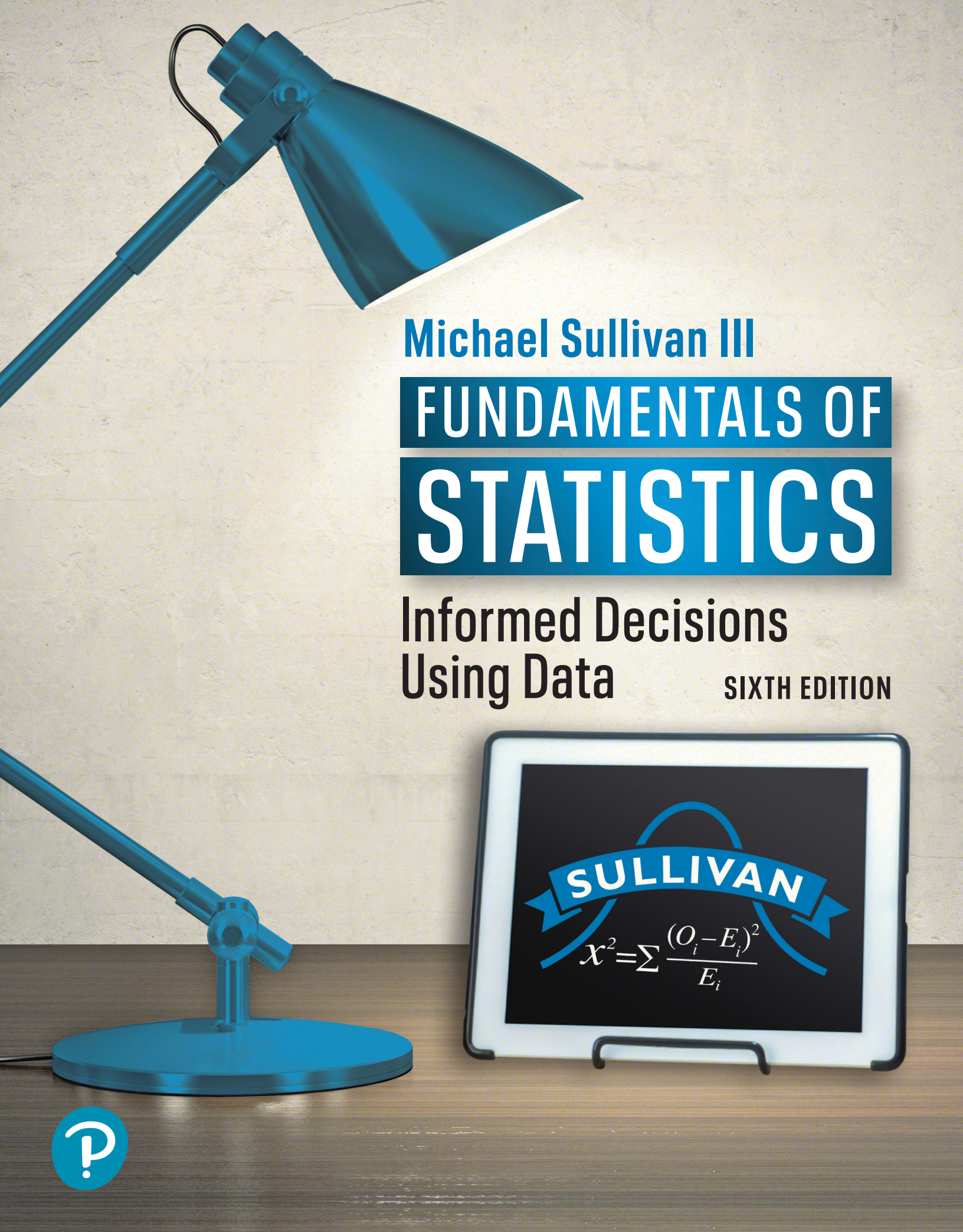

Michael Sullivan III
FUNDAMENTALS OF
STATISTICS
Informed Decisions
Using Data
SIXTH EDITION
SULLIVAN
$x^2=\sum \frac{(O_i-E_i)^2}{E_i}$

Sullivan's Pathway to Making an Informed Decision

Begin your journey . . .

- **Making an Informed Decision** projects at the start of each chapter allow you to work with data in order to make informed decisions that impact your life.
- **Putting It Together** overviews show how material you are about to cover relates to prior material.

Preparation is key . . .

- **Preparing for This Section** lists all of the skills needed to be successful.
- **Preparing for This Section Quizzes** are available as a digital MyLab assignment or as a print quiz to help you check your mastery.
- **Each Objective** is listed at the beginning of the section and then repeated in the text for easy reference.

Look at the model then practice, practice, practice . . .

- **Step-by-Step Annotated Examples** illustrate new concepts and methods in 3 steps:
  1. Problem
  2. Approach
  3. Solution
- **Examples** point to **Now Work Exercises** so you can solve similar exercises on your own.

Exercise Sets . . .

- **Putting It Together** exercises use skills you've acquired in various chapters. (*See facing page*)
- **You Explain It!** exercises ask you to provide an interpretation of statistical results.
- **Threaded Tornado Problems** allow you to analyze a single data set throughout the entire semester. (*See facing page*)
- **Retain Your Knowledge** exercises help you to maintain the skills you have acquired earlier in the course.

Check where you've been and test your mastery . . .

- **Putting It Together Sections** require you to decide which technique to use. (*See facing page*)
- **End-of-Chapter Objectives** are listed with page references for easy review.
- **Chapter Tests** provide an opportunity to test your knowledge.

Apply yourself . . .

- **In-Class Activities** in the Student Activity Workbook allow you to experience statistics in a fun and exciting way by experiencing the process firsthand.
- **Making an Informed Decision** projects require you to use data and statistical techniques learned in the chapter to make important life decisions.
- **End-of-Chapter Case Studies** tie statistical concepts together within an interesting application.

Sullivan's Guide to Putting It Together

Putting It Together Sections	Objective	Page(s)
5.7 Putting It Together: Which Method Do I Use?	1 Determine the appropriate probability rule to use	287–289
	2 Determine the appropriate counting technique to use	289–290
9.3 Putting It Together: Which Method Do I Use?	1 Determine the appropriate confidence interval to construct	425–426
10.4 Putting It Together: Which Method Do I Use?	1 Determine the appropriate hypothesis test to perform (one sample)	470
11.4 Putting It Together: Which Method Do I Use?	1 Determine the appropriate hypothesis test to perform (two samples)	513

Putting It Together Exercises	Skills Utilized	Section(s) Covered	Page(s)
1.2.26 Passive Smoke	Variables, observational studies, designed experiments	1.1, 1.2	23
1.4.37 Comparing Sampling Methods	Simple random sampling and other sampling techniques	1.3, 1.4	38
1.4.38 Thinking about Randomness	Random sampling	1.3, 1.4	38
2.1.29 Online Homework	Variables, designed experiments, bar graphs	1.1, 1.2, 1.6, 2.1	75
2.2.42 Time Viewing a Webpage	Graphing data	2.2	91
2.2.43 Red Light Cameras	Variables, population vs. sample, histograms, dot plots	1.1, 2.2	91
2.2.44 Which Graphical Summary?	Choosing the best graphical summary	2.1, 2.2	91
2.2.45 Shark!	Graphing data	2.3	91–92
3.1.42 Shape, Mean, and Median	Discrete vs. continuous data, histograms, shape of a distribution, mean, median, mode, bias	1.1, 1.4, 2.2, 3.1	120
3.5.18 Paternal Smoking	Observational studies, designed experiments, lurking variables, mean, median, standard deviation, quartiles, boxplots	1.2, 1.6, 3.1, 3.2, 3.4, 3.5	162–163
3.5.19 Taxi Ride	Bar graphs, histograms, boxplots, range, standard deviation	2.1, 2.2, 3.2, 3.5	163
4.2.29 Housing Prices	Scatter diagrams, correlation, linear regression	4.1, 4.2	199
4.2.30 Smoking and Birth Weight	Observational study vs. designed experiment, prospective studies, scatter diagrams, linear regression, correlation vs. causation, lurking variables	1.2, 4.1, 4.2	199–200
4.3.16 Exam Scores	Building a linear model	4.1, 4.2, 4.3	205–206
4.3.17 Cigarette Smuggling	Scatter diagrams, correlation, least-squares regression	4.1, 4.2, 4.3	206
4.4.15 Sullivan Survey II	Relative frequency distributions, bar graphs, pie charts, contingency tables, conditional distributions	2.1, 4.4	216–217
5.1.52 Drug Side Effects	Variables, graphical summaries of data, experiments, probability	1.1, 1.6, 2.1, 5.1	241
5.2.44 Speeding Tickets	Contingency tables, marginal distributions, empirical probabilities	4.4, 5.1	252
5.2.45 Red Light Cameras	Variables, relative frequency distributions, bar graphs, mean, standard deviation, probability, Simpson's Paradox	1.1, 2.1, 3.1, 3.2, 4.4, 5.1, 5.2	252–253
6.1.37 Sullivan Statistics Survey I	Mean, standard deviation, probability, probability distributions	3.1, 3.2, 5.1, 6.1	311
6.2.55 A Drug Study	Types of variables, experimental design; binomial probabilities	1.1, 1.2, 1.6, 6.2	327
6.2.56 Beating the Stock Market	Expected value, binomial probabilities	6.1, 6.2	327
7.2.52 Birth Weights	Relative frequency distribution, histograms, mean and standard deviation from grouped data, normal probabilities	2.1, 2.2, 3.3, 7.2	354
7.3.13 Disney's Dinosaur Ride	Histograms, distribution shape, normal probability plots	2.2, 7.3	359
8.1.34 Bike Sharing	Histograms, mean, standard deviation, distribution shape, sampling distribution of the mean	2.2, 3.1, 3.2, 8.1	383
8.1.35 Playing Roulette	Probability distributions, mean and standard deviation of a random variable, sampling distributions	6.1, 8.1	383
9.1.47 Hand Washing	Observational studies, bias, confidence intervals	1.2, 1.5, 9.1	409–410
9.2.47 Smoking Cessation Study	Experimental design, confidence intervals	1.6, 9.1, 9.2	424
10.2.40 Lupus	Observational studies, retrospective vs. prospective studies, bar graphs, confidence intervals, hypothesis testing	1.2, 2.1, 9.1, 10.2	457
10.2.41 Naughty or Nice?	Experimental design, determining null and alternative hypotheses, binomial probabilities, interpreting *P*-values	1.6, 6.2, 10.1, 10.2	458
11.1.36 Salk Vaccine	Completely randomized design, hypothesis testing	1.6, 11.1	490

(continued)

Putting It Together Exercises	Skills Utilized	Section(s) Covered	Page(s)
11.2.19 Glide Testing	Matched pairs design, hypothesis testing	1.6, 11.2	500–501
11.3.23 Online Homework	Completely randomized design, confounding, hypothesis testing	1.6, 11.3	512
12.1.27 The V-2 Rocket in London	Mean of discrete data, expected value, Poisson probability distribution, goodness-of-fit	6.1, 6.3, 12.1	537
12.1.28 Weldon's Dice	Addition Rule for Disjoint Events, classical probability, goodness-of-fit	5.1, 5.2, 12.1	537
12.2.22 Women, Aspirin, and Heart Attacks	Population, sample, variables, observational study vs. designed experiment, experimental design, compare two proportions, chi-square test of homogeneity	1.1, 1.2, 1.6, 11.1, 12.2	552–553
12.2.23 Corequisite College Algebra	Comparing two independent means, comparing two independent proportions, chi-square test for independence	11.1, 11.3, 12.2	553
12.4.19 Predicting Intelligence	Scatter diagrams, linear correlation coefficient, least-squares regression, normal probability plots, inference on least-squares regression, confidence and prediction intervals	4.1, 4.2, 4.3. 7.3, 12.3, 12.4	574
B.3.27 Psychological Profiles	Standard deviation, sampling methods, two-sample *t*-test, Central Limit Theorem, one-way Analysis of Variance	1.4, 3.2, 8.1, 11.2, B.3	B-28

Threaded Tornado Problems

Throughout the text a single, large data set that measures various variables on all tornadoes that struck the United States in 2017 is utilized. The problems are marked with a icon. The table below shows the sections, problems, topics covered, and page for the Threaded Tornado Problems.

Section	Problem(s)	Topics	Page(s)
1.1	47, 48	Types of variables; types of data	13
2.1	25	Frequency & relative frequency distributions; bar charts; pie charts	74–75
2.2	41	Frequency & relative frequency distributions; histogram; dot plots	91
3.1	41	Mean, median, distribution shape	120
3.2	51	Range, standard deviation	138
3.4	29	Quartiles, interquartile range, outliers	155
3.5	20	Boxplots	163
4.3	15	Scatter diagrams, correlation, least-squares regression, coefficient of determination, residual analysis	205
5.1	49	Probability models; unusual events	240
8.1	33	Describe the distribution of the sample mean from a non-normal population	382–383
9.1	33	Confidence interval for a population proportion	408
9.2	37	Confidence interval for a population mean	422–423
10.2	33	Hypothesis test for a population proportion	456
10.2B	25	Hypothesis test for a population proportion	10.2AB.24
10.3	35	Hypothesis test for a population mean	468
11.1	29	Compare two population proportions (independent samples)	489
11.3	17	Compare two population means (independent samples)	511
12.3	17	Inference on least-squares regression; prediction intervals	573–574
B.3	29	One-way Analysis of Variance (ANOVA)	B-28

FUNDAMENTALS OF STATISTICS

INFORMED DECISIONS USING DATA 6E

Michael Sullivan, III

Joliet Junior College

Pearson

Content Management: Suzanna Bainbridge
Content Production: Robert Carroll, Jean Choe, Peggy McMahon
Product Management: Karen Montgomery
Product Marketing: Alicia Wilson
Rights and Permissions: Pavithra Gunasekaran, Integra

Please contact https://support.pearson.com/getsupport/s/contactsupport with any queries on this content.

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.

MICROSOFT® WINDOWS®, AND MICROSOFT OFFICE® ARE REGISTERED TRADEMARKS OF THE MICROSOFT CORPORATION IN THE U.S.A. AND OTHER COUNTRIES. THIS BOOK IS NOT SPONSORED OR ENDORSED BY OR AFFILIATED WITH THE MICROSOFT CORPORATION.

Copyright © 2022, 2018, 2014 by Pearson Education, Inc. or its affiliates, 221 River Street, Hoboken, NJ 07030. All Rights Reserved. Manufactured 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/.

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

PEARSON, ALWAYS LEARNING, MYLAB(TM) are exclusive trademarks owned by Pearson Education, Inc. or its affiliates in the United States 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.

Library of Congress Cataloging-in-Publication Data

Names: Sullivan, Michael, III, 1967- author.
Title: Fundamentals of statistics : informed decisions using data / Michael Sullivan, III, Joliet Junior College.
Description: 6th edition. | Hoboken : Pearson, 2022. | Includes index. | Summary: "Popular college introductory text to the subject of Statistics that follows GAISE guidelines"-- Provided by publisher.
Identifiers: LCCN 2020035728 | ISBN 9780136807346 (hardback)
Subjects: LCSH: Statistics--Textbooks.
Classification: LCC QA276.12 .S846 2022 | DDC 519.5--dc23
LC record available at https://lccn.loc.gov/2020035728

ScoutAutomatedPrintCode

ISBN 10: 0-13-680734-8
ISBN 13: 978-0-13-680734-6

To My Wife Yolanda
and My Children
Michael, Kevin, and Marissa

This page intentionally left blank

Contents

Preface to the Instructor ix
Resources for Success xv
Applications Index xvii

PART 1 Getting the Information You Need 1

1 Data Collection 2

1.1 Introduction to the Practice of Statistics 3
1.2 Observational Studies versus Designed Experiments 14
1.3 Simple Random Sampling 23
1.4 Other Effective Sampling Methods 30
1.5 Bias in Sampling 38
1.6 The Design of Experiments 44
Chapter 1 Review 55
Chapter Test 59
Making an Informed Decision: What College Should I Attend? 60

PART 2 Descriptive Statistics 61

2 Summarizing Data in Tables and Graphs 62

2.1 Organizing Qualitative Data 63
2.2 Organizing Quantitative Data 76
2.3 Graphical Misrepresentations of Data 92
Chapter 2 Review 100
Chapter Test 104
Making an Informed Decision: Tables or Graphs? 106

3 Numerically Summarizing Data 107

3.1 Measures of Central Tendency 108
3.2 Measures of Dispersion 121
3.3 Measures of Central Tendency and Dispersion from Grouped Data 139
3.4 Measures of Position and Outliers 145
3.5 The Five-Number Summary and Boxplots 155
Chapter 3 Review 164
Chapter Test 167
Making an Informed Decision: What Car Should I Buy? 169

4 **Describing the Relation between Two Variables** 170

4.1 Scatter Diagrams and Correlation 171
4.2 Least-Squares Regression 187
4.3 The Coefficient of Determination 201
4.4 Contingency Tables and Association 206
Chapter 4 Review 217
Chapter Test 222
Making an Informed Decision: Relationships among Variables on a World Scale 224

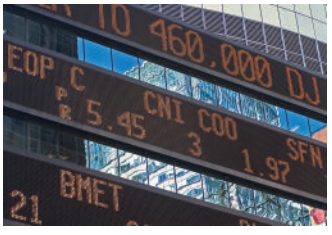
PART 3 Probability and Probability Distributions 225

5 **Probability** 226

5.1 Probability Rules 227
5.2 The Addition Rule and Complements 242
5.3 Independence and the Multiplication Rule 253
5.4 Conditional Probability and the General Multiplication Rule 260
5.5 Counting Techniques 269
5.6 Simulating Probability Experiments 282
5.7 Putting It Together: Which Method Do I Use? 287
Chapter 5 Review 293
Chapter Test 297
Making an Informed Decision: The Effects of Drinking and Driving 298

6 **Discrete Probability Distributions** 299

6.1 Discrete Random Variables 300
6.2 The Binomial Probability Distribution 312
Chapter 6 Review 327
Chapter Test 330
Making an Informed Decision: Should We Convict? 331



7 **The Normal Probability Distribution** 332

7.1 Properties of the Normal Distribution 333
7.2 Applications of the Normal Distribution 343
7.3 Assessing Normality 355
7.4 The Normal Approximation to the Binomial Probability Distribution 360
Chapter 7 Review 365
Chapter Test 367
Making an Informed Decision: Stock Picking 368

PART 4 Inference: From Samples to Population 369

8 Sampling Distributions 370

8.1 Distribution of the Sample Mean 371
8.2 Distribution of the Sample Proportion 384
Chapter 8 Review 392
Chapter Test 393
Making an Informed Decision: How Much Time Do You Spend in a Day . . . ? 394

9 Estimating the Value of a Parameter 395

9.1 Estimating a Population Proportion 396
9.2 Estimating a Population Mean 410
9.3 Putting It Together: Which Method Do I Use? 425
9.4 Estimating a Population Standard Deviation (eText)
9.5 Estimating with Bootstrapping (eText)
Chapter 9 Review 429
Chapter Test 432
Making an Informed Decision: How Much Should I Spend for this House? 433

10 Hypothesis Tests Regarding a Parameter 434

10.1 The Language of Hypothesis Testing 435
10.2 Hypothesis Tests for a Population Proportion 443
10.2A Using Simulation to Perform Hypothesis Tests on a Population Proportion (eText)
10.2B Hypothesis Tests for a Population Proportion Using the Normal Model (eText)
10.3 Hypothesis Tests for a Population Mean 458
10.3A Using Simulation and the Bootstrap to Perform Hypothesis Tests on a Population Mean (eText)
10.4 Putting It Together: Which Method Do I Use? 470
10.5 Hypothesis Tests for a Population Standard Deviation (eText)
Chapter 10 Review 472
Chapter Test 475
Making an Informed Decision: Selecting a Mutual Fund 476

11 Inference on Two Population Parameters 477

11.1 Inference about Two Population Proportions 478
11.1A Using Randomization Techniques to Compare Two Proportions (eText)
11.2 Inference about Two Means: Dependent Samples 490
11.2A Using Bootstrapping to Conduct Inference on Two Dependent Means (eText)
11.3 Inference about Two Means: Independent Samples 501
11.3A Using Randomization Techniques to Compare Two Independent Means (eText)
11.4 Putting It Together: Which Method Do I Use? 513
11.5 Inference about Two Population Standard Deviations (eText)
Chapter 11 Review 517
Chapter Test 520
Making an Informed Decision: Which Car Should I Buy? 522

12 Additional Inferential Methods 523

12.1 Goodness-of-Fit Test 524
12.2 Tests for Independence and the Homogeneity of Proportions 538
12.3 Testing the Significance of the Least-Squares Regression Model 554
12.3A Using Randomization Techniques on the Slope of the Least-Squares Regression Line (eText)
12.4 Confidence and Prediction Intervals 569
Chapter 12 Review 575
Chapter Test 579
Making an Informed Decision: Benefits of College 581

Credits C-1

Appendix A Tables A-1

Appendix B (eText) B-1

B.1 Lines B-1

B.2 Inference about Two Population Proportions: Dependent Samples B-6

B.3 Comparing Three or More Means (One-Way Analysis of Variance) B-12

Answers ANS-1

Subject Index I-1

Tables and Formulas for Sullivan, *Fundamentals of Statistics*, 6e

Preface to the Instructor

Capturing a Powerful and Exciting Discipline in a Textbook

Statistics is a powerful subject, and it is one of my passions. Bringing my passion for the subject together with my desire to create a text that would work for me, my students, and my school led me to write the first edition of this textbook. It continues to motivate me as I reflect on changes in students, in the statistics community, and in the world around us.

