GLOBAL EDITION

Biostatistics

for the Biological and Health Sciences

THIRD EDITION

Marc M. Triola • Mario F. Triola • Jason Roy

Symbol Table

f	frequency with which a value occurs
Σ	capital sigma; summation
Σx	sum of the values
Σx^2	sum of the squares of the values
$(\Sigma x)^2$	square of the sum of all values
Σxy	sum of the products of each x value multiplied by the corresponding y value
n	number of values in a sample
N	number of values in a finite population; also used as the size of all samples combined
$n!$	n factorial
k	number of samples or populations or categories
$\overline{x}$	mean of the values in a sample
$\overline{R}$	mean of the sample ranges
μ	mu; mean of all values in a population
s	standard deviation of a set of sample values
σ	lowercase sigma; standard deviation of all values in a population
s^2	variance of a set of sample values
σ^2	variance of all values in a population
z	standard score
$z_{\alpha/2}$	critical value of z
t	t distribution
$t_{\alpha/2}$	critical value of t
df	number of degrees of freedom
F	F distribution
χ^2	chi-square distribution
χ^2_R	right-tailed critical value of chi-square
χ^2_L	left-tailed critical value of chi-square
p	probability of an event or the population proportion
q	probability or proportion equal to $1 - p$
$\hat{p}$	sample proportion
$\hat{q}$	sample proportion equal to $1 - \hat{p}$
$\overline{p}$	proportion obtained by pooling two samples
$\overline{q}$	proportion or probability equal to $1 - \overline{p}$
$P(A)$	probability of event A
$P(A\|B)$	probability of event A, assuming event B has occurred
${}_nP_r$	number of permutations of n items selected r at a time
${}_nC_r$	number of combinations of n items selected r at a time
$\overline{A}$	complement of event A
H_0	null hypothesis
H_1	alternative hypothesis
α	alpha; probability of a type I error or the area of the critical region
β	beta; probability of a type II error
r	sample linear correlation coefficient
ρ	rho; population linear correlation coefficient
r^2	coefficient of determination
R^2	multiple coefficient of determination
r_s	Spearman's rank correlation coefficient
b_1	point estimate of the slope of the regression line
b_0	point estimate of the y-intercept of the regression line
$\hat{y}$	predicted value of y
d	difference between two matched values
$\overline{d}$	mean of the differences d found from matched sample data
s_d	standard deviation of the differences d found from matched sample data
s_e	standard error of estimate
T	rank sum; used in the Wilcoxon signed-ranks test

continued

Symbol Table

H	Kruskal-Wallis test statistic
R	sum of the ranks for a sample; used in the Wilcoxon rank-sum test
μ_R	expected mean rank; used in the Wilcoxon rank-sum test
σ_R	expected standard deviation of ranks; used in the Wilcoxon rank-sum test
$\mu_{\overline{x}}$	mean of the population of all possible sample means $\overline{x}$
$\sigma_{\overline{x}}$	standard deviation of the population of all possible sample means $\overline{x}$
E	margin of error of the estimate of a population parameter, or expected value
Q_1, Q_2, Q_3	quartiles
$D_1, D_2, \ldots, D_9$	deciles
$P_1, P_2, \ldots, P_{99}$	percentiles
x	data value

THIRD EDITION
GLOBAL EDITION

BIOSTATISTICS

FOR THE BIOLOGICAL AND HEALTH SCIENCES

MARC M. TRIOLA, MD, FACP
New York University School of Medicine

MARIO F. TRIOLA
Dutchess Community College

JASON ROY, PHD
Rutgers School of Public Health

Pearson

Product Management: Shabnam Dohutia, Shahana Bhattacharya, Aaditya Bugga, Amrita Dutta, and Priya Mishra

Product Marketing: Ellie Nicholls

Rights and Permissions: Anjali Singh and Ashish Vyas

Content Production: Abhilasha Watsa

Cover image: Pablo Hidalgo/Shutterstock

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

Pearson Education Limited

KAO Two
KAO Park
Hockham Way
Harlow, Essex
CM17 9SR
United Kingdom

and Associated Companies throughout the world

Visit us on the World Wide Web at: www.pearsonglobaleditions.com

ISBN 10 (Print): 1-292-45201-3
ISBN 13 (Print): 978-1-292-45201-2
ISBN 13 (uPDF eBook): 978-1-292-45203-6

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

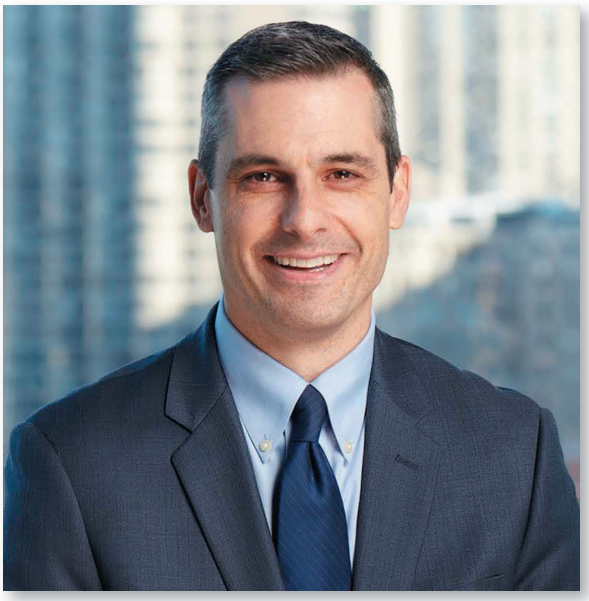
eBook formatted by B2R Technologies Pvt. Ltd.

To Ginny
Dushana and Marisa
Trevor and Mitchell

This third edition of Biostatistics for the Biological and Health Sciences is dedicated to the memory of
Paul Lorczac
who has been so helpful in confirming the accuracy of text, examples, and exercises.

This page is intentionally left blank

ABOUT THE AUTHORS



Marc Triola, MD, FACP, is the Associate Dean for Educational Informatics at NYU School of Medicine, the founding director of the NYU Langone Medical Center Institute for Innovations in Medical Education (IIME), and an Associate Professor of Medicine. Dr. Triola's research focuses on precision education and the use of AI tools to efficiently personalize medical education for individual learners and give new insights to their programs and coaches. His lab develops new learning technologies and AI-driven educational interventions and also works to define educationally sensitive patient and system outcomes that can be used to assess the impact of training. Dr. Triola and IIME have been funded by the National Institutes of Health, the Josiah Macy Jr. Foundation, the Department of Education, the Department of Defense, and the American Medical Association's Accelerating Change in Medical Education program.

Mario F. Triola is a Professor Emeritus of Mathematics at Dutchess Community College, where he has taught statistics for over 30 years. Marty is the author of *Elementary Statistics,* 14th edition; *Essentials of Statistics*, 7th edition; *Elementary Statistics Using Excel*, 7th edition; and *Elementary Statistics Using the TI-83/84 Plus Calculator*, 5th edition; and he is a co-author of *Statistical Reasoning for Everyday Life*, 5th edition. *Elementary Statistics* is currently available as an International Edition, and it has been translated into several foreign languages. Marty designed the original Statdisk statistical software, and he has written several manuals and workbooks for technology supporting statistics education. He has been a speaker at many conferences and colleges. Marty's consulting work includes the design of

casino slot machines and the design of fishing rods, and he has worked with attorneys in determining probabilities in paternity lawsuits, analyzing data in medical malpractice lawsuits, identifying salary inequities based on gender, and analyzing disputed election results. He has also used statistical methods to analyze medical school surveys, survey results for the New York City Transit Authority, and COVID-19 virus data for government officials. Marty has testified as an expert witness in the New York State Supreme Court. As of this writing, Marty's *Elementary Statistics* has been the #1 statistics text in the United States for 27 consecutive years.

Jason Roy, PhD, is a Professor of Biostatistics and Chair of the Department of Biostatistics and Epidemiology at Rutgers University. He is director of Rutgers University Biostatics and Epidemiology Services and co-director of Biostatistics, Epidemiology, and Research Design core, NJ ACTS. Previously, he was Professor of Biostatistics in the Department of Biostatistics, Epidemiology, and Informatics at the University of Pennsylvania. He received his PhD in Biostatistics in 2000 from the University of Michigan. He was the recipient of the 2002 David P. Byar Young Investigator Award from the American Statistical Association Biometrics Section. Dr. Roy is interested in methodological research in developing flexible Bayesian methods for large, observational studies, especially data from EHR and mobile health. He is particularly interested in causal inference problems, where Bayesian nonparametric methods can be used in conjunction with g-computation. He is also interested in functional clustering methods, which can be very useful for extracting features from intensively collected data (such as from mobile devices). Much of his collaborative research is in pharmacoepidemiology.

CONTENTS

PREFACE

The ancient Chinese philosopher Lao Tzu famously wrote: *A journey of a thousand miles must begin with a single step*. This text will lead you, step by step, on a journey through the important concepts of biostatistics. If you're reading this, you've already taken the first step! Thankfully, our journey will be much less physically taxing than "a journey of a thousand miles" and will require only the use of your feet for determining skewness (see page 75).

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

Goals of This Third Edition

- Incorporate the latest and best methods used by professional statisticians
- Include features that address all of the recommendations included in the *Guidelines for Assessment and Instruction in Statistics Education (GAISE)* as recommended by the American Statistical Association
- Provide an abundance of new and interesting data sets, examples, and exercises, such as those involving clinical trials, COVID-19, biometrics, and anthropometrics.
- Foster personal growth of students through critical thinking, use of technology, collaborative work, and development of communication skills
- Enhance teaching and learning with the most extensive and best set of supplements and digital resources

Audience/Prerequisites

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

Hallmark Features

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

Real Data

Hundreds of hours have been devoted to finding data that are real, meaningful, and interesting to students. Fully 92% of the examples are based on real data, and 92% of the exercises are based on real data. Some exercises refer to the 28 data sets provided in Appendix B, and 10 of those data sets are new to this edition. Exercises requiring

use of the Appendix B data sets are located toward the end of each exercise set and are marked with a special data set icon  .

Real data sets are included throughout the text to provide relevant and interesting real-world statistical applications, including COVID-19 clinical trials and tracking, biometric security, body measurements, brain sizes and IQ scores, and data on births. Appendix B includes descriptions of the 28 data sets that can be downloaded from the companion website www.pearsonglobaleditions.com or from Pearson MyLab™ Statistics. The data sets are also included in the free Statdisk software, which is available at www.statdisk.com. Please be advised that *Biostatistics for the Biological and Health Sciences,* 3rd edition, includes examples from the COVID-19 pandemic. Readers may find this a sensitive topic.

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

Readability

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

Website

This text is supported by www.pearsonglobaleditions.com, which includes:

- Statdisk, a free and robust browser-based statistical program designed specifically for this text. This is the only biostatistics text with dedicated and comprehensive statistics software.
- Downloadable Appendix B data sets in a variety of technology formats.
- Downloadable text supplements including Section 1-4 *Ethics in Statistics*, Section 6-6 *Normal as Approximation to Binomial*, the *Glossary of Statistical Terms*, and *Formulas and Tables*.
- Interactive flowcharts for key statistical procedures.
- Online instructional videos that provide step-by-step technology instructions.

Chapter Features

Chapter Opening Features

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

Exercises

Many exercises require the *interpretation* of results. Great care has been taken to ensure their usefulness, relevance, and accuracy. Exercises are arranged in order of increasing difficulty, and exercises are also divided into two groups: (1) *Basic Skills and Concepts* and (2) *Beyond the Basics*. *Beyond the Basics* exercises address more difficult concepts or require a stronger mathematical background. In a few cases, these exercises introduce a new concept.

End-of-Chapter Features

- ***Chapter Quick Quiz*** provides 10 review questions that require brief answers.
- ***Review Exercises*** offer practice on the chapter concepts and procedures.

- ***Cumulative Review Exercises*** reinforce earlier material.
- ***Technology Project*** provides an activity that can be used with a variety of technologies.
- ***Big (or Very Large) Data Projects*** encourage the use of large data sets.
- ***From Data to Decision*** is a capstone problem that requires critical thinking and writing.
- ***Cooperative Group Activities*** encourage active learning in groups.

Other Features

Margin Essays There are 83 margin essays designed to highlight real-world topics and foster student interest. 40 of them are new to this edition. There are also *Go Figure* items that briefly describe interesting numbers or statistics.

Flowcharts The text includes flowcharts that simplify and clarify more complex concepts and procedures. Animated versions of the flowcharts are available within MyLab Statistics.

Formulas and Tables This summary of key formulas, organized by chapter, gives students a quick reference for studying or can be printed for use when taking tests (if allowed by the instructor). It also includes the most commonly used tables. This is available for download in MyLab Statistics or www.pearsonglobaleditions.com.

Technology Integration

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

The Statdisk statistical software package is designed specifically for this text and contains all Appendix B data sets. Statdisk is free to users of this text and can be accessed at www.statdisk.com.

Changes to This 3rd Edition

New Features

New Content: This 3rd edition includes an abundance of new exercises and new examples, as summarized in the following table:

	Number	New to This Edition	Use Real Data
Exercises	1764	54%	92%
Examples	220	58%	92%

New Data Sets: This text includes a rich data set library in Appendix B so that professors and students have ready access to real and interesting data. Appendix B has been expanded from 18 data sets to 28 data sets. Ten of those data sets are new.

Larger Data Sets: The largest data set in the previous edition had 600 cases. The data set library in this 3rd edition includes data sets with 465,506, 70,942, 22,385, 6068, 5755, and 3982 cases. Working with such large data sets is essential in this age of big data and data science.

New Types of Exercises: To foster the development of critical thinking, the Cumulative Review Exercises near the end of Chapters 9, 10, and 11 include open-ended questions in which students are presented with a data set, and then asked to pose a key question relevant to the data, identify a procedure for addressing that question, and analyze the data to form a conclusion.

Big (or Very Large) Data Projects: New to this 3rd edition, these projects are located near the end of each chapter and ask students to think critically while using large data sets.

New Chapter Problem Icon: Examples that relate to the Chapter Problem are now highlighted with this icon (CP) to show how different statistical concepts and procedures can be applied to the real-world issue highlighted in the chapter.

Organization Changes

New Technology: The previous edition of *Biostatistics for the Biological and Health Sciences* introduced the resampling method of bootstrapping in Section 7-4. This 3rd edition of *Biostatistics for the Biological and Health Sciences* includes these methods of resampling using bootstrapping and randomization:

Bootstrap One Proportion
Bootstrap Two Proportions
Bootstrap One Mean
Bootstrap Two Means
Bootstrap Matched Pairs

Randomization One Proportion
Randomization Two Proportions
Randomization One Mean
Randomization Two Means
Randomization Matched Pairs
Randomization Correlation

New Methods: Resampling methods are new to Sections 8-2, 8-3, 8-4, 8-5, 9-5, and 10-1.

New Subsection 1-3, Part 3: *Clinical Trials*

New Content in Section 4-4: *Efficacy Versus Efficiency*

New Section 8-5: *Resampling: Using Technology for Hypothesis Testing*

New Section 9-5: *Resampling: Using Technology for Inferences*

New Subsection 10-1, Part 3: *Randomization Test (for Correlation)*

New Section: *Ethics in Statistics* is available for download (MyLab Statistics or www.pearsonglobaleditions.com).

Removed Section: The content of Section 6-6 *(Normal as Approximation to Binomial)* has been removed from the text and is now available for download (MyLab Statistics or www.pearsonglobaleditions.com).

Technology Changes

New to Statdisk: The previous version of Statdisk for *Biostatistics for the Biological and Health Sciences* included bootstrap resampling, but the new version of Statdisk for the 3rd edition also includes all of the bootstrapping and randomization methods listed above under "*New Technology*."

Statdisk Online: Statdisk is now a browser-based program that can be used on any device with a modern web browser, including laptops (Windows, macOS), Chromebooks, tablets, and smartphones. Statdisk Online includes all of the statistical functions from earlier versions of Statdisk and is continually adding new functions and features.

New Technology: Where it is appropriate, online technology instructions now include R as an additional technology.

MyLab Statistics Resources for Success

MyLab Statistics is available to accompany Pearson's market-leading text options, including ***Biostatistics for the Biological and Health Sciences,* 3rd edition**.

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®, an web-based statistical software program, students learn the skills they need to interact with data in the real world.

MyLab Statistics supports all learners, regardless of their ability and background, to provide an equal opportunity for success. Accessible resources support learners for a more equitable experience no matter their abilities. And options to personalize learning and address individual gaps helps to provide each learner with the specific resources they need to achieve success.

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® Integrated directly into MyLab Statistics, **StatCrunch** is a powerful web-based statistical software that allows users to perform complex analyses, share data sets, and generate compelling reports of their data. The vibrant online community offers tens of thousands shared data sets for students to analyze.

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

StatCrunch can be accessed on your laptop, smartphone, or tablet when you visit the StatCrunch website from your device's browser. For more information, visit the StatCrunch website, or contact your Pearson representative.

Exercises with Immediate Feedback – The exercises in MyLab Statistics reflect the approach and learning style of this text, and regenerate algorithmically to give student unlimited opportunity for practice and mastery. Most exercises include learning aids, such as guided solutions, sample problems, and they offer helpful feedback when students enter incorrect answers.

Integrated Review – Integrated Review MyLab courses provide the full suite of supporting resources for the Biostatistics course, plus additional assignments and study aids from select intermediate algebra topics for students who will benefit from remediation. Assignments for the integrated review content are preassigned in MyLab, making it easier than ever to create your course.

Biostatistics at Work – In these videos, Marty Triola interviews professionals about the use of statistics in practice, providing students with useful context for the concepts they're learning.

R Guidebook – This Guidebook provides students with getting started instructions, step-by-step walkthroughs, and support for using R to perform data analysis with the examples in the text.

Personalized Homework – With Personalized Homework, students take a quiz or test and receive a subsequent homework assignment that is personalized based on their performance. This way, students can focus on just the topics they have not yet mastered.

Mindset videos and assignable, open-ended **exercises** foster a growth mindset in students. This material encourages them to maintain a positive attitude about learning, value their own ability to grow, and view mistakes as learning opportunities—so often a hurdle for math students.

Personal Inventory Assessments are a collection of online exercises designed to promote self-reflection and metacognition in students. These 33 assessments include topics such as a Stress Management Assessment, Diagnosing Poor Performance and Enhancing Motivation, and Time Management Assessment.

StatTalk Videos Hosted by fun-loving statistician Andrew Vickers, the StatTalk video series demonstrates important statistical concepts through interesting stories and real-life events. Videos include assessment questions and an instructor's guide.

StatCrunch Projects provide opportunities for students to analyze big data, practice statistical thinking, and make data-informed decisions. Each project consists of a series of questions about a large data set to provide a deeper dive on business statistics topics.

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.

MyLab Features

Performance Analytics enable instructors to see and analyze student performance across multiple courses. Based on their current course progress, a student's performance is identified above, at, or below expectations through a variety of graphs and visualizations.

Conceptual Question Library There are 1,000 questions in the Assignment Manager that require students to apply their statistical understanding.

PowerPoint Presentations include lecture content and key graphics from the textbook. Accessible PowerPoint slides are also available and are built to align with WCAG 2.0 AA standards and Section 508 guidelines.

TestGen® enables instructors to build, edit, print, and admin-ister tests using a computerized bank of questions developed to cover the objectives of the text.

Test Bank features printable PDF containing all the test exercises available in TestGen.

Accessibility– Pearson works continuously to ensure our products are as accessible as possible to all students. Currently we work toward achieving WCAG 2.0 AA for our existing products (2.1 AA for future products) and Section 508 standards, as expressed in the Pearson Guidelines for Accessible Educational Web Media https://www.pearson.com/uk/accessibility.html.

Flexible Syllabus

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

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

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

1. ***Emphasize statistical literacy and develop statistical thinking:*** Each section exercise set begins with *Statistical Literacy and Critical Thinking* exercises. Many of the text's exercises are designed to encourage statistical thinking rather than the blind use of mechanical procedures.
2. ***Use real data:*** 92% of the examples and 92% of the exercises use real data.
3. ***Stress conceptual understanding rather than mere knowledge of procedures:*** Instead of seeking simple numerical answers, most exercises and examples involve conceptual understanding through questions that encourage practical interpretations of results. Also, each chapter includes a *From Data to Decision* project.
4. ***Foster active learning in the classroom:*** Each chapter ends with several *Cooperative Group Activities*.
5. ***Use technology for developing conceptual understanding and analyzing data:*** Computer software displays are included throughout the text. Special online *Tech Center* subsections include instruction for using the software. Each chapter includes a *Technology Project*. When there are discrepancies between answers based on tables and answers based on technology, Appendix D provides both answers. The website www.pearsonglobaleditions.com includes free text-specific software (Statdisk), data sets formatted for several different technologies, and instructional videos for technologies. MyLab Statistics also includes support videos for different statistical software applications.
6. ***Use assessments to improve and evaluate student learning:*** Assessment tools include an abundance of section exercises, *Chapter Quick Quizzes, Review Exercises, Cumulative Review Exercises, Technology Projects, Big (or Very Large) Data Projects, From Data to Decision* projects, and *Cooperative Group Activities.*

Acknowledgments

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

Marta Piva, Alcorn State University
Samantha Tran, Arizona State University
Susan Whitehead, Becker College
Bryan James, Pennsylvania College of Technology
Jessica Gray, University of North Carolina at Wilmington
Joan Brenneman, University of Central Oklahoma
Rachel Carroll, University of North Carolina at Wilmington
Keith Blount, University of Arkansas at Monticello
Delray Schultz, Milllersville University
Susan Serrano, Florida Southern College
Dan Jelsovsky, Florida Southern College
Mingan Yang, San Diego State University
Mahbobeh Vezvaei, Kent State University
Ryan Pohlig, University of Delaware
Anushka Karkelanova, University of Kentucky
Andrew Hooyman, University of Nevada at Las Vegas
Darrin Rasberry, Mercy College

We would also thank Dirk Tempelaar and the late Paul Lorczak for their help in checking the accuracy of the text and answers.

