Business Statistics

A First Course

8TH EDITION

david LEVINE
kathryn SZABAT
david STEPHAN

A Roadmap for Selecting a Statistical Method

Data Analysis Task	For Numerical Variables	For Categorical Variables
Describing a group or several groups	Ordered array, stem-and-leaf display, frequency distribution, relative frequency distribution, percentage distribution, cumulative percentage distribution, histogram, polygon, cumulative percentage polygon **(Sections 2.2, 2.4)** Mean, median, mode, geometric mean, quartiles, range, interquartile range, standard deviation, variance, coefficient of variation, skewness, kurtosis, boxplot, normal probability plot **(Sections 3.1, 3.2, 3.3, 6.3)** Dashboards **(Section 14.2)**	Summary table, bar chart, pie chart, doughnut chart, Pareto chart **(Sections 2.1 and 2.3)**
Inference about one group	Confidence interval estimate of the mean **(Sections 8.1 and 8.2)** *t* test for the mean **(Section 9.2)**	Confidence interval estimate of the proportion **(Section 8.3)** *Z* test for the proportion **(Section 9.4)**
Comparing two groups	Tests for the difference in the means of two independent populations **(Section 10.1)** Paired *t* test **(Section 10.2)** *F* test for the difference between two variances **(Section 10.4)**	*Z* test for the difference between two proportions **(Section 10.3)** Chi-square test for the difference between two proportions **(Section 12.1)**
Comparing more than two groups	One-way analysis of variance for comparing several means **(Section 11.1)**	Chi-square test for differences among more than two proportions **(Section 12.2)**
Analyzing the relationship between two variables	Scatter plot, time series plot **(Section 2.5)** Covariance, coefficient of correlation **(Section 3.5)** Simple linear regression **(Chapter 13)** *t* test of correlation **(Section 13.7)** Sparklines **(Section 2.7)**	Contingency table, side-by-side bar chart, PivotTables **(Sections 2.1, 2.3, 2.6)** Chi-square test of independence **(Section 12.3)**
Analyzing the relationship between two or more variables	Colored scatter plots, bubble chart, treemap **(Section 2.7)** Multiple regression **(Chapters 14)** Dynamic bubble charts **(Section 14.2)** Regression trees **(Section 14.3)** Cluster analysis **(Section 14.5)** Multidimensional scaling **(Section 14.6)**	Multidimensional contingency tables **(Section 2.6)** Drilldown and slicers **(Section 2.7)** Classification trees **(Section 14.4)** Multiple correspondence analysis **(Section 14.6)**

This page intentionally left blank

Business Statistics
A First Course

This page intentionally left blank

Business Statistics
A First Course

EIGHTH EDITION

David M. Levine

Department of Information Systems and Statistics

Zicklin School of Business, Baruch College, City University of New York

Kathryn A. Szabat

Department of Business Systems and Analytics

School of Business, La Salle University

David F. Stephan

Two Bridges Instructional Technology

Senior VP, Courseware Portfolio Management: Marcia Horton
Director, Portfolio Management: Deirdre Lynch
Courseware Portfolio Manager: Suzanna Bainbridge
Courseware Portfolio Management Assistant: Morgan Danna
Managing Producer: Karen Wernholm
Content Producer: Kathleen A. Manley
Senior Producer: Aimee Thorne
Associate Content Producer: Sneh Singh
Manager, Courseware QA: Mary Durnwald
Manager, Content Development: Robert Carroll
Product Marketing Manager: Kaylee Carlson
Product Marketing Assistant: Shannon McCormack
Field Marketing Manager: Thomas Hayward
Field Marketing Assistant: Derrica Moser
Senior Author Support/Technology Specialist: Joe Vetere
Manager, Rights and Permissions: Gina Cheselka
Manufacturing Buyer: Carol Melville, LSC Communications
Composition and Production Coordination: Pearson Content Services Centre, Julie Kidd
Senior Designer and Cover Design: Barbara T. Atkinson
Cover Image: ArtisticPhoto/Shutterstock

Cataloging-in-Publication Data on file with the Library of Congress

1 19

Pearson

Instructor's Review Copy:
ISBN 13: 978-0-13-524459-3
ISBN 10: 0-13-524459-5
Student Edition:
ISBN 13: 978-0-13-517778-5
ISBN 10: 0-13-517778-2

To our spouses and children,
Marilyn, Mary, Sharyn, and Mark

and to our parents, in loving memory,
Lee, Reuben, Mary, William, Ruth and Francis J.

This page intentionally left blank

About the Authors

Kathryn Szabat, David Levine, and David Stephan

David M. Levine, Kathryn A. Szabat, and David F. Stephan are all experienced business school educators committed to innovation and improving instruction in business statistics and related subjects.

David Levine, Professor Emeritus of Statistics and CIS at Baruch College, CUNY, is a nationally recognized innovator in statistics education for more than three decades. Levine has coauthored 14 books, including several business statistics textbooks; textbooks and professional titles that explain and explore quality management and the Six Sigma approach; and, with David Stephan, a trade paperback that explains statistical concepts to a general audience. Levine has presented or chaired numerous sessions about business education at leading conferences conducted by the Decision Sciences Institute (DSI) and the American Statistical Association, and he and his coauthors have been active participants in the annual DSI Data, Analytics, and Statistics Instruction (DASI) mini-conference. During his many years teaching at Baruch College, Levine was recognized for his contributions to teaching and curriculum development with the College's highest distinguished teaching honor. He earned B.B.A. and M.B.A. degrees from CCNY. and a Ph.D. in industrial engineering and operations research from New York University.

As Associate Professor of Business Systems and Analytics at La Salle University, **Kathryn Szabat** has transformed several business school majors into one interdisciplinary major that better supports careers in new and emerging disciplines of data analysis including analytics. Szabat strives to inspire, stimulate, challenge, and motivate students through innovation and curricular enhancements, and shares her coauthors' commitment to teaching excellence and the continual improvement of statistics presentations. Beyond the classroom she has provided statistical advice to numerous business, nonbusiness, and academic communities, with particular interest in the areas of education, medicine, and nonprofit capacity building. Her research activities have led to journal publications, chapters in scholarly books, and conference presentations. Szabat is a member of the American Statistical Association (ASA), DSI, Institute for Operation Research and Management Sciences (INFORMS), and DSI DASI. She received a B.S. from SUNY-Albany, an M.S. in statistics from the Wharton School of the University of Pennsylvania, and a Ph.D. degree in statistics, with a cognate in operations research, from the Wharton School of the University of Pennsylvania.

Advances in computing have always shaped **David Stephan's** professional life. As an undergraduate, he helped professors use statistics software that was considered advanced even though it could compute *only* several things discussed in Chapter 3, thereby gaining an early appreciation for the benefits of using software to solve problems (and perhaps positively influencing his grades). An early advocate of using computers to support instruction, he developed a prototype of a mainframe-based system that anticipated features found today in Pearson's MathXL and served as special assistant for computing to the Dean and Provost at Baruch College. In his many years teaching at Baruch, Stephan implemented the first computer-based *classroom*, helped redevelop the CIS curriculum, and, as part of a FIPSE project team, designed and implemented a multimedia learning environment. He was also nominated for teaching honors. Stephan has presented at SEDSI and DSI DASI (formerly MSMESB) mini-conferences, sometimes with his coauthors. Stephan earned a B.A. from Franklin & Marshall College and an M.S. from Baruch College, CUNY, and completed the instructional technology graduate program at Teachers College, Columbia University.

For all three coauthors, continuous improvement is a natural outcome of their curiosity about the world. Their varied backgrounds and many years of teaching experience have come together to shape this book in ways discussed in the Preface.

This page intentionally left blank

Brief Contents

This page intentionally left blank

Contents

8 Confidence Interval Estimation 279

9 Fundamentals of Hypothesis Testing: One-Sample Tests 314

This page intentionally left blank

Preface

As business statistics evolves and becomes an increasingly important part of one's business education, how business statistics gets taught and what gets taught becomes all the more important.

We, the authors, think about these issues as we seek ways to continuously improve the teaching of business statistics. We actively participate in Decision Sciences Institute (DSI), American Statistical Association (ASA), and Data, Analytics, and Statistics Instruction and Business (DASI) conferences. We use the ASA's Guidelines for Assessment and Instruction (GAISE) reports and combine them with our experiences teaching business statistics to a diverse student body at several universities.

When writing for introductory business statistics students, five principles guide us.

Help students see the relevance of statistics to their own careers by using examples from the functional areas that may become their areas of specialization. Students need to learn statistics in the context of the functional areas of business. We present each statistics topic in the context of areas such as accounting, finance, management, and marketing and explain the application of specific methods to business activities.

Emphasize interpretation and analysis of statistical results over calculation. We emphasize the interpretation of results, the evaluation of the assumptions, and the discussion of what should be done if the assumptions are violated. We believe that these activities are more important to students' futures and will serve them better than focusing on tedious manual calculations.

Give students ample practice in understanding how to apply statistics to business. We believe that both classroom examples and homework exercises should involve actual or realistic data, using small and large sets of data, to the extent possible.

Familiarize students with **the use of data analysis software.** We integrate using Microsoft Excel, JMP, and Minitab into all statistics topics to illustrate how software can assist the business decision making process. In this edition, we also integrate using Tableau into selected topics, where such integration makes best sense. (Using software in this way also supports our second point about emphasizing interpretation over calculation).

Provide clear instructions to students that facilitate their use of data analysis software. We believe that providing such instructions assists learning and minimizes the chance that the software will distract from the learning of statistical concepts.

What's New in This Edition?

This eighth edition of *Business Statistics*: *A First Course* features many passages rewritten in a more concise style that emphasize definitions as the foundation for understanding statistical concepts. In addition to changes that readers of past editions have come to expect, such as new examples and Using Statistics case scenarios and an extensive number of new end-of-section or end-of-chapter problems, the edition debuts:

- **A First Things First Chapter** that builds on the previous edition's novel Important Things to Learn First Chapter by using real-world examples to illustrate how developments such as the increasing use of business analytics and "big data" have made knowing

and understanding statistics that much more critical. This chapter is available as complimentary online download, allowing students to get a head start on learning.

- **Tabular Summaries** that state hypothesis test and regression example results along with the conclusions that those results support now appear in Chapters 10 through 13.
- Updated Excel and Minitab Guides that reflect the most recent editions of these programs.
- **New JMP Guides** that provide detailed, hands-on instructions for using JMP to illustrate the concepts that this book teaches. JMP provides a starting point for continuing studies in business statistics and business analytics and features visualizations that are easy to construct and that summarize data in innovative ways.
- For selected chapters, **Tableau Guides** that make best use of this software for basic and advanced visualizations and regression analysis.
- **An All-New Business Analytics Chapter (Chapter 14)** that makes extensive use of JMP, Minitab, and Tableau to illustrate predictive analytics for prediction, classification, clustering, and association as well as explaining what text analytics does and how descriptive and prescriptive analytics relate to predictive analytics. This chapter benefits from the insights the coauthors have gained from teaching and lecturing on business analytics as well as research the coauthors have done for a forthcoming companion title on business analytics.

Continuing Features that Readers Have Come to Expect

This edition of *Business Statistics*: *A First Course* continues to incorporate a number of distinctive features that has led to its wide adoption over the previous editions. Table 1 summaries these carry-over features:

TABLE 1
Distinctive Features Continued in the Eighth Edition

Feature	Details
Using Statistics Business Scenarios	A Using Statistics scenario that highlights how statistics is used in a business functional area begins each chapter. Each scenario provides an applied context for learning in its chapter. End-of-chapter "Revisited" sections reinforces the statistical methods that a chapter discusses and apply those methods to the questions raised in the scenario. *In this edition, four chapters have new or revised Using Statistics scenarios*.
Emphasis on Data Analysis and Interpretation of Results	*Basic Business Statistics* was among the first business statistics textbooks to focus on interpretation of the results of a statistical method and not on the mathematics of a method. This tradition continues, now supplemented by JMP results complimenting the Excel and Minitab results of recent prior editions.
Software Integration	Software instructions in this book feature chapter examples and were personally written by the authors, who collectively have over one hundred years experience teaching the application of software to business. Software usage also features templates and applications developed by the authors that minimize the frustration of using software while maximizing statistical learning
Opportunities for Additional Learning	Student Tips, LearnMore bubbles, and Consider This features extend student-paced learning by reinforcing important points or examining side issues or answering questions that arise while studying business statistics such as "What is so 'normal' about the normal distribution?"
Highly Tailorable Context	With an extensive library of separate online topics, sections, and even two full chapters, instructors can combine these materials and the opportunities for additional learning to meet their curricular needs.
Software Flexibility	With modularized software instructions, instructors and students can switch among Excel, Excel with PHStat, JMP, Minitab, and Tableau as they use this book, taking advantage of the strengths of each program to enhance learning.

TABLE 1 Distinctive Features Continued in the Eighth Edition (*continued*)

Feature	Details
End-of-Section and End-of-Chapter Reinforcements	"Exhibits" summarize key processes throughout the book. "Key Terms" provides an index to the definitions of the important vocabulary of a chapter. "Learning the Basics" questions test the basic concepts of a chapter. "Applying the Concepts" problems test the learner's ability to apply those problems to business problems. For the more quantitatively-minded, "Key Equations" list the boxed number equations that appear in a chapter.
Innovative Cases	End-of-chapter cases include a case that continues through many chapters as well as "Digital Cases" that require students to examine business documents and other information sources to sift through various claims and discover the data most relevant to a business case problem as well as common misuses of statistical information. (Instructional tips for these cases and solutions to the Digital Cases are included in the Instructor's Solutions Manual.)
Answers to Even-Numbered Problems	An appendix provides additional self-study opportunities by provides answers to the "Self-Test" problems and most of the even-numbered problems in this book.
Unique Excel Integration	Many textbooks feature Microsoft Excel, but *Business Statistics*: *A First Course* comes from the authors who originated both the Excel Guide workbooks that illustrate model solutions, developed Visual Explorations that demonstrate selected basic concepts, and designed and implemented PHStat, the Pearson statistical add-in for Excel that places the focus on statistical learning. (See Appendix H for a complete summary of PHStat.)