When I started writing, I used the manuscript of this text in class. My students provided valuable, insightful feedback, and I made adjustments based on their comments. In many respects, this text was written by students and for students. I also received constructive feedback from a wide range of statistics faculty, which has refined ideas in the book and in my teaching. I continue to receive valuable feedback from both faculty and students, and this text continues to evolve with the goal of providing clear, concise, and readable explanations, while challenging students to think statistically.

In writing this edition, I continue to make a special effort to abide by the Guidelines for Assessment and Instruction in Statistics Education (GAISE) for the college introductory course endorsed by the American Statistical Association (ASA). The GAISE Report gives six recommendations for the course:

1. Emphasize statistical literacy and develop statistical thinking
2. Use real data in teaching statistics
3. Stress conceptual understanding
4. Foster active learning
5. Use technology for developing conceptual understanding
6. Use assessments to improve and evaluate student learning

Changes to this edition and the hallmark features of the text reflect a strong adherence to these important GAISE guidelines.

New to This Edition

- **Over 350 New and Updated Exercises** The sixth edition makes a concerted effort to require students to write a few sentences that explain the results of their statistical analysis. To reflect this effort, the answers in the back of the text provide recommended explanations of the statistical results. Not all the exercises are computational or require statistical analysis. Many of the exercises have been written to require students to explain statistical concepts or understand pitfalls in faulty statistical analysis.
- **Over 100 New and Updated Examples** The examples continue to engage and provide clear, concise explanations for the students while following the *Problem, Approach, Solution* presentation. Problem lays out the scenario of the example, Approach provides insight into the thought process behind the methodology used to solve the problem, and Solution goes through the solution utilizing the methodology suggested in the approach.
- **Threaded Tornado Problems** Throughout the text a single, large data set that measures various variables on all tornadoes that struck the United States in 2017 is utilized. The problems are marked with a icon. The table on the front inside cover shows the sections, problems, topics covered and pages for the Threaded Tornado Problems. In addition, the author wrote corresponding MyLab problems around this data set. The problems may serve as a semester-long project for your students.
- **Updated MyLab Problems** New MyLab problems written by Michael Sullivan utilize real data that is randomly generated from a larger data set. He also wrote new applet exercises that allow students to explore statistical concepts.
- **Optional Simulation & Randomization Sections** Simulation and randomization methods are a new approach to hypothesis testing. New to this edition are optional sections on using simulation to test hypotheses for a population proportion (Section 10.2A) and population mean (Section 10.3A), and randomization methods for testing hypotheses on two independent proportions (Section 11.1A), two independent means (Section 11.3A), and the slope of the least-squares regression model (Section 12.3A).
- **Classroom Notes** Written by Heidi Lyne and Michael Sullivan, new to this edition are classroom notes, which may be used by the instructor to deliver lectures to students. Students may print these notes out and bring them to the classroom, which facilitates good note-taking and allows them to focus on the concepts. The examples and activities in the classroom notes are different from those in the text and Instructor's Resource Guide.
- **Videos** New lightboard videos featuring the author, Michael Sullivan, develop statistical concepts for students. New animated videos explain concepts or tie material learned earlier in the course with the upcoming chapter or section. And finally, new Excel video solutions for any example in which Excel may be used to obtain statistical results are available.
- **R Technology Guide** Written by Patrick Murphy (nephew of the author) and Michael Sullivan, the R Technology Guide provides a chapter-by-chapter discussion of R commands needed for each topic. The R Technology Guide may be found under Learning Tools in MyLab or at www.sullystats.com/statistics-6e/r-guidebook.
- **Learning Catalytics** Learning Catalytics allows students to use their own mobile devices in the classroom for real-time engagement. Search "SullivanStats" in Learning Catalytics to add pre-made questions written by Michael Sullivan for Sullivan's *Statistics* series.

Hallmark Features

- **Putting It Together** When students are learning statistics, they often struggle with seeing the big picture of how it all fits together. One of my goals is to help students learn not just the important concepts and methods of statistics but also how to put them together and see how the methods work together. On the inside front cover, you'll see a pathway that provides a guide for students as they navigate through the process of learning statistics. The features and chapter organization in the sixth edition reinforce this important process. There are two categories of "Putting It Together."
  - **Putting It Together Sections** appear in Chapters 5, 9, 10, and 11. The problems in these sections are meant to help students identify the correct approach to solving a problem. Many exercises in these sections mix in inferential techniques from earlier sections. Plus, there are problems that require students to identify the inferential technique that may be used to answer the research objective (but no analysis is required). For example, see Problems 20 to 25 in Section 10.4.
  - **Putting It Together Problems** appear throughout the text. The purpose of these problems is to tie concepts together and see the entire statistical process. For example, problems on hypothesis testing may require students to first identify the data collection method (such as observational study or designed experiment, the explanatory and response variables, the role of randomization, the role of control) prior to completing the data analysis.
- **Student Activity Workbook** The student activity workbook now contains an outline for a semester-long project and suggestions for how to use the StatCrunch survey tool to develop a survey that could result in a semester-long project. Plus, there are ten new activities included in the activity workbook along with suggested answers in the corresponding instructor's guide.
- **Retain Your Knowledge** These problems occur periodically at the end of section exercises and are meant to assist students in retaining skills learned earlier in the course. This way, the material is fresh for the final exam.
- **MyLab Technology Help** Online homework problems that may be analyzed using statistical packages now have an updated technology help feature. Marked with a icon, this feature provides step-by-step instructions on how to obtain results using StatCrunch, TI-84 Plus/TI-84 Plus CE, and Excel.
- **Instructor's Resource Guide** Written by Michael Sullivan, the Instructor's Resource Guide provides an overview of the chapter. It also details points to emphasize within each section and suggestions for presenting the material. In addition, the guide provides examples that may be used in the classroom. Many new examples have been added to this edition.
- Because the use of **Real Data** piques student interest and helps show the relevance of statistics, great efforts have been made to extensively incorporate real data in the exercises and examples.
- **Step-by-Step Annotated Examples** guide a student from problem to solution in three easy-to-follow steps.
- **"Now Work"** problems follow most examples so students can practice the concepts shown.
- Multiple types of **Exercises** are used at the end of sections and chapters to test varying skills with progressive levels of difficulty. These exercises include **Vocabulary and Skill Building**, **Applying the Concepts**, and **Explaining the Concepts**.
- **Chapter Review** sections include:
  - **Chapter Summary**.
  - A list of key chapter **Vocabulary**.
  - A list of **Formulas** used in the chapter.
  - **Chapter Objectives** listed with corresponding review exercises.
  - **Review Exercises** with all answers available in the back of the book.
  - **Chapter Test** with all answers available in the back of the book. In addition, the Chapter Test problems have **video solutions** available.
- Each chapter has **Case Studies** available at www.pearsonhighered.com/sullivanstats that help students apply their knowledge and promote active learning.

Integration of Technology

This book can be used with or without technology. Should you choose to integrate technology in the course, the following resources are available for your students:

- Technology Step-by-Step guides are included in applicable sections that show how to use Minitab®, Excel®, the TI-83/84, and StatCrunch to complete statistics processes. The Technology Step-by-Step for StatCrunch was written by Michael Sullivan.
- Any problem that has 12 or more observations in the data set has a icon indicating that data set is included on the companion website (http://www.pearsonhighered.com/sullivanstats) in various formats.
- Where applicable, exercises and examples incorporate output screens from various software including Minitab, the TI-83/84 Plus CE, Excel, and StatCrunch.
- Applets are included on the companion website and connected with certain activities from the Student Activity Workbook, allowing students to manipulate data and interact with animations.
- A technology manual is available that contains detailed tutorial instructions and worked out examples and exercises for the TI-83/84. There is also a new R Technology Manual should you choose to incorporate R into your class.

Companion Website Contents

The companion website is http://www.pearsonhighered.com/sullivanstats.

- Data Sets
- Applets
- Formula Cards and Tables in PDF format
- Additional Topics Folder including:
  - Sections 9.4, 9.5, 10.2A, 10.2B, 10.3A, 10.5, 11.1A, 11.2A, 11.3A, 11.5, 12.3A
  - Appendix A
  - Appendix B
    - Lines
    - Inference about Two Proportions (Dependent Samples)
    - Comparing Three or More Means (One-Way Analysis of Variance)
- **Case Studies** for each chapter in the text.
- A copy of the questions asked on the Sullivan Statistics Survey I and Survey II
- Consumer Reports projects that were formerly in the text
- The author has also created a website at https://www.sullystats.com. This site has chapter-by-chapter suggestions for teaching the material, links to interesting data sets, and much more.

Key Chapter Content Changes

Chapter 1 Data Collection

Section 1.2 now includes a discussion of obtaining data through web scraping and how to obtain data from the Internet. Section 1.6 expands on the discussion of the placebo effect.

Chapter 5

Section 5.1 now distinguishes the Law of Large Numbers from the nonexistent Law of Averages. There is a new Section 5.6 on simulating probability experiments. This material is very helpful in allowing students to see the role of randomness in probability experiments. It also foreshadows topics such as sampling distributions and inference.

Chapter 9

There is an expanded discussion on the normality condition for constructing confidence intervals for the population mean using Student's *t*-distribution in Section 9.2.

Chapter 10

Chapter 10 now contains optional sections on simulation methods for conducting inference. The organization of Chapter 10 allows for presenting simulation along with traditional inference, or simply presenting traditional inference. Should you decide to present only the traditional approach to inference, simply cover Section 10.2 from the text. If you decide to present hypothesis testing using simulation, skip Section 10.2 in the text and cover Sections 10.2A and 10.2B (available in MyLab or the companion website as pdfs). Section 10.3A (MyLab) presents hypothesis testing on a mean using simulation and bootstrapping. This section is optional and may be skipped without loss of continuity.

Chapter 11

Chapter 11 has new optional sections on randomization methods. Section 11.1A (available in MyLab or the companion website as a pdf) presents randomization tests for two independent proportions. If you choose to present randomization methods, we recommend presenting Section 11.1A prior to Section 11.1. Section 11.2A (MyLab) presents hypothesis tests on dependent means using bootstrapping. This section is optional and may be skipped without loss of continuity. Section 11.3A (MyLab) presents randomization tests for two independent means. We recommend covering this section prior to Section 11.3, if you choose to discuss this approach to hypothesis testing.

Chapter 12

Chapter 12 has a new optional section on randomization. Section 12.3A (available in MyLab or the companion website as a pdf) presents randomization tests for the slope of the least-squares regression model. If you choose to cover this section, do so prior to Section 12.3.

Flexible to Work with Your Syllabus

To meet the varied needs of diverse syllabi, this book has been organized to be flexible.

You will notice the "Preparing for This Section" material at the beginning of each section, which will tip you off to dependencies within the course. The two most common variations within an introductory statistics course are the treatment of regression analysis and the treatment of probability.

- **Coverage of Correlation and Regression** The text was written with the descriptive portion of bivariate data (Chapter 4) presented after the descriptive portion of univariate data (Chapter 3). Instructors who prefer to postpone the discussion of bivariate data can skip Chapter 4 and return to it before covering Sections 12.3 and 12.4.
- **Coverage of Probability** The text allows for light to extensive coverage of probability. Instructors wishing to minimize probability may cover Section 5.1 and skip the remaining sections. A mid-level treatment of probability can be accomplished by covering Sections 5.1 through 5.3. Instructors who will cover the chi-square test for independence will want to cover Sections 5.1 through 5.3. In addition, an instructor who will cover binomial probabilities will want to cover independence in Section 5.3 and combinations in Section 5.5.

Acknowledgments

Textbooks evolve into their final form through the efforts and contributions of many people. First and foremost, I would like to thank my family, whose dedication to this project was just as much as mine: my wife, Yolanda, whose words of encouragement and support were unabashed, and my children, Michael, Kevin, and Marissa, who have been supportive throughout their childhood and now into adulthood (my how time flies). I owe each of them my sincerest gratitude. I would also like to thank the entire Mathematics Department at Joliet Junior College and my colleagues who provided support, ideas, and encouragement to help me complete this project. From Pearson Education: I thank Suzanna Bainbridge, whose editorial expertise has been an invaluable asset; Peggy McMahon, who provided organizational skills that made this project go smoothly; Alicia Wilson and Demetrius Hall, for their marketing savvy and dedication to getting the word out; Vicki Dreyfus and Jean Choe, for their dedication in organizing all the media; Rose Kernan for her ability to control the production process; Dana Bettez for her editorial skill with the Instructor's Resource Guide and Student Activity Workbook; and the Pearson sales team, for their confidence and support of this book.

I also want to thank Ryan Cromar, Susan Herring, Craig Johnson, Kathleen McLaughlin, Patrick Murphy, Heidi Lyne, and Dorothy Wakefield for their help in creating supplements. A big thank-you goes to Cindy Trimble and Associates, who assisted in verifying answers for the back of the text and helped in proofreading. I would also like to acknowledge Kathleen Almy and Heather Foes for their help and expertise in developing the Student Activity Workbook. Finally, I would like to thank George Woodbury for helping me with the incredible suite of videos that accompanies the text. Many thanks to all the reviewers, whose insights and ideas form the backbone of this text. I apologize for any omissions.

CALIFORNIA *Charles Biles, Humboldt State University • Carol Curtis, Fresno City College • Jacqueline Faris, Modesto Junior College • Freida Ganter, California State University–Fresno • Jessica Kramer, Santiago Canyon College • Sherry Lohse, Napa Valley College • Craig Nance, Santiago Canyon College • Diane Van Deusen, Napa Valley College* **COLORADO** *Roxanne Byrne, University of Colorado–Denver • Monica Geist, Front Range Community College* **CONNECTICUT** *Kathleen McLaughlin, Manchester Community College • Dorothy Wakefield, University of Connecticut • Cathleen M. Zucco Teveloff, Trinity College* **DISTRICT OF COLUMBIA** *Monica Jackson, American University • Jill McGowan, Howard University* **FLORIDA** *Randall Allbritton, Daytona Beach Community College • Greg Bloxom, Pensacola State College • Anthony DePass, St. Petersburgh College Clearwater • Kelcey Ellis, University of Central Florida • Franco Fedele, University of West Florida • Laura Heath, Palm Beach Community College • Perrian Herring, Okaloosa Walton College • Marilyn Hixson, Brevard Community College • Daniel Inghram, University of Central Florida • Philip Pina, Florida Atlantic University • Mike Rosenthal, Florida International University • James Smart, Tallahassee Community College* **GEORGIA** *Virginia Parks, Georgia Perimeter College • Chandler Pike, University of Georgia • Jill Smith, University of Georgia • John Weber, Georgia Perimeter College* **HAWAII** *Eric Matsuoka at Leeward Community College • Leslie Rush, University of Hawaii* **IDAHO** *K. Shane Goodwin, Brigham Young University • Craig Johnson, Brigham Young University • Brent Timothy, Brigham Young University • Kirk Trigsted, University of Idaho* **ILLINOIS** *Grant Alexander, Joliet Junior College • Kathleen Almy, Rock Valley College • John Bialas, Joliet Junior College • Linda Blanco, Joliet Junior College • Kevin Bodden, Lewis & Clark Community College • Rebecca Goad, Joliet Junior College • Joanne Brunner, Joliet Junior College • Robert Capetta, University of Illinois at Chicago • Elena Catoiu, Joliet Junior College • Faye Dang, Joliet Junior College • Laura Egner, Joliet Junior College • Jason Eltrevoog, Joliet Junior College • Erica Egizio, Lewis University • Heather Foes, Rock Valley College • Randy Gallaher, Lewis & Clark Community College • Melissa Gaddini, Robert Morris University • Iraj Kalantari, Western Illinois University • Donna Katula, Joliet Junior College • Diane Koenig, Rock Valley College • Diane Long, College of DuPage • Heidi Lyne, Joliet Junior College • Jean McArthur, Joliet Junior College • Patricia McCarthy, Robert Morris University • David McGuire, Joliet Junior College • Angela McNulty, Joliet Junior College • James Morgan, Joliet Junior College • Andrew Neath, Southern Illinois University-Edwardsville • Linda Padilla, Joliet Junior College • David Ruffato, Joliet Junior College • Patrick Stevens, Joliet Junior College • Robert Tuskey, Joliet Junior College • Stephen Zuro, Joliet Junior College* **INDIANA** *Susitha Karunaratne, Purdue University North Central • Jason Parcon, Indiana University–Purdue University Ft. Wayne • Henry Wakhungu, Indiana University* **KANSAS** *Donna Gorton, Butler Community College • Ingrid Peterson, University of Kansas* **LOUISIANA** *Melissa Myers, University of Louisiana at Lafayette* **MARYLAND** *Nancy Chell, Anne Arundel Community College • John Climent, Cecil Community College • Rita Kolb, The Community College of Baltimore County • Jignasa Rami, Community College of Baltimore County • Mary Lou Townsend, Wor-Wic Community College* **MASSACHUSETTS** *Susan McCourt, Bristol Community College • Daniel Weiner, Boston University • Pradipta Seal, Boston University of Public Health* **MICHIGAN** *Margaret M. Balachowski, Michigan Technological University • Diane Krasnewich, Muskegon Community College • Susan Lenker, Central Michigan University • Timothy D. Stebbins, Kalamazoo Valley Community College • Sharon Stokero, Michigan Technological University • Alana Tuckey, Jackson Community College* **MINNESOTA** *Mezbhur Rahman, Minnesota State University* **MISSOURI** *Farroll Tim*

Wright, University of Missouri–Columbia **NEBRASKA** *Jane Keller, Metropolitan Community College* **NEW YORK** *Jacob Amidon, Finger Lakes Community College • Stella Aminova, Hunter College • Jennifer Bergamo, Onondaga Community College • Kathleen Cantone, Onondaga Community College • Pinyuen Chen, Syracuse University • Sandra Clarkson, Hunter College of CUNY • Rebecca Daggar, Rochester Institute of Technology • Bryan Ingham, Finger Lakes Community College • Anne M. Jowsey, Niagara County Community College • Maryann E. Justinger, Erie Community College–South Campus • Bernadette Lanciaux, Rochester Institute of Technology • Kathleen Miranda, SUNY at Old Westbury • Robert Sackett, Erie Community College–North Campus • Sean Simpson, Westchester Community College • Bill Williams, Hunter College of CUNY* **NORTH CAROLINA** *Fusan Akman, Coastal Carolina Community College • Mohammad Kazemi, University of North Carolina–Charlotte • Janet Mays, Elon University • Marilyn McCollum, North Carolina State University • Claudia McKenzie, Central Piedmont Community College • Said E. Said, East Carolina University • Karen Spike, University of North Carolina–Wilmington • Jeanette Szwec, Cape Fear Community College* **NORTH DAKOTA** *Myron Berg, Dickinson State University • Ronald Degges, North Dakota State University* **OHIO** *Richard Einsporn, The University of Akron • Michael McCraith, Cuyaghoga Community College* **OREGON** *Daniel Kim, Southern Oregon University • Jong Sung Kin, Portland State University* **PENNSYLVANIA** *Mary Brown, Harrisburg Area Community College • LeAnne Conaway, Harrisburg Area Community College* **SOUTH CAROLINA** *Diana Asmus, Greenville Technical College • Dr. William P. Fox, Francis Marion University • Cheryl Hawkins, Greenville Technical College • Rose Jenkins, Midlands Technical College • Lindsay Packer, College of Charleston • Laura Shick, Clemson University • Erwin Walker, Clemson University* **TENNESSEE** *Tim Britt, Jackson State Community College • Nancy Pevey, Pellissippi State Technical Community College • David Ray, University of Tennessee–Martin* **TEXAS** *Edith Aguirre, El Paso Community College • Ivette Chuca, El Paso Community College • Aaron Gutknecht, Tarrant County College • Jada Hill, Richland College • David Lane, Rice University • Alma F. Lopez, South Plains College • Shanna Moody, University of Texas at Arlington • Jasdeep Pannu, Lamar University* **UTAH** *Joe Gallegos, Salt Lake City Community College • Alia Maw, Salt Lake City Community College* **VIRGINIA** *Kim Jones, Virginia Commonwealth University • Vasanth Solomon, Old Dominion University* **WEST VIRGINIA** *Mike Mays, West Virginia University* **WISCONSIN** *William Applebaugh, University of Wisconsin–Eau Claire • Carolyn Chapel, Western Wisconsin Technical College • Beverly Dretzke, University of Wisconsin–Eau Claire • Jolene Hartwick, Western Wisconsin Technical College • Thomas Pomykalski, Madison Area Technical College • Walter Reid, University of Wisconsin-Eau Claire*

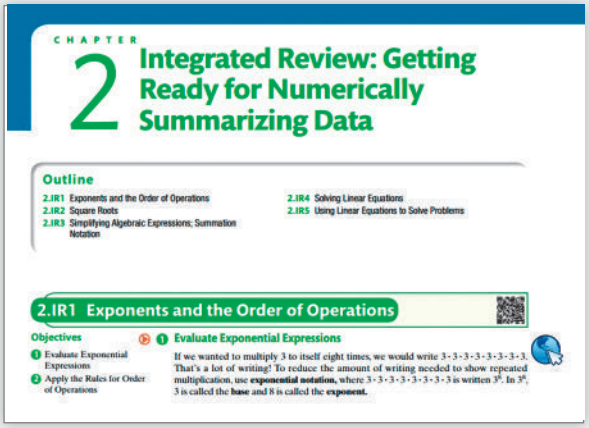
Michael Sullivan, III
Joliet Junior College

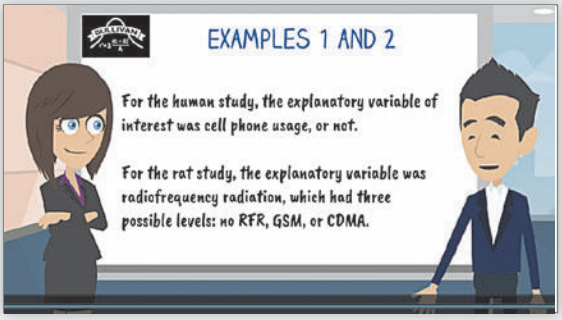
MyLab Statistics is available to accompany Pearson's market-leading text options, including Fundamentals of Statistics, 6e by Michael Sullivan (access code required).

