

2ND EDITION

# BIOSTATISTICS

for the Biological and Health Sciences



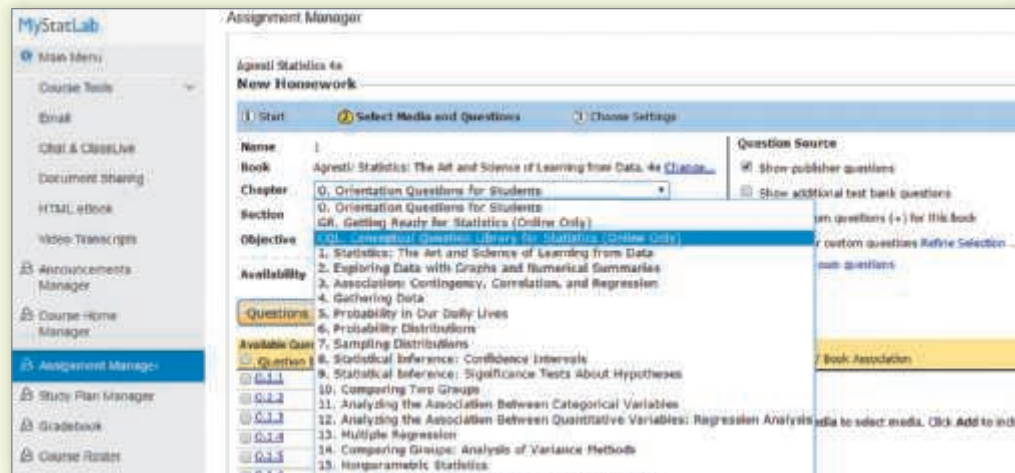
TRIOLA | TRIOLA | ROY



Pearson

## Conceptual Understanding

Students need to be equipped with both the methods and conceptual understanding of statistics. MyStatLab offers a full question library of over 1,000 conceptual-based questions to help tighten the comprehension of statistical concepts.



## Real-World Statistics

MyStatLab video resources help foster conceptual understanding. StatTalk Videos, hosted by fun-loving statistician, Andrew Vickers, demonstrate important statistical concepts through interesting stories and real-life events. This series of 24 videos includes assignable questions built in MyStatLab and an instructor's guide.



Visit [www.mystatlab.com](http://www.mystatlab.com) and click Get Trained to make sure you're getting the most out of MyStatLab.

SECOND EDITION

# BIOSTATISTICS

FOR THE BIOLOGICAL  
AND HEALTH SCIENCES

**MARC M. TRIOLA, MD, FACP**

New York University School of Medicine

**MARIO F. TRIOLA**

Dutchess Community College

**JASON ROY, PHD**

University of Pennsylvania  
Perelman School of Medicine



*To Ginny  
Dushana and Marisa  
Trevor and Mitchell*

*Director, Portfolio Management* Deirdre Lynch  
*Senior Portfolio Manager* Suzy Bainbridge  
*Portfolio Management Assistant* Justin Billing  
*Content Producer* Peggy McMahon  
*Managing Producer* Karen Wernholm  
*Courseware QA Manager* Mary Durnwald  
*Senior Producer* Vicki Dreyfus  
*Product Marketing Manager* Yvonne Vannatta  
*Field Marketing Manager* Evan St. Cyr

*Product Marketing Assistant* Jennifer Myers  
*Field Marketing Assistant* Erin Rush  
*Senior Author Support/Technology Specialist* Joe Vetere  
*Manager, Rights and Permissions* Gina M. Cheselka  
*Text and Cover Design, Illustrations, Production Coordination,  
Composition* Cenveo Publisher Services  
*Cover Image* Robert Essel NYC/Getty Images

Copyright © 2018, 2006 by Pearson Education, Inc. All Rights Reserved. Printed in the United States of America. This publication is protected by copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise. For information regarding permissions, request forms and the appropriate contacts within the Pearson Education Global Rights & Permissions department, please visit [www.pearsoned.com/permissions/](http://www.pearsoned.com/permissions/).

Attributions of third party content appear on page 683–684, which constitutes an extension of this copyright page.

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

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

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

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

**Library of Congress Cataloging-in-Publication Data**

Names: Triola, Marc M. | Triola, Mario F. | Roy, Jason (Jason Allen)

Title: Biostatistics for the biological and health sciences.

Description: Second edition / Marc M. Triola, New York University,  
Mario F. Triola, Dutchess Community College, Jason Roy, University of  
Pennsylvania. | Boston : Pearson, [2018] | Includes bibliographical  
references and index.

Identifiers: LCCN 2016016759 | ISBN 9780134039015 (hardcover) | ISBN  
0134039017 (hardcover)

Subjects: LCSH: Biometry. | Medical statistics.

Classification: LCC QH323.5 .T75 2018 | DDC 570.1/5195–dc23

LC record available at <https://lccn.loc.gov/2016016759>

1 16



ISBN 13: 978-0-13-403901-5

ISBN 10: 0-13-403901-7

# ABOUT THE AUTHORS



**Marc Triola, MD, FACP** is the Associate Dean for Educational Informatics at NYU School of Medicine, the founding director of the NYU Langone Medical Center Institute for Innovations in Medical Education (IIME), and an Associate Professor of Medicine. Dr. Triola's research experience and expertise focus on the disruptive effects of the present revolution in education, driven by technological advances, big data, and learning analytics. Dr. Triola has worked to create a “learning

ecosystem” that includes interconnected computer-based e-learning tools and new ways to effectively integrate growing amounts of electronic data in educational research. Dr. Triola and IIME have been funded by the National Institutes of Health, the Integrated Advanced Information Management Systems program, the National Science Foundation Advanced Learning Technologies program, the Josiah Macy, Jr. Foundation, the U.S. Department of Education, and the American Medical Association Accelerating Change in Medical Education program. He chairs numerous committees at the state and national levels focused on the future of health professions educational technology development and research.



**Mario F. Triola** is a Professor Emeritus of Mathematics at Dutchess Community College, where he has taught statistics for over 30 years. Marty is the author of *Elementary Statistics*, 13th edition, *Essentials of Statistics*, 5th edition, *Elementary Statistics Using Excel*, 6th edition, and *Elementary Statistics Using the TI-83/84 Plus Calculator*, 4th edition, and he is a co-author of *Statistical Reasoning for Everyday Life*, 5th edition. *Elementary Statistics* is currently available as an

International Edition, and it has been translated into several foreign languages. Marty designed the original Statdisk statistical software, and he has written several manuals and workbooks for technology supporting statistics education.

He has been a speaker at many conferences and colleges. Marty's consulting work includes the design of casino slot machines and the design of fishing rods. He has worked with attorneys in determining probabilities in paternity lawsuits, analyzing data in medical malpractice lawsuits, identifying salary inequities based on gender, and analyzing disputed election results. He has also used statistical methods in analyzing medical school surveys and in analyzing survey results for the New York City Transit Authority. Marty has testified as an expert witness in the New York State Supreme Court.



**Jason Roy, PhD**, is Associate Professor of Biostatistics in the Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania. He received his PhD in Biostatistics in 2000 from the University of Michigan. He was recipient of the 2002 David P. Byar Young Investigator Award from the American Statistical Association Biometrics Section. His statistical research interests are in the areas of causal inference, missing data, and prediction

modeling. He is especially interested in the statistical challenges with analyzing data from large health care databases. He collaborates in many different disease areas, including chronic kidney disease, cardiovascular disease, and liver diseases. Dr Roy is Associate Editor of *Biometrics*, *Journal of the American Statistical Association*, and *Pharmacoepidemiology & Drug Safety*, and has over 90 peer-reviewed publications.