Chapter-by-Chapter Changes Made for This Edition

Because the authors believe in continuous quality improvement, *every* chapter of *Business Statistics*: *A First Course* contains changes to enhance, update, or just freshen this book. Table 2 provides a chapter-by-chapter summary of these changes.

TABLE 2
Chapter-by-Chapter Change Matrix

Chapter	Using Statistics Changed	JMP/ Tableau Guide	Problems Changed	Selected Chapter Changes
FTF	•	J, T	n.a.	Think Differently About Statistics Starting Point for Learning Statistics
1	•	J, T	40%	Data Cleaning Other Data Preprocessing Tasks
2		J, T	60%	Organizing a Mix of Variables Visualizing A Mix of Variables Filtering and Querying Data Reorganized categorical variables discussion. Expanded data visualization discussion. New samples of 379 retirement funds and 100 restaurant meal costs for examples.
3		J, T	50%	New samples of 379 retirement funds and 100 restaurant meal costs for examples. Updated NBA team values data set.

TABLE 2
Chapter-by-Chapter Change Matrix (*continued*)

Chapter	Using Statistics Changed	JMP/ Tableau Guide	Problems Changed	Selected Chapter Changes
4		J	43%	Basic Probability Concepts rewritten. Bayes' theorem example moved online.
5		J	60%	Section 5.1 and Binomial Distribution revised.
6	•	J	33%	Normal Distribution rewritten.
7		J	47%	Sampling Distribution of the Proportion rewritten.
8		J	40%	Confidence Interval Estimate for the Mean revised. Revised "Managing Ashland Multi-Comm Services" continuing case.
9		J	20%	Chapter introduction revised. Section 9.1 rewritten. New Section 9.4 example.
10	•	J	43%	New paired *t* test and the difference between two proportions examples.
11		J	43%	Extensive use of new tabular summaries. Revised "Managing Ashland Multi-Comm Services" continuing case.
12		J, T	46%	Chapter introduction revised. Section 12.2 revised.
13		J	30%	Section 13.1 revised. Section 13.3 reorganized and revised. New dummy variable example.
14	n.a.	J, T	n.a.	All-new chapter that introduces business analytics. Software Guide explains using Excel with Power BI Desktop, JMP, Minitab, and Tableau, for various descriptive and predictive analytics methods.

Serious About Writing Improvements

Ever review a textbook that reads the same as an edition from years ago? Or read a preface that claims writing improvements but offers no evidence? Among the writing improvements in this edition of *Business Statistics: A First Course,* the authors have turned to tabular summaries to guide readers to reaching conclusions and making decisions based on statistical information. The authors believe that this writing improvement, which appears in Chapters 9 through 13, not only adds clarity to the purpose of the statistical method being discussed but better illustrates the role of statistics in business decision-making processes. Judge for yourself using the sample from Chapter 10 Example 10.1.

Previously, part of the solution to Example 10.1 was presented as:
You do not reject the null hypothesis because $t_{STAT} = -1.6341 > -1.7341$. The *p*-value (as computed in Figure 10.5) is 0.0598. This *p*-value indicates that the probability that $t_{STAT} < -1.6341$ is equal to 0.0598. In other words, if the population means are equal, the probability that the sample mean delivery time for the local pizza restaurant is at least

2.18 minutes faster than the national chain is 0.0598. Because the *p*-value is greater than a = 0.05, there is insufficient evidence to reject the null hypothesis. Based on these results, there is insufficient evidence for the local pizza restaurant to make the advertising claim that it has a faster delivery time.

In this edition, we present the equivalent solution (on page 360):
Table 10.4 summarizes the results of the pooled-variance *t* test for the pizza delivery data using the calculation above (*not shown in this sample*) and Figure 10.5 results. Based on the conclusions, local branch of the national chain and a local pizza restaurant have similar delivery times. Therefore, as part of the last step of the DCOVA framework, you and your friends exclude delivery time as a decision criteria when choosing from which store to order pizza.

TABLE 10.4
Pooled-variance t test summary for the delivery times for the two pizza restaurants

Result	Conclusions
The $t_{STAT} = -1.6341$ is greater than -1.7341. The *t* test *p*-value = 0.0598 is greater than the level of significance, $\alpha = 0.05$.	**1.** Do not reject the null hypothesis H_0. **2.** Conclude that insufficient evidence exists that the mean delivery time is lower for the local restaurant than for the branch of the national chain. **3.** There is a probability of 0.0598 that $t_{STAT} < -1.6341$.

A Note of Thanks

Creating a new edition of a textbook is a team effort, and we thank our Pearson Education editorial, marketing, and production teammates: Suzanna Bainbridge, Kathy Manley, Kaylee Carlson, Thomas Hayward, Deirdre Lynch, Aimee Thorne, and Morgan Danna. And we would be remiss not to note the continuing work of Joe Vetere to prepare our screen shot illustrations and the efforts of Julie Kidd of Pearson CSC to ensure that this edition meets the highest standard of book production quality that is possible.

We also thank Gail Illich of McLennan Community College for preparing instructor resources for this edition and thank the following people whose comments helped us improve this edition: Mohammad Ahmadi, University of Tennessee-Chattanooga; Sung Ahn, Washington State University; Kelly Alvey, Old Dominion University; Al Batten, University of Colorado-Colorado Springs; Alan Chesen, Wright State University; Gail Hafer, St. Louis Community College-Meramec; Chun Jin, Central Connecticut State University; Benjamin Lev, Drexel University; Lilian Prince, Kent State University; Bharatendra Rai, University of Massachusetts Dartmouth; Ahmad Vakil, St. John's University (NYC); and Shiro Withanachchi, Queens College (CUNY).

We thank the RAND Corporation and the American Society for Testing and Materials for their kind permission to publish various tables in Appendix E, and to the American Statistical Association for its permission to publish diagrams from the *American Statistician*. Finally, we would like to thank our families for their patience, understanding, love, and assistance in making this book a reality.

Contact Us!

Please email us at **authors@davidlevinestatistics.com** or tweet us **@BusStatBooks** with your questions about the contents of this book. Please include the hashtag #BSAFC8 in your tweet or in the subject line of your email. We also welcome suggestions you may have for a future edition of this book. And while we have strived to make this book as error-free as possible, we also appreciate those who share with us any perceived problems or errors that they encounter.

If you need assistance using software, please contact your academic support person or Pearson Support at **support.pearson.com/getsupport/.** They have the resources to resolve and walk you through a solution to many technical issues in a way we do not.

As you use this book, be sure to make use of the "Resources for Success" that Pearson Education supplies for this book (described on the following pages). We also invite you to visit **bsafc8.davidlevinestatistics.com (bit.ly/2Apx1xH),** where we may post additional information or new content as necessary.

David M. Levine
Kathryn A. Szabat
David F. Stephan

Resources for Success

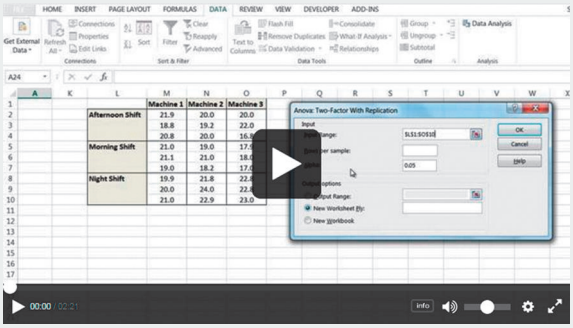
MyLab Statistics Online Course for *Business Statistics: A First Course*, 8th Edition by Levine, Szabat, Stephan

(access code required)

MyLab Statistics is the teaching and learning platform that empowers instructors 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.

MyLab makes learning and using a variety of statistical programs as seamless and intuitive as possible. Download the data files that this book uses (see Appendix C) in Excel, JMP, and Minitab formats. Download supplemental files that support in-book cases or extend learning.

- Download the Excel Data Workbooks that contain the data used in chapter examples or named in problems and end-of-chapter cases.
- Download the Excel Guide Workbooks that contain the model templates and solutions for statistical methods discussed in the textbook.
- Download the JMP Data Tables and Projects that contain the data used in chapter examples or named in problems and end-of-chapter cases.
- Download the Minitab Worksheets and Projects that contain the data used in chapter examples or named in problems and end-of-chapter cases.
- Download the PHStat readme pdf that explains the technical requirements and getting started instructions for using this Microsoft Excel add-in. To download PHStat, visit the PHStat download page. (Download requires an access code as explained on that page.)
- Download the Visual Explorations Workbooks that interactively demonstrate various key statistical concepts.

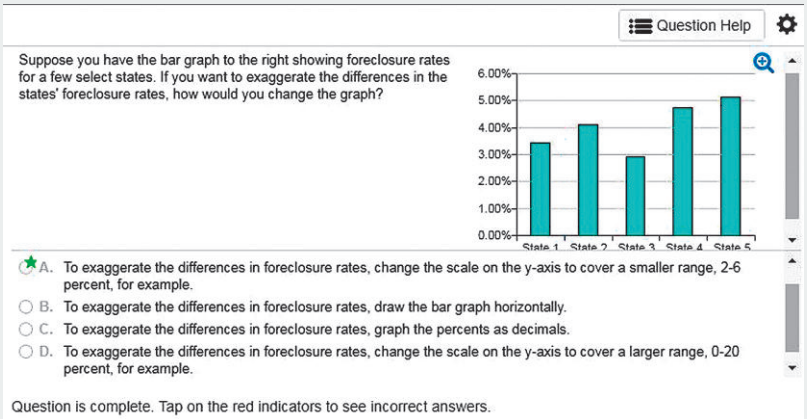


Instructional Videos

Access instructional support videos including Pearson's Business Insight and StatTalk videos, available with assessment questions. Reference technology study cards and instructional videos for Excel, JMP, Minitab, StatCrunch, and R software.

Diverse Question Libraries

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.



Resources for Success

Instructor Resources

Instructor's Solutions Manual, presents solutions for end-of-section and end-of-chapter problems and answers to case questions, and provides teaching tips for each chapter. The Instructor's Solutions Manual is available for download at **www.Pearson.com** or in MyLab Statistics.

Lecture PowerPoint Presentations, by Patrick Schur, Miami University (Ohio), are available for each chapter. These presentations provide instructors with individual lecture notes to accompany the text. The slides include many of the figures and tables from the textbook. Instructors can use these lecture notes as is or customize them in Microsoft PowerPoint. The PowerPoint presentations are available for download at **www.Pearson.com** or in MyLab Statistics.

Test Bank, contains true/false, multiple-choice, fill-in, and problem-solving questions based on the definitions, concepts, and ideas developed in each chapter of the text. The Test Bank is available for download at **www.Pearson.com** or in MyLab Statistics.

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 Education's online catalog.

Student Resources

Student's Solutions Manual, provides detailed solutions to virtually all the even-numbered exercises and worked-out solutions to the self-test problems.
(ISBN-10: 0-13-518243-3;
ISBN-13: 978-0-13-518243-7)

Online resources complement and extend the study of business statistics and support the content of this book. These resources include data files for in-chapter examples and problems, **templates and model solutions**, and **optional topics and chapters**. (See Appendix C for a complete description of the online resources.)

PHStat helps create Excel worksheet solutions to statistical problems. PHStat uses Excel building blocks to create worksheet solutions. These worksheet solutions illustrate Excel techniques and students can examine them to gain new Excel skills. Additionally, many solutions are what-if templates in which the effects of changing data on the results can be explored. Such templates are fully reusable on any computer on which Excel has been installed. PHStat requires an access code and separate download for use. PHStat access codes can be bundled with this textbook using
ISBN-10: 0-13-399058-3;
ISBN-13: 978-0-13-399058-4.

Minitab® More than 4,000 colleges and universities worldwide use Minitab software to help students learn quickly and to provide them with a skill-set that's in demand in today's data-driven workforce. Minitab® includes a comprehensive collection of statistical tools to teach beginning through advanced courses. Bundling Minitab software ensures students have the software they need for the duration of their course work.
(ISBN-10: 0-13-445640-8;
ISBN-13: 978-0-13-445640-9)

JMP® Student Edition software is statistical discovery software from SAS Institute Inc., the leader in business analytics software and services. JMP® Student Edition is a streamlined version of JMP that provides all the statistics and graphics covered in introductory and intermediate statistics courses. Available for bundling with this textbook.
(ISBN-10: 0-13-467979-2;
ISBN-13: 978-0-13-467979-2)

First Things First

USING **STATISTICS**
"The Price of Admission"

It's the year 1900 and you are a promoter of theatrical productions, in the business of selling seats for individual performances. Using your knowledge and experience, you establish a selling price for the performances, a price you hope represents a good trade-off between maximizing revenues and avoiding driving away demand for your seats. You print up tickets and flyers, place advertisements in local media, and see what happens. After the event, you review your results and consider if you made a wise trade-off.

Tickets sold very quickly? Next time perhaps you can charge more. The event failed to sell out? Perhaps next time you could charge less or take out more advertisements to drive demand. If you lived over 100 years ago, that's about all you could do.