MyLab™ is the teaching and learning platform that empowers you to reach every student. MyLab Statistics combines trusted author content—including full eText and assessment with immediate feedback—with digital tools and a flexible platform to personalize the learning experience and improve results for each student. Integrated with StatCrunch®, a web-based statistical software program, students learn the skills they need to interact with data in the real world.

Integrated Review for Corequisite Courses

This MyLab™ includes an additional eText written by the author on prerequisite skills and concepts. There are also prebuilt (and editable) MyLab quizzes that populate personalized homework assignments for gaps in skills for that chapter. These resources may be used in a corequisite course model, or simply to help underprepared students master prerequisite skills and concepts.





NEW! Videos

In addition to existing Author in the Classroom, StatTalk, and Example Videos, the following **video types** were added to this edition.

- **Innovative lightboard videos** featuring Mike Sullivan guide students towards deeper conceptual understanding of certain key topics.
- **Excel video solutions** were added to the existing suite of StatCrunch, TI-83/84 Plus, and by-hand videos for examples in the text.
- **Animation videos** remind students of where they have been and where they are going in the upcoming objectives.

pearson.com/mylab/statistics

Resources for Success

Pearson MyLab

Student Resources

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

StatCrunch

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

Data Sets

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

Statistical Software Support

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

Student Solutions Manual

This manual provides detailed, worked-out solutions to all odd-numbered text exercises, as well as all solutions for the Chapter Reviews and Chapter Tests. It is available in print and can be downloaded from MyLab Statistics. (ISBN-13: 9780136969761)

Instructor Resources

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

Annotated Instructor's Edition

Includes answers to all text exercises, as well as teaching tips and common student errors. (ISBN-13: 9780136807315; ISBN-10:0136807313)

Instructor Solutions Manual

Contains worked-out solutions to all text exercises and case study answers. It can be downloaded from MyLab Statistics or from www.pearson.com.

PowerPoint Lecture Slides

PowerPoint Lecture Slides include key graphics and follow the sequence and philosophy of the text. They can be downloaded from MyLab Statistics or from www.pearson.com.

TestGen

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

Data Analytics / Early Alerts

Instructors have a comprehensive gradebook with enhanced reporting functionality that makes it easier to understand which students are struggling, and which topics they struggle with most. **New Early Alerts** use predictive analytics to identify struggling students as early as possible—even if their assignment scores are not a cause for concern.

Resources for Success

Pearson MyLab

Instructor Resources (continued)

Question Libraries

MyLab Statistics includes a number of question libraries that instructors can incorporate into their regular assignments. StatCrunch Projects consist of questions about large data sets in StatCrunch. The Conceptual Question Library offers 1,000 questions to help students learn concepts and how to think critically. Finally, the StatTalk Video Library includes questions associated to the video series by statistician Andrew Vickers.

Minitab and Minitab Express™

Bundling Minitab software with educational materials ensures students have access to the software they need in the classroom, around campus, and at home. And having 12-month access to Minitab and Minitab Express ensures students can use the software for the duration of their course. (ISBN-13: 9780134456409; ISBN-10: 0134456408) (access card only; not sold as stand alone.)

JMP Student Edition

An easy-to-use, streamlined version of JMP desktop statistical discovery software from SAS Institute, Inc. is available for bundling with the text. (ISBN-13: 9780134679792; ISBN-10: 0134679792)

XLSTAT™

An Excel add-in that enhances the analytical capabilities of Excel. XLSTAT is used by leading businesses and universities around the world. It is available to bundle with this text. For more information go to www.pearsonhighered.com/xlstatupdate. (ISBN-13: 9780321759320; ISBN-10: 032175932X)

Accessibility

Pearson works continuously to ensure our products are as accessible as possible to all students. We are working toward achieving WCAG 2.0 Level AA and Section 508 standards, as expressed in the Pearson Guidelines for Accessible Educational Web Media, www.pearson.com/mylab/statistics/accessibility.

Applications Index

Accounting
client satisfaction, 25–27

Aeronautics
moonwalkers, 11
O-ring failures on Columbia, 114
Spacelab, 503

Agriculture
corn production, B24
optimal level of fertilizer, 48–49
orchard damage, 59
yield
  of orchard, 37
  soybean, 134, B24

Animals/Nature
American black bears, weight and length of, 182–183, 185, 205, 567, 573
shark attacks, 91–92, 98, 219

Anthropometrics
upper arm length, 341
upper leg length, 341, 380

Astronomy
life on Mars, 298

Banking
ATM withdrawals, 382, 465
credit-card debt, 409, 474
credit cards, 390, 471

Biology
alcohol effects, 53, 58
alopecia, 497
blood types, 237
cholesterol level, 38, 520
DNA sequences, 280, 298
growth plates, 352
HDL cholesterol, 423, 568
hemoglobin
  in cats, 154
LDL cholesterol, B27
reaction time, 53, 58, 497–498, 510
testosterone levels, 472, 509

Business. *See also* Work
acceptance sampling, 267, 281, 390
advertising
  campaign, 37
  effective commercial, 101
  humor in, 58
airline customer opinion, 36
bolts production, 57, 152–153
buying new cars, 537
car rentals, 499
car sales, 86
CEO performance, 183, 198, 567, 573
coffee sales, 296
customer satisfaction, 32–33
customer service, 359
Disney World statistics conference, 240
employee morale, 37
entrepreneurship, 393
name vs. store brand, 488
new store opening decision, 42
oil change time, 381
online groceries, 74
packaging error, 267, 281, 296
quality control, 36, 37, 59, 258, 393–394, 471
salaries, 120, 136
shopping habits of customers, 42
shopping online, 73, 118–119, 154
Speedy Lube, 354
stocks on the NASDAQ, 280
stocks on the NYSE, 280
Target demographic information gathering, 38
traveling salesperson, 280
unemployment and inflation, 87
union membership, 98
waiting in line, 309, 498
waiting time for restaurant seating, 89
worker injury, 99
worker morale, 30

Chemistry
calcium in rainwater, 467
diversity and pH, B28
pH in rain, 421–422, 521, B27
pH in water, 117, 134
reaction time, 340, 497–498, 510

Combinatorics
arranging flags, 296
clothing option, 280
combination locks, 280
committee, 267
committee formation, 280
committee selection, 281
license plate numbers, 280, 296
passwords, 298
seating arrangements, 291
starting lineups, 291

Communication(s)
cell phone, 58
  in bathroom, 408
  body mass index and, 22–23
  brain tumors and, 14–15
  conversations, 488
  crime rate and, 186
  servicing, 441
  screen time, 421
do-not-call registry, 43
e-mail, 430
high-speed Internet service, 57, 408
length of phone calls, 340
newspaper article analysis, 13–14, 22, 221–222
social media, 252, 267
teen, 268
text messaging
  number of texts, 73
  while driving, 471
voice-recognition systems, B10–B11

Computer(s). *See also* Internet
defective chips, 269
download time, 37
DSL Internet connection speed, 37
e-mail, 430
FBI ID numbering, 281–282
fingerprint identification, 259
passwords, 282
toner cartridges, 168
user names, 280

Construction
concrete, 205
concrete mix, 116, 133
new road, 104

Consumers
Coke or Pepsi preferences, 54
taste test, 21

Crime(s)
burglaries, 92–93
fingerprints, 259
fraud identity, 71–72
larceny theft, 238–239
murder conviction by race, 216
rate of cell phones, 186
robberies, 98
speeding, 38
weapons used in murder/homicide, 101

Criminology
FBI ID numbering, 281–282
fraud detection, 154, 155

Demographics
births
  live, 101–102, 310
  per capita income and rates of, 105
  per woman, 91
  proportion born each day of week, B25
childless women, 390
family size, 100
handedness and mortality, 23
households speaking foreign language as primary language, 42
left-handedness, 410
life expectancy, 9, 91, 185, 258
living alone, 455, 537
marital status and happiness, 216
number of live births, 50- to 54-year-old mothers, 310
population
  age of, 145
  of selected countries, 9
  shifts in, 536
southpaws, 257

Drugs. *See also* Pharmaceuticals
AndroGel, 408
Aspirin, 552–553
Celebrex, 551
experimental, 324
marijuana use, 291
Nexium, 454
thalidomide on TEN, 327
Viagra, 241
Zoloft, 521

Economy
abolishing the penny, 408
health care expenditures, 99
poverty, 71
unemployment and inflation, 87
unemployment rates, 220

Education. *See also* Test(s)
advanced degrees, 393
bachelor's degree, elapsed time to earn, 509

birthrate and, 181
board work, 267
bullying in schools, 11
calculus exam score, 11
class average, 144
college
application, 168
campus safety, 31
complete rate, 441, 490
drug use among students, 11
exam skills, 512
literature selection, 29
survey, 72–73, 238
textbook packages required, 42
corequisite remediation and study skills, 552, 553
course redesign, 471
course selection, 29
day care, 3-year-old, 237
delivery methods, B24–B25
developmental math, 51
dropping course, 552
exam grades/scores, 58, 117, 119, 133, 205–206, 253
study time, 196
exam time, 116, 133
faculty evaluation, 186
faculty opinion poll, 29
gender bias in grades, 516–517
GPA, 103, 144, 167
vs. seating choice, 577–578
grade distribution, 536
grade inflation, 103
graduation rates, 186, 199, 432, 512
health and, 550
illicit drug use among students, 11, 42
income and, 196, 215
invest in, 567, 573
journal costs, 118
learning community curriculum, 455–456
level of (educational attainment), 96–97, 215
marriage and, 292
mathematics
studying college, 428, 475
teaching, 455
TIMMS exam, 182
TIMS report and Kumon, 514
music's impact on learning, 49–50
online homework, 75, 512
premature birth and, 577
quality of, 455
reading and math ability of fourth-graders, 407
school
admissions, 152
dropouts, 266
e-cigs usage, 442, 551
enrollment, 88
illegal drug use in, 11
multitasking, 515–516
National Honor Society, 292
seat selection in classroom, 296
student loans, 359, 471
seating arrangements, 534–535, 577–578
self-injurious behaviors, 330
self-study format, 475
sleep disorders and academic performance, 521
student age, 469
student loans, 90
student opinion poll/survey, 29, 37, 38
student retention, 475
student services fees, 38
study time, 468
teacher evaluations, 515
teaching reading, 51
time spent on homework, 105
typical student, 106
visual vs. textual learners, 511

Electricity

Christmas lights, 257
light bulbs, 166, 281

Electronics

televisions in the household, 89

Energy

carbon dioxide emissions and energy production, 196–197
consumption, 466
gas price hike, 100
oil reserves, 99
during pregnancy, 392–393

Engineering

batteries and temperature, 54
battery charge life, 393
bearing failure rate, 168
bolts production, 57
concrete strength, 566, 573, B27
driving under the influence (DUI) simulator, 499
glide testing, 500–501
hardness testing, 498–499
linear rotary bearing, 474
O-ring thickness, 359
Prolong engine treatment, 442
ramp metering, 510
tire design, 54
triple modular redundancy, 258
valve pressure, 441

Entertainment. *See also* Leisure and recreation

Academy Award winners, 382
Disney's Dinosaur Ride, 359
movie popcorn, 510
movie ratings, 56
neighborhood party, 267
People Meter measurement, 35
raffle, 13
social drinking, 59
student survey, 154
television
in bedroom, obesity and, 21
hours of watching, 359, 423
luxury or necessity, 407
number of, 393
watching, 382
theme park spending, 427
tickets to concert, 24–25

Environment

Flint water crisis, 162
pH in rain, 421–422, B27
Secchi disk, 432, 498

Exercise

gender differences in, 490
routines, 292

Family

gender income inequality, 456
ideal number of children, 90, 102, 310, 431, 472
imprisoned members of, 432
infidelity among married men, 455
smarter kids, 471
spanking, 326
structure, 250, 550
values, 407

Farming. *See also* Agriculture

incubation times for hen eggs, 340–341, 352, 353

Fashion

women's preference for shoes, 103

Finance. *See also* Investment(s)

ATM withdrawals, 382, 465–466
cash/credit, 514–515
cigarette tax rates, 90, 144, 206
cost
of kids, 99
credit-card debt, 409, 474
credit cards, 311, 390, 471
credit scores, 205, 552, 565, 573
deficit reduction, 408, 489
derivatives, 258
estate tax returns, 427
federal debt, 90
FICO credit score, 182, 197, 466, 565
Gini index, 89
health care expenditures, 99
income
adjusted gross income, 103
age vs., 186
average, 89
children vs. parents, 258
distribution, 103, 169
educational attainment and, 215
gender inequality in, 456
household, 33–34, 42
median, 89, 98, 181
per capita, 105
by region, 266
student survey, 154
taxes, 93–94, B10
IRS audits, 259
loan application, 552
net worth, 120, 394
property tax, 428
retirement savings, 442–443
stock analysis, 454
stock market, 183–184, 198
student loans, 90
taxing, 422
tax rates, 93–94, 423, 511
tax revenue, 99

Firearms

gun laws, B11
muzzle velocity, 166, 427, 497

Food. *See also* Nutrition

accuracy of drive thru orders, 455
allergies, 394
candy weight, 118, 134, 162
cauliflowers, 409
cheeseburgers, fat and calories, 218–219
chewing number and consumption, 50–51, 428
chocolates, 144

consumption of popcorn, 442
cookies
Chips Ahoy, 359
chocolate chip, 162, 352, 353–354
diameter of, 102
dining out, 73
fast-food restaurants, 354
insect fragments, 381
McDonald's drive-through traffic, 104
meatloaf, 168
M&M, 533–534
number of drinks, 424–425
nut mix, 144
para-nonylphenol in processing and packaging of, 466
peanuts, 534
pizza, 441
popcorn, 510
priming for healthy food, 53–54
quality control, 442
soda preferences, 457
takeout, 239
tea tasting, 259
time spent eating and drinking, 421
Tootsie Pop, 421
Yelp ratings for restaurant, 104

Gambling. *See also* Game(s)
betting on sports, 280
craps, 286
lotteries, 280
double jackpot, 257
instant winner, 292
Pick 3, 296
Pick 4, 296
Pick 5, 296
Powerball, 311
state, 296
roulette, 238, 251, 295, 311, 383, 576
video poker, 311

Game(s). *See also* Gambling
Blackjack, 310, 353
card drawing, 250, 266, 281, 296, 324
coin toss, 237, 257
Dictator Game, 54
die/dice, 537
loaded, 240
rolling, 86, 237, 240, 257, 287
five-card stud, 296
Lingo, 292
Little Lotto, 281
Mega Millions, 281
poker
flush, 268–269
royal flush, 269
seven-card stud, 237
three-card, 329
winning, 381
The Price Is Right, 286
Solitaire, 296
Text Twist, 292

Gardening
planting tulips, 268

Gender
lupus and, 457
wage gap, 517
weight gain and, 259

Genetics
Huntington's disease, 238
sickle-cell anemia, 238

Geography
highest elevation for continents, 75

Geology
density of Earth, 366
earthquakes, 88
Old Faithful geyser (Calistoga, California), 118, 135, 162, 205
Old Faithful geyser (Yellowstone Park), 168, 381

Government
federal debt, 90
IRS audits, 259
New Deal policies, 407
Social Security numbers, 280
Social Security reform, 390
state, 11
trust in, 326
type of, 9
waste, 12

Health. *See also* Exercise; Medicine
alcohol abstention, 488
alcohol dependence treatment, 51
alcohol effects on brain, 466
allergy sufferers, 325, 326
blood alcohol concentration, 119
blood types, 237
body mass index, 488
bone mineral density and cola consumption, 60, 198
brain tumors and cell phones, 14–15
burning calories, 98
calories vs. sugar, 578
cancer, 21
breast, 268
cell phones and brain tumors, 14–15
cholesterol, 38
death in, 266
lung, 17, 23
passive smoke and lung cancer, 23
power lines and, 22
skin, coffee consumption and, 21
survival rates, 119
cardiac arrest, 342–343
cholesterol levels and green tea, 53
commuting time and, 136, 166, 182, 197, 205
doctor visits, 251
drug side effects, 241
education and, 550
effect of Lipitor on cardiovascular disease, 46, 47
emergency room visit, 329, 474
exercises, 87
fitness club member satisfaction, 37
flu shots for seniors, 16
ginkgo and memory, 52
handwashing behavior, 409–410
happiness and, 21, 216, 550
hazardous activities, B10
headache, 161
health care expenditures, 99
hearing/vision problems, 251
heart attacks, 552–553
heart disease and baldness, 21
HIV test false positives, 257
hospital-acquired conditions, 250
hospital admissions, 119, 516
hygiene habits, 11
hypertension, 12, 54, 431
insomnia, 52
LDL cholesterol, B27
life expectancy, 185
Lipitor, 454
live births, 101–102, 144
lung cancer and, 17, 23
Lyme disease vs. drownings, 185
marriage/cohabitation and weight gain, 21
migraine, 442
obesity, 185
social well-being and, 550–551
television in the bedroom and, 21
osteoporosis treatment, 520
overweight, 42, 100, 441
pulse rates, 117, 134
self-injurious behaviors, 330
shrinking stomach and diet, 52
skinfold thickness procedure, 167
sleep disorders, 521
sleeping habits, 42, 471
smoking, 12, 268
birth weight, 199–200
cessation program, 424, 521
cigar, 251
e-cig study, B29
educational attainment and, 522
lung cancer and, 17, 23
paternal, 162–163
during pregnancy, 456
profile of, 551
survival rates, 216
tar and nicotine levels in cigarettes, 566, 573
weight gain and, 522
sneezing habits, 325, 364, 474
television stations and life expectancy, 185
testosterone levels, 472
tooth whitener, 51, 57
vitamins, 162
weight loss, role of diet and drugs in, 54
weight of college students, 42
women, aspirin, and heart attacks, 552–553

Height(s)
arm span vs., 519–520
father and son, 498
females
five-year-old, 342
20 years of age, 466
vs. males, 152
head circumference vs., 182, 197, 205, 565–566, 573
10-year-old males, 341

Houses and housing
apartments, 219, 291, 578
females living at home, 364
garage door code, 280
home ownership, 102, 408
household winter temperature, 144
increase in assessments, 391
males living at home, 364
number of people living in, 87
pricing, 199
rents, 310, 428
rooms, 250

single-family home price, 441
square footage, 143
Zestimate, 568

Insurance
collision claims, 514
credit scores and, 552
life, 310
Medicare fines, 250

Intelligence
brain size and, 184
IQ scores, 86, 119, 135, 136–137, 153, 184, 222, 430, 469, 472
predictions, 574

Internet
download time, 118
frequency of use of, 73
high speed access, 408
linear transformations, 120, 136
online dating, 287
online homework, 75, 512
online search, 58
time viewing a Web page, 91
Web page design, 22, 517

