EDITION 5

AP® Edition
Introduction to Statistics and Data Analysis

Roxy Peck
California Polytechnic State University, San Luis Obispo

Chris Olsen
Grinnell College, Grinnell, Iowa

Jay L. Devore
California Polytechnic State University, San Luis Obispo

CENGAGE Learning

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

Brief Contents

Sections and/or chapter numbers shaded in color can be found at www.cengagebrain.com

Contents

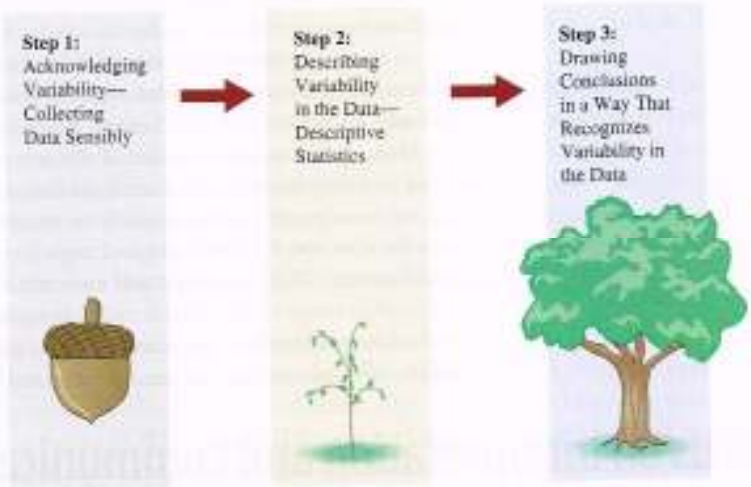
Preface

In a nutshell, statistics is about understanding the role that variability plays in drawing conclusions based on data. *Introduction to Statistics and Data Analysis*, AP® Edition Fifth Edition, develops this crucial understanding of variability through its focus on the data analysis process.

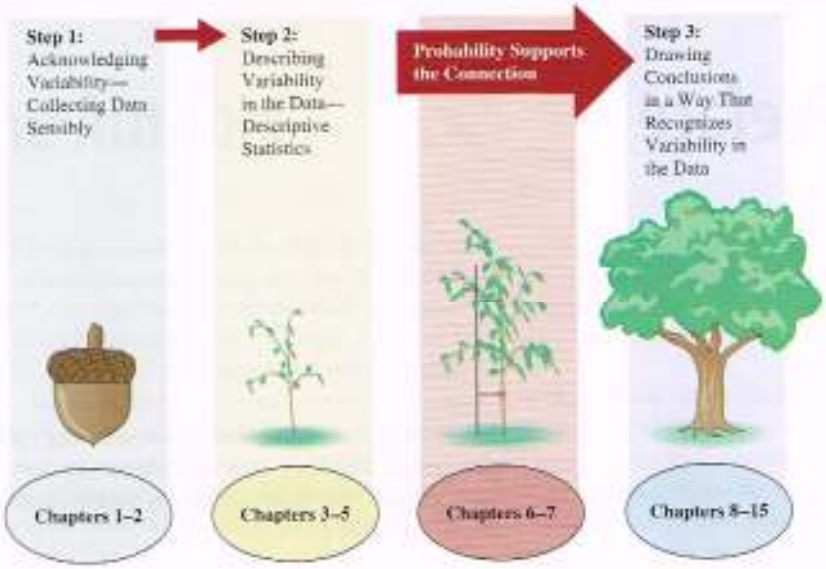
An Organization That Reflects the Data Analysis Process

Students are introduced early to the idea that data analysis is a process that begins with careful planning, followed by data collection, data description using graphical and numerical summaries, data analysis, and finally interpretation of results. This process is described in detail in Chapter 1, and the ordering of topics in the first ten chapters of the book mirrors this process: data collection, then data description, then statistical inference.

The logical order in the data analysis process can be pictured as shown in the following figure.



Unlike many introductory texts, *Introduction to Statistics and Data Analysis*, Fifth Edition, is organized in a manner consistent with the natural order of the data analysis process:



The Importance of Context and Real Data

Statistics is not about numbers; it is about data—numbers in context. It is the context that makes a problem meaningful and something worth considering. For example, exercises that ask students to compute the mean of 10 numbers or to construct a dotplot or boxplot of 20 numbers without context are arithmetic and graphing exercises. They become statistics problems only when a context gives them meaning and allows for interpretation. While this makes for a text that may appear "wordy" when compared to traditional mathematics texts, it is a critical and necessary component of a modern statistics text.

Examples and exercises with overly simple settings do not allow students to practice interpreting results in authentic situations or give students the experience necessary to be able to use statistical methods in real settings. We believe that the exercises and examples are a particular strength of this text, and we invite you to compare the examples and exercises with those in other introductory statistics texts.

Many students are skeptical of the relevance and importance of statistics. Contrived problem situations and artificial data often reinforce this skepticism. A strategy that we have employed successfully to motivate students is to present examples and exercises that involve data extracted from journal articles, newspapers, and other published sources. Most examples and exercises in the book are of this nature; they cover a very wide range of disciplines and subject areas. These include, but are not limited to, health and fitness, consumer research, psychology and aging, environmental research, law and criminal justice, and entertainment.

A Focus on Interpretation and Communication

Most chapters include a section titled "Interpreting and Communicating the Results of Statistical Analyses." These sections include advice on how to best communicate the results of a statistical analysis and also consider how to interpret statistical

CHAPTER 1

The Role of Statistics and the Data Analysis Process

We encounter data and make conclusions based on data every day. **Statistics** is the scientific discipline that provides methods to help us make sense of data. Statistical methods, used intelligently, offer a set of powerful tools for gaining insight into the world around us. The widespread use of statistical analyses in diverse fields such as business, medicine, agriculture, social sciences, natural sciences, and engineering has led to increased recognition that statistical literacy—a familiarity with the goals and methods of statistics—should be a basic component of a well-rounded educational program.

The field of statistics helps us to make intelligent judgments and informed decisions in the presence of uncertainty and variation. In this chapter, we consider the nature and role of variability in statistical settings, introduce some basic terminology, and look at some simple graphical displays for summarizing data.

Chapter 1: Learning Objectives

STUDENTS WILL UNDERSTAND:

- the steps in the data analysis process.

STUDENTS WILL BE ABLE TO:

- distinguish between a population and a sample.
- distinguish between categorical, discrete numerical, and continuous numerical data.
- construct a frequency distribution and a bar chart and describe the distribution of a categorical variable.
- construct a dotplot and describe the distribution of a numerical variable.