This third edition is truly a team effort, and we are thankful for the dedication and commitment of the Pearson team. We thank Laura Briskman, Karen Montgomery, Peggy McMahon, Demetrius Hall, Robert Carroll, Joe Vetere, Dawn Murrin, Deirdre Lynch, and Rose Kernan of RPK Editorial Services. We also thank Pearson's Diverisity, Equity, and Inclusion team for their help and recommendations for making this text bias-free, inclusive, and accessible.

Marc Triola
Mario Triola
Jason Roy
September 2022

Acknowledgments for the Global Edition

Pearson would like to thank the following for their contribution to the Global Edition:

Contributors for the Third Edition
Hatice Yağmur Zengin, Hacettepe University
Raghib Abu-Saris, King Saud bin Abdulaziz University for Health Sciences

Reviewers for the Third Edition
Amit Kumar Misra, Babasaheb Bhimrao Ambedkar University
Ümit Işlak, Boğaziçi University

Contributor for the Second Edition
Vikas Arora

Reviewers for the Second Edition
Hemant Kumar, Vardhman Mahavir Medical College
Santhosh Kumar, Christ University
Kiran Paul

Pearson's Commitment to Diversity, Equity, and Inclusion

Pearson is dedicated to creating bias-free content that reflects the diversity, depth, and breadth of all learners' lived experiences.

We embrace the many dimensions of diversity, including but not limited to race, ethnicity, gender, sex, sexual orientation, socioeconomic status, ability, age, and religious or political beliefs.

Education is a powerful force for equity and change in our world. It has the potential to deliver opportunities that improve lives and enable economic mobility. As we work with authors to create content for every product and service, we acknowledge our responsibility to demonstrate inclusivity and incorporate diverse scholarship so that everyone can achieve their potential through learning. As the world's leading learning company, we have a duty to help drive change and live up to our purpose to help more people create a better life for themselves and to create a better world.

Our ambition is to purposefully contribute to a world where:

- Everyone has an equitable and lifelong opportunity to succeed through learning.
- Our educational content accurately reflects the histories and lived experiences of the learners we serve.
- Our educational products and services are inclusive and represent the rich diversity of learners.
- Our educational content prompts deeper discussions with students and motivates them to expand their own learning (and worldview).

Accessibility

We are also committed to providing products that are fully accessible to all learners. As per Pearson's guidelines for accessible educational Web media, we test and retest the capabilities of our products against the highest standards for every release, following the WCAG guidelines in developing new products for copyright year 2022 and beyond.

You can learn more about Pearson's commitment to accessibility at **https://www.pearson.com/uk/accessibility.html**

Contact Us

While we work hard to present unbiased, fully accessible content, we want to hear from you about any concerns or needs with this Pearson product so that we can investigate and address them.

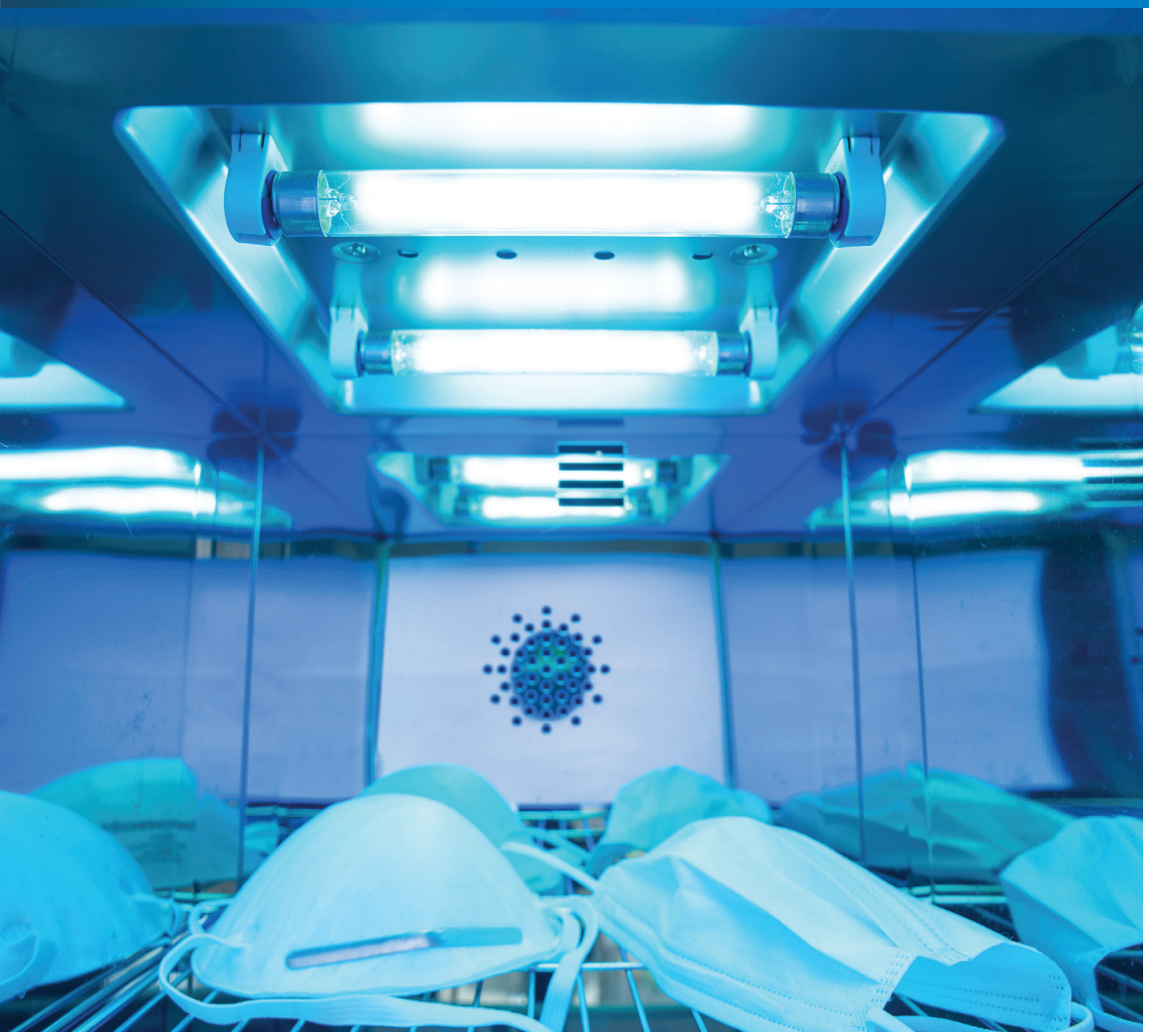
Please contact us with concerns about any potential bias at **https://www.pearson.com/report-bias.html**.

For accessibility-related issues, such as using assistive technology with Pearson products, alternative text requests, or accessibility documentation, email the Pearson Disability Support team at **disability.support@pearson.com.**

Pearson

This page is intentionally left blank

1 Introduction to Statistics



Credit: Nor Gal/Shutterstock

CHAPTER PROBLEM

Which Face Masks Are Better: N95 Respirators or Medical Masks?

Because of mask-wearing requirements during the COVID-19 pandemic, there was much interest in the effectiveness of masks. A four-year experiment was conducted with the objective of comparing the effectiveness of two different types of face masks: N95 respirators and medical masks. Results were based on the rates of influenza among subjects who used the two different types of face masks. Figure 1-1 on the next page is a graph that depicts the percentage of influenza cases in each of the two groups (based on data from "N95 Respirators vs Medical Masks for Preventing Influenza Among Health Care Personnel," by Radonovich et al., *Journal of the American Medical Association*, Vol. 322, No. 9).

Critical Thinking Figure 1-1 makes it appear that the N95 masks are *less effective* because we see about twice as many cases of influenza as with medical masks. But wait! Let's not rush to judgment. Look *carefully* at Figure 1-1 and see that it is *misleading*. Instead of using a scale that begins with 0%,

Figure 1-1 uses a vertical scale that ranges from 6.5% to 8.5%, with the result that the difference between the heights of the two bars is visually *exaggerated*. Figure 1-1 and Figure 1-2 show the same data, but Figure 1-2 shows that the rate of influenza with the N95 masks is actually only slightly higher than the rate with the medical masks. In fact, the experiment concluded there was no *significant* difference in the rate of influenza between the N95 mask and medical mask groups. Figure 1-1 is misleading, whereas Figure 1-2 depicts the same data fairly.

The flaw in Figure 1-1 is among the most commonly used tricks to present misleading arguments, so it is especially important to recognize. Here are brief descriptions of common flaws:

Flaw 1: Misleading Graphs The bar chart in Figure 1-1 is very deceptive. By using a vertical scale that does not start at 0%, the difference between the two percentages is grossly exaggerated. Deceptive graphs are discussed in more detail in Section 2-3.

Flaw 2: Bad Sampling Method Figure 1-1 and Figure 1-2 are based on a clinical trial that was carefully planned and executed. The sampling method appears to be sound based on its description given in the published journal article. However, many other surveys obtain subjects by using methods that are inappropriate and may lead to bad results, such as these:

- **Voluntary Response Sample** In a voluntary response sample, participants decide themselves whether to participate. *Example:* A survey question is posted on a website, and then Internet users decide whether to respond. With a voluntary response sample, it often happens that those who have a strong interest in the topic are more likely to participate, so the results are very questionable.
- **Convenience Sample** With a convenience sample, participants are selected because they are easy to reach and are readily available. *Example:* A cardiac surgeon conducts a study that includes only her patients.

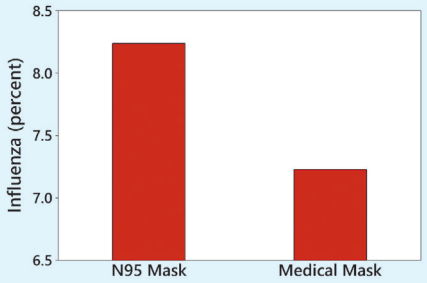


FIGURE 1-1 Effectiveness of N95 Masks and Medical Masks

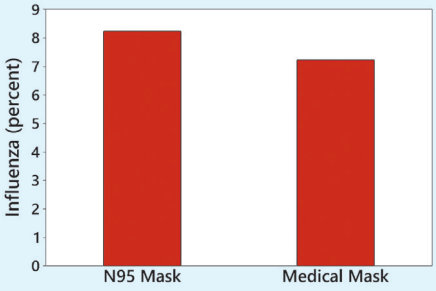


FIGURE 1-2 Effectiveness of N95 Masks and Medical Masks

When using sample data to learn something about a population, it is *extremely* important to obtain sample data that are representative of the population from which the data are drawn. As we proceed through this chapter and discuss types of data and sampling methods, we should focus on these key concepts:

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

It is all too easy to analyze sample data without thinking critically about how the data were collected. We could then develop conclusions that are fundamentally wrong and misleading. Instead, we should develop skills in statistical thinking and critical thinking so that we can distinguish between collections of sample data that are good and those that are seriously flawed.

CHAPTER OBJECTIVES

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

1-1 Statistical and Critical Thinking

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

1-2 Types of Data

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

1-3 Collecting Sample Data

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

1-4 Ethics in Statistics (available at www.pearsonglobaleditions.com)

- Analyze ethical issues in statistics, including those related to data collection, analysis, and reporting.

1-1 Statistical and Critical Thinking

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

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

We begin with some very basic definitions.

Census Results Affect Hospitals

Credit: Juan Camilo Bernal/Shutterstock

The United States Constitution requires a census every ten years. One of the factors affected by census results is the distribution of federal funds to hospitals.

Although accuracy of census results is extremely important, it is becoming more difficult to collect accurate census data due to the growing diversity of cultures and languages and increased distrust of the government. No amount of statistical analysis can salvage poor data, so it is critical that the census data are collected in an appropriate manner.

DEFINITIONS

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

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

A **population** is the complete collection of *all* measurements or data that are being considered.

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

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

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

CP **EXAMPLE 1** **N95 Masks Versus Medical Masks**

The Chapter Problem included graphs based on these results from the study cited: Among a sample of 2512 subjects wearing N95 masks, 207 presented with influenza. In this case, the population and sample are as follows:

Population: All users of N95 masks

Sample: The 2512 subjects included in the study

The objective is to use the sample as a basis for drawing a conclusion about the population of all users of N95 masks, and methods of statistics are helpful in drawing such conclusions.

We now proceed to consider the process involved in a statistical study. See Figure 1-3 for a summary of this process and note that the focus is on critical thinking, not mathematical calculations. Thanks to wonderful developments in technology, we have powerful tools that effectively do the number crunching so that we can focus on understanding and interpreting results.

Prepare

Context Figure 1-3 suggests that we begin our preparation by considering the *context* of the data, so let's start with context by considering the data in Table 1-1. Table 1-1 shows the systolic and diastolic blood pressure measurements (mmHg) of eight females. The format of Table 1-1 suggests the goal of determining whether there is a *relationship* between those two variables.

TABLE 1-1 Systolic and Diastolic Blood Pressure Measurements of Females

Systolic	100	134	138	114	110	100	92	100
Diastolic	70	94	80	66	72	80	58	50

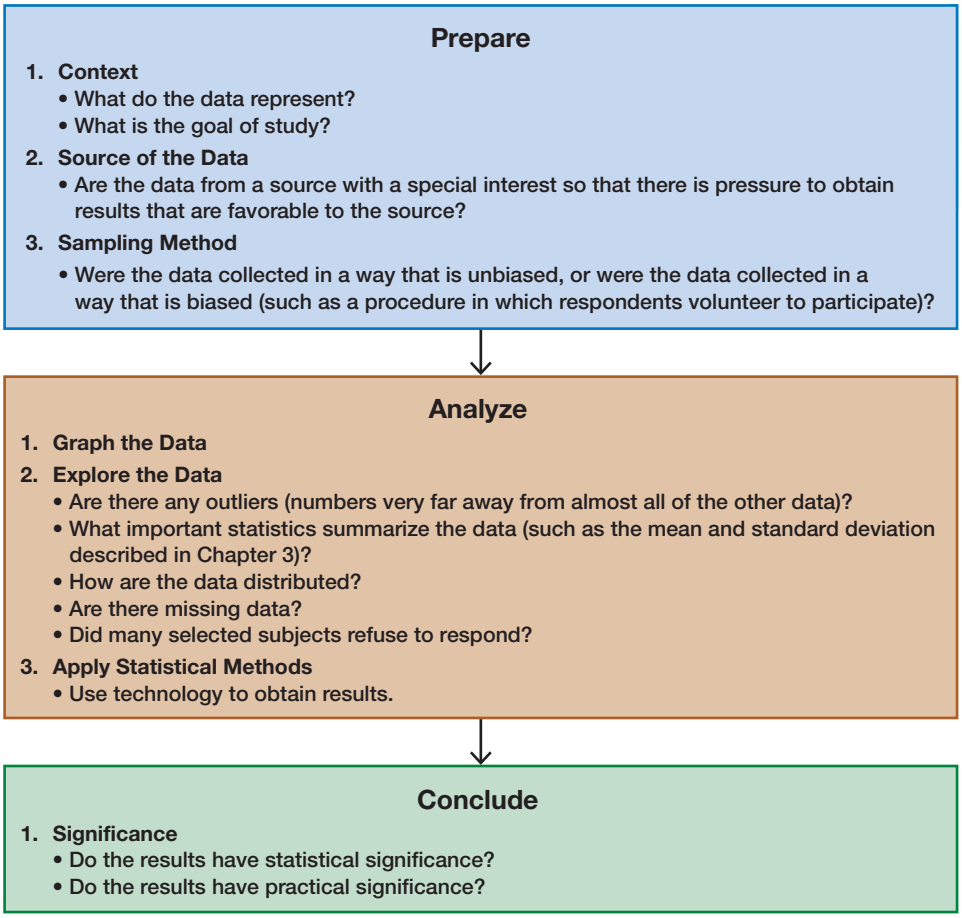


FIGURE 1-3 Statistical and Critical Thinking

Source of the Data The second step is to consider the source (as indicated in Figure 1-3). The data in Table 1-1 are from Data Set 1 "Body Data" in Appendix B, where the source is identified. The data are from the National Center for Health Statistics, which is a source that certainly appears to be reputable and unbiased.

Sampling Method Figure 1-3 suggests that we conclude our preparation by considering the sampling method. The specific sampling method used is fairly complicated, but it is designed to yield a representative sample of the population. Close analysis of the sampling method reveals that it is sound and unbiased.

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

DEFINITION

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

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

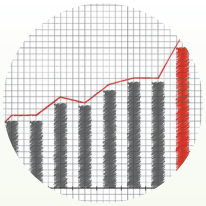
continued

Survivorship Bias

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

Credit: Gary Blakeley/ Shutterstock

Origin of "Statistics"



Credit: USBFCO/ Shutterstock

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

- Internet polls, in which people online decide whether to respond
- Mail-in polls, in which people decide whether to reply
- Telephone call-in polls, in which newspaper, radio, or television announcements ask that you voluntarily call a special number to register your opinion

EXAMPLE 2 **Voluntary Response Sample**

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

YOUR TURN. Do Exercise 1 "Online Medical Info."

Analyze

Figure 1-3 indicates that after completing our preparation by considering the context, source, and sampling method, we begin to *analyze* the data.

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

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

Conclude

Figure 1-3 shows that the final step in our statistical process involves conclusions, and we should develop an ability to distinguish between *statistical significance* and *practical significance*.

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

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

CAUTION An outcome can be statistically significant, and it may or may not be *important*. Don't associate statistical significance with importance.

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

EXAMPLE 3 Statistical Significance Versus Practical Significance

In a trial of weight loss programs, 21 subjects on the Atkins program lost an average (mean) of 2.1 kg (or 4.6 lb) after one year (based on data from "Comparison of the Atkins, Ornish, Weight Watchers, and Zone Diets for Weight Loss and Heart Disease Risk Reduction," by Dansinger et al., *Journal of the American Medical Association*, Vol. 293, No. 1). The results show that this loss is *statistically significant* and is not likely to occur by chance. However, many dieters believe that after following this diet for a year, a loss of only 2.1 kg is not worth the time, cost, and effort so that for these people, this diet does not have *practical significance*.

YOUR TURN. Do Exercise 13 "Diet and Exercise Program."

Example 3 includes a small sample of only 21 subjects, but with very large data sets (e.g., "big data"), statistically significant differences can often be found with very small differences. We should be careful to avoid the mistake of thinking that those small differences have practical significance.

Analyzing Data: Potential Pitfalls

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

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

Correlation does not imply causation.

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

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

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

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

Order of Questions Sometimes survey questions are unintentionally loaded by such factors as the order of the items being considered. See the following two questions from a poll conducted in Germany, along with the very different response rates:

continued

Publication Bias

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

Credit: Wavebreakmedia/Shutterstock

Credit: From "Registering Clinical Trials" by Kay Dickersin and Drummond Rennie in *Journal of American Medical Association,* Vol 290, No: 4, pp: 516-523. Published by Journal of American Medical Association, © 2003.

Statistician Jobs

Credit: Stephen VanHorn/Shutterstock

In a recent year, *U.S. News & World Report* provided a list of the top 10 best jobs. Here are the first two jobs at the top of the list: (1) Software developer; (2) Statistician. It was noted that one reason for this high ranking is that the unemployment rate for statisticians is only 0.9%. That unemployment rate is lower than 1 person in 100. Not to mention how cool the contemporary discipline of statistics has become!

Go Figure

Five out of four people have some difficulty with statistics.

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

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

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

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

> **People who refuse to talk to pollsters are likely to be different from those who do not. Some may be fearful of strangers and others jealous of their privacy, but their refusal to talk demonstrates that their view of the world around them is markedly different from that of those people who will let poll-takers into their homes.**
>
> Credit: From *Lies, Damn Lies, And Statistics: The Manipulation Of Public Opinion In America.* Published by Dell Publishing, © 1976

Low Response Rates Related to the preceding item of nonresponses is the issue of low response rates. If a survey has a low response rate, the reliability of the results decreases. In addition to having a smaller sample size, there is an increased likelihood of having a bias among those who do respond. Some steps to help prevent a low response rate: (1) A survey should present an engaging argument for its importance; (2) a survey should not be very time consuming; (3) it is helpful to provide a reward for completing a survey, such as cash or a chance to win a prize. There are not definitive guidelines for acceptable response rates. A very good response rate is 80% or higher. Some suggest that response rates of at least 40% are acceptable. Pew Research Center reports that its typical telephone surveys have a response rate around 9%, but their surveys tend to be quite good. Sections 7-1, 7-2, and 7-3 include procedures for determining the sample size needed to estimate characteristics (proportion, mean, standard deviation) of a population, and those methods require sound sampling methods.

Straightlining Straightlining occurs when survey respondents give the same answers (such as choice "b") to all or almost all of the questions. Survey responses that result from straightlining should be deleted because it is obvious that they do not truly reflect the opinions of the respondents.

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

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

> **Percentage of:** To find the percentage of an amount, replace the % symbol with division by 100, and then interpret "of" to be multiplication. For example:

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

1-1 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

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

2. Reported Versus Measured In a survey of 10,000 adults conducted by GlaxoSmithKline, subjects were asked if they brushed or flossed their teeth at night, and 45% of the respondents said "no."

a. Identify the sample and the population.

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

3. Statistical Significance Versus Practical Significance When testing a new treatment, what is the difference between statistical significance and practical significance? Can a treatment have statistical significance, but not practical significance?

4. Correlation One study showed that for a recent period of 10 years, there was a strong correlation (or association) between the per capita consumption of margarine and the divorce rate in Maine (based on data from National Vital Statistics reports and the U.S. Department of Agriculture). Does this imply that increasing margarine consumption is the cause of an increase in the divorce rate in Maine? Why or why not?

Consider the Source. ***In Exercises 5–8, determine whether the given source has the potential to create a bias in a statistical study.***

5. Physicians Committee for Responsible Medicine The Physicians Committee for Responsible Medicine tends to oppose the use of meat and dairy products in our diets, and that organization has received hundreds of thousands of dollars in funding from the Foundation to Support Animal Protection.

6. Nicotine in Cigarettes Amounts of nicotine in samples of Camel cigarettes produced by R.J. Reynolds Tobacco Company were measured by William Esty Co., an advertising agency working for the tobacco company.

7. Brain Size A data set in Appendix B includes brain volumes from 10 pairs of monozygotic (identical) twins. The data were collected by researchers at Harvard University, Massachusetts General Hospital, Dartmouth College, and the University of California at Davis.