Investment(s)
attitudes toward, 13
bear markets, 567–568, 573
bull markets, 384
comparing stock sectors, 516
dispersion in the market, 490
diversification, 137, 185
in education, 567, 573
hot stock tips, 310
mutual funds, 135
rate of return on, 135, 138, 259, 327, 381, 511, B25
return on, 90
savings, 143
stock price, 74, 239, 408, 471–472
Super Bowl effect, 456
trade volume, 422
volume of stock
  Altria Group, 90
  PepsiCo, 422
  Starbucks, 468

Landscaping
golf course, 280

Language
foreign, 390
spoken at home, 251

Law(s)
chief justices, 166
death penalty, 409, 488
driver's license, 12
fair packaging and labeling, 441
gun control, 42
jury selection, 281, 325

Law enforcement
age of death-row inmates, 466
racial profiling, 535

Leisure and recreation. *See also* Entertainment
Boy Scouts merit badge requirement, 29
dining out, 73, 74
kids and, 511
Six Flags over Mid-America, 240

Manufacturing
ball bearings, 353
bolts production, 152–153
copper tubing, 393
products made in America, 72, 215, 266–267
Prolong engine treatment, 442
steel rods, 353
tire production, 366

Marriage
age and, 88, 196, 309
age difference, married couples, 514
couples at work, 251
divorce, 71
education and, 292
happiness and, 216
longevity, 266
infidelity/extramarital affairs, 42, 455, 515
testosterone's influence on, 509
unemployment rates, 220

Math
Benford's Law of frequency of digits, 534

Media
death penalty, 488

Medicine. *See also* Drugs; Health; Pharmaceuticals
abortion, 215–216
alcohol dependence treatment, 51
alcohol effects on brain, 466
allergy sufferers, 325
Alzheimer's disease treatment, 56
AndroGel, 408
baby delivery methods, 21
bacteria in hospital, 510
blood alcohol concentration, 119
blood types, 74, 237
Cancer Prevention Study II, 57
cardiac arrest, 342–343
carpal tunnel syndrome, 21
cholesterol level, 38, 105
cortical blindness, 455
Covid-19 vaccine, 54
drug side effects, 241
Ebola vaccine, 54
effect of Lipitor on cardiovascular disease, 46, 47
flu season, 71
folate and hypertension, 12
gum disease, 427
hair growth and platelet-rich plasma, 52–53
HDL cholesterol, 423, 568
heart attacks, 552–553
LDL cholesterol, B27
Lipitor, 454
live births, 144
lupus and, 457
Medicare fines, 250
metastatic melanoma, 489
migraine, 442
placebo effect, 54, 251–252
poison ivy ointments, B11
Salk vaccine, 490
side effects, 488–489
sleep apnea, 427
wart treatment, 12

Military
atomic bomb, protection from, 471
Iraq War, 489
night vision goggles, 57
peacekeeping missions, 37
satellite defense system, 259
V-2 rocket hits in London, 537

Miscellaneous
aluminum bottle, 512
birthdays, 238, 250, 268, 286
diameter of Douglas fir trees, 431
filling bottles, 467–468
fingerprints, 259
journal article results, B28
random-number generator, 340
relationship deal-breakers, 489–490
reproducibility of primary studies, 455
selling yourself, 70–71
sleeping, 384, 421
tattoos, 488
toilet flushing, 95, 325–326, 364
wet suits, 515, 517

Money. *See also* Finance; Investment(s)
abolishing the penny, 408
cash/credit, 514–515
credit-card debt, 474
FICO credit score, 182, 197, 466
income taxes, B10
retirement and, 442–443

Morality
state of, in U.S., 287, 325, 364, 428
unwed women having children, 515

Mortality
bicycle deaths, 535
Gallup Organization, 325
pedestrian death, 536
Titanic disaster, 577

Motor vehicle(s). *See also* Transportation
accident
  fatal traffic, 454–455
  red-light camera programs, 252–253
autonomous vehicles, 427
blood alcohol concentration (BAC) for drivers involved in, 421, 454–455
BMWs, 12
braking distance, 499
buying car, 135, 137, 537
  discrimination in, B26
car accidents, 98
car color, 74, 325
carpoolers, 161
car prices, 166
car rentals, 499
collision coverage claims, 514
collision data, 521
crash data, B26–B27
crash test results, 422
driving under influence, 258–259
fatalities
  alcohol-related, 87
  driver, 252, 267
  traffic, 295
flight time, 116, 133
gas mileage/fuel economy, 471

gas prices, 100, 136
male vs. female drivers, 184, 199
miles per gallon, 116, 133, 342, 416–417
minimum driving age, 329–330
new cars, 292
new vs. used car, 219
octane in fuel, 499–500
oil change, 381
seat belts, 427
SMART car, 154
speeding tickets, 252, 267
SUV vs. car, 498
wearing helmets, 534

Music
arranging songs, 280
effect on learning, 49–50
playing songs, 267, 280, 281

Nutrition. *See also* Food
bone mineral density and cola consumption, 60, 198
caffeinated sports drinks, 431
calories
  burning of, 98
  cheeseburgers, 218–219
  vs. sugar, 578
dissolving rates of vitamins, 162
eating together, 455
fat in
  cheeseburgers, 218–219
green tea and cholesterol levels, 53, 428
overweight, 100, 441
salt and hypertension, 54

Obstetrics. *See also* Pediatrics
birth(s)
  by day of week, B25
  gestation period, 152, 341–342, 352–353, 380
  multiple, 239
  premature, 577
  weight, 341
prenatal care, 550
sleeping patterns during pregnancy, 475

Pediatrics. *See also* Obstetrics
age of mother at childbirth, 145, 161
birth weight, 144, 152, 199–200, 341, 354
  gestation period vs., 295
  maternal age and, 537
  36 weeks of, 340
crawling babies, 422
energy during pregnancy, 392–393
head circumference vs. heights, 182, 197, 205

Pets
talking to, 455

Pharmaceuticals. *See also* Drugs; Medicine
alcohol dependence treatment, 51
Aspirin, 552–553
Celebrex, 551
cholesterol research, 38
cold medication, 51
drug effectiveness, 53
Lipitor, 46, 454
memory drug, 52
Nexium, 454
Prevnar, 488
skin ointment, 60

Physics
muzzle velocity, 166, 427, 497

Politics
affiliation, 102, 268, 552
age and, B27
capitalism, 552
decisions, 471
elections
  county, 37
  predictions, 390, 515
  Senate, 312, 409
  simulation, 286–287, 326
estate taxes, 36
exit polls, 43
Future Government Club, 30, 37
health care and health insurance, 38
House of Representative gender composition, 410
mayor and small business owners, 57
poll, 38
presidents
  age at inauguration, 100, 161
  inaugural addresses, 167
  random sample of, 29
public knowledge about, 11
public policy survey, 59
pundit predictions, 454
questionnaire wording, 489
Republican voters, 456–457
Roosevelt vs. Landon, 577
socialism, 551
views, 119
village poll, 30
voter polls, 37, 38

Polls and surveys
abortion, 215–216
about gun-control laws, 42
annoying behavior, 427
blood donation, 407
boys are preferred, 364
children and childcare, B11
of city residents, 37
college, 72–73, 238
Current Population Survey, 43
on desirability attributes, 72, 215
dream job, 73
dropping course, 552
election, 44, 390
e-mail survey, 42
exit, 43
faculty opinion, 29
on family values, 407
on frequency of having sex, 42
gender of children in family, 280
gun control, 409
happiness and health, 216
on hazardous activities, B10
on high-speed Internet service, 57
informed opinion, 44
liars, 364
on life satisfaction, 390
on long life, 430, 469
on marriage being obsolete, 390
number of drinks, 424–425
order of the questions, 43–44
police department, 42
political, 38
population, 43
random digit dialing, 43
reading number of books, 423
registered voters, 324
response rate, 42, 43
retirement planning, 427
robocalling, 43
rotating choices, 43
sample independence, 269
seat belts, 427
speaker evaluation, 37
student opinion, 29, 30, 37
student sample for, 29
tattoos, 488
on televisions in the household, 89, 90, 309
TVaholics, 466
village, 30
wording of questions, 43
working hours, 420–421

Psychiatry
attention deficit-hyperactivity disorder, 390

Psychology
ESP and, 454
ideomotor action, 51–52
insomnia relief, 52
profiles and, B28
rationalized lies, 535–536
reaction time, B25–B26
risk handling, B11
stressful commute, 407, 565, 573

Psychometrics
IQ scores, 86, 119, 136, 153, 184, 222, 430, 469, 472

Reading
America reads, 366
at bedtime, 469
number of books read, 408, 410, 421, 423
rates, 352, 353, 380–381, 467
SAT scores, 442

Religion
in Congress, 534
teen prayer, 475
trust in, 295–296

Research
level of measurement in, 13

Sex and sexuality
family structure and, 550
sexual intercourse frequency, 42

Society
abortion issue, 215–216
affirmative action, 409
civic duty, 577
death penalty, 409, 488
divorce
  opinion regarding, 71
dog ownership, 259
marijuana use, 291
online dating, 287
path to success, 269
poverty, 71, 94–95, 144
racial profiling, 326
reincarnation belief, 390–391
social well-being and obesity, 550–551

superstition, 407
unwed women having children, 515
Valentine's Day, 407
volunteers and, 237

Sports

athletics participation, 326
baseball
batting averages, 11, 136, 152, 241
cold streaks, 258
ERA, 152
factory production, 366
fastball, 427, 467, 515
homeruns, 89, 183, 198, 205, 237, 296, 467, 553, 574
Ichiro's Hit Parade, 309
injuries, 249–250
jersey numbers, 13
most valuable player, 71, 100
no-hitter, 167
safety, 324
sprint speed of players, 88
starting lineup, 281
variability, 393
winning percentage, 185
World Series, 309, 576
basketball
free throws, 88, 237, 324
point spread, 218, 353
salaries, 121
betting on, 259
bowling, 258
caffeinated sports drinks, 431
car racing, INDY 500, 280
football
completion rate for passes, 56
extra point, 311
fans, 409
fumbles, 391
National Football League combine, 240, B27
spread accuracy, 456
Super Bowl effect on investing, 456
golf
balls, 268, 474
pitching wedge, 342
hockey
National Hockey League, 535
Stanley Cup, 329
human growth hormone (HGH) use among high school athletes, 36
organized play, 237
soccer, 96, 291
swimming, 152
team captains, 29
television commentator, 409
tennis, Wimbledon tournament, 432–433
triathlon, 152

Statistics

classifying probability, 240
coefficient of skewness, 137
coefficient of variation, 138
Fish Story, 134
geometric probability distribution, 326
in media, 456
midrange, 120
negative binomial probability distribution, 326–327
number of tickets issued, 166
practical significance, 468
probability, 237
shape, mean and median, 120
simulation, 286–287, 296, 311, 326, 409, 424, 457, 469
trimmed mean, 120

Temperature

household winter, 144
human, 474

Test(s)

ACT scores, 466
crash results, 422
essay, 292
FICO score, 120, 466, 565
IQ scores, 86, 119, 136, 153, 184, 222, 430, 472
multiple-choice, 58, 298
SAT scores, 135–136, 153, 184, 292, 311, 442, 466, 468
Wechsler Intelligence Scale, 366

Time

cab ride average, 163
drive-through service, 381, 382, 421, 466–467, 520
eating and drinking, 421
eruptions vs. length of eruption, 205
exam, 116, 133
flight, 116, 133, 324–325, 364
oil change, 381
online, 102–103
on phone, 421
reaction, 53, 58, 497–498, 510, B25–B26
study, 468
travel, 117–118, 134, 155
waiting, 90, 118, 154, 340, 422, 511, 520

Transportation. *See also* Motor vehicle(s)

alcohol-related traffic fatalities, 87
bike sharing, 383
fear of flying, 390
flight overbooking, 326
flight time, 324–325, 364
moving violations, 91
on-time flights, 324–325, 364
parking and camera violation fines, 79
red light cameras, 499
roundabout vs. four-way stop, 512
time spent in drive-through, 381
traffic lights, 239–240

Travel

airline reservations, 391
creative thinking during, 367
on-time flights, 324–325
taxes, 422
text while driving, 471
Titanic survivors, 577
walking in airport, 509–510

Weather

forecast, 292
hurricanes, 91, 183, 198, 205, 566, 573
Memphis snowfall, 359
temperatures, 135, 474
tornadoes, 13, 74–75, 91, 120, 138, 155, 163, 205, 240, 382, 408, 422–423, 456, 468, 489, 511, 573–574, B28

Weight(s)

American Black Bears, 185, 197–198, 205, 567, 573
birth, 340, 354
smoking and, 199–200
body mass index, 488
car vs. miles per gallon, 184–185, 198, 205
coins, 168
gaining, 259, 428
gestation period vs., 295
male vs. female, 167

Work. *See also* Business

commuting time, 136, 166, 182, 197, 205, 239
employee morale, 30
getting to, 251
married couples, 251
multiple jobs, 268
rate of unemployment, 220
unemployment, 87
walk to, 409

PART 1

Getting the Information You Need

Statistics is a process—a series of steps that leads to a goal. This text is divided into four parts to help see the process of statistics.

Part 1 is focused on the first step in the process, which is to determine the research objective or question to be answered. Then information is obtained to answer the questions stated in the research objective.

CHAPTER 1 Data Collection

Data Collection

Outline

1.1 Introduction to the Practice of Statistics

1.2 Observational Studies versus Designed Experiments

1.3 Simple Random Sampling

1.4 Other Effective Sampling Methods

1.5 Bias in Sampling

1.6 The Design of Experiments

Making an Informed Decision

It is your senior year of high school. You will have a lot of exciting experiences in the upcoming year, plus a major decision to make—which college should I attend? The choice you make may affect many aspects of your life—your career, where you live, your significant other, and so on, so you don't want to simply choose the college that everyone else picks. You need to design a questionnaire to help you make an informed decision about college. In addition, you want to know how well the college you are considering educates its students. See Making an Informed Decision on page 60.

Putting It Together

Statistics plays a major role in many aspects of our lives. It is used in sports, for example, to help a general manager decide which player might be the best fit for a team. It is used in politics to help candidates understand how the public feels about various policies. And statistics is used in medicine to help determine the effectiveness of new drugs.

Used appropriately, statistics can enhance our understanding of the world around us. Used inappropriately, it can lend support to inaccurate beliefs. Understanding statistical methods will provide you with the ability to analyze and critique studies and the opportunity to become an informed consumer of information. Understanding statistical methods will also enable you to distinguish solid analysis from bogus "facts."

To help you understand the features of this text and for hints to help you study, read the ***Pathway to Success*** on the front inside cover of the text.

1.1 Introduction to the Practice of Statistics

Objectives

1. Define statistics and statistical thinking
2. Explain the process of statistics
3. Distinguish between qualitative and quantitative variables
4. Distinguish between discrete and continuous variables
5. Determine the level of measurement of a variable

1 Define Statistics and Statistical Thinking

What is statistics? Many people say that statistics is numbers. After all, we are bombarded by numbers that supposedly represent how we feel and who we are. For example, we hear on the radio that 50% of first marriages, 67% of second marriages, and 74% of third marriages end in divorce (Forest Institute of Professional Psychology, Springfield, MO).

Another interesting consideration about the "facts" we hear or read is that two different sources can report two different results. For example, a January 12, 2019 poll by Rasmussen Reports indicated that 45% of Americans believed the country was on the right track. However, a January 16, 2019 Monmouth poll indicated that 37% of Americans believed the country was on the right track. Is it possible that the percent of Americans who believe the country is on the right track could decrease by 8% in four days, or is something else going on? Statistics helps to provide the answer.

Certainly, statistics has a lot to do with numbers, but this definition is only partially correct. Statistics is also about where the numbers come from (that is, how they were obtained) and how closely the numbers reflect reality.

Definition

Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

Let's break this definition into four parts. The first part states that statistics involves the collection of information. The second refers to the organization and summarization of information. The third states that the information is analyzed to draw conclusions or answer specific questions. The fourth part states that results should be reported using some measure that represents how convinced we are that our conclusions reflect reality.

What is the information referred to in the definition? The information is **data**, which the *American Heritage Dictionary* defines as "a fact or proposition used to draw a conclusion or make a decision." Data can be numerical, as in height, or nonnumerical, as in gender. In either case, data describe characteristics of an individual.

IN OTHER WORDS

Anecdotal means that the information being conveyed is based on casual observation, not scientific research.

Analysis of data can lead to powerful results. Data can be used to offset anecdotal claims, such as the suggestion that cellular telephones cause brain cancer. After carefully collecting, summarizing, and analyzing data regarding this phenomenon, it was determined that there is no link between cell phone usage and brain cancer. See Examples 1 and 2 in Section 1.2.

Because data are powerful, they can be dangerous when misused. The misuse of data usually occurs when data are incorrectly obtained or analyzed. For example, radio or television talk shows regularly ask poll questions for which respondents must call in or use the Internet to supply their vote. Most likely, the individuals who are going to call in are those who have a strong opinion about the topic. This group is not likely to be representative of people in general, so the results of the poll are not meaningful. Whenever we look at data, we should be mindful of where the data come from.

Even when data tell us that a relation exists, we need to investigate. For example, a study showed that breast-fed children have higher IQs than those who were not breast-fed. Does this study mean that a mother who breast-feeds her child will increase the child's IQ? Not necessarily. It may be that some factor other than breast-feeding contributes to the IQ of the children. In this case, it turns out that mothers who breast-feed generally have higher IQs than those who do not. Therefore, it may be genetics that leads to the higher IQ, not breast-feeding.* This illustrates an idea in statistics known as the *lurking variable*. A good statistical study will have a way of dealing with lurking variables.

A key aspect of data is that they vary. Consider the students in your classroom. Is everyone the same height? No. Does everyone have the same color hair? No. So, within groups there is variation. Now consider yourself. Do you eat the same amount of food each day? No. Do you sleep the same number of hours each day? No. So even considering an individual there is variation. Data vary. One goal of statistics is to describe and understand the sources of variation. Variability in data may help to explain the different results obtained by the Rasmussen Reports and Monmouth polls described at the beginning of this section.

Because of this variability, the results that we obtain using data can vary. In a mathematics class, if Bob and Jane are asked to solve $3x + 5 = 11$, they will both obtain $x = 2$ as the solution when they use the correct procedures. In a statistics class, if Bob and Jane are asked to estimate the average commute time for workers in Dallas, Texas, they will likely get different answers, even though both use the correct procedure. The different answers occur because they likely surveyed different individuals, and these individuals have different commute times. Bob and Jane would get the same result if they both asked *all* commuters or the same commuters about their commutes, but how likely is this?

So, in mathematics when a problem is solved correctly, the results can be reported with 100% certainty. In statistics, when a problem is solved, the results do not have 100% certainty. In statistics, we might say that we are 95% confident that the average commute time in Dallas, Texas, is between 20 and 23 minutes. Uncertain results may seem disturbing now but will feel more comfortable as we proceed through the course.

Without certainty, how can statistics be useful? Statistics can provide an understanding of the world around us because recognizing where variability in data comes from can help us to control it. Understanding the techniques presented in this text will provide you with powerful tools that will give you the ability to analyze and critique media reports, make investment decisions, or conduct research on major purchases. This will help to make you an informed citizen, consumer of information, and critical and statistical thinker.

2 Explain the Process of Statistics

Consider the following scenario.

> You are walking down the street and notice that a person walking in front of you drops $100. Nobody seems to notice the $100 except you. Since you could keep the money without anyone knowing, would you keep the money or return it to the owner?

NOTE

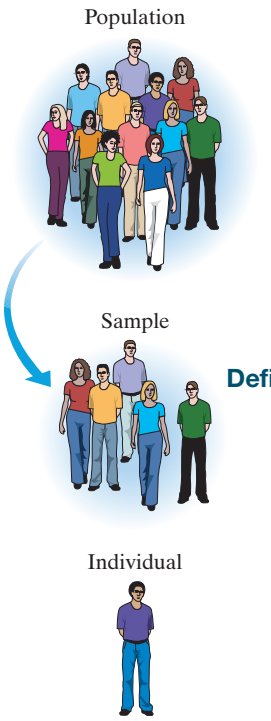
Obtaining a truthful response to a question such as this is challenging. In Section 1.5, we present some techniques for obtaining truthful responses to sensitive questions.

Suppose you wanted to use this scenario as a gauge of the morality of students at your school by determining the percent of students who would return the money. How might you do this? You could attempt to present the scenario to every student at the school, but this would be difficult or impossible if the student body is large. A second possibility is to present the scenario to 50 students and use the results to make a statement about all the students at the school.

*In fact, a study found that a gene called FADS2 is responsible for higher IQ scores in breast-fed babies. *Source:* Duke University, "Breastfeeding Boosts IQ in Infants with 'Helpful' Genetic Variant," *Science Daily* 6 November 2007.

