ISBN-13: 978-0-13-417305-4
ISBN-10: 0-13-417305-8
EAN
9 780134 173054 90000

www.pearsonhighered.com

PEARSON

ALWAYS LEARNING

LEVINE
STEPHAN
SZABAT

Statistics for Managers

Using Microsoft® Excel

8TH
EDITION

PEARSON

Statistics for Managers

Using Microsoft® Excel

8TH EDITION

David M.
LEVINE

David F.
STEPHAN

Kathryn A.
SZABAT

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, sparklines, gauges, treemaps **(Sections 2.2, 2.4, 2.6, 17.4)** 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)** Index numbers **(online Section 16.8)**	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)** Chi-square test for a variance or standard deviation **(online Section 12.7)**	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)** Wilcoxon rank sum test **(Section 12.4)** 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)** McNemar test for two related samples **(online Section 12.6)**
Comparing more than two groups	One-way analysis of variance for comparing several means **(Section 11.1)** Kruskal-Wallis test **(Section 12.5)** Two-way analysis of variance **(Section 11.2)** Randomized block design **(online Section 11.3)**	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)** Time-series forecasting **(Chapter 16)** Sparklines **(Section 2.6)**	Contingency table, side-by-side bar chart, doughnut 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	Multiple regression **(Chapters 14 and 15)** Regression trees **(Section 17.5)**	Multidimensional contingency tables **(Section 2.6)** Drilldown and slicers **(Section 2.6)** Logistic regression **(Section 14.7)** Classification trees **(Section 17.5)**

Statistics for Managers Using Microsoft® Excel

8th Edition

David M. Levine

Department of Statistics and Computer Information Systems

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

David F. Stephan

Two Bridges Instructional Technology

Kathryn A. Szabat

Department of Business Systems and Analytics

School of Business, La Salle University

PEARSON

Boston Columbus Indianapolis New York San Francisco
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editorial Director: Chris Hoag
Editor in Chief: Deirdre Lynch
Acquisitions Editor: Suzanna Bainbridge
Senior Editorial Assistant: Justin Billing
Program Manager: Chere Bemelmans
Project Manager: Sherry Berg
Program Management Team Lead: Karen Wernholm
Project Management Team Lead: Peter Silvia
Media Producer: Jean Choe
TestGen Content Manager: John Flanagan
QA Manager, Assessment Content: Marty Wright
Senior MathXL Content Developer: Robert Carroll
Marketing Manager: Tiffany Bitzel
Field Marketing Manager: Adam Goldstein
Marketing Assistant: Jennifer Myers
Senior Author Support/Technology Specialist: Joe Vetere
Manager, Rights Management: Gina Cheselka
Senior Procurement Specialist: Carol Melville
Associate Director of Design: Andrea Nix
Program Design Lead and Cover Design: Barbara Atkinson
Text Design: Cenveo Publisher Services
Production Coordination, Composition, and Illustrations: Lumina Datamatics
Cover Image: Serp/Shutterstock

MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS MAKE NO REPRESENTATIONS ABOUT THE SUITABILITY OF THE INFORMATION CONTAINED IN THE DOCUMENTS AND RELATED GRAPHICS PUBLISHED AS PART OF THE SERVICES FOR ANY PURPOSE. ALL SUCH DOCUMENTS AND RELATED GRAPHICS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND. MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS HEREBY DISCLAIM ALL WARRANTIES AND CONDITIONS WITH REGARD TO THIS INFORMATION, INCLUDING ALL WARRANTIES AND CONDITIONS OF MERCHANTABILITY, WHETHER EXPRESS, IMPLIED OR STATUTORY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. IN NO EVENT SHALL MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF INFORMATION AVAILABLE FROM THE SERVICES.

THE DOCUMENTS AND RELATED GRAPHICS CONTAINED HEREIN COULD INCLUDE TECHNICAL INACCURACIES OR TYPOGRAPHICAL ERRORS. CHANGES ARE PERIODICALLY ADDED TO THE INFORMATION HEREIN. MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS MAY MAKE IMPROVEMENTS AND/OR CHANGES IN THE PRODUCT(S) AND/OR THE PROGRAM(S) DESCRIBED HEREIN AT ANY TIME. PARTIAL SCREEN SHOTS MAY BE VIEWED IN FULL WITHIN THE SOFTWARE VERSION SPECIFIED.

Library of Congress Cataloging-in-Publication Data

Levine, David M., 1946– author.
Statistics for managers using Microsoft Excel / David M. Levine, David F. Stephan, Kathryn A. Szabat.—8th edition.
pages cm
Includes bibliographical references and index.
ISBN 978-0-13-417305-4 (hardcover)
1. Microsoft Excel (Computer file) 2. Management—Statistical methods. 3. Commercial statistics. 4. Electronic spreadsheets.
5. Management—Statistical methods—Computer programs. 6. Commercial statistics—Computer programs. I. Stephan, David F., author. II. Szabat, Kathryn A., author. III. Title.
HD30.215.S73 2017
519.50285'554—dc23
2015027139

1 2 3 4 5 6 7 8 9 10—CRK—19 18 17 16 15

PEARSON

www.pearsonhighered.com

ISBN-10: 0-13-417305-8
ISBN-13: 978-0-13-417305-4

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

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

About the Authors

Kathryn Szabat, David Levine, and David Stephan

David M. Levine, David F. Stephan, and Kathryn A. Szabat 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 Making Statistics More Effective in Schools and Business (MSMESB) 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.

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 the SEDSI conference and the DSI 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 he studied instructional technology at Teachers College, Columbia 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 MSMESB. 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.

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.

Brief Contents

Contents

8 Confidence Interval Estimation 237

9 Fundamentals of Hypothesis Testing: One-Sample Tests 270

10 Two-Sample Tests 307

11 Analysis of Variance 348

12 Chi-Square and Nonparametric Tests 386

15 Multiple Regression Model Building 521

16 Time-Series Forecasting 553

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 coauthors, 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 Making Statistics More Effective in Schools and Business (MSMESB) 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. We also benefit from the interests and efforts of our past coauthors, Mark Berenson and Timothy Krehbiel.

Our Educational Philosophy

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 into all statistics topics to illustrate how software can assist the business decision making process. (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 and Innovative in This Edition?

This eighth edition of *Statistics for Managers Using Microsoft Excel* contains these new and innovative features.

First Things First Chapter This new chapter provides an orientation that helps students start to understand the importance of business statistics and get ready to use Microsoft Excel even before they obtain a full copy of this book. Like its predecessor "Getting Started: Important Things to Learn First," this chapter has been developed and published to allow distribution online even before a first class meeting. Instructors teaching online or hybrid course sections may find this to be a particularly valuable tool to get students thinking about business statistics and learning the necessary foundational concepts.

Getting Ready to Analyze Data in the Future This newly expanded version of Chapter 17 adds a second Using Statistics scenario that serves as an introduction to business analytics methods. That introduction, in turn, explains several advanced Excel features while familiarizing students with the fundamental concepts and vocabulary of business analytics. As such, the chapter provides students with a path for further growth and greater awareness about applying business statistics and analytics in their other courses and their business careers.

Expanded Excel Coverage *Workbook* instructions replace the *In-Depth Excel* instructions in the Excel Guides and discuss more fully OS X Excel ("Excel for Mac") differences when they occur. Because the many current versions of Excel have varying capabilities, Appendix B begins by sorting through the possible confusion to ensure that students understand that not all Excel versions are alike.

In the Worksheet Notes that help explain the worksheet illustrations that in-chapter examples use as model solutions.

Many More Exhibits Stand-alone summaries of important procedures that serve as a review of chapter passages. Exhibits range from identifying best practices, such "Best Practices for Creating Visualizations" in Chapter 2, to serving as guides to data analysis such as the pair of "Questions to Ask" exhibits in Chapter 17.

New Visual Design This edition uses a new visual design that better organizes chapter content and provides a more uncluttered, streamlined presentation.

Revised and Enhanced Content

This eighth edition of *Statistics for Managers Using Microsoft Excel* contains the following revised and enhanced content.

Revised End-of-Chapter Cases The Managing Ashland MultiComm Services case that reoccurs throughout the book has several new or updated cases. The Clear Mountain State Student Survey case, also recurring, uses new data collected from a survey of undergraduate students to practice and reinforce statistical methods learned in various chapters.

Many New Applied Examples and Problems Many of the applied examples throughout this book use new problems or revised data. Approximately 43% of the problems are new to this edition. Many of the new problems in the end-of-section and end-of-chapter problem sets contain data from *The Wall Street Journal*, *USA Today*, and other news media as well as from industry and marketing surveys from leading consultancies and market intelligence firms.

New or Revised Using Statistics Scenarios This edition contains six all-new and three revised Using Statistics scenarios. Several of the scenarios form a larger narrative when considered together even as they can all be used separately and singularly.

New "Getting Started Learning Statistics" and "Preparing to Use Microsoft Excel for Statistics" sections Included as part of the First Things First chapter, these new sections replace the "Making Best Use" section of the previous editions. The sections prepare students for learning with this book by discussing foundational statistics and Excel concepts together and explain the various ways students can work with Excel while learning business statistics with this book.

Revised Excel Appendices These appendices review the foundational skills for using Microsoft Excel, review the latest technical and relevant setup information, and discuss optional but useful knowledge about Excel.

Software FAQ Appendix This appendix provides answers to commonly-asked questions about PHStat and using Microsoft Excel and related software with this book.

Distinctive Features

This eighth edition of *Statistics for Managers Using Microsoft Excel* continues the use of the following distinctive features.

Using Statistics Business Scenarios Each chapter begins with a Using Statistics scenario, an example that highlights how statistics is used in a functional area of business such as finance, information systems, management, and marketing. Every chapter uses its scenario throughout to provide an applied context for learning concepts. Most chapters conclude with a Using Statistics, Revisited section that reinforces the statistical methods and applications that a chapter discusses.

Emphasis on Data Analysis and Interpretation of Excel Results Our focus emphasizes analyzing data by interpreting results while reducing emphasis on doing calculations. For example, in the coverage of tables and charts in Chapter 2, we help students interpret various charts and explain when to use each chart discussed. Our coverage of hypothesis testing in Chapters 9 through 12 and regression and multiple regression in Chapters 13–15 include extensive software results so that the *p*-value approach can be emphasized.

Student Tips In-margin notes that reinforce hard-to-master concepts and provide quick study tips for mastering important details.

Other Pedagogical Aids We use an active writing style, boxed numbered equations, set-off examples that reinforce learning concepts, problems divided into “Learning the Basics” and “Applying the Concepts,” key equations, and key terms.

Digital Cases These cases ask students to examine interactive PDF documents to sift through various claims and information and discover the data most relevant to a business case scenario. In doing so, students determine whether the data support the conclusions and claims made by the characters in the case as well as learn how to identify 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 A special section at the end of this book provides answers to most of the even-numbered exercises of this book.

Flexibility Using Excel For almost every statistical method discussed, students can use Excel Guide model workbook solutions with the *Workbook* instructions or the *PHStat* instructions to produce the worksheet solutions that the book discusses and presents. And, whenever possible, the book provides *Analysis ToolPak* instructions to create similar solutions.

Extensive Support for Using Excel For readers using the *Workbook* instructions, this book explains operational differences among current Excel versions and provides alternate instructions when necessary.

PHStat PHStat is the Pearson Education Statistics add-in that makes operating Excel as distraction-free as possible. PHStat executes for you the low-level menu selection and worksheet entry tasks that are associated with Excel-based solutions. Students studying statistics can focus solely on mastering statistical concepts and not worry about having to become expert Excel users simultaneously.