1.1 Why Study Statistics?

There is an old saying that "without data, you are just another person with an opinion." While anecdotes and coincidences may make for interesting stories, you wouldn't want to make important decisions on the basis of anecdotes alone. For example, just because a friend of a friend ate 16 apricots and then experienced relief from joint pain doesn't mean that this is all you need to know to help one of your parents choose a treatment for arthritis! Before recommending apricots, you would definitely want to consider relevant data—that is, data that would allow you to investigate the effectiveness of apricots as a treatment for arthritis.

It is difficult to function in today's world without a basic understanding of statistics. For example, here are just a few headlines from articles that draw conclusions based on data that all appeared over two days in *USA Today* (December 19 and 20, 2013):

- The article "American Attitudes Toward Global Warming" summarized data from a nationwide survey of adults. A variety of graphs and charts provide information on opinions regarding the existence of global warming, the impact of global warming, and what actions should be taken in response to global warming.
- "Standardized Testing Fails College Students" is the title of an article describing the use of standardized tests to place college students into appropriate college mathematics courses. The article concludes that many students are not aware of the importance of these exams and do not prepare for them. This results in many students being placed in developmental mathematics courses that slow their progress toward getting their degree.
- The article "Shoppers Say Ho-Hum to Discounts" describes conclusions drawn from a study of how consumers respond to e-mail advertising campaigns that offer discounts. The article concludes that this practice has become so widespread that shoppers largely ignore these e-mails and delete them without even reading them. This is information that retailers should consider when planning future advertising campaigns.
- "Older Americans Could Opt Out of Blood Pressure Meds" is the title of an article describing a study of the effect of high blood pressure on health for those over age 60. Data from this study lead to the conclusion that for older Americans, there is no further benefit to reducing blood pressure below 150/90. This is of interest to doctors because the previous recommendation was that blood pressure should be 140/90 or lower.
- The article "College Coaching Gender Gap Persists" reported data from a study of colleges in six large NCAA sports conferences. The study found that only 39.6% of the coaches of women's sports teams in 2013 were women, which is even lower than in previous years. The article concluded that although Title IX increased opportunities for participation of women in collegiate sports, it has not yet resulted in increased opportunities for women as coaches.

To be an informed consumer of reports such as those described above, you must be able to do the following:

1. Extract information from tables, charts, and graphs.
2. Follow numerical arguments.
3. Understand the basics of how data should be gathered, summarized, and analyzed to draw statistical conclusions.

Your statistics course will help prepare you to perform these tasks.

Studying statistics will also enable you to collect data in a sensible way and then use the data to answer questions of interest. In addition, studying statistics will allow you to

critically evaluate the work of others by providing you with the tools you need to make informed judgments.

Throughout your personal and professional life, you will need to understand and use data to make decisions. To do this, you must be able to

1. Decide whether existing data is adequate or whether additional information is required.
2. If necessary, collect more information in a reasonable and thoughtful way.
3. Summarize the available data in a useful and informative manner.
4. Analyze the available data.
5. Draw conclusions, make decisions, and assess the risk of an incorrect decision.

These are the steps in the data analysis process. These steps will be considered in more detail in Section 1.3.

We hope that this textbook will help you to understand the logic behind statistical reasoning, prepare you to apply statistical methods appropriately, and enable you to recognize when statistical arguments are faulty.

1.2 The Nature and Role of Variability

Statistical methods allow us to collect, describe, analyze, and draw conclusions from data. If we lived in a world where all measurements were identical for every individual, these tasks would be simple. Imagine a population consisting of all students at a particular university. Suppose that *every* student was enrolled in the same number of courses, spent exactly the same amount of money on textbooks this semester, and favored increasing student fees to support expanding library services. For this population, there is *no* variability in number of courses, amount spent on books, or student opinion on the fee increase. A researcher studying students from this population in order to draw conclusions about these three variables would have a particularly easy task. It would not matter how many students the researcher studied or how the students were selected. In fact, the researcher could collect information on number of courses, amount spent on books, and opinion on the fee increase by just stopping the next student who happened to walk by the library. Because there is no variability in the population, this one individual would provide complete and accurate information about the population. The researcher could draw conclusions with no risk of error.

The situation just described is obviously unrealistic. Populations with no variability are rare. In fact, variability is almost universal. We need to understand variability to be able to collect, describe, analyze, and draw conclusions from data in a sensible way.

Examples 1.1 and 1.2 illustrate how describing and understanding variability are important.

EXAMPLE 1.1 If the Shoe Fits

Understand the context

The graphs in Figure 1.1 are examples of a type of graph called a histogram. (The construction and interpretation of such graphs is discussed in Chapter 3.) Figure 1.1(a) shows the distribution of the heights of female basketball players who played at a particular university between 2005 and 2013. The height of each bar in the graph indicates how many players' heights were in the corresponding interval. For example, 40 basketball players had heights between 72 inches and 74 inches, whereas only 2 players had heights between 66 inches and 68 inches. Figure 1.1(b) shows the distribution of heights for members of the women's gymnastics team. Both histograms are based on the heights of 100 women.

Consider the data

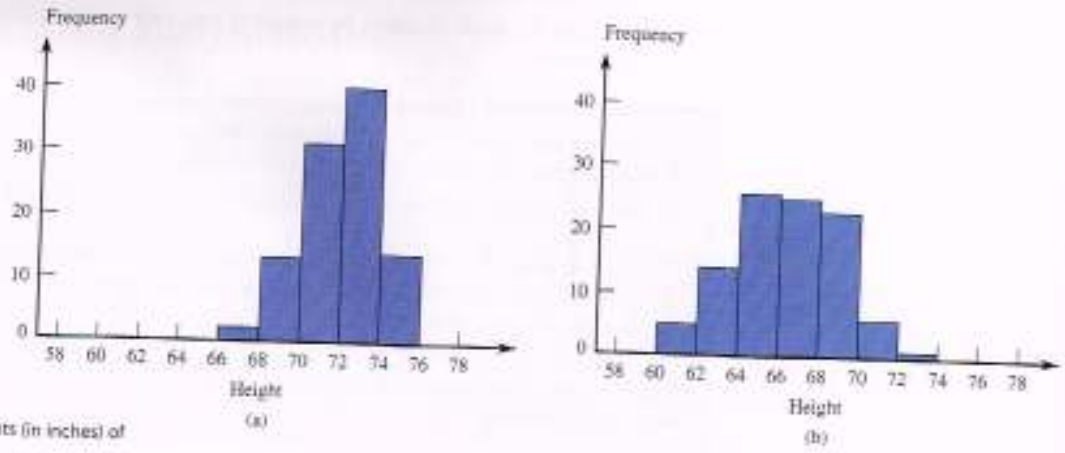


FIGURE 1.1
Histograms of heights (in inches) of female athletes:
(a) 100 basketball players;
(b) 100 gymnasts.

