Introductory Business

Statistics

Introductory Business Statistics

SENIOR CONTRIBUTING AUTHORS
ALEXANDER HOLMES, THE UNIVERSITY OF OKLAHOMA
BARBARA ILLOWSKY, DE ANZA COLLEGE
SUSAN DEAN, DE ANZA COLLEGE

OpenStax
Rice University
6100 Main Street MS-375
Houston, Texas 77005

To learn more about OpenStax, visit https://openstax.org.
Individual print copies and bulk orders can be purchased through our website.

PRINT BOOK ISBN-10	**1-947172-46-8**
PRINT BOOK ISBN-13	**978-1-947172-46-3**
PDF VERSION ISBN-10	**1-947172-47-6**
PDF VERSION ISBN-13	**978-1-947172-47-0**
Revision Number	**IBS-2017-001(03/18)-LC**
Original Publication Year	**2017**

OpenStax

OpenStax provides free, peer-reviewed, openly licensed textbooks for introductory college and Advanced Placement® courses and low-cost, personalized courseware that helps students learn. A nonprofit ed tech initiative based at Rice University, we're committed to helping students access the tools they need to complete their courses and meet their educational goals.

Rice University

OpenStax, OpenStax CNX, and OpenStax Tutor are initiatives of Rice University. As a leading research university with a distinctive commitment to undergraduate education, Rice University aspires to path-breaking research, unsurpassed teaching, and contributions to the betterment of our world. It seeks to fulfill this mission by cultivating a diverse community of learning and discovery that produces leaders across the spectrum of human endeavor.

Foundation Support

OpenStax is grateful for the tremendous support of our sponsors. Without their strong engagement, the goal of free access to high-quality textbooks would remain just a dream.

Laura and John Arnold Foundation (LJAF) actively seeks opportunities to invest in organizations and thought leaders that have a sincere interest in implementing fundamental changes that not only yield immediate gains, but also repair broken systems for future generations. LJAF currently focuses its strategic investments on education, criminal justice, research integrity, and public accountability.

The William and Flora Hewlett Foundation has been making grants since 1967 to help solve social and environmental problems at home and around the world. The Foundation concentrates its resources on activities in education, the environment, global development and population, performing arts, and philanthropy, and makes grants to support disadvantaged communities in the San Francisco Bay Area.

Calvin K. Kazanjian was the founder and president of Peter Paul (Almond Joy), Inc. He firmly believed that the more people understood about basic economics the happier and more prosperous they would be. Accordingly, he established the Calvin K. Kazanjian Economics Foundation Inc, in 1949 as a philanthropic, nonpolitical educational organization to support efforts that enhanced economic understanding.

Guided by the belief that every life has equal value, the Bill & Melinda Gates Foundation works to help all people lead healthy, productive lives. In developing countries, it focuses on improving people's health with vaccines and other life-saving tools and giving them the chance to lift themselves out of hunger and extreme poverty. In the United States, it seeks to significantly improve education so that all young people have the opportunity to reach their full potential. Based in Seattle, Washington, the foundation is led by CEO Jeff Raikes and Co-chair William H. Gates Sr., under the direction of Bill and Melinda Gates and Warren Buffett.

The Maxfield Foundation supports projects with potential for high impact in science, education, sustainability, and other areas of social importance.

Our mission at The Michelson 20MM Foundation is to grow access and success by eliminating unnecessary hurdles to affordability. We support the creation, sharing, and proliferation of more effective, more affordable educational content by leveraging disruptive technologies, open educational resources, and new models for collaboration between for-profit, nonprofit, and public entities.

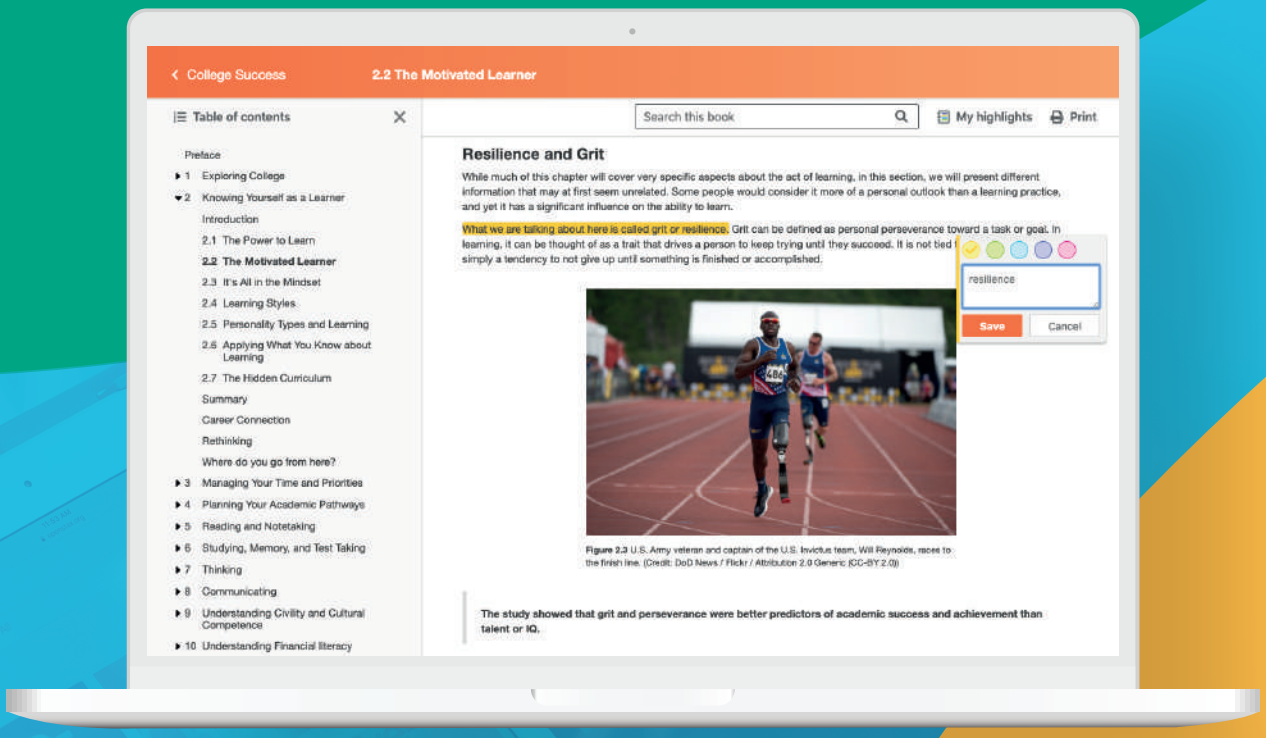
The Bill and Stephanie Sick Fund supports innovative projects in the areas of Education, Art, Science and Engineering.

openstax™

Study where you want, what you want, when you want.

When you access College Success in our web view, you can use our new online highlighting and note-taking features to create your own study guides.

Our books are free and flexible, forever.
Get started at openstax.org/details/books/introductory-business-statistics



Access. The future of education.
openstax.org

RICE

Table of Contents

PREFACE

Welcome to *Introductory Business Statistics*, an OpenStax resource. This textbook was written to increase student access to high-quality learning materials, maintaining highest standards of academic rigor at little to no cost.

About OpenStax

OpenStax is a nonprofit based at Rice University, and it's our mission to improve student access to education. Our first openly licensed college textbook was published in 2012, and our library has since scaled to over 25 books for college and AP® courses used by hundreds of thousands of students. OpenStax Tutor, our low-cost personalized learning tool, is being used in college courses throughout the country. Through our partnerships with philanthropic foundations and our alliance with other educational resource organizations, OpenStax is breaking down the most common barriers to learning and empowering students and instructors to succeed.

About OpenStax resources

Customization

Introductory Business Statistics is licensed under a Creative Commons Attribution 4.0 International (CC BY) license, which means that you can distribute, remix, and build upon the content, as long as you provide attribution to OpenStax and its content contributors.

Because our books are openly licensed, you are free to use the entire book or pick and choose the sections that are most relevant to the needs of your course. Feel free to remix the content by assigning your students certain chapters and sections in your syllabus, in the order that you prefer. You can even provide a direct link in your syllabus to the sections in the web view of your book.

Instructors also have the option of creating a customized version of their OpenStax book. The custom version can be made available to students in low-cost print or digital form through their campus bookstore. Visit the Instructor Resources section of your book page on OpenStax.org for more information.

Errata

All OpenStax textbooks undergo a rigorous review process. However, like any professional-grade textbook, errors sometimes occur. Since our books are web based, we can make updates periodically when deemed pedagogically necessary. If you have a correction to suggest, submit it through the link on your book page on OpenStax.org. Subject matter experts review all errata suggestions. OpenStax is committed to remaining transparent about all updates, so you will also find a list of past errata changes on your book page on OpenStax.org.

Format

You can access this textbook for free in web view or PDF through OpenStax.org, and for a low cost in print.

About *Introductory Business Statistics*

Introductory Business Statistics is designed to meet the scope and sequence requirements of the one-semester statistics course for business, economics, and related majors. Core statistical concepts and skills have been augmented with practical business examples, scenarios, and exercises. The result is a meaningful understanding of the discipline which will serve students in their business careers and real-world experiences.

Coverage and scope

Introductory Business Statistics began as a customized version of OpenStax *Introductory Statistics* by Barbara Illowsky and Susan Dean. Statistics faculty at The University of Oklahoma have used the business statistics adaptation for several years, and the author has continually refined it based on student success and faculty feedback.

The book is structured in a similar manner to most traditional statistics textbooks. The most significant topical changes occur in the latter chapters on regression analysis. Discrete probability density functions have been reordered to provide a logical progression from simple counting formulas to more complex continuous distributions. Many additional homework assignments have been added, as well as new, more mathematical examples.

Introductory Business Statistics places a significant emphasis on the development and practical application of formulas so that students have a deeper understanding of their interpretation and application of data. To achieve this unique approach, the author included a wealth of additional material and purposely de-emphasized the use of the scientific calculator. Specific changes regarding formula use include:

- Expanded discussions of the combinatorial formulas, factorials, and sigma notation
- Adjustments to explanations of the acceptance/rejection rule for hypothesis testing, as well as a focus on terminology regarding confidence intervals
- Deep reliance on statistical tables for the process of finding probabilities (which would not be required if probabilities relied on scientific calculators)
- Continual and emphasized links to the Central Limit Theorem throughout the book; *Introductory Business Statistics* consistently links each test statistic back to this fundamental theorem in inferential statistics

Another fundamental focus of the book is the link between statistical inference and the scientific method. Business and economics models are fundamentally grounded in assumed relationships of cause and effect. They are developed to both test hypotheses and to predict from such models. This comes from the belief that statistics is the gatekeeper that allows some theories to remain and others to be cast aside for a new perspective of the world around us. This philosophical view is presented in detail throughout and informs the method of presenting the regression model, in particular.

The correlation and regression chapter includes confidence intervals for predictions, alternative mathematical forms to allow for testing categorical variables, and the presentation of the multiple regression model.

Pedagogical features

- **Examples** are placed strategically throughout the text to show students the step-by-step process of interpreting and solving statistical problems. To keep the text relevant for students, the examples are drawn from a broad spectrum of practical topics; these include examples about college life and learning, health and medicine, retail and business, and sports and entertainment.
- **Practice, Homework, and Bringing It Together** give the students problems at various degrees of difficulty while also including real-world scenarios to engage students.

Additional resources

Student and instructor resources

We've compiled additional resources for both students and instructors, including Getting Started Guides, an instructor solution manual, and PowerPoint slides. Instructor resources require a verified instructor account, which you can apply for when you log in or create your account on OpenStax.org. Take advantage of these resources to supplement your OpenStax book.

Community Hubs

OpenStax partners with the Institute for the Study of Knowledge Management in Education (ISKME) to offer Community Hubs on OER Commons – a platform for instructors to share community-created resources that support OpenStax books, free of charge. Through our Community Hubs, instructors can upload their own materials or download resources to use in their own courses, including additional ancillaries, teaching material, multimedia, and relevant course content. We encourage instructors to join the hubs for the subjects most relevant to your teaching and research as an opportunity both to enrich your courses and to engage with other faculty.

To reach the Community Hubs, visit www.oercommons.org/hubs/OpenStax.

Technology partners

As allies in making high-quality learning materials accessible, our technology partners offer optional low-cost tools that are integrated with OpenStax books. To access the technology options for your text, visit your book page on OpenStax.org.

About the authors

Senior contributing authors

Alexander Holmes, The University of Oklahoma

Barbara Illowsky, DeAnza College

Susan Dean, DeAnza College

Contributing authors

Kevin Hadley, Analyst, Federal Reserve Bank of Kansas City

Reviewers

Birgit Aquilonius, West Valley College
Charles Ashbacher, Upper Iowa University - Cedar Rapids
Abraham Biggs, Broward Community College

Daniel Birmajer, Nazareth College
Roberta Bloom, De Anza College
Bryan Blount, Kentucky Wesleyan College
Ernest Bonat, Portland Community College
Sarah Boslaugh, Kennesaw State University
David Bosworth, Hutchinson Community College
Sheri Boyd, Rollins College
George Bratton, University of Central Arkansas
Franny Chan, Mt. San Antonio College
Jing Chang, College of Saint Mary
Laurel Chiappetta, University of Pittsburgh
Lenore Desilets, De Anza College
Matthew Einsohn, Prescott College
Ann Flanigan, Kapiolani Community College
David French, Tidewater Community College
Mo Geraghty, De Anza College
Larry Green, Lake Tahoe Community College
Michael Greenwich, College of Southern Nevada
Inna Grushko, De Anza College
Valier Hauber, De Anza College
Janice Hector, De Anza College
Jim Helmreich, Marist College
Robert Henderson, Stephen F. Austin State University
Mel Jacobsen, Snow College
Mary Jo Kane, De Anza College
John Kagochi, University of Houston - Victoria
Lynette Kenyon, Collin County Community College
Charles Klein, De Anza College
Alexander Kolovos
Sheldon Lee, Viterbo University
Sara Lenhart, Christopher Newport University
Wendy Lightheart, Lane Community College
Vladimir Logvenenko, De Anza College
Jim Lucas, De Anza College
Suman Majumdar, University of Connecticut
Lisa Markus, De Anza College
Miriam Masullo, SUNY Purchase
Diane Mathios, De Anza College
Robert McDevitt, Germanna Community College
John Migliaccio, Fordham University
Mark Mills, Central College
Cindy Moss, Skyline College
Nydia Nelson, St. Petersburg College
Benjamin Ngwudike, Jackson State University
Jonathan Oaks, Macomb Community College
Carol Olmstead, De Anza College
Barbara A. Osyk, The University of Akron
Adam Pennell, Greensboro College
Kathy Plum, De Anza College
Lisa Rosenberg, Elon University
Sudipta Roy, Kankakee Community College
Javier Rueda, De Anza College
Yvonne Sandoval, Pima Community College
Rupinder Sekhon, De Anza College
Travis Short, St. Petersburg College
Frank Snow, De Anza College
Abdulhamid Sukar, Cameron University
Jeffery Taub, Maine Maritime Academy
Mary Teegarden, San Diego Mesa College

John Thomas, College of Lake County
Philip J. Verrecchia, York College of Pennsylvania
Dennis Walsh, Middle Tennessee State University
Cheryl Wartman, University of Prince Edward Island
Carol Weideman, St. Petersburg College
Kyle S. Wells, Dixie State University
Andrew Wiesner, Pennsylvania State University

1 | SAMPLING AND DATA

Figure 1.1 We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

Introduction

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

1.1 | Definitions of Statistics, Probability, and Key Terms

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a **fair** coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is $\frac{1}{2}$ or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl **Pearson** who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction $\frac{996}{2000}$ is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter, in this case the mean. A **parameter** is a numerical characteristic of the whole population that can be estimated by a statistic. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, or random variable, usually notated by capital letters such as X and Y, is a characteristic or measurement that can be determined for each member of a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let X equal the number of points earned by one math student at the end of a term, then X is a numerical variable. If we let Y be a person's party affiliation, then some examples of Y include Republican, Democrat, and Independent. Y is a categorical variable. We could do some math with values of X (calculate the average number of points earned, for example), but it makes no sense to do math with values of Y (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is $\frac{22}{40}$ and the proportion of women students is $\frac{18}{40}$. Mean and proportion are discussed in more detail in later chapters.

NOTE

The words " **mean**" and " **average**" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example 1.1

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Solution 1.1

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term: the population mean.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Try It

1.1 Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent $65, $75, and $95, respectively.

Example 1.2

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1. Population ____ 2. Statistic ____ 3. Parameter ____ 4. Sample ____ 5. Variable ____ 6. Data ____

a) all students who attended the college last year
b) the cumulative GPA of one student who graduated from the college last year
c) 3.65, 2.80, 1.50, 3.90
d) a group of students who graduated from the college last year, randomly selected
e) the average cumulative GPA of students who graduated from the college last year
f) all students who graduated from the college last year
g) the average cumulative GPA of students in the study who graduated from the college last year

Solution 1.2

1. f; 2. g; 3. e; 4. d; 5. b; 6. c

Example 1.3

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of “drive” (i.e. dummies)
35 miles/hour	Front Seat

Table 1.1

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver’s seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution 1.3

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

Example 1.4

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Solution 1.4

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.

The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

1.2 | Data, Sampling, and Variation in Data and Sampling

Data may come from a population or from a sample. Lowercase letters like x or y generally are used to represent data

values. Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. **Qualitative data** are also often called categorical data. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative(categorical) data. Qualitative(categorical) data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative(categorical) data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

Data that are not only made up of counting numbers, but that may include fractions, decimals, or irrational numbers, are called **quantitative continuous data**. Continuous data are often the results of measurements like lengths, weights, or times. A list of the lengths in minutes for all the phone calls that you make in a week, with numbers like 2.4, 7.5, or 11.0, would be quantitative continuous data.

Example 1.5 Data Sample of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are the quantitative discrete data.

Try It

1.5 The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

Example 1.6 Data Sample of Quantitative Continuous Data

The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data.

Try It

1.6 The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

Example 1.7

You go to the supermarket and purchase three cans of soup (19 ounces) tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces pistachio ice cream and 32 ounces chocolate chip cookies).

Name data sets that are quantitative discrete, quantitative continuous, and qualitative(categorical).

Solution 1.7

One Possible Solution:

- The three cans of soup, two packages of nuts, four kinds of vegetables and two desserts are quantitative discrete data because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are qualitative(categorical) data because they are categorical.

Try to identify additional data sets in this example.

Example 1.8

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative(categorical) data.

Try It

1.8 The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

NOTE

You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Example 1.9

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

a. the number of pairs of shoes you own
b. the type of car you drive
c. the distance from your home to the nearest grocery store
d. the number of classes you take per school year
e. the type of calculator you use
f. weights of sumo wrestlers

g. number of correct answers on a quiz

h. IQ scores (This may cause some discussion.)

Solution 1.9

Items a, d, and g are quantitative discrete; items c, f, and h are quantitative continuous; items b and e are qualitative, or categorical.

Try It

1.9 Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

Example 1.10

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart **Figure 1.1**. What type of data does this graph show?

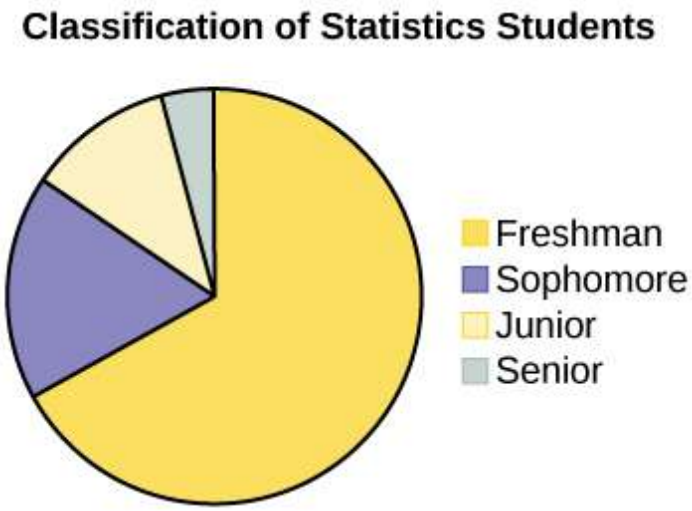


Figure 1.2

Solution 1.10

This pie chart shows the students in each year, which is **qualitative (or categorical) data**.

Try It

1.10 The registrar at State University keeps records of the number of credit hours students complete each semester. The data he collects are summarized in the histogram. The class boundaries are 10 to less than 13, 13 to less than 16, 16 to less than 19, 19 to less than 22, and 22 to less than 25.