PHStat creates the “live,” dynamic worksheets and chart sheets that match chapter illustrations and from which students can learn more about Excel. PHStat includes over 60 procedures including:

Descriptive Statistics: boxplot, descriptive summary, dot scale diagram, frequency distribution, histogram and polygons, Pareto diagram, scatter plot, stem-and-leaf display, one-way tables and charts, and two-way tables and charts

Probability and probability distributions: simple and joint probabilities, normal probability plot, and binomial, exponential, hypergeometric, and Poisson probability distributions

Sampling: sampling distributions simulation

Confidence interval estimation: for the mean, sigma unknown; for the mean, sigma known, for the population variance, for the proportion, and for the total difference

Sample size determination: for the mean and the proportion

One-sample tests: *Z* test for the mean, sigma known; *t* test for the mean, sigma unknown; chi-square test for the variance; and *Z* test for the proportion

Two-sample tests (unsummarized data): pooled-variance *t* test, separate-variance *t* test, paired *t* test, *F* test for differences in two variances, and Wilcoxon rank sum test

Two-sample tests (summarized data): pooled-variance *t* test, separate-variance *t* test, paired *t* test, *Z* test for the differences in two means, *F* test for differences in two variances, chi-square test for differences in two proportions, *Z* test for the difference in two proportions, and McNemar test

Multiple-sample tests: chi-square test, Marascuilo procedure Kruskal-Wallis rank test, Levene test, one-way ANOVA, Tukey-Kramer procedure, randomized block design, and two-way ANOVA with replication

Regression: simple linear regression, multiple regression, best subsets, stepwise regression, and logistic regression

Control charts: *p* chart, *c* chart, and *R* and *Xbar* charts

Decision-making: covariance and portfolio management, expected monetary value, expected opportunity loss, and opportunity loss

Data preparation: stack and unstack data

To learn more about PHStat, see Appendix C.

Visual Explorations The Excel workbooks allow students to interactively explore important statistical concepts in the normal distribution, sampling distributions, and regression analysis. For the normal distribution, students see the effect of changes in the mean and standard deviation on the areas under the normal curve. For sampling distributions, students use simulation to explore the effect of sample size on a sampling distribution. For regression analysis, students fit a line of regression and observe how changes in the slope and intercept affect the goodness of fit.

Chapter-by-Chapter Changes Made for This Edition

As authors, we take pride in updating the content of our chapters *and* our problem sets. Besides incorporating the new and innovative features that the previous section discusses, each chapter of the eighth edition of *Statistics for Managers Using Microsoft Excel* contains specific changes that refine and enhance our past editions as well as many new or revised problems.

The new **First Things First** chapter replaces the seventh edition's Let's Get Started chapter, keeping that chapter's strength while immediately drawing readers into the changing face of statistics and business analytics with a new opening Using Statistics scenario. And like the previous edition's opening chapter, Pearson Education openly posts this chapter so students can get started learning business statistics even before they obtain their textbooks.

Chapter 1 builds on the opening chapter with a new Using Statistics scenario that offers a cautionary tale about the importance of defining and collecting data. Rewritten Sections 1.1 ("Defining Variables") and 1.2 ("Collecting Data") use lessons from the scenario to underscore important points. Over one-third of the problems in this chapter are new or updated.

Chapter 2 features several new or updated data sets, including a new data set of 407 mutual funds that illustrate a number of descriptive methods. The chapter now discusses doughnut charts and sparklines and contains a reorganized section on organizing and visualizing a mix of variables. Section 2.7 ("The Challenge in Organizing and Visualizing Variables") expands on previous editions' discussions that focused solely on visualization issues. This chapter uses an updated Clear Mountain State student survey as well. Over half of the problems in this chapter are new or updated.

Chapter 3 also uses the new set of 407 mutual funds and uses new or updated data sets for almost all examples that the chapter presents. Updated data sets include the restaurant meal cost samples and the NBA values data. This chapter also uses an updated Clear Mountain State student survey. Just under one-half of the problems in this chapter are new or updated.

Chapter 4 uses an updated Using Statistics scenario while preserving the best features of this chapter. The chapter now starts a section on Bayes' theorem which completes as an online section, and 43% of the problems in the chapter are new or updated.

Chapter 5 has been streamlined with the sections "Covariance of a Probability Distribution and Its Application in Finance" and "Hypergeometric Distribution" becoming online sections. Nearly 40% of the problems in this chapter are new or updated.

Chapter 6 features an updated Using Statistics scenario and the section "Exponential Distribution" has become an online section. This chapter also uses an updated Clear Mountain State student survey. Over one-third of the problems in this chapter are new or updated.

Chapter 7 now contains an additional example on sampling distributions from a larger population, and one-in-three problems are new or updated.

Chapter 8 has been revised to provide enhanced explanations of Excel worksheet solutions and contains a rewritten "Managing Ashland MultiComm Services" case. This chapter also uses an updated Clear Mountain State student survey, and new or updated problems comprise 39% of the problems.

Chapter 9 contains refreshed data for its examples and enhanced Excel coverage that provides greater details about the hypothesis test worksheets that the chapter uses. Over 40% of the problems in this chapter are new or updated.

Chapter 10 contains a new Using Statistics scenario that relates to sales of streaming video players and that connects to Using Statistics scenarios in Chapters 11 and 17. This chapter gains a new online section on effect size. The Clear Mountain State survey has been updated, and over 40% of the problems in this chapter are new or updated.

Chapter 11 expands on the Chapter 10 Using Statistics scenario that concerns the sales of mobile electronics. The Clear Mountain State survey has been updated. Over one-quarter of the problems in this chapter are new or updated.

Chapter 12 now incorporates material that was formerly part of the "Short Takes" for the chapter. The chapter also includes updated "Managing Ashland MultiComm Services" and Clear Mountain State student survey cases and 41% of the problems in this chapter are new or updated.

Chapter 13 features a brand new opening passage that better sets the stage for the discussion of regression that continues in subsequent chapters. Chapter 13 also features substantially revised and expanded Excel coverage that describes more fully the details of regression results worksheets. Nearly one-half of the problems in this chapter are new or updated.

Chapter 14 likewise contains expanded Excel coverage, with some Excel Guides sections completely rewritten. As with Chapter 13, nearly one-half of the problems in this chapter are new or updated.

Chapter 15 contains a revised opening passage, and the "Using Transformations with Regression Models" section has been greatly expanded with additional examples. Over 40% of the problems in this chapter are new or updated.

Chapter 16 contains updated chapter examples concerning movie attendance data and Coca-Cola Company and Wal-Mart Stores revenues. Two-thirds of the problems in this chapter are new or updated.

Chapter 17 has been retitled "Getting Ready to Analyze Data in the Future" and now includes sections on Business Analytics that return to issues that the First Things First Chapter scenario raises and that provide students with a path to future learning and application of business statistics. The chapter presents several Excel-based descriptive analytics techniques and illustrates how advanced statistical programs can work with worksheet data created in Excel. One-half of the problems in this chapter are new or updated.

A Note of Thanks

Creating a new edition of a textbook is a team effort, and we would like to thank our Pearson Education editorial, marketing, and production teammates: Suzanna Bainbridge, Chere Bemelmans, Sherry Berg, Tiffany Bitzel, Deirdre Lynch, Jean Choe, and Joe Vetere. We also thank our statistical readers and accuracy checkers James Lapp, Susan Herring, Dirk Tempelaar, Paul Lorczak, Doug Cashing, and Stanley Seltzer for their diligence in checking our work and Nancy Kincade of Lumina Datamatics. We also thank the following people for their helpful comments that we have used to improve this new edition: Anusua Datta, Philadelphia University; Doug Dotterweich, East Tennessee State University; Gary Evans, Purdue University; Chris Maurer, University of Tampa; Bharatendra Rai, University of Massachusetts Dartmouth; Joseph Snider and Keith Stracher, Indiana Wesleyan University; Leonie Stone, SUNY Geneseo; and Patrick Thompson, University of Florida.

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 #SMUME8 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.

We are happy to answer all types of questions, but if you need assistance using Excel or PHStat, please contact your local support person or Pearson Technical Support at **247pearsoned.custhelp.com**. They have the resources to resolve and walk you through a solution to many technical issues in a way we do not.

We invite you to visit us at **smume8.davidlevinestatistics.com** (**bit.ly/1I8Lv2K**), where you will find additional information and support for this book that we furnish in addition to all the resources that Pearson Education offers you on our book's behalf (see pages xxiii and xxiv).

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

BREAKTHROUGH
To improving results

Resources for Success

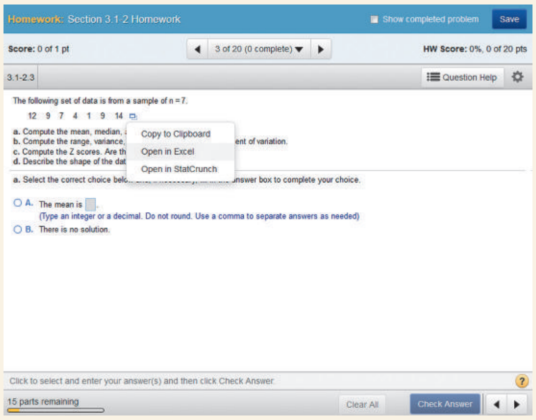
MyStatLab™ Online Course for Statistics for Managers Using Microsoft® Excel by Levine/Stephan/Szabat

(access code required)

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

New! Launch Exercise Data in Excel

Students are now able to quickly and seamlessly launch data sets from exercises within MyStatLab into a Microsoft Excel spreadsheet for easy analysis. As always, students may also copy and paste exercise data sets into most other software programs.



The following set of data is from a sample of n = 7.

12 9 7 4 1 9 14

a. Compute the mean, median,

b. Compute the range, variance, ent of variation.

c. Compute the Z scores. Are th

d. Describe the shape of the dat

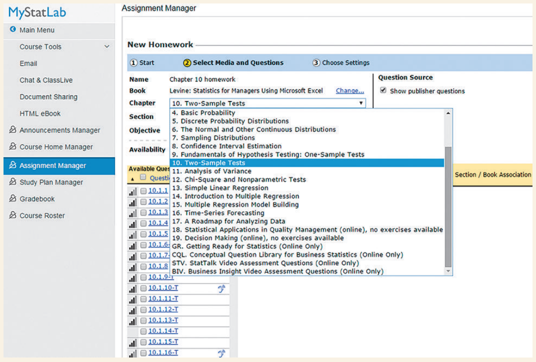
a. Select the correct choice belo answer box to complete your choice.

○ A. The mean is □.
(Type an integer or a decimal. Do not round. Use a comma to separate answers as needed)

○ B. There is no solution.

Click to select and enter your answer(s) and then click Check Answer.

15 parts remaining



Name: Chapter 10 homework

Book: Levine: Statistics for Managers Using Microsoft Excel Change...

Chapter: 10. Two-Sample Tests

Section

Objective

Availability

Question Source

☑ Show publisher questions

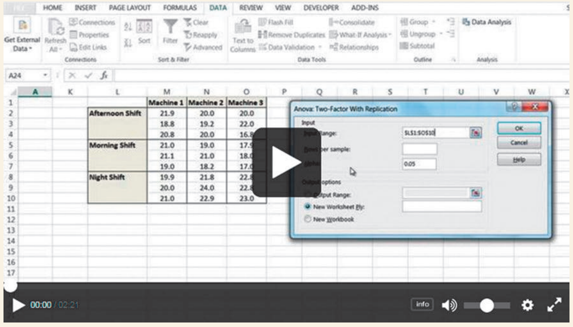
4. Basic Probability
5. Discrete Probability Distributions
6. The Normal and Other Continuous Distributions
7. Sampling Distributions
8. Confidence Interval Estimation
9. Fundamentals of Hypothesis Testing: One-Sample Tests
10. Two-Sample Tests
11. Analysis of Variance
12. Chi-Square and Nonparametric Tests
13. Simple Linear Regression
14. Introduction to Multiple Regression
15. Multiple Regression Model Building
16. Time-Series Forecasting
17. A Roadmap for Analyzing Data
18. Statistical Applications in Quality Management (online), no exercises available
19. Decision Making (online), no exercises available
GR. Getting Ready for Statistics (Online Only)
CQL. Conceptual Question Library for Business Statistics (Online Only)
STV. StatTalk Video Assessment Questions (Online Only)
BIV. Business Insight Video Assessment Questions (Online Only)

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.