The first histogram shows that the heights of female basketball players varied, with most heights falling between 68 inches and 76 inches. In the second histogram we see that the heights of female gymnasts also varied, with most heights in the range of 60 inches to 72 inches. It is also clear that there is more variation in the heights of the gymnasts than in the heights of the basketball players, because the gymnast histogram spreads out more about its center than does the basketball histogram.

Interpret the results }

Now suppose that a tall woman (5 feet 11 inches) tells you she is looking for her sister who is practicing with her team at the gym. Would you direct her to where the basketball team is practicing or to where the gymnastics team is practicing? What reasoning would you use to decide? If you found a pair of size 6 shoes left in the locker room, would you first try to return them by checking with members of the basketball team or the gymnastics team?

You probably answered that you would send the woman looking for her sister to the basketball practice and that you would try to return the shoes to a gymnastics team member. To reach these conclusions, you informally used statistical reasoning that combined your own knowledge of the relationship between heights of siblings and between shoe size and height with the information about the distributions of heights presented in Figure 1.1. You might have reasoned that heights of siblings tend to be similar and that a height as great as 5 feet 11 inches, although not impossible, would be unusual for a gymnast. On the other hand, a height as tall as 5 feet 11 inches would be a common occurrence for a basketball player.

Similarly, you might have reasoned that tall people tend to have bigger feet and that short people tend to have smaller feet. The shoes found were a small size, so it is more likely that they belong to a gymnast than to a basketball player, because small heights are usual for gymnasts and unusual for basketball players. ■

EXAMPLE 1.2 Monitoring Water Quality

Understand the context }

As part of its regular water quality monitoring efforts, an environmental control board selects five water specimens from a particular well each day. The concentration of contaminants in parts per million (ppm) is measured for each of the five specimens, and then the average of the five measurements is calculated. The histogram in Figure 1.2 summarizes the average contamination values for 200 days.

Now suppose that a chemical spill has occurred at a manufacturing plant 1 mile from the well. It is not known whether a spill of this nature would contaminate groundwater in the area

of the spill and, if so, whether a spill this distance from the well would affect the quality of well water.

Consider the data

One month after the spill, five water specimens are collected from the well, and the average contamination is 15.5 ppm. Considering the variation before the spill, would you interpret this as convincing evidence that the well water was affected by the spill? What if the calculated average was 17.4 ppm? 22.0 ppm? How is your reasoning related to the histogram in Figure 1.2?

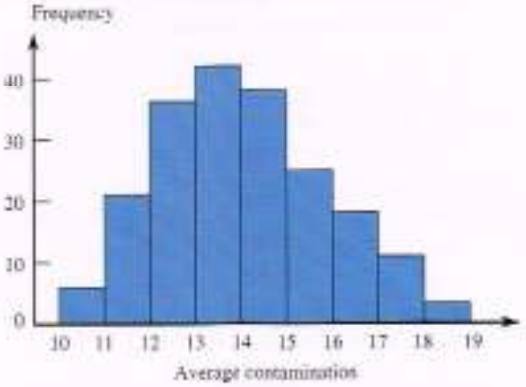


FIGURE 1.2
Average contamination concentration (in parts per million) measured each day for 200 days.

Interpret the results

Before the spill, the average contaminant concentration varied from day to day. An average of 15.5 ppm would not have been an unusual value, so seeing an average of 15.5 ppm after the spill isn't necessarily an indication that contamination has increased. On the other hand, an average as large as 17.4 ppm is less common, and an average as large as 22.0 ppm is not at all typical of the pre-spill values. In this case, we would probably conclude that the well contamination level has increased. ■

In these two examples, reaching a conclusion required an understanding of variability. Understanding variability allows us to distinguish between usual and unusual values. The ability to recognize unusual values in the presence of variability is an important aspect of most statistical procedures. It also enables us to quantify the chance of being incorrect when a conclusion is based on data. These concepts will be developed further in subsequent chapters.

1.3 Statistics and the Data Analysis Process

Statistics involves collecting, summarizing, and analyzing data. All three tasks are critical. Without summarization and analysis, raw data are of little value. Even sophisticated analyses can't produce meaningful information from data that were not collected in a sensible way.

Statistical studies are undertaken to answer questions about our world. Is a new flu vaccine effective in preventing illness? Is the use of bicycle helmets on the rise? Are injuries that result from bicycle accidents less severe for riders who wear helmets than for those who do not? Do engineering students pay more for textbooks than psychology students? Data collection and analysis allow researchers to answer such questions.

The data analysis process can be viewed as a sequence of steps that lead from planning to data collection to making informed conclusions based on the resulting data. The process can be organized into six steps described in the following box.

The Data Analysis Process

1. **Understanding the nature of the problem.** Effective data analysis requires an understanding of the research problem. We must know the goal of the research and what questions we hope to answer. It is important to have a clear direction before gathering data to ensure that we will be able to answer the questions of interest using the data collected.
2. **Deciding what to measure and how to measure it.** The next step in the process is deciding what information is needed to answer the questions of interest. In some cases, the choice is obvious. For example, in a study of the relationship between the weight of a Division I football player and position played, you would need to collect data on player weight and position. In other cases the choice of information is not as straightforward. For example, in a study of the relationship between preferred learning style and intelligence, how would you define learning style and measure it? What measure of intelligence would you use? It is important to carefully define the variables to be studied and to develop appropriate methods for determining their values.
3. **Data collection.** The data collection step is crucial. The researcher must first decide whether an existing data source is adequate or whether new data must be collected. If a decision is made to use existing data, it is important to understand how the data were collected and for what purpose, so that any resulting limitations are also fully understood. If new data are to be collected, a careful plan must be developed, because the type of analysis that is appropriate and the subsequent conclusions that can be drawn depend on how the data are collected.
4. **Data summarization and preliminary analysis.** After the data are collected, the next step is usually a preliminary analysis that includes summarizing the data graphically and numerically. This initial analysis provides insight into important characteristics of the data and can provide guidance in selecting appropriate methods for further analysis.
5. **Formal data analysis.** The data analysis step requires the researcher to select and apply statistical methods. Much of this textbook is devoted to methods that can be used to carry out this step.
6. **Interpretation of results.** Several questions should be addressed in this final step. Some examples are: What can we learn from the data? What conclusions can be drawn from the analysis? How can our results guide future research? The interpretation step often leads to the formulation of new research questions. These new questions lead back to the first step. In this way, good data analysis is often an iterative process.