# CONTENTS

<b>1</b>	<b>INTRODUCTION TO STATISTICS</b>	<b>1</b>
	1-1 Statistical and Critical Thinking 4	
	1-2 Types of Data 13	
	1-3 Collecting Sample Data 24	
<b>2</b>	<b>EXPLORING DATA WITH TABLES AND GRAPHS</b>	<b>40</b>
	2-1 Frequency Distributions for Organizing and Summarizing Data 42	
	2-2 Histograms 51	
	2-3 Graphs That Enlighten and Graphs That Deceive 56	
	2-4 Scatterplots, Correlation, and Regression 65	
<b>3</b>	<b>DESCRIBING, EXPLORING, AND COMPARING DATA</b>	<b>75</b>
	3-1 Measures of Center 77	
	3-2 Measures of Variation 89	
	3-3 Measures of Relative Standing and Boxplots 102	
<b>4</b>	<b>PROBABILITY</b>	<b>118</b>
	4-1 Basic Concepts of Probability 120	
	4-2 Addition Rule and Multiplication Rule 131	
	4-3 Complements, Conditional Probability, and Bayes' Theorem 144	
	4-4 Risks and Odds 153	
	4-5 Rates of Mortality, Fertility, and Morbidity 162	
	4-6 Counting 167	
<b>5</b>	<b>DISCRETE PROBABILITY DISTRIBUTIONS</b>	<b>180</b>
	5-1 Probability Distributions 182	
	5-2 Binomial Probability Distributions 193	
	5-3 Poisson Probability Distributions 206	
<b>6</b>	<b>NORMAL PROBABILITY DISTRIBUTIONS</b>	<b>216</b>
	6-1 The Standard Normal Distribution 218	
	6-2 Real Applications of Normal Distributions 231	
	6-3 Sampling Distributions and Estimators 241	
	6-4 The Central Limit Theorem 252	
	6-5 Assessing Normality 261	
	6-6 Normal as Approximation to Binomial 269	
<b>7</b>	<b>ESTIMATING PARAMETERS AND DETERMINING SAMPLE SIZES</b>	<b>282</b>
	7-1 Estimating a Population Proportion 284	
	7-2 Estimating a Population Mean 299	
	7-3 Estimating a Population Standard Deviation or Variance 315	
	7-4 Bootstrapping: Using Technology for Estimates 324	
<b>8</b>	<b>HYPOTHESIS TESTING</b>	<b>336</b>
	8-1 Basics of Hypothesis Testing 338	
	8-2 Testing a Claim About a Proportion 354	
	8-3 Testing a Claim About a Mean 366	
	8-4 Testing a Claim About a Standard Deviation or Variance 377	
<b>9</b>	<b>INFERENCES FROM TWO SAMPLES</b>	<b>392</b>
	9-1 Two Proportions 394	
	9-2 Two Means: Independent Samples 406	
	9-3 Two Dependent Samples (Matched Pairs) 418	
	9-4 Two Variances or Standard Deviations 428	

<b>10</b>	<b>CORRELATION AND REGRESSION</b>	<b>442</b>
	10-1 Correlation 444	
	10-2 Regression 462	
	10-3 Prediction Intervals and Variation 474	
	10-4 Multiple Regression 481	
	10-5 Dummy Variables and Logistic Regression 489	
<b>11</b>	<b>GOODNESS-OF-FIT AND CONTINGENCY TABLES</b>	<b>502</b>
	11-1 Goodness-of-Fit 503	
	11-2 Contingency Tables 514	
<b>12</b>	<b>ANALYSIS OF VARIANCE</b>	<b>531</b>
	12-1 One-Way ANOVA 533	
	12-2 Two-Way ANOVA 547	
<b>13</b>	<b>NONPARAMETRIC TESTS</b>	<b>560</b>
	13-1 Basics of Nonparametric Tests 562	
	13-2 Sign Test 564	
	13-3 Wilcoxon Signed-Ranks Test for Matched Pairs 575	
	13-4 Wilcoxon Rank-Sum Test for Two Independent Samples 581	
	13-5 Kruskal-Wallis Test for Three or More Samples 586	
	13-6 Rank Correlation 592	
<b>14</b>	<b>SURVIVAL ANALYSIS</b>	<b>603</b>
	14-1 Life Tables 604	
	14-2 Kaplan-Meier Survival Analysis 614	
<b>APPENDIX A</b>	<b>TABLES</b>	<b>625</b>
<b>APPENDIX B</b>	<b>DATA SETS</b>	<b>638</b>
<b>APPENDIX C</b>	<b>WEBSITES AND BIBLIOGRAPHY OF BOOKS</b>	<b>645</b>
<b>APPENDIX D</b>	<b>ANSWERS TO ODD-NUMBERED SECTION EXERCISES</b>	<b>646</b>
	(and all Quick Quizzes, all Review Exercises, and all Cumulative Review Exercises)	
	<b>Credits 683</b>	
	<b>Index 685</b>	



# PREFACE

Statistics permeates nearly every aspect of our lives, and its role has become particularly important in the biological, life, medical, and health sciences. From opinion polls to clinical trials in medicine and analysis of big data from health applications, statistics influences and shapes the world around us. *Biostatistics for the Health and Biological Sciences* forges the relationship between statistics and our world through extensive use of a wide variety of real applications that bring life to theory and methods.

## Goals of This Second Edition

- Incorporate the latest and best methods used by professional statisticians.
- Include features that address all of the recommendations included in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE)* as recommended by the American Statistical Association.
- Provide an abundance of new and interesting data sets, examples, and exercises.
- Foster personal growth of students through critical thinking, use of technology, collaborative work, and development of communication skills.
- Enhance teaching and learning with the most extensive and best set of supplements and digital resources.


## Audience / Prerequisites

*Biostatistics for the Health and Biological Sciences* is written for students majoring in the biological and health sciences, and it is designed for a wide variety of students taking their first statistics course. Algebra is used minimally, and calculus is not required. It is recommended that students have completed at least an elementary algebra course or that students should learn the relevant algebra components through an integrated or co-requisite course. In many cases, underlying theory is included, but this book does not require the mathematical rigor more appropriate for mathematics majors.

## Hallmark Features

Great care has been taken to ensure that each chapter of *Biostatistics for the Health and Biological Sciences* will help students understand the concepts presented. The following features are designed to help meet that objective.

### Real Data

Hundreds of hours have been devoted to finding data that are real, meaningful, and interesting to students. Fully 87% of the examples are based on real data, and 89% of the exercises are based on real data. Some exercises refer to the 18 data sets listed in Appendix B, and 12 of those data sets are new to this edition. Exercises requiring use of the Appendix B data sets are located toward the end of each exercise set and are marked with a special data set icon .

Real data sets are included throughout the book to provide relevant and interesting real-world statistical applications, including biometric security, body measurements, brain sizes and IQ scores, and data from births. Appendix B includes descriptions of

the 18 data sets that can be downloaded from the companion website [www.pearson-highered.com/triola](http://www.pearson-highered.com/triola), the author maintained [www.TriolaStats.com](http://www.TriolaStats.com) and MyStatLab.