Technology Tutorials and Study Cards

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



	Machine 1	Machine 2	Machine 3
Afternoon Shift	21.9	20.0	20.0
	18.8	19.2	22.0
	20.8	20.0	16.8
Morning Shift	21.0	19.0	17.9
	21.1	21.0	18.0
	19.0	18.2	17.0
Night Shift	19.9	21.8	22.8
	20.0	24.0	22.8
	21.0	22.9	23.0

Anova: Two-Factor With Replication

Input

Input Range: $L$1:$O$10

Rows per sample:

Alpha: 0.05

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

www.mystatlab.com

Resources for Success

Instructor Resources

Instructor's Solutions Manual, by Professor Pin Tian Ng of Northern Arizona University, includes solutions for end-of-section and end-of-chapter problems, answers to case questions, where applicable, and teaching tips for each chapter. The Instructor's Solutions Manual is available at the Instructor's Resource Center (**www.pearsonhighered.com/irc**) or in MyStatLab.

Lecture PowerPoint Presentations, by Professor Patrick Schur of Miami University (Ohio), are available for each chapter. The PowerPoint slides provide an instructor with individual lecture outlines to accompany the text. The slides include many of the figures and tables from the text. Instructors can use these lecture notes as is or can easily modify the notes to reflect specific presentation needs. The PowerPoint slides are available at the Instructor's Resource Center (**www.pearsonhighered.com/irc)** or in MyStatLab.

Test Bank, by Professor Pin Tian Ng of Northern Arizona University, 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. New to this edition are specific test questions that use Excel datasets. The Test Bank is available at the Instructor's Resource Center (**www.pearsonhighered.com/irc)** or in MyStatLab.

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, by Professor Pin Tian Ng of Northern Arizona University, provides detailed solutions to virtually all the even-numbered exercises and worked-out solutions to the self-test problems (ISBN-13: 978-0-13-417382-5).

Online resources

The complete set of online resources are discussed fully in Appendix C. For adopting instructors, the following resources are among those available at the Instructor's Resource Center (**www.pearsonhighered.com/irc)** or in MyStatLab.

www.mystatlab.com

First Things First

CONTENTS

EXCEL GUIDE

OBJECTIVES

- Statistics is a way of thinking that can lead to better decision making
- Statistics requires analytics 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

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.

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, sure enough, weekly grosses for *The Lion King* had dropped about 25% by the year 2009. But, for 2013, grosses were up 67% from 2009 and weekly grosses for 2013 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 17 will introduce you. The end result was a system that outperformed the ticket-selling methods others used.

studentTIP
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.)

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** provide one such framework.

DCOVA Framework

- **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, to reach conclusions and present those 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

You have already read that business statistics shows a preference for using software to perform calculations. You can perform calculations *faster and more accurately* using software than you can if you performed those calculations by hand.

When you use software, you do more than just enter data. You need to review and modify, and possibly create, solutions. In Microsoft Excel, you use worksheet solutions that contain a mix of *organized* data and instructions that perform calculations on that data. Being able to review and modify worksheet solutions requires analytical skills more than arithmetic skills.

Allowing individuals to create new solutions from scratch in business can create risk. For example, in the aftermath of the 2012 "London Whale" trading debacle, JP Morgan Chase discovered a worksheet that could greatly miscalculate the volatility of a trading portfolio (see reference 4). To avoid this unnecessary risk, businesses prefer to use **templates**, *reusable* worksheet solutions that have been previously audited and verified.

When templates prove impractical, businesses seek to use *model worksheet solutions*. These solutions provide employees a basis for modification that is more extensive than changes one would make to a template. Whether you use the Excel Guide workbooks or PHStat with this book, you will reflect business practice by working with templates and model solutions as you use this book to learn statistics. You will not find many from-scratch construction tasks other than for the tasks of organizing and visualizing data in this book.

student TIP

Examining the structure of worksheet templates and models can also be helpful if learning more about Excel is one of your secondary learning goals.

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 17 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.

student TIP

Because you cannot "download" a big data collection, this book uses conventional structured (worksheet) files, both small and large, to demonstrate some of the principles and methods of business analytics in selected chapters, including Chapter 17, which introduce you to business analytics.

"Big Data"

Big data are collections 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 other than the traditional structured forms such as data processing records, files, and tables and worksheets. These attributes of volume, velocity, and variety (see reference 5) distinguish big data from a set of data that contains a large number of similarly structured records or rows that you can place into a file or worksheet for browsing. In contrast, you cannot directly view big data; information system and statistical methods typically combine and summarize big data for you and then present the results of that processing.

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).

Structured Versus Unstructured Data

Statistics has traditionally used **structured data**, data that exist in repeating records or rows of similar format, such as the data found in the worksheet data files that this book describes in Appendix C. In contrast, **unstructured data** has very little or no repeating internal structure.

For example, to deeply analyze a group of companies, you might collect structured data in the form of published tables of financial data and the contents of fill-in-the-blank documents that record information from surveys you distributed. However, you might also collect unstructured data such as social media posts and tweets that do not have an internal repeating structure.

Typically, you preprocess or filter unstructured data before performing deep analysis. For example, to analyze social media posts you could use business analytics methods that determine whether the content of each post is a positive, neutral, or negative comment. The "type of comment" can become a new variable that can be inserted into a *structured* record, along with other attributes of the post, such as the number of words, and demographic data about the writer of the post.

Unstructured data can form part of a big data collection. When analyzed as part of a big data collection, you typically see the results of the preprocessing and not the unstructured data itself. Because unstructured data usually has some (external) structure, some authorities prefer to use the term *semistructured data*. If you are familiar with that term, undertand that this book's use of the phrase *unstructured data* incorporates that category.

FTF.3 Getting Started Learning Statistics

Learning the **operational definitions**, precise definitions and explanations that all can understand clearly, of several basic terms is a good way to get started learning statistics. Previously, you learned that *data* are the facts about the world that one seeks to study and explore. A related term, *variable of interest*, commonly shortened to *variable*, can be used to precisely define data in its statistical sense.

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 title and number of chapters, as these facts can vary from book to book. For a given item, variables have a specific value. For this book, the value of the variable title would be "Statistics for Managers Using Microsoft Excel," and "17" would be the value for the variable number of chapters.

Using the definition of variable, you can state the definition of data, in its statistical sense, 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 when referring to this book? Without this rule, you might say that the single list "Levine, Stephan, Szabat" is the value. However, applying this rule, we say that the variable author has the 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 your data into a computer system for analysis.

student TIP

Business analytics, discussed in Chapter 17, combine mostly inferential methods with methods from other disciplines.

Using the definitions of data and variable, you can restate the definition of statistics 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.

Do not confuse this use of the word statistics with the noun *statistic*, the plural of which is, confusingly, *statistics*.

Statistic

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, invalid statistics can be produced if any tasks in the DCOVA framework are applied incorrectly. As discussed in later chapters, 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 discussed later in this book, 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 Preparing to Use Microsoft Excel for Statistics

As Section FTF.1 explains, the proper use of business statistics requires a framework to apply statistics correctly, analytic skills, and software to automate calculation. This book uses Microsoft Excel to demonstrate the integral role of software in applying statistics to decision making, and preparing to use Microsoft Excel is one of the first things you can do to prepare yourself to learn business statistics from this book.

Microsoft Excel is the data analysis component of Microsoft Office that evolved from earlier electronic spreadsheets used in accounting and financial applications. In Excel, you use **worksheets** (also known as spreadsheets) that organize data in tabular blocks as well as store **formulas**, instructions to process that data. You make entries in worksheet **cells** that are formed by the intersections of worksheet rows and columns. You refer to individual cells by their column letter and row number address, such as A1 for the uppermost left top cell (in column A and row 1). Into each cell, you place a single data value or a formula. With the proper design, worksheets can also present summaries of data and results of applying a particular statistical method.

"Excel files" are not single worksheets but **workbooks**, collections of one or more worksheets and **chart sheets**, sheets that display visualizations of data. Because workbooks contain collections, you can clearly present information in more than one way on different "slides" (sheets), much like a slide show. For example, you can present on separate sheets the summary table and appropriate chart for the data for a variable. (These tasks are discussed in Chapter 2.) When designing model solutions, workbooks allow you to segregate the parts of the solution that users may change frequently, such as problem-specific data. For example, the typical model solution files that this book uses and calls **Excel Guide workbooks** have a "Data" worksheet, one or more worksheets and chart sheets that present the results, and one or more worksheets that document the formulas that a template or model solution uses.

Reusability Through Recalculation

Earlier in this chapter, you learned that businesses prefer using templates and model worksheet solutions. You can reuse templates and model solutions, applying a previously constructed and verified worksheet solution to another, similar problem. When you work with templates, you never enter or edit formulas, thereby greatly reducing the chance that the worksheet will produce erroneous results. When you work with a model worksheet solution, you need only to edit or copy certain formulas. By not having to enter your own formulas from scratch, you also minimize the chance of errors. (Recall an analyst's entering of his own erroneous formulas was uncovered in the London Whale investigation mentioned on page 3.)

Templates and model solutions are reusable because worksheets are capable of recalculation. In worksheet **recalculation**, results displayed by formulas can automatically change as the data that the formulas use change, but only if the formulas properly refer to the cells that contain the data that might change.

student TIP

Recalculation is always a basis for goal-seeking and what-if analyses that you may encounter in other business courses.

Practical Matters: Skills You Need

To use Excel effectively with this book, you will need to know how to make cell entries, how to navigate to, or open to, a particular worksheet in a workbook, how to print a worksheet, and how to open and save files. If these skills are new to you, review the introduction to these skills that starts later in this chapter and continues in Appendix B.

You *may* need to modify model worksheet solutions, especially as you progress into the later chapters of this book. However, this book does not *require* you to learn this additional (information systems) skill. You can choose to use PHStat, which performs those modifications for you. By automating the necessary modifications, PHStat reduces your chance of making errors.

student TIP

PHStat also automates the correction of errors that Excel sometimes makes in formatting charts, saving you time.

PHStat creates worksheet solutions that are identical to the solutions found in the Excel Guide workbooks and that are shown and annotated throughout this book. You will not learn anything less if you use PHStat, as you will be using and studying from the same solutions as those who decide not to use PHStat. If the information systems skill of modifying worksheets is one of your secondary goals, you can use PHStat to create solutions to several similar problems and then examine the modifications made in each solution.

PHStat uses a menu-driven interface and is an example of an **add-in**, a programming component designed to extend the abilities of Excel. Unlike add-ins such as the Data Analysis ToolPak that Microsoft packages with Excel, PHStat creates actual worksheets with working formulas. (The ToolPak and most add-ins produce a text-based report that is pasted into a worksheet.) Consider both PHStat and the set of Excel Guide workbooks as stand-ins for the template and model solution library that you would encounter in a well-run business.

Ways of Working with Excel

With this book, you can work with Excel by either using PHStat or making manual changes directly to the Excel Guide workbooks. Readers that are experienced Excel users may prefer making manual changes, and readers who use Excel in organizations that restrict the use of Microsoft Office add-ins may be forced to make such changes. Therefore, this book provides detailed instructions for using Excel with PHStat, which are labeled **PHStat**, *and* instructions for making manual changes to templates and model worksheet solutions, which are labeled **Workbook**.

studentTIP
For several advanced topics, PHStat produces solutions that cannot be easily produced from modified model worksheet solutions. For those topics, you will need to use PHStat.