8. Chocolate An article in *Journal of Nutrition* (Vol. 130, No. 8) noted that chocolate is rich in flavonoids. The article notes "regular consumption of foods rich in flavonoids may reduce the risk of coronary heart disease." The study received funding from Mars, Inc., the candy company, and the Chocolate Manufacturers Association.

Credit: Based on Teresa L. Dillinger et al., 2000, "Food of the Gods: Cure for Humanity? A Cultural History of the Medicinal and Ritual Use of Chocolate," *Journal of Nutrition,* Vol. 130, No. 8

Sampling Method. ***In Exercises 9–12, determine whether the sampling method appears to be sound or is flawed.***

9. Nuclear Power Plants In a survey of 1368 subjects, the following question was posted on the *USA Today* website: "In your view, are nuclear plants safe?" The survey subjects were Internet users who chose to respond to the question posted on the electronic edition of *USA Today*.

10. Clinical Trials Researchers at Yale University conduct a wide variety of clinical trials by using subjects who volunteer after reading advertisements soliciting paid volunteers.

11. NHANES Examinations In a recent year, the National Health and Nutrition Examination Survey (NHANES), sponsored by the National Center for Health Statistics, selected more than 9000 subjects who were given physical exams. The subjects were selected through a somewhat complicated procedure designed to obtain results that are representative of the population.

12. Health In a survey of 3014 randomly selected U.S. adults, 45% reported that they have at least one chronic health condition, such as diabetes or high blood pressure. The survey was conducted by Princeton Survey Research Associates International.

Statistical Significance and Practical Significance. ***In Exercises 13–20, determine whether the results appear to have statistical significance, and also determine whether the results appear to have practical significance.***

13. Diet and Exercise Program In a study of the Ornish weight loss program, 40 subjects lost a mean of 3.3 lb after 12 months (based on data from "Comparison of the Atkins, Ornish, Weight Watchers, and Zone Diets for Weight Loss and Heart Disease Risk Reduction," by Dansinger et al., *Journal of the American Medical Association*, Vol. 293, No. 1). Methods of statistics can be used to show that if this diet had no effect, the likelihood of getting these results is roughly 3 chances in 1000.

14. Surgery Versus Splints A study compared surgery and splinting for subjects suffering from carpal tunnel syndrome. It was found that among 73 patients treated with surgery, there was a 92% success rate. Among 83 patients treated with splints, there was a 72% success rate. Calculations using those results showed that if there really is no difference in success rates between surgery and splints, then there is about one chance in a thousand of getting success rates like the ones obtained in this study.

15. Mendel's Genetics Experiments One of Gregor Mendel's famous hybridization experiments with peas yielded 580 offspring with 152 of those peas (or 26%) having yellow pods. According to Mendel's theory, 25% of the offspring peas should have yellow pods.

16. IQ Scores Most people have IQ scores between 70 and 130. For $39.99, you can purchase a PC or Mac program from HighIQPro that is claimed to increase your IQ score by 10 to 20 points. The program claims to be "the only proven IQ increasing software in the brain training market," but the authors of your text could find no substantial data supporting that claim, so let's suppose that these results were obtained: In a study of 12 subjects using the program, the average increase in IQ score is 3 IQ points. There is a 25% chance of getting such results if the program has no effect.

17. Births A random sample of 860 births in New York State included 426 males.

18. Systolic Blood Pressure High systolic blood pressure is 140 mmHg or higher. (Normal values are lower than 120 mmHg, and prehypertension levels are between 120 and 139 mmHg.) Subjects with high blood pressure are encouraged to take action to lower it. A pharmaceutical company develops a new medication designed to lower blood pressure, and tests on 25 subjects result in an average (mean) decrease of 2 mmHg. Analysis of the results shows that there is a 15% chance of getting such results if the medication has no effect.

19. Clinical Trials of OxyContin OxyContin (oxycodone) is a drug used to treat pain, but it is well known for being dangerous and leading to addiction. In a clinical trial, among 227 subjects treated with OxyContin, 52 developed nausea (based on data from Purdue Pharma L.P.). Among 45 other subjects given placebos, 5 developed nausea. Calculations show that if there really is no difference between the rates of nausea for the OxyContin treatment group and the placebo group, then there is an 8% chance of getting such results.

20. Cell Phones and Handedness A study was conducted to investigate the association between cell phone use and hemispheric brain dominance. Among 216 subjects who prefer to use their left ear for cell phones, 166 were right-handed. Among 452 subjects who prefer to use their right ear for cell phones, 436 were right-handed (based on data from "Hemispheric Dominance and Cell Phone Use," by Seidman et al., *JAMA Otolaryngology – Head & Neck Surgery*, Vol. 139, No. 5). Calculations show that if there really is no difference between those two rates, then there is less than a 1% chance of getting the results obtained in this study.

In Exercises 21–24, refer to the sample of body temperatures (degrees Fahrenheit) in the table below. (The body temperatures are from Data Set 5 in Appendix B.)

	Subject				
	1	2	3	4	5
8 AM	97.0	98.5	97.6	97.7	98.7
12 AM	97.6	97.8	98.0	98.4	98.4

21. Context of the Data Refer to the table of body temperatures. Is there some meaningful way in which each body temperature recorded at 8 AM is matched with the 12 AM temperature?

22. Source The listed body temperatures were obtained from Dr. Steven Wasserman, Dr. Philip Mackowiak, and Dr. Myron Levine, who were researchers at the University of Maryland. Is the source of the data likely to be biased?

23. Conclusion Given the body temperatures in the table, what issue can be addressed by conducting a statistical analysis of the data?

24. Conclusion If we analyze the listed body temperatures with suitable methods of statistics, we conclude that when the differences are found between the 8 AM body temperatures and the 12 AM body temperatures, there is a 64% chance that the differences can be explained by random results obtained from populations that have the same 8 AM and 12 AM body temperatures. What should we conclude about the statistical significance of those differences?

In Exercises 25–28, refer to the data in the table below. The entries are white blood cell counts (1000 cells/μL) and red blood cell counts (million cells/μL) from male subjects examined as part of a large health study conducted by the National Center for Health Statistics. The data are matched, so that the first subject has a white blood cell count of 8.7 and a red blood cell count of 4.91, and so on.

	Subject				
	1	2	3	4	5
White	8.7	5.9	7.3	6.2	5.9
Red	4.91	5.59	4.44	4.80	5.17

25. Context Given that the data are matched and considering the units of the data values, does it make sense to use the difference between each white blood cell count and the corresponding red blood cell count? Why or why not?

26. Analysis Given the context of the data in the table, what issue can be addressed by conducting a statistical analysis of the measurements?

27. Source of the Data Does the source of the data appear to be biased in any way?

28. Conclusion If we were to use the sample data to conclude that there is a correlation or association between white blood cell counts and red blood cell counts, does it follow that higher white blood cell counts are the cause of higher red blood cell counts?

What's Wrong? ***In Exercises 29–36, identify what is wrong.***

29. Potatoes In a poll sponsored by the Idaho Potato Commission, 1000 adults were asked to select their favorite vegetables, and the favorite choice was potatoes, which were selected by 26% of the respondents.

30. Healthy Water In a *USA Today* online poll, 951 Internet users chose to respond, and 57% of them said that they prefer drinking bottled water instead of tap water.

31. Motorcycles Deaths and Sour Cream In recent years, there has been a strong correlation between per capita consumption of sour cream and the numbers of motorcycle riders killed in noncollision accidents. Therefore, consumption of sour cream causes motorcycle fatalities.

32. Smokers The electronic cigarette maker V2 Cigs sponsored a poll showing that 55% of smokers surveyed say that they feel ostracized "sometimes," "often," or "always."

33. Cheese and Bedsheet Deaths In recent years, there has been a strong correlation between the per capita consumption of cheese in the United States and the numbers of people who died from being tangled in their bedsheets. Really. Therefore, consumption of cheese causes bedsheet entanglement fatalities.

34. Storks and Babies In the years following the end of World War II, it was found that there was a strong correlation, or association, between the number of human births and the stork population. It therefore follows that storks cause babies.

35. Age and Weight In an online survey, respondents were asked to submit their age and weight. For 87 respondents aged 8 years to 14 years, it was found that there is a correlation between age and weight.

36. Diet Research Twelve nutritionists are each paid $100,000 to try a new celebrity diet on ten of their clients and then write a report summarizing the results. Based on the sample results, it is found that the diet is effective for 118 of the 120 people.

Percentages. ***In Exercises 37–38, answer the given questions, which are related to percentages.***

37. Health In a survey of 3014 randomly selected U.S. adults, 45% reported that they have at least one chronic health condition, such as diabetes or high blood pressure (based on data from Princeton Survey Research Associates International).

a. What is 45% of 3014?

b. Could the result from part (a) be the actual number of survey subjects who have at least one chronic condition?

c. What is the actual number of survey subjects who have at least one chronic condition?

d. Among those surveyed, 1808 were called by landline and 1206 were called by cell phone. What percentage of the survey subjects were called by cell phone?

38. Smoking Cessation In a program designed to help patients stop smoking, 198 patients were given "sustained" care, and 82.8% of those were no longer smoking after one month (based on data from "Sustained Care Intervention and Postdischarge Smoking Cessation Among Hospitalized Adults," by Rigotti et al., *Journal of the American Medical Association,* Vol. 312, No. 7).

a. What is 82.8% of 198?

b. Could the result from part (a) be the actual number of patients who were no longer smoking after one month?

c. What is the actual number of patients who were no longer smoking after one month?

d. The study included 198 patients who were given sustained care and 199 other patients who were given standard care. What is the percentage of patients given sustained care?

1-1 Beyond the Basics

39. What's Wrong with This Picture? The *Newport Chronicle* ran a survey by asking readers to call in their response to this question: "Do you support the development of atomic weapons that could kill millions of innocent people?" It was reported that 20 readers responded and that 87% said "no," while 13% said "yes." Identify four major flaws in this survey.

40. Falsifying Data A researcher at the Sloan-Kettering Cancer Research Center was once criticized for falsifying data. Among his data were figures obtained from 6 groups of mice, with 20 individual mice in each group. The following values were given for the percentage of successes in each group: 53%, 58%, 63%, 46%, 48%, 67%. What's wrong with those values?

1-2 Types of Data

Key Concept Because a major use of statistics is to collect and use sample data to make conclusions about populations, we should know and understand the meanings of the terms *statistic* and *parameter*, as defined below. In this section we describe a few different types of data. The type of data is one of the key factors that determine the statistical methods we use in our analysis.

In Part 1 of this section we describe the basics of different types of data, and then in Part 2 we consider "big data" and missing data.

Go Figure

7 billion: The world population that was exceeded in early 2012, which is 13 years after it passed 6 billion.

PART 1 Basic Types of Data

Parameter / Statistic

DEFINITIONS

A **parameter** is a numerical measurement describing some characteristic of a *population*.

A **statistic** is a numerical measurement describing some characteristic of a *sample*.

HINT The alliteration in "population parameter" and "sample statistic" helps us remember the meanings of these terms.

If we have more than one statistic, we have "statistics." Another meaning of "statistics" was given in Section 1-1, where we defined *statistics* to be the science of planning studies and experiments; obtaining data; organizing, summarizing, presenting, analyzing, and interpreting those data; and then drawing conclusions based on them. We now have two different definitions of statistics, but we can determine which of these two definitions applies by considering the context in which the term *statistics* is used. The following example uses the first meaning of *statistics* as given on this page.

EXAMPLE 1 Parameter/Statistic

There are 258,664,729 adults in the United States. Data Set 4 "Measured and Reported" in Appendix B includes a sample of 5755 adults. Their measured mean height is 166.47 cm.

1. **Parameter:** The population size of 258,664,729 adults is a *parameter* because it is the entire population of all adults in the United States.
2. **Statistic:** The mean of 166.47 cm from Data Set 4 is a *statistic* because it is based on a sample, not the entire population of the United States.

YOUR TURN. Do Exercise 1 "Parameter and Statistic."

Quantitative / Categorical

Some data are numbers representing counts or measurements (such as heights of adults), whereas others are attributes (such as eye color of green or brown) that are not counts or measurements. The terms *quantitative data* and *categorical data* distinguish between these types.

DEFINITIONS

Quantitative (or **numerical) data** consist of *numbers* representing counts or measurements.

Categorical (or **qualitative** or **attribute) data** consist of names or labels (not numbers that represent counts or measurements).

Validation Question

Credit: Dmitry Naumov/Shutterstock

A question is sometimes used in a survey to confirm that a subject is attempting to seriously complete the survey questions instead of just mindlessly checking off answers. Here is an example:

This question is unlike the others. To confirm that you have read this question carefully, please select "Don't know" from the following list.

- *Definitely will*
- *Probably will*
- *Probably will not*
- *Definitely will not*
- *Don't know*

CAUTION Categorical data are sometimes coded with numbers, with those numbers replacing names. Although such numbers might appear to be quantitative, they are actually categorical data. See the third part of Example 2 that follows.

Include Units of Measurement With quantitative data, it is important to use the appropriate units of measurement, such as dollars, hours, feet, or meters. We should carefully observe information given about the units of measurement, such as "all amounts are in *thousands of dollars*" or "all units are in *kilograms*." Ignoring such units of measurement can be very costly. The National Aeronautics and Space Administration (NASA) lost its $125 million Mars Climate Orbiter when the orbiter crashed because the controlling software had acceleration data in *English* units, but they were incorrectly assumed to be in *metric* units.

Hopefully, the day will soon come when the United States adopts the metric system and joins almost all of the rest of the countries on planet Earth.

EXAMPLE 2 **Quantitative/Categorical**

1. **Quantitative Data:** The ages (in years) of subjects enrolled in a clinical trial
2. **Categorical Data as Labels:** The sex (male/female) of subjects enrolled in a clinical trial
3. **Categorical Data as Numbers:** The identification numbers 1, 2, 3, . . . , 25 are assigned randomly to the 25 subjects in a clinical trial. Those numbers are substitutes for names. They don't measure or count anything, so they are categorical data.

YOUR TURN. Do Exercise 2 "Quantitative/Categorical Data."

Discrete/Continuous

Quantitative data can be further described by distinguishing between *discrete* and *continuous* types.

DEFINITIONS

Discrete data result when the data values are quantitative and the number of values is finite, or "countable." (If there are infinitely many values, the collection of values is countable if it is possible to count them individually, such as the number of tosses of a coin before getting tails.)

Continuous (numerical) data result from infinitely many possible quantitative values, where the collection of values is not countable. (That is, it is impossible to count the individual items because at least some of them are on a continuous scale, such as the lengths of distances from 0 cm to 12 cm.)

Continuous Data

Credit: Suppakij1017/Shutterstock

Discrete Data

Credit: Khamidulin Sergey/Shutterstock

CAUTION The concept of countable data plays a key role in the preceding definitions, but it is not a particularly easy concept to understand. Continuous data can be measured, but not counted. If you select a particular data value from continuous data, there is no "next" data value. See Example 3.

EXAMPLE 3 **Discrete/Continuous**

1. **Discrete Data of the Finite Type:** Bellevue Hospital counts the number of patients admitted for emergency care each day. The numbers are discrete because they are finite numbers that result from a counting process.
2. **Discrete Data of the Infinite Type:** A psychologist tests the memory of subjects by giving them 50 random letters and asking them to memorize those letters. She counts the number of subjects who respond until one of them can memorize all 50 letters. It is possible that she could count indefinitely without ever getting a subject who can memorize all 50 letters. Because the numbers result from a counting process, the numbers are discrete.
3. **Continuous Data:** Burmese pythons are invading Florida. Researchers capture pythons and measure their lengths. So far, the largest python captured in Florida was 17 feet long. If the python lengths are between 0 feet and 17 feet, there are infinitely many values between 0 feet and 17 feet. Because it is impossible to count the number of different possible values on such a continuous scale, these lengths are continuous data.

YOUR TURN. Do Exercise 3 "Discrete/Continuous Data."

GRAMMAR: FEWER VERSUS LESS When describing smaller amounts, it is correct grammar to use "fewer" for discrete amounts and "less" for continuous amounts. It is correct to say that we drank *fewer* cans of cola and that, in the process, we drank *less* cola. The numbers of cans of cola are discrete data, whereas the volume amounts of cola are continuous data.

Levels of Measurement

Another common way of classifying data is to use four levels of measurement: nominal, ordinal, interval, and ratio, all defined below. (Also see Table 1-2 on page 37 for brief descriptions of the four levels of measurements.) When we are applying statistics to real problems, the level of measurement of the data helps us decide which procedure to use. There will be references to these levels of measurement in this text, but the important point here is based on common sense: *Don't do computations and don't use statistical methods that are not appropriate for the data.* For example, it would not make sense to compute an average (mean) of Social Security numbers, because those numbers are data that are used for identification, and they don't represent measurements or counts of anything.

DEFINITION

The **nominal level of measurement** is characterized by data that consist of names, labels, or categories only. The data cannot be arranged in some order (such as low to high).

Measuring Disobedience

Credit: DeiMosz/ Shutterstock

How are data collected about something that doesn't seem to be measurable, such as people's level of disobedience? Psychologist Stanley Milgram devised the following experiment: A researcher instructed a volunteer subject to operate a control board that gave increasingly painful "electrical shocks" to a third person. Actually, no real shocks were given, and the third person was an actor. The volunteer began with 15 volts and was instructed to increase the shocks by increments of 15 volts. The disobedience level was the point at which the subject refused to increase the voltage. Surprisingly, two-thirds of the subjects obeyed orders even when the actor screamed and faked a heart attack.

Big Data Study of Measles Vaccine and Autism

Credit: Sherry Yates Young/123RF

In 2019, there were measles outbreaks in U.S. geographic regions with large numbers of children who did not receive the MMR (measles, mumps, rubella) vaccine. Many parents opposed those vaccinations because they believed that they were associated with autism. Much of that belief was fueled by a 1998 "study" of 12 subjects showing an autism and MMR link that was reported in *The Lancet*, but that article was later retracted. Based on a new ten-year study of 657,461 children, the *Annals of Internal Medicine* reported that the "MMR vaccination does not increase the risk for autism, does not trigger autism in susceptible children, and is not associated with clustering of autism cases after vaccination." An article in *The New York Times* reported about this study and emphasized the key point with this headline: "One More Time, With Big Data: Measles Vaccine Doesn't Cause Autism." In this case, the use of big data is being used to help overcome misunderstandings that result in unnecessary measles outbreaks.

Credit: *Annals of Internal Medicine,* 2019

EXAMPLE 4 **Nominal Level**

Here are examples of sample data at the nominal level of measurement.

1. **Yes/No/Undecided:** Survey responses of *yes*, *no*, and *undecided*
2. **Coded Survey Responses:** For an item on a survey, respondents are given a choice of possible answers, and they are coded as follows: "I agree" is coded as 1; "I disagree" is coded as 2; "I don't care" is coded as 3; "I refuse to answer" is coded as 4; "Go away and stop bothering me" is coded as 5. The numbers 1, 2, 3, 4, 5 don't measure or count anything.

YOUR TURN. Do Exercise 21 "Births."

Because nominal data lack any ordering or numerical significance, they should not be used for calculations. Numbers such as 1, 2, 3, and 4 are sometimes assigned to the different categories (especially when data are coded for computers), but these numbers have no real computational significance and any average (mean) calculated from them is meaningless and possibly misleading.

DEFINITION

Data are at the **ordinal level of measurement** if they can be arranged in some order, but differences (obtained by subtraction) between data values either cannot be determined or are meaningless.

EXAMPLE 5 **Ordinal Level**

Here is an example of sample data at the ordinal level of measurement.

Course Grades: A biostatistics professor assigns grades of A, B, C, D, or F. These grades can be arranged in order, but we can't determine differences between the grades. For example, we know that A is higher than B (so there is an ordering), but we cannot subtract B from A (so the difference cannot be found).

YOUR TURN. Do Exercise 22 "Medical School Rankings."

Ordinal data provide information about relative comparisons but not the *magnitudes* of the differences. Ordinarily, ordinal data (such as course grades of A, B, C, D, F) should not be used for calculations such as the average (mean), but calculations are commonly used for some ordinal data, such as data from a survey question with a rating scale of 0 to 10. (A *Likert scale* is used to measure attitudes or opinions with a scale used for the level of agreement, usually with five to ten choices ranging from one extreme opinion to the opposite extreme.)

DEFINITION

Data are at the **interval level of measurement** if they can be arranged in order, and differences between data values can be found and are meaningful. *Data at the interval level do not have a natural zero starting point at which none of the quantity is present.*

EXAMPLE 6 Interval Level

These examples illustrate the interval level of measurement.

1. **Body Temperatures:** The body temperatures of 98.0°F and 97.0°F listed in Data Set 5 "Body Temperatures" in Appendix B are examples of data at the interval level of measurement. Those values are ordered, and we can determine that their difference is 1.0°F. However, there is no natural starting point. The value of 0°F is arbitrary and does not represent the total absence of heat.
2. **Pandemic Years:** Major pandemics occurred in the years 1918 and 2020. Those numbers can be arranged in order and the difference of 102 years can be found and is meaningful. However, time did not begin in the year 0, so the year 0 is arbitrary instead of being a natural zero starting point representing "no time." The years 1918 and 2020 are therefore at the interval level of measurement.
3. **Shoe Sizes:** The shoe sizes of 10 and 5 can be arranged in order, and the difference is the same as the difference in shoe sizes of 8 and 13. However, size 0 is arbitrary.

YOUR TURN. Do Exercise 25 "Manatee Boat Deaths."

DEFINITION

Data are at the **ratio level of measurement** if they can be arranged in order, differences can be found and are meaningful, and *there is a natural zero starting point* (where zero indicates that none of the quantity is present). For data at this level, differences and ratios are both meaningful.

EXAMPLE 7 Ratio Level

The following are examples of data at the ratio level of measurement. Note the presence of the natural zero value, and also note the use of meaningful ratios of "twice" and "three times."

1. **Heights of Students:** Heights of 180 cm and 90 cm for a high school student and a preschool student (0 cm represents no height, and 180 cm is *twice* as tall as 90 cm.)
2. **Class Times:** The times of 50 min and 100 min for a biostatistics class (0 min represents no class time, and 100 min is *twice* as long as 50 min.)