TriolaStats.com includes downloadable data sets in formats for technologies including Excel, Minitab, JMP, SPSS, and TI-83/84 Plus calculators. The data sets are also included in the free Statdisk software, which is also available on the website.

### Readability

Great care, enthusiasm, and passion have been devoted to creating a book that is readable, understandable, interesting, and relevant. Students pursuing any major in the biological, life, medical, or health fields are sure to find applications related to their future work.

### Website

This textbook is supported by [www.TriolaStats.com](http://www.TriolaStats.com), and [www.pearsonhighered.com/triola](http://www.pearsonhighered.com/triola) which are continually updated to provide the latest digital resources, including:

- Statdisk: A free, robust statistical software package designed for this book.
- Downloadable Appendix B data sets in a variety of technology formats.
- Downloadable textbook supplements including *Glossary of Statistical Terms* and *Formulas and Tables*.
- Online instructional videos created specifically for this book that provide step-by-step technology instructions.
- Triola Blog, which highlights current applications of statistics, statistics in the news, and online resources.

### Chapter Features

#### *Chapter Opening Features*

- Chapters begin with a *Chapter Problem* that uses real data and motivates the chapter material.
- *Chapter Objectives* provide a summary of key learning goals for each section in the chapter.

#### *Exercises*

Many exercises require the *interpretation* of results. Great care has been taken to ensure their usefulness, relevance, and accuracy. Exercises are arranged in order of increasing difficulty, and they begin with *Basic Skills and Concepts*. Most sections include additional *Beyond the Basics* exercises that address more difficult concepts or require a stronger mathematical background. In a few cases, these exercises introduce a new concept.

#### *End-of-Chapter Features*

- *Chapter Quick Quiz* provides review questions that require brief answers.
- *Review Exercises* offer practice on the chapter concepts and procedures.
- *Cumulative Review Exercises* reinforce earlier material.
- *Technology Project* provides an activity that can be used with a variety of technologies.
- *From Data to Decision* is a capstone problem that requires critical thinking and writing.
- *Cooperative Group Activities* encourage active learning in groups.

### Other Features

**Margin Essays** There are 57 margin essays designed to highlight real-world topics and foster student interest.

**Flowcharts** The text includes flowcharts that simplify and clarify more complex concepts and procedures. Animated versions of the text's flowcharts are available within MyStatLab and MathXL.

**Quick-Reference Endpapers** Tables A-2 and A-3 (the normal and  $t$  distributions) are reproduced on the rear inside cover pages.

**Detachable Formula and Table Card** This insert, organized by chapter, gives students a quick reference for studying, or for use when taking tests (if allowed by the instructor). It also includes the most commonly used tables. This is also available for download at [www.TriolaStats.com](http://www.TriolaStats.com), [www.pearsonhighered.com/triola](http://www.pearsonhighered.com/triola) and in MyStatLab.

### Technology Integration

As in the preceding edition, there are many displays of screens from technology throughout the book, and some exercises are based on displayed results from technology. Where appropriate, sections include a reference to an online *Tech Center* subsection that includes detailed instructions for Statdisk, Minitab®, Excel®, StatCrunch, or a TI-83/84 Plus® calculator. (Throughout this text, “TI-83/84 Plus” is used to identify a TI-83 Plus or TI-84 Plus calculator). The end-of-chapter features include a *Technology Project*.

The Statdisk statistical software package is designed specifically for this textbook and contains all Appendix B data sets. Statdisk is free to users of this book, and it can be downloaded at [www.statdisk.org](http://www.statdisk.org).

## Changes in This Edition

### New Features

**Chapter Objectives** provide a summary of key learning goals for each section in the chapter.

**Larger Data Sets:** Some of the data sets in Appendix B are much larger than in the previous edition. It is no longer practical to print all of the Appendix B data sets in this book, so the data sets are *described* in Appendix B, and they can be downloaded at [www.TriolaStats.com](http://www.TriolaStats.com), [www.pearsonhighered.com/triola](http://www.pearsonhighered.com/triola), and MyStatLab.

**New Content:** New examples, new exercises, and Chapter Problems provide relevant and interesting real-world statistical applications, including biometric security, drug testing, gender selection, and analyzing ultrasound images.

	Number	New to This Edition	Use Real Data
Exercises	1600	85%	89%
Examples	200	84%	87%

### Major Organization Changes

#### All Chapters

- **New Chapter Objectives:** All chapters now begin with a list of key learning goals for that chapter. *Chapter Objectives* replaces the former *Overview* numbered sections. The first numbered section of each chapter now covers a major topic.

#### Chapter 1

- **New Section 1-1: Statistical and Critical Thinking**
- **New Subsection 1-3, Part 2: Big Data and Missing Data: Too Much and Not Enough**

### Chapters 2 and 3

- **Chapter Partitioned:** Chapter 2 (Describing, Exploring, and Comparing Data) from the first edition has been partitioned into Chapter 2 (Summarizing and Graphing) and Chapter 3 (Statistics for Describing, Exploring, and Comparing Data).
- **New Section 2-4: Scatterplots, Correlation, and Regression** This new section includes scatterplots in Part 1, the linear correlation coefficient  $r$  in Part 2, and linear regression in Part 3. These additions are intended to greatly facilitate coverage for those professors who prefer some early coverage of correlation and regression concepts. Chapter 10 includes these topics discussed with much greater detail.

### Chapter 4

- **Combined Sections:** Section 3-3 (Addition Rule) and Section 3-4 (Multiplication Rule) from the first edition are now combined into one section: 4-2 (Addition Rule and Multiplication Rule).
- **New Subsection 4-3, Part 3: Bayes' Theorem**

### Chapter 5

- **Combined Sections:** Section 4-3 (Binomial Probability Distributions) and Section 4-4 (Mean, Variance, and Standard Deviation for the Binomial Distribution) from the first edition are now combined into one section: 5-2 (Binomial Probability Distributions).

### Chapter 6

- **Switched Sections:** Section 6-5 (Assessing Normality) now precedes Section 6-6 (Normal as Approximation to Binomial).

### Chapter 7

- **Combined Sections:** Sections 6-4 (Estimating a Population Mean:  $\sigma$  Known) and 6-5 (Estimating a Population Mean:  $\sigma$  Not Known) from the first edition have been combined into one section: 7-2 (Estimating a Population Mean). The coverage of the  $\sigma$  known case has been substantially reduced and it is now limited to Part 2 of Section 7-2.
- **New Section 7-4: Bootstrapping: Using Technology for Estimates**

### Chapter 8

- **Combined Sections:** Sections 7-4 (Testing a Claim About a Population Mean:  $\sigma$  Known) and 7-5 (Testing a Claim About a Population Mean:  $\sigma$  Not Known) from the first edition have been combined into one section: 8-3 (Testing a Claim About a Mean). Coverage of the  $\sigma$  known case has been substantially reduced and it is now limited to Part 2 of Section 8-3.