In practice, if you face no restrictions on using add-ins, you may want to use a mix of both approaches if you have had some previous exposure to Excel. In this mix, you open the Excel Guide workbooks that contain the simpler templates and fill them in, and you use PHStat to modify the more complex solutions associated with statistical methods found in later chapters. You may also want to use PHStat when you construct charts, as PHStat automates the correction of chart formatting mistakes that Excel sometimes makes. (Making these corrections can be time-consuming and a distraction from learning statistics.)

This book also includes instructions for using the Data Analysis ToolPak, which is labeled **ToolPak**, for readers who prefer using this Microsoft-supplied add-in. Certain model worksheet solutions found in the Excel Guide workbooks, used by PHStat, and shown in this book mimic the appearance of ToolPak solutions to accommodate readers used to ToolPak results. Do not be fooled, though—while the worksheets mimic those solutions, the worksheets are fundamentally different, as they contain active formulas and not the pasted-in text of the ToolPak solutions.

Excel Guides

Excel Guides contain the detailed **PHStat**, **Workbook**, and, when applicable, **ToolPak** instructions. Guides present instructions by chapter section numbers, so, for example, Excel Guide Section EG2.3 provides instructions for the methods discussed in Section 2.3. Most Guide sections begin with a *key technique* that presents the most important or critical Excel feature that the templates and model solutions use and cite the *example* that the instructions that follow use. (The example is usually the example that has been worked out in the chapter section.)

For some methods, the Guides present separate instructions for summarized or unsummarized data. In such cases, you will see either **(summarized)** or **(unsummarized)** as part of the instruction label. When minor variations among current Excel versions affect the **Workbook** instructions, special sentences or separate instructions clarify the differences. (The minor variations do not affect either the **PHStat** or **ToolPak** instructions.)

Which Excel Version to Use?

studentTIP
OS X Excel is also known as “Excel for Mac.”

Use a current version of Microsoft Excel for Microsoft Windows or (Mac) OS X when working with the examples, problems, and Excel Guide instructions in this book. A current version is a version that receives “mainstream support” from Microsoft that includes updates, refinements, and online support. As this book went to press, current versions included Microsoft Windows Excel 2016, 2013, and 2010, and the OS X Excel 2016 and 2011. If you have an Office 365 subscription, you always have access to the most current version of Excel.

studentTIP
If you are currently using Excel 2007, consider upgrading to a newer version that will maximize your learning with this book and minimize your problems using Excel.

If you use Microsoft Windows Excel 2007, you should know that Microsoft has already ended mainstream support and will end all (i.e., security update) support for this version during the expected in-print lifetime of this book. Excel 2007 does not include all of the features of Excel used in this book and has a number of significant differences that affect various worksheet solutions. (This is further explained in Appendix F.) Note that many Excel Guide workbooks contain special worksheet solutions for use with Excel 2007. If you use PHStat with Excel 2007, PHStat will produce these special worksheet solutions automatically.

If you use a mobile Excel version such as Excel for Android Tablets, you will need an Office 365 subscription to open, examine, edit, and save Excel Guide workbooks and data workbooks. (Without a subscription, you can only open and examine those workbooks.) As this book went to press, the current version of Excel for Android Tablets did not support all of the Excel features discussed or used in this book and did not support the use of add-ins such as PHStat.

Conventions Used

In both the annotated worksheet solutions shown as chapter figures and in the Excel Guide instructions, this book uses the following conventions:

- What to type as an entry and where to type the entry appear in boldface: Enter **450** in cell **B5**.
- Names of special keys appear capitalized and in boldface: Press **Enter**.

- For improved readability, Excel ribbon tabs appear in mixed case (File, Insert), not capitalized (FILE, INSERT) as they appear in certain Excel versions.
- Menu and ribbon selections appear in boldface, and sequences of consecutive selections are linked using the ➔ symbol: Select **File ➔ New**. Select **PHStat ➔ Descriptive Statistics ➔ Boxplot**.
- Key combinations, two or more keys that you press at the same time, are shown in boldface: Press **Ctrl+C**. Press **Command+Enter**.
- Names of specific Excel functions, worksheets, or workbooks appear in boldface.

Placeholders that express the general case appear in italics and boldface, such as **AVERAGE (*cell range of variable*)**. When you encounter a placeholder, you replace it with an actual value. For example, you would replace ***cell range of variable*** with an actual variable cell range. By special convention in this book, PHStat menu sequences always begin with **PHStat,** even though in some Excel versions you must first select the Add-Ins tab to display the PHStat menu.

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." *The 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. Liberatore, M., and W. Luo. "The Analytics Movement." *Interfaces* 40 (2010): 313–324.
8. "What Is Big Data?" IBM Corporation, **www.ibm.com/big-data/us/en/**.

KEY TERMS

add-in	7
big data	4
cells	6
chart sheet	6
data	2
business analytics	4
DCOVA framework	3
descriptive statistics	5
formula	6
formula bar	10
inferential statistics	5
logical causality	6
operational definition	5
recalculation	7
statistic	5
statistics	2
structured data	4
summarized data	2
template	3
unstructured data	4
variable	5
workbook	6
worksheet	6

▼ EXCEL GUIDE

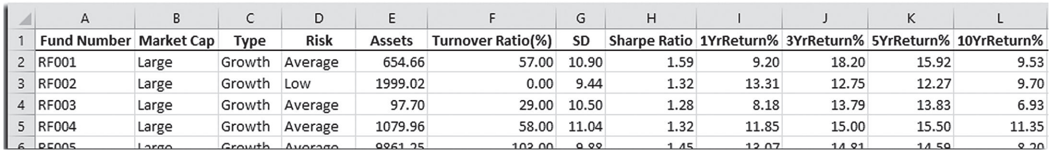
As explained earlier in this chapter, Excel Guides contain the detailed instructions for using Microsoft Excel with this book. Whether you choose to use the **PHStat**, **Workbook**, or, when applicable, **ToolPak** instructions (see page 8), you should know how to enter data for variables into a worksheet and how to review and inspect worksheets before applying them to a problem.

EG.1 ENTERING DATA

You should enter the data for variables using the style that the DATA worksheets of the Excel Guide workbooks and the Excel data files (see Appendix C) use. Those DATA worksheets use the business convention of entering the data for each variable in separate columns, and using the cell entry in the first row in each column as a heading to identify a variable by name. These worksheets also begin with column A, row 1 (cell A1) and do not skip any rows when entering data for a variable into a column.

To enter 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. As you type, what you type appears in a space above the worksheet called the **formula bar**. Complete your data entry by pressing **Tab** or **Enter** or by clicking the checkmark button in the formula bar.

When you enter data, never skip any rows in a column, and as a general rule, also avoid skipping any columns. Also try to avoid using numbers as row 1 variable headings; if you cannot avoid their use, precede such headings with apostrophes. When you create a new data worksheet, begin the first entry in cell A1, as the sample below shows. Pay attention to special instructions in this book that note specific orderings of the columns that hold your variables. For some statistical methods, entering variables in a column order that Excel does not expect will lead to incorrect results.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Fund Number	Market Cap	Type	Risk	Assets	Turnover Ratio(%)	SD	Sharpe Ratio	1YrReturn%	3YrReturn%	5YrReturn%	10YrReturn%
2	RF001	Large	Growth	Average	654.66	57.00	10.90	1.59	9.20	18.20	15.92	9.53
3	RF002	Large	Growth	Low	1999.02	0.00	9.44	1.32	13.31	12.75	12.27	9.70
4	RF003	Large	Growth	Average	97.70	29.00	10.50	1.28	8.18	13.79	13.83	6.93
5	RF004	Large	Growth	Average	1079.96	58.00	11.04	1.32	11.85	15.00	15.50	11.35
6	RF005	Large	Growth	Average	9861.25	103.00	9.88	1.45	13.07	14.81	14.59	8.20

student TIP

At the time this book went to press, some mobile versions of Excel could not display worksheets in formula view. If you are using such a version, you can select a cell and view the formula in the formula bar.

EG.2 REVIEWING WORKSHEETS

You should 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, you 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.

Whether you use PHStat or the Excel Guide workbooks, you will notice cell formatting operations that change the background color of cells, change text attributes such as boldface of cell entries, and round values to a certain number of decimal places (typically four). Because cells in PHStat worksheets and Excel Guide workbooks have been already formatted for you, using this book does not require that you know how to apply these formatting operations. However, if you want to learn more about cell formatting, Appendix B includes a summary of common formatting operations, including those used in the worksheet solutions presented in this book.

EG.3 IF YOU PLAN TO USE THE *WORKBOOK* INSTRUCTIONS

The *Workbook* instructions in the Excel Guides help you to modify model worksheet solutions by directly operating a current version of Microsoft Excel. For most statistical methods, the *Workbook* instructions will be identical for all current versions. In some cases, especially in the instructions for constructing tabular and visual summaries discussed in Chapter 2, the *Workbook* instructions can greatly vary from one version to another. In those cases, the Excel Guides express instructions in the most universal way possible. Many instructions ask you to select (click on) an item from a gallery of items and identify that item 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* to refer to either a task pane or a two-panel dialog box that contains similar or identical choices. A **task pane**, found in more recent versions, opens to the side of the worksheet and can remain onscreen indefinitely. Some parts of a pane may be initially hidden and you may need to click on an icon or label to reveal a hidden part to complete a command sequence. A **two-panel dialog box** opens over the worksheet and must be closed before you can continue your work. These dialog boxes contain one left panel, always visible, and a series of right panels, only one of which is visible at any given time. To reveal a hidden right panel, you click on a left panel entry, analogous to clicking an icon or label in a task pane. (To close either a task pane or a dialog box, click the system close button.)

Current Excel versions can vary in their command sequences. Excel Guide instructions show these variations as parenthetical phrases. For example, the command 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 Excels use Design and the current OS X Excel uses **Chart Design**.)

student TIP

The authors discourage you from using OS X Excel 2011 if you plan to use the Chapter 2 *Workbook* instructions for creating visual summaries.

In some cases, OS X Excel 2016 instructions differ so much that an Excel Guide presents an alternate instruction using this color and font. In addition, OS X Excel 2011 has significantly different command sequences for creating visual and some tabular summaries. If you plan to use OS X Excel 2011 with this book, take note of the Student Tip to the left. If you must use this older OS X Excel, be sure to download and use the **OS X Excel 2011 Supplement** that provides notes and instructions for creating visual and tabular summaries in OS X Excel 2011. (For methods other than visual and tabular summaries, OS X Excel 2011 uses the same or similar sequences that other Excel versions use.)

Again, if only one set of *Workbook* instructions appears, that set applies to *all* current versions. You do not need to be concerned about command sequence differences if you use the *PHStat* (or *Analysis ToolPak*) instructions. Those instructions are always the same for all current versions.

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, and you decide to use that new formulation going forward, having statistically shown that people prefer the new taste formulation. *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.

studentTIP

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.

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 6).

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.

Remember the lessons of these cases as you study the rest of this book. Due to the necessities of instruction, examples and problems presented in all chapters but the last one 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

When a proper business problem or objective has been identified, you can begin to define your data. You define data by defining variables. You assign an **operational definition** to each variable you identify and specify the type of variable and the *scale*, or type of measurement, the variable uses (the latter two concepts are discussed later in this section).