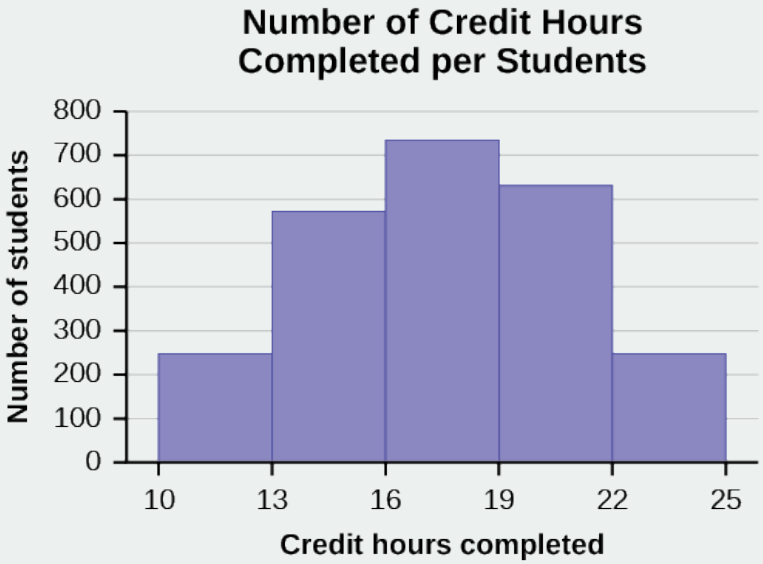


Figure 1.3

What type of data does this graph show?

Qualitative Data Discussion

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

De Anza College				**Foothill College**		
	Number	Percent			Number	Percent
Full-time	9,200	40.9%		Full-time	4,059	28.6%
Part-time	13,296	59.1%		Part-time	10,124	71.4%
Total	22,496	100%		Total	14,183	100%

Table 1.2 Fall Term 2007 (Census day)

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative(categorical) data are pie charts and bar graphs.

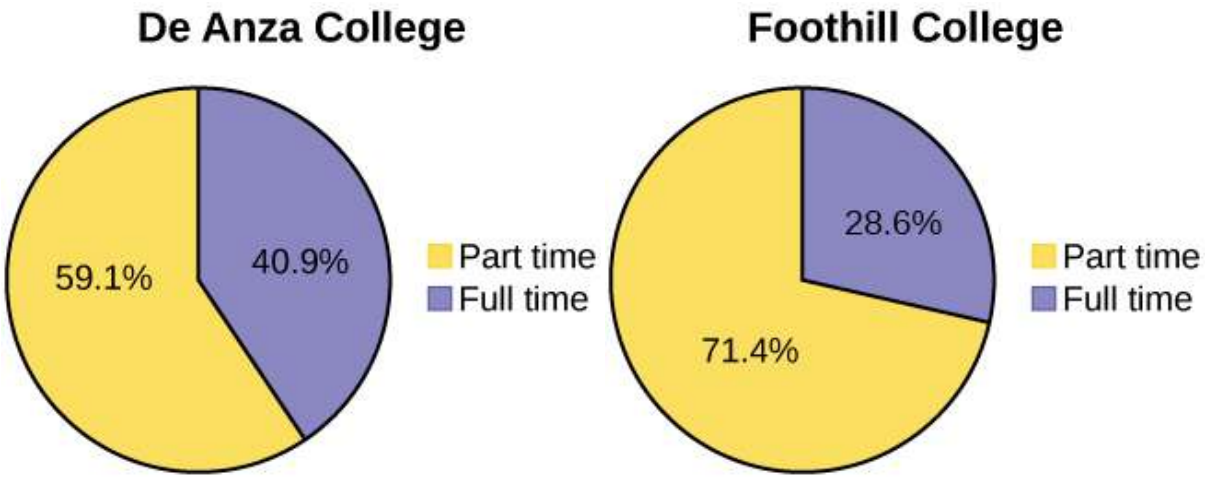
In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.

In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at **Figure 1.4** and **Figure 1.5** and determine which graph (pie or bar) you think displays the comparisons better.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the "best" graph depending on the data and the context. Our choice also depends on what we are using the data for.



(a) (b)

Figure 1.4

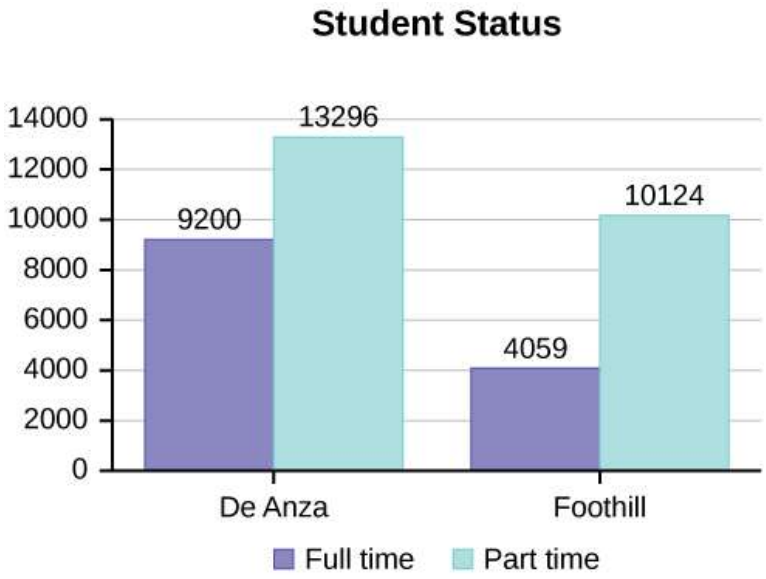


Figure 1.5

Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Characteristic/Category	Percent
Full-Time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%

Table 1.3 De Anza College Spring 2010

Characteristic/Category	Percent
Students under age 25	61.0%
TOTAL	150.5%

Table 1.3 De Anza College Spring 2010

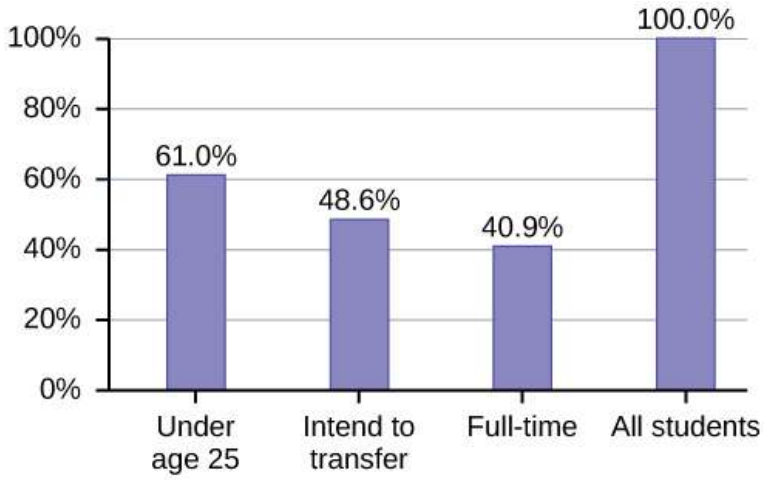


Figure 1.6

Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

Table 1.4 Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)

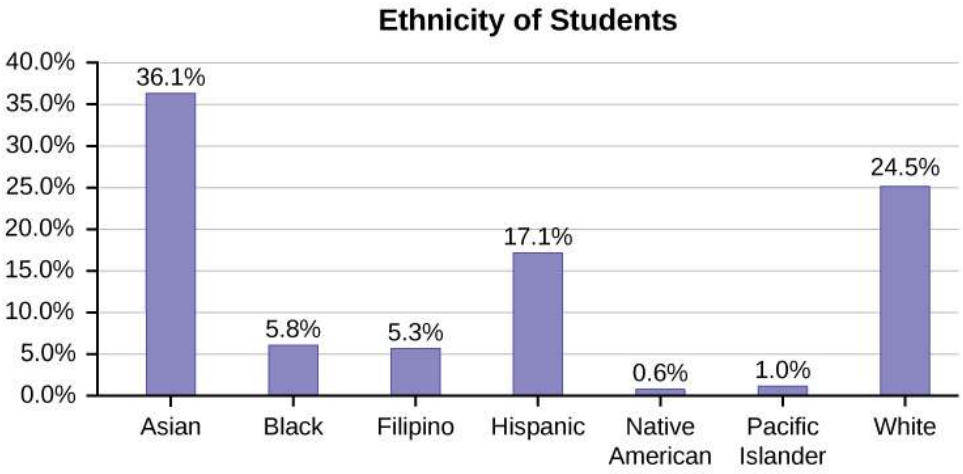


Figure 1.7

The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been included. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us.

This particular bar graph in **Figure 1.8** can be difficult to understand visually. The graph in **Figure 1.9** is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

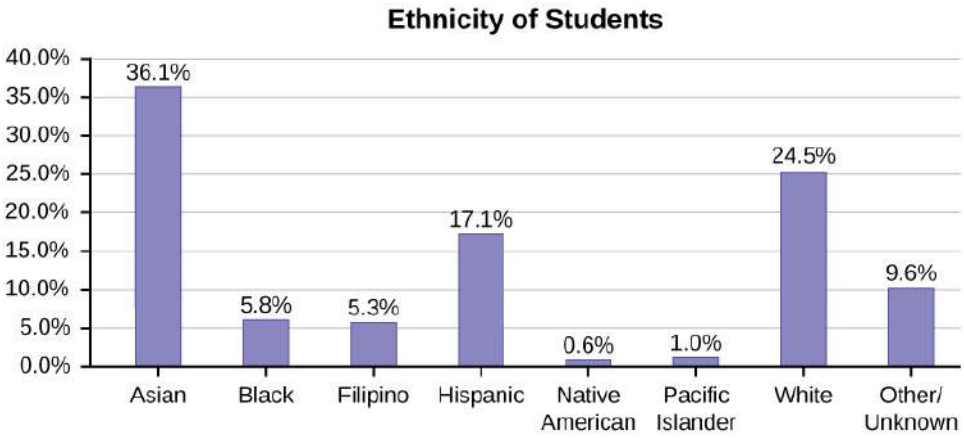


Figure 1.8 Bar Graph with Other/Unknown Category

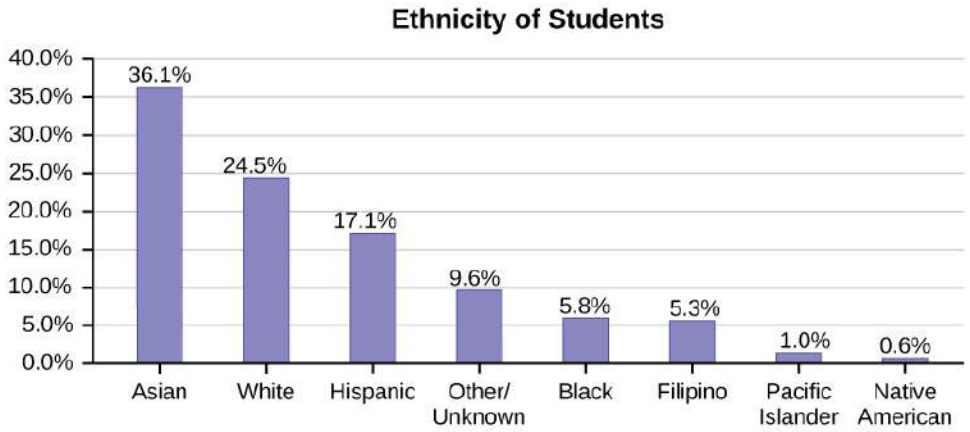


Figure 1.9 Pareto Chart With Bars Sorted by Size

Pie Charts: No Missing Data

The following pie charts have the "Other/Unknown" category included (since the percentages must add to 100%). The chart in **Figure 1.10b** is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in **Figure 1.10a**.

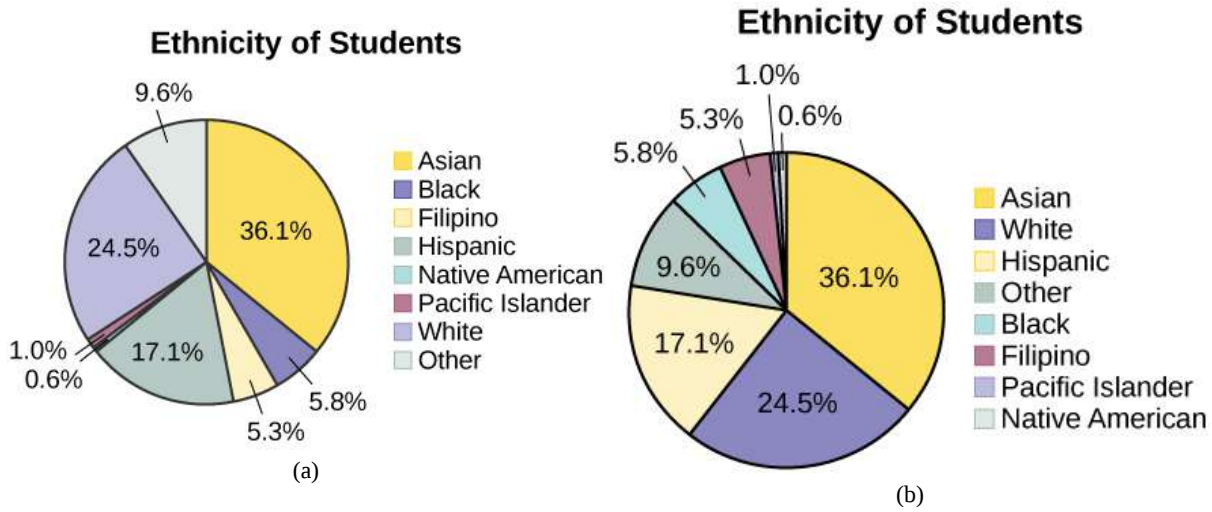


Figure 1.10

Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of *n* individuals is equally likely to be chosen as any other group of *n* individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected.

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a

proportionate simple random sample from each stratum (each department) to get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose you want to pick a sample of 1,000 randomly for a survey. **For any particular sample of 1,000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling **with replacement for any particular sample**, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To four decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling

process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions
- Undue influence: collecting data or asking questions in a way that influences the response
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

Example 1.11

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

a. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.

b. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.

c. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.

d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.

e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

Solution 1.11
a. stratified; b. systematic; c. simple random; d. cluster; e. convenience

Example 1.12

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
b. A pollster interviews all human resource personnel in five different high tech companies.
c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution 1.12
a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f.convenience

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

Example 1.13

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.

Suppose we take two different samples.

First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

$128; $87; $173; $116; $130; $204; $147; $189; $93; $153

The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:

$50; $40; $36; $15; $50; $100; $40; $53; $22; $22

It is unlikely that any student is in both samples.

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Solution 1.13
a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average parttime student. Both samples are biased. Also, in both cases,

not all students have a chance to be in either sample.

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Solution 1.13

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

$180; $50; $150; $85; $260; $75; $180; $200; $200; $150

c. Is the sample biased?

Solution 1.13

c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Try It

1.13 A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8; 16.1; 15.2; 14.8; 15.8; 15.9; 16.0; 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more **samples** from the same **population**, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither

would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This **variability in samples** cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations, usually given the symbol n) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. Later we will find that even much smaller sample sizes will give very good results. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

1.3 | Levels of Measurement

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

Levels of Measurement

The way a set of data is measured is called its **level of measurement**. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- **Nominal scale level**
- **Ordinal scale level**
- **Interval scale level**
- **Ratio scale level**

Data that is measured using a **nominal scale** is **qualitative (categorical)**. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. The data are the names of the companies that make smartphones, but there is no agreed upon order of these brands, even though people may have personal preferences. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are "excellent," "good," "satisfactory," and "unsatisfactory." These responses are ordered from the most desired response to the least desired. But the differences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40° is equal to 100° minus 60°. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations, but one type of comparison cannot be done. 80° C is not four times as hot as 20° C (nor is 80° F four times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or four to one).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. For example, four multiple choice statistics

final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams are machine-graded.

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So 80 is four times 20. The score of 80 is four times better than the score of 20.

Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Table 1.5 lists the different data values in ascending order and their frequencies.

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2
7	1

Table 1.5 Frequency Table of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to **Table 1.5**, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample–in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

Table 1.6 Frequency Table of Student Work Hours with Relative Frequencies

The sum of the values in the relative frequency column of **Table 1.6** is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative

frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in **Table 1.7**.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

Table 1.7 Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

NOTE

Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

Table 1.8 represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100}$ = 0.05	0.05
61.95–63.95	3	$\frac{3}{100}$ = 0.03	0.05 + 0.03 = 0.08
63.95–65.95	15	$\frac{15}{100}$ = 0.15	0.08 + 0.15 = 0.23
65.95–67.95	40	$\frac{40}{100}$ = 0.40	0.23 + 0.40 = 0.63
67.95–69.95	17	$\frac{17}{100}$ = 0.17	0.63 + 0.17 = 0.80
69.95–71.95	12	$\frac{12}{100}$ = 0.12	0.80 + 0.12 = 0.92
71.95–73.95	7	$\frac{7}{100}$ = 0.07	0.92 + 0.07 = 0.99

Table 1.8 Frequency Table of Soccer Player Height

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
73.95–75.95	1	$\frac{1}{100} = 0.01$	0.99 + 0.01 = 1.00
	Total = 100	**Total = 1.00**	

Table 1.8 Frequency Table of Soccer Player Height

The data in this table have been **grouped** into the following intervals:

- 59.95 to 61.95 inches
- 61.95 to 63.95 inches
- 63.95 to 65.95 inches
- 65.95 to 67.95 inches
- 67.95 to 69.95 inches
- 69.95 to 71.95 inches
- 71.95 to 73.95 inches
- 73.95 to 75.95 inches

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Example 1.14

From **Table 1.8**, find the percentage of heights that are less than 65.95 inches.

Solution 1.14

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are 5 + 3 + 15 = 23 players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.

Try It

1.14 **Table 1.9** shows the amount, in inches, of annual rainfall in a sample of towns.

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
2.95–4.97	6	$\frac{6}{50} = 0.12$	0.12
4.97–6.99	7	$\frac{7}{50} = 0.14$	0.12 + 0.14 = 0.26
6.99–9.01	15	$\frac{15}{50} = 0.30$	0.26 + 0.30 = 0.56
9.01–11.03	8	$\frac{8}{50} = 0.16$	0.56 + 0.16 = 0.72
11.03–13.05	9	$\frac{9}{50} = 0.18$	0.72 + 0.18 = 0.90
13.05–15.07	5	$\frac{5}{50} = 0.10$	0.90 + 0.10 = 1.00
	Total = 50	Total = 1.00	

Table 1.9

From **Table 1.9**, find the percentage of rainfall that is less than 9.01 inches.

Example 1.15

From **Table 1.8**, find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution 1.15
Add the relative frequencies in the second and third rows: 0.03 + 0.15 = 0.18 or 18%.

Try It

1.15 From **Table 1.9**, find the percentage of rainfall that is between 6.99 and 13.05 inches.

Example 1.16

Use the heights of the 100 male semiprofessional soccer players in **Table 1.8**. Fill in the blanks and check your answers.

a. The percentage of heights that are from 67.95 to 71.95 inches is: ____.
b. The percentage of heights that are from 67.95 to 73.95 inches is: ____.
c. The percentage of heights that are more than 65.95 inches is: ____.
d. The number of players in the sample who are between 61.95 and 71.95 inches tall is: ____.
e. What kind of data are the heights?

f. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

Solution 1.16

a. 29%

b. 36%

c. 77%

d. 87

e. quantitative continuous

f. get rosters from each team and choose a simple random sample from each

Example 1.17

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. **Table 1.10** was produced:

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{4}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

Table 1.10 Frequency of Commuting Distances

a. Is the table correct? If it is not correct, what is wrong?

b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.

c. What fraction of the people surveyed commute five or seven miles?

d. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Solution 1.17

a. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.

b. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.

c. $\frac{5}{19}$

d. $\frac{7}{19}, \frac{12}{19}, \frac{7}{19}$

Try It

1.17 **Table 1.9** represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Example 1.18

Table 1.11 contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Table 1.11

Answer the following questions.

a. What is the frequency of deaths measured from 2006 through 2009?

b. What percentage of deaths occurred after 2009?

c. What is the relative frequency of deaths that occurred in 2003 or earlier?

d. What is the percentage of deaths that occurred in 2004?

e. What kind of data are the numbers of deaths?

f. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Solution 1.18

a. 97,118 (11.8%)

b. 41.6%

c. 67,092/823,356 or 0.081 or 8.1 %

d. 27.8%

e. Quantitative discrete

f. Quantitative continuous

Try It Σ

1.18 **Table 1.12** contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Table 1.12

Answer the following questions.

a. What is the frequency of deaths measured from 2000 through 2004?
b. What percentage of deaths occurred after 2006?
c. What is the relative frequency of deaths that occurred in 2000 or before?
d. What is the percentage of deaths that occurred in 2011?
e. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

1.4 | Experimental Design and Ethics

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **independent variable** or **explanatory variable**. The affected variable is called the **dependent variable** or **response variable**: stimulus, response. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An **experimental unit** is a single object or individual to be measured.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is

accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

*Results showed that believing one had taken the substance resulted in [*performance*] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.*[1]

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment–a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

Example 1.19

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

a. Describe the explanatory and response variables in this study.
b. What are the treatments?
c. Identify any lurking variables that could interfere with this study.
d. Is it possible to use blinding in this study?