YOUR TURN. Do Exercise 24 "Pulse Rates."

TABLE 1-2 Levels of Measurement

Level of Measurement	Brief Description	Example
Ratio	There is a natural zero starting point and ratios make sense.	Heights, lengths, distances, volumes
Interval	Differences are meaningful, but there is no natural zero starting point and ratios are meaningless.	Body temperatures in degrees Fahrenheit or Celsius
Ordinal	Data can be arranged in order, but differences either can't be found or are meaningless.	Ranks of medical schools in *U.S. News & World Report*
Nominal	Categories only. Data cannot be arranged in order.	Eye colors

Six Degrees of Separation

Credit: Allstar Picture Library/Alamy Stock Photo

Social psychologists, historians, political scientists, and communications specialists are interested in "The Small World Problem": Given any two people in the world, how many intermediate links are necessary to connect the two original people? In the 1950s and 1960s, social psychologist Stanley Milgram conducted an experiment in which subjects tried to contact other target people by mailing an information folder to an acquaintance who they thought would be closer to the target. Among 160 such chains that were initiated, only 44 were completed, so the failure rate was 73%. Among the successes, the number of intermediate acquaintances varied from 2 to 10, with a median of 6 (hence "six degrees of separation"). The experiment has been criticized for its high failure rate and its disproportionate inclusion of subjects with above-average incomes. A more recent study conducted by Microsoft researcher Eric Horvitz and Stanford Assistant Professor Jure Leskovec involved 30 billion instant messages and 240 million people. This study found that for instant messages that used Microsoft, the mean length of a path between two individuals is 6.6, suggesting "seven degrees of separation." Work continues in this important and interesting field.

HINT The distinction between the interval and ratio levels of measurement can be a bit tricky. Here are two tools to help with that distinction:

1. **Ratio Test** Focus on the term "ratio" and know that the term "twice" describes the ratio of one value to be double the other value. To distinguish between the interval and ratio levels of measurement, use a "ratio test" by asking this question: Does use of the term "twice" make sense? "Twice" makes sense for data at the ratio level of measurement, but it does not make sense for data at the interval level of measurement.
2. **True Zero** For ratios to make sense, there must be a value of "true zero," where the value of zero indicates that none of the quantity is present, and zero is not simply an arbitrary value on a scale. The temperature of 0°F is arbitrary and does not indicate that there is no heat, so temperatures on the Fahrenheit scale are at the interval level of measurement, not the ratio level.

EXAMPLE 8 Distinguishing Between the Ratio Level and Interval Level

For each of the following, determine whether the data are at the ratio level of measurement or the interval level of measurement:

a. Times (minutes) it takes biostatistics students to complete a statistics test

b. Body temperatures (Celsius) of biostatistics students

SOLUTION

a. Apply the "ratio test" described in the preceding hint. If one student completes the test in 40 minutes and another student completes the test in 20 minutes, does it make sense to say that the first student used *twice* as much time? Yes! So the times are at the ratio level of measurement. We could also apply the "true zero" test. A time of 0 minutes does represent "no time," so the value of 0 is a true zero indicating that no time was used.

b. Apply the "ratio test" described in the preceding hint. If one student has a body temperature of 40°C and another student has a body temperature of 20°C, does it make sense to say that the first student is *twice* as hot as the second student? No! So the body temperatures are not at the ratio level of measurement. Because the difference between 40°C and 20°C is the same as the difference between 90°C and 70°C, the differences are meaningful, but because ratios do not make sense, the body temperatures are at the interval level of measurement. Also, the temperature of 0°C does not represent "no heat" so the value of 0 is not a true zero indicating that no heat is present.

YOUR TURN. Do Exercise 27 "Arsenic."

PART 2 Big Data and Missing Data: Too Much and Not Enough

When working with data, we might encounter some data sets that are ginormous, and we might also encounter some data sets with individual elements missing. Here in Part 2 we briefly discuss both cases.

Big Data

Based on a study of 61.9 million electronic medical records of U.S. adults aged 18 and older, it was discovered that people with dementia are twice as likely to contract coronavirus. (For another advantage of using big data, see the margin essay "Big Data Instead of a Clinical Trial.") The need to analyze such large data sets has led to the birth of *data science*. There is not universal agreement on the following definitions, and various other definitions can be easily found elsewhere.

DEFINITIONS

Big data refers to data sets so large and so complex that their analysis is beyond the capabilities of traditional software tools. Analysis of big data may require software simultaneously running in parallel on many different computers.

Data science involves applications of statistics, computer science, and software engineering, along with some other relevant fields (such as genetics).

Examples of Data Set Magnitudes We can see from the definition of big data that there isn't a fixed number that serves as an exact boundary for determining whether a data set qualifies as being big data, but big data typically involves amounts of data such as the following:

- Terabytes (10^{12} or 1,000,000,000,000 bytes) of data
- Petabytes (10^{15} bytes) of data
- Exabytes (10^{18} bytes) of data
- Zettabytes (10^{21} bytes) of data
- Yottabytes (10^{24} bytes) of data

Examples of Applications of Big Data The following are a few other examples involving big data:

- Attempts to forecast flu epidemics are made by analyzing Internet searches of flu symptoms.
- Kaiser Permanente, a healthcare network based in California, has roughly 40,000,000,000,000,000 bytes of data from electronic health records.
- Flatiron Health analyzes billions of data points from cancer patients so that care can be enhanced.
- Correlations and patterns from millions of medical records are identified so that cures for diseases can be found.
- The company Tempus is compiling a massive database of molecular data so that physicians can personalize treatment of patients.

Examples of Jobs According to Analytic Talent, there are 6000 companies hiring data scientists, and here are some job posting examples:

- Facebook: Data Scientist
- IBM: Data Scientist
- PayPal: Data Scientist
- The College Board: SAS Programmer/Data Scientist
- Netflix: Senior Data Engineer/Scientist

Big Data Instead of a Clinical Trial

Credit: 18percentgrey/Shutterstock

Nicholas Tatonetti of Columbia University searched Food and Drug Administration databases for adverse reactions in patients that resulted from different pairings of drugs. He discovered that the Paxil (paroxetine) drug for depression and the pravastatin drug for high cholesterol interacted to create increases in glucose (blood sugar) levels. When taken separately by patients, neither drug raised glucose levels, but the increase in glucose levels occurred when the two drugs were taken together. This finding resulted from a general database search of interactions from many pairings of drugs, not from a clinical trial involving patients using Paxil and pravastatin.

Hawthorne and Experimenter Effects

Credit: Triff/ Shutterstock

The well-known placebo effect occurs when untreated subjects incorrectly believe that they are receiving a real treatment and report an improvement in symptoms. The Hawthorne effect occurs when treated subjects somehow respond differently simply because they are part of an experiment. (This phenomenon was called the "Hawthorne effect" because it was first observed in a study of factory workers at Western Electric's Hawthorne plant.) An experimenter effect (sometimes called a Rosenthal effect) occurs when the researcher or experimenter unintentionally influences subjects through such factors as facial expression, tone of voice, or attitude.

It was noted in the Preface that we are experiencing a new major revolution in technology that uses artificial intelligence, machine learning, and deep learning—topics studied in Data Science, which requires a study of statistics. Data Science and statistics are now experiencing unprecedented growth.

Statistics in Data Science The modern data scientist has a solid background in statistics and computer systems as well as expertise in fields that extend beyond statistics. The modern data scientist might be skilled with software of *R*, Python, or Hadoop. The modern data scientist might also have a strong background in some other field such as psychology, biology, medicine, chemistry, or economics. Because of the wide range of disciplines required, a data science project might typically involve a team of collaborating individuals with expertise in different fields. An introductory statistics course is a great first step in becoming a data scientist.

Missing Data

When collecting sample data, it is quite common to find that some values are missing. Ignoring missing data can sometimes create misleading results. If you make the mistake of skipping over a few different sample values when you are manually typing them into a statistics software program, the missing values are not likely to have a serious effect on the results. However, if a survey includes many missing salary entries because those with very low incomes are reluctant to reveal their salaries, those missing low values will have the serious effect of making salaries appear higher than they really are.

For an example of missing data, see the following table. The body temperature for Subject 2 at 12 AM on Day 2 is missing. (The table below includes the first three rows of data from Data Set 5 "Body Temperatures" in Appendix B.)

Body Temperatures (in degrees Fahrenheit) of Healthy Adults

			Temperature Day 1		Temperature Day 2	
Subject	Sex	Smoke	8 AM	12 AM	8 AM	12 AM
1	M	Y	98.0	98.0	98.0	98.6
2	M	Y	97.0	97.6	97.4	----
3	M	Y	98.6	98.8	97.8	98.6

There are different categories of missing data, as described in the following definitions.

DEFINITION

A data value is **missing completely at random** if the likelihood of its being missing is independent of its value or any of the other values in the data set. That is, any data value is just as likely to be missing as any other data value.

(*NOTE:* More complete discussions of missing data distinguish between *missing completely at random* and *missing at random*, which means that the likelihood of a value being missing is independent of its value after controlling for another variable. There is no need to know this distinction in this text.)

Example of Missing Data—Random When using a keyboard to manually enter ages of survey respondents, the operator is distracted by a colleague singing "Fever" and makes the mistake of failing to enter the age of 37 years. This data value is missing completely at random.

DEFINITION

A data value is **missing not at random** if the missing value is related to the reason that it is missing.

Example of Missing Data—Not at Random A survey question asks respondents to enter their annual income, but respondents with very low incomes skip this question because they find it embarrassing.

Biased Results? Based on the preceding two definitions and examples, it makes sense to conclude that if we ignore data *missing completely at random*, the remaining values are not likely to be biased and good results should be obtained. However, if we ignore data that are *missing not at random*, it is very possible that the remaining values are biased and results will be misleading.

Correcting for Missing Data There are different methods for dealing with missing data.

1. **Delete Cases:** One very common method for dealing with missing data is to delete all subjects having any missing values.
   - If the data are missing completely at random, the remaining values are not likely to be biased and good results can be obtained, but with a smaller sample size.
   - If the data are missing not at random, deleting subjects that have any missing values can easily result in a bias among the remaining values, so results can be misleading.
2. **Impute Missing Values:** We "impute" missing data values when we substitute values for them. There are different methods of determining the replacement values, such as using the mean of the other values, or using a randomly selected value from other similar cases, or using a method based on regression analysis (which will make more sense after studying Chapter 10).

In this text we do not work much with missing data, but it is important to understand this:

> **When analyzing sample data with missing values, try to determine *why* they are missing, then decide whether it makes sense to treat the remaining values as being representative of the population. If it appears that there are missing values that are *missing not at random* (that is, their values are related to the reasons why they are missing), know that the remaining data may well be biased and any conclusions based on those remaining values may well be misleading.**

1-2 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Parameter and Statistic A Quest Diagnostics analysis of 10 million drug tests of adults in the United States revealed that 4.2% of the tests were positive for illegal drugs. Identify the population and the sample. Is the value of 4.2% a statistic or a parameter?

Declining Response Rate

Credit: 123 RF GB Limited

The Pew Research Center is now using the Internet for most of its surveys conducted in the United States. One major factor precipitating that change is the low and declining response rate of telephone surveys. The response rate for telephone surveys was 36% in 1997, but it has now dropped to only 6%. A major cause of this declining response rate is the high and growing use of robocalls. Public opinion surveys conducted by telephone usually appear as an unknown source, so potential respondents are much more likely to reject such calls. However, Pew research has shown that low response rates do not cause inaccurate results. But low response rates for telephone surveys do result in higher survey costs.

2. Quantitative / Categorical Data Identify each of the following as quantitative data or categorical data.

a. The platelet counts of exam subjects in Data Set 1 "Body Data" in Appendix B

b. The names of the pharmaceutical companies that manufacture acetaminophen tablets

c. The colors of acetaminophen tablets

d. The weights of acetaminophen tablets

e. The prices of acetaminophen tablets

3. Discrete / Continuous Data Which of the following describe discrete data?

a. The numbers of people surveyed in each of the next several National Health and Nutrition Examination Surveys

b. The exact foot lengths (cm) of a random sample of biostatistics students

c. The exact times that randomly selected surgeons perform appendectomies

4. E-Cigarette Survey In a survey of 36,000 adults, 3.7% said that they regularly use E-cigarettes (based on data from the National Center for Health Statistics).

a. Identify the sample and population.

b. Is the value of 3.7% a statistic or parameter?

c. What is the level of measurement of the value of 3.7%? (nominal, ordinal, interval, ratio)

d. Are the numbers of subjects in such surveys discrete or continuous?

In Exercises 5–12, identify whether the given value is a statistic or a parameter.

5. Reported Weight From the sample of 5755 *reported* weights listed in Data Set 4 "Measured and Reported" from Appendix B, the average (mean) is 79.95 kg.

6. Measured Weight From the sample used in the preceding exercise, the average (mean) *measured* weight is 80.89 kg.

7. Height of Presidents Data Set 19 "Presidents" in Appendix B lists the heights of all Presidents as of this writing. The average (mean) is 180.0 cm.

8. Cuckoo Egg Length The average (mean) length of the hedge sparrow eggs included in Data Set 17 "Cuckoo Egg Lengths" in Appendix B is 23.12 mm.

9. Cardiac Surgery In a study comparing inhaled anesthetics and intravenous anesthetics used in patients undergoing coronary-artery bypass grafting, 154 of the patients did not survive (based on "Volatile Anesthetics versus Total Intravenous Anesthesia for Cardiac Surgery," by Landoni et al., *The New England Journal of Medicine,* Vol. 383, No. 13).

10. Motor Vehicle Fatalities In a recent year, the number of deaths from road accidents worldwide was 1.35 million (based on data from the Centers for Disease Control and Prevention).

11. *Titanic* A study was conducted of all 2223 passengers aboard the *Titanic* when it sank.

12. Periodic Table The average (mean) atomic weight of all elements in the periodic table is 134.355 unified atomic mass units.

In Exercises 13–20, determine whether the data are from a discrete or continuous data set.

13. Brain Size Data Set 12 "IQ and Brain Size" in Appendix B includes the volumes (cm^3) of human brains.

14. Manatee Boat Deaths Data Set 15 "Manatee Boat Deaths" in Appendix B includes the annual numbers of manatee deaths caused by boats.

15. Nicotine in Cigarettes Data Set 22 in Appendix B includes the amounts of nicotine (mg) in a sample of cigarettes.

16. Births The Centers for Disease Control and Prevention publishes vital statistics that include the numbers of births each year in the United States.

17. BMI From Data Set 1 "Body Data" in Appendix B, we see that the average BMI of males is 28.16.

18. Arsenic in Rice Data Set 23 "Arsenic in Rice" in Appendix B lists the weights of arsenic in rice measured in micrograms (μg) per serving.

19. Biostatistics Classes In each of her classes, a biostatistics professor records the number of students who earned a grade of A.

20. Criminal Forensics When studying the relationship between lengths of feet and heights so that footprint evidence at a crime scene can be used to estimate the height of the suspect, a researcher records the exact lengths of feet from a large sample of random subjects.

In Exercises 21–28, determine which of the four levels of measurement (nominal, ordinal, interval, ratio) best describes the given data.

21. Births Data Set 6 "Births" in Appendix B includes the sex of newborn babies. Those values are 0's and 1's, where 0 represents female and 1 represents male.

22. Medical School Rankings *U.S. News & World Report* provides rankings of medical schools. In a recent year, the ranks for Harvard, New York University, and Duke were 1, 2, and 3, respectively.

23. Pain Scale Physicians routinely use the Numeric Rating Scale (NRS-11), which consists of the numbers 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 used by patients to report their level of pain. 0 represents no pain and 10 represents the worst possible pain (such as statistics calculations done without technology).

24. Pulse Rates Data Set 1 "Body Data" in Appendix B includes pulse rates measured in beats per minute.

25. Manatee Boat Deaths Data Set 15 "Manatee Boat Deaths" in Appendix B includes the years in which the deaths were recorded, beginning with 1991.

26. Hospitals Data Set 6 "Births" in Appendix B includes the names of the hospitals in which the births occurred.

27. Arsenic Data Set 23 "Arsenic in Rice" in Appendix B includes the weights (μg) of arsenic in the servings of rice.

28. Body Temperatures Body temperatures (in degrees Fahrenheit) listed in Data Set 5 "Body Temperatures" in Appendix B

In Exercises 29–32, identify the level of measurement of the data as nominal, ordinal, interval, or ratio. Also, explain what is wrong with the given calculation.

29. Clinical Trial In order to maintain the privacy of patients in a clinical trial, the patients are assigned identification numbers randomly selected between 1 and 279. The mean of those numbers is 140.

30. Medical School Rankings From the medical school rankings in Exercise 22, the difference between Harvard and New York University is the same as the difference between New York University and Duke.

31. Pain Scale Using the pain scale described in Exercise 23, one patient reports a pain level of 8 while a second patient reports a pain level of 4, so the first patient has twice as much pain as the second patient.

32. Body Temperatures One patient has a body temperature measured to be 100°F and another patient has a body temperature measured to be 95°F, so the second patient is 5% cooler than the first patient.

1-2 Beyond the Basics

33. Countable For each of the following, categorize the nature of the data using one of these three descriptions: (1) discrete because the number of possible values is finite; (2) discrete because the number of possible values is infinite but countable; (3) continuous because the number of possible values is infinite and not countable.

a. Exact lengths of the feet of members of the rock band The Monkees

b. Shoe sizes of members of the rock band The Monkees (such as 9, 9½, and so on)

c. The number of albums sold by The Monkees rock band

d. The numbers of primate monkeys typing at keyboards before one of them randomly types the lyrics for the song "Daydream Believer."

34. Directions in Degrees Standard navigation systems used for aviation and boating are based on directions measured in degrees, with north represented by 0°. Relative to north, east is 90°, south is 180°, and west is 270°. What is the level of measurement of such directions measured in degrees?

1-3 Collecting Sample Data

Key Concept When analyzing sample data, it is essential to use an appropriate method for collecting those sample data. This section includes comments about various methods and sampling procedures. Of particular importance is the method of using a *simple random sample*. We will make frequent use of this sampling method throughout the remainder of this text.

As you read this section, remember this:

If sample data are not collected in an appropriate way, the data may be so utterly useless that no amount of statistical torturing can salvage them.

PART 1 Basics of Design of Experiments and Collecting Sample Data

The Gold Standard Randomness with placebo/treatment groups is sometimes called the "gold standard" because it is so effective.

DEFINITION

A **placebo** is a harmless and ineffective pill, medicine, or procedure sometimes used for psychological benefit or sometimes used by researchers for comparison to other treatments.

The following example describes how the gold standard was used in the largest health experiment ever conducted.

EXAMPLE 1 The Salk Vaccine Experiment

In 1954, an experiment was designed to test the effectiveness of the Salk vaccine in preventing polio, which had killed or paralyzed thousands of children. By random selection, 401,974 children were randomly assigned to two groups: (1) 200,745 children were given a *treatment* consisting of Salk vaccine injections; (2) 201,229 children were injected with a *placebo* that contained no drug. Children were assigned to the treatment or placebo group through a process of random selection, equivalent to flipping a coin. Among the children given the Salk vaccine, 33 later developed paralytic polio, and among the children given a placebo, 115 later developed paralytic polio.

YOUR TURN. Do Exercise 1 "Magnet Treatment of Pain."

Example 1 describes an *experiment* because subjects were given a treatment, but ethical, cost, time, and other considerations sometimes prohibit the use of an experiment. We would never want to conduct a driving/texting experiment in which we ask subjects to text while driving—some of them could die. It would be far better to observe past crash results to understand the effects of driving while texting. See the following definitions.

DEFINITIONS

In an **experiment**, we apply some *treatment* and then proceed to observe its effects on the individuals. (The individuals in experiments are called **experimental units**, and they are often called **subjects** when they are people.)

In an **observational study**, we observe and measure specific characteristics, but we don't attempt to *modify* the individuals being studied.

Experiments are often better than observational studies because well-planned experiments typically reduce the chance of having the results affected by some variable that is not part of a study. A *lurking variable* is one that affects the variables included in the study, but it is not included in the study.

EXAMPLE 2 Ice Cream and Drownings

Observational Study: Observe past data to incorrectly conclude that ice cream causes drownings (based on data showing that increases in ice cream sales are associated with increases in drownings). The mistake is to miss the lurking variable of temperature and the failure to see that as the temperature increases, ice cream sales increase and drownings increase because more people swim.

Experiment: Conduct an *experiment* with one group treated with ice cream while another group gets no ice cream. We would see that the rate of drowning victims is about the same in both groups, so ice cream consumption has no effect on drownings.

Here, the experiment is clearly better than the observational study.

YOUR TURN. Do Exercise 6 "Experiment or Observational Study."

Clinical Trials Versus Observational Studies

Credit: Andersen Ross/Stockbyte/Getty Images

In a *New York Times* article about hormone therapy for females, reporter Denise Grady wrote about randomized clinical trials that involve subjects who were randomly assigned to a treatment group and another group not given the treatment. Such randomized clinical trials are often referred to as the "gold standard" for medical research. In contrast, observational studies can involve patients who decide themselves to undergo some treatment. Subjects who decide themselves to undergo treatments are often healthier than other subjects, so the treatment group might appear to be more successful simply because it involves healthier subjects, not necessarily because the treatment is effective. Researchers criticized observational studies of hormone therapy for females by saying that results might appear to make the treatment more effective than it really is.

Pew Surveys

Credit: andreypopov/123rf

In the recent past, Pew Research Center conducted surveys by randomly calling phone numbers selected from a nationwide list of landline and cell phone numbers. Largely because of robocalls, Americans are now unlikely to answer calls from unknown numbers. Currently, with 93% of adults using the Internet, Pew Research Center does most of its polling online. Pew starts with a master list of residential addresses and mails printed invitations to randomly selected subjects who are invited to take several online surveys. Mailed surveys are used for the 7% of the population who don't use the Internet, and Pew also sends tablets and data plans so that some of them can use the Internet. As an incentive, participants are paid between $5 and $20 for each survey.

Design of Experiments

Good design of experiments includes *replication, blinding,* and *randomness*.