EXAMPLE 1.1

Defining Data at GT&M

You 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

You need to know the type of data that a variable defines in order to choose statistical methods that are appropriate for that data. Broadly, all variables are either **numerical**, variables whose data represent a counted or measured quantity, or **categorical**, variables whose data represent categories. Gender 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, you must further specify numerical variables 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 some data, you might define a numerical variable for one problem that you wish to study, but define the same data as a categorical variable for another. For example, a person's age might seem to always be a numerical variable, but what if you are interested in comparing the buying habits of children, young adults, middle-aged persons, and retirement-age people? In that case, defining age as categorical variable would make better sense.

studentTIP

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

Measurement Scales

You identify the **measurement scale** that the data for a variable represent, as 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 you can express one value in terms of another. Table1.1 presents examples of measurement scales, some of which are used in the rest of this section.

You 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. If a numerical variable has a ratio scale, you can characterize one value in terms of another. You can say that the item cost (ratio) $2 is twice as expensive as the item cost $1. However, because Fahrenheit temperatures use an interval scale, 2°F does not represent twice the heat of 1°F. For both interval and ratio scales, what the difference of 1 unit represents remains the same among pairs of values, so that the difference between $11 and $10 represents the same difference as the difference between $2 and $1 (and the difference between 11°F and 10°F represents the same as the difference between 2°F and 1°F).

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 give you 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 you to know that "good" is better than poor or fair and not better than excellent. But unlike interval and ratio scales, you do not know that the difference from poor to fair is the same as fair to good (or good to excellent).

learnMORE

Read the Short Takes for Chapter 1 for more examples of classifying variables as either categorical or numerical.

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

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.

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

b. Explain why the type of beverage is an example of a nominal-scaled variable.

1.2 U.S. businesses are listed by size: small, medium, and large. Explain why business size is an example of an ordinal-scaled variable.

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

a. Explain why the download time is a continuous numerical variable.

b. Explain why the download time is a ratio-scaled variable.

APPLYING THE CONCEPTS

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

a. Number of cellphones 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 cellphone is used for email

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 and determine its measurement scale.

1.6 For each of the following variables, determine whether the variable is categorical or numerical and determine its measurement scale. 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 connect to the Internet

d. Number of online purchases made in a month

e. Where the individual uses social networks to find sought-after information

1.7 For each of the following variables, determine whether the variable is categorical or numerical and determine its measurement scale. 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 and measurement scale.

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 score a 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 in (a) 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 12) 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

You collect your data from either a population or a sample. A **population** contains all the items or individuals of interest that you seek 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. You analyze a sample to estimate characteristics of an entire population. You might select a sample of 200 GT&M sales transactions, a sample of 50 full-time students selected for a marketing study, or a sample of 500 registered voters in Ohio in lieu of analyzing the populations identified in this paragraph.

You collect data from 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 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 a designed experiment and recording the outcomes of the experiment
- Conducting an observational study and recording the results of the study

When you perform the activity that collects the data, you are using a **primary data source**. When the data collection part of these activities is done by someone else, you are using 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 such as GT&M, 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. The other three sources are either primary or secondary, depending on whether you (your organization) are doing the activity. For example, if you oversee the distribution of a survey and the compilation of its results, the survey is a primary data source.

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 (the variable of interest) that a new electronic payment method might cause, you would take a sample of visitors, 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. In a designed experiment, you permit only those you select to use the new electronic payment method and then discover if the people you selected have different theme park usage patterns (from those who you did not select to use the new payment method). Choosing to use an observational study (or experiment) affects the statistical methods you apply and the decision-making processes that use the results of those methods, as later chapters (10, 11, and 17) will further explain.

learnMORE

Read the Short Takes for Chapter 1 for a further discussion about data sources.

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 **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 you collect data by selecting a sample, you 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.

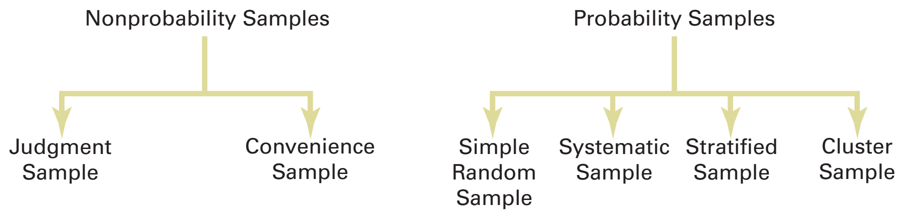
Using your frame, you select either a nonprobability sample or a probability sample. In a **nonprobability sample**, you select the items or individuals without knowing their probabilities of selection. In a **probability sample**, you select items based on known probabilities. Whenever possible, you should use a probability sample as such a sample will allow you 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.

Figure1.1 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**, you 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 tops 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 24 for a related story.)

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

FIGURE 1.1
Types of samples



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.

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, you use n to represent the sample size and N to represent the frame size. You number every item in the frame from 1 to N. The chance that you will select any particular member of the frame on the first selection is $1/N$.

You select samples with replacement or without replacement. **Sampling with replacement** means that after you select an item, you return it to the frame, where it has the same probability of being selected again. Imagine that you have a fishbowl containing N business cards, one card for each person. On the first selection, you select the card for Grace Kim. You record pertinent information and replace the business card in the bowl. You then mix up the cards in the bowl and select a second card. On the second selection, Grace Kim has the same probability of being selected again, $1/N$. You repeat this process until you have selected the desired sample size, n.

Typically, you do not want the same item or individual to be selected again in a sample. **Sampling without replacement** means that once you select an item, you cannot select it again. The chance that you will select any particular item in the frame—for example, the business card for Grace Kim—on the first selection is $1/N$. The chance that you will select any card not previously chosen on the second selection is now 1 out of $N - 1$. This process continues until you have selected the desired sample of size n.

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

learnMORE

Learn to use a table of random numbers to select a simple random sample in an online section of this chapter.

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, you first need to assign code numbers to the individual items of the frame. Then you 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 the number system uses 10 digits $(0, 1, 2, \ldots, 9)$, the chance that you will randomly generate any particular digit is equal to the probability of generating any other digit. This probability is 1 out of 10. Hence, if you generate a sequence of 800 digits, you would expect about 80 to be the digit 0, 80 to be the digit 1, and so on. 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. The digits themselves are grouped into sequences of five in order to make reading the table easier.

Systematic Sample

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

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

You round k to the nearest integer. To select a systematic sample, you choose the first item to be selected at random from the first k items in the frame. Then, you 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, you partition the frame of 800 into 40 groups, each of which contains 20 employees. You 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 you select is 008, your subsequent selections are 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, you can use either stratified sampling methods or cluster sampling methods.

Stratified Sample

In a **stratified sample**, you 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. You 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 you are ensured of the representation of items across the entire population. The homogeneity of items within each stratum provides greater precision in the estimates of underlying population parameters. In addition, stratified sampling enables you to reach conclusions about each strata in the frame. However, using a stratified sample requires that you can determine the variable(s) on which to base the stratification and can also be expensive to implement.

learnMORE

Learn how to select a stratified sample in the online section of this chapter.

Cluster Sample

In a **cluster sample**, you 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, households, or sales territories. You then take a random sample of one or more clusters and study all items in each selected cluster.

Cluster sampling is often more cost-effective than simple random sampling, particularly if the population is spread over a wide geographic region. However, cluster sampling often requires a larger sample size to produce results as precise as those from simple random sampling or stratified sampling. A detailed discussion of systematic sampling, stratified sampling, and cluster sampling procedures can be found in references 2, 4, and 5.

PROBLEMS FOR SECTION 1.3

LEARNING THE BASICS

1.17 For a population containing $N = 902$ individuals, what code number would you assign for

a. the first person on the list?
b. the fortieth person on the list?
c. the last person on the list?

1.18 For a population of $N = 902$, verify that by starting in row 05, column 01 of the table of random numbers (Table E.1), you need only six rows to select a sample of $N = 60$ *without* replacement.

1.19 Given a population of $N = 93$, starting in row 29, column 01 of the table of random numbers (Table E.1), and reading across the row, select a sample of $N = 15$

a. *without* replacement.
b. *with* replacement.

APPLYING THE CONCEPTS

1.20 For a study that consists of personal interviews with participants (rather than mail or phone surveys), explain why simple random sampling might be less practical than some other sampling methods.

1.21 You want to select a random sample of $n = 1$ from a population of three items (which are called *A*, *B*, and *C*). The rule for selecting the sample is as follows: Flip a coin; if it is heads, pick item *A*; if it is tails, flip the coin again; this time, if it is heads, choose *B*; if it is tails, choose *C*. Explain why this is a probability sample but not a simple random sample.

1.22 A population has four members (called *A*, *B*, *C*, and *D*). You would like to select a random sample of $n = 2$, which you decide to do in the following way: Flip a coin; if it is heads, the sample will be items *A* and *B*; if it is tails, the sample will be items *C* and *D*. Although this is a random sample, it is not a simple random sample. Explain why. (Compare the procedure described in Problem 1.21 with the procedure described in this problem.)

1.23 The registrar of a university with a population of $N = 4{,}000$ full-time students is asked by the president to conduct a survey to measure satisfaction with the quality of life on campus. The following table contains a breakdown of the 4,000 registered full-time students, by gender and class designation:

	CLASS DESIGNATION				
GENDER	**Fr.**	**So.**	**Jr.**	**Sr.**	**Total**
Female	700	520	500	480	2,200
Male	560	460	400	380	1,800
Total	1,260	980	900	860	4,000

The registrar intends to take a probability sample of $n = 200$ students and project the results from the sample to the entire population of full-time students.

a. If the frame available from the registrar's files is an alphabetical listing of the names of all $N = 4{,}000$ registered full-time students, what type of sample could you take? Discuss.

b. What is the advantage of selecting a simple random sample in (a)?

c. What is the advantage of selecting a systematic sample in (a)?

d. If the frame available from the registrar's files is a list of the names of all $N = 4{,}000$ registered full-time students compiled from eight separate alphabetical lists, based on the gender and class designation breakdowns shown in the class designation table, what type of sample should you take? Discuss.

e. Suppose that each of the $N = 4{,}000$ registered full-time students lived in one of the 10 campus dormitories. Each dormitory accommodates 400 students. It is college policy to fully integrate students by gender and class designation in each dormitory. If the registrar is able to compile a listing of all students by dormitory, explain how you could take a cluster sample.

✓SELF TEST **1.24** Prenumbered sales invoices are kept in a sales journal. The invoices are numbered from 0001 to 5000.

a. Beginning in row 16, column 01, and proceeding horizontally in a table of random numbers (Table E.1), select a simple random sample of 50 invoice numbers.

b. Select a systematic sample of 50 invoice numbers. Use the random numbers in row 20, columns 05–07, as the starting point for your selection.

c. Are the invoices selected in (a) the same as those selected in (b)? Why or why not?

1.25 Suppose that 10,000 customers in a retailer's customer database are categorized by three customer types: 3,500 prospective buyers, 4,500 first time buyers, and 2,000 repeat (loyal) buyers. A sample of 1,000 customers is needed.

a. What type of sampling should you do? Why?

b. Explain how you would carry out the sampling according to the method stated in (a).

c. Why is the sampling in (a) not simple random sampling?

1.4 Data Preparation

As you collect data, you may need to perform actions to help prepare the data for processing. This preparation both reviews the data for errors and inconsistencies as well as places the data into a format and structure required by the software you use to perform calculations (Microsoft Excel in this book).

Given the time constraints of the typical business statistics course, you will likely not practice data preparation tasks very much. With the exception of several examples designed for use with this section, data for the problems and examples in this book have already been properly "prepared" so that you can focus on the statistical concepts and methods that you are learning. That said, you should study these tasks for the occasions when you independently collect your own data and be generally knowledgeable about the importance of such tasks.

studentTIP

Data preparation considerations are also critical when using business analytics, as Chapter 17 explains.

Data Cleaning

Perhaps the most critical data preparation task is **data cleaning**, the finding and possible fixing of irregularities in the data you collect. Irregularities that you find include undefined or impossible values as well as so-called **missing values**, data that were not able to be collected for a variable (and therefore not available for analysis).