Solution 1.19

a. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
b. There are two treatments: a floral-scented mask and an unscented mask.
c. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
d. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

1. McClung, M. Collins, D. "Because I know it will!": placebo effects of an ergogenic aid on athletic performance. Journal of Sport & Exercise Psychology. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

KEY TERMS

Average also called mean or arithmetic mean; a number that describes the central tendency of the data

Blinding not telling participants which treatment a subject is receiving

Categorical Variable variables that take on values that are names or labels

Cluster Sampling a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

Control Group a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

Convenience Sampling a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Cumulative Relative Frequency The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Data a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Discrete Random Variable a random variable (RV) whose outcomes are counted

Double-blinding the act of blinding both the subjects of an experiment and the researchers who work with the subjects

Experimental Unit any individual or object to be measured

Explanatory Variable the **independent variable** in an experiment; the value controlled by researchers

Frequency the number of times a value of the data occurs

Informed Consent Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Institutional Review Board a committee tasked with oversight of research programs that involve human subjects

Lurking Variable a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Mathematical Models a description of a phenomenon using mathematical concepts, such as equations, inequalities, distributions, etc.

Nonsampling Error an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Numerical Variable variables that take on values that are indicated by numbers

Observational Study a study in which the independent variable is not manipulated by the researcher

Parameter a number that is used to represent a population characteristic and that generally cannot be determined easily

Placebo an inactive treatment that has no real effect on the explanatory variable

Population all individuals, objects, or measurements whose properties are being studied

Probability a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion the number of successes divided by the total number in the sample

Qualitative Data See **Data**.

Quantitative Data See **Data**.

Random Assignment the act of organizing experimental units into treatment groups using random methods

Random Sampling a method of selecting a sample that gives every member of the population an equal chance of being selected.

Relative Frequency the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes

Representative Sample a subset of the population that has the same characteristics as the population

Response Variable the **dependent variable** in an experiment; the value that is measured for change at the end of an experiment

Sample a subset of the population studied

Sampling Bias not all members of the population are equally likely to be selected

Sampling Error the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample.

Statistic a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Statistical Models a description of a phenomenon using probability distributions that describe the expected behavior of the phenomenon and the variability in the expected observations.

Stratified Sampling a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Survey a study in which data is collected as reported by individuals.

Systematic Sampling a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

Treatments different values or components of the explanatory variable applied in an experiment

Variable a characteristic of interest for each person or object in a population

CHAPTER REVIEW

1.1 Definitions of Statistics, Probability, and Key Terms

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

1.2 Data, Sampling, and Variation in Data and Sampling

Data are individual items of information that come from a population or sample. Data may be classified as qualitative (categorical), quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

1.3 Levels of Measurement

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

In addition to rounding your answers, you can measure your data using the following four levels of measurement.

- **Nominal scale level:** data that cannot be ordered nor can it be used in calculations
- **Ordinal scale level:** data that can be ordered; the differences cannot be measured
- **Interval scale level:** data with a definite ordering but no starting point; the differences can be measured, but there is no such thing as a ratio.
- **Ratio scale level:** data with a starting point that can be ordered; the differences have meaning and ratios can be calculated.

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.

1.4 Experimental Design and Ethics

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

“An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule.”[2] Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

---

2. Andrew Gelman, “Open Data and Open Methods,” Ethics and Statistics, http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics1.pdf (accessed May 1, 2013).

HOMEWORK

1.1 Definitions of Statistics, Probability, and Key Terms

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

1. A fitness center is interested in the mean amount of time a client exercises in the center each week.

2. Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

3. A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

4. Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

5. A politician is interested in the proportion of voters in his district who think he is doing a good job.

6. A marriage counselor is interested in the proportion of clients she counsels who stay married.

7. Political pollsters may be interested in the proportion of people who will vote for a particular cause.

8. A marketing company is interested in the proportion of people who will buy a particular product.

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

9. What is the population she is interested in?

a. all Lake Tahoe Community College students
b. all Lake Tahoe Community College English students
c. all Lake Tahoe Community College students in her classes
d. all Lake Tahoe Community College math students

10. Consider the following:

X = number of days a Lake Tahoe Community College math student is absent

In this case, X is an example of a:

a. variable.
b. population.
c. statistic.
d. data.

11. The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

a. parameter.
b. data.
c. statistic.
d. variable.

1.2 Data, Sampling, and Variation in Data and Sampling

For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.

12. number of tickets sold to a concert

13. percent of body fat

14. favorite baseball team

15. time in line to buy groceries

16. number of students enrolled at Evergreen Valley College

17. most-watched television show

18. brand of toothpaste

19. distance to the closest movie theatre

20. age of executives in Fortune 500 companies

21. number of competing computer spreadsheet software packages

Use the following information to answer the next two exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

22. "Number of times per week" is what type of data?
a. qualitative (categorical)
b. quantitative discrete
c. quantitative continuous

23. "Duration (amount of time)" is what type of data?
a. qualitative (categorical)
b. quantitative discrete
c. quantitative continuous

24. Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.
a. Using complete sentences, list three things wrong with the way the survey was conducted.
b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

25. Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

26. Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

27. List some practical difficulties involved in getting accurate results from a telephone survey.

28. List some practical difficulties involved in getting accurate results from a mailed survey.

29. With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.

30. The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is
a. cluster sampling
b. stratified sampling
c. simple random sampling
d. convenience sampling

31. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was:
a. simple random
b. systematic
c. stratified
d. cluster

32. Name the sampling method used in each of the following situations:

a. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.
b. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.
e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

33. A "random survey" was conducted of 3,274 people of the "microprocessor generation" (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2,000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users.

a. Do you consider the sample size large enough for a study of this type? Why or why not?
b. Based on your "gut feeling," do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?
   Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."
c. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

34. The Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below.

Identify the type of data obtained from each question used in this survey: qualitative(categorical), quantitative discrete, or quantitative continuous.

a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
c. In the last seven days, on how many days did you exercise for 30 minutes or more?
d. Do you have health insurance coverage?

35. In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.

a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
b. What effect does the low response rate have on the reliability of the sample?
c. Are these problems examples of sampling error or nonsampling error?
d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. These researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

36. Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates.

Which of the potential problems with samples discussed in **Section 1.2** could explain this connection?

37. YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks:

"Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?"[3]

As of April 25, 11 people responded to this question. Each participant answered "NO!"

Which of the potential problems with samples discussed in this module could explain this connection?

38. A scholarly article about response rates begins with the following quote:

"Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research."[4]

The Pew Research Center for People and the Press admits:

"The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more."[5]

a. What are some reasons for the decline in response rate over the past decade?
b. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

1.3 Levels of Measurement

39. Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		
3			

Table 1.13 Part-time Student Course Loads

a. Fill in the blanks in **Table 1.13**.
b. What percent of students take exactly two courses?
c. What percent of students take one or two courses?

3. lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: http://www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).
4. Scott Keeter et al., "Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey," Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759.full (http://poq.oxfordjournals.org/content/70/5/759.full) (accessed May 1, 2013).
5. Frequently Asked Questions, Pew Research Center for the People & the Press, http://www.people-press.org/methodology/frequently-asked-questions/#dont-you-have-trouble-getting-people-to-answer-your-polls (accessed May 1, 2013).

40. Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in **Table 1.14**.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Freq.
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

Table 1.14 Flossing Frequency for Adults with Gum Disease

a. Fill in the blanks in **Table 1.14**.
b. What percent of adults flossed six times per week?
c. What percent flossed at most three times per week?

41. Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows: 2; 5; 7; 2; 2; 10; 20; 15; 0; 7; 0; 20; 5; 12; 15; 12; 4; 5; 10.

Table 1.15 was produced.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	$\frac{2}{19}$	0.1053
2	3	$\frac{3}{19}$	0.2632
4	1	$\frac{1}{19}$	0.3158
5	3	$\frac{3}{19}$	0.4737
7	2	$\frac{2}{19}$	0.5789
10	2	$\frac{2}{19}$	0.6842
12	2	$\frac{2}{19}$	0.7895
15	1	$\frac{1}{19}$	0.8421
20	1	$\frac{1}{19}$	1.0000

Table 1.15 Frequency of Immigrant Survey Responses

a. Fix the errors in **Table 1.15**. Also, explain how someone might have arrived at the incorrect number(s).
b. Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the U.S. for 5 years.”
c. Fix the statement in **b** to make it correct.
d. What fraction of the people surveyed have lived in the U.S. five or seven years?
e. What fraction of the people surveyed have lived in the U.S. at most 12 years?
f. What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
g. What fraction of the people surveyed have lived in the U.S. from five to 20 years, inclusive?

42. How much time does it take to travel to work? **Table 1.16** shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

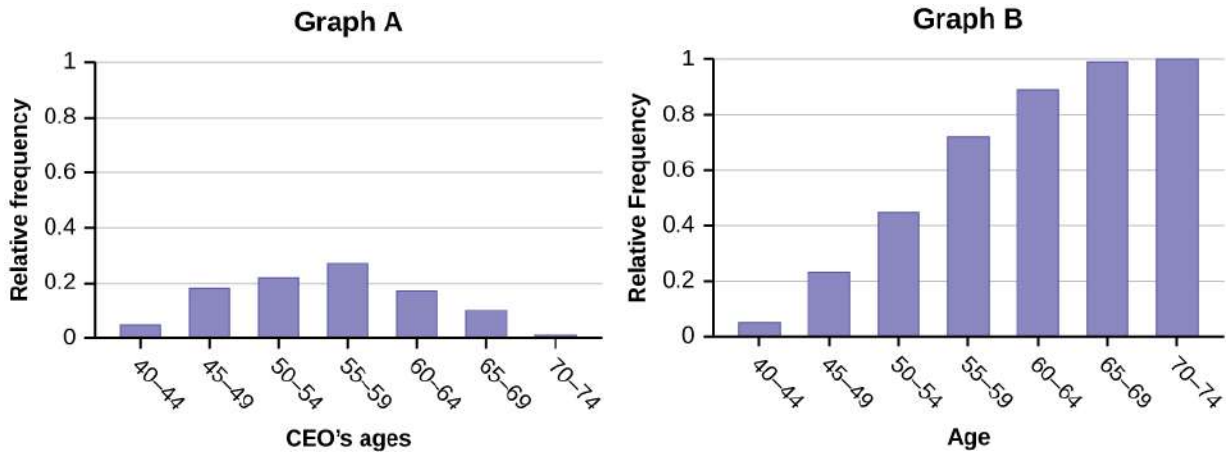
Table 1.16

43. *Forbes* magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. **Table 1.17** shows the ages of the chief executive officers for the first 60 ranked firms.

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

Table 1.17

a. What is the frequency for CEO ages between 54 and 65?
b. What percentage of CEOs are 65 years or older?
c. What is the relative frequency of ages under 50?
d. What is the cumulative relative frequency for CEOs younger than 55?
e. Which graph shows the relative frequency and which shows the cumulative relative frequency?



(a) (b)

Figure 1.11

Use the following information to answer the next two exercises: **Table 1.18** contains data on hurricanes that have made direct hits on the U.S. Between 1851 and 2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

Table 1.18 Frequency of Hurricane Direct Hits

44. What is the relative frequency of direct hits that were category 4 hurricanes?
a. 0.0768
b. 0.0659
c. 0.2601
d. Not enough information to calculate

45. What is the relative frequency of direct hits that were AT MOST a category 3 storm?
a. 0.3480
b. 0.9231
c. 0.2601
d. 0.3370

REFERENCES

1.1 Definitions of Statistics, Probability, and Key Terms

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html (accessed May 1, 2013).

1.2 Data, Sampling, and Variation in Data and Sampling

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx (accessed May 1, 2013).

Data from http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President

Dominic Lusinchi, "'President' Landon and the 1936 *Literary Digest* Poll: Were Automobile and Telephone Owners to Blame?" Social Science History 36, no. 1: 23-54 (2012), http://ssh.dukejournals.org/content/36/1/23.abstract (accessed May 1, 2013).

"The Literary Digest Poll," Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).

"Gallup Presidential Election Trial-Heat Trends, 1936–2008," Gallup Politics http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).

LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus (accessed May 1, 2013).

Data from San Jose Mercury News

1.3 Levels of Measurement

"State & County QuickFacts," U.S. Census Bureau. http://quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).

"State & County QuickFacts: Quick, easy access to facts about people, business, and geography," U.S. Census Bureau. http://quickfacts.census.gov/qfd/index.html (accessed May 1, 2013).

"Table 5: Direct hits by mainland United States Hurricanes (1851-2004)," National Hurricane Center, http://www.nhc.noaa.gov/gifs/table5.gif (accessed May 1, 2013).

"Levels of Measurement," http://infinity.cos.edu/faculty/woodbury/stats/tutorial/Data_Levels.htm (accessed May 1, 2013).

Courtney Taylor, "Levels of Measurement," about.com, http://statistics.about.com/od/HelpandTutorials/a/Levels-Of-Measurement.htm (accessed May 1, 2013).

David Lane. "Levels of Measurement," Connexions, http://cnx.org/content/m10809/latest/ (accessed May 1, 2013).

1.4 Experimental Design and Ethics

"Vitamin E and Health," Nutrition Source, Harvard School of Public Health, http://www.hsph.harvard.edu/nutritionsource/vitamin-e/ (accessed May 1, 2013).

Stan Reents. "Don't Underestimate the Power of Suggestion," athleteinme.com, http://www.athleteinme.com/ArticleView.aspx?id=1053 (accessed May 1, 2013).

Ankita Mehta. "Daily Dose of Aspiring Helps Reduce Heart Attacks: Study," International Business Times, July 21, 2011. Also available online at http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html (accessed May 1, 2013).

M.L. Jacskon et al., "Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors," Accident Analysis and Prevention Journal, Jan no. 50 (2013), http://www.ncbi.nlm.nih.gov/pubmed/22721550 (accessed May 1, 2013).

"Earthquake Information by Year," U.S. Geological Survey. http://earthquake.usgs.gov/earthquakes/eqarchives/year/ (accessed May 1, 2013).

"Fatality Analysis Report Systems (FARS) Encyclopedia," National Highway Traffic and Safety Administration. http://www-fars.nhtsa.dot.gov/Main/index.aspx (accessed May 1, 2013).

Data from www.businessweek.com (accessed May 1, 2013).

Data from www.forbes.com (accessed May 1, 2013).

"America's Best Small Companies," http://www.forbes.com/best-small-companies/list/ (accessed May 1, 2013).

U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.

"April 2013 Air Travel Consumer Report," U.S. Department of Transportation, April 11 (2013), http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report (accessed May 1, 2013).

Lori Alden, "Statistics can be Misleading," econoclass.com, http://www.econoclass.com/misleadingstats.html (accessed May 1, 2013).

Maria de los A. Medina, "Ethics in Statistics," Based on "Building an Ethics Module for Business, Science, and Engineering Students" by Jose A. Cruz-Cruz and William Frey, Connexions, http://cnx.org/content/m15555/latest/ (accessed May 1, 2013).

SOLUTIONS

2

a. all children who take ski or snowboard lessons
b. a group of these children
c. the population mean age of children who take their first snowboard lesson
d. the sample mean age of children who take their first snowboard lesson
e. X = the age of one child who takes his or her first ski or snowboard lesson
f. values for X, such as 3, 7, and so on

4

a. the clients of the insurance companies
b. a group of the clients
c. the mean health costs of the clients
d. the mean health costs of the sample
e. X = the health costs of one client
f. values for X, such as 34, 9, 82, and so on

6

a. all the clients of this counselor
b. a group of clients of this marriage counselor
c. the proportion of all her clients who stay married
d. the proportion of the sample of the counselor's clients who stay married
e. X = the number of couples who stay married
f. yes, no

8

a. all people (maybe in a certain geographic area, such as the United States)
b. a group of the people
c. the proportion of all people who will buy the product
d. the proportion of the sample who will buy the product
e. X = the number of people who will buy it
f. buy, not buy

10 a

12 quantitative discrete, 150

14 qualitative, Oakland A's

16 quantitative discrete, 11,234 students

18 qualitative, Crest

20 quantitative continuous, 47.3 years

22 b

24

a. The survey was conducted using six similar flights.
The survey would not be a true representation of the entire population of air travelers.
Conducting the survey on a holiday weekend will not produce representative results.
b. Conduct the survey during different times of the year.

Conduct the survey using flights to and from various locations.
Conduct the survey on different days of the week.

26 Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

28 Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can't be sure who is responding. In addition, mailing lists can be incomplete.

30 b

32 convenience; cluster; stratified ; systematic; simple random

34

a. qualitative(categorical)
b. quantitative discrete
c. quantitative discrete
d. qualitative(categorical)

36 Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate. Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

38

a. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed
b. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.

40

a.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	0.4500	0.4500
1	18	0.3000	0.7500
3	11	0.1833	0.9333
6	3	0.0500	0.9833
7	1	0.0167	1

Table 1.19

b. 5.00%
c. 93.33%

42 The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state's travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

44 b

2 | DESCRIPTIVE STATISTICS

Figure 2.1 When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

Introduction

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics."** You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we

will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

2.1 | Display Data

Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

One simple graph, the **stem-and-leaf graph** or **stemplot**, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example 2.1

For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):
33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Stem	Leaf
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

Table 2.1 Stem-and-Leaf Graph

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% $\left(\frac{8}{31}\right)$ were in the 90s or 100, a fairly high number of As.

Try It

2.1 For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):
32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61
Construct a stem plot for the data.

The stemplot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers, so we will cover them in more detail later.

Example 2.2

The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data: 1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3

Do the data seem to have any concentration of values?

NOTE

The leaves are to the right of the decimal.

Solution 2.2

The value 12.3 may be an outlier. Values appear to concentrate at three and four kilometers.

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

Table 2.2

Try It

2.2 The following data show the distances (in miles) from the homes of off-campus statistics students to the college. Create a stem plot using the data and identify any outliers:

0.5; 0.7; 1.1; 1.2; 1.2; 1.3; 1.3; 1.5; 1.5; 1.7; 1.7; 1.8; 1.9; 2.0; 2.2; 2.5; 2.6; 2.8; 2.8; 2.8; 3.5; 3.8; 4.4; 4.8; 4.9; 5.2; 5.5; 5.7; 5.8; 8.0

Example 2.3

A **side-by-side stem-and-leaf plot** allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. **Table 2.4** and **Table 2.5** show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Solution 2.3

Ages at Inauguration		Ages at Death
9 9 8 7 7 7 6 3 2	4	6 9
8 7 7 7 7 6 6 6 5 5 5 5 4 4 4 4 4 2 2 1 1 1 1 1 0	5	3 6 6 7 7 8
9 8 5 4 4 2 1 1 1 0	6	0 0 3 3 4 4 5 6 7 7 7 8
	7	0 0 1 1 1 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

Table 2.3

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51		

Table 2.4 Presidential Ages at Inauguration

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46

Table 2.5 Presidential Age at Death

President	Age	President	Age	President	Age
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Table 2.5 Presidential Age at Death

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in **Example 2.4**, the ***x*-axis** (horizontal axis) consists of **data values** and the ***y*-axis** (vertical axis) consists of **frequency points**. The frequency points are connected using line segments.

Example 2.4

In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in **Table 2.6** and in **Figure 2.2**.