To illustrate these steps, consider the following example. The admissions director at a large university might be interested in learning why some applicants who were accepted for the fall 2014 term failed to enroll at the university. The **population** of interest to the director consists of all accepted applicants who did not enroll in the fall 2014 term. Because this population is large and it may be difficult to contact all the individuals, the director might decide to collect data from only 300 selected students. These 300 students constitute a **sample**.

DEFINITION

Population: The entire collection of individuals or objects about which information is desired is called the **population** of interest.

Sample: A **sample** is a subset of the population, selected for study.

Deciding how to select the 300 students and what data should be collected from each student are steps 2 and 3 in the data analysis process. Step 4 in the process involves organizing and summarizing data. Methods for organizing and summarizing data, such as the use of tables, graphs, or numerical summaries, make up the branch of statistics called **descriptive statistics**. The second major branch of statistics, **inferential statistics**, involves generalizing from a sample to the population from which it was selected. When we generalize in this way, we run the risk of an incorrect conclusion, because a conclusion about the population is based on incomplete information. An important aspect in the development of inferential techniques involves quantifying the chance of an incorrect conclusion.

DEFINITION

Descriptive statistics: The branch of statistics that includes methods for organizing and summarizing data.

Inferential statistics: The branch of statistics that involves generalizing from a sample to the population from which the sample was selected and assessing the reliability of such generalizations.

Example 1.3 illustrates the steps in the data analysis process.

EXAMPLE 1.3 The Benefits of Acting Out

Understand the context)

A number of studies have reached the conclusion that stimulating mental activities can lead to improved memory and psychological wellness in older adults. The article **"A Short-Term Intervention to Enhance Cognitive and Affective Functioning in Older Adults"** (*Journal of Aging and Health* [2004]: 562–585) describes a study to investigate whether training in acting has similar benefits. Acting requires a person to consider the goals of the characters in the story, to remember lines of dialogue, to move on stage as scripted, and to do all of this at the same time. The researchers conducting the study wanted to see if participation in this type of complex multitasking would lead to an improvement in the ability to function independently.

Participants in the study were assigned to one of three groups. One group took part in an acting class for 4 weeks. One group spent a similar amount of time in a class on visual arts. The third group was a comparison group (called the "no-treatment group") that did not take either class. A total of 124 adults age 60 to 86 participated in the study.

Interpret the results)

At the beginning of the 4-week study period and again at the end of the 4-week study period, each participant took several tests designed to measure problem-solving ability, memory span, self-esteem, and psychological well-being. After analyzing the data from this study, the researchers concluded that those in the acting group showed greater gains than both the visual arts group and the no-treatment group in both problem solving and psychological well-being.

Several new areas of research were suggested in the discussion that followed the analysis. The researchers wondered whether the effect of studying writing or music would be similar to what was observed for acting and described plans to investigate this further. They also noted that the participants in this study were generally well educated and recommended study of a more diverse group before generalizing conclusions about the benefits of studying acting to the larger population of all older adults.

This study illustrates the nature of the data analysis process. A clearly defined research question and an appropriate choice of how to measure the variables of interest (the tests used to measure problem solving, memory span, self-esteem, and psychological well-being) preceded the data collection. Assuming that a reasonable method was used to collect the data (we will see how this can be evaluated in Chapter 2) and that appropriate methods of analysis were employed, the investigators reached the conclusion that the study of acting showed promise. However, they recognized the limitations of the study, which in turn led to plans for further research. As is often the case, the data analysis cycle led to new research questions, and the process began again. ■

EXERCISES 1.1 - 1.11

1.1 Give a brief definition of the terms *descriptive statistics* and *inferential statistics*.

1.2 Give a brief definition of the terms *population* and *sample*.

1.3 The following conclusion from a study appeared in the article **"Smartphone Nation" (*AARP Bulletin*, September 2009)**: "If you love your smart phone, you are not alone. Half of all boomers sleep with their cell phone within arm's length. Two of three people age 50 to 64 use a cell phone to take photos, according to a 2010 Pew Research Center report." Are the given proportions (half and two of three) population values, or were they calculated from a sample?

1.4 Based on a study of 2121 children between the ages of 1 and 4, researchers at the Medical College of Wisconsin concluded that there was an association between iron deficiency and the length of time that a child is bottle-fed **(*Milwaukee Journal Sentinel*, November 26, 2005)**. Describe the sample and the population of interest for this study.

1.5 The student senate at a university with 15,000 students is interested in the proportion of students who favor a change in the grading system to allow for plus and minus grades (e.g., B+, B, B−, rather than just B). Two hundred students are interviewed to determine their attitude toward this proposed change.

a. What is the population of interest?

b. What group of students constitutes the sample in this problem?

1.6 The increasing popularity of online shopping has many consumers using Internet access at work to browse and shop online. In fact, the Monday after Thanksgiving has been nicknamed "Cyber Monday" because of the large increase in online purchases that occurs on that day. Data from a large-scale survey by a market research firm **(*Detroit Free Press*, November 26, 2005)** was used to compute estimates of the percent of men and women who shop online while at work. The resulting estimates probably won't make most employers happy—42% of the men and 32% of the women in the sample were shopping online at work!

Are the estimates given computed using data from a sample or for the entire population?

1.7 The supervisors of a rural county are interested in the proportion of property owners who support the construction of a sewer system. Because it is too costly to contact all 7000 property owners, a survey of 500 owners is undertaken. Describe the population and sample for this problem.

1.8 A consumer group conducts crash tests of new model cars. To determine the severity of damage to 2014 Toyota Camrys resulting from a 10-mph crash into a concrete wall, the research group tests six cars of this type and assesses the amount of damage. Describe the population and sample for this problem.

1.9 A building contractor has a chance to buy an odd lot of 5000 used bricks at an auction. She is interested in determining the proportion of bricks in the lot that are cracked and therefore unusable for her current project, but she does not have enough time to inspect all 5000 bricks. Instead, she checks 100 bricks to determine whether each is cracked. Describe the population and sample for this problem.