For a categorical variable, an undefined value would be a category that is not one of the categories defined for the variable. For a numerical variable, an impossible value would be a value that falls outside a defined range of possible values for the variable. For a numerical variable without a defined range of possible values, you might also find **outliers**, values that seem excessively different from most of the other values. Such values may or may not be errors, but they demand a second review.

Missing values most frequently arise when you record a *nonresponse* to a survey question, but also occur with other data sources for a variety of reasons. Proper statistical analysis requires that you exclude missing values, and most specialized statistical software has ways of performing that exclusion, a type of data cleaning, for you. However, Microsoft Excel does not. When you use Excel, you must find and then exclude missing values manually as part of your data cleaning.

You may be able to fix some typographical irregularities that can occur with the data for a categorical variable. For example, given the variable gender with the defined categories male and female, data improperly recorded as mail or mael or as MALE or maLe can all be reasonably fixed to male, so long as you preserve a copy of the original data for later audit purposes. For other undefined or impossible values that you cannot fix, you may be tempted to exclude the data, but such exclusions violate proper sampling methods. Instead, unfixable irregularities should be recorded as missing values.

learnMORE

Excel Data Cleaning.pdf discusses techniques that you can use in Excel to clean data.

Data Formatting

You may need to reformat your data when you collect your data. Reformatting can mean rearranging the structure of the data or changing the electronic encoding of the data or both. For example, suppose that you seek to collect financial data about a sample of companies. You might find these data structured as tables of data, as the contents of standard forms, in a continuous stock ticker stream, or as messages or blog entries that appear on various websites. These data sources have various levels of structure which affect the ease of reformatting them for use.

Because tables of data are highly structured and are similar to the structure of a worksheet, tables would require the least reformatting. In the best case, you could make the rows and columns of a table the rows and columns of a worksheet. Unstructured data sources, such as messages and blog entries, often represent the worst case. You may have to paraphrase or characterize the message contents in a way that does not involve a direct transfer. As the use of business analytics grows (see Chapter 17), the use of automated ways to paraphrase or characterize these and other types of unstructured data grows, too.

Independent of the structure, the data you collect may exist in an electronic form that needs to be changed in order to be analyzed. For example, data presented as a digital picture of Excel worksheets would need to be changed into an actual Excel worksheet before that data could be analyzed. In this example, you are changing the electronic encoding of all the data, from a picture format such as jpeg to an Excel workbook format such as xlsx. Sometimes, individual numerical values that you have collected may need to changed, especially if you collect values that result from a computational process. You can demonstrate this issue in Excel by entering a formula that is equivalent to the expression 1 × (0.5 – 0.4 – 0.1), which should evaluate as 0 but in Excel evaluates to a very small negative number. You would want to alter that value to 0 as part of your data cleaning.

Stacked and Unstacked Variables

When collecting data for a numerical variable, you sometimes want to subdivide that data into two or more groups for analysis. For example, if you were collecting data about the cost of a restaurant meal in an urban area, you might want to consider the cost of meals at restaurants in the center city district separately from the meal costs at metro area restaurants. When you want to consider two or more groups, you can arrange your data as either unstacked or stacked.

To use an **unstacked** arrangement, you create separate numerical variables for each group. In the example, you would create a center city meal cost variable and a second variable to hold the meal costs at metro area restaurants. To use a **stacked** arrangement format, you pair the single numerical variable meal cost with a second, categorical variable that contains two categories, such as center city and metro area. (The DATA and UNSTACKED worksheets of the Restaurants workbook present stacked and unstacked versions of the meal cost data that Chapter 2 uses for several examples.)

When you use software to analyze data, you may discover that a particular procedure requires data to be stacked (or unstacked). When such cases arise using Microsoft Excel, they are noted in the Excel Guide instructions. If you collect several numerical variables, each of which you want to subdivide in the same way, stacking your data will probably be the more efficient choice for you. Otherwise, it makes little difference whether your collected data is stacked or unstacked.

studentTIP

If you use PHStat, you can automate the stacking or unstacking of data as discussed in Section 1.4.

Recoding Variables

After you have collected data, you may need to reconsider the categories that you defined for a categorical variable or transform a numerical variable into a categorical variable by assigning the individual numeric values to one of several groups. In either case, you can define a **recoded variable** that supplements or replaces the original variable in your analysis.

For example, having already defined the variable class standing with the categories freshman, sophomore, junior, and senior, you decide that you want to investigate the differences between lowerclassmen (freshmen or sophomores) and upperclassmen (juniors or seniors). You can define a recoded variable UpperLower and assign the value Upper if a student is a junior or senior and assign the value Lower if the student is a freshman or sophomore.

When recoding variables, make sure that one and only one of the new categories can be assigned to any particular value being recoded and that each value can be recoded successfully by one of your new categories. You must ensure that your recoding has these properties of being **mutually exclusive** and **collectively exhaustive**.

When recoding numerical variables, pay particular attention to the operational definitions of the categories you create for the recoded variable, especially if the categories are not self-defining ranges. For example, while the recoded categories Under 12, 12–20, 21–34, 35–54, and 55-and-over are self-defining for age, the categories child, youth, young adult, middle aged, and senior each need to be further defined in terms of mutually exclusive and collectively exhaustive numerical ranges.

PROBLEMS FOR SECTION 1.4

APPLYING THE CONCEPTS

1.26 The cellphone brands owned by a sample of 20 respondents were:

Apple, Samsung, Appel, Nokia, Blackberry, HTC, Apple, , Samsung, HTC, LG, Blueberry, Samsung, Samsung, APPLE, Motorola, Apple, Samsun, Apple, Samsung

a. Clean these data and identify any irregularities in the data.
b. Are there any missing values in this set of 20 respondents? Identify the missing values.

1.27 The amount of monthly data usage by a sample of 10 cellphone users (in MB) was:

0.4, 2.7MB, 5.6, 4.3, 11.4, 26.8, 1.6, 1,079, 8.3, 4.2

Are there any potential irregularities in the data?

1.28 An amusement park company owns three hotels on an adjoining site. A guest relations manager wants to study the time it takes for shuttle buses to travel from each of the hotels to the amusement park entrance. Data were collected on a particular day that recorded the travel times in minutes.

a. Explain how the data could be organized in an unstacked format.
b. Explain how the data could be organized in a stacked format.

1.29 A hotel management company runs 10 hotels in a resort area. The hotels have a mix of pricing—some hotels have budget-priced rooms, some have moderate-priced rooms, and some have deluxe-priced rooms. Data are collected that indicate the number of rooms that are occupied at each hotel on each day of a month. Explain how the 10 hotels can be recoded into these three price categories.

1.5 Types of Survey Errors

When you collect data using the compiled responses from a survey, you must verify two things about the survey in order to make sure you have results that can be used in a decision-making process. You must evaluate the validity of the survey to make sure the survey does not lack objectivity or credibility. To do this, you evaluate the purpose of the survey, the reason the survey was conducted, and for whom the survey was conducted.

Having validated the objectivity and credibility of such a sample, you then determine if the survey was based on a probability sample (see Section 1.3). Surveys that use nonprobability samples are subject to serious biases that make their results useless for decision making purposes. In the case of the Coca-Cola managers concerned about the "Pepsi Challenge" results (see page 13), the managers failed to reflect on the subjective nature of the challenge as well as the nonprobability sample that this survey used. Had the mangers done so, they might not have been so quick to make the reformulation blunder that was reversed just weeks later.

Even when you verify these two things, surveys can suffer from any combination of the following types of survey errors: coverage error, nonresponse error, sampling error, or measurement error. Developers of well-designed surveys seek to reduce or minimize these types of errors, often at considerable cost.

Coverage Error

The key to proper sample selection is having an adequate frame. **Coverage error** occurs if certain groups of items are excluded from the frame so that they have no chance of being selected in the sample or if items are included from outside the frame. Coverage error results in a **selection bias**. If the frame is inadequate because certain groups of items in the population were not properly included, any probability sample selected will provide only an estimate of the characteristics of the frame, not the *actual* population.

Nonresponse Error

Not everyone is willing to respond to a survey. **Nonresponse error** arises from failure to collect data on all items in the sample and results in a **nonresponse bias**. Because you cannot always assume that persons who do not respond to surveys are similar to those who do, you need to follow up on the nonresponses after a specified period of time. You should make several attempts to convince such individuals to complete the survey and possibly offer an incentive to participate. The follow-up responses are then compared to the initial responses in order to make valid inferences from the survey (see references 2, 4, and 5). The mode of response you use, such as face-to-face interview, telephone interview, paper questionnaire, or computerized questionnaire, affects the rate of response. Personal interviews and telephone interviews usually produce a higher response rate than do mail surveys—but at a higher cost.

Sampling Error

When conducting a probability sample, chance dictates which individuals or items will or will not be included in the sample. **Sampling error** reflects the variation, or "chance differences," from sample to sample, based on the probability of particular individuals or items being selected in the particular samples.

When you read about the results of surveys or polls in newspapers or on the Internet, there is often a statement regarding a margin of error, such as "the results of this poll are expected to be within ±4 percentage points of the actual value." This **margin of error** is the sampling error. You can reduce sampling error by using larger sample sizes. Of course, doing so increases the cost of conducting the survey.

Measurement Error

In the practice of good survey research, you design surveys with the intention of gathering meaningful and accurate information. Unfortunately, the survey results you get are often only a proxy for the ones you really desire. Unlike height or weight, certain information about behaviors and psychological states is impossible or impractical to obtain directly.

When surveys rely on self-reported information, the mode of data collection, the respondent to the survey, and or the survey itself can be possible sources of **measurement error**. Satisficing, social desirability, reading ability, and/or interviewer effects can be dependent on the mode of data collection. The social desirability bias or cognitive/memory limitations of a respondent can affect the results. And vague questions, double-barreled questions that ask about multiple issues but require a single response, or questions that ask the respondent to report something that occurs over time but fail to clearly define the extent of time about which the question asks (the reference period) are some of the survey flaws that can cause errors.

To minimize measurement error, you need to standardize survey administration and respondent understanding of questions, but there are many barriers to this (see references 1, 3, and 10).

Ethical Issues About Surveys

Ethical considerations arise with respect to the four types of survey error. Coverage error can result in selection bias and becomes an ethical issue if particular groups or individuals are purposely excluded from the frame so that the survey results are more favorable to the survey's sponsor. Nonresponse error can lead to nonresponse bias and becomes an ethical issue if the sponsor knowingly designs the survey so that particular groups or individuals are less likely than others to respond. Sampling error becomes an ethical issue if the findings are purposely presented without reference to sample size and margin of error so that the sponsor can promote a viewpoint that might otherwise be inappropriate. Measurement error can become an ethical issue in one of three ways: (1) a survey sponsor chooses leading questions that guide the respondent in a particular direction; (2) an interviewer, through mannerisms and tone, purposely makes a respondent obligated to please the interviewer or otherwise guides the respondent in a particular direction; or (3) a respondent willfully provides false information.

Ethical issues also arise when the results of nonprobability samples are used to form conclusions about the entire population. When you use a nonprobability sampling method, you need to explain the sampling procedures and state that the results cannot be generalized beyond the sample.

CONSIDER THIS

New Media Surveys/Old Survey Errors

Software company executives decide to create a "customer experience improvement program" to record how customers use the company's products, with the goal of using the collected data to make product enhancements. Product marketers decide to use social media websites to collect consumer feedback. These people risk making the same type of survey error that led to the quick demise of a very successful magazine nearly 80 years ago.

By 1935, "straw polls" conducted by the magazine *Literary Digest* had successfully predicted five consecutive U.S. presidential elections. For the 1936 election, the magazine promised its largest poll ever and sent about 10 million ballots to people all across the country. After tabulating more than 2.3 million ballots, the *Digest* confidently proclaimed that Alf Landon would be an easy winner over Franklin D. Roosevelt. The actual results: FDR won in a landslide and Landon received the fewest electoral votes in U.S. history.