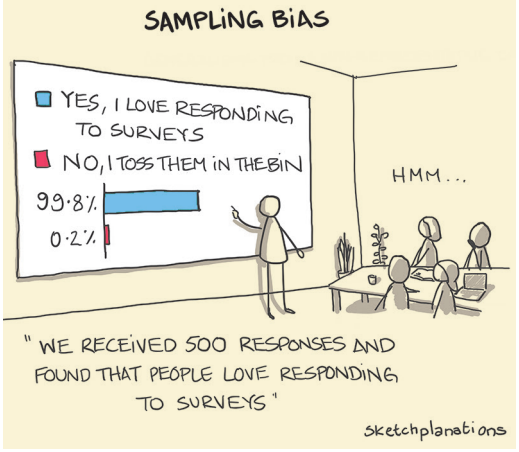
- **Replication** is the repetition of an experiment on more than one individual. Good use of replication requires sample sizes that are large enough so that we can see effects of treatments. The Salk experiment in Example 1 used sufficiently large sample sizes, so the researchers could see that the Salk vaccine was effective.
- **Blinding** is used when the subjects don't know whether they are receiving a treatment or a placebo. Blinding is a way to get around the **placebo effect,** which occurs when an untreated subject reports an improvement in symptoms. (The reported improvement in the placebo group may be real or imagined.) The Salk experiment in Example 1 was **double-blind**, which means that blinding occurred at two levels: (1) The children being injected didn't know whether they were getting the Salk vaccine or a placebo, and (2) the doctors who gave the injections and evaluated the results did not know either. Codes were used so that the researchers could objectively evaluate the effectiveness of the Salk vaccine.
- **Randomness** is used when individuals are assigned to different groups through a process of random selection, as in the Salk vaccine experiment in Example 1. The logic behind randomness is to use chance as a way to create two groups that are similar. The following definition refers to one common and effective way to collect sample data in a way that uses randomness.

DEFINITION

A **simple random sample** of n subjects is selected in such a way that every possible *sample of the same size n* has the same chance of being chosen. (A simple random sample is often called a random sample, but strictly speaking, a *random sample* has the weaker requirement that all members of the population have the same chance of being selected. That distinction is not so important in this text. See Exercise 38 "Simple Random Sample vs. Random Sample".)

Throughout, we will use various statistical procedures, and we often have a requirement that we have collected a *simple random sample*, as defined above.

Unlike careless or haphazard sampling, random sampling usually requires very careful planning and execution.



Clever illustration of sampling bias. Image is by Sketchplanations.

Other Sampling Methods In addition to simple random sampling, here are some other sampling methods commonly used for surveys. Figure 1-4 illustrates these different sampling methods.

DEFINITIONS

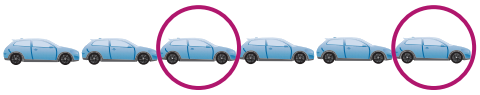
With **systematic sampling**, we select some starting point and then select every *k*th (such as every 50th) element in the population.

With **convenience sampling**, we simply use data that are very easy to get.

With **stratified sampling**, we subdivide the population into at least two different subgroups (or strata) so that subjects within the same subgroup share the same characteristics (such as sex). Then we draw a sample from each subgroup (or stratum).

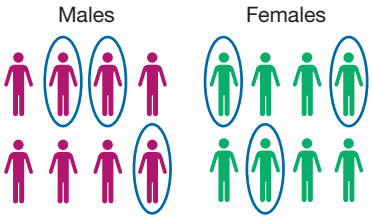
With **cluster sampling**, we first divide the population area into sections (or clusters). Then we randomly select some of those clusters and choose *all* the members from those selected clusters.

Simple Random Sample
Select a sample of *n* subjects so that every sample of the same size *n* has the same chance of being selected.



Systematic Sample
Select every *k*th subject.

Convenience Sample
Use data that are very easy to get.



Stratified Sample
Subdivide population into strata (groups) with the same characteristics, then randomly sample within those strata.

Cluster Sample
Partition the population in clusters (groups), then randomly select some clusters, then select all members of the selected clusters.

FIGURE 1-4 Common Sampling Methods
Credit: Pixsooz/Shutterstock

Value of a Statistical Life

Credit: Fujji/ Shutterstock

The *value of a statistical life* (VSL) is a measure routinely calculated and used for making decisions in fields such as medicine, insurance, environmental health, and transportation safety. As of this writing, the value of a statistical life is $11.8 million (source: U.S. Department of Transportation).

Many people oppose the concept of putting a value on a human life, but the word *statistical* in the "value of a statistical life" is used to ensure that we don't equate it with the true worth of a human life. Some people legitimately argue that every life is priceless, but others argue that there are conditions in which it is impossible or impractical to save every life, so a value must be somehow assigned to a human life in order that sound and rational decisions can be made.

HINT Because it's difficult to remember the distinction between stratified sampling and cluster sampling, picture your entire class as one cluster among all classes at your college. Remember the alliteration of "cluster class" to recall that with cluster sampling, you choose *all* of the members of selected clusters. Associate "cluster" with "all." Then, stratified sampling is the other method of choosing samples from selected classes or subgroups.

Multistage Sampling Professional pollsters and government researchers often collect data by using some combination of the preceding sampling methods. In a multistage sample design, pollsters select a sample in different stages, and each stage might use different methods of sampling, as in the following example.

EXAMPLE 3 **Multistage Sample Design**

The U.S. government's unemployment statistics are based on surveys of households. It is impractical to personally survey each household in a simple random sample, because they would be scattered all over the country, making it nearly impossible to contact each of them. Instead, the U.S. Census Bureau and the Bureau of Labor Statistics collaborate to conduct a survey called the Current Population Survey. A recent survey incorporates a multistage sample design, roughly following these steps:

1. The entire United States is partitioned into 2025 different regions called *primary sampling units* (PSUs). The primary sampling units are metropolitan areas, large counties, or combinations of smaller counties. The 2025 primary sampling units are then grouped into 824 different strata.
2. In each of the 824 different strata, one of the primary sampling units is selected so that the probability of selection is proportional to the size of the population in each primary sampling unit.
3. Among the 824 selected primary sampling units, census data are used to randomly select about 60,000 households.
4. A responsible person in each of the 60,000 selected households is interviewed about the employment status of each household member of age 16 or older.

This multistage sample design includes a combination of random, stratified, and cluster sampling at different stages. The end result is a very complicated sampling design, but it is much more practical, less expensive, and faster than using a simpler design, such as a simple random sample. (Using a simple random sample would result in households that are far apart and difficult to contact.)

PART 2 Beyond the Basics of Design of Experiments and Collecting Sample Data

In Part 2 of this section, we discuss different types of observational studies and different ways of designing experiments.

Observational Studies The following definitions identify the standard terminology used in professional journals for different types of observational studies. These definitions are illustrated in Figure 1-5.

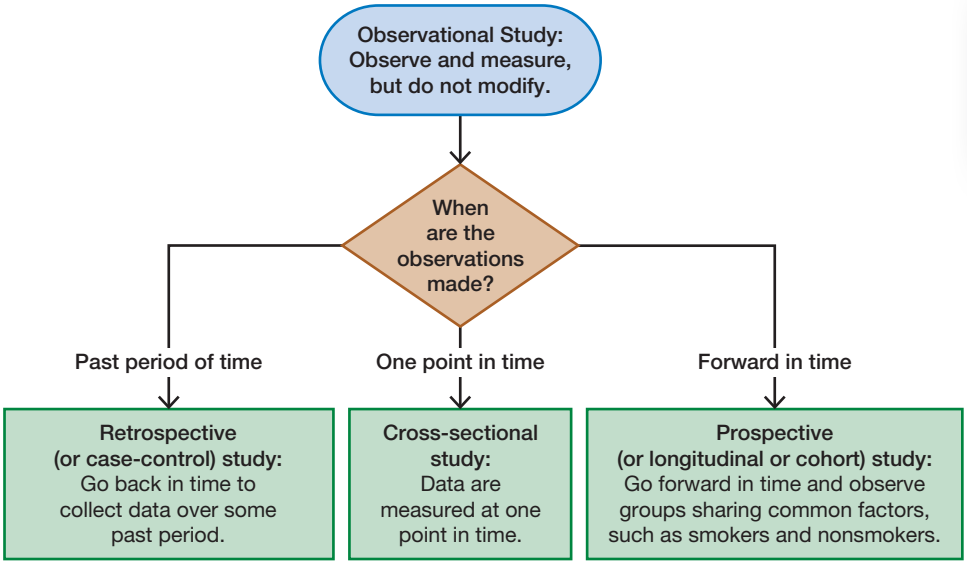


FIGURE 1-5 Types of Observational Studies

DEFINITIONS

In a **cross-sectional study**, data are observed, measured, and collected at *one point in time*, not over a period of time.

In a **retrospective (**or **case-control) study**, data are collected from a *past time period* by going back in time (through examination of records, interviews, and so on).

In a **prospective (**or **longitudinal** or **cohort) study**, data are collected in the *future* from groups that share common factors (such groups are called *cohorts*).

Experiments In an experiment, confounding occurs when we can see some effect, but we can't identify the specific factor that caused it, as in the ice cream and drowning observational study in Example 2. See also the bad experimental design illustrated in Figure 1-6(a) on the next page, where confounding can occur when the treatment group of females shows strong positive results. Because the treatment group consists of females and the placebo group consists of males, confounding has occurred because we cannot determine whether the positive results are attributable to the treatment or to the sex of the subjects. The Salk vaccine experiment in Example 1 illustrates one method for controlling the effect of the treatment variable: Use a *completely randomized experimental design*, whereby randomness is used to assign subjects to the treatment group and the placebo group. A completely randomized experimental design is just one of the following methods that are used to control effects of variables.

Completely Randomized Experimental Design: Assign subjects to different treatment groups through a process of *random selection*, as illustrated in Figure 1-6(b) on the next page.

Randomized Block Design: See Figure 1-6c on the next page. A **block** is a group of subjects that are similar, but blocks differ in ways that might affect the outcome of the experiment. Use the following procedure, as illustrated in Figure 1-6(c):

1. Form blocks (or groups) of subjects with similar characteristics.
2. Randomly assign treatments to subjects within each block.

The Human Project

Credit: Pavlo Vakhrushev/123RF

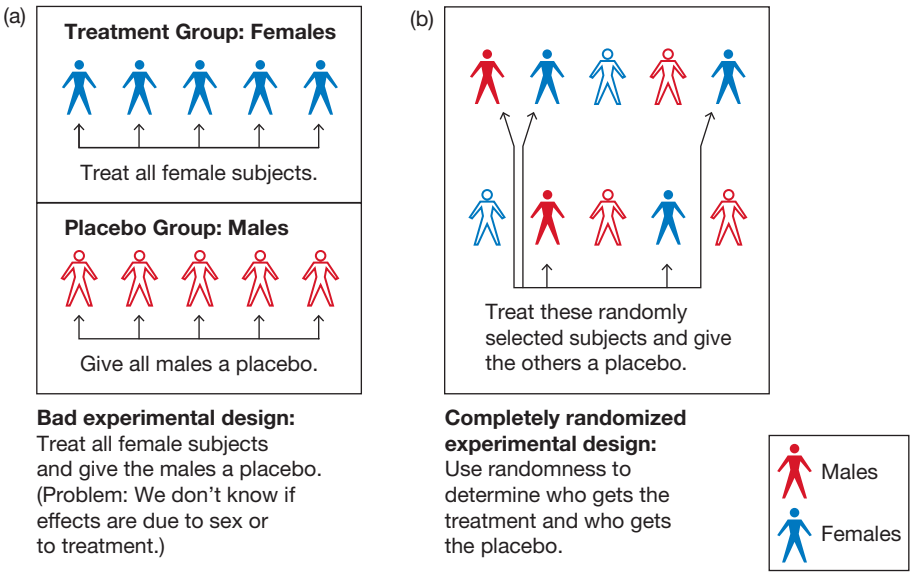
Started in 2014, the Human Project is a *prospective study* in which 10,000 New Yorkers will be followed for decades. The goal of this ambitious study is to "*provide new insights and understanding into how our biology, our environment, and our behavior interact to determine our health.*" The subjects in the study will be the basis for collecting medical records, education records, data from physical examinations, patterns of physical activity, environmental measurements, and a wide variety of other measurements. The hope is that big data analysis will enable researchers to generate new insights into the biological, behavioral, and environmental factors that influence our health. The Human Project was started by the Kavli Foundation and the New York University Institute for Interdisciplinary Study of Decision Making.

For example, in designing an experiment to test the effectiveness of aspirin treatments on heart disease, we might form a block of males and a block of females, because it is known that the hearts of males and females can behave differently. By controlling for sex, this randomized block design eliminates sex as a possible source of confounding.

A randomized block design uses the same basic idea as stratified sampling, but randomized block designs are used when designing experiments, whereas stratified sampling is used for surveys.

Matched Pairs Design: Compare two treatment groups (such as treatment and placebo) by using subjects matched in pairs, as in the following cases.

- Before/After: Matched pairs might consist of measurements from subjects before and after some treatment, as illustrated in Figure 1-6(d). Each subject yields a "before" measurement and an "after" measurement, and each before/after pair of measurements is a matched pair.
- Twins: A test of Crest toothpaste used matched pairs of twins, where one twin used Crest and the other used another toothpaste.



Bad experimental design: Treat all female subjects and give the males a placebo. (Problem: We don't know if effects are due to sex or to treatment.)

Completely randomized experimental design: Use randomness to determine who gets the treatment and who gets the placebo.

Randomized block design:
1. Form a block of females and a block of males.
2. Within each block, randomly select subjects to be treated.

Matched pairs design: Get measurements from the same subjects before and after some treatment.

FIGURE 1-6 Designs of Experiments

Rigorously Controlled Design: Carefully assign subjects to different treatment groups, so that those given each treatment are similar in the ways that are important to the experiment. This can be extremely difficult to implement, and often we can never be sure that we have accounted for all of the relevant factors.

Sampling Errors

In statistics, you could use a good sampling method and do everything correctly, and yet it is possible to get wrong results. No matter how well you plan and execute the sample collection process, there is likely to be some error in the results. The different types of sampling errors are described here.

DEFINITIONS

A **sampling error** (or **random sampling error**) occurs when the sample has been selected with a random method, but there is a discrepancy between a sample result and the true population result; such an error results from chance sample fluctuations.

A **nonsampling error** is the result of human error, including such factors as wrong data entries, computing errors, questions with biased wording, false data provided by respondents, forming biased conclusions, or applying statistical methods that are not appropriate for the circumstances.

A **nonrandom sampling error** is the result of using a sampling method that is not random, such as using a convenience sample or a voluntary response sample.

Experimental design requires much more thought and care than we can describe in this relatively brief section. Taking a complete course in the design of experiments is a good start in learning so much more about this important topic.

PART 3 Clinical Trials

A clinical trial is a prospective experiment designed to test the safety and effectiveness of a new drug, device, surgery technique, counseling method, and so on. Years of research with animals and/or human cells are typically conducted *before* a clinical trial begins. In the U.S., the clinical trial itself typically takes about six years at a cost of millions of dollars.

A "warp speed" clinical trial that is a notable exception:

Pfizer and BioNTech worked together in 2020 to develop a vaccine for the prevention of the COVID-19 virus. It took only *eight months* to develop and gain emergency approval for their vaccine. The official FDA approval was granted on August 23, 2021, when marketing of the "Pfizer-BioNTech COVID-19 Vaccine" began under the new name of "Comirnaty."

Note: All of the following lengths of time and costs are estimates that can vary substantially and are based on the U.S. approval process. See "Estimating Research and

Development Investment Needed to Bring a New Medicine to Market," by Wouters, McKee, and Luyten, *Journal of the American Medical Association,* Vol. 323, No. 9.

If research conducted prior to the clinical trial is successful, the results are sent to the Food and Drug Administration (FDA) for approval in the U.S. so that testing with humans can begin. About 14% of all drugs successfully complete a clinical trial and gain FDA approval. The median cost of a clinical trial is about $19 million.

A clinical trial is typically conducted with the following four phases.

Phase 1: Test for Safety

Goal: The main goal of this first phase is to assess the safety of the treatment and to identify the correct dosage.
Sample Size: This phase usually involves a small number of subjects, such as 20 to 100.
Time and Cost: Phase 1 can take several months and cost around $4 million.
Success Rate: About 70% of medications proceed to phase 2.

Phase 2: Test for Effectiveness

Goal: The main goal of phase 2 is to assess the effectiveness of the treatment.
Sample Size: Phase 2 uses a larger group, such as 100 to 300 subjects.
Time and Cost: Phase 2 can take several months up to two years at a cost of $13 million.
Success Rate: About 50% of medications proceed to phase 3.

Phase 3: Test for Effectiveness with Different Populations

Goal: The main goal of phase 3 is to collect data about safety and effectiveness with a larger and more diverse group of subjects.
Sample Size: Phase 3 uses a much larger group, such as several hundred to 3000 subjects.
Time and Cost: Phase 3 can last years and can cost around $20 million.
Success Rate: About 60% of medications successfully complete phase 3.

Phase 4: Monitor Results from Widespread Use

Phase 4 occurs when the medication receives widespread use. The effectiveness of the drug is monitored along with its safety. The safety is reflected through observations of adverse reactions.

Checkpoints

- Before human testing begins with a medication, prior research must be conducted with animals and/or human cells. Based on the results, the FDA may or may not grant approval.
- An Institutional Review Board reviews the data, monitors the trials, and ensures that the trials are ethical and that they protect the rights and safety of the subjects.
- Upon completion of phase 3, FDA approval is required before the medication can be made available to the general public.
- In phase 4, the effectiveness and safety are monitored as the medication experiences widespread use.

1-3 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Magnet Treatment of Pain Researchers conducted a study to determine whether magnets are effective in treating back pain. Pain was measured using the visual analog scale, and the results given below are among the results obtained in the study (based on data from "Bipolar Permanent Magnets for the Treatment of Chronic Lower Back Pain: A Pilot Study," by Collacott, Zimmerman, White, and Rindone, *Journal of the American Medical Association*, Vol. 283, No. 10). Higher scores correspond to greater pain levels.

Reduction in Pain Level After Magnet Treatment: $n = 20$, $\bar{x} = 0.49$, $s = 0.96$

Reduction in Pain Level After Sham Treatment: $n = 20$, $\bar{x} = 0.44$, $s = 1.4$

Is this study an experiment or an observational study? Explain.

2. Blinding What does it mean when we say that the study cited in Exercise 1 was "double-blind"?

3. Replication In what specific way was replication applied in the study cited in Exercise 1?

4. Sampling Method The patients were recruited among those at a Veterans Affairs hospital. What type of sampling best describes the way in which the subjects were chosen: simple random sample, systematic sample, convenience sample, stratified sample, cluster sample? Does the method of sampling appear to adversely affect the quality of the results?

Exercises 5–8 refer to the study of an association between which ear is used for cell phone calls and whether the subject is left-handed or right-handed. The study is reported in "Hemispheric Dominance and Cell Phone Use," by Seidman et al.,* JAMA Otolaryngology—Head & Neck Surgery, *Vol. 139, No. 5. The study began with a survey e-mailed to 5000 people belonging to an otology online group, and 717 surveys were returned.* (Otology *relates to the ear and hearing.)

5. Sampling Method What type of sampling best describes the way in which the 717 subjects were chosen: simple random sample, systematic sample, convenience sample, stratified sample, cluster sample? Does the method of sampling appear to adversely affect the quality of the results?

6. Experiment or Observational Study Is the study an experiment or an observational study? Explain.

7. Response Rate What percentage of the 5000 surveys were returned? Does that response rate appear to be low? In general, what is a problem with a very low response rate?

8. Sampling Method Assume that the population consists of all students currently in your biostatistics class. Describe how to obtain a sample of six students so that the result is a sample of the given type.

a. Simple random sample

b. Systematic sample

c. Stratified sample

d. Cluster sample

e. Convenience sample

In Exercises 9–20, identify which of these types of sampling is used: random, systematic, convenience, stratified, or cluster.

9. Cormorant Density Cormorant bird population densities were studied by using the "line transect method" with aircraft observers flying along the shoreline of Lake Huron and collecting sample data at intervals of every 20 km (based on data from Brian S. Dorr et al., 2010, "Management Effects on Breeding and Foraging Numbers and Movements of Double-Crested Comorants in the Les Cheneaux Islands, Lake Huron, Michigan," *Journal of Great Lakes Research,* Vol. 36, No.)

10. Survey of Exercise One of the authors surveyed students by asking them how many minutes they exercise in a typical day.

11. Vaccine Survey A Pew Research Center poll used telephone calls to 12,648 randomly selected adults to ask them about their willingness to get vaccinations.

12. Reported and Observed Results A Harris Interactive study involved 1013 adults who were interviewed about washing their hands in restrooms and another 6336 adults who were observed in public restrooms.

13. Motor Vehicle Injuries A medical researcher randomly selected 25 hospitals and obtained data from all of the patients being treated for injuries resulting from motor vehicle crashes.

14. Acupuncture Study In a study of treatments for back pain, 641 subjects were randomly assigned to the four different treatment groups of individualized acupuncture, standardized acupuncture, simulated acupuncture, and usual care (based on data from "A Randomized Trial Comparing Acupuncture, Simulated Acupuncture, and Usual Care for Chronic Low Back Pain," by Cherkin et al., *Archives of Internal Medicine*, Vol. 169, No. 9).

15. Medical Research Researchers obtained treatment data from 75 patients in each of the three categories of major causes of death: heart disease, cancer, unintentional injury.

16. Deforestation Rates Satellites are used to collect sample data for estimating deforestation rates. The Forest Resources Assessment of the United Nations (UN) Food and Agriculture Organization uses a method of selecting a sample of a 10-km-wide square at every 1° intersection of latitude and longitude.

17. Testing Lipitor In a clinical trial of the cholesterol drug Lipitor (atorvastatin), subjects were partitioned into groups given a placebo or Lipitor doses of 10 mg, 20 mg, 40 mg, or 80 mg. The subjects were randomly assigned to the different treatment groups (based on data from Pfizer, Inc.).

18. Schools of Dentistry Five of the 67 accredited schools of dentistry were randomly selected and then all of the enrolled students were surveyed.