Jump ahead about 70 years. You're still a promoter but now using a computer system that allows your customers to buy tickets over the phone. You can get summary reports of advance sales for future events and adjust your advertising on radio and on TV and, perhaps, add or subtract performance dates using the information in those reports.

Jump ahead to today. You're still a promoter but you now have a fully computerized sales system that allows you to constantly adjust the price of tickets. You also can manage many more categories of tickets than just the near-stage and far-stage categories you might have used many years ago. You no longer have to wait until after an event to make decisions about changing your sales program. Through your sales system you have gained insights about your customers such as where they live, what other tickets they buy, and their appropriate demographic traits. Because you know more about your customers, you can make your advertising and publicity more efficient by aiming your messages at the types of people more likely to buy your tickets. By using social media networks and other online media, you can also learn almost immediately who is noticing and responding to your advertising messages. You might even run experiments online presenting your advertising in two different ways and seeing which way sells better.

Your current self has capabilities that allow you to be a more effective promoter than any older version of yourself. Just how much better? Turn the page.

CONTENTS

OBJECTIVES

- Statistics is a way of thinking that can lead to better decision making
- Statistics requires analytical skills and is an important part of your business education
- Recent developments such as the use of business analytics and "big data" have made knowing statistics even more critical
- The DCOVA framework guides your application of statistics
- The opportunity business analytics represents for business students

Now Appearing on Broadway ... *and* Everywhere Else

In early 2014, Disney Theatrical Productions woke up the rest of Broadway when reports revealed that its *17*-year-old production of *The Lion King* had been the top-grossing Broadway show in 2013. How could such a long-running show, whose most expensive ticket was less than half the most expensive ticket on Broadway, earn so much while being so old? Over time, grosses for a show decline, and weekly grosses for *The Lion King* had dropped about 25% by the year 2009. But, in 2013, grosses were up 67% from 2009, and weekly grosses typically exceeded the grosses of opening weeks in 1997, adjusted for inflation!

Heavier advertising and some changes in ticket pricing helped, but the major reason for this change was something else: combining business acumen with the systematic application of *business statistics and analytics* to the problem of selling tickets. As a producer of the newest musical at the time said, "We make educated predictions on price. Disney, on the other hand, has turned this into a science" (see reference 3).

Disney had followed the plan of action that this book presents. It had collected its daily and weekly results and summarized them, using techniques this book introduces in the next three chapters. Disney then analyzed those results by performing experiments and tests on the data collected (using techniques that later chapters introduce). In turn, those analyses were applied to a new interactive seating map that allowed customers to buy tickets for specific seats and permitted Disney to adjust the pricing of each seat for each performance. The whole system was constantly reviewed and refined, using the semiautomated methods to which Chapter 14 will introduce you. The end result was a system that outperformed the ticket-selling methods others used.

student TIP

From other business courses, you may recognize that Disney's system uses dynamic pricing.

FTF.1 Think Differently About Statistics

The "Using Statistics" scenario suggests, and the Disney example illustrates, that modern-day information technology has allowed businesses to apply statistics in ways that could not be done years ago. This scenario and example reflect how this book teaches you about statistics. In these first two pages, you may notice

- the lack of calculation details and "math."
- the emphasis on enhancing business methods and management decision making.
- that none of this seems like the content of a middle school or high school statistics class you may have taken.

You may have had some prior knowledge or instruction in *mathematical statistics*. This book discusses *business statistics*. While the boundary between the two can be blurry, business statistics emphasizes business problem solving and shows a preference for using software to perform calculations.

One similarity that you might notice between these first two pages and any prior instruction is *data*. **Data** are the facts about the world that one seeks to study and explore. Some data are unsummarized, such as the facts about a single ticket-selling transaction, whereas other facts, such as weekly ticket grosses, are **summarized**, derived from a set of unsummarized data. While you may think of data as being numbers, such as the cost of a ticket or the percentage that weekly grosses have increased in a year, do not overlook that data can be non-numerical as well, such as ticket-buyer's name, seat location, or method of payment.

Statistics: A Way of Thinking

Statistics are the methods that allow you to work with data effectively. Business statistics focuses on interpreting the results of applying those methods. You interpret those results to help you enhance business processes and make better decisions. Specifically, business statistics provides you with a formal basis to summarize and visualize business data, reach conclusions about that data, make reliable predictions about business activities, and improve business processes.

You must apply this way of thinking correctly. Any "bad" things you may have heard about statistics, including the famous quote "there are lies, damned lies, and statistics" made famous by Mark Twain, speak to the errors that people make when either misusing statistical methods or mistaking statistics as a substitution for, and not an enhancement of, a decision-making process. (Disney Theatrical Productions' success was based on *combining* statistics with business acumen, not *replacing* that acumen.)

DCOVA Framework To minimize errors, you use a framework that organizes the set of tasks that you follow to apply statistics properly. The five tasks that comprise the **DCOVA framework** are

- **D**efine the data that you want to study to solve a problem or meet an objective.
- **C**ollect the data from appropriate sources.
- **O**rganize the data collected, by developing tables.
- **V**isualize the data collected, by developing charts.
- **A**nalyze the data collected, reach conclusions, and present the results.

You must always do the **D**efine and **C**ollect tasks before doing the other three. The order of the other three varies, and sometimes all three are done concurrently. In this book, you will learn more about the **D**efine and **C**ollect tasks in Chapter 1 and then be introduced to the **O**rganize and **V**isualize tasks in Chapter 2. Beginning with Chapter 3, you will learn methods that help complete the **A**nalyze task. Throughout this book, you will see specific examples that apply the DCOVA framework to specific business problems and examples.

Analytical Skills More Important Than Arithmetic Skills The business preference for using software to automate statistical calculations maximizes the importance of having analytical skills while it minimizes the need for arithmetic skills. With software, you perform calculations faster and more accurately than if you did those calculations by hand, minimizing the need for advanced arithmetic skills. However, with software you can *also* generate inappropriate or meaningless results if you have not fully understood a business problem or goal under study or if you use that software without a proper understanding of statistics.

Therefore, using software to create results that help solve business problems or meet business goals is *always* intertwined with using a framework. And using software does not mean memorizing long lists of software commands or how-to operations, but knowing how to review, modify, and possibly create software solutions. If you can analyze what you need to do and have a general sense of what you need, you can always find instructions or illustrative sample solutions to guide you. (This book provides detailed instructions *as well as* sample solutions for every statistical activity discussed in end-of-chapter software guides and through the use of various downloadable files and sample solutions.)

If you were introduced to using software in an application development setting or an introductory information systems class, do not mistake building applications from scratch as being a necessary skill. A "smart" smartphone user knows how to use apps such as Facebook, Instagram, YouTube, Google Maps, and Gmail effectively to communicate or discover and use information and has no idea how to construct a social media network, create a mapping system, or write an email program. Your approach to using the software in this book should be the same as that smart user. Use your analytical skills to focus on being an effective user and to understand *conceptually* what a statistical method or the software that implements that method does.

Statistics: An Important Part of Your Business Education

Until you read these pages, you may have seen a course in business statistics solely as a required course with little relevance to your overall business education. In just two pages, you have learned that statistics is a way of thinking that can help enhance your effectiveness in business—that is, applying statistics correctly is a fundamental, global skill in your business education.

In the current data-driven environment of business, you need the general analytical skills that allow you to work with data and interpret analytical results regardless of the discipline in which you work. No longer is statistics only for accounting, economics, finance, or other disciplines that directly work with numerical data. As the Disney example illustrates, the decisions you make will be increasingly based on data and not on your gut or intuition supported by past experience. Having a well-balanced mix of statistics, modeling, and basic technical skills as well as managerial skills, such as business acumen and problem-solving and communication skills, will best prepare you for the workplace today … *and* tomorrow (see reference 1).

FTF.2 Business Analytics: The Changing Face of Statistics

Of the recent changes that have made statistics an important part of your business education, the emergence of the set of methods collectively known as business analytics may be the most significant change of all. **Business analytics** combine traditional statistical methods with methods from management science and information systems to form an interdisciplinary tool that supports fact-based decision making. Business analytics include

- statistical methods to analyze and explore data that can uncover previously unknown or unforeseen relationships.
- information systems methods to collect and process data sets of all sizes, including very large data sets that would otherwise be hard to use efficiently.
- management science methods to develop optimization models that support all levels of management, from strategic planning to daily operations.

In the Disney Theatrical Productions example, statistical methods helped determine pricing factors, information systems methods made the interactive seating map and pricing analysis possible, and management science methods helped adjust pricing rules to match Disney's goal of sustaining ticket sales into the future. Other businesses use analytics to send custom mailings to their customers, and businesses such as the travel review site tripadvisor.com use analytics to help optimally price advertising as well as generate information that makes a persuasive case for using that advertising.

Generally, studies have shown that businesses that actively use business analytics and combine that use with data-guided management see increases in productivity, innovation, and competition (see reference 1). Chapter 14 introduces you to the statistical methods typically used in business analytics and shows how these methods are related to statistical methods that the book discusses in earlier chapters.

"Big Data"

Big data is a collection of data that cannot be easily browsed or analyzed using traditional methods. Big data implies data that are being collected in huge volumes, at very fast rates or velocities (typically in near real time), and in a variety of forms that can differ from the structured forms such as records stored in files or rows of data stored in worksheets that businesses use every day. These attributes of volume, velocity, and variety (see reference 5) distinguish big data from a "big" (large) set of data that contains numerous records or rows of similar data. When combined with business analytics and the basic statistical methods discussed in this book, big data presents opportunities to gain new management insights and extract value from the data resources of a business (see reference 8).

Unstructured Data Big data may also include **unstructured data**, data that has an irregular pattern and contain values that are not comprehensible without additional automated or manual interpretation. Unstructured data takes many forms such as unstructured text, pictures, videos, and audio tracks, with unstructured text, such as social media comments,

getting the most immediate attention today for its possible use in customer, branding, or marketing analyses. Unstructured data can be adapted for use with a number of methods, such as regression, which this book illustrates with conventional, structured files and worksheets. Unstructured data may require performing data collection and preparation tasks beyond those tasks that Chapter 1 discusses. While describing all such tasks is beyond the scope of this book, Section 14.1 includes an example of the additional interpretation that is necessary when working with unstructured text.

FTF.3 Starting Point for Learning Statistics

Statistics has its own vocabulary and learning the precise meanings, or **operational definitions**, of basic terms provides the basis for understanding the statistical methods that this book discusses. For example, *in statistics*, a **variable** defines a characteristic, or property, of an item or individual that can vary among the occurrences of those items or individuals. For example, for the item "book," variables would include the title and number of chapters, as these facts can vary from book to book. For a given book, these variables have a specific value. For *this* book, the value of the title variable would be "Business Statistics: A First Course," and "15" would be the value for the number of chapters variable. Note that a statistical variable is not an algebraic variable, which serves as a stand-in to represent one value in an algebraic statement and could never take a non-numerical value such as "Business Statistics: A First Course."

Using the definition of variable, data, in its statistical sense, can be defined as the set of values associated with one or more variables. In statistics, each value for a specific variable is a single fact, not a list of facts. For example, what would be the value of the variable author for this book? Without this rule, you might say that the single list "Levine, Szabat, Stephan" is the value. However, applying this rule, one would say that the variable has three separate values: "Levine", "Stephan", and "Szabat". This distinction of using only *single-value data* has the practical benefit of simplifying the task of entering data for software analysis.

Using the definitions of data and variable, the definition of statistics can be restated as the methods that analyze the data of the variables of interest. The methods that primarily help summarize and present data comprise **descriptive statistics**. Methods that use data collected from a small group to reach conclusions about a larger group comprise **inferential statistics**. Chapters 2 and 3 introduce descriptive methods, many of which are applied to support the inferential methods that the rest of the book presents.

Statistic

The previous section uses *statistics* in the sense of a collective noun, a noun that is the name for a collection of things (methods in this case). The word statistics also serves as the plural form of the noun statistic, as in "one uses methods of descriptive statistics (collective noun) to generate descriptive statistics (plural of the singular noun)." In this sense, a **statistic** refers to a value that summarizes the data of a particular variable. (More about this in coming chapters.) In the Disney Theatrical Productions example, the statement "for 2013, weekly grosses were up 67% from 2009" cites a statistic that summarizes the variable weekly grosses using the 2013 data—all 52 values.

When someone warns you of a possible unfortunate outcome by saying, "Don't be a statistic!" you can always reply, "I can't be." *You* always represent one value and a *statistic* always summarizes multiple values. For the statistic "87% of our employees suffer a workplace accident," you, as an employee, will either have suffered or have not suffered a workplace accident. The "have" or "have not" value contributes to the statistic but cannot be the statistic. A statistic can facilitate preliminary decision making. For example, would you immediately accept a position at a company if you learned that 87% of their employees suffered a workplace accident? (Sounds like this might be a dangerous place to work and that further investigation is necessary.)

Can Statistics (*pl.*, statistic) Lie?

The famous quote "lies, damned lies, and statistics" actually refers to the plural form of *statistic* and does not refer to statistics, the field of study. Can any statistic "lie"? No, faulty or invalid statistics can only be produced through willful misuse of statistics or when DCOVA framework tasks are done incorrectly. For example, many statistical methods are valid only if the data being analyzed have certain properties. To the extent possible, you test the assertion that the data have those properties, which in statistics are called *assumptions*. When an assumption is *violated*, shown to be invalid for the data being analyzed, the methods that require that assumption should not be used.