Number of times teenager is reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4

Table 2.6

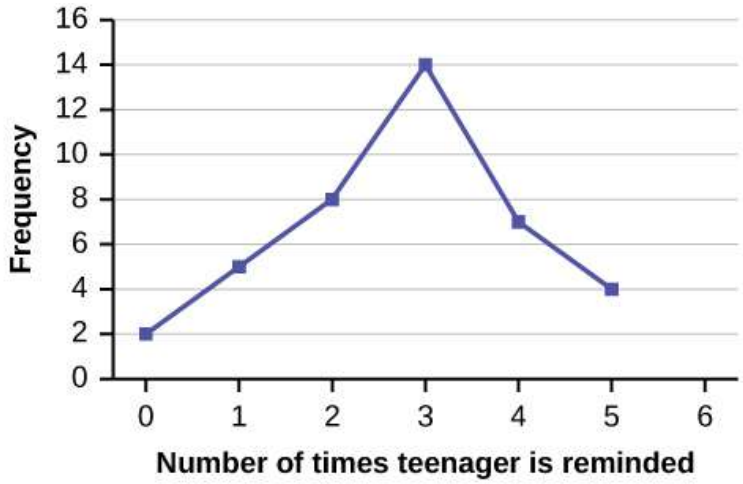


Figure 2.2

Try It

2.4 In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in **Table 2.7**. Construct a line graph.

Number of times in shop	Frequency
0	7
1	10
2	14
3	9

Table 2.7

Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graph** shown in **Example 2.5** has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

Example 2.5

By the end of 2011, Facebook had over 146 million users in the United States. **Table 2.7** shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

Table 2.8

Solution 2.5

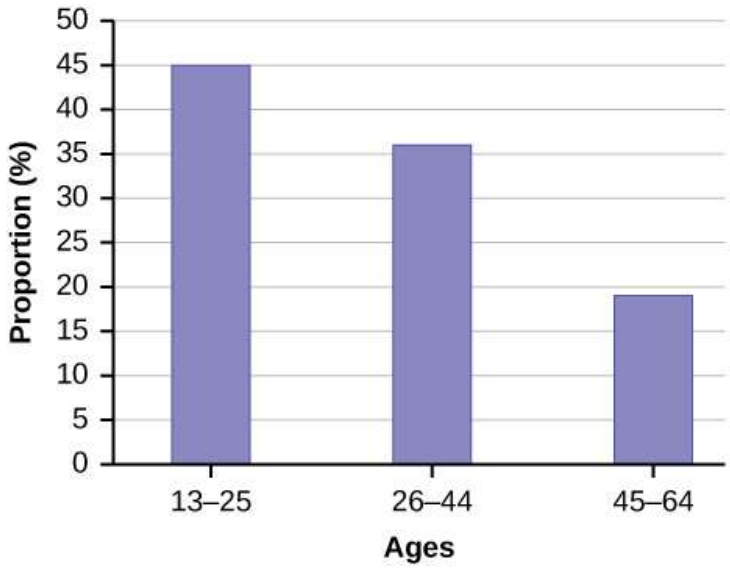


Figure 2.3

Try It

2.5 The population in Park City is made up of children, working-age adults, and retirees. **Table 2.9** shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

Table 2.9

Example 2.6

The columns in **Table 2.9** contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the *x*-axis, and the Advanced Placement examinee population percentages on the *y*-axis.

Race/Ethnicity	**AP Examinee Population**	**Overall Student Population**
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

Table 2.10

Solution 2.6

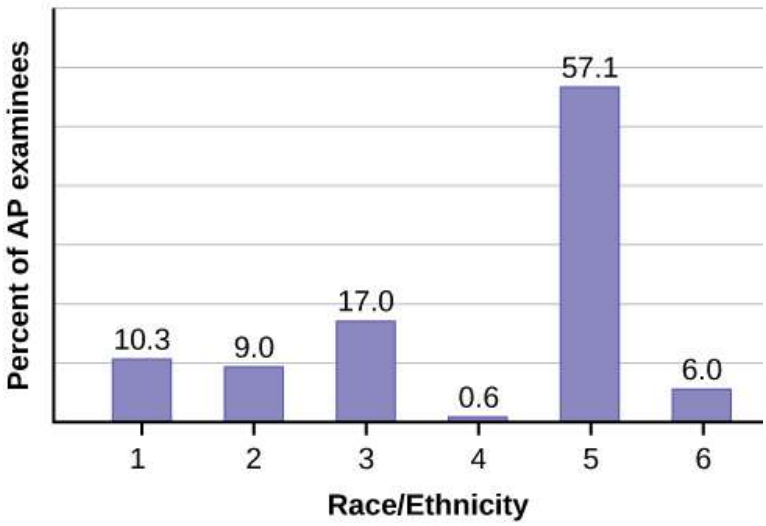


Figure 2.4

Try It

2.6 Park city is broken down into six voting districts. The table shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

Table 2.11

Example 2.7

Below is a two-way table showing the types of pets owned by men and women:

	Dogs	Cats	Fish	Total
Men	4	2	2	8
Women	4	6	2	12
Total	8	8	4	20

Table 2.12

Given these data, calculate the conditional distributions for the subpopulation of men who own each pet type.

Solution 2.7

Men who own dogs = 4/8 = 0.5

Men who own cats = 2/8 = 0.25

Men who own fish = 2/8 = 0.25

Note: The sum of all of the conditional distributions must equal one. In this case, 0.5 + 0.25 + 0.25 = 1; therefore, the solution "checks".

Histograms, Frequency Polygons, and Time Series Graphs

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either **frequency** or **relative frequency** (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample.(Remember, frequency is defined as the number of times an answer occurs.) If:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then:

$$RF = \frac{f}{n}$$

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, $f = 3$, $n = 40$, and $RF = \frac{f}{n} = \frac{3}{40} = 0.075$. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 – 0.05 = 6.05). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 (1.5 – 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 – 0.0005 = 0.9995). If all the data happen to be integers and the smallest value is two, then a convenient starting point is 1.5 (2 – 0.5 = 1.5). Also, when the starting point and other boundaries are carried to one additional decimal place, no data

value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

Example 2.8

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data, since height is measured.
60; 60.5; 61; 61; 61.5
63.5; 63.5; 63.5
64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5
66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5
68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5
70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71
72; 72; 72; 72.5; 72.5; 73; 73.5
74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

60 – 0.05 = 59.95 which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

The largest value is 74, so 74 + 0.05 = 74.05 is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose eight bars.

$$\frac{74.05 - 59.95}{8} = 1.76$$

NOTE

We will round up to two and make each bar or class interval two units wide. Rounding up to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. A guideline that is followed by some for the width of a bar or class interval is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through 71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is

in the interval 73.95–75.95.

The following histogram displays the heights on the *x*-axis and relative frequency on the *y*-axis.

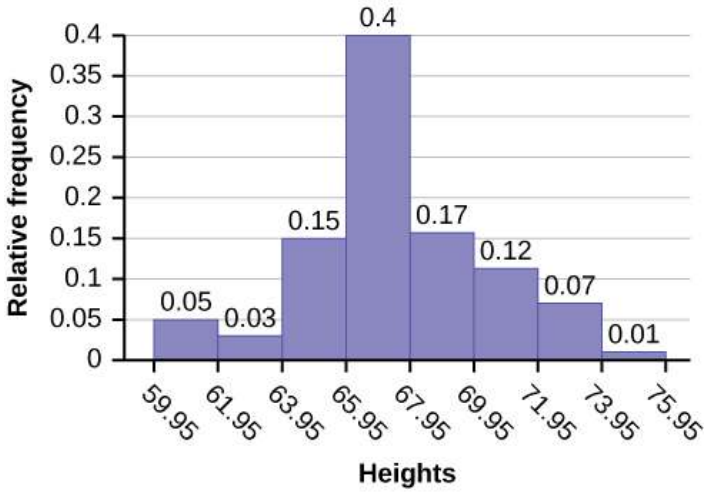


Figure 2.5

Try It

2.8 The following data are the shoe sizes of 50 male students. The sizes are continuous data since shoe size is measured. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars.
9; 9; 9.5; 9.5; 10; 10; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5
11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5
12; 12; 12; 12; 12; 12; 12; 12.5; 12.5; 12.5; 12.5; 14

Example 2.9

Create a histogram for the following data: the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.
1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4
5; 5; 5; 5; 5
6; 6

Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.

Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6, and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _____ to _____, the 5 in the middle of the interval from _____ to _____, and the _____ in the middle of the interval from _____ to _____ .

Solution 2.9

- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Calculate the number of bars as follows:

$$\frac{6.5 - 0.5}{\text{number of bars}} = 1$$

where 1 is the width of a bar. Therefore, bars = 6.

The following histogram displays the number of books on the *x*-axis and the frequency on the *y*-axis.

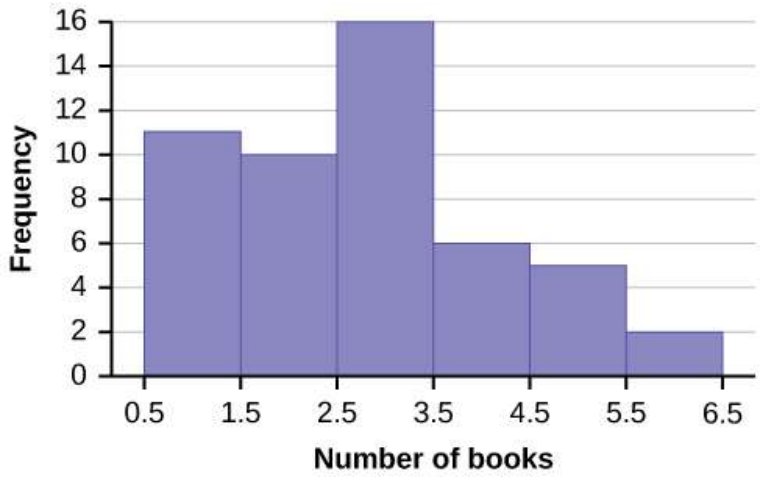


Figure 2.6

Example 2.10

Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends				
9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Table 2.13

Solution 2.10

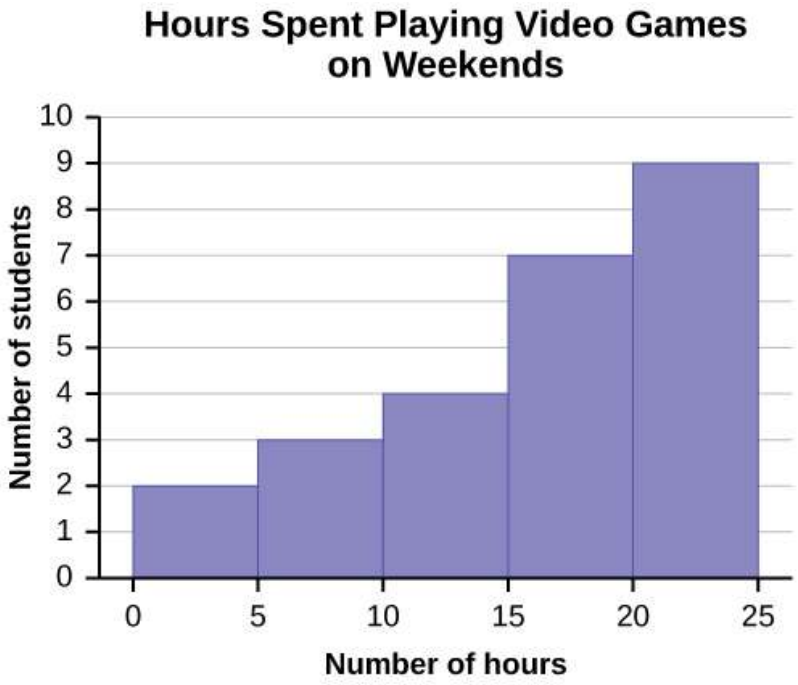


Figure 2.7

Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary, but not if it falls on the right boundary. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

Frequency Polygons

Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons.

To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the *x*-axis and *y*-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

Example 2.11

A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Table 2.14

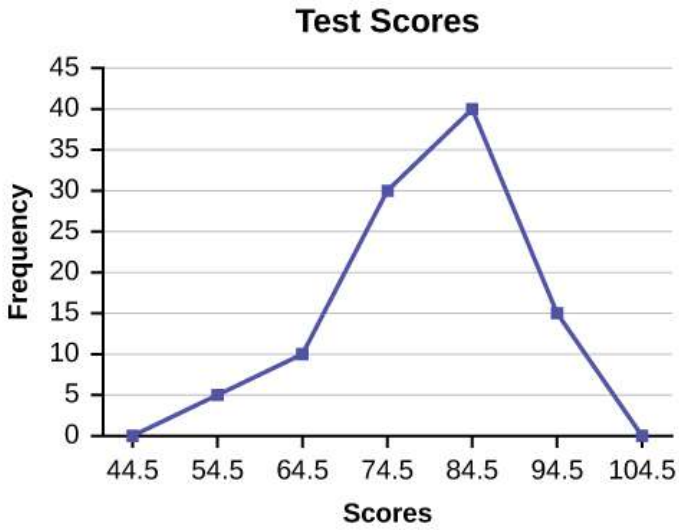


Figure 2.8

The first label on the *x*-axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 54.5 represents the next interval, or the first "real" interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

Try It

2.11 Construct a frequency polygon of U.S. Presidents' ages at inauguration shown in **Table 2.15**.

Age at Inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	2

Table 2.15

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

Example 2.12

We will construct an overlay frequency polygon comparing the scores from **Example 2.11** with the students' final numeric grade.

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Table 2.16

Frequency Distribution for Calculus Final Grades			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100

Table 2.17

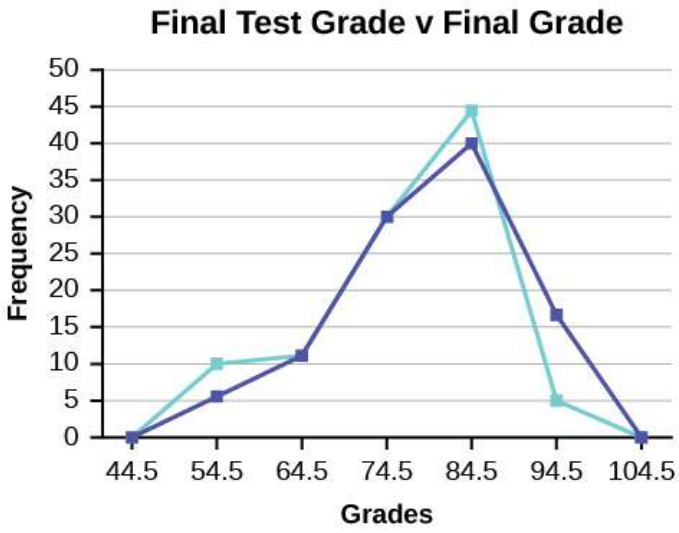


Figure 2.9

Constructing a Time Series Graph

Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with these data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we don‘t have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

To construct a time series graph, we must look at both pieces of our **paired data set**. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

Example 2.13

The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Table 2.18

Year	Aug	Sep	Oct	Nov	Dec	Annual
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Table 2.19

Solution 2.13

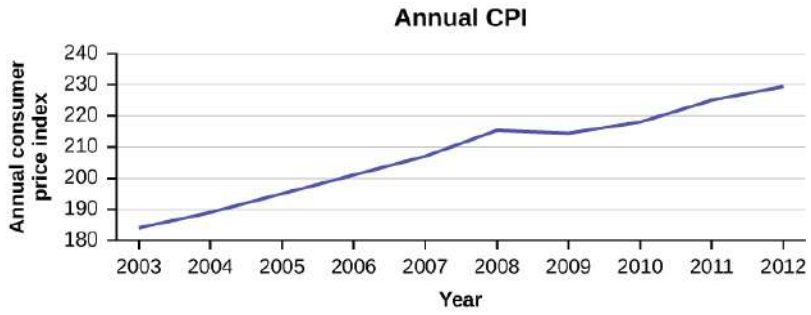


Figure 2.10

Try It

2.13 The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO_2 emissions for the United States.

CO2 Emissions			
	Ukraine	United Kingdom	United States
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

Table 2.20

Uses of a Time Series Graph

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

How NOT to Lie with Statistics

It is important to remember that the very reason we develop a variety of methods to present data is to develop insights into the subject of what the observations represent. We want to get a "sense" of the data. Are the observations all very much alike or are they spread across a wide range of values, are they bunched at one end of the spectrum or are they distributed evenly and so on. We are trying to get a visual picture of the numerical data. Shortly we will develop formal mathematical measures of the data, but our visual graphical presentation can say much. It can, unfortunately, also say much that is distracting, confusing and simply wrong in terms of the impression the visual leaves. Many years ago Darrell Huff wrote the book *How to Lie with Statistics*. It has been through 25 plus printings and sold more than one and one-half million copies. His perspective was a harsh one and used many actual examples that were designed to mislead. He wanted to make people

aware of such deception, but perhaps more importantly to educate so that others do not make the same errors inadvertently.

Again, the goal is to enlighten with visuals that tell the story of the data. Pie charts have a number of common problems when used to convey the message of the data. Too many pieces of the pie overwhelm the reader. More than perhaps five or six categories ought to give an idea of the relative importance of each piece. This is after all the goal of a pie chart, what subset matters most relative to the others. If there are more components than this then perhaps an alternative approach would be better or perhaps some can be consolidated into an "other" category. Pie charts cannot show changes over time, although we see this attempted all too often. In federal, state, and city finance documents pie charts are often presented to show the components of revenue available to the governing body for appropriation: income tax, sales tax motor vehicle taxes and so on. In and of itself this is interesting information and can be nicely done with a pie chart. The error occurs when two years are set side-by-side. Because the total revenues change year to year, but the size of the pie is fixed, no real information is provided and the relative size of each piece of the pie cannot be meaningfully compared.

Histograms can be very helpful in understanding the data. Properly presented, they can be a quick visual way to present probabilities of different categories by the simple visual of comparing relative areas in each category. Here the error, purposeful or not, is to vary the width of the categories. This of course makes comparison to the other categories impossible. It does embellish the importance of the category with the expanded width because it has a greater area, inappropriately, and thus visually "says" that that category has a higher probability of occurrence.

Time series graphs perhaps are the most abused. A plot of some variable across time should never be presented on axes that change part way across the page either in the vertical or horizontal dimension. Perhaps the time frame is changed from years to months. Perhaps this is to save space or because monthly data was not available for early years. In either case this confounds the presentation and destroys any value of the graph. If this is not done to purposefully confuse the reader, then it certainly is either lazy or sloppy work.

Changing the units of measurement of the axis can smooth out a drop or accentuate one. If you want to show large changes, then measure the variable in small units, penny rather than thousands of dollars. And of course to continue the fraud, be sure that the axis does not begin at zero, zero. If it begins at zero, zero, then it becomes apparent that the axis has been manipulated.

Perhaps you have a client that is concerned with the volatility of the portfolio you manage. An easy way to present the data is to use long time periods on the time series graph. Use months or better, quarters rather than daily or weekly data. If that doesn't get the volatility down then spread the time axis relative to the rate of return or portfolio valuation axis. If you want to show "quick" dramatic growth, then shrink the time axis. Any positive growth will show visually "high" growth rates. Do note that if the growth is negative then this trick will show the portfolio is collapsing at a dramatic rate.

Again, the goal of descriptive statistics is to convey meaningful visuals that tell the story of the data. Purposeful manipulation is fraud and unethical at the worst, but even at its best, making these type of errors will lead to confusion on the part of the analysis.

2.2 | Measures of the Location of the Data

The common measures of location are **quartiles** and **percentiles**

Quartiles are special percentiles. The first quartile, Q_1, is the same as the 25^{th} percentile, and the third quartile, Q_3, is the same as the 75^{th} percentile. The median, M, is called both the second quartile and the 50^{th} percentile.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90^{th} percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75^{th} percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.

1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1
Ordered from smallest to largest:
1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.

$$\frac{6.8 + 7.2}{2} = 7$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:
1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.
1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the **first quartile**. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The **third quartile**, $Q3$, is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$IQR = Q_3 - Q_1$

The IQR can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than (1.5)(IQR) below the first quartile or more than (1.5)(IQR) above the third quartile**. Potential outliers always require further investigation.

NOTE

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

Example 2.14

For the following 13 real estate prices, calculate the IQR and determine if any prices are potential outliers. Prices are in dollars.
389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Solution 2.14

Order the data from smallest to largest.
114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

$M = 488{,}800$

$$Q_1 = \frac{230{,}500 + 387{,}000}{2} = 308{,}750$$

$$Q_3 = \frac{639{,}000 + 659{,}000}{2} = 649{,}000$$

$IQR = 649{,}000 - 308{,}750 = 340{,}250$

(1.5)(*IQR*) = (1.5)(340,250) = 510,375

Q_1 – (1.5)(*IQR*) = 308,750 – 510,375 = –201,625

Q_3 + (1.5)(*IQR*) = 649,000 + 510,375 = 1,159,375

No house price is less than –201,625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Example 2.15

For the two data sets in the **test scores example**, find the following:

a. The interquartile range. Compare the two interquartile ranges.

b. Any outliers in either set.