Definitions The entire group to be studied is called the **population**. An **individual** is a person or object that is a member of the population being studied. A **sample** is a subset of the population that is being studied. See Figure 1.

Figure 1



In the $100 study presented, the population is all the students at the school. Each student is an individual. The sample is the 50 students selected to participate in the study.

Suppose 39 of the 50 students stated that they would return the money to the owner. We could present this result by saying that the percent of students in the survey who would return the money to the owner is 78%. This is an example of a *descriptive statistic* because it describes the results of the sample without making any general conclusions about the population.

Definitions A **statistic** is a numerical summary of a sample. **Descriptive statistics** consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs.

So 78% is a statistic because it is a numerical summary based on a sample. Descriptive statistics make it easier to get an overview of what the data are telling us.

If we extend the results of our sample to the population, we are performing *inferential statistics*.

Definition **Inferential statistics** uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.

The generalization contains uncertainty because a sample cannot tell us everything about a population. Therefore, inferential statistics includes a level of confidence in the results. So rather than saying that 78% of all students would return the money, we might say that we are 95% confident that between 74% and 82% of all students would return the money. Notice how this inferential statement includes a *level of confidence* (measure of reliability) in our results. It also includes a range of values to account for the variability in our results.

One goal of inferential statistics is to use statistics to estimate *parameters*.

Definition A **parameter** is a numerical summary of a population.

EXAMPLE 1 Parameter versus Statistic

(a) Suppose 48.2% of all students on your campus own a car. This value represents a parameter because it is a numerical summary of a population. Suppose a sample of 100 students is obtained, and from this sample we find that 46% own a car. This value represents a statistic because it is a numerical summary of a sample.

(b) Suppose the average salary of all employees in the City of Joliet is $78,302. This value represents a parameter because it is a numerical summary of a population. Suppose a sample of 30 employees is obtained, and from this sample we find the average salary is $75,038. This value represents a statistic because it is a numerical summary of a sample.

NW **Now Work Problem 3**

The methods of statistics follow a process.

CAUTION!
Many nonscientific studies are based on *convenience samples*, such as Internet surveys or phone-in polls. The results of any study performed using this type of sampling method are not reliable.

The Process of Statistics

1. *Identify the research objective.* A researcher must determine the question(s) he or she wants answered. The question(s) must clearly identify the population that is to be studied.
2. *Collect the data needed to answer the question(s) posed in (1).* Conducting research on an entire population is often difficult and expensive, so we typically look at a sample. This step is vital to the statistical process, because if the data are not collected correctly, the conclusions drawn are meaningless. Do not overlook the importance of appropriate data collection. We discuss this step in detail in Sections 1.2 through 1.6.
3. *Describe the data.* Descriptive statistics allow the researcher to obtain an overview of the data and can help determine the type of statistical methods the researcher should use. We discuss this step in detail in Chapters 2 through 4.
4. *Perform inference.* Apply the appropriate techniques to extend the results obtained from the sample to the population and report a level of reliability of the results. We discuss techniques for measuring reliability in Chapters 5 through 8 and inferential techniques in Chapters 9 through 15.

EXAMPLE 2 The Process of Statistics: Trust Your Neighbor

Pew Research conducted a poll and asked, "Do you trust all or most of your neighbors?" The following process allowed the researchers to conduct their study.

1. *Identify the Research Objective* The researchers wanted to determine the percentage of adult Americans who trust all or most of their neighbors. Therefore, the population was adult Americans.
2. *Collect the Data Needed to Answer the Question Posed in (1)* It is unreasonable to expect to survey the more than 200 million adult Americans to determine whether they trust all or most of their neighbors. So, the researchers surveyed a sample of 1628 adult Americans. Of those surveyed, 847 stated they trust all or most of their neighbors.
3. *Describe the Data* Of the 1628 individuals in the survey, 52% $(= 847/1628)$ stated they trust all or most of their neighbors. This is a descriptive statistic because it is a summary of the sample data.
4. *Perform Inference* Pew Research wanted to extend the results of the survey to all adult Americans. When generalizing results from a sample to a population, the results are uncertain. To account for this uncertainty, Pew reported a 2.5% *margin of error.* This means Pew feels fairly certain (in fact, 95% certain) that the percentage of *all* adult Americans who trust all or most of their neighbors is somewhere between 49.5% $(= 52\% - 2.5\%)$ and 54.5% $(= 52\% + 2.5\%)$.

NW **Now Work Problem 45**

3 Distinguish between Qualitative and Quantitative Variables

Once a research objective is stated, a list of the information we want to learn about the individuals must be created. **Variables** are the characteristics of the individuals within the population. For example, recently my son and I planted a tomato plant in our backyard. We collected information about the tomatoes harvested from the plant. The individuals we studied were the tomatoes. The variable that interested us was the weight of a tomato. My son noted that the tomatoes had different weights even though they came from the same plant. He discovered that variables such as weight may vary.

If variables did not vary, they would be constants, and statistical inference would not be necessary. Think about it this way: If each tomato had the same weight, then knowing the weight of one tomato would allow us to determine the weights of all tomatoes. However, the weights of the tomatoes vary. One goal of research is to learn the causes of the variability so that we can learn to grow plants that yield the best tomatoes.

Variables can be classified into two groups: *qualitative* or *quantitative*.

Definitions

Qualitative, or **categorical, variables** allow for classification of individuals based on some attribute or characteristic.

Quantitative variables provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.

Many examples in this text will include a suggested **approach**, or a way to look at and organize a problem so that it can be solved. The approach will be a suggested method of *attack* toward solving the problem. This does not mean that the approach given is the only way to solve the problem, because many problems have more than one approach leading to a correct solution.

EXAMPLE 3 Distinguishing between Qualitative and Quantitative Variables

Problem Determine whether the following variables are qualitative or quantitative.

(a) Gender
(b) Temperature
(c) Number of days during the past week that a college student studied
(d) Zip code

Approach Quantitative variables are numeric measures such that meaningful arithmetic operations can be performed on the values of the variable. Qualitative variables describe an attribute or characteristic of the individual that allows researchers to categorize the individual.

Solution

(a) Gender is a qualitative variable because it allows a researcher to categorize the individual as male or female. Notice that arithmetic operations cannot be performed on these attributes.
(b) Temperature is a quantitative variable because it is numeric, and operations such as addition and subtraction provide meaningful results. For example, 70°F is 10°F warmer than 60°F.
(c) Number of days during the past week that a college student studied is a quantitative variable because it is numeric, and operations such as addition and subtraction provide meaningful results.
(d) Zip code is a qualitative variable because it categorizes a location. Notice that, even though zip codes are numeric, adding or subtracting zip codes does not provide meaningful results.

NW **Now Work Problem 11**

Example 3(d) shows us that a variable may be qualitative while having numeric values. Just because the value of a variable is numeric does not mean that the variable is quantitative.

4 Distinguish between Discrete and Continuous Variables

We can further classify quantitative variables into two types: *discrete* or *continuous*.

Definitions

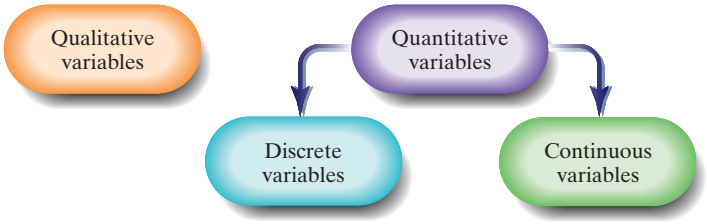
A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of possible values. The term *countable* means that the values result from counting, such as 0, 1, 2, 3, and so on. A discrete variable cannot take on every possible value between any two possible values.

A **continuous variable** is a quantitative variable that has an infinite number of possible values that are not countable. A continuous variable may take on every possible value between any two values.

IN OTHER WORDS
If you count to get the value of a quantitative variable, it is discrete. If you measure to get the value of a quantitative variable, it is continuous.

Figure 2 illustrates the relationship among qualitative, quantitative, discrete, and continuous variables.

Figure 2



EXAMPLE 4 Distinguishing between Discrete and Continuous Variables

Problem Determine whether the quantitative variables are discrete or continuous.

(a) The number of heads obtained after flipping a coin five times.

(b) The number of cars that arrive at a McDonald's drive-thru between 12:00 P.M. and 1:00 P.M.

(c) The distance a 2020 Toyota Prius can travel in city driving conditions with a full tank of gas.

Approach A variable is discrete if its value results from counting. A variable is continuous if its value is measured.

Solution

(a) The number of heads obtained by flipping a coin five times is a discrete variable because we can count the number of heads obtained. The possible values of this discrete variable are 0, 1, 2, 3, 4, 5.

(b) The number of cars that arrive at a McDonald's drive-thru between 12:00 P.M. and 1:00 P.M. is a discrete variable because we find its value by counting the cars. The possible values of this discrete variable are 0, 1, 2, 3, 4, and so on. Notice that this number has no upper limit.

(c) The distance traveled is a continuous variable because we measure the distance (miles, feet, inches, and so on).

NW **Now Work Problem 19**

Continuous variables are often rounded. For example, if a certain make of car gets 24 miles per gallon (mpg) of gasoline, its miles per gallon must be greater than or equal to 23.5 and less than 24.5, or $23.5 \leq \text{mpg} < 24.5$.

The type of variable (qualitative, discrete, or continuous) dictates the methods that can be used to analyze the data.

The list of observed values for a variable is **data**. Gender is a variable; the observations male and female are data. **Qualitative data** are observations corresponding to a qualitative variable. **Quantitative data** are observations corresponding to a quantitative variable. **Discrete data** are observations corresponding to a discrete variable. **Continuous data** are observations corresponding to a continuous variable.

EXAMPLE 5 Distinguishing between Variables and Data

Problem Table 1 presents a group of selected countries and information regarding these countries as of July, 2020. Identify the individuals, variables, and data in Table 1.

Table 1

Country	Government Type	Life Expectancy (years)	Population (in millions)
Australia	Federal parliamentary democracy	82.7	25.5
Canada	Constitutional monarchy	83.4	37.7
France	Republic	82.2	67.8
Morocco	Constitutional monarchy	73.3	35.6
Poland	Republic	78.3	38.3
Senegal	Presidential republic	63.2	15.7
United States	Federal republic	80.3	332.6

Source: CIA *World Factbook*

Approach An individual is an object or person for whom we wish to obtain data. The variables are the characteristics of the individuals, and the data are the specific values of the variables.

Solution The individuals in the study are the countries: Australia, Canada, and so on. The variables measured for each country are *government type*, *life expectancy*, and *population*. The variable *government type* is qualitative because it categorizes the individual. The variables *life expectancy* and *population* are quantitative.

The quantitative variable *life expectancy* is continuous because it is measured. The quantitative variable *population* is discrete because we count people. The observations are the data. For example, the data corresponding to the variable *life expectancy* are 82.7, 83.4, 82.2, 73.3, 78.3, 63.2, and 80.3. The following data correspond to the individual Poland: a republic government with residents whose life expectancy is 78.3 years and population is 38.3 million people. Republic is an instance of qualitative data that results from observing the value of the qualitative variable *government type*. The life expectancy of 78.3 years is an instance of quantitative data that results from observing the value of the quantitative variable *life expectancy*.

NW **Now Work Problem 41**

5 Determine the Level of Measurement of a Variable

Rather than classify a variable as qualitative or quantitative, we can assign a level of measurement to the variable.

Definitions

A variable is at the **nominal level of measurement** if the values of the variable name, label, or categorize. In addition, the naming scheme does not allow for the values of the variable to be arranged in a ranked or specific order.

A variable is at the **ordinal level of measurement** if it has the properties of the nominal level of measurement, however, the naming scheme allows for the values of the variable to be arranged in a ranked or specific order.

IN OTHER WORDS
The word ***nominal*** comes from the Latin word **nomen**, which means to name. When you see the word **ordinal**, think order.

(continued)

A variable is at the **interval level of measurement** if it has the properties of the ordinal level of measurement and the differences in the values of the variable have meaning. A value of zero does not mean the absence of the quantity. Arithmetic operations such as addition and subtraction can be performed on values of the variable.

A variable is at the **ratio level of measurement** if it has the properties of the interval level of measurement and the ratios of the values of the variable have meaning. A value of zero means the absence of the quantity. Arithmetic operations such as multiplication and division can be performed on the values of the variable.

Nominal or ordinal variables are also qualitative variables. Interval or ratio variables are also quantitative variables.

EXAMPLE 6 Determining the Level of Measurement of a Variable

Problem For each of the following variables, determine the level of measurement.
(a) Gender
(b) Temperature
(c) Number of days during the past week that a college student studied
(d) Letter grade earned in your statistics class

Approach For each variable, we ask the following: Does the variable simply categorize each individual? If so, the variable is nominal. Does the variable categorize *and* allow ranking of each value of the variable? If so, the variable is ordinal. Do differences in values of the variable have meaning, but a value of zero does not mean the absence of the quantity? If so, the variable is interval. Do ratios of values of the variable have meaning *and* there is a natural zero starting point? If so, the variable is ratio.

Solution
(a) Gender is a variable measured at the nominal level because it only allows for categorization of male or female. Plus, it is not possible to rank gender classifications.
(b) Temperature is a variable measured at the interval level because differences in the value of the variable make sense. For example, 70°F is 10°F warmer than 60°F. Notice that the ratio of temperatures does not represent a meaningful result. For example, 60°F is not twice as warm as 30°F. In addition, 0°F does not represent the absence of heat.
(c) Number of days during the past week that a college student studied is measured at the ratio level, because the ratio of two values makes sense and a value of zero has meaning. For example, a student who studies four days studies twice as many days as a student who studies two days.
(d) Letter grade is a variable measured at the ordinal level because the values of the variable can be ranked, but differences in values have no meaning. For example, an A is better than a B, but A − B has no meaning.

NW **Now Work Problem 27**

When classifying variables according to their level of measurement, it is extremely important that we recognize what the variable is intended to measure. For example, suppose we want to know whether cars with 4-cylinder engines get better gas mileage than cars with 6-cylinder engines. Here, engine size represents a category of data and so the variable is nominal. On the other hand, if we want to know the average number of cylinders in cars in the United States, the variable is classified as ratio (an 8-cylinder engine has twice as many cylinders as a 4-cylinder engine).

1.1 Assess Your Understanding

Vocabulary and Skill Building

1. Match each word or phrase with its definition.

Word/Phrase	Definition
(a) Statistics	I. A numerical summary of a sample.
(b) Population	II. Organizing and summarizing data through tables, graphs, and numerical summaries.
(c) Sample	III. The science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. It is also about providing a measure of confidence in any conclusions.
(d) Parameter	IV. A subset of the group of individuals that is being studied.
(e) Statistic	V. Uses methods that take results from a sample and extends them to the population, and measures the reliability of the result.
(f) Individual	VI. A person or object that is a member of the group being studied.
(g) Descriptive Statistics	VII. A numerical summary of a population.
(h) Inferential Statistics	VIII. The entire group of individuals to be studied.

2. Match each word or phrase with its definition.

Word/Phrase	Definition
(a) Discrete Variable	I. Provide numerical measures of individuals. The measures can be added or subtracted, and provide meaningful results.
(b) Data	II. Allow for classification of individuals based on some attribute or characteristic.
(c) Continuous Variable	III. The characteristics of the individuals within the population.
(d) Qualitative Variable	IV. Information that describes characteristics of an individual.
(e) Quantitative Variable	V. Has either a finite number of possible values or countable number of possible values. The values of these variables typically result from counting.
(f) Variable	VI. Has an infinite number of possible values that are not countable. The values of these variables typically result from measurement.

In Problems 3–10, determine whether the underlined value is a parameter or a statistic.

NW **3. State Government** Following the 2018 national midterm election, 18% of the governors of the 50 United States were female. *Source:* National Governors Association

4. Calculus Exam The average score for a class of 28 students taking a calculus midterm exam was 72%.

5. School Bullies In a national survey of 1300 high school students (grades 9 to 12), 32% of respondents reported that someone had bullied them at school. *Source:* Bureau of Justice Statistics

6. Drug Use In a national survey on substance abuse, 13.3% of 12th graders reported using illicit drugs within the past month. *Source:* National Institute on Drug Abuse

7. Batting Average Ty Cobb is one of Major League Baseball's greatest hitters of all time, with a career batting average of 0.366. *Source:* baseball-almanac.com

8. Moonwalkers Only 12 men have walked on the moon. The average time these men spent on the moon was 43.92 hours. *Source:* www.theguardian.com

9. Hygiene Habits A study of 6076 adults in public rest rooms (in Atlanta, Chicago, New York City, and San Francisco) found that 23% did not wash their hands before exiting. *Source:* American Society for Microbiology and the Soap and Detergent Association.

10. Public Knowledge Interviews of 100 adults 18 years of age or older, conducted nationwide, found that 44% could state the minimum age required for the office of U.S. president. *Source: Newsweek Magazine.*

In Problems 11–18, classify the variable as qualitative or quantitative.

NW **11.** Nation of origin

12. Number of siblings

13. Grams of carbohydrates in a doughnut

14. Number on a football player's jersey

15. Number of unpopped kernels in a bag of microwave popcorn

16. Assessed value of a house

17. Phone number

18. Student ID number

In Problems 19–26, determine whether the quantitative variable is discrete or continuous.

NW **19.** Goals scored in a season by a soccer player

20. Volume of water lost each day through a leaky faucet

21. Length (in minutes) of a country song

22. Number of Sequoia trees in a randomly selected acre of Yosemite National Park

23. High temperature on a randomly selected day in Memphis, Tennessee

24. Internet connection speed in kilobytes per second

25. Points scored in an NCAA basketball game

26. Air pressure in pounds per square inch in an automobile tire

In Problems 27–34, determine the level of measurement of each variable.

NW **27.** Nation of origin

28. Movie ratings of one star through five stars

29. Volume of water used by a household in a day

30. Year of birth of college students

31. Highest degree conferred (high school, bachelor's, and so on)

32. Eye color

33. Assessed value of a house

34. Time of day measured in military time

In Problems 35–40, a research objective is presented. For each, identify the population and sample in the study.

35. The Gallup Organization contacts 1028 teenagers who are 13 to 17 years of age and live in the United States and asks whether or not they had been prescribed medications for any mental disorders, such as depression or anxiety.

36. A quality-control manager randomly selects 50 bottles of Coca-Cola that were filled on October 15 to assess the calibration of the filling machine.

37. A farmer interested in the weight of his soybean crop randomly samples 100 plants and weighs the soybeans on each plant.

38. Every year the U.S. Census Bureau releases the *Current Population Report* based on a survey of 50,000 households. The goal of this report is to learn the demographic characteristics, such as income, of all households within the United States.

39. Folate and Hypertension Researchers want to determine whether or not higher folate intake is associated with a lower risk of hypertension (high blood pressure) in women (27 to 44 years of age). To make this determination, they look at 7373 cases of hypertension in these women and find that those who consume at least 1000 micrograms per day (μg/d) of total folate had a decreased risk of hypertension compared with those who consume less than 200 μg/d. *Source:* John P. Forman, MD; Eric B. Rimm, ScD; Meir J. Stampfer, MD; Gary C. Curhan, MD, ScD, "Folate Intake and the Risk of Incident Hypertension among US Women," *Journal of the American Medical Association* 293:320–329, 2005.

40. A community college notices that an increasing number of full-time students are working while attending the school. The administration randomly selects 128 students and asks how many hours per week each works.

In Problems 41 and 42, identify the individuals, variables, and data corresponding to the variables. Determine whether each variable is qualitative, continuous, or discrete.

NW **41. Driver's License Laws** The following data represent driver's license laws for various states.

State	Minimum Age for Driver's License (unrestricted)	Mandatory Belt Use Seating Positions	Maximum Allowable Speed Limit (cars on rural interstate), mph
Alabama	17	Front	70
Colorado	17	Front	75
Indiana	18	All	70
North Carolina	16	All	70
Wisconsin	18	All	65

Source: Governors Highway Safety Association.

42. BMW Cars The following information relates to the 2019 model year product line of BMW automobiles.

Model	Body Style	Weight (lb)	Number of Seats
3 Series	Sedan	3489	4
4 Series	Coupe	3574	4
5 Series	Sedan	3790	5
7 Series	Sedan	4244	5
X3	Sport utility	4034	5
Z4	Roadster Coupe	3287	2

Source: www.motortrend.com

Applying the Concepts

43. Smoker's IQ A study was conducted in which 20,211 18-year-old Israeli male military recruits were given an exam to measure IQ. In addition, the recruits were asked to disclose their smoking status. An individual was considered a smoker if he smoked at least one cigarette per day. The goal of the study was to determine whether adolescents aged 18 to 21 who smoke have a lower IQ than nonsmokers. It was found that the average IQ of the smokers was 94, while the average IQ of the nonsmokers was 101. The researchers concluded that lower IQ individuals are more likely to choose to smoke, not that smoking makes people less intelligent.