1.10 The article **"Brain Shunt Tested to Treat Alzheimer's" (*San Francisco Chronicle*, October 23, 2002)** summarizes the findings of a study that appeared in the journal *Neurology*. Doctors at Stanford Medical Center were interested in determining whether a new surgical approach to treating Alzheimer's disease results in improved memory functioning. The surgical procedure involves implanting a thin tube, called a shunt, which is designed to drain toxins from the fluid-filled space that cushions the brain. Eleven patients had shunts implanted and were followed for a year, receiving quarterly tests of memory function. Another sample of Alzheimer's patients was used as a comparison group. Those in the comparison group received the standard care for Alzheimer's disease. After analyzing the data from this study, the investigators concluded that the "results suggested the treated patients essentially held their own in the cognitive tests while the patients in the control group steadily declined. However, the study was too small to produce conclusive statistical evidence."

a. What were the researchers trying to learn? What questions motivated their research?

b. Do you think that the study was conducted in a reasonable way? What additional information would you want in order to evaluate this study? (Hint: See Example 1.3.)

1.11 In a study of whether taking a garlic supplement reduces the risk of getting a cold, participants were assigned to either a garlic supplement group or to a group that did not take a garlic supplement

("Garlic for the Common Cold," Cochrane Database of Systematic Reviews, 2009). Based on the study, it was concluded that the proportion of people taking a garlic supplement who get a cold is lower than the proportion of those not taking a garlic supplement who get a cold.

a. What were the researchers trying to learn? What questions motivated their research?

b. Do you think that the study was conducted in a reasonable way? What additional information would you want in order to evaluate this study?

Bold exercises answered in back • Data set available online ▼ Video Solution available

1.4 Types of Data and Some Simple Graphical Displays

Every discipline has its own particular way of using common words, and statistics is no exception. You will recognize some of the terminology from previous math and science courses, but much of the language of statistics will be new to you. In this section, you will learn some of the terminology used to describe data.

Types of Data

The individuals or objects in any particular population typically possess many characteristics that might be studied. Consider a group of students currently enrolled in a statistics course at a particular college. One characteristic of the students in the population is the brand of calculator owned (Casio, Hewlett-Packard, Sharp, Texas Instruments, and so on). Another characteristic is the number of textbooks purchased that semester, and yet another is the distance from the college to each student's permanent residence. A **variable** is any characteristic whose value may change from one individual or object to another. For example, *calculator brand* is a variable, and so are *number of textbooks purchased* and *distance to the college*. **Data** result from making observations either on a single variable or simultaneously on two or more variables.

DEFINITION

Variable: A characteristic whose value may change from one observation to another.

Data: A collection of observations on one or more variables.

A **univariate data set** consists of observations on a single variable made on individuals in a sample or population. There are two types of univariate data sets: **categorical** and **numerical**. In the previous example, *calculator brand* is a categorical variable, because each student's response to the query, "What brand of calculator do you own?" is a category. The collection of responses from all these students forms a categorical data set. The other two variables, *number of textbooks purchased* and *distance to the college*, are both numerical in nature. Determining the value of such a numerical variable (by counting or measuring) for each student results in a numerical data set.

DEFINITION

Univariate data set: A data set consisting of observations on a single characteristic is a **univariate data set**.

Categorical data set: A univariate data set is **categorical** (or **qualitative**) if the individual observations are categorical responses.

Numerical data set: A univariate data set is **numerical** (or **quantitative**) if each observation is a number.

EXAMPLE 1.4 College Choice Do-Over?

Understand the context

The Higher Education Research Institute at UCLA surveys over 20,000 college seniors each year. One question on the 2008 survey asked seniors the following question: If you could make your college choice over, would you still choose to enroll at your current college? Possible responses were definitely yes (DY), probably yes (PY), probably no (PN), and definitely no (DN). Responses for 20 students were:

DY PN DN DY PY PY PN PY PY DY
DY PY DY DY PY PY DY DY PN DY

Consider the context

(These data are just a small subset of the data from the survey. For a description of the full data set, see Exercise 1.18). Because the response to the question about college choice is categorical, this is a univariate categorical data set. ■

In Example 1.4, the data set consisted of observations on a single variable (college choice response), so this is a univariate data set. In some studies, attention focuses simultaneously on two different characteristics. For example, both height (in inches) and weight (in pounds) might be recorded for each individual in a group. The resulting data set consists of pairs of numbers, such as (68, 146). This is called a **bivariate data set. Multivariate data** result from obtaining a category or value for each of two or more attributes (so bivariate data are a special case of multivariate data). For example, multivariate data would result from determining height, weight, pulse rate, and systolic blood pressure for each individual in a group. Example 1.5 illustrates a bivariate data set.

EXAMPLE 1.5 How Safe Are College Campuses?

Understand the context

● Consider the accompanying data on violent crime on college campuses in Florida during 2012. http://www.fbi.gov/

Consider the data

University/College	Student Enrollment	Number of Violent Crimes Reported in 2012
Edison State College	17,107	4
Florida A&M University	13,204	14
Florida Atlantic University	25,246	4
Florida Gulf Coast University	12,851	3
Florida International University	44,616	9
Florida State University	41,057	31
New College of Florida	845	1
Pensacola State College	11,531	3
Santa Fe College	15,493	1
Tallahassee Community College	15,090	2
University of Central Florida	58,465	26
University of Florida	49,589	18
University of North Florida	16,198	2
University of South Florida	4,310	2
University of West Florida	11,982	2

Here two variables—*student enrollment* and *number of violent crimes reported*—were recorded for each of the 15 schools. Because this data set consists of values of two variables for each school, it is a bivariate data set. Each of the two variables considered here is numerical (rather than categorical). ■

● Data set available online

Two Types of Numerical Data

There are two different types of numerical data: **discrete** and **continuous**. Consider a number line (Figure 1.3) for locating values of the numerical variable being studied. Each possible number (2, 3.125, 8.12976, etc.) corresponds to exactly one point on the number line.

Suppose that the variable of interest is the number of courses in which a student is enrolled. If no student is enrolled in more than eight courses, the possible values are 1, 2, 3, 4, 5, 6, 7, and 8. These values are identified in Figure 1.4(a) by the dots at the points marked 1, 2, 3, 4, 5, 6, 7, and 8. These possible values are isolated from one another on the number line. We can place an interval around any possible value that is small enough that no other possible value is included in the interval. On the other hand, the line segment in Figure 1.4(b) identifies a plausible set of possible values for the time (in seconds) it takes for the first kernel in a bag of microwave popcorn to pop. Here the possible values make up an entire interval on the number line, and no possible value is isolated from other possible values.