Being so wrong ruined the reputation of *Literary Digest* and it would cease publication less than two years after it made its erroneous claim. A review much later found that the low response rate (less than 25% of the ballots distributed were returned) and nonresponse error (Roosevelt voters were less likely to mail in a ballot than Landon voters) were significant reasons for the failure of the *Literary Digest* poll (see reference 9).

The *Literary Digest* error proved to be a watershed event in the history of sample surveys. First, the error disproved the assertion that the larger the sample is, the better the results will be—an assertion some people still mistakenly make today. The error paved the way for the modern methods of sampling discussed in this chapter and gave prominence to the more "scientific" methods that George Gallup and Elmo Roper both used to correctly predict the 1936 elections. (Today's Gallup Polls and Roper Reports remember those researchers.)

In more recent times, Microsoft software executives overlooked that experienced users could easily opt out of participating in their improvement program. This created another case of nonresponse error which may have led to the improved product (Microsoft Office) being so poorly received initially by experienced Office users who, by being more likely to opt out of the improvement program, biased the data that Microsoft used to determine Office "improvements."

And while those product marketers may be able to collect a lot of customer feedback data, those data also suffer from nonresponse error. In collecting data from social media websites, the marketers cannot know who chose *not* to leave comments. The marketers also cannot verify if the data collected suffer from a selection bias due to a coverage error

That you might use media newer than the mailed, dead-tree form that *Literary Digest* used does not mean that you automatically avoid the old survey errors. Just the opposite—the accessibility and reach of new media makes it much easier for unknowing people to commit such errors.

PROBLEMS FOR SECTION 1.5

APPLYING THE CONCEPTS

1.30 A survey indicates that the vast majority of college students own their own personal computers. What information would you want to know before you accepted the results of this survey?

1.31 A simple random sample of $n = 300$ full-time employees is selected from a company list containing the names of all $N = 5,000$ full-time employees in order to evaluate job satisfaction.

a. Give an example of possible coverage error.
b. Give an example of possible nonresponse error.
c. Give an example of possible sampling error.
d. Give an example of possible measurement error.

✓SELF TEST **1.32** Results of a 2015 Contact Solutions study reveal insights on perceptions and attitudes toward mobile shopping, providing direction to retailers for developing strategies for investing their dollars in the mobile app experience (**bit.ly/1OKdmqH**). Increased consumer interest in using shopping applications means retailers must adapt to meet the rising expectations for specialized mobile shopping experiences.

The results show that 23% of consumers indicate that in-app recommendations would drive them to add more items to their cart and that 33% would spend more time in the app. But shopper priorities change the moment they need help; they expect to get it immediately and effortlessly. If forced to stop what they're doing and leave the app to get help, 1 out of 4 shoppers would likely not make a purchase with the brand at all. The research is based on an online survey with a sample of 1,600 U.S. adults.

Identify *potential* concerns with coverage, nonresponse, sampling, and measurement errors.

1.33 A recent PwC survey of 1,322 CEOs in a wide range of industries representing a mix of company sizes from Asia, Europe, and the Americas indicated that CEOs no longer question the need to embrace technology at the core of their business in order to create value for customers (**pwc.to/1C5bG9D**). Eighty percent (80%) of CEOs see data mining and analysis as strategically important for their organization. Effectively leveraging data analytics tools, however, presents challenges for companies. Companies are not using analytics enough; there are issues about data quality, information overload, and a continuing lack of trust in the value of digital data. When companies do invest in digital technologies to deliver what customers want, that commitment would appear to pay off. CEOs are seeing the best return on digital investment in the area of operational efficiency; 88% percent think value has been created in this area.

What additional information would you want to know about the survey before you accepted the results for the study?

1.34 A recent survey points to growing consumer demand for state-of-the-art technology in vehicles. The 2015 KPMG Global Automotive Executive study found that automobile executives believe that consumers are still fixated on traditional product issues and not yet on innovative concepts and online services (**bit.ly/1GlkZ4Q**). Sixty-seven percent of executives rated fuel efficiency and 53% rated enhanced vehicle life span as extremely important to consumer vehicle purchases, while only 24% rated vehicle-bound connectivity and built-in technologies as extremely important. What additional information would you want to know about the survey before you accepted the results of the study?

USING STATISTICS
Defining Moments, Revisited

The New Coke and airline quality cases illustrate missteps that can occur during the define and collect tasks of the DCOVA framework. To use statistics effectively, you must properly define a business problem or goal and then collect data that will allow you to make observations and reach conclusions that are relevant to that problem or goal.

In the New Coke case, managers failed to consider that data collected about a taste test would not necessary provide useful information about the sales issues they faced. The managers also did not realize that the test used improper sampling techniques, deliberately introduced biases, and were subject to coverage and nonresponse errors. Those mistakes invalidated the test, making the conclusion that New Coke tasted better than Pepsi an invalid claim.

In the airline quality case, no mistakes in defining and collecting data were made. The results that fliers like quality was a valid one, but decision makers overlooked that quality was not the most significant factor for people buying seats on transatlantic flights (price was). This case illustrates that no matter how well you apply statistics, if you do not properly analyze the business problem or goal being considered, you may end up with valid results that lead you to invalid management decisions.

SUMMARY

In this chapter, you learned the details about the Define and Collect tasks of the DCOVA framework which are important first steps to applying statistics properly to decision making. You learned that defining variables means developing an operational definition that includes establishing the type of variable and the measurement scale that the variable uses. You learned important details about data collection as well some new basic vocabulary terms (sample, population, and parameter) and as a more precise definition of statistic. You specifically learned about sampling and the types of sampling methods available to you. Finally, you surveyed data preparation considerations and learned about the type of survey errors you can encounter.

REFERENCES

1. Biemer, P. B., R. M. Graves, L. E. Lyberg, A. Mathiowetz, and S. Sudman. *Measurement Errors in Surveys*. New York: Wiley Interscience, 2004.

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

3. Fowler, F. J. *Improving Survey Questions: Design and Evaluation*, *Applied Special Research Methods Series*, Vol. 38, Thousand Oaks, CA: Sage Publications, 1995.

4. Groves R. M., F. J. Fowler, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey Methodology*, 2nd ed. New York: John Wiley, 2009.

5. Lohr, S. L. *Sampling Design and Analysis*, 2nd ed. Boston, MA: Brooks/Cole Cengage Learning, 2010.

6. Polaris Marketing Research. “Brilliant Marketing Research or What? The New Coke Story.” **bit.ly/1DofHSM**, posted 20 Sep 2011.

7. Rosenbaum, D. “The New Big Data Magic.” CFO.com, 29 Aug 2011, **bit.ly/1DUMWzv**

8. Osbourne, J. *Best Practices in Data Cleaning*. Thousand Oaks, CA: Sage Publications, 2012.

9. Squire, P. “Why the 1936 *Literary Digest* Poll Failed.” *Public Opinion Quarterly* 52 (1988): 125–133.

10. Sudman, S., N. M. Bradburn, and N. Schwarz. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass, 1993.

KEY TERMS

categorical variable 14
cluster 19
cluster sample 19
collectively exhaustive 22
continuous variable 14
convenience sample 17
coverage error 23
data cleaning 20
discrete variable 14

frame 17
interval scale 14
judgment sample 17
margin of error 23
measurement error 24
measurement scale 14
missing value 20
mutually exclusive 22
nominal scale 14
nonprobability sample 17
nonresponse bias 23
nonresponse error 23
numerical variable 14
operational definition 13
ordinal scale 14
outlier 21
parameter 16
population 16
primary data source 16
probability sample 17
qualitative variable 14
quantitative variable 14
ratio scale 14
recoded variable 22
sample 16
sampling error 23
sampling with replacement 18
sampling without replacement 18
secondary data source 16
selection bias 23
simple random sample 18
stacked 21
statistic 16
strata 19
stratified sample 19
systematic sample 18
table of random numbers 18
treatment 16
unstacked 21

CHECKING YOUR UNDERSTANDING

1.35 What is the difference between a sample and a population?

1.36 What is the difference between a statistic and a parameter?

1.37 What is the difference between a categorical variable and a numerical variable?

1.38 What is the difference between a discrete numerical variable and a continuous numerical variable?

1.39 What is the difference between a nominal scaled variable and an ordinal scaled variable?

1.40 What is the difference between an interval scaled variable and a ratio scaled variable?

1.41 What is the difference between probability sampling and nonprobability sampling?

1.42 What is the difference between a missing value and an outlier?

1.43 What is the difference between unstack and stacked variables?

CHAPTER REVIEW PROBLEMS

1.44 Visit the official website for Microsoft Excel, **products.office.com/excel**. Review the features of Excel and then state the ways the program could be useful in statistical analysis.

1.45 Results of a 2015 Contact Solutions study reveal insights on perceptions and attitudes toward mobile shopping, providing direction to retailers for developing strategies for investing their dollars in the mobile app experience (**bit.ly/1OKdmqH**). Increased consumer interest in using shopping applications means retailers must adapt to meet the rising expectations for specialized mobile shopping experiences.

The study results show that 23% of consumers indicate that in-app recommendations would drive them to add more items to their cart and that 33% would spend more time in the app. But shopper priorities change the moment they need help; they expect to get it immediately and effortlessly. If forced to stop what they're doing and leave the app to get help, 1 out of 4 shoppers would likely not make a purchase with the brand at all. The research is based on an online survey with a sample of 1,600 U.S. adults who have made purchases via their mobile device.

a. Describe the population of interest.
b. Describe the sample that was collected.
c. Describe a parameter of interest.
d. Describe the statistic used to estimate the parameter in (c).

1.46 The Gallup organization releases the results of recent polls at its website, **www.gallup.com**. Visit this site and read an article of interest.

a. Describe the population of interest.
b. Describe the sample that was collected.
c. Describe a parameter of interest.
d. Describe the statistic used to estimate the parameter in (c).

1.47 A recent PwC survey of 1,322 CEOs in a wide range of industries representing a mix of company sizes from Asia, Europe, and the Americas indicated that CEOs no longer question the need to embrace technology at the core of their business in order to create value for customers (**pwc.to/1C5bG9D**). Eighty percent (80%) of CEOs see data mining and analysis as strategically important for their organization. Effectively leveraging data analytics tools, however, presents challenges for companies. Companies are not using analytics enough; there are issues about data quality, information overload, and a continuing lack of trust in the value of digital data.

When companies do invest in digital technologies to deliver what customers want, that commitment would appear to pay off. CEOs are seeing the best return on digital investment in the area of operational efficiency; 88% percent think value has been created in this area.

a. Describe the population of interest.
b. Describe the sample that was collected.

c. Describe a parameter of interest.
d. Describe the statistic used to estimate the parameter in (c).

1.48 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. Describe a variable for which data is collected.
b. Is the variable categorical or numerical?
c. If the variable is numerical, is it discrete or continuous?

1.49 Download and examine Zarca Interactive's *Association Salary Survey*, available on Zarca's "Sample Surveys, Online Survey Example, Sample Association Surveys, Sample Customer Service Survey" web page, **www.zarca.com/Online-Survey-Resource/Sample-Surveys.html**.
a. Give an example of a categorical variable included in the survey.
b. Give an example of a numerical variable included in the survey.

1.50 Three professors examined awareness of four widely disseminated retirement rules among employees at the University of Utah. These rules provide simple answers to questions about retirement planning (R. N. Mayer, C. D. Zick, and M. Glaittle, "Public Awareness of Retirement Planning Rules of Thumb," *Journal of Personal Finance*, 2011 10(1), 12–35). At the time of the investigation, there were approximately 10,000 benefited employees, and 3,095 participated in the study. Demographic data collected on these 3,095 employees included gender, age (years), education level (years completed), marital status, household income ($), and employment category.
a. Describe the population of interest.
b. Describe the sample that was collected.
c. Indicate whether each of the demographic variables mentioned is categorical or numerical.