For the inferential methods that this book discusses in later chapters, you must always look for logical causality. **Logical causality** means that you can plausibly claim something directly causes something else. For example, you wear black shoes today and note that the weather is sunny. The next day, you again wear black shoes and notice that the weather continues to be sunny. The third day, you change to brown shoes and note that the weather is rainy. The fourth day, you wear black shoes again and the weather is again sunny. These four days seem to suggest a strong pattern between your shoe color choice and the type of weather you experience. You begin to think if you wear brown shoes on the fifth day, the weather will be rainy. Then you realize that your shoes cannot plausibly influence weather patterns, that your shoe color choice cannot *logically cause* the weather. What you are seeing is mere coincidence. (On the fifth day, you do wear brown shoes and it happens to rain, but that is just another coincidence.)

You can easily spot the lack of logical causality when trying to correlate shoe color choice with the weather, but in other situations the lack of logical causality may not be so easily seen. Therefore, relying on such correlations by themselves is a fundamental misuse of statistics. When you look for patterns in the data being analyzed, you must *always* be thinking of logical causes. Otherwise, you are misrepresenting your results. Such misrepresentations sometimes cause people to wrongly conclude that all statistics are "lies." Statistics (*pl.*, statistic) are not lies or "damned lies." They play a significant role in *statistics*, the way of thinking that can enhance your decision making and increase your effectiveness in business.

FTF.4 Starting Point for Using Software

This book uses Microsoft Excel, JMP, Minitab, and Tableau to help explain and illustrate statistical concepts and methods as well as demonstrate how such applications can help facilitate business decision making. To begin using the software that this book uses requires only the knowledge of basic user interface skills, operations, and vocabulary that Table FTF.1 summarizes and which the supplemental, online **Basic Computing Skills** document reviews. (Learn more about online supplemental files in Appendix C.)

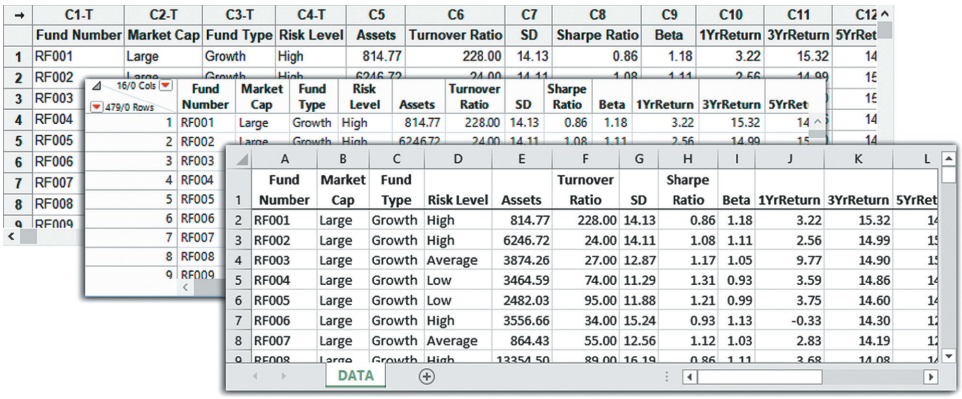
TABLE FTF.1
Basic Computing Knowledge

Skill or Operation	Specifics
Identify and use standard window objects	Title bar, minimize/resize/close buttons, scroll bars, mouse pointer, menu bars or ribbons, dialog box, window subdivisions such as areas, panes, or child windows
Identify and use common dialog box items	Command button, list box, drop-down list, edit box, option button, check box, tabs (tabbed panels)
Mouse operations	Click, called select in some list or menu contexts and check or clear in some check box contexts; double-click; right-click; drag and drag-and-drop

Excel, JMP, and Minitab all use **worksheets** to display the contents of a data set and as the means to enter or edit data. (JMP calls its worksheets **data tables**.) Worksheets are containers that present tabular arrangements of data, in which the intersections of rows and columns form **cells**, boxes into which individual entries are made. One places the data for a variable into the

cells of a column such that each column contains the data for a different variable, if more than one variable is under study. By convention, one uses the cell in the initial row to enter names of the variables (variable columns).

Shown below, from back to front, are a Minitab worksheet, a JMP data table, and an Excel worksheet. The JMP and Minitab containers contain a special unnumbered row into which column variable names can be entered. In Excel, variable names are entered in row 1 of the worksheet, which can sometimes lead to inadvertent errors.



	Fund Number	Market Cap	Fund Type	Risk Level	Assets	Turnover Ratio	SD	Sharpe Ratio	Beta	1YrReturn	3YrReturn	5YrRet
1	RF001	Large	Growth	High	814.77	228.00	14.13	0.86	1.18	3.22	15.32	14
2	RF002	Large	Growth	High	6246.72	24.00	14.11	1.08	1.11	2.56	14.99	15
3	RF003	Large	Growth	Average	3874.26	27.00	12.87	1.17	1.05	9.77	14.90	15
4	RF004	Large	Growth	Low	3464.59	74.00	11.29	1.31	0.93	3.59	14.86	14
5	RF005	Large	Growth	Low	2482.03	95.00	11.88	1.21	0.99	3.75	14.60	14
6	RF006	Large	Growth	High	3556.66	34.00	15.24	0.93	1.13	-0.33	14.30	12
7	RF007	Large	Growth	Average	864.43	55.00	12.56	1.12	1.03	2.83	14.19	12

DATA

studentTIP

Selected Excel, JMP, and Minitab solutions that this book presents exist as templates that simplify the production of results and serve as models for learning more about using formulas.

Generally, entries in each cell are single data values that can be text or numbers. All three programs also permit **formulas**, instructions to process data, to compute cell values. Formulas can include **functions** that simplify certain arithmetic tasks or provide access to advanced processing or statistical features. Formulas play an important role in designing **templates**, *reusable* solutions that have been previously audited and verified. However, JMP and Minitab allow only *column* formulas that define calculations for all the cells in a column, whereas Excel allows only *cell* formulas that define calculations for individual cells.

These three programs save worksheet data and results as one file, called a **workbook** in Excel and a **project** in JMP and Minitab. JMP and Minitab also allow the saving of individual worksheets or results as separate files, whereas Excel always saves a workbook even if the workbook contains (only) one worksheet. Both JMP and Minitab can open the data worksheets of an Excel workbook, making the Excel workbook a universal format for sharing files that contain only data, such as the set of data files for use with this book that Appendix C documents.

Appendix Section G.5 explains the limitations on using Tableau workbooks that Tableau Public users face.

Tableau Differences Like Excel, Tableau uses workbooks to store one or more worksheets, but Tableau defines the concept of worksheet differently. A Tableau worksheet stores tabular and visual summaries that are associated with a separately defined *data source* that can be a complex collection of data or be equivalent to an Excel data worksheet (as are the data sources that this book uses for examples). Data sources can be viewed and column formulas can be used to define new columns, but individual values cannot be edited. Data sources can be unique to a specific Tableau worksheet or shared by several Tableau worksheets. Tableau workbooks can also store dashboards, a concept that Chapter 14 discusses. Table FTF.2 summarizes some of the various file formats that the four programs use. (Appendix D discusses macro and add-in files.)

TABLE FTF.2 Common File Formats for Excel, JMP, Minitab, and Tableau

File Type	Excel	JMP	Minitab	Tableau
All-in-one-file	**.xlsx** (workbook)	**.jmpprj** (project)	**.mpj** (project)	**.twbx** (workbook)
Single worksheet	**.xlsx**	**.jmp**	**.mtw**	
Results only	n.a.	**.jrp** (report)	**.mgf** (graph)	
Macro or add-in	**.xlsm**, **.xlam**	**.jsl**, **.jmpaddin**	**.mtb**, **.mac**	

student TIP

Check the student download web page for this book for more information about PHStat and JMP and Minitab macros and add-ins that may be available for download.

Using Software Properly

Learning to use software *properly* can be hard as software has limited ways to provide feedback for user actions that are invalid operations. In addition, no software will ever know if you are following proper procedures for using that software. The principles that Exhibit FTF.1 lists will assist you and should govern your use of software with this book. These principles will minimize your chance of making errors and lessen the frustration that often occurs when these principles are unknown or overlooked by a user.

EXHIBIT FTF.1

Principles of Using Software Properly

Ensure that software is properly updated. Users who manage their own computers often overlook the importance of ensuring that all installed software is up to date.

Understand the basic operational tasks. Take the time to master the tasks of starting the application, loading and entering data, and how to select or choose commands in a general way.

Understand the statistical concepts that an application uses. Not understanding those concepts can lead to making wrong choices in the application and can make interpreting results difficult.

Know how to review software use for errors. Review and verify that the proper data preparation procedures (see Chapter 1) have been applied to the data before analysis. Verify that the correct procedures, commands, and software options have been selected. For information entered for labeling purposes, verify that no typographical errors exist.

Seek reuse of preexisting solutions to solve new problems. Build solutions from scratch only as necessary, particularly if using Excel in which errors can be most easily made. Some solutions, and almost all Excel solutions that this book presents, exist as models or templates that can *and should* be reused because such reuse models best practice.

Understand how to organize and present information from the results that the software produces. Think about the best ways to arrange and label the data. Consider ways to enhance or reorganize results that will facilitate communication with others.

Use self-identifying names, especially for the files that you create and save. Naming files Document 1, Document 2, and so on will impede the later retrieval and use of those files.

In addition, also look for ways in which you can simplify the user interface of the software you use. If using Excel with this book, consider using PHStat, supplied separately or as part of a bundle by Pearson. PHStat simplifies the user interface by providing a consistent dialog box driven interface that minimizes keystrokes and mouse selections. If using JMP and Minitab, look for macros and add-ins that simplify command sequences or automate repetitive activities.

Software-related Conventions Table FTF.3 on page 9 summarizes the software-related conventions that this book uses. These conventions are used extensively in the end-of-chapter software guides and certain appendices to provide a concise and clear way of expressing specific user activities.

TABLE FTF.3 Conventions That This Book Uses

Convention	Example
Special key names appear capitalized and in boldface	Press **Enter**. Press **Command** or **Ctrl**.
Key combinations appear in boldface, with key names linked using this symbol: **+**	Enter the formula and press **Ctrl+Enter**. Press **Ctrl+C**.
Menu or Ribbon selections appear in boldface and sequences of consecutive selections are shown using this symbol: **➔**	Select **File➔New**. Select **PHStat➔Descriptive Statistics➔Boxplot**.
Target of mouse operations appear in boldface	Click **OK**. Select **Attendance** and then click the **Y button**.
Entries and the location of where entries are made appear in boldface	Enter **450** in cell **B5**. Add **Temperature** to the **Model Effects** list.
Variable or column names sometimes appear capitalized for emphasis	This file contains the Fund Type, Assets, and Expense Ratio for the growth funds.
Placeholders that express a general case appear in italics and may also appear in boldface as part of a function definition	**AVERAGE (*cell range of variable*)** Replace *cell range of variable* with the cell range that contains the Asset variable.
Names of data files mentioned in sections or problems appear in a special font but appear in boldface in end-of-chapter Guide instructions	Retirement Funds Open the **Retirement Funds workbook**.
When current versions of Excel and Minitab differ in their user interface, alternate instructions for older versions appear in a second color immediately following the primary instructions	In the Select Data Source display, click the icon inside the **Horizontal (Category) axis labels** box. Click **Edit** under the **Horizontal (Categories) Axis Labels** heading.

REFERENCES

1. Advani, D. "Preparing Students for the Jobs of the Future." *University Business* (2011), **bit.ly/1gNLTJm**.
2. Davenport, T., J. Harris, and R. Morison. *Analytics at Work*. Boston: Harvard Business School Press, 2010.
3. Healy, P. "Ticker Pricing Puts 'Lion King' atop Broadway's Circle of Life." *New York Times, New York edition*, March 17, 2014, p. A1, and **nyti.ms.1zDkzki**.
4. JP Morgan Chase. "Report of JPMorgan Chase & Co. Management Task Force Regarding 2012 CIO Losses," **bit.ly/1BnQZzY**, as quoted in J. Ewok, "The Importance of Excel," *The Baseline Scenario*, **bit.ly/1LPeQUy**.
5. Laney, D. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Stamford, CT: META Group. February 6, 2001.
6. Levine, D., and D. Stephan. "Teaching Introductory Business Statistics Using the DCOVA Framework." *Decision Sciences Journal of Innovative Education* 9 (Sept. 2011): 393–398.
7. Marr, B. "20 Claims About Big Data and Why They All Are Wrong." **Data-informed.com**, September 28, 2015.
8. "What Is Big Data?" IBM Corporation, **www.ibm.com/big-data/us/en/**.

KEY TERMS

big data 4
cells 6
data 2
data table 6
business analytics 4
DCOVA framework 3
descriptive statistics 5
formula 7
function 7
inferential statistics 5
logical causality 6
operational definition 5
project (JMP, Minitab) 7
statistic 5
statistics 2
summarized data 2
template 7
unstructured data 4
variable 5
workbook 7
worksheet 6

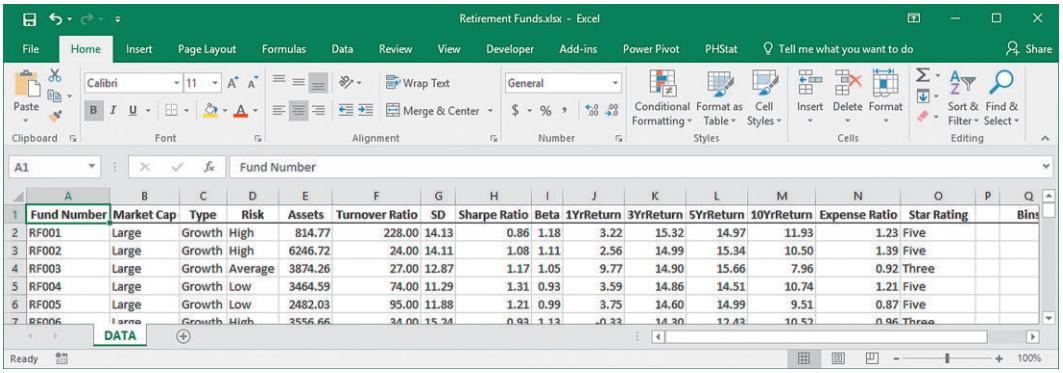
EG.1 GETTING STARTED with EXCEL

student TIP

Excel sometimes displays a task pane in the worksheet area that presents formatting and similar choices.