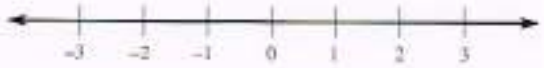


FIGURE 1.3
A number line.

FIGURE 1.4
Possible values of a variable: (a) number of courses; (b) popping time.

DEFINITION

Discrete numerical variable: A numerical variable results in **discrete** data if the possible values of the variable correspond to isolated points on the number line.

Continuous numerical variable: A numerical variable results in **continuous** data if the set of possible values forms an entire interval on the number line.

Discrete data usually arise when observations are determined by counting (for example, the number of roommates a student has or the number of petals on a flower).

EXAMPLE 1.6 Do U Txt?

● The number of text messages sent on a particular day is recorded for each of 12 students. The resulting data set is

23 0 14 13 15 0 60 82 0 40 41 22

Possible values for the variable *number of text messages sent* are 0, 1, 2, 3, ... These are isolated points on the number line, so this data set consists of discrete numerical data.

Suppose that instead of the number of text messages sent, the *time spent texting* had been recorded. Even though time spent may have been reported rounded to the nearest minute, the actual time spent could have been 6 minutes, 6.2 minutes, 6.28 minutes, or any other value in an entire interval. So, recording values of *time spent texting* would result in continuous data. ■

In general, data are continuous when observations involve making measurements, as opposed to counting. In practice, measuring instruments do not have infinite accuracy, so possible measured values, strictly speaking, do not form a continuum on the number line. However, any number in the continuum *could* be a value of the variable. The distinction between discrete and continuous data will be important in our discussion of probability models in Chapter 6.

● Data set available online

Frequency Distributions and Bar Charts for Categorical Data

An appropriate graphical or tabular display of data can be an effective way to summarize and communicate information. When the data set is categorical, a common way to present the data is in the form of a table, called a **frequency distribution**.

DEFINITION

Frequency distribution for categorical data: A table that displays the possible categories along with the associated frequencies and/or relative frequencies.

Frequency: The **frequency** for a particular category is the number of times the category appears in the data set.

Relative frequency: The **relative frequency** for a particular category is calculated as

$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations in the data set}}$$

The relative frequency for a particular category is the proporton of the observations that belong to that category.

Relative frequency distribution: A frequency distribution that includes relative frequencies.

EXAMPLE 1.7 Motorcycle Helmets—Can You See Those Ears?

Understand the context) The U.S. Department of Transportation establishes standards for motorcycle helmets. To ensure a certain degree of safety, helmets should reach the bottom of the motorcyclist's ears. The report "Motorcycle Helmet Use in 2005—Overall Results" (National Highway Traffic Safety Administration, August 2005) summarized data collected in June of 2005 by observing 1700 motorcyclists nationwide at selected roadway locations. Each time a motorcyclist passed by, the observer noted whether the rider was wearing no helmet, a noncompliant helmet, or a compliant helmet. Using the coding

NH = noncompliant helmet
CH = compliant helmet
N = no helmet

Consider the data) a few of the observations were

CH N CH NH N CH CH CH N N

There were also 1690 additional observations, which we didn't reproduce here. In total, there were 731 riders who wore no helmet, 153 who wore a noncompliant helmet, and 816 who wore a compliant helmet.

The corresponding frequency distribution is given in Table 1.1.

Do the work)

TABLE 1.1 Frequency Distribution for Helmet Use

Helmet Use Category	Frequency	Relative Frequency
No helmet	731	0.430 ← 731/1700
Noncompliant helmet	153	0.090 ← 153/1700
Compliant helmet	816	0.480
	1700 ← Total number of observations	1.000 ← Should total 1, but in some cases may be slightly off due to rounding

Interpret the results } From the frequency distribution, we can see that a large percentage of riders (43%) were not wearing a helmet, but most of those who wore a helmet were wearing one that met the Department of Transportation safety standard. ■

A frequency distribution displays a data set in a table. It is also common to display categorical data graphically. A bar chart is one of the most widely used types of graphical displays for categorical data.

Bar Charts

A **bar chart** is a graph of a frequency distribution of categorical data. Each category in the frequency distribution is represented by a bar or rectangle, and the picture is constructed in such a way that the *area* of each bar is proportional to the corresponding frequency or relative frequency.

Bar Charts

When to Use Categorical data.

How to Construct

1. Draw a horizontal axis, and write the category names or labels below the line at regularly spaced intervals.
2. Draw a vertical axis, and label the scale using either frequency or relative frequency.
3. Place a rectangular bar above each category label. The height is determined by the category's frequency or relative frequency, and all bars should have the same width. With the same width, both the height and the area of the bar are proportional to frequency and relative frequency.

What to Look For

- Frequently and infrequently occurring categories.

EXAMPLE 1.8 Revisiting Motorcycle Helmets

Understand the context } Example 1.7 used data on helmet use from a sample of 1700 motorcyclists to construct a frequency distribution (Table 1.1). Figure 1.5 shows the bar chart corresponding to this frequency distribution.

Step-by-step technology instructions available online

Consider the data }

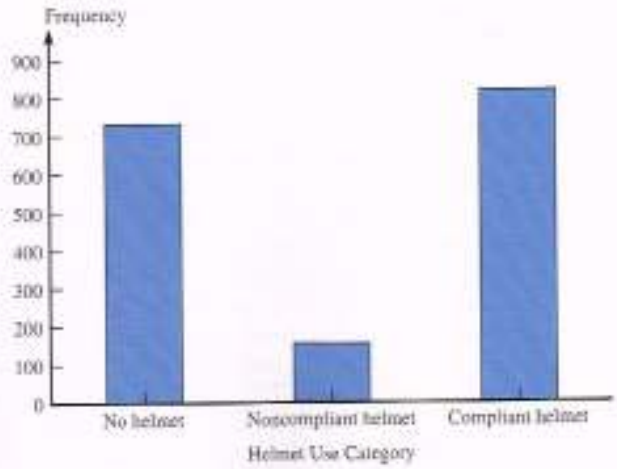


FIGURE 1.5
Bar chart of helmet use.