1.51 Social media provides an enormous amount of data about the activities and habits of people using social platforms like Facebook and Twitter. The belief is that mining that data provides a treasure trove for those who seek to quantify and predict future human behavior. A marketer is planning a survey of Internet users in the United States to determine social media usage. The objective of the survey is to gain insight on these three items: key social media platforms used, frequency of social media usage, and demographics of key social media platform users.
a. For each of the three items listed, indicate whether the variables are categorical or numerical. If a variable is numerical, is it discrete or continuous?
b. Develop five categorical questions for the survey.
c. Develop five numerical questions for the survey.

CHAPTER 1 CASES

Managing Ashland MultiComm Services

Ashland MultiComm Services (AMS) provides high-quality telecommunications services in the Greater Ashland area. AMS traces its roots to a small company that redistributed the broadcast television signals from nearby major metropolitan areas but has evolved into a provider of a wide range of broadband services for residential customers.

AMS offers subscription-based services for digital cable television, local and long-distance telephone services, and high-speed Internet access. Recently, AMS has faced competition from other service providers as well as Internet-based, on demand streaming services that have caused many customers to "cut the cable" and drop their subscription to cable video services.

AMS management believes that a combination of increased promotional expenditures, adjustment in subscription fees, and improved customer service will allow AMS to successfully face these challenges. To help determine the proper mix of strategies to be taken, AMS management has decided to organize a research team to undertake a study.

The managers suggest that the research team examine the company's own historical data for number of subscribers, revenues, and subscription renewal rates for the past few years. They direct the team to examine year-to-date data as well, as the managers suspect that some of the changes they have seen have been a relatively recent phenomena.

1. What type of data source would the company's own historical data be? Identify other possible data sources that the research team might use to examine the current marketplace for residential broadband services in a city such as Ashland.

2. What type of data collection techniques might the team employ?

3. In their suggestions and directions, the AMS managers have named a number of possible variables to study, but offered no operational definitions for those variables. What types of possible misunderstandings could arise if the team and managers do not first properly define each variable cited?

CardioGood Fitness

CardioGood Fitness is a developer of high-quality cardiovascular exercise equipment. Its products include treadmills, fitness bikes, elliptical machines, and e-glides. CardioGood Fitness looks to increase the sales of its treadmill products and has hired The AdRight Agency, a small advertising firm, to create and implement an advertising program. The AdRight Agency plans to identify particular market segments that are most likely to buy their clients' goods and services and then locate advertising outlets that will reach that market group. This activity includes collecting data on clients' actual sales and on the customers who make the purchases, with the goal of determining whether there is a distinct profile of the typical customer for a particular product or service. If a distinct profile emerges, efforts are made to match that profile to advertising outlets known to reflect the particular profile, thus targeting advertising directly to high-potential customers.

CardioGood Fitness sells three different lines of treadmills. The TM195 is an entry-level treadmill. It is as dependable as other models offered by CardioGood Fitness, but with fewer programs and features. It is suitable for individuals who thrive on minimal programming and the desire for simplicity to initiate their walk or hike. The TM195 sells for $1,500.

The middle-line TM498 adds to the features of the entry-level model two user programs and up to 15% elevation upgrade. The TM498 is suitable for individuals who are walkers at a transitional stage from walking to running or midlevel runners. The TM498 sells for $1,750.

The top-of-the-line TM798 is structurally larger and heavier and has more features than the other models. Its unique features include a bright blue backlit LCD console, quick speed and incline keys, a wireless heart rate monitor with a telemetric chest strap, remote speed and incline controls, and an anatomical figure that specifies which muscles are minimally and maximally activated. This model features a nonfolding platform base that is designed to handle rigorous, frequent running; the TM798 is therefore appealing to someone who is a power walker or a runner. The selling price is $2,500.

As a first step, the market research team at AdRight is assigned the task of identifying the profile of the typical customer for each treadmill product offered by CardioGood Fitness. The market research team decides to investigate whether there are differences across the product lines with respect to customer characteristics. The team decides to collect data on individuals who purchased a treadmill at a CardioGood Fitness retail store during the prior three months.

The team decides to use both business transactional data and the results of a personal profile survey that every purchaser completes as their sources of data. The team identifies the following customer variables to study: product purchased—TM195, TM498, or TM798; gender; age, in years; education, in years; relationship status, single or partnered; annual household income ($); mean number of times the customer plans to use the treadmill each week; mean number of miles the customer expects to walk/run each week; and self-rated fitness on a 1-to-5 scale, where 1 is poor shape and 5 is excellent shape. For this set of variables:

1. Which variables in the survey are categorical?
2. Which variables in the survey are numerical?
3. Which variables are discrete numerical variables?

Clear Mountain State Student Survey

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students who attend CMSU. They create and distribute a survey of 14 questions and receive responses from 111 undergraduates (stored in StudentSurvey).

Download (see Appendix C) and review the survey document **CMUndergradSurvey.pdf**. For each question asked in the survey, determine whether the variable is categorical or numerical. If you determine that the variable is numerical, identify whether it is discrete or continuous.

Learning with the Digital Cases

Identifying and preventing misuses of statistics is an important responsibility for all managers. The Digital Cases allow you to practice the skills necessary for this important task.

Each chapter's Digital Case tests your understanding of how to apply an important statistical concept taught in the chapter. As in many business situations, not all of the information you encounter will be relevant to your task, and you may occasionally discover conflicting information that you have to resolve in order to complete the case.

To assist your learning, each Digital Case begins with a learning objective and a summary of the problem or issue at hand. Each case directs you to the information necessary to reach your own conclusions and to answer the case questions. Many cases, such as the sample case worked out next, extend a chapter's Using Statistics scenario. You can download digital case files which are PDF format documents that may contain extended features as interactivity or data file attachments. Open these files with a current version of Adobe Reader, as other PDF programs may not support the extended features. (For more information, see Appendix C.)

To illustrate learning with a Digital Case, open the Digital Case file **WhitneyWireless.pdf** that contains summary information about the Whitney Wireless business. Apparently, from the claim on the title page, this business is celebrating its "best sales year ever."

Review the **Who We Are**, **What We Do**, and **What We Plan to Do** sections on the second page. Do these sections contain any useful information? What *questions* does this passage raise? Did you notice that while many facts are presented, no data that would support the claim of "best sales year ever" are presented? And were those mobile "mobilemobiles" used solely for promotion? Or did they generate any sales? Do you think that a talk-with-your-mouth-full event, however novel, would be a success?

Continue to the third page and the **Our Best Sales Year Ever!** section. How would you support such a claim? With a table of numbers? Remarks attributed to a knowledgeable source? Whitney Wireless has used a chart to present "two years ago" and "latest twelve months" sales data by category. Are there any problems with what the company has done? *Absolutely!*

Take a moment to identify and reflect on those problems. Then turn to pages 4 though 6 that present an annotated version of the first three pages and discusses some of the problems with this document.

In subsequent Digital Cases, you will be asked to provide this type of analysis, using the open-ended case questions as your guide. Not all the cases are as straightforward as this example, and some cases include perfectly appropriate applications of statistical methods. And none have annotated answers!

CHAPTER 1 EXCEL GUIDE

EG1.1 DEFINING VARIABLES

Classifying Variables by Type

Microsoft Excel infers the variable type from the data you enter into a column. If Excel discovers a column that contains numbers, it treats the column as a numerical variable. If Excel discovers a column that contains words or alphanumeric entries, it treats the column as a non-numerical (categorical) variable.

This imperfect method works most of the time, especially if you make sure that the categories for your categorical variables are words or phrases such as "yes" and "no." However, because you cannot explicitly define the variable type, Excel can mistakenly offer or allow you to do nonsensical things such as using a statistical method that is designed for numerical variables on categorical variables. If you must use coded values such as 1, 2, or 3, enter them preceded with an apostrophe, as Excel treats all values that begin with an apostrophe as non-numerical data. (You can check whether a cell entry includes a leading apostrophe by selecting a cell and viewing the contents of the cell in the formula bar.)

EG1.2 COLLECTING DATA

There are no Excel Guide instructions for this section.

EG1.3 TYPES OF SAMPLING METHODS

Simple Random Sample

Key Technique Use the **RANDBETWEEN(*smallest integer, largest integer*)** function to generate a random integer that can then be used to select an item from a frame.

Example 1 Create a simple random sample *with* replacement of size 40 from a population of 800 items.

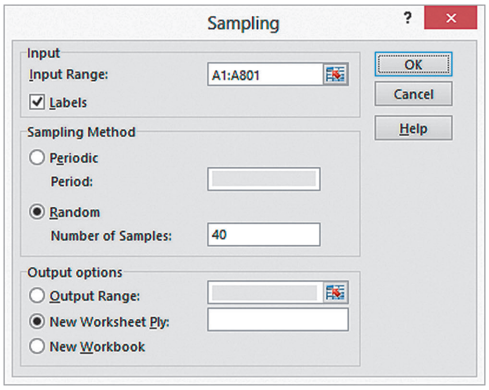
Workbook Enter a formula that uses this function and then copy the formula down a column for as many rows as is necessary. For example, to create a simple random sample with replacement of size 40 from a population of 800 items, open to a new worksheet. Enter **Sample** in cell **A1** and enter the formula **=RANDBETWEEN(1, 800)** in cell **A2**. Then copy the formula down the column to cell **A41**.

Excel contains no functions to select a random sample *without* replacement. Such samples are most easily created using an add-in such as PHStat or the Analysis ToolPak, as described in the following paragraphs.

Analysis ToolPak Use **Sampling** to create a random sample *with replacement*.

For the example, open to the worksheet that contains the population of 800 items in column A and that contains a column heading in cell A1. Select **Data ➔ Data Analysis**. In the Data Analysis dialog box, select **Sampling** from the **Analysis Tools** list and then click **OK**. In the procedure's dialog box (shown below):

1. Enter **A1:A801** as the **Input Range** and check **Labels**.
2. Click **Random** and enter **40** as the **Number of Samples**.
3. Click **New Worksheet Ply** and then click **OK**.



Sampling

Input

Input Range: A1:A801

☑ Labels

Sampling Method

○ Periodic

Period:

● Random

Number of Samples: 40

Output options

○ Output Range:

● New Worksheet Ply:

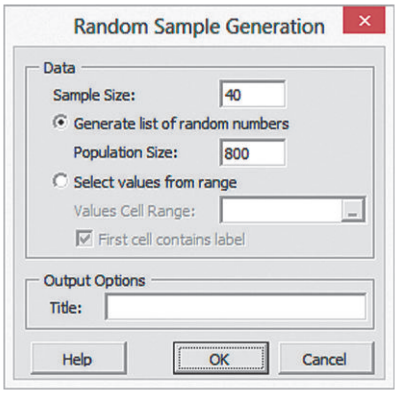
○ New Workbook

Example 2 Create a simple random sample *without* replacement of size 40 from a population of 800 items.

PHStat Use **Random Sample Generation**.

For the example, select **PHStat ➔ Sampling ➔ Random Sample Generation**. In the procedure's dialog box (shown in next column):

1. Enter **40** as the **Sample Size**.
2. Click **Generate list of random numbers** and enter **800** as the **Population Size**.
3. Enter a **Title** and click **OK**.



Random Sample Generation

Data

Sample Size: 40

● Generate list of random numbers

Population Size: 800

○ Select values from range

Values Cell Range:

☑ First cell contains label

Output Options

Title:

Unlike most other PHStat results worksheets, the worksheet created contains no formulas.