Solution 2.15

The five number summary for the day and night classes is

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

Table 2.21

a. The IQR for the day group is $Q_3 – Q_1 = 82.5 – 56 = 26.5$
The IQR for the night group is $Q_3 – Q_1 = 89 – 78 = 11$

The interquartile range (the spread or variability) for the day class is larger than the night class *IQR*. This suggests more variation will be found in the day class's class test scores.

b. Day class outliers are found using the IQR times 1.5 rule. So,
Q_1 - *IQR*(1.5) = 56 – 26.5(1.5) = 16.25
Q_3 + *IQR*(1.5) = 82.5 + 26.5(1.5) = 122.25
Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

Night class outliers are calculated as:

Q_1 – *IQR* (1.5) = 78 – 11(1.5) = 61.5
Q_3 + IQR(1.5) = 89 + 11(1.5) = 105.5
For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

Example 2.16

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Table 2.22

Find the 28th percentile. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the fours, fives, sixes and sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th percentile, then, must be an eight**. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Try It

2.16 Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Table 2.23

Example 2.17

Using **Table 2.22**:

a. Find the 80th percentile.
b. Find the 90th percentile.
c. Find the first quartile. What is another name for the first quartile?

Solution 2.17

Using the data from the frequency table, we have:

a. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th an 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$
b. The 90th percentile will be the 45th data value (location is 0.90(50) = 45) and the 45th data value is nine.
c. Q_1 is also the 25th percentile. The 25th percentile location calculation: P_{25} = 0.25(50) = 12.5 ≈ 13 the 13th data value. Thus, the 25th percentile is six.

A Formula for Finding the *k*th Percentile

If you were to do a little research, you would find several formulas for calculating the k^{th} percentile. Here is one of them.

k = the k^{th} percentile. It may or may not be part of the data.

i = the index (ranking or position of a data value)

n = the total number of data points, or observations

- Order the data from smallest to largest.
- Calculate $i = \frac{k}{100}(n+1)$
- If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Example 2.18

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*
18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 70th percentile.
b. Find the 83rd percentile.

Solution 2.18

a. k = 70
i = the index
n = 29
$i = \frac{k}{100}(n+1) = (\frac{70}{100})(29+1) = 21$. Twenty-one is an integer, and the data value in the 21st position in the ordered data set is 64. The 70th percentile is 64 years.

b. k = 83rd percentile
i = the index
n = 29

$i = \frac{k}{100}(n + 1) =)\frac{83}{100})(29 + 1) = 24.9$, which is NOT an integer. Round it down to 24 and up to 25. The age in the 24th position is 71 and the age in the 25th position is 72. Average 71 and 72. The 83rd percentile is 71.5 years.

Try It

2.18 Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77
Calculate the 20th percentile and the 55th percentile.

A Formula for Finding the Percentile of a Value in a Data Set

- Order the data from smallest to largest.
- x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- y = the number of data values equal to the data value for which you want to find the percentile.
- n = the total number of data.
- Calculate $\frac{x + 0.5y}{n}(100)$. Then round to the nearest integer.

Example 2.19

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*
18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile for 58.
b. Find the percentile for 25.

Solution 2.19

a. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.
$x = 18$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{18 + 0.5(1)}{29}(100) = 63.80$. 58 is the 64th percentile.

b. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.
$x = 3$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{3 + 0.5(1)}{29}(100) = 12.07$. Twenty-five is the 12th percentile.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

NOTE

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

Example 2.20

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Solution 2.20

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Example 2.21

On a 20 question math test, the 70^{th} percentile for number of correct answers was 16. Interpret the 70^{th} percentile in the context of this situation.

Solution 2.21

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

Try It

2.21 On a 60 point written assignment, the 80^{th} percentile for the number of points earned was 49. Interpret the 80^{th} percentile in the context of this situation.

Example 2.22

At a community college, it was found that the 30^{th} percentile of credit units that students are enrolled for is seven units. Interpret the 30^{th} percentile in the context of this situation.

Solution 2.22

- Thirty percent of students are enrolled in seven or fewer credit units.

- Seventy percent of students are enrolled in seven or more credit units.
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Example 2.23

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes

10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;

30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

Determine the following five values.

Min = 0
Q_1 = 20
Med = 40
Q_3 = 60
Max = 300

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the *IQR* is 40 minutes (60 – 20 = 40), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.

However, the principal needs to be careful. The value 300 appears to be a potential outlier.

$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120$.

The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

Min = 0
Q_1 = 20
Q_3 = 60
Max = 120

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

2.3 | Measures of the Center of the Data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. Technically this is the arithmetic mean. We will discuss the geometric mean later. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts meaning an equal number of observations on each side. The weight of 25 people are below this weight and 25 people are heavier than this weight. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

NOTE

The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. Formally, the arithmetic mean is called the first moment of the distribution by mathematicians. However, in practice among non-statisticians, “average" is commonly accepted for “arithmetic mean.”

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the **sample mean** is an x with a bar over it (pronounced “x bar”): $\bar{x}$.

The Greek letter μ (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:
1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7$$

$$\bar{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7$$

In the second calculation, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median is the 49^{th} value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$. The median occurs midway between the 50^{th} and 51^{st} values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example 2.24

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;
Calculate the mean and the median.

Solution 2.24

The calculation for the mean is:

$$\bar{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6$$

To find the median, M, first use the formula for the location. The location is:

$$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$$

Starting at the smallest value, the median is located between the 20^{th} and 21^{st} values (the two 24s):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

$$M = \frac{24 + 24}{2} = 24$$

Example 2.25

Suppose that in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center": the mean or the median?

Solution 2.25

$$\bar{x} = \frac{5,\ 000,\ 000 + 49(30,\ 000)}{50} = 129{,}400$$

M = 30,000

(There are 49 people who earn $30,000 and one person who earns $5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

Example 2.26

Statistics exam scores for 20 students are as follows:

50; 53; 59; 59; 63; 63; 72; 72; 72; 72; 72; 76; 78; 81; 83; 84; 84; 84; 90; 93

Find the mode.

Solution 2.26

The most frequent score is 72, which occurs five times. Mode = 72.

Example 2.27

Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.

When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

NOTE

The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Calculating the Arithmetic Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval

frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean: $mean = \frac{data\ sum}{number\ of\ data\ values}$ We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is $\frac{lower\ boundary + upper\ boundary}{2}$. We can now modify the mean definition to be $Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$ where f = the frequency of the interval and m = the midpoint of the interval.

Example 2.28

A frequency table displaying professor Blount's last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Table 2.24

Solution 2.28

- Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

Table 2.25

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm$

$$53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$$

- $\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$

Try It

2.28 Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	Number of Teenagers
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

Table 2.26

What is the best estimate for the mean number of hours spent playing video games?

2.4 | Sigma Notation and Calculating the Arithmetic Mean

Formula for Population Mean

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Formula for Sample Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This unit is here to remind you of material that you once studied and said at the time "I am sure that I will never need this!"

Here are the formulas for a population mean and the sample mean. The Greek letter μ is the symbol for the population mean and $\bar{x}$ is the symbol for the sample mean. Both formulas have a mathematical symbol that tells us how to make the calculations. It is called Sigma notation because the symbol is the Greek capital letter sigma: Σ. Like all mathematical symbols it tells us what to do: just as the plus sign tells us to add and the x tells us to multiply. These are called mathematical operators. The Σ symbol tells us to add a specific list of numbers.

Let's say we have a sample of animals from the local animal shelter and we are interested in their average age. If we list each value, or observation, in a column, you can give each one an index number. The first number will be number 1 and the second number 2 and so on.

Animal	Age
1	9
2	1
3	8.5
4	10.5
5	10
6	8.5
7	12
8	8
9	1
10	9.5

Table 2.27

Each observation represents a particular animal in the sample. Purr is animal number one and is a 9 year old cat, Toto is animal number 2 and is a 1 year old puppy and so on.

To calculate the mean we are told by the formula to add up all these numbers, ages in this case, and then divide the sum by 10, the total number of animals in the sample.

Animal number one, the cat Purr, is designated as X_1, animal number 2, Toto, is designated as X_2 and so on through Dundee who is animal number 10 and is designated as X_{10}.

The i in the formula tells us which of the observations to add together. In this case it is X_1 through X_{10} which is all of them. We know which ones to add by the indexing notation, the i = 1 and the n or capital N for the population. For this example the indexing notation would be i = 1 and because it is a sample we use a small n on the top of the Σ which would be 10.

The standard deviation requires the same mathematical operator and so it would be helpful to recall this knowledge from your past.

The sum of the ages is found to be 78 and dividing by 10 gives us the sample mean age as 7.8 years.

2.5 | Geometric Mean

The mean (Arithmetic), median and mode are all measures of the “center” of the data, the “average”. They are all in their own way trying to measure the “common” point within the data, that which is “normal”. In the case of the arithmetic mean this is solved by finding the value from which all points are equal linear distances. We can imagine that all the data values are combined through addition and then distributed back to each data point in equal amounts. The sum of all the values is what is redistributed in equal amounts such that the total sum remains the same.

The geometric mean redistributes not the sum of the values but the product of multiplying all the individual values and then redistributing them in equal portions such that the total product remains the same. This can be seen from the formula for the geometric mean, $\tilde{x}$: *(Pronounced x-tilde)*

$$\tilde{x} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} = \sqrt[n]{x_1 * x_2 \bullet\bullet\bullet x_n} = (x_1 * x_2 \bullet\bullet\bullet x_n)^{\frac{1}{n}}$$

where π is another mathematical operator, that tells us to multiply all the x_i numbers in the same way capital Greek sigma tells us to add all the x_i numbers. Remember that a fractional exponent is calling for the nth root of the number thus an exponent of 1/3 is the cube root of the number.

The geometric mean answers the question, "if all the quantities had the same value, what would that value have to be in order to achieve the same product?” The geometric mean gets its name from the fact that when redistributed in this way the sides form a geometric shape for which all sides have the same length. To see this, take the example of the numbers 10, 51.2 and 8. The geometric mean is the product of multiplying these three numbers together (4,096) and taking the cube

root because there are three numbers among which this product is to be distributed. Thus the geometric mean of these three numbers is 16. This describes a cube 16x16x16 and has a volume of 4,096 units.

The geometric mean is relevant in Economics and Finance for dealing with growth: growth of markets, in investment, population and other variables the growth in which there is an interest. Imagine that our box of 4,096 units (perhaps dollars) is the value of an investment after three years and that the investment returns in percents were the three numbers in our example. The geometric mean will provide us with the answer to the question, what is the average rate of return: 16 percent. The arithmetic mean of these three numbers is 23.6 percent. The reason for this difference, 16 versus 23.6, is that the arithmetic mean is additive and thus does not account for the interest on the interest, compound interest, embedded in the investment growth process. The same issue arises when asking for the average rate of growth of a population or sales or market penetration, etc., knowing the annual rates of growth. The formula for the geometric mean rate of return, or any other growth rate, is:

$$r_s = (x_1 * x_2 \bullet\bullet\bullet x_n)^{\frac{1}{n}} - 1$$

Manipulating the formula for the geometric mean can also provide a calculation of the average rate of growth between two periods knowing only the initial value a_0 and the ending value a_n and the number of periods, n. The following formula provides this information:

$$\left(\frac{a_n}{a_0}\right)^{\frac{1}{n}} = \tilde{x}$$

Finally, we note that the formula for the geometric mean requires that all numbers be positive, greater than zero. The reason of course is that the root of a negative number is undefined for use outside of mathematical theory. There are ways to avoid this problem however. In the case of rates of return and other simple growth problems we can convert the negative values to meaningful positive equivalent values. Imagine that the annual returns for the past three years are +12%, -8%, and +2%. Using the decimal multiplier equivalents of 1.12, 0.92, and 1.02, allows us to compute a geometric mean of 1.0167. Subtracting 1 from this value gives the geometric mean of +1.67% as a net rate of population growth (or financial return). From this example we can see that the geometric mean provides us with this formula for calculating the geometric (mean) rate of return for a series of annual rates of return:

$$r_s = \tilde{x} - 1$$

where r_s is average rate of return and $\tilde{x}$ is the geometric mean of the returns during some number of time periods. Note that the length of each time period must be the same.

As a general rule one should convert the percent values to its decimal equivalent multiplier. It is important to recognize that when dealing with percents, the geometric mean of percent values does not equal the geometric mean of the decimal multiplier equivalents and it is the decimal multiplier equivalent geometric mean that is relevant.

2.6 | Skewness and the Mean, Median, and Mode

Consider the following data set.
4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.

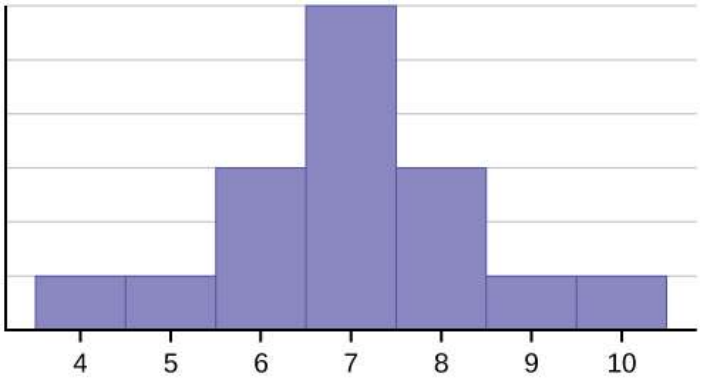


Figure 2.11

The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left. We can formally measure the skewness of a distribution just as we can mathematically measure the center weight of the data or its general "speadness". The mathematical formula for skewness is: $a_3 = \sum \frac{\left(x_i - \bar{x}\right)^3}{ns^3}$. The greater the deviation from zero indicates a greater degree of skewness. If the skewness is negative then the distribution is skewed left as in **Figure 2.12**. A positive measure of skewness indicates right skewness such as **Figure 2.13**.

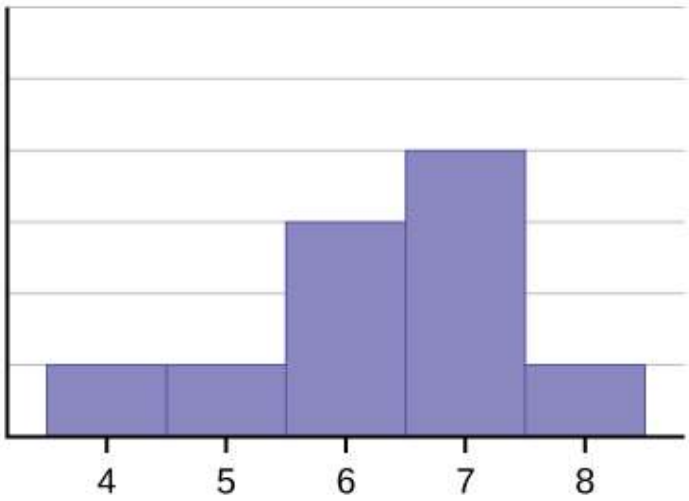


Figure 2.12

The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10, is also not symmetrical. It is **skewed to the right**.

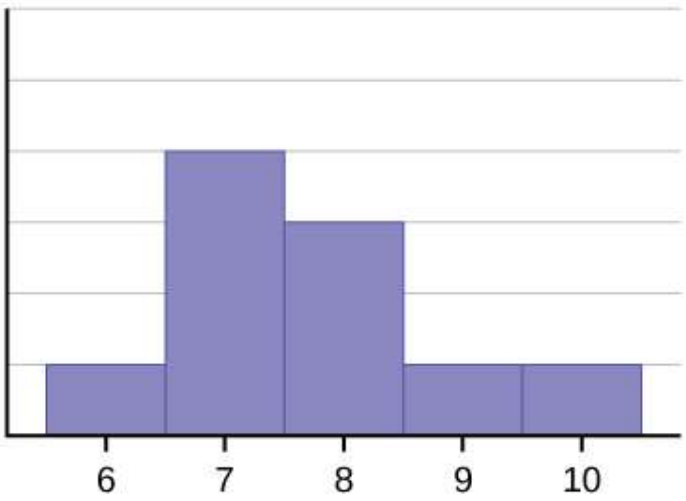


Figure 2.13

The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

As with the mean, median and mode, and as we will see shortly, the variance, there are mathematical formulas that give us precise measures of these characteristics of the distribution of the data. Again looking at the formula for skewness we see that this is a relationship between the mean of the data and the individual observations cubed.

$$a_3 = \sum \frac{\left(x_i - \bar{x}\right)^3}{ns^3}$$

where s is the sample standard deviation of the data, X_i , and $\bar{x}$ is the arithmetic mean and n is the sample size.

Formally the arithmetic mean is known as the first moment of the distribution. The second moment we will see is the variance, and skewness is the third moment. The variance measures the squared differences of the data from the mean and skewness measures the cubed differences of the data from the mean. While a variance can never be a negative number, the measure of skewness can and this is how we determine if the data are skewed right of left. The skewness for a normal distribution is zero, and any symmetric data should have skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. The skewness characterizes the degree of asymmetry of a distribution around its mean. While the mean and standard deviation are *dimensional* quantities (this is why we will take the square root of the variance) that is, have the same units as the measured quantities X_i , the skewness is conventionally defined in such a way as to make it *nondimensional*. It is a pure number that characterizes only the shape of the distribution. A positive value of skewness signifies a distribution with an asymmetric tail extending out towards more positive X and a negative value signifies a distribution whose tail extends out towards more negative X. A zero measure of skewness will indicate a symmetrical distribution.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

2.7 | Measures of the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. The average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B*. The standard deviation for the wait time is four minutes.

Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average.

Calculating the Standard Deviation

If x is a number, then the difference "x minus the mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - \bar{x}$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations** (the $x - \bar{x}$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations. Formally, the variance is the second moment of the distribution or the first moment around the mean. Remember that the mean is the first moment of the distribution.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by $\boldsymbol{n - 1}$, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

- $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\left(\sum_{i=1}^{n} x^2\right) - n\bar{x}^2}{n - 1}}$
- For the sample standard deviation, the denominator is $\boldsymbol{n - 1}$, that is the sample size minus 1.

Formulas for the Population Standard Deviation

- $\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f(x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2}$
- For the population standard deviation, the denominator is N, the number of items in the population.

In these formulas, f represents the frequency with which a value appears. For example, if a value appears once, f is one. If a value appears three times in the data set or population, f is three. Two important observations concerning the variance and standard deviation: the deviations are measured from the mean and the deviations are squared. In principle, the deviations could be measured from any point, however, our interest is measurement from the center weight of the data, what is the "normal" or most usual value of the observation. Later we will be trying to measure the "unusualness" of an observation or a sample mean and thus we need a measure from the mean. The second observation is that the deviations are squared. This does two things, first it makes the deviations all positive and second it changes the units of measurement from that of the mean and the original observations. If the data are weights then the mean is measured in pounds, but the variance

is measured in pounds-squared. One reason to use the standard deviation is to return to the original units of measurement by taking the square root of the variance. Further, when the deviations are squared it explodes their value. For example, a deviation of 10 from the mean when squared is 100, but a deviation of 100 from the mean is 10,000. What this does is place great weight on outliers when calculating the variance.

Types of Variability in Samples

When trying to study a population, a sample is often used, either for convenience or because it is not possible to access the entire population. Variability is the term used to describe the differences that may occur in these outcomes. Common types of variability include the following:

- Observational or measurement variability
- Natural variability
- Induced variability
- Sample variability

Here are some examples to describe each type of variability.

Example 1: Measurement variability

Measurement variability occurs when there are differences in the instruments used to measure or in the people using those instruments. If we are gathering data on how long it takes for a ball to drop from a height by having students measure the time of the drop with a stopwatch, we may experience measurement variability if the two stopwatches used were made by different manufacturers: For example, one stopwatch measures to the nearest second, whereas the other one measures to the nearest tenth of a second. We also may experience measurement variability because two different people are gathering the data. Their reaction times in pressing the button on the stopwatch may differ; thus, the outcomes will vary accordingly. The differences in outcomes may be affected by measurement variability.

Example 2: Natural variability

Natural variability arises from the differences that naturally occur because members of a population differ from each other. For example, if we have two identical corn plants and we expose both plants to the same amount of water and sunlight, they may still grow at different rates simply because they are two different corn plants. The difference in outcomes may be explained by natural variability.

Example 3: Induced variability

Induced variability is the counterpart to natural variability; this occurs because we have artificially induced an element of variation (that, by definition, was not present naturally): For example, we assign people to two different groups to study memory, and we induce a variable in one group by limiting the amount of sleep they get. The difference in outcomes may be affected by induced variability.

Example 4: Sample variability

Sample variability occurs when multiple random samples are taken from the same population. For example, if I conduct four surveys of 50 people randomly selected from a given population, the differences in outcomes may be affected by sample variability.

Example 2.29

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

$$\bar{x} = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.

The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s.

Data	Freq.	Deviations	$Deviations^2$	(Freq.)($Deviations^2$)
x	f	$(x - \bar{x})$	$(x - \bar{x})^2$	$(f)(x - \bar{x})^2$
9	1	9 – 10.525 = –1.525	$(-1.525)^2 = 2.325625$	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(-1.025)^2 = 1.050625$	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = –0.525	$(-0.525)^2 = 0.275625$	4 × 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = –0.025	$(-0.025)^2 = 0.000625$	4 × 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2 = 0.225625$	6 × 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2 = 0.950625$	3 × 0.950625 = 2.851875
				The total is 9.7375

Table 2.28

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$$s^2 = \frac{9.7375}{20 - 1} = 0.5125$$

The **sample standard deviation** s is equal to the square root of the sample variance:

$s = \sqrt{0.5125} = 0.715891$, which is rounded to two decimal places, $s = 0.72$.