Opening Excel displays a window that contains the Office Ribbon tabs above a worksheet area that displays the current worksheet of the current workbook, the name of which appears centered in the title bar. The top of the worksheet area contains a formula bar that allows you to see and edit the contents of the currently selected cell (cell A1 in the illustration). Immediately below the worksheet grid is a sheet tab that identifies the name of current worksheet (DATA). In workbooks with more than one sheet, clicking the sheet tabs navigates through the workbook. In the illustration, the current cell is cell A1, and its content is displayed in the formula bar.

The **Basic Computing Skills** *online document (see Appendix C) discusses the other standard features Microsoft Office seen in the illustration.*



	Fund Number	Market Cap	Type	Risk	Assets	Turnover Ratio	SD	Sharpe Ratio	Beta	1YrReturn	3YrReturn	5YrReturn	10YrReturn	Expense Ratio	Star Rating
2	RF001	Large	Growth	High	814.77	228.00	14.13	0.86	1.18	3.22	15.32	14.97	11.93	1.23	Five
3	RF002	Large	Growth	High	6246.72	24.00	14.11	1.08	1.11	2.56	14.99	15.34	10.50	1.39	Five
4	RF003	Large	Growth	Average	3874.26	27.00	12.87	1.17	1.05	9.77	14.90	15.66	7.96	0.92	Three
5	RF004	Large	Growth	Low	3464.59	74.00	11.29	1.31	0.93	3.59	14.86	14.51	10.74	1.21	Five
6	RF005	Large	Growth	Low	2482.03	95.00	11.88	1.21	0.99	3.75	14.60	14.99	9.51	0.87	Five

The illustration shows the Retirement Funds workbook, one of many data workbooks that support this textbook. Use the Excel data workbooks with either the Excel Guide workbooks or PHStat to create solutions to problems or to recreate results used in examples. See Appendix C for description of these learning resources. (Using PHStat requires a separate download and an access code, which may have been bundled with the purchase of this book, as Appendix H explains.)

EG.2 ENTERING DATA

In Excel, enter data into worksheet columns, starting with the leftmost, first column, using the cells in row 1 to enter variable names. Avoid skipping rows or columns as such skipping can disrupt or alter the way certain Excel procedures work. Complete a cell entry by pressing **Tab** or **Enter**, or, if using the formula bar to make a cell entry, by clicking the **check mark icon** in the formula bar. To enter or edit data in a specific cell, either use the cursor keys to move the cell pointer to the cell or select the cell directly.

Try to avoid using numbers as row 1 variable headings; if you cannot avoid their use, precede such headings with apostrophes. Pay attention to special instructions in this book that note specific orderings of variable columns that are necessary for some Excel operations. When in doubt, use the DATA worksheets of the Excel Guide Workbooks as the guide for entering and arranging variable data.

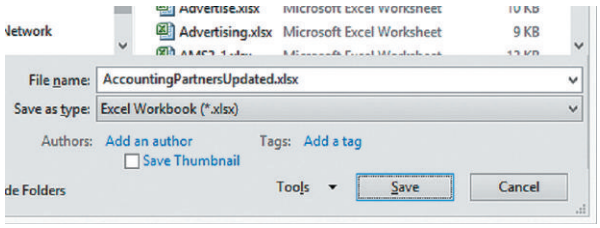
EG.3 OPEN or SAVE a WORKBOOK

Use **File → Open** or **File → Save As**.

Open and **Save As** use similar means to open or save the workbook by name while specifying the physical device or network location and folder for that workbook. Save As dialog boxes enable one to save a file in alternate formats for programs that cannot open Excel workbooks (**.xlsx** files) directly. Alternate formats include a simple text file with values delimited with tab characters, **Text (Tab delimited) (*.txt)** that saves the contents of the current worksheet as a simple text file, and **CSV (Comma delimited) (*.csv)** that saves worksheet cell values as text values that are delimited with commas. Excels for Mac list these choices as **Tab Delimited Text (.txt)** and **Windows Comma Separated (.csv)**.

The illustration on the next page shows the part of the Save As dialog box that contains the **Save as type** drop-down list.

(Open dialog boxes have a similar drop-down list.) In all Windows Excel versions, you can also select a file format in the Open dialog box. Selecting **All Files (*.*)** from the drop-down list can list files that had been previously saved in unexpected formats.



To open a new workbook, select **File → New** (**New Workbook** in Excel for Mac). Excel displays a new workbook with one or more blank worksheets.

EG.4 WORKING with a WORKBOOK

Use **Insert** (or **Insert Sheet**), **Delete**, or **Move or Copy**.

Alter the contents of a workbook by adding a worksheet or by deleting, copying, or rearranging the worksheets and chart sheets that the workbook contains. To perform one of these operations, right-click a sheet tab and select the appropriate choice from the shortcut menu that appears.

To add a worksheet, select **Insert**. In Microsoft Windows Excel, you also click **Worksheet** and then click **OK** in the Insert dialog box. To delete a worksheet or chart sheet, right-click the sheet tab of the worksheet to be deleted and select **Delete**. To copy or rearrange the position of a worksheet or chart sheet, right-click the sheet tab of the sheet and select **Move or Copy**. In the Move or Copy dialog box, first select the workbook and the position in the workbook for the sheet. If copying a sheet, also check **Create a copy**. Then click **OK**.

EG.5 PRINT a WORKSHEET

Use **File → Print**.

Excel prints worksheets and chart sheets, not workbooks. When you select **Print**, Excel displays a preview of the currently opened sheet in a dialog box or pane that allows you to select that sheet or other sheets from the workbook. You can adjust the print formatting of the worksheet(s) to be printed by clicking **Page Setup**. Typically, in the Page Setup dialog box, you might click the **Sheet** tab and then check or clear the **Gridlines** and **Row and column headings** checkboxes to add or remove worksheet cell gridlines and the numbered row and lettered column headings that are similar to how a worksheet is displayed onscreen.

EG.6 REVIEWING WORKSHEETS

Follow the best practice of reviewing worksheets before you use them to help solve problems. When you use a worksheet, what you see displayed in cells may be the result of either the recalculation of formulas or cell formatting. A cell that displays 4 might contain the value 4, might contain a formula calculation that results in the value 4, or might contain a value such as 3.987 that has been formatted to display as the nearest whole number.

To display and review all formulas, press **Ctrl+`** (grave accent). Excel displays the *formula view* of the worksheet, revealing all formulas. (Pressing **Ctrl+`** a second time restores the worksheet to its normal display.) If you use the Excel Guide workbooks, you will discover that each workbook contains one or more FORMULAS worksheets that provide a second way of viewing all formulas.

In the Excel solutions for this book, you will notice cell formatting operations that have changed the background color of cells, changed text attributes such as boldface of cell entries, and rounded values to a certain number of decimal places (typically four). However, if you want to learn more about cell formatting, Appendix B includes a summary of common formatting operations, including those used in the Excel solutions for this book.

EG.7 IF YOU USE the *WORKBOOK* INSTRUCTIONS

Excel Guide *Workbook* instructions enable you to directly modify the template and model worksheet solutions for problems other than the one they help solve. (In contrast, PHStat provides a dialog box interface in which you make entries that PHStat uses to automate such modifications.) *Workbook* instructions express Excel operations in the most universal way possible. For example, many instructions ask you to select (click on) an item from a gallery of items and identify that item selection by name. In some Excel versions, these names may be visible captions for the item; in other versions, you will need to move the mouse over the image to pop up the image name.

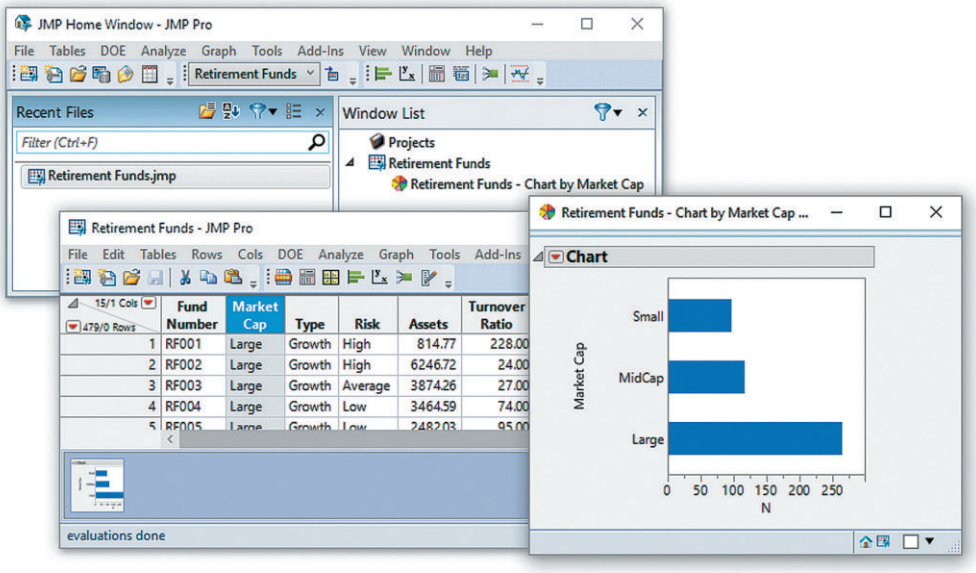
Guides also use the word *display* as in the "Format Axis display" to refer to a user interaction that may be presented by Excel in a **task pane** or a **two-panel dialog box** ("Format Axis task pane" or "Format Axis dialog box"). Task panes open to the side of the worksheet and can remain onscreen indefinitely. Initially, some parts of a pane may be hidden and you may need to click on an icon or label to reveal that hidden part to complete a *Workbook* instruction. Two-panel dialog boxes open over the worksheet and must be *closed* before you can do other Excel activities. The left panel of such dialog boxes are always visible and clicking entries in the left panel makes visible one of the right panels. (Click the system close button at the top right of a task pane or dialog box to close the display and remove it from the screen.)

Current Excel versions can vary in their menu sequences. Excel Guide instructions show these variations as parenthetical phrases. For example, the menu sequence, "select **Design** (or **Chart Design**) → **Add Chart Element**" tells you to first select **Design** *or* **Chart Design** to begin the sequence and then to continue by selecting **Add Chart Element**. (Microsoft Windows Excel uses **Design** and Excel for Mac uses **Chart Design**.)

For the current Excel versions that this book supports (see the FAQs in Appendix G), the *Workbook* Instructions are generally identical. Occasionally, individual instructions may differ significantly for one (or more) versions. In such cases, the instructions that apply for multiple versions appear first, in normal text, and the instructions for the unique version immediately follows in this text color.

JG.1 GETTING STARTED with JMP

Opening JMP displays the JMP Home Window (shown below) that contains the main menu bar and toolbar through which you make JMP command selections, as well as lists of recent files and any other JMP windows that JMP has been set previously to open. In the illustration below, JMP has opened the Retirement Fund data table window and the Retirement Fund - Chart by Market Cap window and displays those two items in the Window List.



Windows that JMP opens or creates display independently of other windows and can be arranged to overlap, as the illustration shows. Note that JMP displays thumbnails of results windows associated with a data table in an evaluations done panel that appears below the data table. In the Windows List, associated results appear as indented list items under the name of the data table window.

In many windows that JMP creates, JMP hides a copy of the home window's menu bar and tool bar under a "thin blue bar" as shown above. Clicking the **thin blue bar**, seen in the Retirement Fund - Chart by Market Cap window, displays a copy of the home window's user interface. Most results windows also contain a right downward-pointing triangle to the left of a result heading (Chart in the illustration). Clicking this red triangle displays a **red triangle menu** of commands and options appropriate for the results that appear under the heading. Red triangle menus also appear in other contexts, such as in the upper left corner of data tables where they hide various row and column selection, data entry, and formatting commands.

A result heading also includes a gray right triangle **disclosure button** that hides or reveals results (to the left of the red triangle in the Chart heading). By using the disclosure button and a combination of red triangle menu selections you can tailor the results, what JMP calls a report, to your specific needs.

Selecting **Help ➔ Books** from the main window's menu bar displays a list of books in PDF format that you can display in JMP or save and read when not using JMP. Consult the books *Discovering JMP* and *Using JMP* as additional sources for getting started with JMP or to discover the JMP features and commands that the instructions in this book do not use.

JG.2 ENTERING DATA

In JMP, enter data into data table (worksheet) columns, starting with the first numbered row and the leftmost, first column. Never skip a cell when entering data because JMP will interpret that skipped cell as a "missing value" (see Section 1.4) that can affect analysis. Complete a cell entry by pressing **Enter**. To enter or edit data in a specific cell, either use the cursor keys to move the cell pointer to the cell or select the cell directly.

As one enters data into columns, JMP assigns default column names in the form *Column 1*, *Column 2*, and so forth. Change these default names to variable names by double-clicking the name or right-clicking and selecting **Column Info** from the shortcut menu. Either action displays the Column dialog box in which you can enter the variable name and set data type and scale, attributes of the data that Chapter 1 explains.

