Slightly Agree __ Strongly Agree
How satisfied are you with this product?	__ Very Unsatisfied __ Unsatisfied __ Satisfied __ Very Satisfied

TABLE 1.3 Some examples of categorical variables.

Categorical or Quantitative?

Dates can be confusing. Depending on how a date is used, it may be categorical or quantitative. For example, *Day of the Week* has no units, and is categorical. What about a date such as October 31, 2017 (which is a string of characters)? Most software will treat this as categorical. However, many statistics programs can add and subtract dates to determine that there are 60 days between 10/30/17 and 12/30/17, and that 11/30/17 falls exactly in the middle of this date range. Most programs can convert any date into the number of seconds, minutes or hours past a given starting date. If this is the case, then dates may be treated as a quantitative variable, but be sure to specify the units.

Many measurements are quantitative. In a purchase record, price, quantity, and time spent on the website are all quantitative values with **units** (dollars, count, and seconds). For quantitative variables, the units tell how each value has been measured. Even more important, units such as yen, cubits, carats, angstroms, nanoseconds, miles per hour, or degrees Celsius tell us the *scale* of measurement, so we know how far apart two values are. Without units, the values of a measured variable have no clear meaning. It does little good to be promised a raise of 5000 a year if you don't know whether it will be paid in euros, dollars, yen, or Estonian krooni. An essential part of a quantitative variable is its units. Some quantitative variables, however, don't have obvious units. The Dow Jones Industrial "Average" has units (points?) but no one talks about them. Percentages are ratios of two quantities and so the units "cancel out," but they are still percentages of something. So, although it isn't imperative that a quantitative variable have explicit units, when they are not explicit, be careful to think about whether adding their values, averaging them, or otherwise treating them as numerical, makes sense.

The distinction between categorical and quantitative variables seems clear, but there are reasons to be careful. First, some variables can be considered as either categorical or quantitative, depending on the kind of questions we ask about them. For example, the variable *Age* would be considered quantitative if the responses were numerical and they had units. A doctor would certainly consider *Age* to be

quantitative. The units could be years, or for infants, the doctor would want even more precise units, like months, or even days. On the other hand, a retailer might lump together the values into categories like "Child (12 years or less)," "Teen (13 to 19)," "Adult (20 to 64)," or "Senior (65 or over)." For many purposes, like knowing which song download coupon to send you, that might be all the information needed. Then *Age* would be a categorical variable.

How to classify some variables as categorical or quantitative may seem obvious. But be careful. Area codes may look quantitative, but are really categories. What about ZIP codes? They are categories too, but the numbers do contain information. If you look at a map of the United States with ZIP codes, you'll see that as you move West, the first digit of ZIP codes increases, so treating them as quantitative might make sense for some questions. Area codes present a similar set of issues; see the sidebar.

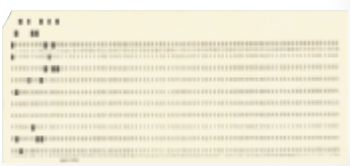
Another reason to be careful about classifying variables comes from the analysis of Big Data. When analysts want to decide what advertisement to send to the web page you're looking at, or what the probability is that you'll renew your phone contract, they use automatic methods involving dozens or even hundreds of variables. Usually the software used to do the analysis has to guess the type of variable from its values. When the variable contains symbols other than numbers, the software will correctly type the variable as categorical, but just because a variable has numbers doesn't mean it is quantitative. We've seen examples (area code, order number) where that's just not the case. Data miners spend much of their time going back through data sets to correctly retype variables as categorical or quantitative to avoid silly mistakes of misuse.

Area Codes?

When area codes were first introduced all phones had dials. To reduce wear and tear on the dials and to speed calls, the lowest-digit codes (the fastest to dial—those for which the dial spun the least) were assigned to the largest cities. So, New York City was given 212, Chicago 312, LA 213, and Philadelphia 215, but rural upstate New York was 607, Joliet was 815, and San Diego 619. Back then, the numerical value of an area code could be used to guess something about the population of its region. But after dials gave way to push buttons, new area codes were assigned without regard to population and area codes are now just categories, with no quantitative information.

Variable Names That Make Sense

A tradition still hangs on in some places to name variables with cryptic abbreviations in uppercase letters. This can be traced back to the 1960s, when computer programs were controlled with instructions punched on cards. The earliest punch card equipment used only uppercase letters, and statistics programs limited variable names to six or eight characters, so variables had names like PRSRF3. Modern programs don't have such restrictive limits, so there is no reason not to use names that make sense.



Identifiers

A special kind of categorical variable is worth mentioning. **Identifier variables** are categorical variables whose only purpose is to assign a unique identifier code to each individual in the data set. Your student ID number, social security number, and phone number are all identifiers.

Identifier variables are crucial in this era of Big Data because, by uniquely identifying the cases, they make it possible to combine data from different sources and provide unique labels. Your school's grade transcripts and your bursar bill records are kept separately, but both refer to you. Your student ID is what links them. Most companies keep such relational databases. The identifier is crucial to linking one data table to another in a relational database. The identifiers in Table 1.2 are the *Customer Number*, *Product ID*, and *Transaction Number*. Variables like *UPS Tracking Number* and *Social Security Number* are other examples of identifiers.

Other Data Types

AirBnB, like many travel sites, uses stars to rate their listings. But are star ratings categorical or quantitative? There is certainly an *order* of perceived worth; more stars indicate higher perceived worth. An AirBnB property whose customer responses average around 4 stars is better than one whose average is around 2, but is it *twice* as good? These values are not quantitative, so we can't really answer that question. When the values of a categorical variable have an intrinsic order, we can say that the variable is **ordinal**. By contrast, a categorical variable with unordered categories is sometimes called **nominal**. Values can be individually ordered (e.g., the ranks of employees based on the number of days they've worked for the company) or ordered in classes (e.g., Freshman, Sophomore, Junior, Senior). Ordering is not absolute; how the values are ordered depends on the purpose of the ordering. For example, are the categories Infant, Youth, Teen, Adult, and Senior ordinal?