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. (For **Example 2.29**, there are $n = 20$ deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation. By squaring the deviations we are placing an extreme penalty on observations that are far from the mean; these observations get greater weight in the calculations of the variance. We will see later on that the variance (standard deviation) plays the critical role in determining our conclusions in inferential statistics. We can begin now by using the standard deviation as a measure of "unusualness." "How did you do on the test?" "Terrific! Two standard deviations above the mean." This, we will see, is an unusually good exam grade.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation divided by $n - 1 = 20 - 1 = 19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one ($n - 1$). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** This estimate requires us to use an estimate of the population mean rather than the actual population mean. Based on the theoretical mathematics that lies behind these calculations, dividing by ($n - 1$) gives a better estimate of the population variance.

The standard deviation, s or σ, is either zero or larger than zero. Describing the data with reference to the spread is called "variability". The variability in data depends upon the method by which the outcomes are obtained; for example, by measuring or by random sampling. When the standard deviation is zero, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

Example 2.30

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.

b. Calculate the following to one decimal place:

   i. The sample mean

   ii. The sample standard deviation

   iii. The median

   iv. The first quartile

   v. The third quartile

   vi. *IQR*

Solution 2.30

a. See **Table 2.29**

b. i. The sample mean = 73.5

   ii. The sample standard deviation = 17.9

   iii. The median = 73

   iv. The first quartile = 61

   v. The third quartile = 90

   vi. *IQR* = 90 – 61 = 29

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612

Table 2.29

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1? ANSWER: Rounding)

Table 2.29

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula: $Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$

where $f =$ interval frequencies and m = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how "unusual" individual data is compared to the mean.

Example 2.31

Find the standard deviation for the data in **Table 2.30**.

Class	Frequency, f	Midpoint, m	$f*m$	$f(m - \bar{x})^2$
0–2	1	1	$1*1=1$	$1(1 - 7.58)^2 = 43.26$
3–5	6	4	$6*4=24$	$6(4 - 7.58)^2 = 76.77$
6-8	10	7	$10*7=70$	$10(7 - 7.58)^2 = 3.33$
9-11	7	10	$7*10=70$	$7(10 - 7.58)^2 = 41.10$
12-14	0	13	$0*13=0$	$0(13 - 7.58)^2 = 0$
	26=n		$\bar{x} = \frac{197}{26} = 7.58$	$s^2 = \frac{306.35}{26 - 1} = 12.25$

Table 2.30

For this data set, we have the mean, $\bar{x} = 7.58$ and the standard deviation, $s_x = 3.5$. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since 7.58 – 3.5 – 3.5 = 0.58. While the formula for calculating the standard deviation is not complicated, $s_x = \sqrt{\frac{\Sigma(m - \bar{x})^2 f}{n - 1}}$ where

s_x = sample standard deviation, $\bar{x}$ = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value x, calculate how many standard deviations away from its mean the value is.
- Use the formula: x = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#\ of STDEVs = \frac{\text{x – mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "*z*-score"; we can use the symbol *z*. In symbols, the formulas become:

Sample	$x = \bar{x} + zs$	$z = \frac{x - \bar{x}}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Table 2.31

Example 2.32

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Table 2.32

Solution 2.32

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$z = \#\text{ of STDEVs} = \frac{\text{value – mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$

For John, $z = \#\ of STDEVs = \frac{2.85 - 3.0}{0.7} = -0.21$

For Ali, $z = \#\ of STDEVs = \frac{77 - 80}{10} = -0.3$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's *z*-score of –0.21 is higher than Ali's *z*-score of –0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Try It

2.32 Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Table 2.33

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a Normal Distribution, which we will examine in great detail later:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

Coefficient of Variation

Another useful way to compare distributions besides simple comparisons of means or standard deviations is to adjust for differences in the scale of the data being measured. Quite simply, a large variation in data with a large mean is different than the same variation in data with a small mean. To adjust for the scale of the underlying data the Coefficient of Variation (CV) has been developed. Mathematically:

$$CV = \frac{s}{\bar{x}} * 100 \text{ conditioned upon } \bar{x} \neq 0, \text{ where } s \text{ is the standard deviation of the data and } \bar{x} \text{ is the mean.}$$

We can see that this measures the variability of the underlying data as a percentage of the mean value; the center weight of the data set. This measure is useful in comparing risk where an adjustment is warranted because of differences in scale of two data sets. In effect, the scale is changed to common scale, percentage differences, and allows direct comparison of the two or more magnitudes of variation of different data sets.

KEY TERMS

Frequency the number of times a value of the data occurs

Frequency Table a data representation in which grouped data is displayed along with the corresponding frequencies

Histogram a graphical representation in *x*-*y* form of the distribution of data in a data set; *x* represents the data and *y* represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

Interquartile Range or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Mean (arithmetic) a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Mean (geometric) a measure of central tendency that provides a measure of average geometric growth over multiple time periods.

Median a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Midpoint the mean of an interval in a frequency table

Mode the value that appears most frequently in a set of data

Outlier an observation that does not fit the rest of the data

Percentile a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50^{th} percentile. The first and third quartiles are the 25^{th} and the 75^{th} percentiles, respectively.

Quartiles the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

Relative Frequency the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes

Standard Deviation a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation.

Variance mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - \bar{x}$ where *x* is a value of the data and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

CHAPTER REVIEW

2.1 Display Data

A **stem-and-leaf plot** is a way to plot data and look at the distribution. In a stem-and-leaf plot, all data values within a class are visible. The advantage in a stem-and-leaf plot is that all values are listed, unlike a histogram, which gives classes of data values. A **line graph** is often used to represent a set of data values in which a quantity varies with time. These graphs are useful for finding trends. That is, finding a general pattern in data sets including temperature, sales, employment, company profit or cost over a period of time. A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis

represents a discrete value. Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulative effect (stacked bar graphs). Bar graphs are especially useful when categorical data is being used.

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually goes on *y*-axis with the frequency being graphed on the *x*-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.

2.2 Measures of the Location of the Data

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50^{th} percentile would be greater than 50 percent of the other obeservations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25^{th} percentile,the second quartile (Q_2 or median) is 50^{th} percentile, and the third quartile (Q_3) is the the 75^{th} percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3, and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

2.3 Measures of the Center of the Data

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occuring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

2.6 Skewness and the Mean, Median, and Mode

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of distributions. A **right (or positive) skewed** distribution has a shape like **Figure 2.12**. A **left (or negative) skewed** distribution has a shape like **Figure 2.13**. A **symmetrical** distrubtion looks like **Figure 2.11**.

2.7 Measures of the Spread of the Data

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ or $s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n-1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}}$.

FORMULA REVIEW

2.2 Measures of the Location of the Data

$$i = \left(\frac{k}{100}\right)(n+1)$$

where i = the ranking or position of a data value,

k = the kth percentile,

n = total number of data.

Expression for finding the percentile of a data value: $\left(\frac{x + 0.5y}{n}\right)(100)$

where x = the number of values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile,

y = the number of data values equal to the data value for which you want to find the percentile,

n = total number of data

2.3 Measures of the Center of the Data

$\mu = \frac{\sum fm}{\sum f}$ Where f = interval frequencies and m = interval midpoints.

The arithmetic mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$

The arithmetic mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$

2.5 Geometric Mean

The Geometric Mean: $\tilde{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 * x_2 \bullet\bullet\bullet x_n} = (x_1 * x_2 \bullet\bullet\bullet x_n)^{\frac{1}{n}}$

2.6 Skewness and the Mean, Median, and Mode

Formula for skewness: $a_3 = \sum \frac{\left(x_i - \bar{x}\right)^3}{ns^3}$

Formula for Coefficient of Variation: $CV = \frac{s}{\bar{x}} * 100$ conditioned upon $\bar{x} \neq 0$

2.7 Measures of the Spread of the Data

$s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2}$ where

$s_x =$ sample standard deviation

$\bar{x} =$ sample mean

Formulas for Sample Standard Deviation $s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{n - 1}}$ or $s = \sqrt{\frac{\left(\sum_{i=1}^{n} x^2\right) - n\bar{x}^2}{n - 1}}$ For the sample standard deviation, the denominator is ***n*** **- 1**, that is the sample size - 1.

Formulas for Population Standard Deviation $\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\Sigma f(x - \mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N} - \mu^2}$ For the population standard deviation, the denominator is N, the number of items in the population.

PRACTICE

2.1 Display Data

For the next three exercises, use the data to construct a line graph.

1. In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown in **Table 2.34**.

Number of times in store	Frequency
1	4
2	10
3	16
4	6
5	4

Table 2.34

2. In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown in **Table 2.35**.

Years since last purchase	Frequency
0	2
1	8
2	13
3	22
4	16
5	9

Table 2.35

3. Several children were asked how many TV shows they watch each day. The results of the survey are shown in **Table 2.36**.

Number of TV Shows	Frequency
0	12
1	18
2	36
3	7
4	2

Table 2.36

4. The students in Ms. Ramirez's math class have birthdays in each of the four seasons. **Table 2.37** shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the number of students.

Seasons	Number of students	Proportion of population
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

Table 2.37

5. Using the data from Mrs. Ramirez's math class supplied in **Exercise 2.4**, construct a bar graph showing the percentages.

6. David County has six high schools. Each school sent students to participate in a county-wide science competition. **Table 2.38** shows the percentage breakdown of competitors from each school, and the percentage of the entire student population of the county that goes to each school. Construct a bar graph that shows the population percentage of competitors from each school.

High School	Science competition population	Overall student population
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%
Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%
Tynneson	24.2%	10.1%
West End	8.7%	28.8%

Table 2.38

7. Use the data from the David County science competition supplied in **Exercise 2.6**. Construct a bar graph that shows the county-wide population percentage of students at each school.

8. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Complete the table.

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Table 2.39

9. What does the frequency column in **Table 2.39** sum to? Why?

10. What does the relative frequency column in **Table 2.39** sum to? Why?

11. What is the difference between relative frequency and frequency for each data value in **Table 2.39**?

12. What is the difference between cumulative relative frequency and relative frequency for each data value?

13. To construct the histogram for the data in **Table 2.39**, determine appropriate minimum and maximum x and y values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.

Figure 2.14

14. Construct a frequency polygon for the following:

a.

Pulse Rates for Women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

Table 2.40

b.

Actual Speed in a 30 MPH Zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

Table 2.41

c.

Tar (mg) in Nonfiltered Cigarettes	Frequency
10–13	1
14–17	0
18–21	15
22–25	7
26–29	2

Table 2.42

15. Construct a frequency polygon from the frequency distribution for the 50 highest ranked countries for depth of hunger.

Depth of Hunger	Frequency
230–259	21
260–289	13
290–319	5
320–349	7
350–379	1
380–409	1
410–439	1

Table 2.43

16. Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlayed frequency polygon and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?

Life Expectancy at Birth – Women	Frequency
49–55	3
56–62	3
63–69	1
70–76	3
77–83	8
84–90	2

Table 2.44

Life Expectancy at Birth – Men	Frequency
49–55	3
56–62	3
63–69	1
70–76	1
77–83	7
84–90	5

Table 2.45

17. Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.

Sex/Year	1855	1856	1857	1858	1859	1860	1861
Female	45,545	49,582	50,257	50,324	51,915	51,220	52,403
Male	47,804	52,239	53,158	53,694	54,628	54,409	54,606
Total	93,349	101,821	103,415	104,018	106,543	105,629	107,009

Table 2.46

Sex/Year	1862	1863	1864	1865	1866	1867	1868	1869
Female	51,812	53,115	54,959	54,850	55,307	55,527	56,292	55,033
Male	55,257	56,226	57,374	58,220	58,360	58,517	59,222	58,321
Total	107,069	109,341	112,333	113,070	113,667	114,044	115,514	113,354

Table 2.47

Sex/Year	1870	1871	1872	1873	1874	1875
Female	56,431	56,099	57,472	58,233	60,109	60,146
Male	58,959	60,029	61,293	61,467	63,602	63,432
Total	115,390	116,128	118,765	119,700	123,711	123,578

Table 2.48

18. The following data sets list full time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1973.

Year	1961	1962	1963	1964	1965	1966	1967
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Table 2.49

Year	1968	1969	1970	1971	1972	1973
Police	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	28.03	31.49	37.39	46.26	47.24	52.33

Table 2.50

a. Construct a double time series graph using a common *x*-axis for both sets of data.
b. Which variable increased the fastest? Explain.
c. Did Detroit's increase in police officers have an impact on the murder rate? Explain.

2.2 Measures of the Location of the Data

19. Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 40^{th} percentile.
b. Find the 78^{th} percentile.

20. Listed are 32 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 18; 21; 22; 25; 26; 27; 29; 30; 31; 31; 33; 36; 37; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile of 37.
b. Find the percentile of 72.

21. Jesse was ranked 37^{th} in his graduating class of 180 students. At what percentile is Jesse's ranking?

22.

a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
b. The 20^{th} percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20^{th} percentile in the context of the situation.
c. A bicyclist in the 90^{th} percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90^{th} percentile in the context of the situation.

23.

a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
b. The 40^{th} percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40^{th} percentile in the context of the situation.

24. On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

25. Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85^{th} percentile of wait times. Is that good or bad? Write a sentence interpreting the 85^{th} percentile in the context of this situation.

26. In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78^{th} percentile. Should Li be pleased or upset by this result? Explain.

27. In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1,700 in damage and was in the 90^{th} percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90^{th} percentile in the context of this problem.

28. The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
b. Students whose GPAs are at or above the 96^{th} percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are "eligible in the local context"?

29. Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34^{th} percentile. The 34^{th} percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Use the following information to answer the next six exercises. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

30. First quartile = _____

31. Second quartile = median = 50^{th} percentile = _____

32. Third quartile = _____

33. Interquartile range (*IQR*) = _____ – _____ = _____

34. 10^{th} percentile = _____

35. 70^{th} percentile = _____

2.3 Measures of the Center of the Data

36. Find the mean for the following frequency tables.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

Table 2.51

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

Table 2.52

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Table 2.53

Use the following information to answer the next three exercises: The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest: 16; 17; 19; 20; 20; 21; 23; 24; 25; 25; 25; 26; 26; 27; 27; 27; 28; 29; 30; 32; 33; 33; 34; 35; 37; 39; 40

37. Calculate the mean.

38. Identify the median.

39. Identify the mode.

Use the following information to answer the next three exercises: Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Calculate the following:

40. sample mean = $\bar{x}$ = _____

41. median = _____

42. mode = _____

2.4 Sigma Notation and Calculating the Arithmetic Mean

43. A group of 10 children are on a scavenger hunt to find different color rocks. The results are shown in the **Table 2.54** below. The column on the right shows the number of colors of rocks each child has. What is the mean number of rocks?

Child	Rock Colors
1	5
2	5
3	6
4	2
5	4
6	3
7	7
8	2
9	1
10	10

Table 2.54

44. A group of children are measured to determine the average height of the group. The results are in **Table 2.55** below. What is the mean height of the group to the nearest hundredth of an inch?

Child	Height in Inches
Adam	45.21
Betty	39.45
Charlie	43.78
Donna	48.76
Earl	37.39
Fran	39.90
George	45.56
Heather	46.24

Table 2.55

45. A person compares prices for five automobiles. The results are in **Table 2.56**. What is the mean price of the cars the person has considered?

Price
$20,987
$22,008
$19,998
$23,433
$21,444

Table 2.56

46. A customer protection service has obtained 8 bags of candy that are supposed to contain 16 ounces of candy each. The candy is weighed to determine if the average weight is at least the claimed 16 ounces. The results are in given in **Table 2.57**. What is the mean weight of a bag of candy in the sample?

Weight in Ounces
15.65
16.09
16.01
15.99
16.02
16.00
15.98
16.08

Table 2.57

47. A teacher records grades for a class of 70, 72, 79, 81, 82, 82, 83, 90, and 95. What is the mean of these grades?

48. A family is polled to see the mean of the number of hours per day the television set is on. The results, starting with Sunday, are 6, 3, 2, 3, 1, 3, and 7 hours. What is the average number of hours the family had the television set on to the nearest whole number?

49. A city received the following rainfall for a recent year. What is the mean number of inches of rainfall the city received monthly, to the nearest hundredth of an inch? Use **Table 2.58**.

Month	Rainfall in Inches
January	2.21
February	3.12
March	4.11
April	2.09
May	0.99
June	1.08
July	2.99
August	0.08
September	0.52
October	1.89
November	2.00
December	3.06

Table 2.58

50. A football team scored the following points in its first 8 games of the new season. Starting at game 1 and in order the scores are 14, 14, 24, 21, 7, 0, 38, and 28. What is the mean number of points the team scored in these eight games?

2.5 Geometric Mean

51. What is the geometric mean of the data set given? 5, 10, 20

52. What is the geometric mean of the data set given? 9.000, 15.00, 21.00

53. What is the geometric mean of the data set given? 7.0, 10.0, 39.2

54. What is the geometric mean of the data set given? 17.00, 10.00, 19.00

55. What is the average rate of return for the values that follow? 1.0, 2.0, 1.5

56. What is the average rate of return for the values that follow? 0.80, 2.0, 5.0

57. What is the average rate of return for the values that follow? 0.90, 1.1, 1.2

58. What is the average rate of return for the values that follow? 4.2, 4.3, 4.5

2.6 Skewness and the Mean, Median, and Mode

Use the following information to answer the next three exercises: State whether the data are symmetrical, skewed to the left, or skewed to the right.