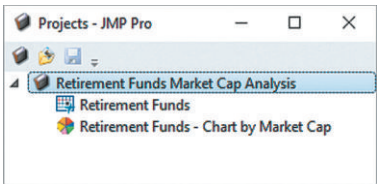
JG.3 CREATE NEW PROJECT or DATA TABLE

Use **File → New → Project.**

Use **File → New → Data Table.**

While report windows that contain results can be opened and saved separately from the data table that provides the data for those results (see Section JG.4), best practice groups a set of report window and data table files into one project file. To create a project file, select **File → New → Project**. JMP opens a new Projects window with the name Untitled. Right-click "Untitled" and rename the project. In the illustration below, the project has been renamed Retirement Funds Market Cap Analysis.

To add a window to a project, right-click the project name and select **Add Window**. In the dialog box that appears, select the window to be included and click **OK**. In lieu of selecting **Add Window**, select **Add All Windows** to add all onscreen JMP windows excluding the home window. In the illustration below, the Retirement Funds data table and Chart by Market Cap windows has been added to the renamed project. Project files can be opened and saved as Section JG.4 explains.



The data table New command opens a blank data table in its own window. Any new data table is not automatically added to the currently open project, and you must use the Add Window command if you want a new data table to be part of a project.

JG.4 OPEN or SAVE FILES

Use **File → Open**.

Use **File → Save As**.

In JMP, all displayed windows can be opened or saved as separate files, as well as open and save special grouping files such as projects. By default, JMP lists all JMP file types in open operations and properly assigns the file type in all save operations. To import an Excel workbook, select **Excel Files (*.xls, *.xlsx, *.xlsm)** from the pull-down list in the Open Data File dialog box. To export a JMP data table as an Excel file, change the **Save as type** in the Save JMP File As dialog box to **Excel Workbook (*.xlsx, *.xls)**.

Report windows can be saved as "interactive HTML" files that allow others to use systems on which JMP has not been installed to explore results in an interactive way, using a subset of JMP functionality. To save this type of file, change the **Save as type** in the Save JMP File As dialog box to **Interactive HTML with Data (*.htm;*.html)**.

JG.5 PRINT DATA TABLES or REPORT WINDOWS

Use **File → Print** or **File → Print Preview**.

To print a JMP object, select these File commands from the window that contains the object you want to print. For results (report) windows, first click the thin blue bar to reveal the menu bar that contains File. If using Print Preview, JMP opens a new window in which preview output and printing options can be adjusted before printing. [To print, click the leftmost (Print) icon in the window.]

JG.6 JMP SCRIPT FILES

JMP script files record many user interface actions and construct or modify JMP objects such as data tables. Using its own JSL scripting language, JMP records your actions as you analyze data in a script file that you can optionally save and play back later to recreate the analysis. Saved script files are text files that can be viewed, edited, and run in their own JMP window or edited by word or texting processing applications.

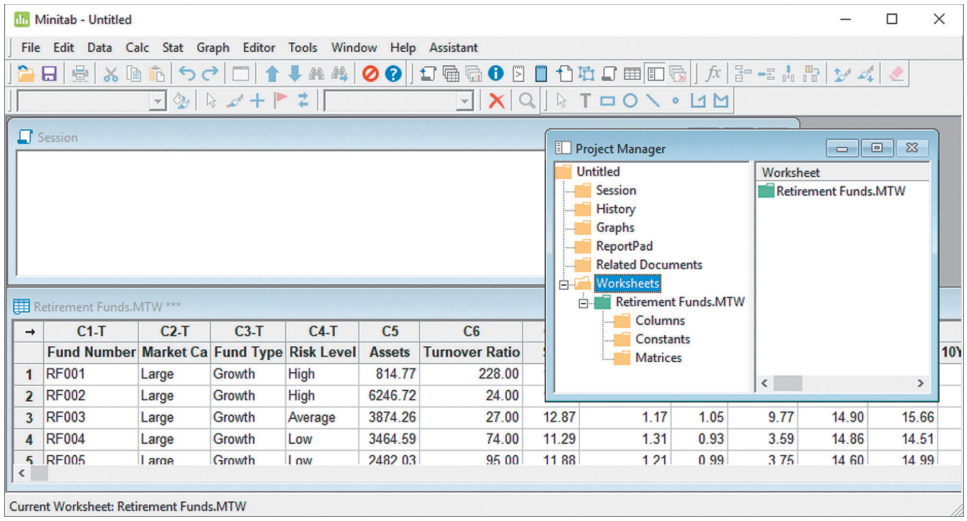
JSL also includes user interface commands and directives allowing one to construct scripts that simplify and customize the use of the JMP Home window menu bar and toolbar. For selected chapters, JMP scripts created especially for this book can facilitate your use of JMP for those chapters (see Appendix C). JMP scripts are sometimes packaged as a *JMP add-in* that can be "installed" in JMP and directly selected from the JMP Home window menu bar, eliminating the need to open a script and then run the script from inside the script window.

MG.1 GETTING STARTED with MINITAB

student TIP

To view a window that may be obscured or hidden, select **Window** from the Minitab menu bar, and then select the name of the window you want to view.

When you open Minitab, you see a main window and a number of child windows that cannot be moved outside the boundaries of the main window. You will normally see a blank worksheet and the Session window that records commands and displays results as the child windows. Pictured below is a project with one opened worksheet. Besides the slightly obscured DATA worksheet window and **Session** window, this figure also shows a **Project Manager** that lists the contents of the current project. (Use the keyboard shortcut **Ctrl+I** to display the Project Manager if it is not otherwise visible in the main window.)



→	C1-T	C2-T	C3-T	C4-T	C5	C6
	Fund Number	Market Ca	Fund Type	Risk Level	Assets	Turnover Ratio
1	RF001	Large	Growth	High	814.77	228.00
2	RF002	Large	Growth	High	6246.72	24.00
3	RF003	Large	Growth	Average	3874.26	27.00
4	RF004	Large	Growth	Low	3464.59	74.00
5	RF005	Large	Growth	Low	2482.03	95.00

12.87	1.17	1.05	9.77	14.90	15.66	
11.29	1.31	0.93	3.59	14.86	14.51	
11.88	1.21	0.99	3.75	14.60	14.99	

MG.2 ENTERING DATA

In Minitab, enter data into worksheet columns, starting with the first numbered row and leftmost, first column. Minitab names columns using the form C*n*, such that the first column is named C1, the second column is C2, and the tenth column is C10. Use the first, unnumbered and shaded row to enter variable names that can be used as a second way to refer to a column by name. If a variable name contains spaces or other special characters, such as **Market Cap**, Minitab will display that name in dialog boxes using a pair of single quotation marks (**'Market Cap'**). You must include those quotation marks any time you enter such a variable name in a dialog box.

If a column contains non-numerical data, Minitab displays the column name with an appended **-T** such as C1-T, C2-T, and C3-T in the worksheet shown above. If a column contains data that Minitab interprets as either dates or times, Minitab displays the column name with an appended **-D**. If a column contains data that a column formula (see Chapter 1) computes, Minitab displays a small green check mark above and to the right of the C*n* name. (Neither the appended -D nor the check mark are shown in the worksheet above.)

To enter or edit data in a specific cell, either use the cursor keys to move the cell pointer to the cell or use your mouse to select the cell directly. Never skip a cell in a numbered row when entering data because Minitab will interpret a skipped cells as a "missing value" (see Section 1.4).

MG.3 OPEN or SAVE FILES

Use **File → Open Worksheet** or **File → Open Project** and **File → Save Current Worksheet** or **File → Save Project As**.

In Minitab, open and save individual worksheets or entire projects, collections of worksheets, Session results, and graphs. To save data in a form readable by Excel, select **Excel** from the **Save as type drop-down list** before clicking Save. Data can also be saved as a simple text file, **Text**, or as simple text with values delimited with commas, **CSV**.

In Minitab, individual graphs and a project's session window can be opened or saved, although these operations are never used in the Minitab Guides in this book.

MG.4 INSERT or COPY WORKSHEETS

Use **File → New** or **File → Open Worksheet**.

To insert a new worksheet, select **File → New** and in the New dialog box click **Minitab Worksheet** and then click **OK**. To insert a copy of a worksheet, select **File → Open Worksheet** and select worksheet to be copied.

MG.5 PRINT WORKSHEETS

Use **File → Print Worksheet** (or **Print Graph** or **Print Session Window**).

Selecting Print Worksheet displays the Data Window Print Options dialog box. In this dialog box, specify the title and formatting options for printing. Selecting Print Graph or Print Session Window displays a dialog box that allows you to change the default printer settings.

If you need to change printing attributes, first select **File → Print Setup** and make the appropriate selections in the Print dialog box before you select the Print command.

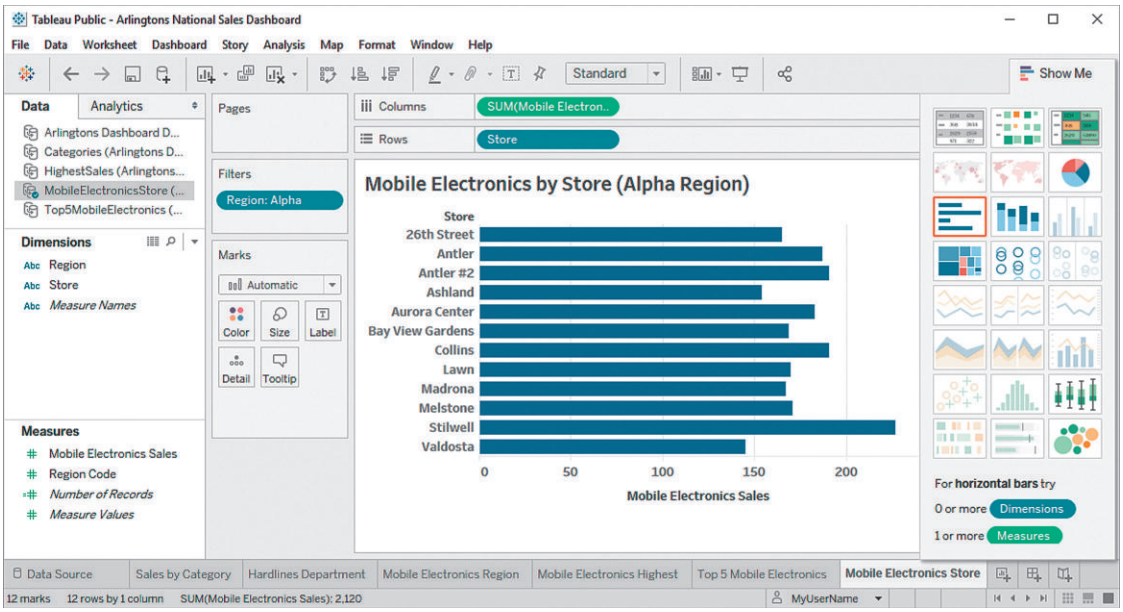
TABLEAU GUIDE

First Things First

TG.1 GETTING STARTED with TABLEAU

The Tableau Guides for this book feature Tableau Public, version 2018, also known as the Tableau Desktop Public Edition. Tableau Public allows users to share data and create, in worksheets, tabular and visual summaries that can be building blocks for a dashboard or a story, a slide-based presentation. Tableau Public uses a drag-and-drop interface that will be most familiar to users of the JMP Graph Builder or the Excel PivotTable feature. Tableau Public consists of three different editions, as Appendix Section G.5 explains. Except where noted in this book, instructions apply to all three versions.

Tableau Public uses workbooks to store tabular and visual worksheet summaries, dashboards, and stories. Opening Tableau Public displays the main Tableau window in which the contents of a workbook can be viewed and edited. The main window shown below displays the Mobile Electronics Store worksheet of the Arlingtons National Sales Dashboard Tableau workbook that is used in a Chapter 14 example.



The main Tableau window contains a menu bar and toolbar, a tabbed left area that presents data and formatting details, several special areas that Section TG.4 explains, the worksheet display area, and the Show Me gallery that displays tabular and visual summaries appropriate for the data in the worksheet area. Worksheet tabs appear under these areas, along with a tab for the current data source to the left of the tabs and icon shortcuts for a new worksheet, new dashboard,

and new story, respectively, to the right of the tabs. The bottom of the window displays status information at the left, the current signed-in user ("MyUserName"), a drop-down button to sign out of Tableau Public, and four media controls that permit browsing through the worksheet tabs of the current workbook.

The Data tab shown on the previous page displays the workbook data sources (five) and lists the "Dimensions" and "Measures," the variable columns of the data source for the worksheet, summary measures, and calculated values for the MobileElectronicsStore data source associated with the displayed worksheet. (Section TG1.1 explains more about the significance of dimensions and measures.)

For the worksheet shown on the previous page, the variable Store was dragged-and-dropped in the Rows *shelf* and the variable Mobile Electronics Sales was dragged-and-dropped in the Column *shelf*. The worksheet also contains a *filter* for the Region column that selects only those rows in the source worksheet in which the value in the Region column is Alpha. (To create the bar chart, **horizontal bars icon** was selected from the Show Me gallery.)