### Chapter 10

- **New Section: 10-5 Dummy Variables and Logistic Regression**

### Chapter 11

- **New Subsection: Section 11-2, Part 2 Test of Homogeneity, Fisher's Exact Test, and McNemar's Test for Matched Pairs**

### Chapter 14

- **Combined Sections:** Section 13-2 (Elements of a Life Table) and Section 13-3 (Applications of Life Tables) from the first edition have been combined into Section 14-1 (Life Tables).
- **New Section: 14-2 Kaplan-Meier Survival Analysis**

## Flexible Syllabus

This book's organization reflects the preferences of most statistics instructors, but there are two common variations:

- **Early Coverage of Correlation and Regression:** Some instructors prefer to cover the basics of correlation and regression early in the course. Section 2-4 now includes basic concepts of scatterplots, correlation, and regression without the use of formulas and greater depth found in Sections 10-1 (*Correlation*) and 10-2 (*Regression*).
- **Minimum Probability:** Some instructors prefer extensive coverage of probability, while others prefer to include only basic concepts. Instructors preferring minimum coverage can include Section 4-1 while skipping the remaining sections of Chapter 4, as they are not essential for the chapters that follow. Many instructors prefer to cover the fundamentals of probability along with the basics of the addition rule and multiplication rule (Section 4-2).

## GAISE

This book reflects recommendations from the American Statistical Association and its *Guidelines for Assessment and Instruction in Statistics Education* (GAISE). Those guidelines suggest the following objectives and strategies.

1. **Emphasize statistical literacy and develop statistical thinking:** Each section exercise set begins with *Statistical Literacy and Critical Thinking* exercises. Many of the book's exercises are designed to encourage statistical thinking rather than the blind use of mechanical procedures.
2. **Use real data:** 87% of the examples and 89% of the exercises use real data.
3. **Stress conceptual understanding rather than mere knowledge of procedures:** Instead of seeking simple numerical answers, most exercises and examples involve conceptual understanding through questions that encourage practical interpretations of results. Also, each chapter includes a *From Data to Decision* project.
4. **Foster active learning in the classroom:** Each chapter ends with several *Cooperative Group Activities*.
5. **Use technology for developing conceptual understanding and analyzing data:** Computer software displays are included throughout the book. Special *Tech Center* subsections are available online, and they include instruction for using the software. Each chapter includes a *Technology Project*. When there are discrepancies between answers based on tables and answers based on technology, Appendix D provides *both* answers. The websites [www.TriolaStats.com](http://www.TriolaStats.com) and [www.pearsonhighered.com/triola](http://www.pearsonhighered.com/triola) as well as MyStatLab include free text-specific software (Statdisk), data sets formatted for several different technologies, and instructional videos for technologies.
6. **Use assessments to improve and evaluate student learning:** Assessment tools include an abundance of section exercises, *Chapter Quick Quizzes*, *Review Exercises*, *Cumulative Review Exercises*, *Technology Projects*, *From Data to Decision* projects, and *Cooperative Group Activities*.

## Acknowledgments

We would like to thank the many statistics professors and students who have contributed to the success of this book. We thank the reviewers for their suggestions for this second edition:

James Baldone, Virginia College  
Naomi Brownstein, Florida State University  
Christina Caruso, University of Guelph  
Erica A. Corbett, Southeastern Oklahoma State University  
Xiangming Fang, East Carolina University  
Phil Gona, UMASS Boston  
Sharon Homan, University of North Texas  
Jackie Milton, Boston University  
Joe Pick, Palm Beach State College  
Steve Rigdon, St. Louis University  
Brian Smith, Black Hills State University  
Mahbobeh Vezvaei, Kent State University  
David Zeitler, Grand Valley State University

We also thank Paul Lorzak, Joseph Pick and Erica Corbett for their help in checking the accuracy of the text and answers.

Marc Triola  
Mario Triola  
Jason Roy  
*September 2016*



Pearson

# Resources for Success

**MyStatLab® Online Course** for Biostatistics: For the Biological and Health Sciences, 2e by Marc M. Triola, Mario F. Triola and Jason Roy (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.

**MathXL coverage** - MathXL is a market-leading text-specific autograded homework system built to improve student learning outcomes.



**Enhanced video program to meet Introductory Statistics needs:**

- **New! Tech-Specific Video Tutorials** - These short, topical videos address how to use varying technologies to complete exercises.
- **Updated! Section Lecture Videos** - Watch author, Marty Triola, work through examples and elaborate on key objectives of the chapter.

**Real-World Data Examples** - Help understand how statistics applies to everyday life through the extensive current, real-world data examples and exercises provided throughout the text.

Year	L.L.A.S.	MSE	SEEM	THAAA	THAAA	SEV	SEV	SEV	SEV
1	11	1	25	25	25	25	25	25	25
2	11	1	25	25	25	25	25	25	25
3	11	1	25	25	25	25	25	25	25
4	11	1	25	25	25	25	25	25	25
5	11	1	25	25	25	25	25	25	25
6	11	1	25	25	25	25	25	25	25
7	11	1	25	25	25	25	25	25	25
8	11	1	25	25	25	25	25	25	25
9	11	1	25	25	25	25	25	25	25
10	11	1	25	25	25	25	25	25	25
11	11	1	25	25	25	25	25	25	25
12	11	1	25	25	25	25	25	25	25
13	11	1	25	25	25	25	25	25	25
14	11	1	25	25	25	25	25	25	25
15	11	1	25	25	25	25	25	25	25
16	11	1	25	25	25	25	25	25	25
17	11	1	25	25	25	25	25	25	25
18	11	1	25	25	25	25	25	25	25
19	11	1	25	25	25	25	25	25	25
20	11	1	25	25	25	25	25	25	25
21	11	1	25	25	25	25	25	25	25
22	11	1	25	25	25	25	25	25	25

## Supplements

### For the Student

**Student's Solutions Manual**, by James Lapp (Colorado Mesa University) provides detailed, worked-out solutions to all odd-numbered text exercises.

(ISBN-13: 978-0-13-403909-1; ISBN-10: 0-13-403909-2)

**Student Workbook for the Triola Statistics Series**, by Laura Iossi (Broward College) offers additional examples, concept exercises, and vocabulary exercises for each chapter.

(ISBN-13: 978-0-13-446423-7; ISBN 10: 0-13-446423-0)

The following technology manuals, available in MyStatLab, include instructions, examples from the main text, and interpretations to complement those given in the text.

**Excel Student Laboratory Manual and Workbook (Download Only)**, by Laurel Chiappetta (University of Pittsburgh).

(ISBN-13: 978-0-13-446427-5; ISBN-10: 0-13-446427-3)

**MINITAB Student Laboratory Manual and Workbook (Download Only)**, by Mario F. Triola.

(ISBN-13: 978-0-13-446418-3; ISBN-10: 0-13-446418-4)

**Graphing Calculator Manual for the TI-83 Plus, TI-84 Plus, TI-84 Plus C and TI-84 Plus CE (Download Only)**, by Kathleen McLaughlin (University of Connecticut) & Dorothy Wakefield (University of Connecticut Health Center).

(ISBN-13: 978-0-13-446414-5; ISBN 10: 0-13-446414-1)

**Statdisk Student Laboratory Manual and Workbook (Download Only)**, by Mario F. Triola. These files are available to instructors and students through the Triola Statistics Series website, [www.pearsonhighered.com/triola](http://www.pearsonhighered.com/triola), and MyStatLab.

**SPSS Student Laboratory Manual and Workbook (Download Only)**, by James J. Ball (Indiana State University). These files are available to instructors and students through the Triola Statistics Series website, [www.pearsonhighered.com/triola](http://www.pearsonhighered.com/triola), and MyStatLab.