Interpret the results

The bar chart provides a visual representation of the information in the frequency distribution. From the bar chart, it is easy to see that the compliant helmet use category occurred most often in the data set. The bar for compliant helmets is about five times as tall (and therefore has five times the area) as the bar for noncompliant helmets because approximately five times as many motorcyclists wore compliant helmets than wore noncompliant helmets. ■

Dotplots for Numerical Data

A dotplot is a simple way to display numerical data when the data set is reasonably small. Each observation is represented by a dot above the location corresponding to its value on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence and these dots are stacked vertically.

Dotplots

When to Use Small numerical data sets.

How to Construct

1. Draw a horizontal line and mark it with an appropriate measurement scale.
2. Locate each value in the data set along the measurement scale, and represent it by a dot. If there are two or more observations with the same value, stack the dots vertically.

What to Look for

Dotplots convey information about:

- A representative or typical value in the data set.
- The extent to which the data values spread out.
- The nature of the distribution of values along the number line.
- The presence of unusual values in the data set.

EXAMPLE 1.9 Making It to Graduation . . .

Understand the context

● The article **"Keeping Score When It Counts: Graduation Success and Academic Progress Rates for the 2013 NCAA Men's Division I Basketball Tournament Teams" (The Institute for Diversity and Ethics in Sport, University of Central Florida, March 2013)** compared graduation rates of basketball players to those of all student athletes for the universities and colleges that sent teams to the 2013 Division I playoffs. The graduation rates in the accompanying table represent the percentage of athletes who started college in 2006 who had graduated by the end of 2012. Also shown are the differences between the graduation rate for all student athletes and the graduation rate for basketball student athletes.

Minitab, a computer software package for statistical analysis, was used to construct a dotplot of the graduation rates for basketball players (see Figure 1.6). From this dotplot, we see that basketball graduation rates varied a great deal from school to school, ranging from a low of 17% to a high of 100%.

Consider the data

We can also see that the graduation rates seem to cluster in several groups, denoted by the colored ovals that have been added to the dotplot. There are quite a few schools with graduation rates of 100% (excellent!) and another group of 17 schools with graduation rates that are higher than most. The majority of schools are in the large cluster with graduation rates from about 40% to about 76%. And then there is that bottom group of four schools with embarrassingly low graduation rates for basketball players: University of Florida (17%), North Carolina A&T (25%), Southern University (27%), and New Mexico State (29%).

● Data set available online

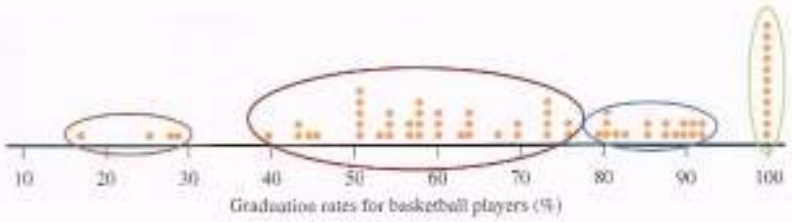


FIGURE 1.6
Minitab dotplot of graduation rates for basketball players.

GRADUATION RATES (%)

Basketball	All Athletes	Difference (All − BB)	Basketball	All Athletes	Difference (All − BB)	Basketball	All Athletes	Difference (All − BB)
100	92	−8	29	70	41	75	80	5
79	76	−3	25	55	30	50	81	31
100	99	−1	73	77	4	87	93	6
80	83	3	75	68	−7	64	84	20
53	82	29	50	77	27	54	83	29
91	94	3	64	87	23	56	76	20
100	97	−3	92	92	0	67	84	17
100	98	−2	62	73	11	73	80	7
73	73	0	44	83	39	92	76	−16
80	94	14	27	51	24	50	75	25
90	96	6	58	87	29	91	88	−3
100	98	−2	43	78	35	100	99	−1
43	80	37	45	85	40	70	72	2
60	81	21	60	75	15	85	80	−5
50	80	30	57	73	16	54	78	24
60	83	23	82	82	0	100	84	−16
58	77	19	54	68	14	40	83	43
64	91	27	70	84	14	80	94	14
58	74	16	50	80	30	100	94	−6
85	83	−2	56	78	22	73	80	7
87	91	4	17	82	65	100	79	−21
89	85	−4	100	89	−11	90	83	−7
83	78	−5	100	85	−15			

Figure 1.7 shows two dotplots of graduation rates—one for basketball players and one for all student athletes. There are some striking differences that are easy to see when the data is displayed in this way. The graduation rates for all student athletes tend to be higher and to vary less from school to school than the graduation rates for basketball players.

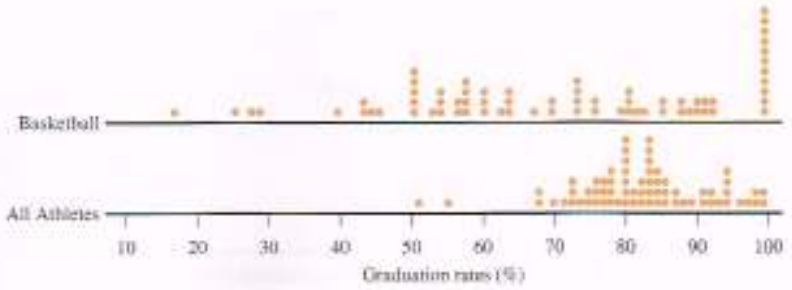


FIGURE 1.7
Minitab dotplot of graduation rates for basketball players and for all athletes.

The dotplots in Figure 1.7 are informative, but we can do even better. The data given here are an example of *paired data*. Each basketball graduation rate is paired with a graduation rate for all student athletes from the same school. When data are paired in this way, it is usually more informative to look at differences—in this case, the difference between the graduation rate for all student athletes and for basketball players for each school. These differences (all − basketball) are also shown in the data table.

Interpret the results

Figure 1.8 gives a dotplot of the differences. Notice that three differences are equal to 0. This corresponds to schools for which the basketball graduation rate is equal to the graduation rate of all student athletes. There are 20 schools for which the difference is negative. Negative differences correspond to schools that have a graduation rate for basketball players that is higher than the graduation rate for all student athletes.

The most interesting features of the difference dotplot are the very large number of positive differences and the wide spread. Positive differences correspond to schools that have a lower graduation rate for basketball players. There is a lot of variability in the graduation rate difference from school to school, and one school has a difference that is noticeably higher than the rest. (In case you were wondering, this school is the University of Florida with a difference of 65%.)