Source: Weiser, M., Zarka, S., Werbeloff, N., Kravitz, E. and Lubin, G. (2010). "Cognitive Test Scores in Male Adolescent Cigarette Smokers Compared to Non-smokers: A Population-Based Study." *Addiction*. 105:358–363. doi: 10.1111/j.1360-0443.2009.02740.x).

(a) What is the research objective?
(b) What is the population being studied? What is the sample?
(c) What are the descriptive statistics?
(d) What are the conclusions of the study?

44. A Cure for the Common Wart A study was designed "to determine if application of duct tape is as effective as cryotherapy (liquid nitrogen applied to the wart for 10 seconds every 2 to 3 weeks) in the treatment of common warts." The researchers randomly divided 51 patients into two groups. The 26 patients in group 1 had their warts treated by applying duct tape to the wart for 6.5 days and then removing the tape for 12 hours, at which point the cycle was repeated for a maximum of 2 months. The 25 patients in group 2 had their warts treated by cryotherapy for a maximum of six treatments. Once the treatments were complete, it was determined that 85% of the patients in group 1 and 60% of the patients in group 2 had complete resolution of their warts. The researchers concluded that duct tape is significantly more effective in treating warts than cryotherapy.

Source: Dean R. Focht III, Carole Spicer, Mary P. Fairchok. "The Efficacy of Duct Tape vs. Cryotherapy in the Treatment of Verruca Vulgaris (The Common Wart)," *Archives of Pediatrics and Adolescent Medicine*, 156(10), 2002.

(a) What is the research objective?
(b) What is the population being studied? What is the sample?
(c) What are the descriptive statistics?
(d) What are the conclusions of the study?

NW **45. Government Waste** Gallup News Service conducted a survey of 1017 American adults aged 18 years or older. The respondents were asked, "Of every tax dollar that goes to the federal government in Washington, D.C., do you believe 51 cents or more are wasted?" Of the 1017 individuals surveyed, 35% indicated that 51 cents or more is wasted. Gallup reported that 35% of all adult Americans 18 years or older believe the federal government wastes at least 51 cents of each dollar spent, with a margin of error of 4% and a 95% level of confidence.

(a) What is the research objective?
(b) What is the population?
(c) What is the sample?
(d) List the descriptive statistics.
(e) What can be inferred from this survey?

46. Investment Decision The Gallup Organization conducted a survey of 1018 adults, aged 18 and older, living in the United States and asked, "If you had a thousand dollars to spend, do you think investing it in the stock market would be a good or bad idea?" Of the 1018 adults, 46% said it would be a bad idea. The Gallup Organization reported that 46% of all adults, aged 18 and older, living in the United States thought it was a bad idea to invest $1000 in the stock market with a 4% margin of error with 95% confidence.

(a) What is the research objective?
(b) What is the population?
(c) What is the sample?
(d) List the descriptive statistics.
(e) What can be inferred from this survey?

47. Threaded Problem: Tornado The data set "Tornadoes_2017" located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017. For each of the following variables in the data set, indicate whether the variable is qualitative or quantitative. For those that are quantitative, indicate whether the variable is discrete or continuous.

(a) State
(b) F Scale (this is the Fujita scale for rating tornadoes based on wind speed, where 0 is a tornado whose wind speed is less than 73 mph; 1 is a tornado whose wind speed is 73–112 mph; up through 5, which is a tornado whose wind speed is 261–318 mph).
(c) Fatalities
(d) Length

48. Threaded Problem: Tornado The data set "Tornadoes_2017" located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017. For each of the following variables in the data set, indicate the level of measurement of each variable.

(a) State
(b) F Scale (this is the Fujita scale for rating tornadoes based on wind speed, where 0 is a tornado whose wind speed is less than 73 mph; 1 is a tornado whose wind speed is 73–112 mph; up through 5, which is a tornado whose wind speed is 261–318 mph).
(c) Fatalities
(d) Length

49. What Level of Measurement? It is extremely important for a researcher to clearly define the variables in a study because this helps to determine the type of analysis that can be performed on the data. For example, if a researcher wanted to describe baseball players based on jersey number, what level of measurement would the variable *jersey number* be? Now suppose the researcher felt that certain players who were of lower caliber received higher numbers. Does the level of measurement of the variable change? If so, how?

50. Interpreting the Variable Suppose a fundraiser holds a raffle for which each person who enters the room receives a ticket numbered 1 to N, where N is the number of people at the fundraiser. The first person to arrive receives ticket number 1, the second person receives ticket number 2, and so on. Determine the level of measurement for each of the following interpretations of the variable *ticket number*.

(a) The winning ticket number.
(b) The winning ticket number was announced as 329. An attendee noted his ticket number was 294 and stated, "I guess I arrived too early."
(c) The winning ticket number was announced as 329. An attendee looked around the room and commented, "It doesn't look like there are 329 people in attendance."

51. Analyze the Article Read the newspaper article and answer the following questions:

(a) What is the research question the study addresses?
(b) What is the sample?
(c) What type of variable is season in which you were born?
(d) What can be said (in general) about individuals born in summer? Winter?
(e) What conclusion was drawn from the study?

Season of Birth Affects Your Mood Later In Life by Nicola Fifield

Babies born in the summer are much more likely to suffer from mood swings when they grow up while those born in the winter are less likely to become irritable adults, scientists claim.

Researchers studied 400 people and matched their personality type to when in the year they were born.

They claim that people born at certain times of the year have a far greater chance of developing certain types of temperaments, which can lead to mood disorders.

The scientists, from Budapest, said this was because the seasons had an influence on certain monoamine neurotransmitters, such as dopamine and serotonin, which control mood, however more research was needed to find out why.

They discovered that the number of people with a "cyclothymic" temperament, characterized by rapid, frequent swings between sad and cheerful moods, was significantly higher in those born in the summer.

Those with a hyperthymic temperament, a tendency to be excessively positive, was significantly higher among those born in the spring and summer.

The study also found that those born in the autumn were less likely to be depressive, while those born in winter were less likely to be irritable.

Lead researcher, assistant professor Xenia Gonda, said: "Biochemical studies have shown that the season in which you are born has an influence on certain monoamine neurotransmitters, such as dopamine and serotonin, which is detectable even in adult life. This led us to believe that birth season may have a longer-lasting effect.

"Our work looked at 400 subjects and matched their birth season to personality types in later life.

"Basically, it seems that when you are born may increase or decrease your chance of developing certain mood disorders.

Professor Gonda added: "We can't yet say anything about the mechanisms involved.

What we are now looking at is to see if there are genetic markers which are related to season of birth and mood disorder".

The study may well provide a clue as to why some of the nation's best known personalities are good natured, while others are slightly grumpier.

The Duchess of Cambridge was born in winter, on January 9, which according to the study, means she is less likely to be irritable while Roy Keane, the famously hot-headed former Manchester United footballer, was born in August, when the scientists say people are more likely to have mood swings.

Mary Berry, the ever-cheerful presenter of the Great British Bake Off, was born in the Spring, when, according to the study, people are more likely to be excessively positive.

The study is being presented at the annual conference of the European College of Neuropsychopharmacology (ECNP) in Berlin, Germany, on Sunday.

Professor Eduard Vieta, from the ECNP, said: "Although both genetic and environmental factors are involved in one's temperament, now we know that the season at birth plays a role too.

"And the finding of "high mood" tendency (hyperthymic temperament) for those born in summer is quite intriguing."
The Telegraph, October 19, 2014

Source: Season of Birth Affects Your Mood Later In Life by Nicola Fifield from The Telegraph. Copyright © 2014 by Telegraph Media Group Limited.

Explaining the Concepts

52. Explain the difference between a population and a sample.

53. Contrast the differences between qualitative and quantitative variables. Discuss the differences between discrete and continuous variables.

54. In your own words, define the four levels of measurement of a variable. Give an example of each.

55. Explain what is meant when we say "data vary." How does this variability affect the results of statistical analysis?

56. Explain the process of statistics.

57. The age of a person is commonly considered to be a continuous random variable. Could it be considered a discrete random variable instead? Explain.

1.2 Observational Studies versus Designed Experiments

Objectives
1. Distinguish between an observational study and an experiment
2. Explain the various types of observational studies

1 Distinguish between an Observational Study and an Experiment

Once a research objective is determined, the researcher develops the method for obtaining the data that can be used to answer the questions posed in our research objective. There are two methods for collecting data: *observational studies* and *designed experiments*. To see the difference between these two methods, read the following two studies.

EXAMPLE 1 Cellular Phones and Brain Tumors

Researchers wanted to determine whether there is an association between mobile phone use and brain tumors. To do so, 791,710 middle-aged women in the United Kingdom were followed over a period of 7 years. During this time, there were 1261 incidences of brain tumors. The researchers compared the women who never used a mobile phone to those who used mobile phones and found no significant difference in the incidence rate of brain tumors between the two groups.

Source: Benson, V. S. et al. "Mobile Phone Use and Risk of Brain Neoplasms and Other Cancers: Prospective Study," *International Journal of Epidemiology* 2013 Jun; 42(3); 792–802.

EXAMPLE 2 Cellular Phones and Brain Tumors

Researchers from the United States National Toxicology Program conducted a study to address the concern that radio-frequency radiation (RFR) may be associated with an increased likelihood of developing brain tumors in humans. Certainly, it is unethical

to purposely expose humans to a potential carcinogen, so rats were used instead. The researchers randomly assigned 90 rats to one of three possible groups. Each of the groups was housed in a reverberation chamber that allowed the rats to be exposed to the RFR. Group 1 rats served as a control and were not exposed to any RFR in their chamber. Group 2 rats were exposed to Global System for Mobile Communications (GSM)-modulated RFR, and Group 3 rats were exposed to Code Division Multiple Access (CDMA)-modulated RFR. GSM and CDMA are the modulations primarily used in the United States. The rats in Groups 2 and 3 were exposed to RFR using a continuous cycle of 10 minutes on (exposed) and 10 minutes off (not exposed) for a total daily exposure time of about 9 hours a day, 7 days per week for approximately two years. Each chamber was maintained on a 12-hour light/dark cycle with a temperature range of 72 degrees Fahrenheit (plus or minus 3 degrees), a humidity range of 50 plus/minus 15%, and with at least 10 air changes per hour. All the rats had the same access to food and water. The researchers found low incidences of brain tumors in rats exposed to RFR for both GSM and CDMA modulations, while there were no cases of brain tumors in the control group. However, the incidence rate was not statistically significant.

Source: M. Wyde, et al. bioRxiv 055699; doi: https://doi.org/10.1101/055699, June 23, 2016. "Report of Partial Findings from the National Toxicology Program Carcinogenesis Studies of Cell Phone Radiofrequency Radiation in Hsd: Sprague Dawley® SD rats (Whole Body Exposures)."

In both studies, the goal was to determine if radio frequencies from cell phones increase the risk of contracting brain tumors. Whether or not brain cancer was contracted is the *response variable*. The level of cell phone usage is the *explanatory variable*. In research, we wish to determine how varying the amount of an **explanatory variable** affects the value of a **response variable**.

What are the differences between the studies in Examples 1 and 2? Obviously, in Example 1 the study was conducted on humans, while the study in Example 2 was conducted on rats. However, there is a bigger difference. In Example 1, no attempt was made to influence the individuals in the study. The researchers simply followed the women over time to determine their use of cell phones. In other words, no attempt was made to influence the value of the explanatory variable, radio-frequency exposure (cell phone use). Because the researchers simply recorded the behavior of the participants, the study in Example 1 is an *observational study*.

Definition

An **observational study** measures the value of the response variable without attempting to influence the value of either the response or explanatory variables. That is, in an observational study, the researcher observes the behavior of the individuals without trying to influence the outcome of the study.

In the study in Example 2, the researchers obtained 90 rats and divided the rats into three groups. Each group was *intentionally* exposed to various levels of radiation. The researchers then compared the number of rats in each group that had brain tumors. Clearly, there was an attempt to influence the individuals in this study because the value of the explanatory variable (exposure to radio frequency) was manipulated to three levels. Because the researchers manipulated the value of an explanatory variable (radiation) and controlled others (temperature, food), we call the study in Example 2 a *designed experiment*.

Definition

If a researcher randomly assigns the individuals in a study to groups, intentionally manipulates the value of an explanatory variable, controls other explanatory variables at fixed values, and then records the value of the response variable for each individual, the study is a **designed experiment**.

NW **Now Work Problem 9**

Which Is Better? A Designed Experiment or an Observational Study?
To answer this question, let's consider another study.

EXAMPLE 3 Do Flu Shots Benefit Seniors?

Researchers wanted to determine the long-term benefits of the influenza vaccine on seniors aged 65 years and older by looking at records of over 36,000 seniors for 10 years. The seniors were divided into two groups. Group 1 were seniors who chose to get a flu vaccination shot, and group 2 were seniors who chose not to get a flu vaccination shot. After observing the seniors for 10 years, it was determined that seniors who get flu shots are 27% less likely to be hospitalized for pneumonia or influenza and 48% less likely to die from pneumonia or influenza.

Source: Kristin L. Nichol, MD, MPH, MBA, James D. Nordin, MD, MPH, David B. Nelson, PhD, John P. Mullooly, PhD, Eelko Hak, PhD. "Effectiveness of Influenza Vaccine in the Community-Dwelling Elderly," *New England Journal of Medicine* 357:1373–1381, 2007.

Wow! The results of this study sound great! All seniors should go out and get a flu shot. Right? Not necessarily. The authors were concerned about *confounding*. They were concerned that lower hospitalization and death rates may have been due to something other than the flu shot. Could it be that seniors who get flu shots are more health conscious or are able to get to the clinic more easily? Does race, income, or gender play a role in whether one might contract (and possibly die from) influenza?

Definition

Confounding in a study occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable or variables not accounted for in the study.

Confounding is potentially a major problem with observational studies. Often, the cause of confounding is a *lurking variable*.

Definition

A **lurking variable** is an explanatory variable that was not considered in a study, but that affects the value of the response variable in the study. In addition, lurking variables are typically related to explanatory variables considered in the study.

In the influenza study, possible lurking variables might be age, health status, or mobility of the senior. How can we manage the effect of lurking variables? One possibility is to look at the individuals in the study to determine if they differ in any significant way. For example, it turns out in the influenza study that the seniors who elected to get a flu shot were actually *less* healthy than those who did not. The researchers also accounted for race and income. The authors identified another potential lurking variable, *functional status*, meaning the ability of the seniors to conduct day-to-day activities on their own. The authors were able to adjust their results for this variable as well.

Even after accounting for all the potential lurking variables in the study, the authors were still careful to conclude that getting an influenza shot is *associated* with a lower risk of being hospitalized or dying from influenza. The authors used the term *associated*, instead of saying the influenza shots *caused* a lower risk of death, because the study was observational.

Observational studies do not allow a researcher to claim causation, only association.

Designed experiments, on the other hand, are used whenever control of certain variables is possible and desirable. This type of research allows the researcher to identify certain cause and effect relationships among the variables in the study.

So why ever conduct an observational study if we can't claim causation? Often, it is unethical to conduct an experiment. Consider the link between smoking and lung cancer. In a designed experiment to determine if smoking causes lung cancer in humans, a researcher would divide a group of volunteers into groups. Group 1 individuals would smoke a pack of cigarettes every day for the next 10 years, and Group 2 individuals would not smoke. In addition, eating habits, sleeping habits, and exercise would be controlled so that the only difference between the two groups was smoking. After 10 years the experiment's researcher would compare the proportion of participants in the study who contract lung cancer in the smoking group to the nonsmoking group. If the two proportions differ significantly, it could be said that smoking causes cancer. This designed experiment is able to control many of the factors that might affect whether one contracts lung cancer that would not be controlled in an observational study, however, it is a very unethical study.

Other reasons exist for conducting observational studies over designed experiments. An article in support of observational studies states, "observational studies have several advantages over designed experiments, including lower cost, greater timeliness, and a broader range of patients." (*Source:* Kjell Benson, BA, and Arthur J. Hartz, MD, PhD. "A Comparison of Observational Studies and Randomized, Controlled Trials," *New England Journal of Medicine* 342:1878–1886, 2000.)

One final thought regarding confounding. In designed experiments, it is possible to have two explanatory variables in a study that are related to each other and related to the response variable. For example, suppose Professor Egner wanted to conduct an experiment in which she compared student success using online homework versus traditional textbook homework. To do the study, she taught her morning statistics class using the online homework and her afternoon class using traditional textbook homework. At the end of the semester, she compared the final exam scores for the online homework section to the textbook homework section. If the morning section had higher scores, could Professor Egner conclude that online homework is the cause of higher exam scores? Not necessarily. It is possible that the morning class had students who were more motivated. It is impossible to know whether the outcome was due to the online homework or to the time at which the class was taught. In this sense, we say that the time of day the class is taught is a *confounding variable*.

Definition

A **confounding variable** is an explanatory variable that was considered in a study whose effect cannot be distinguished from a second explanatory variable in the study.

The big difference between lurking variables and confounding variables is that lurking variables are not considered in the study (for example, we did not consider lifestyle in the pneumonia study) whereas confounding variables are measured in the study (for example, we measured morning versus afternoon classes).

So lurking variables are related to both the explanatory and response variables, and this relation is what creates the apparent association between the explanatory and response variable in the study. For example, lifestyle (healthy or not) is associated with the likelihood of getting an influenza shot as well as the likelihood of contracting pneumonia or influenza.

A confounding variable is a variable in a study that does not necessarily have any association with the other explanatory variable, but does have an effect on the response variable. Perhaps morning students are more motivated, and this is what led to the higher final exam scores, not the homework delivery system.

The bottom line is that both lurking variables and confounding variables can confound the results of a study, so a researcher should be mindful of their potential existence.

We will continue to look at obtaining data through various types of observational studies until Section 1.6. In Section 1.6, we will look at designed experiments.

2 Explain the Various Types of Observational Studies

There are three major categories of observational studies: (1) cross-sectional studies, (2) case-control studies, and (3) cohort studies.

Cross-sectional Studies These observational studies collect information about individuals at a specific point in time or over a very short period of time.

For example, a researcher might want to assess the risk associated with smoking by looking at a group of people, determining how many are smokers, and comparing the rate of lung cancer of the smokers to the nonsmokers.

An advantage of cross-sectional studies is that they are cheap and quick to do. However, they have limitations. For our lung cancer study, individuals might develop cancer after the data are collected, so our study will not give the full picture.

Case-control Studies These studies are **retrospective**, meaning that they require individuals to look back in time or require the researcher to look at existing records. In case-control studies, individuals who have a certain characteristic may be matched with those who do not.

For example, we might match individuals who smoke with those who do not. When we say "match" individuals, we mean that we would like the individuals in the study to be as similar (homogeneous) as possible in terms of demographics and other variables that may affect the response variable. Once homogeneous groups are established, we would ask the individuals in each group how much they smoked over the past 25 years. The rate of lung cancer between the two groups would then be compared.

A disadvantage to this type of study is that it requires individuals to recall information from the past. It also requires the individuals to be truthful in their responses. An advantage of case-control studies is that they can be done relatively quickly and inexpensively.

Cohort Studies A cohort study first identifies a group of individuals to participate in the study (the cohort). The cohort is then observed over a long period of time. During this period, characteristics about the individuals are recorded and some individuals will be exposed to certain factors (not intentionally) and others will not. At the end of the study the value of the response variable is recorded for the individuals.

Typically, cohort studies require many individuals to participate over long periods of time. Because the data are collected over time, cohort studies are **prospective**. Another problem with cohort studies is that individuals tend to drop out due to the long time frame. This could lead to misleading results. That said, cohort studies are the most powerful of the observational studies.

One of the largest cohort studies is the Framingham Heart Study. In this study, more than 10,000 individuals have been monitored since 1948. The study continues to this day, with the grandchildren of the original participants taking part in the study. This cohort study is responsible for many of the breakthroughs in understanding heart disease. Its cost is in excess of $10 million.

Some Concluding Remarks about Observational Studies versus Designed Experiments

Is a designed experiment always superior to an observational study? Not necessarily. Plus, observational studies play a role in the research process. For example, because cross-sectional and case-control observational studies are relatively inexpensive, they allow researchers to explore possible associations prior to undertaking large cohort studies or designing experiments.

Also, it is not always possible to conduct an experiment. For example, we could not conduct an experiment to investigate the perceived link between high tension wires and leukemia (on humans). Do you see why?

NW **Now Work Problem 19**

Census Data

Another source of data is a *census*.

Definition A **census** is a list of all individuals in a population along with certain characteristics of each individual.

The United States conducts a census every 10 years to learn the demographic makeup of the United States. Everyone whose usual residence is within the borders of the United States must fill out a questionnaire packet. The cost of obtaining the census in 2010* was approximately $5.4 billion; about 635,000 temporary workers were hired to assist in collecting the data.