### For the Instructor

**Instructor's Solutions Manual (Download Only)**, by James Lapp (Colorado Mesa University) contains solutions to all the exercises. These files are available to qualified instructors through Pearson Education's online catalog at [www.pearsonhighered.com/irc](http://www.pearsonhighered.com/irc) or within MyStatLab.

**Insider's Guide to Teaching with the Triola Statistics Series**, by Mario F. Triola, contains sample syllabi and tips for incorporating projects, as well as lesson overviews, extra examples, minimum outcome objectives, and recommended assignments for each chapter.

(ISBN-13: 978-0-13-446425-1; ISBN-10: 0-13-446425-7)

**TestGen® Computerized Test Bank** ([www.pearsoned.com/testgen](http://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 testbank are available for download from Pearson Education's online catalog at [www.pearsonhighered.com](http://www.pearsonhighered.com). A test bank (Download Only) is also available from the online catalog.

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

## Technology Resources

The following resources can be found on the Triola Statistics Series website (<http://www.pearsonhighered.com/triola>), the author maintained [www.triolastats.com](http://www.triolastats.com), and MyStatLab

- Appendix B data sets formatted for Minitab, SPSS, SAS, Excel, JMP, and as text files. Additionally, these data sets are available as an APP for the TI-83/84 Plus calculators, and supplemental programs for the TI-83/84 Plus calculator are also available.
- Statdisk statistical software instructions for download. New features include the ability to directly use lists of data instead of requiring the use of their summary statistics.
- Extra data sets, an index of applications, and a symbols table.

**Video resources** have been expanded, updated and now supplement most sections of the book, with many topics presented by the author. The videos aim to support both instructors and students through lecture, reinforcing statistical basics through technology, and applying concepts:

- **Section Lecture Videos**



- **New! Technology Video Tutorials** - These short, topical videos address how to use Excel, Statdisk, and the TI graphing calculator to complete exercises.
- **StatTalk Videos: 24 Conceptual Videos to Help You Actually Understand Statistics.** Fun-loving statistician Andrew Vickers takes to the streets of Brooklyn, NY, to demonstrate important statistical concepts through interesting stories and real-life events. These fun and engaging videos will help students actually understand statistical concepts. Available with an instructors user guide and assessment questions.

### MyStatLab™ Online Course (access code required)

MyStatLab is a course management system that delivers proven results in helping individual students succeed.

- MyStatLab can be successfully implemented in any environment—lab-based, hybrid, fully online, traditional—and demonstrates the quantifiable difference that integrated usage has on student retention, subsequent success, and overall achievement.
- MyStatLab’s comprehensive online gradebook automatically tracks students’ results on tests, quizzes, homework, and in the study plan. Instructors can use the gradebook to provide positive feedback or intervene if students have trouble. Gradebook data can be easily exported to a variety of spreadsheet programs, such as Microsoft Excel. You can determine which points of data you want to export, and then analyze the results to determine success.

MyStatLab provides engaging experiences that personalize, stimulate, and measure learning for each student. In addition to the resources below, each course includes a full interactive online version of the accompanying textbook.

- **Tutorial Exercises with Multimedia Learning Aids:** The homework and practice exercises in MyStatLab align with the exercises in the textbook, and they regenerate algorithmically to give students unlimited opportunity for practice and mastery. Exercises offer immediate helpful feedback, guided solutions, sample problems, animations, videos, and eText clips for extra help at point-of-use.
- **Getting Ready for Statistics:** A library of questions now appears within each MyStatLab course to offer the developmental math topics students need for the course. These can be assigned as a prerequisite to other assignments, if desired.
- **Conceptual Question Library:** In addition to algorithmically regenerated questions that are aligned with

your textbook, there is a library of 1000 Conceptual Questions available in the assessment manager that require students to apply their statistical understanding.

- **StatCrunch™:** MyStatLab integrates the web-based statistical software, StatCrunch, within the online assessment platform so that students can easily analyze data sets from exercises and the text. In addition, MyStatLab includes access to [www.StatCrunch.com](http://www.StatCrunch.com), a website where users can access more than 15,000 shared data sets, conduct online surveys, perform complex analyses using the powerful statistical software, and generate compelling reports.
- **Statistical Software Support:** Knowing that students often use external statistical software, we make it easy to copy our data sets, both from the ebook and the MyStatLab questions, into software such as StatCrunch, Minitab, Excel, and more. Students have access to a variety of support tools—Technology Tutorial Videos, Technology Study Cards, and Technology Manuals for select titles—to learn how to effectively use statistical software.

### MathXL® for Statistics Online Course (access code required)

MathXL® is the homework and assessment engine that runs MyStatLab. (MyStatLab is MathXL plus a learning management system.)

With MathXL for Statistics, instructors can:

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

With MathXL for Statistics, students can:

- Take chapter tests in MathXL and receive personalized study plans and/or personalized homework assignments based on their test results.
- Use the study plan and/or the homework to link directly to tutorial exercises for the objectives they need to study.
- Students can also access supplemental animations and video clips directly from selected exercises.
- Knowing that students often use external statistical software, we make it easy to copy our data sets, both from the ebook and the MyStatLab questions, into software like StatCrunch™, Minitab, Excel, and more.

MathXL for Statistics is available to qualified adopters. For more information, visit our website at [www.mathxl.com](http://www.mathxl.com), or contact your Pearson representative.

### **StatCrunch™**

StatCrunch is powerful, web-based statistical software that allows users to perform complex analyses, share data sets, and generate compelling reports. A vibrant online community offers more than 15,000 data sets for students to analyze.

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

Full access to StatCrunch is available with MyStatLab and StatCrunch is available by itself to qualified adopters. StatCrunch Mobile is now available to access from your mobile device. For more information, visit our website at [www.StatCrunch.com](http://www.StatCrunch.com), or contact your Pearson representative.

**Minitab® 17 and Minitab Express™** make learning statistics easy and provide students with a skill-set that's in demand in today's data driven workforce. Bundling Minitab® software with educational materials ensures students have access to the software they need in the classroom, around campus, and at home. And having 12 month versions of Minitab 17 and Minitab Express available ensures students can use the software for the duration of their course.

ISBN 13: 978-0-13-445640-9

ISBN 10: 0-13-445640-8 (Access Card only; not sold as stand alone.)

**JMP Student Edition, Version 12** is an easy-to-use, streamlined version of JMP desktop statistical discovery software from SAS Institute, Inc., and is available for bundling with the text.

(ISBN-13: 978-0-13-467979-2 ISBN-10: 0-13-467979-2)

# 1

# Introduction to Statistics



- 1-1** Statistical and Critical Thinking
- 1-2** Types of Data
- 1-3** Collecting Sample Data

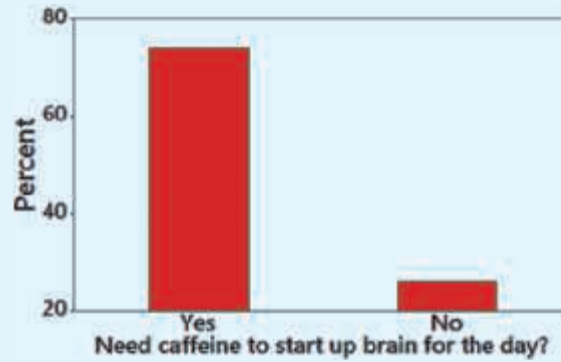


## Survey Question: Do You Need Caffeine to Start Up Your Brain for the Day?

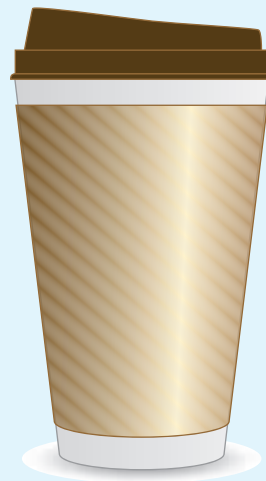
Surveys provide data that enable us to improve products or services. Surveys guide political candidates, shape business practices, identify effective medical treatments, and affect many aspects of our lives. Surveys give us insight into the opinions and behaviors of others. As an example, the National Health and Nutrition Examination Survey (NHANES) is part

of a research program that studies the health and nutrition of thousands of adults and children in the United States.