19. Survey of Nursing Students The Senior Coordinator at Duke University School of Nursing conducted a survey of all enrolled nursing students.

20. Drinking and Driving The Town of Poughkeepsie police implemented a sobriety checkpoint by stopping and testing every tenth car driver.

Critical Thinking: What's Wrong? ***In Exercises 21–28, determine whether the study is an experiment or an observational study, and then identify a major problem with the study.***

21. *People Magazine* Survey *People Magazine* invited visitors to its website to vote for the most beautiful person. At the urging of radio personality Howard Stern, 230,169 votes were cast for his sidekick Hank. Hank won handily.

22. Physicians' Health Study The Physicians' Health Study involved 22,071 male physicians. Based on random selections, 11,037 of them were treated with aspirin and the other 11,034 were given placebos. The study was stopped early because it became clear that aspirin reduced the risk of myocardial infarctions by a substantial amount.

23. Drinking and Driving A researcher for a consortium of insurance companies plans to test for the effects of drinking on driving ability by randomly selecting 1000 drivers and then randomly assigning them to two groups: One group of 500 will drive in New York City after no alcohol consumption, and the second group will drive in New York City after consuming three shots of Jim Beam bourbon whiskey.

24. Survey Too Long A veterinarian constructed a survey containing 50 questions. Because the survey was so long, only eight responses were received from the 250 surveys that were mailed.

25. Sleep Study When designing the study of a new treatment for insomnia in adults, researchers were criticized because their test subjects consisted of 75 college students. They then expanded the study so that 750 college students were given the treatment.

26. Atkins Weight Loss Program An independent researcher tested the effectiveness of the Atkins weight loss program by randomly selecting 1000 subjects using that program. Each of the subjects was called to report their weight before the diet and after the diet.

27. Survey of Hand Washing In a survey, 94% of 2800 respondents said that they wash their hands after using public restrooms (based on "A Nationwide Survey on the Hand Washing Behavior and Awareness," by Jeong et al., *Journal of Preventive Medicine and Public Health,* Vol. 40, No. 3).

28. Medications The Pharmaceutical Research and Manufacturers of America wants information about the consumption of various medications. An independent researcher conducts a survey by mailing 10,000 questionnaires to randomly selected adults in the United States, and the researcher receives 152 responses.

1-3 Beyond the Basics

In Exercises 29–32, indicate whether the observational study used is cross-sectional, retrospective, or prospective.

29. Nurses' Health Study II Phase II of the Nurses' Health Study was started in 1989 with 116,000 female registered nurses. The study is ongoing.

30. Heart Health Study Samples of subjects with and without heart disease were selected, and then researchers looked back in time to determine whether they took aspirin on a regular basis.

31. Marijuana Study Researchers from the National Institutes of Health want to determine the current rates of marijuana consumption among adults living in states that have legalized the use of marijuana. They conduct a survey of 500 adults in those states.

32. Framingham Heart Study The Framingham Heart Study was started in 1948 and is ongoing. Its focus is on heart disease.

In Exercises 33–36, identify which of these designs is most appropriate for the given experiment: completely randomized design, randomized block design, or matched pairs design.

33. Lunesta Lunesta is a drug designed to treat insomnia. In a clinical trial of Lunesta, amounts of sleep each night are measured before and after subjects have been treated with the drug.

34. Lipitor A clinical trial of Lipitor treatments is being planned to determine whether its effects on diastolic blood pressure are different for males and females.

35. West Nile Vaccine Currently, there is no approved vaccine for the prevention of infection by West Nile virus. A clinical trial of a possible vaccine is being planned to include subjects treated with the vaccine while other subjects are given a placebo.

36. HIV Vaccine The HIV Trials Network is conducting a study to test the effectiveness of two different experimental HIV vaccines. Subjects will consist of 80 pairs of twins. For each pair of twins, one of the subjects will be treated with the DNA vaccine and the other twin will be treated with the adenoviral vector vaccine.

37. Sample Design Literacy In "Cardiovascular Effects of Intravenous Triiodothyronine in Patients Undergoing Coronary Artery Bypass Graft Surgery" (*Journal of the American Medical Association,* Vol. 275, No. 9), the authors explain that patients were assigned to one of three groups: (1) a group treated with triiodothyronine, (2) a group treated with normal saline bolus and dopamine, and (3) a placebo group given normal saline. The authors summarize the sample design as a "prospective, randomized, double-blind, placebo-controlled trial." Describe the meaning of each of those terms in the context of this study.

38. Simple Random Sample vs. Random Sample Refer to the definition of *simple random sample* on page 46 and its accompanying definition of *random sample* enclosed within parentheses. Determine whether each of the following is a simple random sample and/or a random sample.

a. In a recent year, the following medical schools had the given enrollments: New York University (103), Columbia (140), and New York Medical College (219). The names of those three colleges are printed on three separate index cards, the cards are shuffled, and one card is drawn. The sample consists of the names of the medical students enrolled in the selected medical school.

b. For the three medical schools in part (a), the 462 names of the students are printed on 462 index cards, then the cards are shuffled. Forty different cards are selected, and the sample consists of the 40 selected medical students.

c. For the three medical schools in part (a), a sample is constructed by selecting the 40 youngest medical students.

1-4 Ethics in Statistics

The website www.pearsonglobaleditions.com includes a downloadable section that discusses ethical issues in statistics, including those related to data collection, analysis, and reporting.

Chapter Quick Quiz

1. Survey In a study of cell phone use and brain hemispheric dominance, an Internet survey was e-mailed to 5000 subjects randomly selected from an online group involved with the study of ears. If the 717 returned surveys are assigned identification codes consisting of the numbers 1–717, does it make sense to calculate the average (mean) of those numbers?

2. Survey Which of the following best describes the level of measurement of the numbers 1, 2, 3, . . . , 717 described in Exercise 1: nominal, ordinal, interval, ratio?

3. Survey In the same survey cited in Exercise 1, are the exact unrounded ages of the 717 subjects discrete data or continuous data?

4. Survey In the same survey cited in Exercise 1, are the exact unrounded ages of the 717 subjects quantitative data or categorical data?

5. Survey Which of the following best describes the level of measurement of the exact unrounded ages of the 717 survey subjects from Exercise 1: nominal, ordinal, interval, ratio?

6. Birth Weights For 100 randomly selected births from Bellevue Hospital Center, the birth weights are added and then divided by 100. The result is 3240 g. Is the value of 3240 g a statistic or a parameter?

7. Birth Weights Refer to the sample described in Exercise 6. Because Bellevue Hospital Center agreed to provide the 100 birth weights, does the sample of birth weights constitute a voluntary response sample?

8. Birth Weights Are the data described in Exercise 6 the result of an observational study or an experiment?

9. Physicians' Health Study In the Physicians' Health Study, some of the subjects were treated with aspirin while others were given a placebo. For the subjects in this experiment, what is *blinding*?

10. Sampling In a statistical study, which of the following types of samples is generally best: convenience sample, voluntary response sample, simple random sample, biased sample?

Review Exercises

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

2. Sampling Seventy-two percent of Americans squeeze their toothpaste tube from the top. This and other not-so-serious findings are included in *The First Really Important Survey of American Habits*. Those results are based on 7000 responses from the 25,000 questionnaires that were mailed.

a. What is wrong with this survey?

b. As stated, the value of 72% refers to all Americans, so is that 72% a statistic or a parameter? Explain.

c. Does the survey constitute an observational study or an experiment?

3. Sample Design Literacy In "High-Flow Oxygen for Treatment of Cluster Headache" (*Journal of the American Medical Association*, Vol. 302, No. 22), the authors explain that 150 patients were treated with oxygen, and 148 patients were given a placebo. The authors summarize the sample design as "randomized and double-blind." Describe the meaning of "randomized" and "double-blind" in the context of this study.

4. Divorces and Margarine One study showed that there is a very high correlation between the divorce rate in Maine and per capita consumption of margarine in the United States. Can we conclude that either one of those two variables is the cause of the other?

5. Sampling For each of the following, identify the term that best describes the type of sample: *systematic*, *convenience*, *stratified*, *cluster*, or *simple random sample*.

a. As Lipitor pills are being manufactured, a quality control plan is to select every 500th pill and test it to confirm that it contains 80 mg of atorvastatin.

b. To test for an age difference in the way that 20–49-year-olds and 50–79-year-olds purchase medications online, Gallup surveys 500 randomly selected 20–49-year-olds and 500 randomly selected 50–79-year-olds.

c. A list of all 1,736,997 adults in Manhattan is obtained; the list is numbered from 1 to 1,736,997; and then a computer is used to randomly generate 500 different numbers between 1 and 1,736,997. The sample consists of the adults corresponding to the selected numbers.

d. A statistics student creates a survey and presents it to fellow statistics students.

e. The Commissioner of Major League Baseball obtains a sample of results from drug tests of baseball players by randomly selecting one team from the American League and one team from the National League, and all players on the selected teams are tested.

6. What's Wrong? A survey sponsored by the American Laser Centers included responses from 575 adults, and 24% of the respondents said that the face is their favorite body part (based on data from *USA Today*). What is wrong with this survey?

7. State Populations Currently, California has the largest population with 39,613,493 residents, and Wyoming has the smallest population with 581,075 residents.

a. Are the population sizes of the different states discrete or continuous?

b. What is the level of measurement for the numbers of residents in the different states? (nominal, ordinal, interval, ratio)

c. What is wrong with surveying state residents by mailing questionnaires to 10,000 of them who are randomly selected?

continued

d. If we randomly select 50 hospitalized patients in each of the 50 states, what type of sample is obtained? (random, systematic, convenience, stratified, cluster)

e. If we randomly select two states and survey all of their hospitalized patients, what type of sample is obtained? (random, systematic, convenience, stratified, cluster)

8. Percentages This exercise is based on data from "Sustained Care Intervention and Postdischarge Smoking Cessation Among Hospitalized Adults," by Rigotti et al., *Journal of the American Medical Association,* Vol. 312, No. 7.

a. In a program designed to help patients stop smoking, 199 patients were given standard care and 62.8% of them were no longer smoking after one month. How many of these patients were no longer smoking after one month?

b. In a program designed to help patients stop smoking, 198 patients were given "sustained care," and 164 of them were no longer smoking after one month. What is the percentage of these patients who were no longer smoking after one month?

9. Types of Data In each of the following, identify the level of measurement of the sample data (nominal, ordinal, interval, ratio) and the type of sampling used to obtain the data (random, systematic, convenience, stratified, cluster).

a. At Albany Medical Center, every 10th newborn baby is selected and the body temperature is measured (degrees Fahrenheit).

b. In each of the 50 states, 40 hospitalized patients are randomly selected and the names of the hospitals are identified.

c. A pollster first selects a local hospital and then stops each discharged patient and asks them to rate the quality of the care that they received (on a scale of 1 star to 4 stars).

10. Statistical Significance and Practical Significance The Genetics and IVF Institute developed a procedure designed to increase the likelihood that a baby would be a male. In a clinical trial of their procedure, 239 males were born among 291 births. If the method has no effect, there is less than a 1% chance that such extreme results would occur. Does the procedure appear to have statistical significance? Does the procedure appear to have practical significance?

Cumulative Review Exercises

For Chapter 2 through Chapter 14, the Cumulative Review Exercises include topics from preceding chapters. For this chapter, we present a few calculator warm-up exercises, with expressions similar to those found throughout this text. Use your calculator to find the indicated values.

1. Cigarette Contents Listed below are the tar contents (mg) in 100 mm cigarettes of some popular brands. What value is obtained when those tar contents are added and the total is divided by the number of brands? (This result, called the *mean*, is discussed in Chapter 3.) What is notable about these values, and what does it tell us about how the tar contents were measured?

5 16 17 13 13 14 15 15

2. Streak of Males Jay and Kateri Schwandt had 13 children—all males! The probability that 13 randomly selected children are all males is found by evaluating 0.5^{13}. Find that value and round the result to six decimal places.

3. LeBron James LeBron James, one of the best professional basketball players ever, has a height of 203 cm. The expression below converts his height of 203 cm (or 6' 8") to a standardized score. Find this value and round the result to two decimal places. Such standardized scores are considered to be significantly high if they are greater than 2 or 3. Is the result significantly high?

$$\frac{203 - 176}{6}$$

4. Body Temperature The given expression is used for determining the likelihood that the average (mean) human body temperature is different from the value of 98.6°F that is commonly used. Find the given value and round the result to two decimal places.

$$\frac{98.2 - 98.6}{\frac{0.62}{\sqrt{106}}}$$

5. Determining Sample Size The given expression is used to determine the size of the sample necessary to estimate the proportion of college students who have the profound wisdom to take a statistics course. Find the value and round the result up to the next larger whole number.

$$\frac{1.95996^2 \cdot 0.25}{0.03^2}$$

6. Standard Deviation One way to get a very rough approximation of the value of a standard deviation of sample data is to find the range, then divide it by 4. The range is the difference between the highest sample value and the lowest sample value. In using this approach, what value is obtained from the sample data listed in Exercise 1 "Cigarette Contents"?

7. Standard Deviation The standard deviation is an extremely important concept introduced in Chapter 3. Using the sample data from Exercise 1 "Cigarette Contents," part of the calculation of the standard deviation is shown in the expression below. Evaluate this expression. (Fortunately, calculators and software are designed to automatically execute such expressions, so our future work with standard deviation will not be burdened with cumbersome calculations.)

$$\frac{(5 - 13.5)^2}{7}$$

8. Standard Deviation The given expression is used to compute the standard deviation of three randomly selected body temperatures. Perform the calculation and round the result to two decimal places.

$$\sqrt{\frac{(98.4 - 98.6)^2 + (98.6 - 98.6)^2 + (98.8 - 98.6)^2}{3 - 1}}$$

Scientific Notation. ***In Exercises 9–12, the given expressions are designed to yield results expressed in a form of scientific notation. For example, the calculator-displayed result of 1.23E5 can be expressed as 123,000, and the result of 1.23E-4 can be expressed as 0.000123. Perform the indicated operation and express the result as an ordinary number that is not in scientific notation.***

9. 0.3^6 **10.** 8^{12} **11.** 85^6 **12.** 0.2^{12}

Technology Project

Missing Data The focus of this project is to download a data set and manipulate it to work around the issue of missing data.

a. First, download Data Set 5 "Body Temperatures" in Appendix B from www.pearsonglobaleditions.com. Choose the download format that matches your technology.

b. Some statistical procedures, such as those involved with correlation and regression (discussed in later chapters), require data that consist of matched pairs of values, and those procedures ignore pairs in which at least one of the data values in a matched pair is missing. Assume that we want to conduct analyses for correlation and regression on the last two columns of data in Data Set 5: body temperatures measured at 8 AM on day 2 and again at 12 AM on day 2. For those last two columns, identify the rows with at least one missing value. Note that in some technologies, such as TI-83/84 Plus calculators, missing data must be represented by a constant such as −9 or 999.

continued

c. Here are two different strategies for reconfiguring the data set to work around the missing data in the last two columns (assuming that we need matched pairs of data with no missing values):

i. Manual Deletion Highlight rows with at least one missing value in the last two columns, then delete those rows. This can be tedious if there are many rows with missing data and those rows are interspersed throughout instead of being adjacent rows.

ii. Sort Most technologies have a Sort feature that allows you to rearrange all rows using one particular column as the basis for sorting (TI-83/84 Plus calculators *do not* have this type of sort feature). The result is that all rows remain the same but they are in a different order. First use the technology's Sort feature to rearrange all rows using the "8 AM day 2" column as the basis for sorting (so that all missing values in the "8 AM day 2" column are at the beginning); then highlight and delete all of those rows with missing values in the "8 AM day 2" column. Next, use the technology's Sort feature to rearrange all rows using the "12 AM day 2" column as the basis for sorting (so that all missing values in the "12 AM day 2" column are at the beginning); then highlight and delete all of those rows with missing values in the "12 AM day 2" column. The remaining rows will include matched pairs of body temperatures, and those rows will be suitable for analyses such as correlation and regression. Print the resulting reconfigured data set that now has no missing data.

Big (or Very Large) Data Project

Refer to Data Set 24 "Births in New York" in Appendix B, with records from 465,506 births. Find the number of males and find the number of females. What percentage of births are males? How does that result compare to the value of 51.2%, which is believed to be the proportion of male births in the population? Based on these results, does it appear that the sample accurately reflects the population?

FROM DATA TO DECISION

Critical Thinking: Is Being a Student More Dangerous or Unhealthy Than Being a Carpenter?

In 1835, the Swiss physician H. C. Lombard compiled longevity data for different professions. Over a period of 50 years, he collected 8496 death certificates from Geneva that included name, age at death, and profession. He calculated the average (mean) length of life for different professions, and some of his results are listed on the right. The table shows that among the five professions listed, being a student is most dangerous or unhealthy with an average (mean) age of death of only 20.2 years. Similar results would be obtained if the same data were collected today. (See "A Selection of Selection Anomalies" by Wainer, Palmer, and Bradlow in *Chance*, Vol. 11, No. 2.) Using common sense, the most indispensable tool for statistical thinking, it should not seem reasonable that being a student is substantially more dangerous or unhealthy than being a carpenter or barber.

Profession	Average (Mean) Age at Death
Carpenters	55.1
Bakers	49.8
Barbers	47.4
Shoemakers	54.2
Students	20.2

Analysis

1. Consider the population of students. About how many people do you know who are students aged 50 years or older?

2. Estimate the age at which people stop being students.

3. Why is it that being a student is actually no more dangerous or unhealthy than being a carpenter or barber?

4. How are samples of carpenters, bakers, barbers, shoemakers, and students fundamentally different?

Cooperative Group Activities

1. In-class activity For each student in the class, collect the number of children in their immediate family, including the student. Combine the results and find the average (mean). Compare the result to the value of 1.9 reported in the *World Factbook* by the Central Intelligence Agency. Use a class discussion to explain the discrepancy.

2. In-class activity Working in groups of three or four, design an experiment to determine whether pulse rates of college students are the same while the students are standing and sitting. Conduct the experiment and collect the data. Save the data so that they can be analyzed with methods presented in the following chapters.

3. Out-of-class activity Conduct a survey using the following questions. Try to detect the effect of the different versions of the fifth question.

1. What is the color of your eyes?

2. Do you exercise vigorously (such as running, swimming, cycling, playing basketball, and so on) for at least 20 minutes a week?

3. How many cigarettes have you smoked in the past 24 hours?

4. Are you left-handed, right-handed, or ambidextrous?

5. *For the fifth question, alternate between the following two versions:*

Version 1: Do you agree with a requirement that all children must receive the measles, mumps, rubella (MMR) vaccine before being allowed to enroll in public schools, with no regard or exception for religious or personal beliefs?

Version 2: Do you agree or disagree with a requirement that for the safety of all, every child must receive the measles, mumps, rubella (MMR) vaccine before being allowed to enroll in public schools?

4. In-class activity Identify problems with a mailing from *Consumer Reports* magazine that included an annual questionnaire about cars and other consumer products. Also included were a request for a voluntary contribution of money and a voting ballot for the board of directors. Responses were to be mailed back in envelopes that required postage stamps.

5. Out-of-class activity Find a report of a survey that used a voluntary response sample. Describe how it is quite possible that the results do not accurately reflect the population.

6. Out-of-class activity Find a professional journal with an article that includes a statistical analysis of an experiment. Describe and comment on the design of the experiment. Identify one particular issue addressed by the study, and determine whether the results were found to be statistically significant. Determine whether those same results have practical significance.

2 Exploring Data with Tables and Graphs

Credit: Josh McCulloch/All Canada Photos/Alamy Stock Photo

CHAPTER PROBLEM

Does Exposure to Lead Affect IQ Scores?

Data Set 11 "IQ and Lead" in Appendix B includes the full IQ scores from three groups of children who lived near a lead smelter. The children in Group 1 had *low* levels of measured lead in their blood (with blood levels lower than 40 μg/100 mL in each of two years). Group 2 had *medium* levels of measured lead in their blood (with blood levels of at least 40 μg/100 mL in exactly one of two years). Group 3 had *high* levels of measured lead in their blood (with blood levels of at least 40 μg/100 mL in each of two years).

Let's consider the measured full IQ scores from Group 1 (low lead level) and Group 3 (high lead level), as listed in Table 2-1. Only an exceptionally rare person could look at

both lists of IQ scores and form meaningful conclusions. Almost all of us must work at describing, exploring, and comparing the two sets of data. In this chapter, we present methods that focus on summarizing the data and using graphs that enable us to understand important characteristics of the data, especially the *distribution* of the data. These methods will help us compare sets of data so that we can determine whether the IQ scores of the *low* lead group are somehow different from the IQ scores of the *high* lead group. Such comparisons will be helpful as we try to address this important and key issue: Does exposure to lead have an effect on IQ score?

TABLE 2-1 Full IQ Scores of the Low Lead Group and the High Lead Group

Low Lead Level (Group 1)															
70	85	86	76	84	96	94	56	115	97	77	128	99	80	118	86
141	88	96	96	107	86	80	107	101	91	125	96	99	99	115	106
105	96	50	99	85	88	120	93	87	98	78	100	105	87	94	89
80	111	104	85	94	75	73	76	107	88	89	96	72	97	76	107
104	85	76	95	86	89	76	96	101	108	102	77	74	92		
High Lead Level (Group 3)															
82	93	85	75	85	80	101	89	80	94	88	104	88	88	83	104
96	76	80	79	75											

CHAPTER OBJECTIVES

This chapter and the following chapter focus on important characteristics of data, including the following:

Important Characteristics of Data

1. **Center:** A representative value that shows us where the middle of the data set is located.
2. **Variation:** A measure of the amount that the data values vary.
3. **Distribution:** The nature or shape of the spread of the data over the range of values (such as bell-shaped).
4. **Outliers:** Sample values that lie very far away from the vast majority of the other sample values.
5. **Time:** Any change in the characteristics of the data over time.

HINT Remember the sentence "**C**omputer **V**iruses **D**estroy **O**r **T**erminate" to recall the first letters of the characteristics (CVDOT).