TG.2 ENTERING DATA

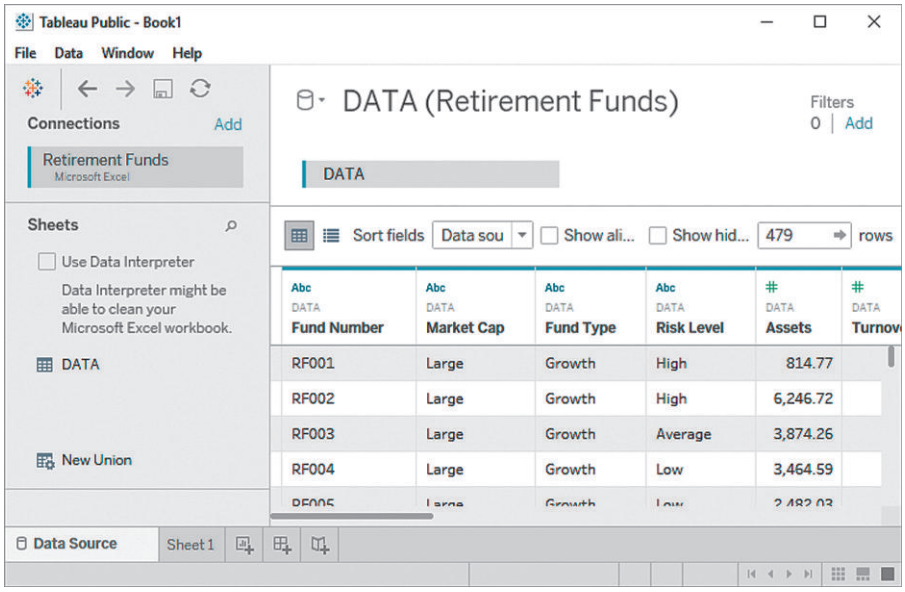
Tableau does not support the direct entry of data values. When using Tableau, one *imports* data from other sources, such as Microsoft Excel workbooks, text files, JSON files, or data retrieved from server-based data systems. Many users create and save data files in another program such as Excel and then open the saved data file in Tableau to import the data.

TG.3 OPEN or SAVE a WORKBOOK

Use **File → Open** or **File → Open from Tableau Public**.
Use **File → Save to Tableau Public As**.

Use Open to import simple data sources such as Microsoft Excel workbooks or text files or to open a *Tableau* workbook that previously downloaded to a local computing device. Most Tableau Guide instructions begin by opening a *Microsoft Excel* workbook to avoid the limitations of Tableau Public that Appendix Section G.5 discusses. (This procedure mimics the typical usage of creating data in another program that Section TG.2 discusses.)

To open a new workbook, select **File → New**. For a new workbook, the Data tab will display the hyperlink **Connect to Data**. Clicking the hyperlink displays the Connect panel from which data sources can be retrieved. When opening an Excel workbook that contains more than one data worksheet, each Excel worksheet being used must be defined as a separate data source. This Connect panel also appears when opening an Excel workbook directly using the file open command. The illustration below shows the data source linked to the Data worksheet of the Excel Retirement Funds workbook.



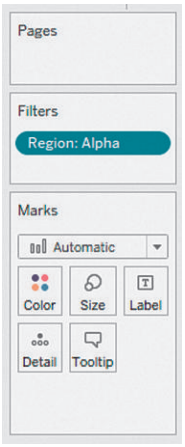
Using the Open from Tableau Public or Save to Tableau Public As command requires a valid Tableau online account. (Accounts are complimentary but require registration.) Using these commands means signing into a Tableau Public account and retrieving (or storing) a Tableau workbook from that account or from an account that has been shared. Save to Tableau Public As stores the Tableau workbook in the account and opens a web browser to display the workbook and to permit its downloading to a local computing device. In paid-subscription editions of Tableau Desktop, these open and save commands appear in the Server menu and not in the File menu.

Although not used in the Tableau Guides, Tableau Public permits *join* and *union* operations to combine columns from two data sources. **Join** operations combine two tables, typically by matching values in a variable column that both original tables share. **Union** operations add rows. Union operations require that tables share columns that hold values for the same variables. Joins and unions can solve problems that arise from seeking to perform an analysis of data on a set of variables stored in two different places, such as two different Excel data worksheets in the same Excel workbook.

TG.4 WORKING with DATA

Tableau Desktop uses the term **data field** to refer to what this book calls a variable or, in some contexts, a column. What an Excel, JMP, or Minitab user would call a formula function, Tableau calls an aggregation. What an Excel, JMP, or MInitab user would call a formula, Tableau calls an aggregate calculation. Tableau also invents its own vocabulary for several user interface elements in the worksheet window that users of Excel, JMP, or Minitab may know under more common names. Knowing this vocabulary can be helpful when consulting the Tableau help system or other references.

Tableau calls the Pages, Filters, Columns, and Rows areas, all shown in the Section TG.1 illustration, *shelves*. The Marks area seems to be a shelf but for reasons that would be only self-evident to a regular user of Tableau, the Marks area is called a *card*. **Shelves** are places into which things can be placed or dropped, such as the *pills* that have been placed in the Filters, Row, and Columns shelves. A **pill** represents some data and is so named because it reminds some of a medicinal capsule. In its simplest form, a pill represents a data field in the data source. However, pills can represent a filtering operation, such as the Filters shelf pill in the illustration below, or a calculated result, similar to a worksheet or data table formula. Pills can be either blue or green, reflecting the type of numerical data, discrete or continuous (see Section TG1.1), or red, reflecting an error condition.



In the Section TG.1 illustration, the Store Dimension has been dragged to the Rows shelf, creating the Store pill and the Mobile Electronics Sales Measure has been dropped on the Columns shelf, creating the SUM(Mobile Electronics Sales) pill (slightly truncated, see figure detail below). Dropping a measure name on a shelf creates an aggregation (formula function). The SUM aggregation pill sums mobile electronic sales values by rows (each store). In the special case where each store is represented by only one row, there is no actual summing of values.

TG.5 PRINT a WORKBOOK

Tableau Public does not contain a print function that would print worksheets. (Commercial versions of Tableau Desktop do.) To print a tabular or visual summary, use a screen capture utility to capture the display for later printing. For online worksheets displayed in a web browser, use the print function of the browser.

1 Defining and Collecting Data

CONTENTS

OBJECTIVES

- Understand issues that arise when defining variables
- How to define variables
- Understand the different measurement scales
- How to collect data
- Identify the different ways to collect a sample
- Understand the issues involved in data preparation
- Understand the types of survey errors

USING STATISTICS
Defining Moments

#1 You're the sales manager in charge of the best-selling beverage in its category. For years, your chief competitor has made sales gains, claiming a better tasting product. Worse, a new sibling product from your company, known for its good taste, has quickly gained significant market share at the expense of your product. Worried that your product may soon lose its number one status, you seek to improve sales by improving the product's taste. You experiment and develop a new beverage formulation. Using methods taught in this book, you conduct surveys and discover that people overwhelmingly like the newer formulation. You decide to use that new formulation going forward, having statistically shown that people prefer it. *What could go wrong?*

#2 You're a senior airline manager who has noticed that your frequent fliers always choose another airline when flying from the United States to Europe. You suspect fliers make that choice because of the other airline's perceived higher quality. You survey those fliers, using techniques taught in this book, and confirm your suspicions. You then design a new survey to collect detailed information about the quality of all components of a flight, from the seats to the meals served to the flight attendants' service. Based on the results of that survey, you approve a costly plan that will enable your airline to match the perceived quality of your competitor. *What could go wrong?*

In both cases, much did go wrong. Both cases serve as cautionary tales that if you choose the wrong variables to study, you may not end up with results that support making better decisions. Defining and collecting data, which at first glance can seem to be the simplest tasks in the DCOVA framework, can often be more challenging than people anticipate.

As the initial chapter notes, statistics is a way of thinking that can help fact-based decision making. But statistics, even properly applied using the DCOVA framework, can never be a substitute for sound management judgment. If you misidentify the business problem or lack proper insight into a problem, statistics cannot help you make a good decision. Case #1 retells the story of one of the most famous marketing blunders ever, the change in the formulation of Coca-Cola in the 1980s. In that case, Coke brand managers were so focused on the taste of Pepsi and the newly successful sibling Diet Coke that they decided only to define a variable and collect data about which drink tasters preferred in a blind taste test. When New Coke was preferred, even over Pepsi, managers rushed the new formulation into production. In doing so, those managers failed to reflect on whether the statistical results about a test that asked people to compare one-ounce samples of several beverages would demonstrate anything about beverage sales. After all, people were asked which beverage tasted better, not whether they would buy that better-tasting beverage in the future. New Coke was an immediate failure, and Coke managers reversed their decision a mere 77 days after introducing their new formulation (see reference 7).

Coke managers also overlooked other issues, such as people's emotional connection and brand loyalty to Coca-Cola, issues better discussed in a marketing book than this book.

Case #2 represents a composite story of managerial actions at several airlines. In some cases, managers overlooked the need to state operational definitions for quality factors about which fliers were surveyed. In at least one case, statistics was applied correctly, and an airline spent great sums on upgrades and was able to significantly improve quality. Unfortunately, their frequent fliers still chose the competitor's flights. In this case, no statistical survey about quality could reveal the managerial oversight that given the same level of quality between two airlines, frequent fliers will almost always choose the cheaper airline. While quality was a significant variable of interest, it was not the most significant.

The lessons of these cases apply throughout this book. Due to the necessities of instruction, the examples and problems in all but the last chapter include preidentified business problems and defined variables. Identifying the business problem or objective to be considered is always a prelude to applying the DCOVA framework.

1.1 Defining Variables

Identifying a proper business problem or objective enables one to begin to identify and define the variables for analysis. For each variable identified, assign an **operational definition** that specifies the type of variable and the *scale*, the type of measurement, that the variable uses.

EXAMPLE 1.1

Defining Data at GT&M

You have been hired by Good Tunes & More (GT&M), a local electronics retailer, to assist in establishing a fair and reasonable price for Whitney Wireless, a privately held chain that GT&M seeks to acquire. You need data that would help to analyze and verify the contents of the wireless company's basic financial statements. A GT&M manager suggests that one variable you should use is monthly sales. What do you do?

SOLUTION Having first confirmed with the GT&M financial team that monthly sales is a relevant variable of interest, you develop an operational definition for this variable. Does this variable refer to sales per month for the entire chain or for individual stores? Does the variable refer to net or gross sales? Do the monthly sales data represent number of units sold or currency amounts? If the data are currency amounts, are they expressed in U.S. dollars? After getting answers to these and similar questions, you draft an operational definition for ratification by others working on this project.

Classifying Variables by Type

The type of data that a variable contains determines the statistical methods that are appropriate for a variable. Broadly, all variables are either **numerical**, variables whose data represent a counted or measured quantity, or **categorical**, variables whose data represent categories. Gender

student TIP

Some prefer the terms **quantitative** and **qualitative** over the terms numerical and categorical when describing variables. These two pairs of terms are interchangeable.

with its categories male and female is a categorical variable, as is the variable preferred-New-Coke with its categories yes and no. In Example 1.1, the monthly sales variable is numerical because the data for this variable represent a quantity.

For some statistical methods, numerical variables must be further specified as either being *discrete* or *continuous*. **Discrete** numerical variables have data that arise from a counting process. Discrete numerical variables include variables that represent a "number of something," such as the monthly number of smartphones sold in an electronics store. **Continuous** numerical variables have data that arise from a measuring process. The variable "the time spent waiting on a checkout line" is a continuous numerical variable because its data represent timing measurements. The data for a continuous variable can take on any value within a continuum or an interval, subject to the precision of the measuring instrument. For example, a waiting time could be 1 minute, 1.1 minutes, 1.11 minutes, or 1.113 minutes, depending on the precision of the electronic timing device used.

For a particular variable, one might use a numerical definition for one problem, but use a categorical definition for another problem. For example, a person's age might seem to always be a numerical age variable, but what if one was interested in comparing the buying habits of children, young adults, middle-aged persons, and retirement-age people? In that case, defining age as a categorical variable would make better sense.

Measurement Scales

Determining the **measurement scale** that the data for a variable represent is part of defining a variable. The measurement scale defines the ordering of values and determines if differences among pairs of values for a variable are equivalent and whether one value can be expressed in terms of another. Table 1.1 presents examples of measurement scales, some of which are used in the rest of this section.

TABLE 1.1
Examples of Different Scales and Types

Data	Scale, Type	Values
Cellular provider	nominal, categorical	AT&T, T-Mobile, Verizon, Other, None
Excel skills	ordinal, categorical	novice, intermediate, expert
Temperature (°F)	interval, numerical	−459.67°F or higher
SAT Math score	interval, numerical	a value between 200 and 800, inclusive
Item cost (in $)	ratio, numerical	$0.00 or higher

student TIP

JMP and Tableau users will benefit the most from understanding measurement scales.

Define numerical variables as using either an **interval scale**, which expresses a difference between measurements that do not include a true zero point, or a **ratio scale**, an ordered scale that includes a true zero point. Categorical variables use measurement scales that provide less insight into the values for the variable. For data measured on a **nominal scale**, category values express no order or ranking. For data measured on an **ordinal scale**, an ordering or ranking of category values is implied. Ordinal scales contain some information to compare values but not as much as interval or ratio scales. For example, the ordinal scale poor, fair, good, and excellent allows one to know that "good" is better than poor or fair and not better than excellent. But unlike interval and ratio scales, one would not know that the difference from poor to fair is the same as fair to good (or good to excellent).

PROBLEMS FOR SECTION 1.1

LEARNING THE BASICS

1.1 Four different beverages are sold at a fast-food restaurant: soft drinks, tea, coffee, and bottled water.

Explain why the type of beverage sold is an example of a categorical variable.

1.2 U.S. businesses are listed by size: small, medium, and large.

Explain why business size is an example of categorical variable.

1.3 The time it takes to download a video from the Internet is measured.

Explain why the download time is a continuous numerical variable.