Let's consider one *USA Today* survey in which respondents were asked if they need caffeine to start up their brain for the day. Among 2,006 respondents, 74% said that they did need the caffeine. Figure 1-1 includes graphs that depict these results.



**FIGURE 1-1(a)** Survey Results



**People Needing Caffeine to Start Up Brain for the Day**



**People Not Needing Caffeine to Start Up Brain for the Day**

**FIGURE 1-1(b)** Survey Results

The survey results suggest that people overwhelmingly need caffeine to start up their brains for the day. The graphs in Figure 1-1 visually depict the survey results. One of the most important objectives of this book is to encourage the use of critical thinking so that such results are not blindly accepted. We might question whether the survey results are valid. Who conducted the survey? How were respondents selected? Do the graphs in Figure 1-1 depict the results well, or are those graphs somehow misleading?

The survey results presented here have major flaws that are among the most common, so they are especially important to recognize. Here are brief descriptions of each of the major flaws:

**Flaw 1: Misleading Graphs** The bar chart in Figure 1-1(a) is very deceptive. By using a vertical scale that does not start at zero, the difference between the two percentages is grossly exaggerated. Figure 1-1(a) makes it appear that approximately eight times as many people need the caffeine. However, with 74% needing caffeine and 26% not needing caffeine, the ratio is actually about 3:1, rather than the 8:1 ratio that is suggested by the graph.

The illustration in Figure 1-1(b) is also deceptive. Again, the difference between the actual response rates of 74% (needing caffeine) and 26% (not needing caffeine) is a difference that is grossly distorted. The picture graph (or “pictograph”) in Figure 1-1(b) makes it appear that

the ratio of people needing caffeine to people not needing caffeine is roughly 9:1 instead of the correct ratio of about 3:1. (Objects with area or volume can distort perceptions because they can be drawn to be disproportionately larger or smaller than the data indicate.) Deceptive graphs are discussed in more detail in Section 2-3, but we see here that the illustrations in Figure 1-1 grossly exaggerate the number of people needing caffeine.

**Flaw 2: Bad Sampling Method** The aforementioned survey responses are from a *USA Today* survey of Internet users. The survey question was posted on a website and Internet users decided whether to respond. This is an example of a *voluntary response sample*—a sample in which respondents themselves decide whether to participate. With a voluntary response sample, it often happens that those with a strong interest in the topic are more likely to participate, so the results are very questionable. For example, people who strongly feel that they cannot function without their morning cup(s) of coffee might be more likely to respond to the caffeine survey than people who are more ambivalent about caffeine or coffee. When using sample data to learn something about a population, it is *extremely* important to obtain sample data that are representative of the population from which the data are drawn. As we proceed through this chapter and discuss types of data and sampling methods, we should focus on these key concepts:

- **Sample data must be collected in an appropriate way, such as through a process of random selection.**
- **If sample data are not collected in an appropriate way, the data may be so completely useless that no amount of statistical torturing can salvage them.**

It would be easy to accept the preceding survey results and blindly proceed with calculations and statistical analyses, but we would miss the critical two flaws described above. We could then develop conclusions that are fundamentally wrong and misleading. Instead, we should develop skills in statistical thinking and critical thinking so that we are better prepared to analyze such data.

## CHAPTER OBJECTIVES

The single most important concept presented in this chapter is this: When using methods of statistics with sample data to form conclusions about a population, it is absolutely essential to collect sample data in a way that is appropriate. Here are the main chapter objectives:

1-1

### Statistical and Critical Thinking

- Analyze sample data relative to context, source, and sampling method.
- Understand the difference between statistical significance and practical significance.
- Define and identify a *voluntary response sample* and know that statistical conclusions based on data from such a sample are generally not valid.



**1-2** Types of Data

- Distinguish between a *parameter* and a *statistic*.
- Distinguish between *quantitative data* and *categorical* (or *qualitative* or *attribute*) *data*.
- Distinguish between *discrete* data and *continuous* data.
- Determine whether basic statistical calculations are appropriate for a particular data set.

**1-3** Collecting Sample Data

- Define and identify a *simple random sample*.
- Understand the importance of sound sampling methods and the importance of good design of experiments.

**1-1****Statistical and Critical Thinking**

**Key Concept** In this section we begin with a few very basic definitions, and then we consider an *overview* of the process involved in conducting a statistical study. This process consists of “prepare, analyze, and conclude.” “Preparation” involves consideration of the *context*, the *source* of data, and *sampling method*. In future chapters we construct suitable graphs, explore the data, and execute computations required for the statistical method being used. In future chapters we also form conclusions by determining whether results have statistical significance and practical significance.

Statistical thinking involves critical thinking and the ability to make sense of results. Statistical thinking demands so much more than the ability to execute complicated calculations. Through numerous examples, exercises, and discussions, this text will help you develop the statistical thinking skills that are so important in today’s world.

We begin with some very basic definitions.

**DEFINITIONS**

**Data** are collections of observations, such as measurements, or survey responses. (A single data value is called a *datum*, a term rarely used. The term “data” is plural, so it is correct to say “data *are*...” not “data *is*...”)

**Statistics** is the science of planning studies and experiments; obtaining data; and organizing, summarizing, presenting, analyzing, and interpreting those data and then drawing conclusions based on them.

A **population** is the complete collection of *all* measurements or data that are being considered. Typically, the population is the complete collection of data that we would like to make inferences about.

A **census** is the collection of data from *every* member of the population.

A **sample** is a *subcollection* of members selected from a population.

Because populations are often very large, a common objective of the use of statistics is to obtain data from a sample and then use those data to form a conclusion about the population.

### EXAMPLE 1 Residential Carbon Monoxide Detectors

In the journal article “Residential Carbon Monoxide Detector Failure Rates in the United States” (by Ryan and Arnold, *American Journal of Public Health*, Vol. 101, No. 10), it was stated that there are 38 million carbon monoxide detectors installed in the United States. When 30 of them were randomly selected and tested, it was found that 12 of them failed to provide an alarm in hazardous carbon monoxide conditions. In this case, the population and sample are as follows:

**Population:** All 38 million carbon monoxide detectors in the United States

**Sample:** The 30 carbon monoxide detectors that were selected and tested

The objective is to use the sample data as a basis for drawing a conclusion about the population of all carbon monoxide detectors, and methods of statistics are helpful in drawing such conclusions.

We now proceed to consider the process involved in a statistical study. See Figure 1-2 for a summary of this process and note that the focus is on critical thinking, not mathematical calculations. Thanks to wonderful developments in technology, we have powerful tools that effectively do the number crunching so that we can focus on understanding and interpreting results.