This chapter provides tools that enable us to gain insight into data by organizing, summarizing, and representing them in ways that enable us to see important characteristics of the data. Here are the chapter objectives:

2-1 Frequency Distributions for Organizing and Summarizing Data

- Develop an ability to summarize data in the format of a frequency distribution and a relative frequency distribution.
- For a frequency distribution, identify values of class width, class midpoint, class limits, and class boundaries.

2-2 Histograms

- Develop the ability to picture the distribution of data in the format of a histogram or relative frequency histogram.
- Examine a histogram and identify common distributions, including a uniform distribution and a normal distribution.

2-3 Graphs That Enlighten and Graphs That Deceive

- Develop an ability to graph data using a dotplot, stemplot, time-series graph, Pareto chart, pie chart, and frequency polygon.
- Determine when a graph is deceptive through the use of a nonzero axis or a pictograph that uses an object of area or volume for one-dimensional data.

2-4 Scatterplots, Correlation, and Regression

- Develop an ability to construct a scatterplot of paired data.
- Analyze a scatterplot to determine whether there appears to be a correlation between two variables.

2-1 Frequency Distributions for Organizing and Summarizing Data

Key Concept When working with large data sets, a *frequency distribution* (or *frequency table*) is often helpful in organizing and summarizing data. A frequency distribution helps us to understand the nature of the *distribution* of a data set. Also, construction of a frequency distribution is often the first step in constructing a histogram, which is a graph used to help visualize the distribution of data.

DEFINITION

A **frequency distribution** (or **frequency table**) shows how data are partitioned among several categories (or *classes*) by listing the categories along with the number (frequency) of data values in each of them.

TABLE 2-2 IQ Scores of the Low Lead Group

IQ Score	Frequency
50–69	2
70–89	33
90–109	35
110–129	7
130–149	1

Consider the IQ scores of the low lead group listed in Table 2-1. Table 2-2 is a frequency distribution summarizing those IQ scores for the low lead group. The **frequency** for a particular class is the number of original values that fall into that class. For example, the first class in Table 2-2 has a frequency of 2, so 2 of the IQ scores are between 50 and 69 inclusive.

DEFINITIONS

Lower class limits are the smallest numbers that can belong to each of the different classes. (Table 2-2 has lower class limits of 50, 70, 90, 110, and 130.)

Upper class limits are the largest numbers that can belong to each of the different classes. (Table 2-2 has upper class limits of 69, 89, 109, 129, and 149.)

Class boundaries are the numbers used to separate the classes, but without the gaps created by class limits. From Figure 2-1 we see that the values of 69.5, 89.5, 109.5, and 129.5 are in the centers of those gaps. Following the pattern of those class boundaries, we see that the lowest class boundary is 49.5 and the highest class boundary is 149.5. Thus the complete list of class boundaries is 49.5, 69.5, 89.5, 109.5, 129.5, and 149.5.

Class midpoints are the values in the middle of the classes. Table 2-2 has class midpoints of 59.5, 79.5, 99.5, 119.5, and 139.5. Each class midpoint is computed by adding the lower class limit to the upper class limit and dividing the sum by 2.

Class width is the difference between two consecutive lower class limits (or two consecutive lower class boundaries) in a frequency distribution. Table 2-2 uses a class width of 20. (The first two lower class boundaries are 50 and 70, and their difference is 20. *Caution:* The class width in Table 2-2 is 20, not 19.)

CAUTION Finding the correct class width can be tricky. For class width, don't make the most common mistake of using the difference between a lower class limit and an upper class limit. See Table 2-2 and note that the class width is 20, not 19.

CAUTION For class boundaries, remember that they split the difference between the end of one class and the beginning of the next class, as shown in Figure 2-1.

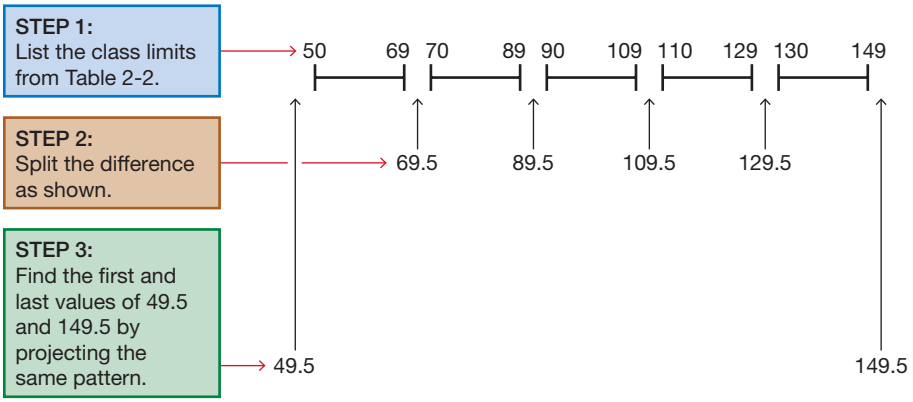


FIGURE 2-1 Finding Class Boundaries from Class Limits in Table 2-2

Procedure for Constructing a Frequency Distribution

We construct frequency distributions to (1) summarize large data sets, (2) see the distribution of the data, (3) identify outliers, and (4) have a basis for constructing graphs (such as *histograms,* introduced in Section 2-2). Technology can generate frequency distributions, but here are the steps for manually constructing them:

1. Select the number of classes, usually between 5 and 20. The number of classes might be affected by the convenience of using round numbers. (According to "Sturges' guideline," the ideal number of classes for a frequency distribution can be approximated by $1 + (\log n)/(\log 2)$, where n is the number of data values. We don't use this guideline in this text.)
2. Calculate the class width.

$$\text{Class width} \approx \frac{(\text{maximum data value}) - (\text{minimum data value})}{\text{number of classes}}$$

   Round this result to get a convenient number. (It's usually best to round *up*.) Using a specific number of classes is not too important, and it's usually wise to change the number of classes so that they use convenient values for the class limits.
3. Choose the value for the first lower class limit by using either the minimum value or a convenient value below the minimum.

continued

No Phones or Bathtubs

Credit: Gallofoto/ Shutterstock

Many statistical analyses must consider changing characteristics of populations over time. Here are some observations of life in the United States from 100 years ago:

- 8% of homes had a telephone.
- 14% of homes had a bathtub.
- The mean life expectancy was 47 years.
- The mean hourly wage was 22 cents.
- There were approximately 230 annual murders in the entire United States.

Although these observations from 100 years ago are in stark contrast to the United States of today, statistical analyses should always consider changing population characteristics that might have more subtle effects.

4. Using the first lower class limit and the class width, list the other lower class limits. (Do this by adding the class width to the first lower class limit to get the second lower class limit. Add the class width to the second lower class limit to get the third lower class limit, and so on.)
5. List the lower class limits in a vertical column and then determine and enter the upper class limits.
6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to find the total frequency for each class.

When constructing a frequency distribution, be sure the classes do not overlap. Each of the original values must belong to exactly one class. Include all classes, even those with a frequency of zero. Try to use the same width for all classes, although it is sometimes impossible to avoid open-ended intervals, such as "65 years or older."

CP **EXAMPLE 1** **IQ Scores of Low Lead Group**

Using the IQ scores of the low lead group in Table 2-1, follow the above procedure to construct the frequency distribution shown in Table 2-2. Use five classes.

SOLUTION

Step 1: Select 5 as the number of desired classes.

Step 2: Calculate the class width as shown below. Note that we round 18.2 up to 20, which is a much more convenient number.

$$\text{Class width} \approx \frac{(\text{maximum data value}) - (\text{minimum data value})}{\text{number of classes}}$$

$$= \frac{141 - 50}{5} = 18.2 \approx 20 (\text{rounded up to a convenient number})$$

Step 3: The minimum data value is 50 and it is a convenient starting point, so use 50 as the first lower class limit. (If the minimum value had been 52 or 53, we would have rounded down to the more convenient starting point of 50.)

Step 4: Add the class width of 20 to 50 to get the second lower class limit of 70. Continue to add the class width of 20 until we have five lower class limits. The lower class limits are therefore 50, 70, 90, 110, and 130.

50–
70–
90–
110–
130–

Step 5: List the lower class limits vertically, as shown in the margin. From this list, we identify the corresponding upper class limits as 69, 89, 109, 129, and 149.

Step 6: Enter a tally mark for each data value in the appropriate class. Then add the tally marks to find the frequencies shown in Table 2-2.

YOUR TURN. Do Exercise 13 "Pulse Rates of Males."

Categorical Data So far we have discussed frequency distributions using only quantitative data sets, but frequency distributions can also be used to summarize categorical (or qualitative or attribute) data, as illustrated in Example 2.

EXAMPLE 2 **Causes of Fatal Plane Crashes**

Table 2-3 on the next page lists data for the causes of fatal plane crashes from 1960 until a recent year. The causes are categorical data at the nominal level of measurement, but we can create the frequency distribution as shown. We can see that pilot error is the major cause of fatal plane crashes. Such information is helpful

to regulatory agencies, such as the Federal Aviation Administration, as they develop strategies for reducing such crashes.

TABLE 2-3 Causes of Fatal Plane Crashes

Cause	Frequency
Pilot Error	640
Mechanical	195
Sabotage	95
Weather	63
Other	111

YOUR TURN. Do Exercise 27 "Causes of Death."

Go Figure

14: The number of different shapes of human noses, from a study by Abrahim Tamir that was published in the *Journal of Craniofacial Surgery.*

Relative Frequency Distribution

A variation of the basic frequency distribution is a **relative frequency distribution** or **percentage frequency distribution**, in which each class frequency is replaced by a relative frequency (or proportion) or a percentage. In this text we use the term "relative frequency distribution" whether we use relative frequencies or percentages. Relative frequencies and percentages are calculated as follows:

$$\text{Relative frequency for a class} = \frac{\text{frequency for a class}}{\text{sum of all frequencies}}$$

$$\text{Percentage for a class} = \frac{\text{frequency for a class}}{\text{sum of all frequencies}} \times 100\%$$

Table 2-4 is an example of a relative frequency distribution. It is a variation of Table 2-2 in which each class frequency is replaced by the corresponding percentage value. Because there are 78 data values, we divide each class frequency by 78 and then multiply by 100%. The first class of Table 2-2 has a frequency of 2, so we divide 2 by 78 to get 0.0256, and then multiply by 100% to get 2.56%, which we round to 2.6%. The sum of the percentages should be 100%, with a small discrepancy allowed for rounding errors, so a sum such as 99% or 101% is acceptable. The sum of the percentages in Table 2-4 is 100.1%.

TABLE 2-4 Relative Frequency Distribution of IQ Scores of Low Lead Group

IQ Score	Frequency
50–69	2.6%
70–89	42.3%
90–109	44.9%
110–129	9.0%
130–149	1.3%

The sum of the percentages in a relative frequency distribution must be very close to 100% (with a little room for rounding errors).

Comparisons Example 3 illustrates this principle:

Combining two or more relative frequency distributions in one table makes comparisons of different data sets much easier.

EXAMPLE 3 Comparing IQ Scores of the Low Lead Group and the High Lead Group

Now let's compare the IQ scores from the low lead group and the high lead group (listed in Table 2-1). Table 2-5 shows the relative frequency distributions for the data from Table 2-1. By comparing the relative frequencies in Table 2-5, we see that there are major differences. The IQ scores of the high lead group appear to vary much less than those of the low lead group. The high lead group has no IQ scores in the low IQ class of 50–69, nor does it have any IQ scores above 109.

continued

Growth Charts Updated

Credit: Valua Vitaly/ Shutterstock

Pediatricians typically use standardized growth charts to compare their patient's weight and height to a sample of other children. Children are considered to be in the normal range if their weight and height fall between the 5th and 95th percentiles. If they fall outside that range, they are often given tests to ensure that there are no serious medical problems. Pediatricians became increasingly aware of a major problem with the charts: Because they were based on children living between 1929 and 1975, the growth charts had become inaccurate. To rectify this problem, the charts were updated in 2000 to reflect the current measurements of millions of children. The weights and heights of children are good examples of populations that change over time. This is the reason for including changing characteristics of data over time as an important consideration for a population.

TABLE 2-5 IQ Scores of Low Lead Group and High Lead Group

IQ Score	Low Lead Group	High Lead Group
50–69	2.6%	
70–89	42.3%	71.4%
90–109	44.9%	28.6%
110–129	9.0%	
130–149	1.3%	

YOUR TURN. Do Exercise 2 "Relative Frequency Distribution."

Cumulative Frequency Distribution

Another variation of a frequency distribution is a **cumulative frequency distribution**, in which the frequency for each class is the sum of the frequencies for that class and all previous classes. Table 2-6 is a cumulative frequency distribution based on Table 2-2. Using the original frequencies of 2, 33, 35, 7, and 1 from Table 2-2, we add $2 + 33$ to get the second cumulative frequency of 35; then we add $2 + 33 + 35$ to get the third; and so on. In Table 2-6, note that in addition to the use of cumulative frequencies, the class limits are replaced by "less than" expressions that describe the new ranges of values.

TABLE 2-6 Cumulative Frequency Distribution of IQ Scores of Low Lead Group

IQ Score	Cumulative Frequency
Less than 70	2
Less than 90	35
Less than 110	70
Less than 130	77
Less than 150	78

Critical Thinking: Using Frequency Distributions to Understand Data

At the beginning of this section we noted that a frequency distribution can help us understand the *distribution* of a data set, which is the nature or shape of the spread of the data over the range of values (such as bell-shaped). In statistics we are often interested in determining whether the data have a *normal distribution*. (Normal distributions are discussed extensively in Chapter 6.) Data that have an approximately normal distribution are characterized by a frequency distribution with the following features.

Normal Distribution

1. The frequencies start low, then increase to one or two high frequencies, and then decrease to a low frequency.
2. The distribution is approximately symmetric: Frequencies that precede the maximum frequency should be roughly a mirror image of those that follow the maximum frequency.

Table 2-7 satisfies these two conditions. The frequencies start low, increase to the maximum of 18, and then decrease to a low frequency. Also, the frequencies of 2, 4, and 10 that precede the maximum are a mirror image of the frequencies 10, 4, and 2 that follow the maximum. Real data sets are usually not so perfect as Table 2-7, and judgment must be used to determine whether the distribution comes close enough to satisfying those two conditions. (There are more objective procedures described later.)

TABLE 2-7 Frequency Distribution Showing a Normal Distribution

Time	Frequency	Normal Distribution
0–14	2	← Frequencies start low, . . .
15–29	4	
30–44	10	
45–59	18	← Increase to this maximum, . . .
60–74	10	
75–89	4	
90–104	2	← Decrease to become low again.

Let's consider the IQ scores from the low lead group in Table 2-1. The frequency distribution of Table 2-2 shows the frequencies of 2, 33, 35, 7, and 1. Those frequencies start low, they increase to a maximum, and then they decrease. Although being far from perfect, those frequencies do suggest that the low lead group has IQ scores with a distribution that is approximately normal. More about normality later.

Analysis of Last Digits Example 4 illustrates this principle:

> **Frequencies of last digits sometimes reveal how the data were collected or measured.**

EXAMPLE 4 **Exploring Data: How Were the Pulse Rates Measured?**

Upon examination of measured pulse rates from 2219 adults included in the National Health and Examination Survey, the last digits of the recorded pulse rates are identified and the frequency distribution for those last digits is as shown in Table 2-8. Here is an important observation about those last digits: All of the last digits are *even* numbers. If the pulse rates were counted for 1 full minute, there would surely be a large number of them ending with an *odd* digit. So what happened?

One reasonable explanation is that even though the pulse rates are the number of heartbeats in 1 minute, they were likely counted for 30 seconds and the number of beats was doubled. (The original pulse rates are not all multiples of 4, so we can rule out a procedure of counting for 15 seconds and then multiplying by 4.) Analysis of these last digits reveals to us the method used to obtain these data.

In many surveys, we can determine that surveyed subjects were asked to *report* some values, such as their heights or weights, because disproportionately many values end in 0 or 5. This is a strong clue that the respondent is rounding instead of being physically measured. Fascinating stuff!

YOUR TURN. Do Exercise 21 "Analysis of Last Digits."

TABLE 2-8 Last Digits of Pulse Rates from the National Health and Examination Survey

Last Digit of Pulse Rate	Frequency
0	455
1	0
2	461
3	0
4	479
5	0
6	425
7	0
8	399
9	0

Gaps Example 5 illustrates this principle:

> **The presence of gaps can suggest that the data are from two or more different populations.**

The converse of this principle is not true, because data from different populations do not necessarily result in gaps.

TABLE 2-9 Body Temperatures of Adults

Temperature (°F)	Frequency	
96.0–96.9	2	Healthy
97.0–97.9	30	
98.0–98.9	66	
99.0–99.9	8	
100.0–100.9	0	Fever
101.0–101.9	5	
102.0–102.9	44	
103.0–103.9	53	
104.0–104.9	4	

EXAMPLE 5 What Does a Gap Tell Us?

Table 2-9 is a frequency distribution of the body temperatures (°F) of randomly selected adults. Examination of the frequencies reveals a *gap* between the lowest temperatures and the highest temperatures. This suggests that we have two different populations: About half of the temperatures appear to be from healthy adults, while the other half appear to be from unhealthy adults who have fevers.

YOUR TURN. Do Exercise 21 "Analysis of Last Digits" and determine whether there is a gap. If so, what is a reasonable explanation for it?

TECH CENTER

Frequency Distributions

Access tech supplements, videos, and data sets at **www.pearsonglobaleditions.com**

Frequency distributions are often easy to obtain after generating a histogram, as described in Section 2-2. With Statdisk, for example, we can generate a histogram with a desired starting point and class width, then move the cursor over the histogram to see the frequency for each class. If histograms are not used, "sort" the data (arrange them in order) so that we can see the maximum data value and the minimum data value used for computing the class width. Once the class limits are established, it is easy to find the frequency for each class using sorted data. Every statistics software package includes a sort feature.

2-1 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

Table for Exercises 1, 2, and 3

Nicotine (mg)	Frequency
1.0–1.1	14
1.2–1.3	4
1.4–1.5	3
1.6–1.7	3
1.8–1.9	1

Table for Exercise 4

Height (cm)	Relative Frequency
130–144	23%
145–159	25%
160–174	22%
175–189	27%
190–204	28%

1. Nicotine in Cigarettes Refer to the accompanying frequency distribution summarizing the amounts of nicotine (mg) in king-size cigarettes (from Data Set 22 "Cigarette Contents" in Appendix B). Does the distribution appear to be a normal distribution? Why or why not?

2. Relative Frequency Distribution Use percentages to construct the relative frequency distribution that corresponds to the accompanying frequency distribution.

3. Class Limits For the accompanying frequency distribution, what would be wrong with using class limits of 1.0–1.2, 1.2–1.4, 1.4–1.6, 1.6–1.8, 1.8–2.0?

4. What's Wrong? Heights of adult males are known to have a normal distribution, as described in this section. A researcher claims to have randomly selected adult males and measured their heights with the resulting relative frequency distribution as shown in the margin. Identify two major flaws with these results.

In Exercises 5–8, identify the class width, class midpoints, and class boundaries for the given frequency distribution. Also identify the number of individuals included in the summary. The frequency distributions are based on real data from Appendix B.

5.

White Blood Cell Count of Females	Frequency
2.0–3.9	7
4.0–5.9	56
6.0–7.9	46
8.0–9.9	29
10.0–11.9	8
12.0–13.9	0
14.0–15.9	1

6.

White Blood Cell Count of Males	Frequency
2.0–3.9	9
4.0–5.9	60
6.0 –7.9	50
8.0–9.9	29
10.0–11.9	3
12.0 –13.9	2

7.

Weight (in kg) of Females	Frequency
30–49	8
50–69	53
70–89	49
90–109	28
110–129	5
130–149	3
150–169	1

8.

Weight (in kg) of Males	Frequency
50–69	30
70–89	66
90–109	43
110–129	12
130–149	2

Normal Distributions. ***In Exercises 9 and 10, using a loose interpretation of the criteria for determining whether a frequency distribution is approximately a normal distribution, determine whether the given frequency distribution is approximately a normal distribution. Give a brief explanation.***

9. White Blood Cell Counts of Females Refer to the frequency distribution from Exercise 5.

10. White Blood Cell Counts of Males Refer to the frequency distribution from Exercise 6.

11. Normal distribution Refer to the frequency distribution given in Exercise 7 and ignore the given frequencies. Assume that the first three frequencies are 7, 9, and 20, respectively. Assuming that the distribution of the 147 sample values is a normal distribution, identify the remaining four frequencies.

12. Normal Distribution Refer to the frequency distribution given in Exercise 8 and determine whether it appears to be a normal distribution. Explain.

Constructing Frequency Distributions. ***In Exercises 13–20, use the indicated data to construct the frequency distribution. (The data for Exercises 17–20 can be downloaded at www.pearsonglobaleditions.com.)***

13. Pulse Rates of Males Listed below are some of the pulse rates (beats per min) of males from Data Set 1 "Body Data" in Appendix B. Begin with a lower class limit of 40 and use a class width of 10. Do the pulse rates appear to be from a normal distribution?

84 74 50 60 52 62 52 76 52 62 72 60 64 78 82 66 66 96 42 86
72 64 72 72 54 66 56 80 72 64 64 96 58 66 58 58 68 70 84 58

14. Pulse Rates of Females Listed below are some of the pulse rates (beats per min) of females from Data Set 1 "Body Data" in Appendix B. Begin with a lower class limit of 30 and use a class width of 10. Do the pulse rates appear to be from a normal distribution?

80 94 58 66 56 82 78 86 88 56 36 66 84 76 78 64 66 78 60 64
84 82 70 74 86 90 88 90 90 94 68 90 82 80 74 56 100 74 76 76

15. Lead and IQ Listed below are some of the verbal IQ scores from the low lead group in Data Set 11 "IQ and Lead" in Appendix B. Begin with a lower class limit of 50 and use a class width of 10. Do these values appear to be from a population having a normal distribution?

61 82 70 72 72 95 89 57 116 95 82 116 99 74 100
72 126 80 86 94 100 72 63 101 85 85 124 105 81 87

16. Lead and IQ Listed below are the verbal IQ scores from the high lead group in Data Set 11 "IQ and Lead" in Appendix B. Begin with a lower class limit of 60 and use a class width of 10. Do these values appear to be from a population having a normal distribution?