APPLYING THE CONCEPTS

SELF TEST **1.4** For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous.

a. Number of cell phones in the household
b. Monthly data usage (in MB)
c. Number of text messages exchanged per month
d. Voice usage per month (in minutes)
e. Whether the cell phone is used for streaming video

1.5 The following information is collected from students upon exiting the campus bookstore during the first week of classes.

a. Amount of time spent shopping in the bookstore
b. Number of textbooks purchased
c. Academic major
d. Gender

Classify each variable as categorical or numerical.

1.6 For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous.

a. Name of Internet service provider
b. Time, in hours, spent surfing the Internet per week
c. Whether the individual uses a mobile phone to stream video
d. Number of online purchases made in a month
e. Where the individual accesses social networks to find sought-after information

1.7 For each of the following variables, determine whether the variable is categorical or numerical. If the variable is numerical, determine whether the variable is discrete or continuous.

a. Amount of money spent on clothing in the past month
b. Favorite department store
c. Most likely time period during which shopping for clothing takes place (weekday, weeknight, or weekend)
d. Number of pairs of shoes owned

1.8 Suppose the following information is collected from Robert Keeler on his application for a home mortgage loan at the Metro County Savings and Loan Association.

a. Monthly payments: $2,227
b. Number of jobs in past 10 years: 1
c. Annual family income: $96,000
d. Marital status: Married

Classify each of the responses by type of data.

1.9 One of the variables most often included in surveys is income. Sometimes the question is phrased, "What is your income (in thousands of dollars)?" In other surveys, the respondent is asked to "Select the circle corresponding to your income level" and is given a number of income ranges to choose from.

a. In the first format, explain why income might be considered either discrete or continuous.
b. Which of these two formats would you prefer to use if you were conducting a survey? Why?

1.10 If two students both score 90 on the same examination, what arguments could be used to show that the underlying variable—test score—is continuous?

1.11 The director of market research at a large department store chain wanted to conduct a survey throughout a metropolitan area to determine the amount of time working women spend shopping for clothing in a typical month.

a. Indicate the type of data the director might want to collect.
b. Develop a first draft of the questionnaire needed by writing three categorical questions and three numerical questions that you feel would be appropriate for this survey.

1.2 Collecting Data

Collecting data using improper methods can spoil any statistical analysis. For example, Coca-Cola managers in the 1980s (see page 19) faced advertisements from their competitor publicizing the results of a "Pepsi Challenge" in which taste testers consistently favored Pepsi over Coke. No wonder—test recruiters deliberately selected tasters they thought would likely be more favorable to Pepsi and served samples of Pepsi chilled, while serving samples of Coke lukewarm (not a very fair comparison!). These introduced biases made the challenge anything but a proper scientific or statistical test. Proper data collection avoids introducing biases and minimizes errors.

Populations and Samples

Data are collected from either a population or a sample. A **population** contains all the items or individuals of interest that one seeks to study. All of the GT&M sales transactions for a specific year, all of the full-time students enrolled in a college, and all of the registered voters in Ohio are examples of populations. A **sample** contains only a portion of a population of interest. One analyzes a sample to estimate characteristics of an entire population. For example, one might select a sample of 200 sales transactions for a retailer or select a sample of 500 registered voters in Ohio in lieu of analyzing the populations of all the sales transactions or all the registered voters.

One uses a sample when selecting a sample will be less time consuming or less cumbersome than selecting every item in the population or when analyzing a sample is less cumbersome or

learnMORE
Read the SHORT TAKES for Chapter 1 for a further discussion about data sources.

more practical than analyzing the entire population. Section FTF.3 defines *statistic* as a "value that summarizes the data of a specific variable." More precisely, a **statistic** summarizes the value of a specific variable for sample data. Correspondingly, a **parameter** summarizes the value of a population for a specific variable.

Data Sources

Data sources arise from the following activities:

- Capturing data generated by ongoing business activities
- Distributing data compiled by an organization or individual
- Compiling the responses from a survey
- Conducting an observational study and recording the results of the study
- Conducting a designed experiment and recording the outcomes of the experiment

Tableau uses the term data source *in a different sense, to refer to the data being presented as a specific tabular or visual summary.*

When the person conducting an analysis performs one of these activities, the data source is a **primary data source**. When one of these activities is done by someone other than the person conducting an analysis, the data source is a **secondary data source**.

Capturing data can be done as a byproduct of an organization's transactional information processing, such as the storing of sales transactions at a retailer, or as result of a service provided by a second party, such as customer information that a social media website business collects on behalf of another business. Therefore, such data capture may be either a primary or a secondary source.

Typically, organizations such as market research firms and trade associations distribute complied data, as do businesses that offer syndicated services, such as The Neilsen Company, known for its TV ratings. Therefore, this source of data is usually a secondary source. (If one supervised the distribution of a survey, compiled its results, and then analyzed those results, the survey would be a primary data source.)

Choosing to conduct an observation study or a designed experiment on a variable of interest affects the statistical methods and the decision-making processes that can be used, as Chapters 10 and 14 further explain.

In both observational studies and designed experiments, researchers that collect data are looking for the effect of some change, called a **treatment**, on a variable of interest. In an observational study, the researcher collects data in a natural or neutral setting and has no direct control of the treatment. For example, in an observational study of the possible effects on theme park usage patterns that a new electronic payment method might cause, one would take a sample of guests, identify those who use the new method and those who do not, and then "observe" if those who use the new method have different park usage patterns. As a designed experiment, one would select guests to use the new electronic payment method and then discover if those guests have theme park usage patterns that are different from the guests not selected to use the new payment method.

PROBLEMS FOR SECTION 1.2

APPLYING THE CONCEPTS

1.12 The American Community Survey (**www.census.gov/acs**) provides data every year about communities in the United States. Addresses are randomly selected, and respondents are required to supply answers to a series of questions.

a. Which of the sources of data best describe the American Community Survey?

b. Is the American Community Survey based on a sample or a population?

1.13 Visit the website of the Gallup organization at **www.gallup.com**. Read today's top story. What type of data source is the top story based on?

1.14 Visit the website of the Pew Research organization at **www.pewresearch.org**. Read today's top story. What type of data source is the top story based on?

1.15 Transportation engineers and planners want to address the dynamic properties of travel behavior by describing in detail the driving characteristics of drivers over the course of a month. What type of data collection source do you think the transportation engineers and planners should use?

1.16 Visit the home page of the Statistics Portal "Statista" at **www.statista.com**. Examine one of the "Popular infographic topics" in the Infographics section on that page. What type of data source is the information presented here based on?

1.3 Types of Sampling Methods

When selecting a sample to collect data, begin by defining the **frame**. The frame is a complete or partial listing of the items that make up the population from which the sample will be selected. Inaccurate or biased results can occur if a frame excludes certain groups, or portions of the population. Using different frames to collect data can lead to different, even opposite, conclusions.

Using the frame, select either a nonprobability sample or a probability sample. In a **nonprobability sample**, select the items or individuals without knowing their probabilities of selection. In a **probability sample**, select items based on known probabilities. Whenever possible, use a probability sample as such a sample will allow one to make inferences about the population being analyzed.

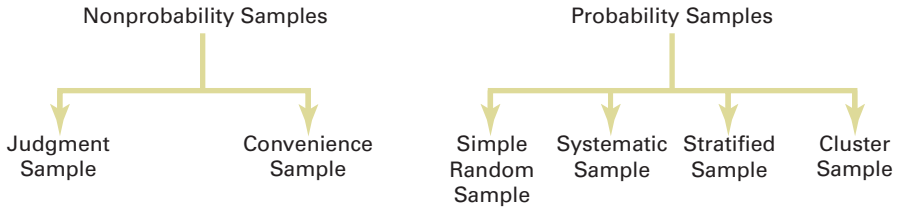
Nonprobability samples can have certain advantages, such as convenience, speed, and low cost. Such samples are typically used to obtain informal approximations or as small-scale initial or pilot analyses. However, because the theory of statistical inference depends on probability sampling, nonprobability samples *cannot be used* for statistical inference and this more than offsets those advantages in more formal analyses.

Figure 1.1 shows the subcategories of the two types of sampling. A nonprobability sample can be either a convenience sample or a judgment sample. To collect a **convenience sample**, select items that are easy, inexpensive, or convenient to sample. For example, in a warehouse of stacked items, selecting only the items located on the top of each stack and within easy reach would create a convenience sample. So, too, would be the responses to surveys that the websites of many companies offer visitors. While such surveys can provide large amounts of data quickly and inexpensively, the convenience samples selected from these responses will consist of self-selected website visitors. (Read the *Consider This* essay on page 31 for a related story.)

To collect a **judgment sample**, collect the opinions of preselected experts in the subject matter. Although the experts may be well informed, one cannot generalize their results to the population.

The types of probability samples most commonly used include simple random, systematic, stratified, and cluster samples. These four types of probability samples vary in terms of cost, accuracy, and complexity, and they are the subject of the rest of this section.

FIGURE 1.1
Types of samples



Simple Random Sample

In a **simple random sample**, every item from a frame has the same chance of selection as every other item, and every sample of a fixed size has the same chance of selection as every other sample of that size. Simple random sampling is the most elementary random sampling technique. It forms the basis for the other random sampling techniques. However, simple random sampling has its disadvantages. Its results are often subject to more variation than other sampling methods. In addition, when the frame used is very large, carrying out a simple random sample may be time consuming and expensive.

With simple random sampling, use n to represent the sample size and N to represent the frame size. Number every item in the frame from 1 to N. The chance that any particular member of the frame will be selected during the first selection is $1/N$.

Select samples with replacement or without replacement. **Sampling with replacement** means that selected items are returned to the frame, where it has the same probability of being selected again. For example, imagine a fishbowl containing N business cards, one card for each person. The first selection selects the card for Grace Kim. After the pertinent information has been recorded, the business card is placed back in the fishbowl. All cards are thoroughly mixed and a second card selected. On the second selection, Grace Kim has the same probability of being selected again, $1/N$.

Most sampling is *sampling without replacement*. **Sampling without replacement** means that once an item has been selected, the item cannot ever again be selected for the sample.

The chance that any particular item in the frame will be selected—for example, the business card for Grace Kim—on the first selection is $1/N$. The chance that any card not previously chosen will be chosen on the second selection becomes 1 out of $N - 1$.

When creating a simple random sample, avoid the "fishbowl" method of selecting a sample because this method lacks the ability to thoroughly mix items and, therefore, randomly select a sample. Instead, use a more rigorous selection method.

learnMORE
Learn to use a table of random numbers to select a simple random sample in the **Section 1.3 LearnMore** online topic.

One such method is to use a **table of random numbers**, such as Table E.1 in Appendix E, for selecting the sample. A table of random numbers consists of a series of digits listed in a randomly generated sequence. To use a random number table for selecting a sample, assign code numbers to the individual items of the frame. Then generate the random sample by reading the table of random numbers and selecting those individuals from the frame whose assigned code numbers match the digits found in the table. Because every digit or sequence of digits in the table is random, the table can be read either horizontally or vertically. The margins of the table designate row numbers and column numbers, and the digits are grouped into sequences of five in order to make reading the table easier.

Because the number system uses 10 digits (0, 1, 2, … , 9), the chance that any particular digit will be randomly generated is equal 1 out of 10 and is equal to the probability of generating any other digit. For a generated sequence of 800 digits, one would expect about 80 to be the digit 0, 80 to be the digit 1, and so on.

Systematic Sample

In a **systematic sample**, partition the N items in the frame into n groups of k items, where

$$k = \frac{N}{n}$$

Round k to the nearest integer. To select a systematic sample, choose the first item to be selected at random from the first k items in the frame. Then, select the remaining $n - 1$ items by taking every kth item thereafter from the entire frame.

If the frame consists of a list of prenumbered checks, sales receipts, or invoices, taking a systematic sample is faster and easier than taking a simple random sample. A systematic sample is also a convenient mechanism for collecting data from membership directories, electoral registers, class rosters, and consecutive items coming off an assembly line.

To take a systematic sample of $n = 40$ from the population of $N = 800$ full-time employees, partition the frame of 800 into 40 groups, each of which contains 20 employees. Then select a random number from the first 20 individuals and include every twentieth individual after the first selection in the sample. For example, if the first random number selected is 008, subsequent selections will be 028, 048, 068, 088, 108, … , 768, and 788.

Simple random sampling and systematic sampling are simpler than other, more sophisticated, probability sampling methods, but they generally require a larger sample size. In addition, systematic sampling is prone to selection bias that can occur when there is a pattern in the frame. To overcome the inefficiency of simple random sampling and the potential selection bias involved with systematic sampling, one can use either stratified sampling methods or cluster sampling methods.

Stratified Sample

learnMORE
Learn how to select a stratified sample in the online in the **Section 1.3 LearnMore** online topic.

In a **stratified sample**, first subdivide the N items in the frame into separate subpopulations, or **strata**. A stratum is defined by some common characteristic, such as gender or year in school. Then select a simple random sample within each of the strata and combine the results from the separate simple random samples. Stratified sampling is more efficient than either simple random sampling or systematic sampling because the representation of items across the entire population is assured. The homogeneity of items within each stratum provides greater precision in the estimates of underlying population parameters. In addition, stratified sampling enables one to reach conclusions about each strata in the frame. However, using a stratified sample requires that one can determine the variable(s) on which to base the stratification and can also be expensive to implement.

Cluster Sample

In a **cluster sample**, divide the N items in the frame into clusters that contain several items. **Clusters** are often naturally occurring groups, such as counties, election districts, city blocks,