Why is the U.S. Census so important? The results of the census are used to determine the number of representatives in the House of Representatives in each state, congressional districts, distribution of funds for government programs (such as Medicaid), and planning for the construction of schools and roads. The first census of the United States was obtained in 1790 under the direction of Thomas Jefferson. It is a constitutional mandate that a census be conducted every 10 years.

Is the United States successful in obtaining a census? Not entirely. Some individuals go uncounted due to illiteracy or language issues. People experiencing homelessness may also go uncounted. Given the political stakes that are based on the census, politicians often debate how to count these individuals. Statisticians have offered solutions to the counting problem. If you wish, go to www.census.gov and in the search box type *count homeless*. You will find many articles on the Census Bureau's attempt to count the homeless. The bottom line is that even census data can have flaws.

Obtaining Data through Web Scraping

Web scraping, or **data mining**, is the process of extracting data from the Internet. Web scraping can be used to extract data from tables on web pages and then upload the data to a file. Or, web scraping can be used to create a data set of words from an online article (that is, fetching unstructured information and transforming it into a structured format through something called *parsing* and *reformatting processes*). Web scraping can also be used to dynamically call information from websites with links. Web scraping of dynamic information is based on the fact that information on web pages is constantly changing, so some data users might want to learn information over time in an automated fashion. For example, researchers may want to learn information about prices, price changes, and sold-out items over a holiday weekend (such as the weekend after Thanksgiving).

Web scraping is becoming an important part of the data science industry. Because data is so valuable, companies scrape the web constantly to gain an edge on their competition and to learn more about their customers. Web scraping tools will search websites for information on pricing of certain products in order to do instant price comparisons. For example, companies scrape airline flight prices to find the best deals (cheapOair.com). Real estate websites scrape housing purchases to develop home price estimate algorithms (Zillow.com). Or, companies that link individuals with jobs will search the Internet for job openings and share new postings with subscribers who have posted skills that match the job (LinkedIn). Investment companies now scrape social media (primarily Twitter) to determine stockholders' sentiments to try to predict the community reaction to quarterly earnings announcements. Sports analytics firms scrape sports play-by-play information to develop ranking algorithms (Statcast). The list goes on and on.

Web scraping packages are being developed continuously within programming languages such as Python and R. Most data scientists prefer to scrape the web using Python's BeautifulSoup package. BeautifulSoup is widely known as the most advanced library for web scraping. In R, rvest and scrapeR are popular web scraping packages. It is important to realize that these tools are incredibly useful for scraping the web, but they are also limited. Without some understanding of programming languages used in building web pages, your web scraping abilities will be limited to HTML tables.

There are ethical issues associated with web scraping. After all, the scraping of data on these pages is often done without the permission of the host. A lot of dialogue is

*Costs of the 2020 census were not available at the time of printing. However, some estimates place its cost at over $15 billion.

currently taking place in the "web capture" community surrounding the legality and ethics of web capturing. We experience many benefits as a result of web scraping (cheaper flights, the right job offers, and so on), but there are also many pitfalls (personal information is potentially sourced from a website you make a purchase from). What is the responsibility of the website host to protect your information (data)? What are your responsibilities to protect your information?

If you are interested in learning more about the methods of web scraping and the ethics surrounding the technique, type "Web Scraping" or "Web Scraping Ethics" in the search engine of your browser.

If you would like to try web scraping, consult the Student Activity Workbook that accompanies this text. There is an activity that introduces StatCrunchThis—a web scraping tool available with StatCrunch.

Downloading Data from the Web

More than ever you will find that government agencies, companies, and sports organizations regularly make data available to the public. Often, this data can be downloaded from their websites as csv (Excel) or txt (text) files. Then, the data can be uploaded into your favorite statistical analysis package (such as StatCrunch, Minitab, or R) or spreadsheet (such as Excel or Google Spreadsheet). The amount of data we have access to right now is vast and growing. For a small sample of some websites that have data available for download, go to https://www.sullystats.com/resources. This page is updated periodically to keep current with the immense amount of data available for analysis.

1.2 Assess Your Understanding

Vocabulary and Skill Building

1. In your own words, define explanatory variable and response variable.

2. Match each word or phrase with its definition.

Word/Phrase	Definition
(a) Designed Experiment	I. Occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable not accounted for in the study.
(b) Observational Study	II. An explanatory variable that was considered in a study whose effect cannot be distinguished from a second explanatory variable in the study.
(c) Lurking Variable	III. A researcher randomly assigns the individuals in a study to groups, intentionally manipulates the value of an explanatory variable, controls other explanatory variables at fixed values, and then records the value of the response variable for each individual.
(d) Confounding	IV. An explanatory variable that was not considered in a study, but that affects the value of the response variable in the study. In addition, this variable is typically related to other explanatory variables in the study.
(e) Confounding Variable	V. A researcher measures the value of the response variable without attempting to influence the value of either the response or explanatory variables. That is, the researcher observes the behavior of individuals in the study and records the values of the explanatory and response variables.

3. Match each type of study to its definition.

Word	Definition
(a) Cohort Study	I. Studies that are retrospective, meaning they require the researcher to look at existing records, or the subject to recall information from the past. Individuals who have certain characteristics are matched with those who don't.
(b) Cross-sectional Study	II. Studies that follow a group of individuals over a long period of time. Characteristics of the individuals are recorded and some individuals will be exposed to certain factors (not intentionally) and others will not. Because the data are collected over time, these studies are prospective.
(c) Case-control Study	III. Studies that collect information about individuals at a specific point in time, or over a short period of time.

4. Which type of study allows the researcher to claim causation between an explanatory variable and a response variable?

5. Given a choice, would you conduct a study using an observational study or a designed experiment? Why?

6. The data used in the influenza study presented in Example 3 were obtained from a cohort study. What does this mean? Why is a cohort study superior to a case-control study?

7. Explain why it would be unlikely to use a designed experiment to answer the research question posed in Example 3.

8. What does it mean when an observational study is retrospective? What does it mean when an observational study is prospective?

In Problems 9–16, determine whether the study depicts an observational study or an experiment.

NW **9. Cancer Study** The American Cancer Society is beginning a study to learn why some people never get cancer. To take part in the study, a person must be 30–65 years of age and never had cancer. The study requires that the participants fill out surveys about their health and habits and give blood samples and waist measurements. These surveys must be filled out every two years. The study is expected to last for the next 20 years.

10. Rats with cancer are divided into two groups. One group receives 5 milligrams (mg) of a medication that is thought to fight cancer, and the other receives 10 mg. After 2 years, the spread of the cancer is measured.

11. Seventh-grade students are randomly divided into two groups. One group is taught math using traditional techniques; the other is taught math using a reform method. After 1 year, each group is given an achievement test to compare proficiency.

12. Hair and Heart Disease A study in which balding men were compared with non-balding men at one point in time found that balding men were 70% more likely to have heart disease.

Source: USA Today, April 4, 2013.

13. A survey is conducted asking 400 people, "Do you prefer Coke or Pepsi?"

14. Two hundred people are asked to perform a taste test in which they drink from two randomly placed, unmarked cups and are asked which drink they prefer.

15. Sixty patients with carpal tunnel syndrome are randomly divided into two groups. One group is treated weekly with both acupuncture and an exercise regimen. The other is treated weekly with the exact same exercise regimen, but no acupuncture. After 1 year, both groups are questioned about their level of pain due to carpal tunnel syndrome.

16. Conservation agents netted 250 large-mouth bass in a lake and determined how many were carrying parasites.

Applying the Concepts

17. Happiness and Your Heart Is there an association between level of happiness and the risk of heart disease? Researchers studied 1739 people over a 10-year period and asked questions about their daily lives and the hassles they face. The researchers also determined which individuals in the study experienced any type of heart disease. After their analysis, they concluded that happy individuals are less likely to experience heart disease.

Source: European Heart Journal 31 (9):1065–1070, February 2010.

(a) What type of observational study is this? Explain.
(b) What is the response variable? What is the explanatory variable?
(c) In the report, the researchers stated that "the research team also hasn't ruled out that a common factor like genetics could be causing both the emotions and the heart disease." Use the language introduced in this section to explain what this sentence means.

18. Daily Coffee Consumption Is there an association between daily coffee consumption and the occurrence of skin cancer? Researchers asked 93,676 women to disclose their coffee-drinking habits and also determined which of the women had nonmelanoma skin cancer. The researchers concluded that consumption of six or more cups of caffeinated coffee per day was associated with a reduction in nonmelanoma skin cancer.

Source: European Journal of Cancer Prevention, 16(5): 446–452, October 2007.

(a) What type of observational study was this? Explain.
(b) What is the response variable in the study? What is the explanatory variable?
(c) In their report, the researchers stated that "After adjusting for various demographic and lifestyle variables, daily consumption of six or more cups was associated with a 30% reduced prevalence of nonmelanoma skin cancer." Why was it important to adjust for these variables?

NW **19. Television in the Bedroom** Is a television (TV) in the bedroom associated with obesity? Researchers questioned 379 twelve-year-old adolescents and concluded that the body mass index (BMI) of the adolescents who had a TV in their bedroom was significantly higher than the BMI of those who did not have a TV in their bedroom.

Source: Christelle Delmas, Carine Platat, Brigitte Schweitzer, Aline Wagner, Mohamed Oujaa, and Chantal Simon. "Association Between Television in Bedroom and Adiposity Throughout Adolescence," *Obesity*, 15:2495–2503, 2007.

(a) Why is this an observational study? What type of observational study is this?
(b) What is the response variable in the study? What is the explanatory variable?
(c) Can you think of any lurking variables that may affect the results of the study?
(d) In the report, the researchers stated, "These results remain significant after adjustment for socioeconomic status." What does this mean?
(e) Can we conclude that a television in the bedroom causes a higher body mass index? Explain.

20. Get Married, Gain Weight Are young couples who marry or cohabitate more likely to gain weight than those who stay single? Researchers followed 8000 men and women for 7 years. At the start of the study, none of the participants were married or living with a romantic partner. The researchers found that women who married or cohabitated during the study gained 9 pounds more than single women, and married or cohabitating men gained, on average, 6 pounds more than single men.

(a) Why is this an observational study? What type of observational study is this?
(b) What is the response variable in the study? What is the explanatory variable?
(c) Identify some potential lurking variables in this study.
(d) Can we conclude that getting married or cohabiting causes one to gain weight? Explain.

21. Midwives Researchers Sally Tracy and associates undertook a cross-sectional study looking at the method of delivery and cost of delivery for first-time "low risk" mothers under three delivery scenarios:

(1) Caseload midwifery
(2) Standard hospital care
(3) Private obstetric care

The results of the study revealed that 58.5% of all births with midwifery were vaginal deliveries compared with 48.2% of standard hospital births and 30.8% of private obstetric care. In addition, the costs of delivery from midwifery was $3903.78 compared with $5279.23 for standard hospital care and $5413.69 for private obstetric care.

Source: Sally K Tracy, Alec Welsh, Bev Hall, Donna Hartz, Anne Lainchbury, Andrew Bisits, Jan White, and Mark Tracy "Caseload midwifery compared to standard or private obstetric care for first time mothers in a public teaching hospital in Australia: a cross sectional study of cost and birth outcomes" *BMC Pregnancy and Childbirth* 2014, 14:46.

(a) Why is this a cross-sectional observational study?
(b) Name the explanatory variable in the study.
(c) Name the two response variables in the study and determine whether each is qualitative or quantitative.

22. Web Page Design Magnum, LLC, is a web page design firm that has two designs for an online hardware store. To determine which is the more effective design, Magnum uses one page in the Denver area and a second page in the Miami area. For each visit, Magnum records the amount of time visiting the site and the amount spent by the visitor.

(a) What is the explanatory variable in this study? Is it qualitative or quantitative?
(b) What are the two response variables? For each response variable, state whether it is qualitative or quantitative.
(c) Explain how confounding might be an issue with this study.

23. Analyze the Article Write a summary of the following opinion. The opinion was posted at abcnews.com. Include the type of study conducted, possible lurking variables, and conclusions. What is the message of the author of the article?

Power Lines and Cancer—To Move or Not to Move

New Research May Cause More Fear Than Warranted, One Physician Explains

OPINION by JOSEPH MOORE, M.D.

A recent study out of Switzerland indicates there might be an increased risk of certain blood cancers in people with prolonged exposure to electromagnetic fields, like those generated from high-voltage power lines.

If you live in a house near one of these high-voltage power lines, a study like this one might make you wonder whether you should move.

But based on what we know now, I don't think that's necessary. We can never say there is no risk, but we can say that the risk appears to be extremely small.

"Scare Science"

The results of studies like this add a bit more to our knowledge of potential harmful environmental exposures, but they should also be seen in conjunction with the results of hundreds of studies that have gone before. It cannot be seen as a definitive call to action in and of itself.

The current study followed more than 20,000 Swiss railway workers over a period of 30 years. True, that represents a lot of people over a long period of time.

However, the problem with many epidemiological studies, like this one, is that it is difficult to have an absolute control group of people to compare results with. The researchers compared the incidence of different cancers of workers with a high amount of electromagnetic field exposure to those workers with lower exposures.

These studies aren't like those that have identified definitive links between an exposure and a disease— like those involving smoking and lung cancer. In those studies, we can actually measure the damage done to lung tissue as a direct result of smoking. But usually it's very difficult for the conclusions of an epidemiological study to rise to the level of controlled studies in determining public policy.

Remember the recent scare about coffee and increased risk of pancreatic cancer? Or the always-simmering issue of cell phone use and brain tumors?

As far as I can tell, none of us have turned in our cell phones. In our own minds, we've decided that any links to cell phone use and brain cancer have not been proven definitively. While we can't say that there is absolutely no risk in using cell phones, individuals have determined on their own that the potential risks appear to be quite small and are outweighed by the benefits.

Findings Shouldn't Lead to Fear

As a society, we should continue to investigate these and other related exposures to try to prove one way or another whether they are disease-causing. If we don't continue to study, we won't find out. It's that simple.

When findings like these come out, and I'm sure there will be more in the future, I would advise people not to lose their heads. Remain calm. You should take the results as we scientists do—as intriguing pieces of data about a problem we will eventually learn more about, either positively or negatively, in the future. It should not necessarily alter what we do right now.

What we can do is take actions that we know will reduce our chances of developing cancer.

Stop smoking and avoid passive smoke. It is the leading cause of cancer that individuals have control over.

Whenever you go outside, put on sunscreen or cover up.

Eat a healthy diet and stay physically active.

Make sure you get tested or screened. Procedures like colonoscopies, mammograms, pap smears and prostate exams can catch the early signs of cancer, when the chances of successfully treating them are the best.

Taking the actions above will go much farther in reducing your risks for cancer than moving away from power lines or throwing away your cell phone.

Dr. Joseph Moore is a medical oncologist at Duke University Comprehensive Cancer Center.

Source: Copyright © by Joseph Moore.

24. Cellular Phones Researchers wanted to determine whether there is an association between mobile phone use and body mass index. To do so, 105,028 men and women aged 18 years or over from the United Kingdom were recruited and their cell-phone use behavior was studied (number of calls per day, number of hours per week, year cell phone was first used) along with other variables (amount of exercise, body mass index) of the individuals. The researchers found a strong positive association between duration of phone calls on a cell phone and body mass index (that is, as the duration of phone calls increases, body mass index tends to increase as well).

Source: Mireille B. Toledano, Rachel B. Smith, Irene Chang, Margaret Douglass, and Paul Elliott, "Cohort Profile: UK COSMOS—a UK cohort for study of environment and health," *International Journal of Epidemiology*, 46(3):775–787, June 1, 2017, https://doi.org/10.1093/ije/dyv203

(a) What type of observational study is this?

(b) Many studies involving cell phones look for a link between cell-phone usage and negative health outcomes (such as stroke or cancer) due to radio-frequency exposure. The following quote is from the article: "Obesity is associated with health outcomes such as stroke and cancers, which are of interest in relation to radio frequency exposure, and therefore is potential for confounding." Explain what this means.

25. A Flawed Retrospective Study In an infamous study, researchers suggested that left-handed individuals died younger than right-handed individuals. In the study, researchers identified 987 individuals who died in 1990 and then used historical records to determine birth year as well as whether the individual was right-handed or not. They found that individuals who were right-handed lived 75 years, on average, while those who were not right-handed (left-handed or ambidextrous) lived 66 years, on average. Explain the flaw in this retrospective study and point out the potential dangers in retrospective studies. *Hint:* In the early 1900s individuals were often pressured to become right-handed at an early age. This pressure subsided, and the percentage of individuals born around 1950 or later that are left-handed is around 10%–12%, the norm.

26. Putting It Together: Passive Smoke? The following abstract appears in *The New England Journal of Medicine*:

> BACKGROUND. The relation between passive smoking and lung cancer is of great public health importance. Some previous studies have suggested that exposure to environmental tobacco smoke in the household can cause lung cancer, but others have found no effect. Smoking by the spouse has been the most commonly used measure of this exposure.
>
> METHODS. In order to determine whether lung cancer is associated with exposure to tobacco smoke within the household, we conducted a case-control study of 191 patients with lung cancer who had never smoked and an equal number of persons without lung cancer who had never smoked. Lifetime residential histories including information on exposure to environmental tobacco smoke were compiled and analyzed. Exposure was measured in terms of "smoker-years," determined by multiplying the number of years in each residence by the number of smokers in the household.
>
> RESULTS. Household exposure to 25 or more smoker-years during childhood and adolescence doubled the risk of lung cancer. Approximately 15 percent of the control subjects who had never smoked reported this level of exposure. Household exposure of less than 25 smoker-years during childhood and adolescence did not increase the risk of lung cancer. Exposure to a spouse's smoking, which constituted less than one third of total household exposure on average, was not associated with an increase in risk.
>
> CONCLUSIONS. The possibility of recall bias and other methodologic problems may influence the results of case-control studies of environmental tobacco smoke. Nonetheless, our findings regarding exposure during early life suggest that approximately 17 percent of lung cancers among nonsmokers can be attributed to high levels of exposure to cigarette smoke during childhood and adolescence.

(a) What is the research objective?
(b) What makes this study a case-control study? Why is this a retrospective study?
(c) What is the response variable in the study? Is it qualitative or quantitative?
(d) What is the explanatory variable in the study? Is it qualitative or quantitative?
(e) Can you identify any lurking variables that may have affected this study?
(f) What is the conclusion of the study? Can we conclude that exposure to smoke in the household causes lung cancer?
(g) Would it be possible to design an experiment to answer the research question in part (a)? Explain.

27. Name three ways that web scraping can be used to obtain data.

28. Discuss the ethics behind scraping data from the Internet. In particular, answer the following questions. What is the responsibility of the website host to protect your information (data)? What are your responsibilities to protect your information? For assistance, type "Web Scraping" or "Web Scraping Ethics" in the search engine of your browser.

1.3 Simple Random Sampling

Objective 1 Obtain a simple random sample

Sampling

Besides the observational studies that we looked at in Section 1.2, observational studies can also be conducted by administering a survey. When administering a survey, the researcher must first identify the population that is to be targeted. For example, the Gallup Organization regularly surveys Americans about various pop-culture and political issues. Often, the population of interest is Americans aged 18 years or older. Of course, the Gallup Organization cannot survey *all* adult Americans (there are over 200 million), so instead the group typically surveys a *random sample* of about 1000 adult Americans.

Definition **Random sampling** is the process of using chance to select individuals from a population to be included in the sample.

For the results of a survey to be reliable, the characteristics of the individuals in the sample must be representative of the characteristics of the individuals in the population. The key to obtaining a sample representative of a population is to let *chance* or *randomness* play a role in dictating which individuals are in the sample, rather than convenience. **If convenience is used to obtain a sample, the results of the survey are meaningless.**

Suppose that Gallup wants to know the proportion of adult Americans who consider themselves to be baseball fans. If Gallup obtained a sample by standing outside of Fenway Park (home of the Boston Red Sox professional baseball team), the survey results are not likely to be reliable. Why? Clearly, the individuals in the sample do not accurately reflect the makeup of the entire population. As another example, suppose you wanted to learn the proportion of students on your campus who work. It might be convenient to survey the students in your statistics class, but do these students represent the overall student body? Does the proportion of freshmen, sophomores, juniors, and seniors in your class mirror the proportion of freshmen, sophomores, juniors, and seniors on campus? Does the proportion of males and females in your class resemble the proportion of males and females across campus? Probably not. For this reason, the convenient sample is not representative of the population, which means any results reported from your survey are misleading.

We will discuss four basic sampling techniques: *simple random sampling*, *stratified sampling*, *systematic sampling*, and *cluster sampling*. These sampling methods are designed so that any selection biases introduced (knowingly or unknowingly) by the surveyor during the selection process are eliminated. In other words, the surveyor does not have a choice as to which individuals are in the study. We will discuss simple random sampling now and the remaining three types of sampling in Section 1.4.

1 Obtain a Simple Random Sample

The most basic sample survey design is *simple random sampling*.