75 87 76 76 76 92 91 82 80 91 81
97 76 85 80 100 86 79 70 84 69

17. Pulse Rates of Males Repeat Exercise 13 using all of the pulse rates of males in Data Set 1 "Body Data" in Appendix B. Begin with a lower class limit of 40 and use a class width of 10. Does the resulting distribution appear to be dramatically different from the one found in Exercise 13?

18. Pulse Rates of Females Repeat Exercise 14 using all of the pulse rates of females in Data Set 1 "Body Data" in Appendix B. Begin with a lower class limit of 30 and use a class width of 10. Does the resulting distribution appear to be dramatically different from the one found in Exercise 14?

19. Freshman 15 Refer to Data Set 13"Freshman 15" in Appendix B and use the weights (kg) of males in September of their freshman year. Begin with a lower class limit of 50 kg and use a class width of 10 kg.

20. Freshman 15 Repeat the preceding exercise using the weights (kg) of males in April. Use the same class width of 10 and begin with a lower class limit of 50 kg. Compare the result to the frequency distribution from the preceding exercise. Does it appear that males gain 15 lb (or 6.8 kg) during their freshman year?

21. Analysis of Last Digits Heights (in.) of statistics students were obtained by one of the authors as part of an experiment conducted for class. The last digits of those heights are listed below. Construct a frequency distribution with 10 classes. Based on the distribution, do the heights appear to be reported or actually measured? Does there appear to be a gap in the frequencies and, if so, how might that gap be explained? What do you know about the accuracy of the results?

0	0	0	0	0	0	0	0	0	1	1	2	3	3	3	4	5	5	5
5	5	5	5	5	5	5	5	5	5	5	5	6	6	8	8	8	9	

22. Analysis of Last Digits Weights of respondents were recorded as part of the California Health Interview Survey. The last digits of weights from 50 randomly selected respondents are listed below. Construct a frequency distribution with 10 classes. Based on the distribution, do the weights appear to be reported or actually measured? Does there appear to be a gap in the frequencies and, if so, how might that gap be explained? What do you know about the accuracy of the results?

5	0	1	0	2	0	5	0	5	0	3	8	5	0	5	0	5	6	0	0	0	0	0	0	8
5	5	0	4	5	0	0	4	0	0	0	0	0	8	0	9	5	3	0	5	0	0	0	5	8

Relative Frequencies for Comparisons. ***In Exercises 23 and 24, construct the relative frequency distributions and answer the given questions.***

23. White Blood Cell Counts Construct one table (similar to Table 2-5 on page 68) that includes relative frequencies from Exercises 5 and 6. Then compare the two distributions. Are there any notable differences?

24. Weights Construct one table (similar to Table 2-5 on page 68) that includes relative frequencies based on the frequency distributions from Exercises 7 and 8, and then compare them. Are there notable differences?

Cumulative Frequency Distributions. ***In Exercises 25 and 26, construct the cumulative frequency distribution that corresponds to the frequency distribution in the exercise indicated.***

25. Exercise 5 "White Blood Cell Count of Females"

26. Exercise 6 "White Blood Cell Count of Males"

Categorical Data. ***In Exercises 27 and 28, use the given categorical data to construct the relative frequency distribution.***

27. Causes of Death Here are the numbers of deaths in the United States for the five leading causes during a recent year: heart disease (659,041), cancer (599,601), accidents (173,040), chronic lower respiratory disease (156,979), and stroke (150,005). (In 2020 and 2021 only, COVID was the third leading cause of death.) What do we know about the other causes of death not listed here?

28. Death Caused by Drowning in Males According to WHO, deaths caused by drowning in males globally, in seven different age groups, below 5, 5–14, 15–29, 30–49, 50–59, 60–69, and 70 and above, are listed as 26,843, 24,012, 29,557, 33,920, 15,368, 13,516, and 18,170, respectively. Does it appear that such deaths occur with equal frequency in the different age groups in males?

2-1 Beyond the Basics

29. Cumulative Frequency Distribution The accompanying cumulative frequency distribution summarizes the systolic blood pressure measurements (mmHg) of females from Data Set 1 "Body Data" in Appendix B. Construct the corresponding frequency distribution, and then determine whether the measurements appear to be from a normal distribution.

Systolic Blood Pressure (mmHg) of Females	Cumulative Frequency
Less than 100	8
Less than 120	70
Less than 140	123
Less than 160	144
Less than 180	146
Less than 200	147

2-2 Histograms

PART 1 Basic Concepts of Histograms

Key Concept While a frequency distribution is a useful tool for summarizing data and investigating the distribution of data, an even better tool is a *histogram*, which is a graph that is easier to understand and interpret than a table of numbers.

DEFINITION

A **histogram** is a graph consisting of bars of equal width drawn adjacent to each other (unless there are gaps in the data). The horizontal scale represents classes of quantitative data values, and the vertical scale represents frequencies. The heights of the bars correspond to frequency values.

Important Uses of a Histogram

- Visually displays the shape of the *distribution* of the data
- Shows the location of the *center* of the data
- Shows the *spread* of the data
- Identifies *outliers*

A histogram is basically a graph of a frequency distribution. For example, Figure 2-2 on the next page shows the Minitab-generated histogram corresponding to the frequency distribution given in Table 2-2 on page 64.

Class frequencies should be used for the vertical scale and that scale should be labeled as in Figure 2-2. There is no universal agreement on the procedure for selecting which values are used for the bar locations along the horizontal scale, but it is common to use class midpoints (as shown in Figure 2-2) or class boundaries or class limits or something else. It is often easier for us mere mortals to use class midpoints for the horizontal scale. Histograms can usually be generated using technology.

Relative Frequency Histogram

A **relative frequency histogram** has the same shape and horizontal scale as a histogram, but the vertical scale uses relative frequencies (as percentages or proportions) instead of actual frequencies. Figure 2-3 is the relative frequency histogram corresponding to Figure 2-2.

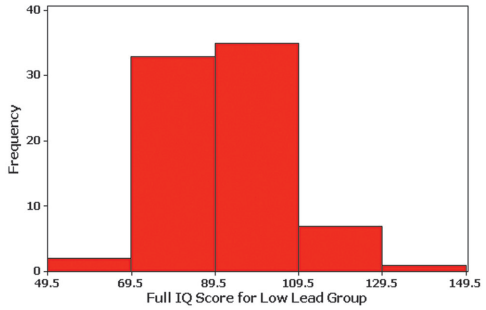


FIGURE 2-2 Histogram

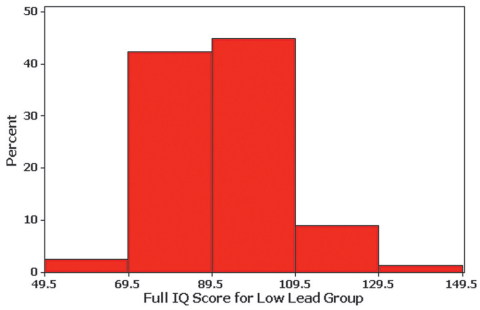


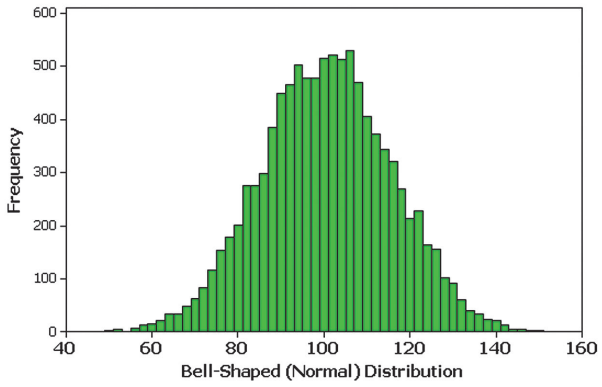
FIGURE 2-3 Relative Frequency Histogram

Critical Thinking: Interpreting Histograms

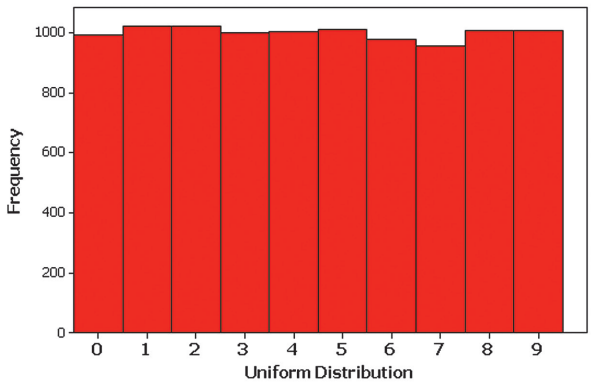
Even though creating histograms is more fun than human beings should be allowed to have, the ultimate objective is to *understand* characteristics of the data. Explore the data by analyzing the histogram to see what can be learned about "CVDOT": the *center* of the data, the *variation* (which will be discussed at length in Section 3-2), the *shape of the distribution*, whether there are any *outliers* (values far away from the other values), and *time* (whether there is any change in the characteristics of the data over time). Examining Figure 2-2, we see that the histogram is centered close to 90, the values vary from around 50 to 150, and the distribution is very roughly bell-shaped. There aren't any outliers, and any changes in time are irrelevant for these data.

Common Distribution Shapes

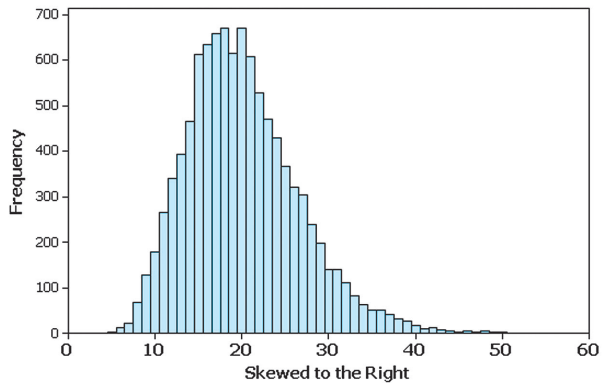
The histograms shown in Figure 2-4 depict four common distribution shapes.



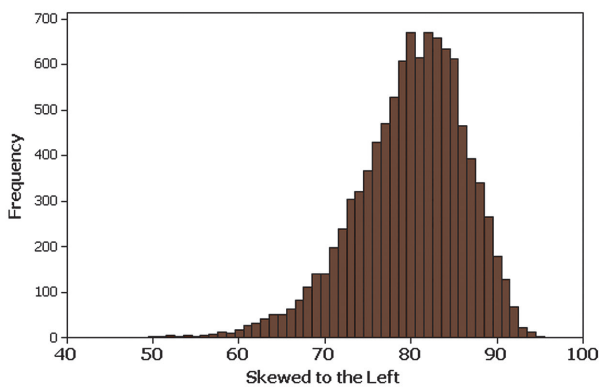
(a)



(b)



(c)



(d)

FIGURE 2-4 Common Distributions

Normal Distribution

When graphed as a histogram, a normal distribution has a "bell" shape similar to the one superimposed in Figure 2-5. Many statistical methods require that sample data come from a population having a distribution that is approximately a normal distribution, and we can often use a histogram to judge whether this requirement is satisfied. There are more advanced and less subjective methods for determining whether the distribution is close to being a normal distribution. Normal quantile plots can be very helpful for assessing normality: see Part 2 of this section.

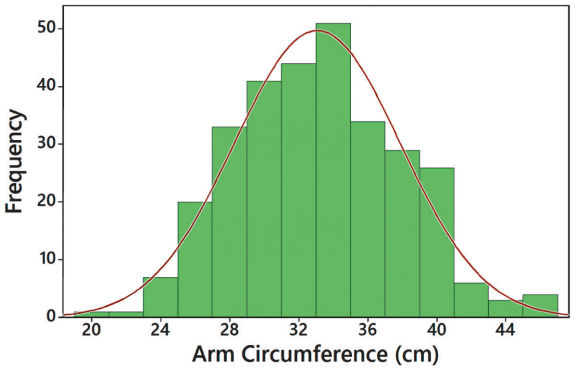


FIGURE 2-5 Bell-Shaped Distribution of Arm Circumferences
Because this histogram is roughly bell-shaped, we say that the data have a *normal distribution*. (A more rigorous definition will be given in Chapter 6.)

Uniform Distribution

With data that have a uniform distribution, the different possible values occur with approximately the same frequency, so the heights of the bars in the histogram are approximately uniform, as in Figure 2-4(b).

Skewness

A distribution of data is **skewed** if it is not symmetric and extends more to one side than to the other. Data **skewed to the right** (also called *positively skewed*) have a longer right tail, as in Figure 2-4(c). Annual incomes of adult Americans are skewed to the right. Data **skewed to the left** (also called *negatively skewed*) have a longer left tail, as in Figure 2-4(d). Life span data in humans are skewed to the left. (Here's a mnemonic for remembering skewness: A distribution skewed to the right resembles the toes on your right foot, and one skewed to the left resembles the toes on your left foot.) Distributions skewed to the right are more common than those skewed to the left because it's often easier to get exceptionally large values than values that are exceptionally small. With annual incomes, for example, it's impossible to get values below zero, but there are a few people who earn millions or billions of dollars in a year. Annual incomes therefore tend to be skewed to the right.

Remembering Skewness:
Skewed Left: Resembles toes on left foot
Skewed Right: Resembles toes on right foot

Go Figure

2.5 quintillion bytes: Amount of data generated each day. (A quintillion is 1 followed by 18 zeroes.)

PART 2 Assessing Normality with Normal Quantile Plots

Some really important methods presented in later chapters have a requirement that sample data must be from a population having a normal distribution. Histograms can be helpful in determining whether the normality requirement is satisfied, but they are not very helpful with small data sets. Section 6-5 discusses methods for *assessing normality*—that is, determining whether the sample data are from a normally distributed population. Section 6-5 includes a procedure for constructing *normal quantile plots*, which are easy to generate using technology such as Statdisk, Minitab, XLSTAT, StatCrunch, or a TI-83/84 Plus calculator. Interpretation of a normal quantile plot is based on the following criteria:

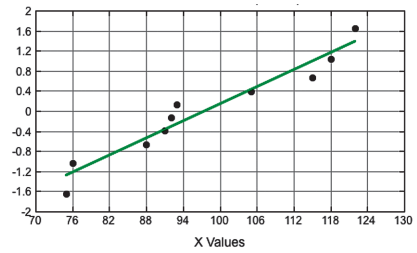
Criteria for Assessing Normality with a Normal Quantile Plot

Normal Distribution: The population distribution is normal if the pattern of the points in the normal quantile plot is reasonably close to a straight line, and the points do not show some other systematic pattern that is not a straight-line pattern.

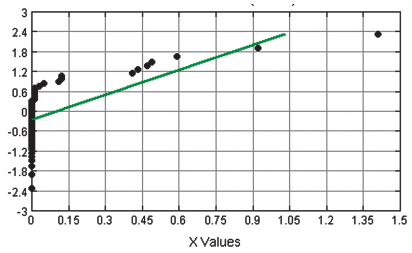
Not a Normal Distribution: The population distribution is *not* normal if the normal quantile plot has either or both of these two conditions:

- The points do not lie reasonably close to a straight-line pattern.
- The points show some *systematic pattern* that is not a straight-line pattern.

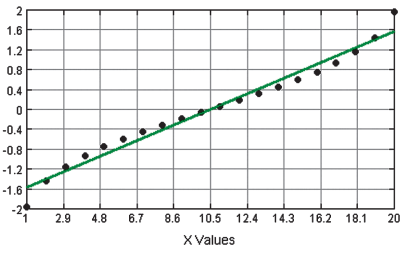
The following are examples of normal quantile plots. Procedures for creating such plots are described in Section 6-5.



Normal Distribution: The points are reasonably close to a straight-line pattern, and there is no other systematic pattern that is not a straight-line pattern.



Not a Normal Distribution: The points do not lie reasonably close to a straight line.



Not a Normal Distribution: The points show a systematic pattern that is not a straight-line pattern.

TECH CENTER

Histograms

Visit **www.pearsonglobaleditions.com** for technology procedures for StatCrunch, Excel, Minitab, Statdisk, TI-83/84 Plus Calculator, and R. Tech supplements, videos, and downloadable data sets for this textbook are also available at this website.

2-2 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. IQ Scores IQ scores of adults are normally distributed. If a large sample of adults is randomly selected and the IQ scores are illustrated in a histogram, what is the shape of that histogram?

2. More IQ Scores The population of IQ scores of adults is normally distributed. If we obtain a voluntary response sample of 5000 of those IQ scores, will a histogram of the sample be bell-shaped?

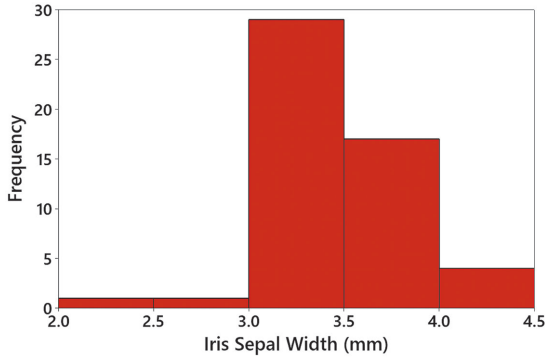
3. Cell Phone Radiation Listed below are radiation levels (W/kg) from Samsung Galaxy cell phones (based on data from Samsung). Why does it *not* make sense to construct a histogram for this data set?

0.87 1.18 1.47 1.25 1.59 1.55

Credit: Based on data from Samsung

4. Cell Phone Radiation If we collect a sample of cell phone radiation amounts much larger than the sample included with Exercise 3, and if our sample includes a single outlier, how will that outlier appear in a histogram?

Interpreting a Histogram. ***In Exercises 5–8, answer the questions by referring to the following Minitab-generated histogram, which depicts the sepal widths (mm) of a sample of irises. (See Data Set 16 "Iris Measurements" in Appendix B.)***



5. Sample Size Based on the histogram, what is the approximate number of irises in the sample?

6. Class Width and Class Limits What is the class width? What are the approximate lower and upper class limits of the first class?

7. Outlier? What is the largest possible value? Is that value an outlier?

8. Normal Distribution Does it appear that the sample is from a population having a normal distribution?

Constructing Histograms. ***In Exercises 9–18, construct the histograms and answer the given questions.***

9. Pulse Rates of Males Refer to Exercise 17 in Section 2-1 and construct the histogram based on the frequency distribution. Does the distribution appear to be normal? Are there any outliers present?

10. Pulse Rates of Females Refer to Exercise 18 in Section 2-1 and construct the histogram based on the frequency distribution. Does the distribution appear to be normal? Are there any outliers present?

11. Freshman 15 Refer to Exercise 19 in Section 2-1 and construct the histogram based on the frequency distribution. Does the distribution appear to be normal? Are there any outliers present?

12. Freshman 15 Refer to Exercise 20 in Section 2-1 and construct the histogram based on the frequency distribution. Does the distribution appear to be normal? Are there any outliers present?

13. Audiometry Refer to Data Set 7 "Audiometry" in Appendix B and construct a histogram based on all of the hearing threshold measurements from the right ear. Use a class width of 10 and a starting point of 5. Does the distribution appear to be normal?

14. Vision Refer to Data Set 8 "Vision" in Appendix B and construct a histogram based on all of the measurements of vision from the right eye. Use a class width of 20 and a starting point of 20. Does the distribution appear to be normal?

15. Vision Refer to Data Set 21 "Passive and Active Smoke" in Appendix B and construct a histogram based on all of the measurements from smokers. Use a class width of 100 and a starting point of 0. Does the distribution appear to be normal?

16. Environmental Tobacco Smoke Refer to Data Set 21 "Passive and Active Smoke" in Appendix B and construct a histogram based on all of the measurements from the sample values of ETS (nonsmokers exposed to environmental tobacco smoke). Use a class width of 100 and a starting point of 0. Does the distribution appear to be normal?

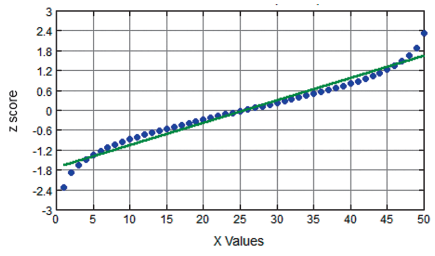
17. Analysis of Last Digits Use the frequency distribution from Exercise 21 in Section 2-1 on page 72 to construct a histogram. What can be concluded from the distribution of the digits? Specifically, do the heights appear to be reported or actually measured?

18. Analysis of Last Digits Use the frequency distribution from Exercise 22 in Section 2-1 on page 72 to construct a histogram. What can be concluded from the distribution of the digits? Specifically, do the weights appear to be reported or actually measured?

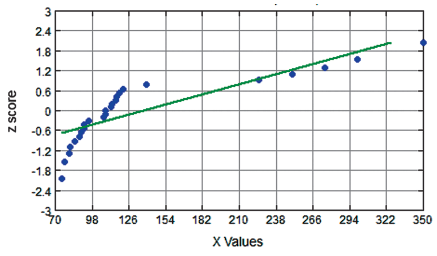
2-2 Beyond the Basics

19. Interpreting Normal Quantile Plots. Which of the following normal quantile plots appear to represent data from a population having a normal distribution? Explain.

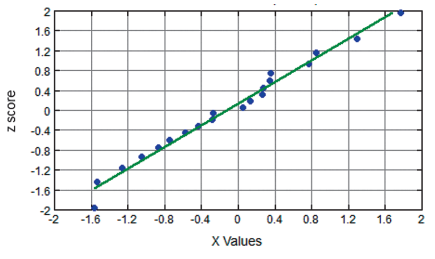
a.



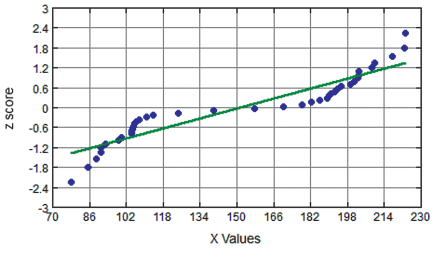
b.



c.



d.



20. Comparing Histograms Refer to the data in Data Set 20 "Alcohol and Tobacco in Movies" in Appendix B and construct histograms for the movie lengths, the times of tobacco use, and the times of alcohol use. Compare the three histograms. Are there any differences in the *distributions* of the data?