59. 1; 1; 1; 2; 2; 2; 2; 3; 3; 3; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5

60. 16; 17; 19; 22; 22; 22; 22; 22; 23

61. 87; 87; 87; 87; 87; 88; 89; 89; 90; 91

62. When the data are skewed left, what is the typical relationship between the mean and median?

63. When the data are symmetrical, what is the typical relationship between the mean and median?

64. What word describes a distribution that has two modes?

65. Describe the shape of this distribution.

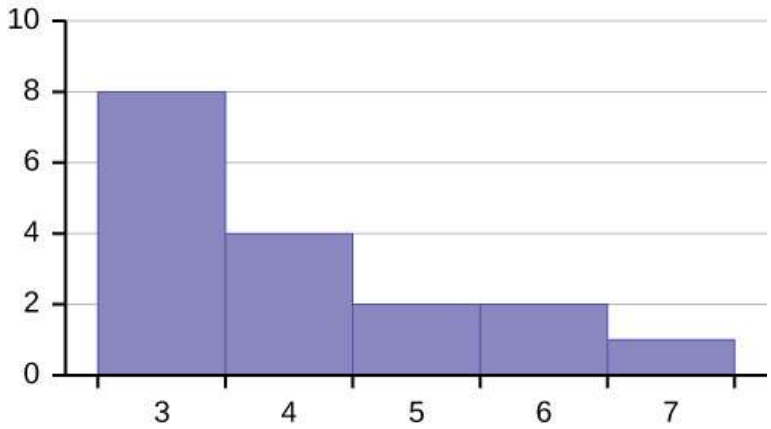


Figure 2.15

66. Describe the relationship between the mode and the median of this distribution.

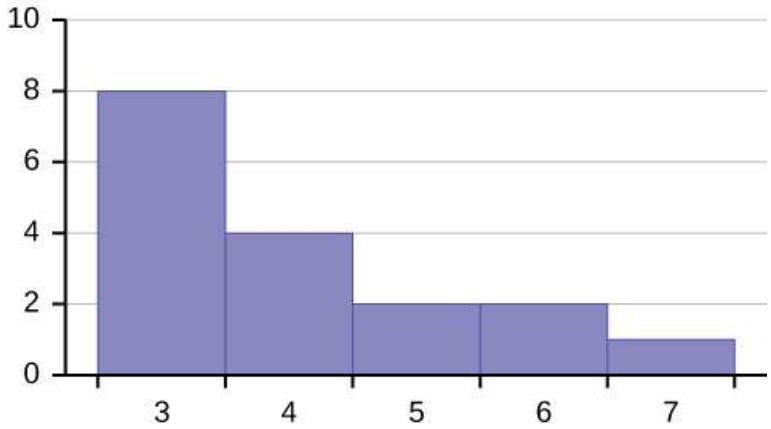


Figure 2.16

67. Describe the relationship between the mean and the median of this distribution.

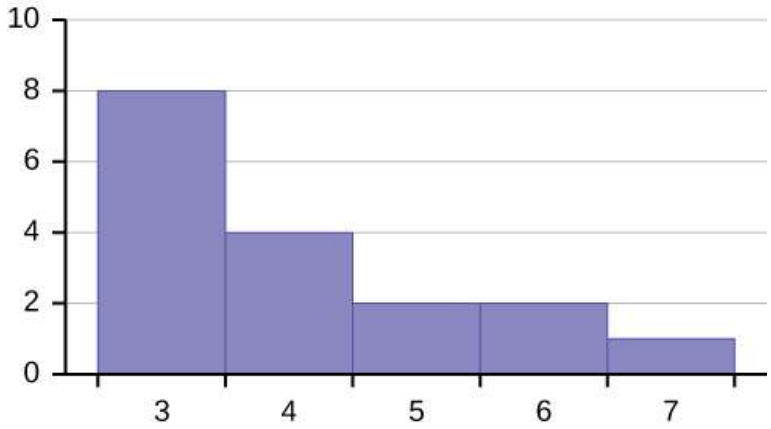


Figure 2.17

68. Describe the shape of this distribution.

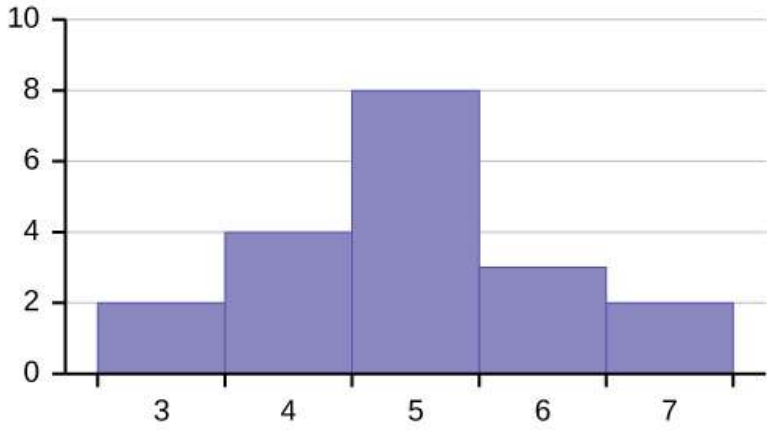


Figure 2.18

69. Describe the relationship between the mode and the median of this distribution.

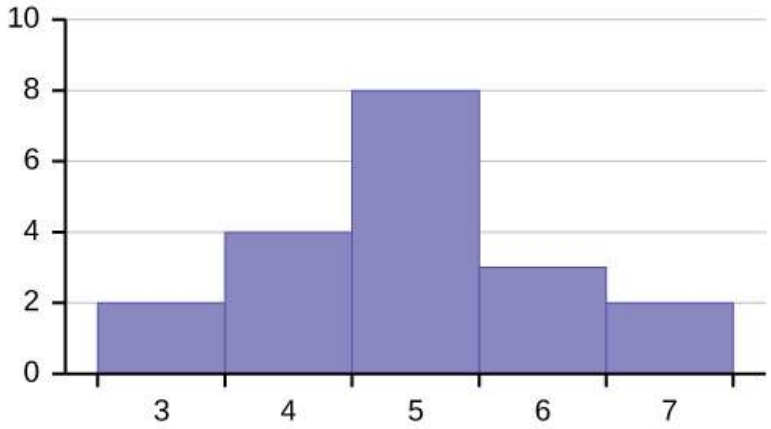


Figure 2.19

70. Are the mean and the median the exact same in this distribution? Why or why not?

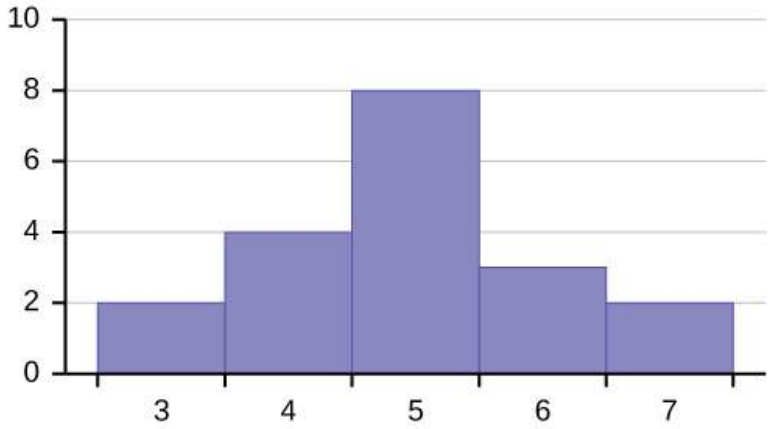


Figure 2.20

71. Describe the shape of this distribution.

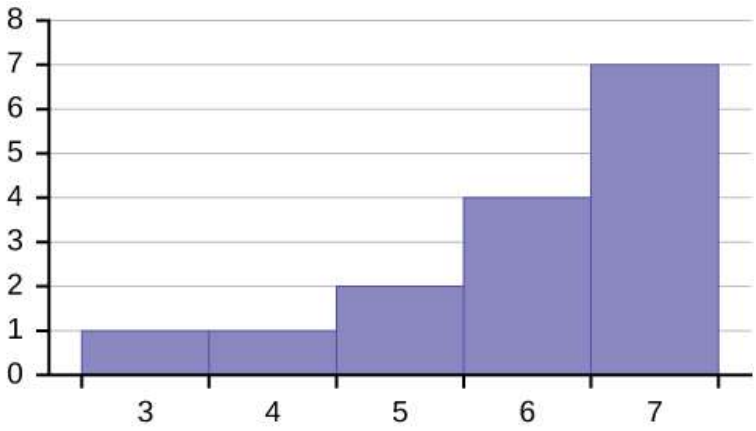


Figure 2.21

72. Describe the relationship between the mode and the median of this distribution.

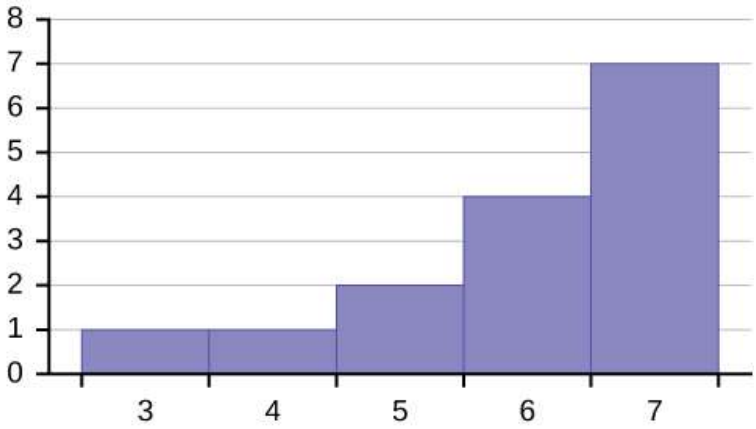


Figure 2.22

73. Describe the relationship between the mean and the median of this distribution.

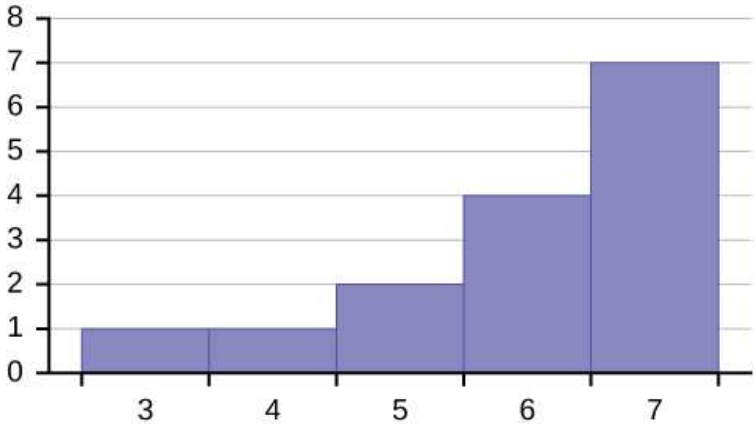


Figure 2.23

74. The mean and median for the data are the same.

3; 4; 5; 5; 6; 6; 6; 6; 7; 7; 7; 7; 7; 7; 7

Is the data perfectly symmetrical? Why or why not?

75. Which is the greatest, the mean, the mode, or the median of the data set?

11; 11; 12; 12; 12; 12; 13; 15; 17; 22; 22; 22

76. Which is the least, the mean, the mode, and the median of the data set?

56; 56; 56; 58; 59; 60; 62; 64; 64; 65; 67

77. Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

78. In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

2.7 Measures of the Spread of the Data

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.
29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

79. Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

80. Find the value that is one standard deviation below the mean.

81. Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Table 2.59

82. Use **Table 2.59** to find the value that is three standard deviations:
a. above the mean
b. below the mean

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

83. Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

Table 2.60

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

Table 2.61

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Table 2.62

HOMEWORK

2.1 Display Data

84. **Table 2.63** contains the 2010 obesity rates in U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Table 2.63

a. Use a random number generator to randomly pick eight states. Construct a bar graph of the obesity rates of those eight states.
b. Construct a bar graph for all the states beginning with the letter "A."
c. Construct a bar graph for all the states beginning with the letter "M."

85. Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Table 2.64 Publisher A

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Table 2.65 Publisher B

# of books	Freq.	Rel. Freq.
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

Table 2.66 Publisher C

a. Find the relative frequencies for each survey. Write them in the charts.
b. Use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of one. For Publisher C, make bar widths of two.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.
f. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

86. Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group.

Amount($)	Frequency	Rel. Frequency
51–100	5	
101–150	10	
151–200	15	
201–250	15	
251–300	10	
301–350	5	

Table 2.67 Singles

Amount($)	Frequency	Rel. Frequency
100–150	5	
201–250	5	
251–300	5	
301–350	5	
351–400	10	
401–450	10	
451–500	10	
501–550	10	
551–600	5	
601–650	5	

Table 2.68 Couples

a. Fill in the relative frequency for each group.
b. Construct a histogram for the singles group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
c. Construct a histogram for the couples group. Scale the *x*-axis by $50 widths. Use relative frequency on the *y*-axis.
d. Compare the two graphs:
    i. List two similarities between the graphs.
    ii. List two differences between the graphs.
    iii. Overall, are the graphs more similar or different?
e. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the *x*-axis by $50, scale it by $100. Use relative frequency on the *y*-axis.
f. Compare the graph for the singles with the new graph for the couples:
    i. List two similarities between the graphs.
    ii. Overall, are the graphs more similar or different?
g. How did scaling the couples graph differently change the way you compared it to the singles graph?
h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

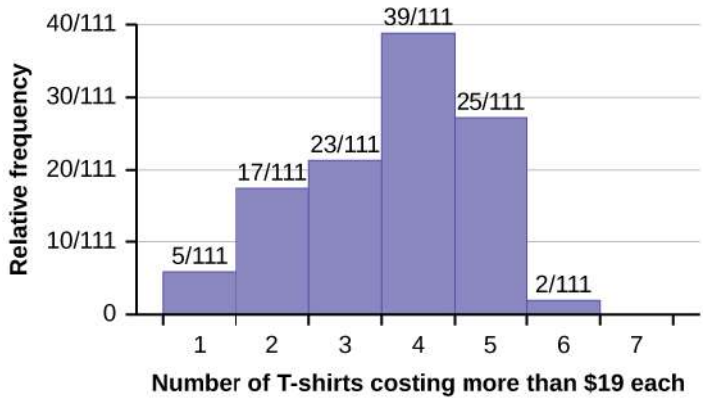
87. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

Table 2.69

a. Construct a histogram of the data.
b. Complete the columns of the chart.

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than $19 each.



88. The percentage of people who own at most three t-shirts costing more than $19 each is approximately:
a. 21
b. 59
c. 41
d. Cannot be determined

89. If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:
a. cluster
b. simple random
c. stratified
d. convenience

90. Following are the 2010 obesity rates by U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Table 2.70

Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the *x*-axis with the states.

2.2 Measures of the Location of the Data

91. The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years.

a. Based upon this information, give two reasons why the black median age could be lower than the white median age.
b. Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?
c. How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

92. Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in **Table 2.71**. Also, include left endpoint, but not the right endpoint.

Salary ($)	Relative Frequency
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

Table 2.71

a. What percentage of the survey answered "not sure"?
b. What percentage think that middle-class is from $25,000 to $50,000?
c. Construct a histogram of the data.
    i. Should all bars have the same width, based on the data? Why or why not?
    ii. How should the <20,000 and the 100,000+ intervals be handled? Why?
d. Find the 40^{th} and 80^{th} percentiles
e. Construct a bar graph of the data

2.3 Measures of the Center of the Data

93. The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

Table 2.72

a. What is the best estimate of the average obesity percentage for these countries?
b. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
c. How does the United States compare to other countries?

94. **Table 2.73** gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

Table 2.73

2.4 Sigma Notation and Calculating the Arithmetic Mean

95. A sample of 10 prices is chosen from a population of 100 similar items. The values obtained from the sample, and the values for the population, are given in **Table 2.74** and **Table 2.75** respectively.

a. Is the mean of the sample within $1 of the population mean?
b. What is the difference in the sample and population means?

Prices of the Sample
$21
$23
$21
$24
$22
$22
$25
$21
$20
$24

Table 2.74

Prices of the Population	Frequency
$20	20
$21	35
$22	15
$23	10
$24	18
$25	2

Table 2.75

96. A standardized test is given to ten people at the beginning of the school year with the results given in **Table 2.76** below. At the end of the year the same people were again tested.

a. What is the average improvement?
b. Does it matter if the means are subtracted, or if the individual values are subtracted?

Student	Beginning Score	Ending Score
1	1100	1120
2	980	1030
3	1200	1208
4	998	1000
5	893	948
6	1015	1030
7	1217	1224
8	1232	1245
9	967	988
10	988	997

Table 2.76

97. A small class of 7 students has a mean grade of 82 on a test. If six of the grades are 80, 82,86, 90, 90, and 95, what is the other grade?

98. A class of 20 students has a mean grade of 80 on a test. Nineteen of the students has a mean grade between 79 and 82, inclusive.

a. What is the lowest possible grade of the other student?
b. What is the highest possible grade of the other student?

99. If the mean of 20 prices is $10.39, and 5 of the items with a mean of $10.99 are sampled, what is the mean of the other 15 prices?

2.5 Geometric Mean

100. An investment grows from $10,000 to $22,000 in five years. What is the average rate of return?

101. An initial investment of $20,000 grows at a rate of 9% for five years. What is its final value?

102. A culture contains 1,300 bacteria. The bacteria grow to 2,000 in 10 hours. What is the rate at which the bacteria grow per hour to the nearest tenth of a percent?

103. An investment of $3,000 grows at a rate of 5% for one year, then at a rate of 8% for three years. What is the average rate of return to the nearest hundredth of a percent?

104. An investment of $10,000 goes down to $9,500 in four years. What is the average return per year to the nearest hundredth of a percent?

2.6 Skewness and the Mean, Median, and Mode

105. The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

a. What does it mean for the median age to rise?
b. Give two reasons why the median age could rise.
c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

2.7 Measures of the Spread of the Data

Use the following information to answer the next nine exercises: The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- μ = 1000 FTES
- median = 1,014 FTES
- σ = 474 FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- n = 29 years

106. A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

107. 75% of all years have an FTES:
a. at or below: _____
b. at or above: _____

108. The population standard deviation = _____

109. What percent of the FTES were from 528.5 to 1447.5? How do you know?

110. What is the *IQR*? What does the *IQR* represent?

111. How many standard deviations away from the mean is the median?

Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Year	2005–06	2006–07	2007–08	2008–09	2009–10	2010–11
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

Table 2.77

112. Calculate the mean, median, standard deviation, the first quartile, the third quartile and the *IQR*. Round to one decimal place.

113. Compare the *IQR* for the FTES for 1976–77 through 2004–2005 with the *IQR* for the FTES for 2005-2006 through 2010–2011. Why do you suppose the *IQR*s are so different?

114. Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	**GPA**	**School Average GPA**	**School Standard Deviation**
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Table 2.78

115. A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.

116. An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain why.

117. The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in **Table 14**.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

Table 2.79

What is the best estimate of the average obesity percentage for these countries? What is the standard deviation for the listed obesity rates? The United States has an average obesity rate of 33.9%. Is this rate above average or below? How "unusual" is the United States' obesity rate compared to the average rate? Explain.

118. **Table 2.80** gives the percent of children under five considered to be underweight.

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

Table 2.80

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

BRINGING IT TOGETHER: HOMEWORK

119. Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

	Javier	Ercilia
$\bar{x}$	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

Table 2.81

a. How can you determine which survey was correct ?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?

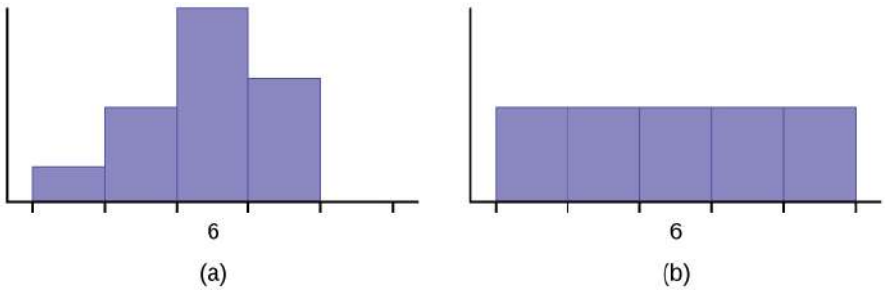


Figure 2.24

Use the following information to answer the next three exercises: We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

Table 2.82

120. What is the *IQR*?
a. 8
b. 11
c. 15
d. 35

121. What is the mode?

a. 19
b. 19.5
c. 14 and 20
d. 22.65

122. Is this a sample or the entire population?

a. sample
b. entire population
c. neither

123. Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

Table 2.83

a. Find the sample mean $\bar{x}$.
b. Find the approximate sample standard deviation, s.

124. Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

X	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

Table 2.84

a. Find the sample mean $\bar{x}$
b. Find the sample standard deviation, s
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. What percent of the students owned at least five pairs?
i. Find the 40^{th} percentile.
j. Find the 90^{th} percentile.
k. Construct a line graph of the data
l. Construct a stemplot of the data

125. Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. The middle 50% of the weights are from _____ to _____.
f. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
g. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
h. Assume the population was the San Francisco 49ers. Find:
    i. the population mean, μ.
    ii. the population standard deviation, σ.
    iii. the weight that is two standard deviations below the mean.
    iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?
i. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

126. One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

3; 8; –1; 2; 0; 5; –3; 1; –1; 6; 5; –2

a. What is the mean change score?
b. What is the standard deviation for this population?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

127. Refer to **Figure 2.25** determine which of the following are true and which are false. Explain your solution to each part in complete sentences.

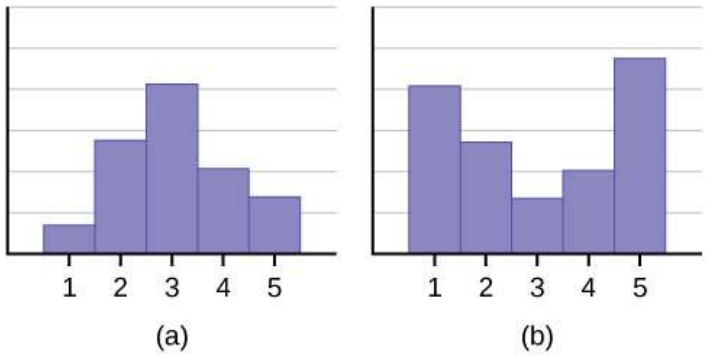


Figure 2.25

a. The medians for both graphs are the same.
b. We cannot determine if any of the means for both graphs is different.
c. The standard deviation for graph b is larger than the standard deviation for graph a.
d. We cannot determine if any of the third quartiles for both graphs is different.

128. In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a chart.
b. Find the median, the first quartile, and the third quartile.
c. Find the 65^{th} percentile.
d. Find the 10^{th} percentile.
e. The middle 50% of the conferences last from _____ days to _____ days.
f. Calculate the sample mean of days of engineering conferences.
g. Calculate the sample standard deviation of days of engineering conferences.
h. Find the mode.
i. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
j. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

129. A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Use the following information to answer the next two exercises. X = the number of days per week that 100 clients use a particular exercise facility.

x	Frequency
0	3
1	12
2	33
3	28
4	11
5	9
6	4

Table 2.85

130. The 80^{th} percentile is _____

a. 5
b. 80
c. 3
d. 4

131. The number that is 1.5 standard deviations BELOW the mean is approximately _____

a. 0.7
b. 4.8
c. –2.8
d. Cannot be determined

132. Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the **Table 2.86**.

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Table 2.86

a. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR* to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts a and c of this problem give the same answer?
e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

REFERENCES

2.1 Display Data

Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics*, 2011. Available online at http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/ (accessed August 21, 2013).