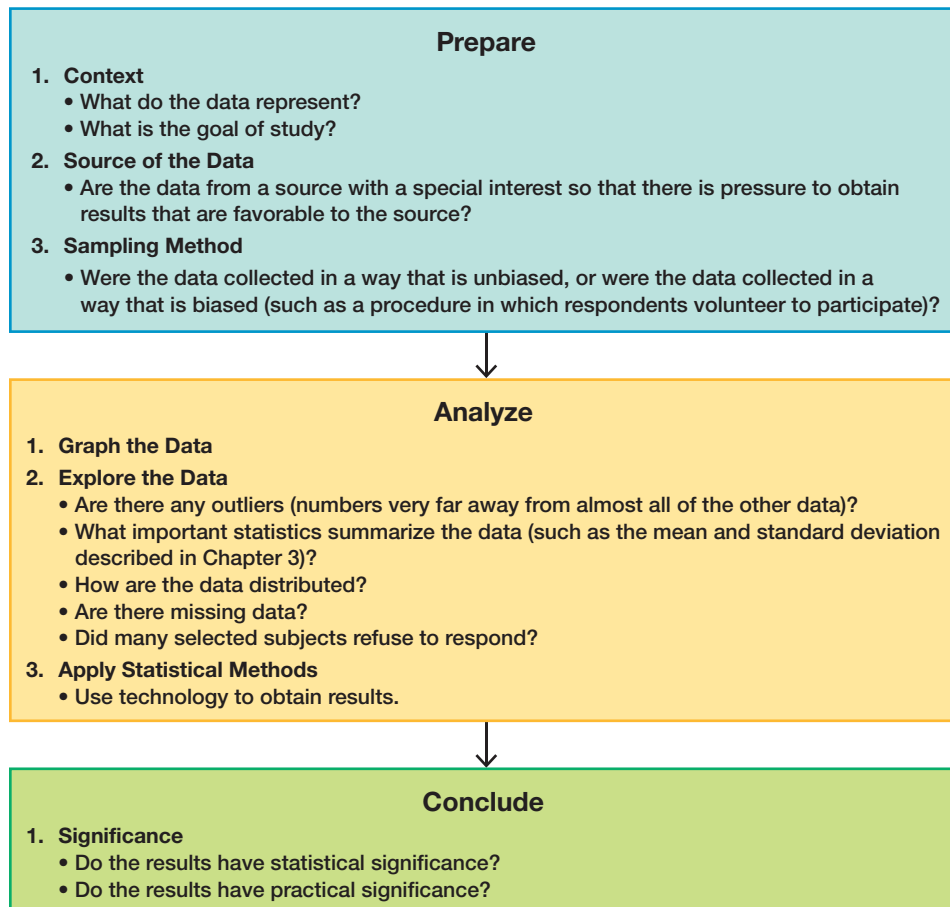


FIGURE 1-2 Statistical Thinking

### Survivorship Bias

In World War II, statistician Abraham Wald saved many lives with his work on the Applied Mathematics Panel. Military leaders asked the panel how they could improve the chances of aircraft bombers returning after missions. They wanted to add some armor for protection, and they recorded locations on the bombers where damaging holes were found. They reasoned that armor should be placed in locations with the most holes, but Wald said that strategy would be a big mistake. He said that armor should be placed where returning bombers were *not* damaged. His reasoning was this: The bombers that made it back with damage were survivors, so the damage they suffered could be survived. Locations on the aircraft that were not damaged were the most vulnerable, and aircraft suffering damage in those vulnerable areas were the ones that did not make it back. The military leaders would have made a big mistake with survivorship bias by studying the planes that survived instead of thinking about the planes that did not survive.



## Origin of “Statistics”



The word *statistics* is derived from the Latin word *status* (meaning “state”).

Early uses of statistics involved compilations of data and graphs describing various aspects of a state or country. In 1662, John Graunt published statistical information about births and deaths. Graunt’s work was followed by studies of mortality and disease rates, population sizes, incomes, and unemployment rates. Households, governments, and businesses rely heavily on statistical data for guidance. For example, unemployment rates, inflation rates, consumer indexes, and birth and death rates are carefully compiled on a regular basis, and the resulting data are used by business leaders to make decisions affecting future hiring, production levels, and expansion into new markets.

**TABLE 1-1** IQ Scores and Brain Volumes (cm<sup>3</sup>)

IQ	96	87	101	103	127	96	88	85	97	124
Brain Volume (cm <sup>3</sup> )	1005	1035	1281	1051	1034	1079	1104	1439	1029	1160

## Prepare

**Context** Figure 1-2 suggests that we begin our preparation by considering the *context* of the data, so let’s start with context by considering the data in Table 1-1. (The data are from Data Set 9 “IQ and Brain Size” in Appendix B.) The data in Table 1-1 consist of measured IQ scores and measured brain volumes from 10 different subjects. The data are matched in the sense that each individual “IQ/brain volume” pair of values is from the same person. The first subject had a measured IQ score of 96 and a brain volume of 1005 cm<sup>3</sup>. The format of Table 1-1 suggests the following goal: Determine whether there is a *relationship* between IQ score and brain volume. This goal suggests a possible hypothesis: People with larger brains tend to have higher IQ scores.

**Source of the Data** The data in Table 1-1 were provided by M. J. Tramo, W. C. Loftus, T. A. Stukel, J. B. Weaver, and M. S. Gazziniga, who discuss the data in the article “Brain Size, Head Size, and IQ in Monozygotic Twins,” *Neurology*, Vol. 50. The researchers are from reputable medical schools and hospitals, and they would not gain by presenting the results in way that is misleading. In contrast, Kiwi Brands, a maker of shoe polish, commissioned a study that resulted in this statement, which was printed in some newspapers: “According to a nationwide survey of 250 hiring professionals, scuffed shoes was the most common reason for a male job seeker’s failure to make a good first impression.”

When physicians who conduct clinical experiments on the efficacy of drugs receive funding from drug companies, they have an incentive to obtain favorable results. Some professional journals, such as the *Journal of the American Medical Association*, now require that physicians report sources of funding in journal articles. We should be skeptical of studies from sources that may be biased.

**Sampling Method** Figure 1-2 suggests that we conclude our preparation by considering the sampling method. The data in Table 1-1 were obtained from subjects whose medical histories were reviewed in an effort to ensure that no subjects had neurologic or psychiatric disease. In this case, the sampling method appears to be sound, but we cannot be sure of that without knowing how the subjects were recruited and whether any payments may have affected participation in the study.

Sampling methods and the use of randomization will be discussed in Section 1-3, but for now, we stress that a sound sampling method is absolutely essential for good results in a statistical study. It is generally a bad practice to use voluntary response (or self-selected) samples, even though their use is common.

### DEFINITION

A **voluntary response sample** (or **self-selected sample**) is one in which the respondents themselves decide whether to be included.

The following types of polls are common examples of voluntary response samples. By their very nature, all are seriously flawed because we should not make conclusions about a population on the basis of samples with a strong possibility of bias:

- Internet polls, in which people online can decide whether to respond
- Mail-in polls, in which people decide whether to reply



- Telephone call-in polls, in which newspaper, radio, or television announcements ask that you voluntarily call a special number to register your opinion

The Chapter Problem involves a *USA Today* survey with a voluntary response sample. See also the following Example 2.