Definition

A sample of size n from a population of size N is obtained through **simple random sampling** if every possible sample of size n has an equally likely chance of occurring. The sample is then called a **simple random sample**.

IN OTHER WORDS
Simple random sampling is like selecting names from a hat.

The number of individuals in the sample is always less than the number of individuals in the population.

EXAMPLE 1 Illustrating Simple Random Sampling

Problem Sophia has four tickets to a concert. Six of her friends, Yolanda, Michael, Kevin, Marissa, Annie, and Katie, have all expressed an interest in going to the concert. Sophia decides to randomly select three of her six friends to attend the concert.

(a) List all possible samples of size $n = 3$ from the population of size $N = 6$. Once an individual is chosen, he or she cannot be chosen again.

(b) Comment on the likelihood of the sample containing Michael, Kevin, and Marissa.

Approach List all possible combinations of three people chosen from the six. Remember, in simple random sampling, each sample of size 3 is equally likely to occur.

Solution

(a) The possible samples of size 3 are listed in Table 2.

Table 2

Yolanda, Michael, Kevin	Yolanda, Michael, Marissa	Yolanda, Michael, Annie	Yolanda, Michael, Katie
Yolanda, Kevin, Marissa	Yolanda, Kevin, Annie	Yolanda, Kevin, Katie	Yolanda, Marissa, Annie
Yolanda, Marissa, Katie	Yolanda, Annie, Katie	Michael, Kevin, Marissa	Michael, Kevin, Annie
Michael, Kevin, Katie	Michael, Marissa, Annie	Michael, Marissa, Katie	Michael, Annie, Katie
Kevin, Marissa, Annie	Kevin, Marissa, Katie	Kevin, Annie, Katie	Marissa, Annie, Katie

From Table 2, we see that there are 20 possible samples of size 3 from the population of size 6. The term *sample* means the individuals in the sample.

(b) Only 1 of the 20 possible samples contains Michael, Kevin, and Marissa, so there is a 1 in 20 chance that the simple random sample will contain these three. In fact, all the samples of size 3 have a 1 in 20 chance of occurring.

NW **Now Work Problem 7**

Obtaining a Simple Random Sample

The results of Example 1 leave one question unanswered: How do we select the individuals in a simple random sample? We could write the names of the individuals in the population on different sheets of paper and then select names from a hat. Often, however, the size of the population is so large that performing simple random sampling in this fashion is not practical. Instead, each individual in the population is assigned a unique number between 1 and N, where N is the size of the population. Then n distinct random numbers from this list are selected, where n represents the size of the sample. To number the individuals in the population, we need a **frame**—a list of all the individuals within the population.

IN OTHER WORDS
A frame lists all the individuals in a population. For example, a list of all registered voters in a particular precinct might be a frame.

EXAMPLE 2 Obtaining a Simple Random Sample Using a Table of Random Numbers

Problem The accounting firm of Senese and Associates has grown. To make sure their clients are still satisfied with the services they are receiving, the company decides to send a survey out to a simple random sample of 5 of its 30 clients.

Approach

Step 1 The clients must be listed (the frame) and numbered from 01 to 30.

Step 2 Five unique numbers will be randomly selected. The clients corresponding to the numbers are sent a survey. This process is called *sampling without replacement*. In a **sample without replacement**, an individual who is selected is removed from the population and cannot be chosen again. In a **sample with replacement**, a selected individual is placed back into the population and could be chosen a second time. We use sampling without replacement so that we don't select the same client twice.

Solution

Step 1 Table 3 shows the list of clients. We arrange them in alphabetical order, although this is not necessary, and number them from 01 to 30.

Table 3

01. ABC Electric	11. Fox Studios	21. R&Q Realty
02. Brassil Construction	12. Haynes Hauling	22. Ritter Engineering
03. Bridal Zone	13. House of Hair	23. Simplex Forms
04. Casey's Glass House	14. John's Bakery	24. Spruce Landscaping
05. Chicago Locksmith	15. Logistics Management, Inc.	25. Thors, Robert DDS
06. DeSoto Painting	16. Lucky Larry's Bistro	26. Travel Zone
07. Dino Jump	17. Moe's Exterminating	27. Ultimate Electric
08. Euro Car Care	18. Nick's Tavern	28. Venetian Gardens Restaurant
09. Farrell's Antiques	19. Orion Bowling	29. Walker Insurance
10. First Fifth Bank	20. Precise Plumbing	30. Worldwide Wireless

Step 2 A table of random numbers can be used to select the individuals to be in the sample. See Table 4 on the next page* We pick a starting place in the table by closing

*Each digit is in its own column. The digits are displayed in groups of five for ease of reading. The digits in row 1 are 893922321274483, and so on. The first digit, 8, is in column 1; the second digit, 9, is in column 2; the ninth digit, 1, is in column 9.

(continued)

Table 4

Row Number	Column Number									
	01–05	06–10	11–15	16–20	21–25	26–30	31–35	36–40	41–45	46–50
01	89392	**23**212	74483	36590	25956	36544	68518	40805	09980	00467
02	61458	17639	96252	95649	73727	33912	72896	66218	52341	97141
03	11452	74197	81962	48433	90360	26480	73231	37740	26628	44690
04	27575	04429	31308	02241	01698	19191	18948	78871	36030	23980
05	36829	59109	88976	46845	28329	47460	88944	08264	00843	84592
06	81902	93458	42161	26099	09419	89073	82849	09160	61845	40906
07	59761	55212	33360	68751	86737	79743	85262	31887	37879	17525
08	46827	25906	64708	20307	78423	15910	86548	08763	47050	18513
09	24040	66449	32353	83668	13874	86741	81312	54185	78824	00718
10	98144	96372	50277	15571	82261	66628	31457	00377	63423	55141
11	14228	17930	30118	00438	49666	65189	62869	31304	17117	71489
12	55366	51057	90065	14791	62426	02957	85518	28822	30588	32798
13	961**01**	30646	35526	90389	73634	79304	96635	06626	94683	16696
14	381**52**	55474	30153	26525	83647	31988	82182	98377	33802	80471
15	850**07**	18416	24661	95581	45868	15662	28906	36392	07617	50248
16	855**44**	15890	80011	18160	33468	84106	40603	01315	74664	20553
17	104**46**	20699	98370	17684	16932	80449	92654	02084	19985	59321
18	672**37**	45509	17638	65115	29757	80705	82686	48565	72612	61760
19	230**26**	89817	05403	82209	30573	47501	00135	33955	50250	72592
20	674**11**	58542	18678	46491	13219	84084	27783	34508	55158	78742

Column 4

Row 13

We skip 52 because it is larger than 30.

our eyes and placing a finger on it. This method accomplishes the goal of being random. Suppose we start in column 4, row 13. Because our data have two digits, we select two-digit numbers from the table using columns 4 and 5. We select numbers between 01 and 30, inclusive, and skip 00, numbers greater than 30, and numbers already selected.

The first number in the list is 01, so the client corresponding to 01 will receive a survey. Reading down, the next number in the list is 52, which is greater than 30, so we skip it. Continuing down the list, the following numbers are selected from the list:

01, 07, 26, 11, 23

We display each of the random numbers used to select the individuals in the sample in boldface type in Table 4 to help you to understand where they came from. The clients corresponding to these numbers are

ABC Electric, Dino Jump, Travel Zone, Fox Studios, Simplex Forms

EXAMPLE 3 Obtaining a Simple Random Sample Using Technology

Problem Find a simple random sample of five clients for the problem presented in Example 2.

Approach The approach is similar to that given in Example 2.

Step 1 Obtain the frame and assign the clients numbers from 01 to 30.

Step 2 Randomly select five numbers using a random number generator. To do this, we must first set the *seed*. The **seed** is an initial point for the generator to start creating

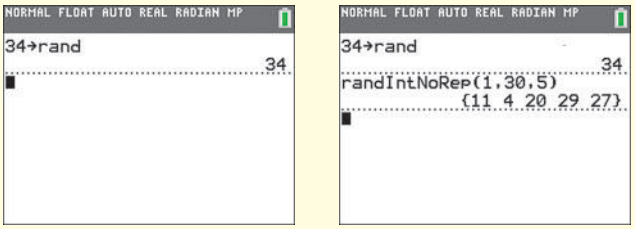
random numbers—like selecting the initial point in the table of random numbers. The seed can be any nonzero number. Statistical software such as StatCrunch, Minitab, or Excel can be used to generate random numbers, but we will use a TI-84 Plus C graphing calculator. The steps for obtaining random numbers using StatCrunch, Minitab, Excel, and the TI-83/84 Plus/84 Plus C graphing calculator can be found in the Technology Step-by-Step shown below.

Solution

Step 1 Table 3 on page 25 shows the list of clients and numbers corresponding to the clients.

Step 2 Figure 3(a) shows the seed set at 34 on a TI-84 Plus C graphing calculator. Now we can generate a list of random numbers, which are shown in Figure 3(b).

Figure 3



(a) (b)

Using Technology

If you are using a different statistical package or type of calculator, the random numbers generated will likely be different. This does not mean you are wrong. There is no such thing as a wrong random sample as long as the correct procedures are followed.

The following numbers are generated by the calculator:

11, 4, 20, 29, 27

The clients corresponding to these numbers are the clients to be surveyed: Fox Studios, Casey's Glass House, Precise Plumbing, Walker Insurance, and Ultimate Electric. ○

NW Now Work Problem 11

CAUTION!

Random-number generators are not truly random, because they are programs, and programs do not act "randomly." The seed dictates the random numbers that are generated.

Notice an important difference in the solutions of Examples 2 and 3. Because both samples were obtained randomly, they resulted in different individuals in the sample! For this reason, each sample will likely result in different descriptive statistics. Any inference based on each sample *may* result in different conclusions regarding the population. This is the nature of statistics. Inferences based on samples will vary because the individuals in different samples vary.

Technology Step-by-Step Obtaining a Simple Random Sample

TI-83/84 Plus

1. Enter any nonzero number (the seed) on the HOME screen.
2. Press the STO ▶ button.
3. Press the MATH button.
4. Highlight the PRB menu and select 1: rand.
5. From the HOME screen press ENTER.
6. Press the MATH button. Highlight the PRB menu and select 5: randInt(.
7. With randInt(on the HOME screen, enter 1, N), where N is the population size. For example, if $N = 500$, enter the following:

randInt(1,500)

Press ENTER to obtain the first individual in the sample. Continue pressing ENTER until the desired sample size is obtained.

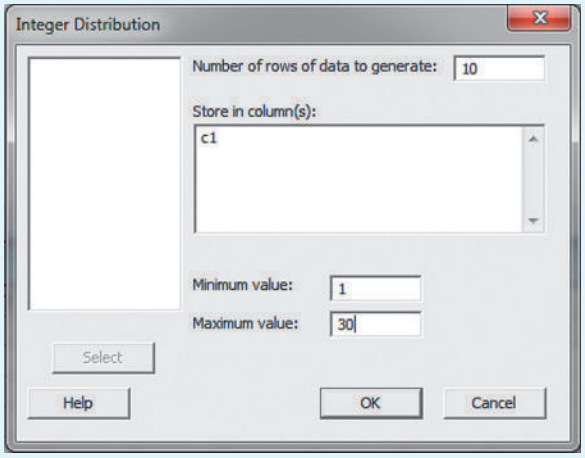
TI-84 Plus C

1. Enter any nonzero number (the seed) on the HOME screen.
2. Press the STO ▶ button.
3. Press the MATH button.
4. Highlight the PROB menu and select 1: rand.
5. From the HOME screen press ENTER.
6. Press the MATH button.
7. Highlight the PROB menu and select 8: randIntNoRep(.

8. Type in the values for lower, upper, and n.
9. Highlight Paste. Press ENTER. Press ENTER a second time from the HOME screen.

Minitab

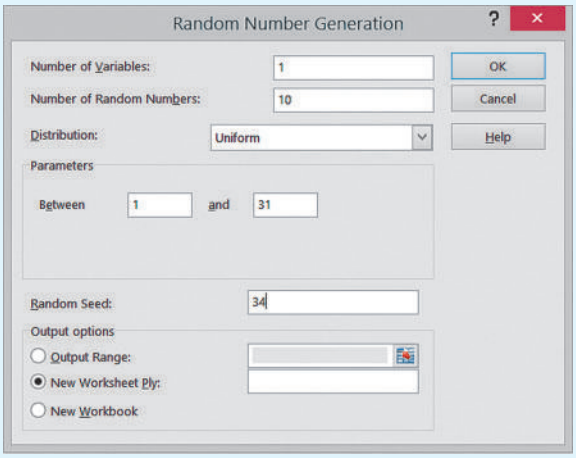
1. Select the **Calc** menu and highlight **Set Base . . .** .
2. Enter any seed number you desire. Note that it is not necessary to set the seed, because Minitab uses the time of day in seconds to set the seed.
3. Select the **Calc** menu, highlight **Random Data**, and select **Integer . . .** .
4. Fill in the following window with the appropriate values. To obtain a simple random sample for the situation in Example 2, we would enter the following:



Generate 10 rows of data (instead of 5) in case any of the random numbers repeat. Select OK, and the random numbers will appear in column 1 (C1) in the spreadsheet.

Excel

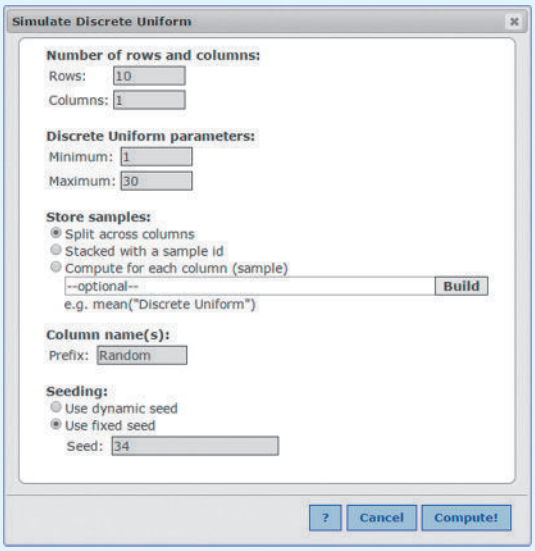
1. Be sure the Data Analysis ToolPak is activated. This is done by selecting **File** and then **Options**. Select **Add-Ins**. Under **Manage**:, select **Excel Add-Ins**. Click **Go** . . . . Check the **Analysis ToolPak** box. Click **OK**.
2. Select **Data** and then **Data Analysis**. Highlight **Random Number Generation** and select **OK**.
3. Fill in the window with the appropriate values. To obtain a simple random sample for the situation in Example 2, fill in the window as shown in the next column. Generate 10 rows of data (instead of 5) in case any of the random numbers repeat. Notice also that the parameter is between 1 and 31, so any value greater than or equal to 1 and less than or equal to 31 is possible. In the unlikely event that 31 appears, simply ignore it. Select **OK** and



the random numbers will appear in column 1 (A1) in the spreadsheet. Ignore any values to the right of the decimal place.

StatCrunch

1. Select **Data**, highlight **Simulate**, then select **Discrete Uniform**.
2. Fill in the window with the appropriate values. To obtain a simple random sample for the situation in Example 2, enter the values shown in the figure. Generate 10 rows of data (instead of 5) in case any of the random numbers repeat. Click Compute!, and the random numbers will appear in the spreadsheet. *Note:* You could also select the dynamic seed radio button, if you like, to set the seed.



1.3 Assess Your Understanding

Vocabulary and Skill Building

1. What is a frame?
2. Define simple random sampling.
3. What does it mean when sampling is done without replacement?
4. What is random sampling? Why is it used and how does it compare with convenience sampling?

5. Literature As part of a college literature course, students must read three classic works of literature from the provided list. Obtain a simple random sample of size 3 from this list. Write a short description of the process you used to generate your sample.

Pride and Prejudice	*The Scarlet Letter*
As I Lay Dying	*The Jungle*
Death of a Salesman	*Huckleberry Finn*
The Sun Also Rises	*Crime and Punishment*
A Tale of Two Cities	

6. Team Captains A coach must select two players to serve as captains. He wants to randomly select two players to be the captains. Obtain a simple random sample of size 2 from the following list: Mady, Breanne, Evin, Tori, Emily, Clair, Caty, Jory, Payton, Jordyn. Write a short description of the process you used to generate your sample.

NW **7. Course Selection** A student entering a doctoral program in educational psychology is required to select two courses from the list of courses provided as part of his or her program.

EPR 616, *Research in Child Development*
EPR 630, *Educational Research Planning and Interpretation*
EPR 631, *Nonparametric Statistics*
EPR 632, *Methods of Multivariate Analysis*
EPR 645, *Theory of Measurement*
EPR 649, *Fieldwork Methods in Educational Research*
EPR 650, *Interpretive Methods in Educational Research*

(a) List all possible two-course selections.
(b) Comment on the likelihood that the pair of courses EPR 630 and EPR 645 will be selected.

8. Merit Badge Requirements To complete the Citizenship in the World merit badge, one must select two of the following seven organizations and describe their role in the world.

Source: Boy Scouts of America

1. The United Nations
2. The World Court
3. World Organization of the Scout Movement
4. The World Health Organization
5. Amnesty International
6. The International Committee of the Red Cross
7. CARE

(a) List all possible pairs of organizations.
(b) Comment on the likelihood that the pair "The United Nations" and "Amnesty International" will be selected.

Applying the Concepts

9. Sampling the Faculty A community college employs 87 full-time faculty members. To gain the faculty's opinions about an upcoming building project, the college president wishes to obtain a simple random sample that will consist of 9 faculty members. He numbers the faculty from 1 to 87.

(a) Using Table I from Appendix A, the president closes his eyes and drops his ink pen on the table. It points to the digit in row 5, column 22. Using this position as the starting point and proceeding downward, determine the numbers for the 9 faculty members who will be included in the sample.
(b) If the president uses technology, determine the numbers for the 9 faculty members who will be included in the sample.

10. Sampling the Students The same community college from Problem 9 has 7656 students currently enrolled in classes. To gain the students' opinions about an upcoming building project, the college president wishes to obtain a simple random sample of 20 students. He numbers the students from 1 to 7656.

(a) Using Table I from Appendix A, the president closes his eyes and drops his ink pen on the table. It points to the digit in row 11, column 32. Using this position as the starting point and proceeding downward, determine the numbers for the 20 students who will be included in the sample.
(b) If the president uses technology, determine the numbers for the 20 students who will be included in the sample.

NW DATA **11. Obtaining a Simple Random Sample** Suppose the president of the United States wants to hold a forum with the governors of ten states. Use the frame shown to answer parts (a) and (b).

1. Alabama	18. Louisiana	35. Ohio
2. Alaska	19. Maine	36. Oklahoma
3. Arizona	20. Maryland	37. Oregon
4. Arkansas	21. Massachusetts	38. Pennsylvania
5. California	22. Michigan	39. Rhode Island
6. Colorado	23. Minnesota	40. South Carolina
7. Connecticut	24. Mississippi	41. South Dakota
8. Delaware	25. Missouri	42. Tennessee
9. Florida	26. Montana	43. Texas
10. Georgia	27. Nebraska	44. Utah
11. Hawaii	28. Nevada	45. Vermont
12. Idaho	29. New Hampshire	46. Virginia
13. Illinois	30. New Jersey	47. Washington
14. Indiana	31. New Mexico	48. West Virginia
15. Iowa	32. New York	49. Wisconsin
16. Kansas	33. North Carolina	50. Wyoming
17. Kentucky	34. North Dakota	

(a) Obtain a simple random sample of size 10 using Table I in Appendix A, a graphing calculator, or computer software.
(b) Obtain a second simple random sample of size 10 using Table I in Appendix A, a graphing calculator, or computer software.

DATA **12. Obtaining a Simple Random Sample** The following table lists the 45 presidents of the United States.

1. Washington	16. Lincoln	31. Hoover
2. J. Adams	17. A. Johnson	32. F. D. Roosevelt
3. Jefferson	18. Grant	33. Truman
4. Madison	19. Hayes	34. Eisenhower
5. Monroe	20. Garfield	35. Kennedy
6. J. Q. Adams	21. Arthur	36. L. B. Johnson
7. Jackson	22. Cleveland	37. Nixon
8. Van Buren	23. B. Harrison	38. Ford
9. W. H. Harrison	24. Cleveland	39. Carter
10. Tyler	25. McKinley	40. Reagan
11. Polk	26. T. Roosevelt	41. G. H. Bush
12. Taylor	27. Taft	42. Clinton
13. Fillmore	28. Wilson	43. G. W. Bush
14. Pierce	29. Harding	44. Obama
15. Buchanan	30. Coolidge	45. Trump

(a) Obtain a simple random sample of size 8 using Table I in Appendix A, a graphing calculator, or computer software.
(b) Obtain a second simple random sample of size 8 using Table I in Appendix A, a graphing calculator, or computer software.