"9th Annual AP Report to the Nation." CollegeBoard, 2013. Available online at http://apreport.collegeboard.org/goals-and-findings/promoting-equity (accessed September 13, 2013).

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker

"Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at http://www.scholastic.com/teachers/article/timeline-guide-us-presidents (accessed April 3, 2013).

"Presidents." Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).

"Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).

"Consumer Price Index." United States Department of Labor: Bureau of Labor Statistics. Available online at http://data.bls.gov/pdq/SurveyOutputServlet (accessed April 3, 2013).

"CO2 emissions (kt)." The World Bank, 2013. Available online at http://databank.worldbank.org/data/home.aspx (accessed

April 3, 2013).

“Births Time Series Data.” General Register Office For Scotland, 2013. Available online at http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html (accessed April 3, 2013).

“Demographics: Children under the age of 5 years underweight.” Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).

Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.

“Overweight and Obesity: Adult Obesity Facts.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

2.2 Measures of the Location of the Data

Cauchon, Dennis, Paul Overberg. “Census data shows minorities now a majority of U.S. births.” USA Today, 2012. Available online at http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1 (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).

“1990 Census.” United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).

Data from *San Jose Mercury News*.

Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

2.3 Measures of the Center of the Data

Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).

“Demographics: Obesity – adult prevalence rate.” Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

2.7 Measures of the Spread of the Data

Data from Microsoft Bookshelf.

King, Bill.“Graphically Speaking.” Institutional Research, Lake Tahoe Community College. Available online at http://www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

SOLUTIONS

1

Figure 2.26

3

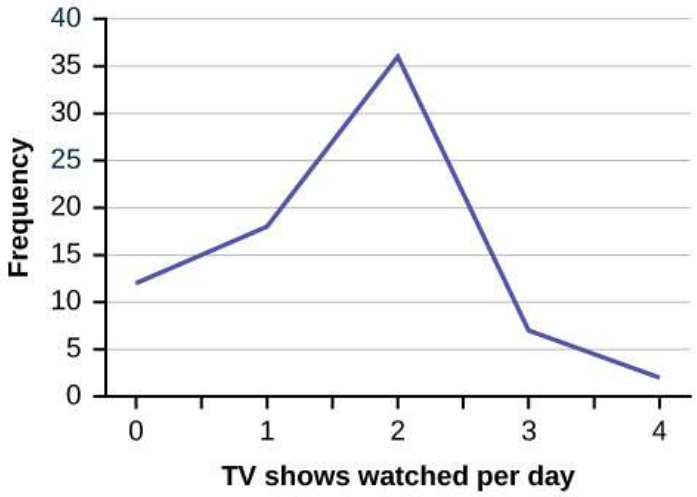


Figure 2.27

5

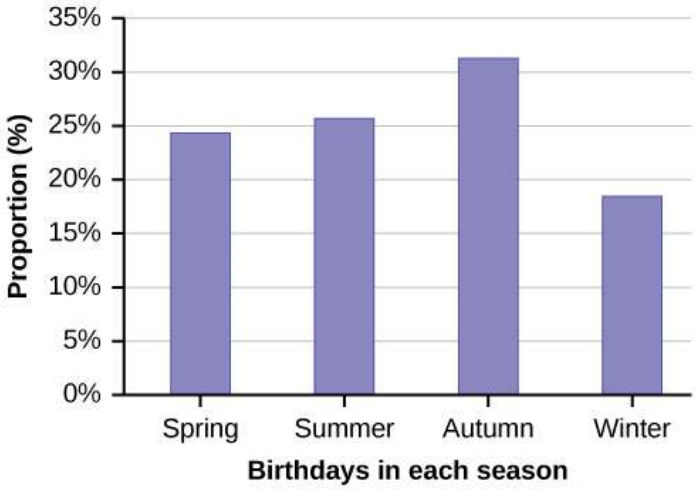


Figure 2.28

7

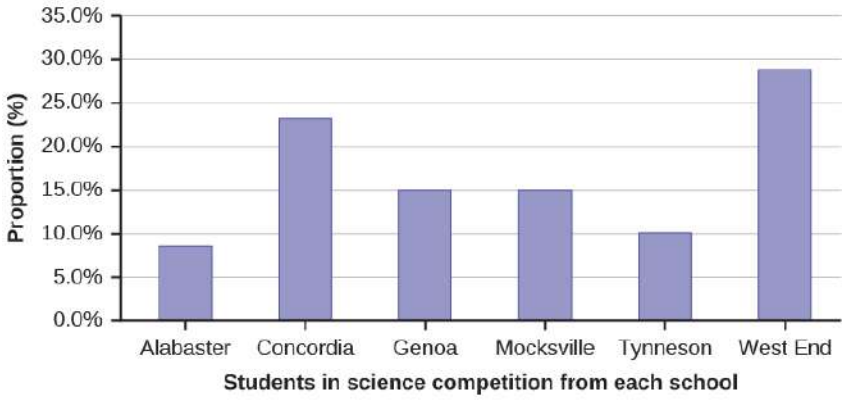


Figure 2.29

9 65

11 The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

13 Answers will vary. One possible histogram is shown:

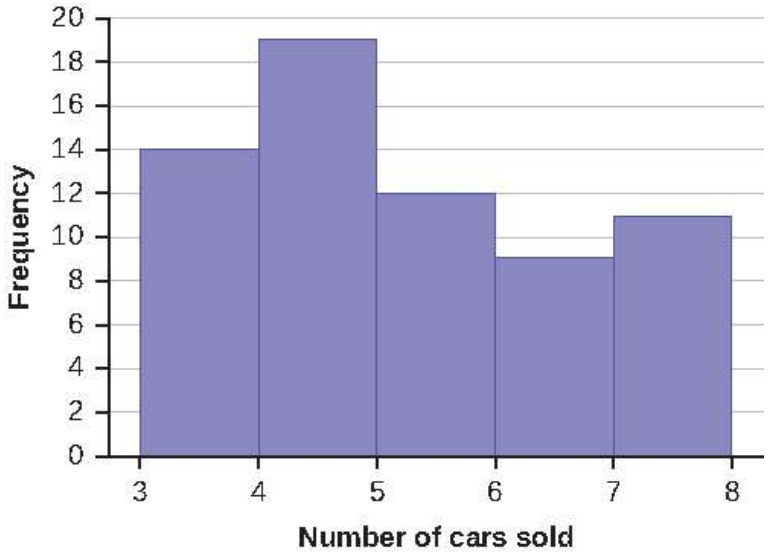


Figure 2.30

15 Find the midpoint for each class. These will be graphed on the *x*-axis. The frequency values will be graphed on the *y*-axis values.

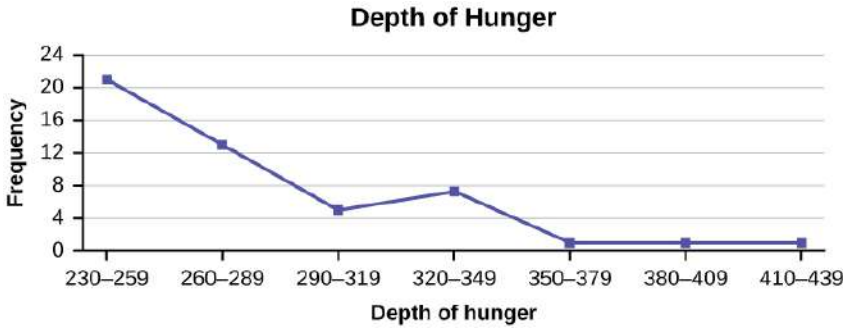


Figure 2.31

17

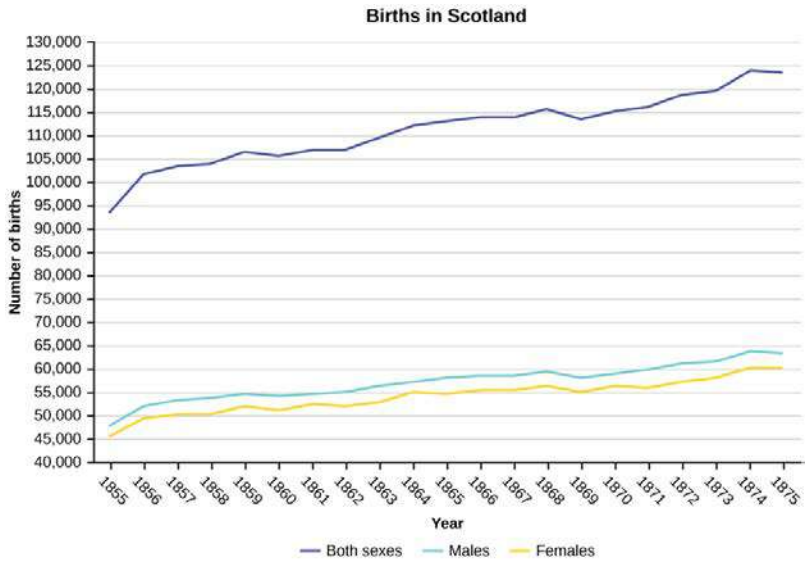


Figure 2.32

19

a. The 40th percentile is 37 years.

b. The 78th percentile is 70 years.

21 Jesse graduated 37th out of a class of 180 students. There are 180 – 37 = 143 students ranked below Jesse. There is one rank of 37. $x = 143$ and $y = 1$. $\frac{x + 0.5y}{n}(100) = \frac{143 + 0.5(1)}{180}(100) = 79.72$. Jesse's rank of 37 puts him at the 80th percentile.

23

a. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed which is faster.

b. 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

25 When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

27 The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of $1700 or less; only 10% had damage repair costs of $1700 or more.

29 You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

31 4

33 6 – 4 = 2

35 6

37 Mean: 16 + 17 + 19 + 20 + 20 + 21 + 23 + 24 + 25 + 25 + 25 + 26 + 26 + 27 + 27 + 27 + 28 + 29 + 30 + 32 + 33 + 33 + 34 + 35 + 37 + 39 + 40 = 738; $\frac{738}{27} = 27.33$

39 The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

41 4

44 39.48 in.

45 $21,574

46 15.98 ounces

47 81.56

48 4 hours

49 2.01 inches

50 18.25

51 10

52 14.15

53 14

54 14.78

55 44%

56 100%

57 6%

58 33%

59 The data are symmetrical. The median is 3 and the mean is 2.85. They are close, and the mode lies close to the middle of the data, so the data are symmetrical.

61 The data are skewed right. The median is 87.5 and the mean is 88.2. Even though they are close, the mode lies to the left of the middle of the data, and there are many more instances of 87 than any other number, so the data are skewed right.

63 When the data are symmetrical, the mean and median are close or the same.

65 The distribution is skewed right because it looks pulled out to the right.

67 The mean is 4.1 and is slightly greater than the median, which is four.

69 The mode and the median are the same. In this case, they are both five.

71 The distribution is skewed left because it looks pulled out to the left.

73 The mean and the median are both six.

75 The mode is 12, the median is 12.5, and the mean is 15.1. The mean is the largest.

77 The mean tends to reflect skewing the most because it is affected the most by outliers.

79 $s = 34.5$

81 For Fredo: $z = \frac{0.158 - 0.166}{0.012} = -0.67$ For Karl: $z = \frac{0.177 - 0.189}{0.015} = -0.8$ Fredo's z-score of –0.67 is higher than Karl's z-score of –0.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

83

a. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$

b. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$

c. $s_x = \sqrt{\frac{\sum fm^2}{n} - \bar{x}^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

84

a. Example solution for using the random number generator for the TI-84+ to generate a simple random sample of 8 states. Instructions are as follows.

Number the entries in the table 1–51 (Includes Washington, DC; Numbered vertically)

Press MATH

Arrow over to PRB

Press 5:randInt(

Enter 51,1,8)

Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}.

Corresponding percents are {30.1, 22.2, 26.5, 27.1, 30.9, 34.0, 26.0, 25.1}.

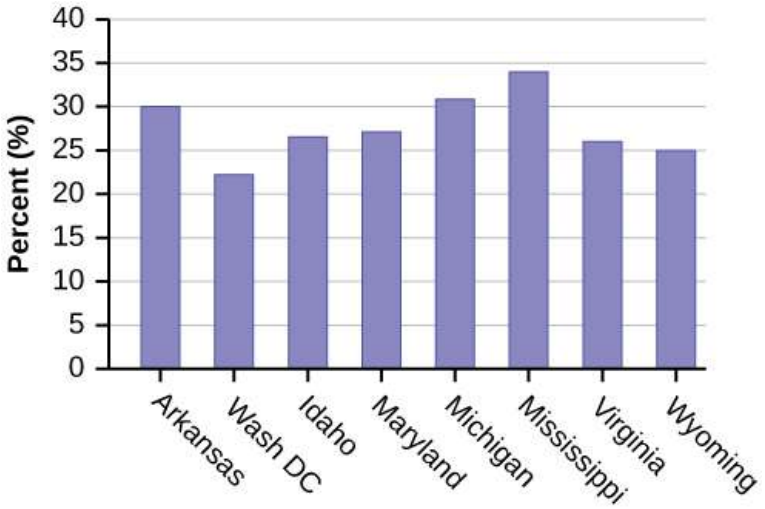


Figure 2.33

b.

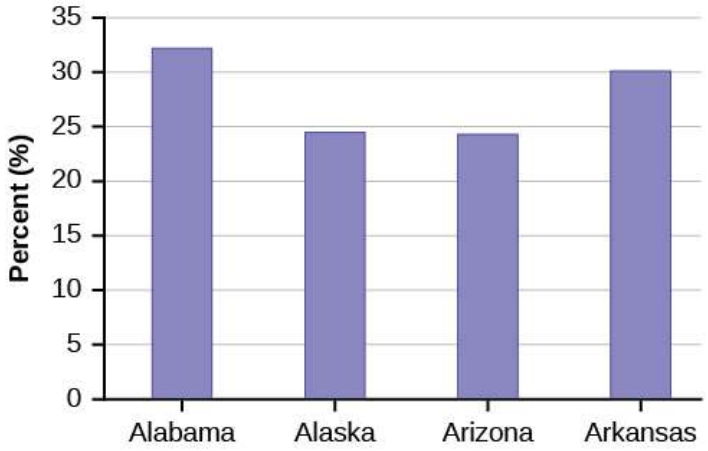
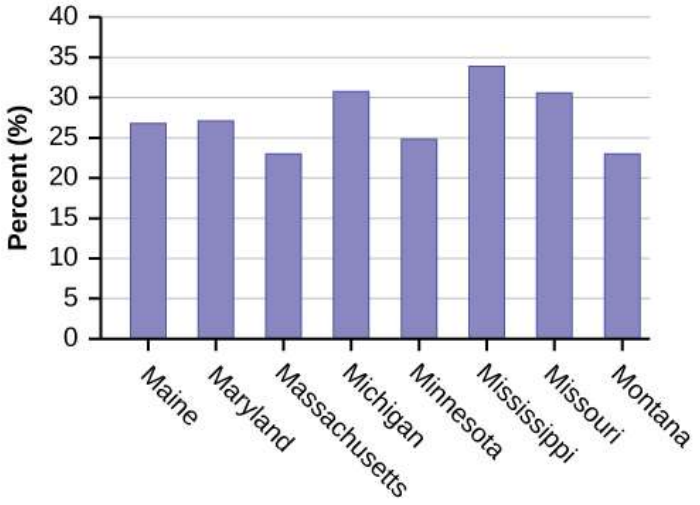


Figure 2.34



c.

Figure 2.35

86

Amount($)	Frequency	Relative Frequency
51–100	5	0.08
101–150	10	0.17
151–200	15	0.25
201–250	15	0.25
251–300	10	0.17
301–350	5	0.08

Table 2.87 Singles

Amount($)	Frequency	Relative Frequency
100–150	5	0.07
201–250	5	0.07
251–300	5	0.07
301–350	5	0.07
351–400	10	0.14
401–450	10	0.14
451–500	10	0.14
501–550	10	0.14
551–600	5	0.07
601–650	5	0.07

Table 2.88 Couples

a. See **Table 2.68** and **Table 2.68**.

b. In the following histogram data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where both boundary values are included).

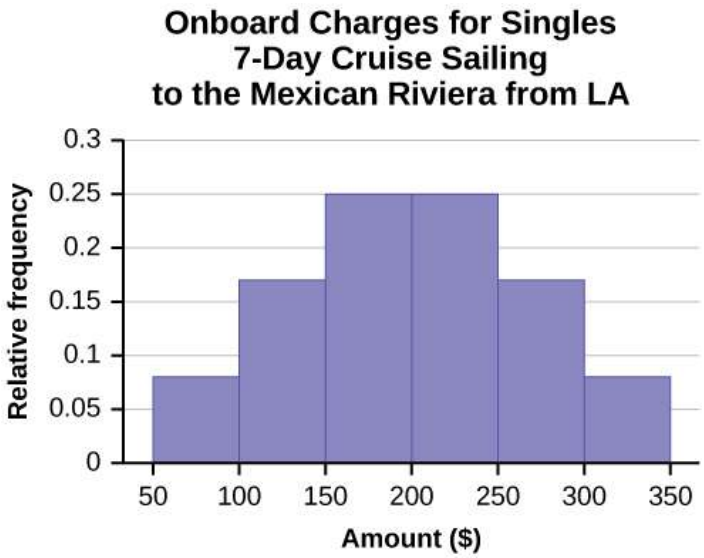


Figure 2.36

c. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where values on both boundaries are included).

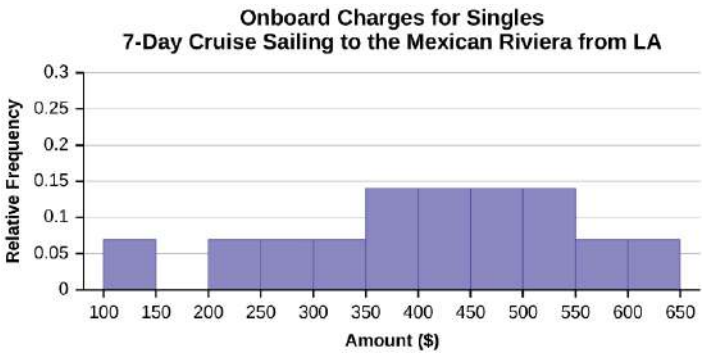


Figure 2.37

d. Compare the two graphs:

i. Answers may vary. Possible answers include:
- Both graphs have a single peak.
- Both graphs use class intervals with width equal to $50.

ii. Answers may vary. Possible answers include:
- The couples graph has a class interval with no values.
- It takes almost twice as many class intervals to display the data for couples.

iii. Answers may vary. Possible answers include: The graphs are more similar than different because the overall

patterns for the graphs are the same.

e. Check student's solution.

f. Compare the graph for the Singles with the new graph for the Couples:

i.
- Both graphs have a single peak.
- Both graphs display 6 class intervals.
- Both graphs show the same general pattern.

ii. Answers may vary. Possible answers include: Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.

g. Answers may vary. Possible answers include: You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the Couples graph. Because a couple represents two individuals, the new scale leads to a more accurate comparison.

h. Answers may vary. Possible answers include: Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

88 c

90 Answers will vary.

92

a. $1 - (0.02+0.09+0.19+0.26+0.18+0.17+0.02+0.01) = 0.06$

b. $0.19+0.26+0.18 = 0.63$

c. Check student's solution.

d. 40^{th} percentile will fall between 30,000 and 40,000

80^{th} percentile will fall between 50,000 and 75,000

e. Check student's solution.

94 The mean percentage, $\bar{x} = \frac{1328.65}{50} = 26.75$

95

a. Yes

b. The sample is 0.5 higher.

96

a. 20

b. No

97 51

98

a. 42

b. 99

99 $10.19

100 17%

101 $30,772.48

102 4.4%

103 7.24%

104 -1.27%

106 The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th

number in order. Six years will have totals at or below the median.

108 474 FTES

110 919

112

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- *IQR* = 245

113 Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

115 For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

117

- $\bar{x} = 23.32$
- Using the TI 83/84, we obtain a standard deviation of: $s_x = 12.95$.
- The obesity rate of the United States is 10.58% higher than the average obesity rate.
- Since the standard deviation is 12.95, we see that 23.32 + 12.95 = 36.27 is the obesity percentage that is one standard deviation from the mean. The United States obesity rate is slightly less than one standard deviation from the mean. Therefore, we can assume that the United States, while 34% obese, does not hav e an unusually high percentage of obese people.

120 a

122 b

123

a. 1.48

b. 1.12

125

a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302

b. 241

c. 205.5

d. 272.5

e. 205.5, 272.5

f. sample

g. population

h.
   i. 236.34
   ii. 37.50
   iii. 161.34
   iv. 0.84 std. dev. below the mean

i. Young

127

a. True

b. True

c. True

d. False

129

a.

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

Table 2.89

b. Check student’s solution.

c. mode

d. 8628.74

e. 6943.88

f. –0.09

131 a