### EXAMPLE 2 Voluntary Response Sample

*USA Today* posted this question on the electronic edition of their newspaper: “Have you ever been bitten by an animal?” Internet users who saw that question then decided themselves whether to respond. Among the 2361 responses, 65% said “yes” and 35% said “no.” Because the 2361 subjects themselves chose to respond, they are a voluntary response sample and the results of the survey are highly questionable. It would be much better to get results through a poll in which the pollster randomly selects the subjects, instead of allowing the subjects to volunteer themselves.

## Analyze

Figure 1-2 indicates that after completing our preparation by considering the context, source, and sampling method, we begin to *analyze* the data.

**Graph and Explore** An analysis should begin with appropriate graphs and explorations of the data. Graphs are discussed in Chapter 2, and important statistics are discussed in Chapter 3.

**Apply Statistical Methods** Later chapters describe important statistical methods, but application of these methods is often made easy with technology (calculators and/or statistical software packages). A good statistical analysis does not require strong computational skills. A good statistical analysis does require using common sense and paying careful attention to sound statistical methods.

## Conclude

Figure 1-2 shows that the final step in our statistical process involves conclusions, and we should develop an ability to distinguish between statistical significance and practical significance.

**Statistical Significance** *Statistical significance* is achieved in a study when we get a result that is very unlikely to occur by chance. A common criterion is that we have statistical significance if the likelihood of an event occurring by chance is 5% or less.

- Getting 98 girls in 100 random births is statistically significant because such an extreme outcome is not likely to result from random chance.
- Getting 52 girls in 100 births is not statistically significant because that event could easily occur with random chance.

**Practical Significance** It is possible that some treatment or finding is effective, but common sense might suggest that the treatment or finding does not make enough of a difference to justify its use or to be practical, as illustrated in Example 3 which follows.

### EXAMPLE 3 Statistical Significance Versus Practical Significance

ProCare Industries once supplied a product named Gender Choice that supposedly increased the chance of a couple having a baby with the gender that they desired. In the absence of any evidence of its effectiveness, the product was banned by the Food and Drug Administration (FDA) as a “gross deception of the consumer.” But suppose that the product was tested with 10,000 couples who wanted to have baby girls, and the results consist of 5200 baby girls born in the 10,000 births. This result is statistically significant because the likelihood of it happening due to chance is only 0.003%, so chance doesn’t seem like a feasible explanation. That 52% rate of girls is statistically significant, but it lacks practical significance because 52% is only slightly above 50%. Couples would not want to spend the time and money to increase the likelihood of a girl from 50% to 52%. (*Note:* In reality, the likelihood of a baby being a girl is about 48.8%, not 50%.)

## Analyzing Data: Potential Pitfalls

Here are a few more items that could cause problems when analyzing data.

**Misleading Conclusions** When forming a conclusion based on a statistical analysis, we should make statements that are clear even to those who have no understanding of statistics and its terminology. We should carefully avoid making statements not justified by the statistical analysis. For example, later in this book we introduce the concept of a correlation, or association between two variables, such as smoking and pulse rate. A statistical analysis might justify the statement that there is a correlation between the number of cigarettes smoked and pulse rate, but it would not justify a statement that the number of cigarettes smoked *causes* a person’s pulse rate to change. Such a statement about causality can be justified by physical evidence, not by statistical analysis.

**Correlation does not imply causation.**

**Sample Data Reported Instead of Measured** When collecting data from people, it is better to take measurements yourself instead of asking subjects to *report* results. Ask people what they weigh and you are likely to get their *desired* weights, not their actual weights. People tend to round, usually down, sometimes *way* down. When asked, someone with a weight of 187 lb might respond that he or she weighs 160 lb. Accurate weights are collected by using a scale to measure weights, not by asking people what they weigh.

**Loaded Questions** If survey questions are not worded carefully, the results of a study can be misleading. Survey questions can be “loaded” or intentionally worded to elicit a desired response. Here are the actual rates of “yes” responses for the two different wordings of a question:

97% yes: “Should the President have the line item veto to eliminate waste?”

57% yes: “Should the President have the line item veto, or not?”

**Order of Questions** Sometimes survey questions are unintentionally loaded by such factors as the order of the items being considered. See the following two

questions from a poll conducted in Germany, along with the very different response rates:

“Would you say that traffic contributes more or less to air pollution than industry?” (45% blamed traffic; 27% blamed industry.)

“Would you say that industry contributes more or less to air pollution than traffic?” (24% blamed traffic; 57% blamed industry.)

In addition to the order of items within a question, as illustrated above, the order of separate questions could also affect responses.

**Nonresponse** A *nonresponse* occurs when someone either refuses to respond to a survey question or is unavailable. When people are asked survey questions, some firmly refuse to answer. The refusal rate has been growing in recent years, partly because many persistent telemarketers try to sell goods or services by beginning with a sales pitch that initially sounds as though it is part of an opinion poll. (This “selling under the guise” of a poll is called *sugging*.) In *Lies, Damn Lies, and Statistics*, author Michael Wheeler makes this very important observation:

**People who refuse to talk to pollsters are likely to be different from those who do not. Some may be fearful of strangers and others jealous of their privacy, but their refusal to talk demonstrates that their view of the world around them is markedly different from that of those people who will let poll-takers into their homes.**

**Percentages** Some studies cite misleading or unclear percentages. Note that 100% of some quantity is *all* of it, but if there are references made to percentages that exceed 100%, such references are often not justified. If a medical researcher claims that she has developed a treatment for migraine headaches and the treatment results in a 150% reduction in those headaches, that researcher cannot be correct, because totally eliminating *all* migraine headaches would be a 100% reduction. It is impossible to reduce the number of migraine headaches by more than 100%.

When working with percentages, we should know that % or “percent” really means “divided by 100.” Here is a principle used often in this book.

**Percentage of:** To find a percentage of an amount, replace the % symbol with division by 100, and then interpret “of” to be multiplication. The following calculation shows that 6% of 1200 is 72:

$$6\% \text{ of } 1200 \text{ responses} = \frac{6}{100} \times 1200 = 72$$

## 1-1 Basic Skills and Concepts

### Statistical Literacy and Critical Thinking

**1. Online Medical Info** *USA Today* posted this question on its website: “How often do you seek medical information online?” Of 1072 Internet users who chose to respond, 38% of them responded with “frequently.” What term is used to describe this type of survey in which the people surveyed consist of those who decided to respond? What is wrong with this type of sampling method?

**2. Reported Versus Measured** In a survey of 1046 adults conducted by Bradley Corporation, subjects were asked how often they wash their hands when using a public restroom, and 70% of the respondents said “always.”

a. Identify the sample and the population.

b. Why would better results be obtained by observing the hand washing instead of asking about it?

### Publication Bias

There is a “publication bias” in professional journals. It is the tendency to publish positive results (such as showing that some treatment is effective) much more often than negative results (such as showing that some treatment has no effect). In the article “Registering Clinical Trials” (*Journal of the American Medical Association*, Vol. 290, No. 4), authors Kay Dickersin and Drummond Rennie state that “the result of not knowing who has performed what (clinical trial) is loss and distortion of the evidence, waste and duplication of trials, inability of funding agencies to plan, and a chaotic system from which only certain sponsors might benefit, and is invariably against the interest of those who offered to participate in trials and of patients in general.” They support a process in which *all* clinical trials are registered in one central system, so that future researchers have access to all previous studies, not just the studies that were published.