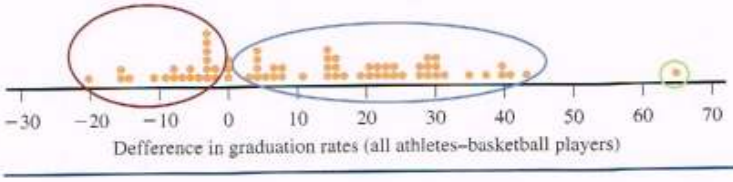


FIGURE 1.8
Dotplot of graduation rate differences (all athletes − basketball players).

EXERCISES 1.12 - 1.31

1.12 Classify each of the following variables as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.

a. Number of students in a class of 35 who turn in a term paper before the due date

b. Gender of the next baby born at a particular hospital

c. Amount of fluid (in ounces) dispensed by a machine used to fill bottles with soda pop

d. Thickness of the gelatin coating of a vitamin E capsule

e. Birth order classification (only child, firstborn, middle child, lastborn) of a math major

1.13 Classify each of the following variables as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.

a. Brand of computer purchased by a customer

b. State of birth for someone born in the United States

c. Price of a textbook

d. Concentration of a contaminant (micrograms per cubic centimeter) in a water sample

e. Zip code (Think carefully about this one.)

f. Actual weight of coffee in a 1-pound can

1.14 For the following numerical variables, state whether each is discrete or continuous.

a. The number of insufficient-funds checks received by a grocery store during a given month

b. The amount by which a 1-pound package of ground beef decreases in weight (because of moisture loss) before purchase

c. The number of New York Yankees during a given year who will not play for the Yankees the next year

d. The number of students in a class of 35 who have purchased a used copy of the textbook

1.15 For the following numerical variables, state whether each is discrete or continuous.

a. The length of a 1-year-old rattlesnake

b. The altitude of a location in California selected randomly by throwing a dart at a map of the state

c. The distance from the left edge at which a 12-inch plastic ruler snaps when bent sufficiently to break

d. The price per gallon paid by the next customer to buy gas at a particular station

1.16 For each of the following situations, give a set of possible data values that might arise from making the observations described.

a. The manufacturer for each of the next 10 automobiles to pass through a given intersection is noted.

b. The grade point average for each of the 15 seniors in a statistics class is determined.

c. The number of gas pumps in use at each of 20 gas stations at a particular time is determined.

d. The actual net weight of each of 12 bags of fertilizer having a labeled weight of 50 pounds is determined.

e. Fifteen different radio stations are monitored during a 1-hour period, and the amount of time devoted to commercials is determined for each.

1.17 In a survey of 100 people who had recently purchased motorcycles, data on the following variables were recorded:

Gender of purchaser
Brand of motorcycle purchased
Number of previous motorcycles owned by purchaser
Telephone area code of purchaser
Weight of motorcycle as equipped at purchase

a. Which of these variables are categorical?

b. Which of these variables are discrete numerical?

c. Which type of graphical display would be an appropriate choice for summarizing the gender data, a bar chart or a dotplot?

d. Which type of graphical display would be an appropriate choice for summarizing the weight data, a bar chart or a dotplot?

1.18 The report **"Findings from the 2008 Administration of the College Senior Survey" (Higher Education Research Institute, UCLA, June 2009)** gave the following relative frequency distribution summarizing student responses to the question "If you could make your college choice over, would you still choose to enroll at your current college?"

Response	Relative Frequency
Definitely yes	0.447
Probably yes	0.373
Probably no	0.134
Definitely no	0.046

a. Use this information to construct a bar chart for the response data.

b. If you were going to use the response data and the bar chart from Part (a) as the basis for an article for your student paper, what would be a good headline for your article?

1.19 ● The article **"Feasting on Protein" (*AARP Bulletin*, September 2009)** gave the cost (in cents per gram) of protein for 19 common food sources of protein.

Food	Cost	Food	Cost
Chicken	1.8	Yogurt	5.0
Salmon	5.8	Milk	2.5
Turkey	1.5	Peas	5.2
Soybeans	3.1	Tofu	6.9
Roast beef	2.7	Cheddar cheese	3.6
Cottage cheese	3.1	Nuts	5.2
Ground beef	2.3	Eggs	5.7
Ham	2.1	Peanut butter	1.8
Lentils	3.3	Ice cream	5.3
Beans	2.9		

a. Construct a dotplot of the cost data. (Hint: See Example 1.9.)

b. Locate the cost for meat and poultry items in your dotplot and highlight them in a different color. Based on the dotplot, do meat and poultry items appear to be a good value? That is, do they appear to be relatively low cost compared to other sources of protein?

1.20 ● Box Office Mojo (www.boxofficemojo.com) tracks movie ticket sales. Ticket sales (in millions of dollars) for each of the top 20 movies in 2007 and 2008 are shown in the accompanying table.

Movie (2007)	2007 Sales (millions of dollars)
Spider-Man 3	336.5
Shrek the Third	322.7
Transformers	319.2
Pirates of the Caribbean: At World's End	309.4
Harry Potter and the Order of the Phoenix	292.0
I Am Legend	256.4
The Bourne Ultimatum	227.5
National Treasure: Book of Secrets	220.0
Alvin and the Chipmunks	217.3
300	210.6
Ratatouille	206.4
The Simpsons Movie	183.1
Wild Hogs	168.3
Knocked Up	148.8
Juno	143.5
Rush Hour 3	140.1
Live Free or Die Hard	134.5

continued

Movie (2007)	2007 Sales (millions of dollars)
Fantastic Four: Rise of the Silver Surfer	131.9
American Gangster	130.2
Enchanted	127.8

Movie (2008)	2008 Sales (millions of dollars)
The Dark Knight	533.3
Iron Man	318.4
Indiana Jones and the Kingdom of the Crystal Skull	317.1
Hancock	227.9
WALL-E	223.8
Kung Fu Panda	215.4
Twilight	192.8
Madagascar: Escape 2 Africa	180.0
Quantum of Solace	168.4
Dr. Suess' Horton Hears a Who!	154.5
Sex and the City	152.6
Gran Torino	148.1
Mamma Mia!	144.1
Marley and Me	143.2
The Chronicles of Narnia: Prince Caspian	141.6
Slumdog Millionaire	141.3
The Incredible Hulk	134.8
Wanted	134.5
Get Smart	130.3
The Curious Case of Benjamin Button	127.5

a. Construct a dotplot of the 2008 ticket sales data. Comment on any interesting features of the dotplot. (Hint: See "What to Look For" in the Dotplots box on page 14.)

b. Construct a dotplot of the 2007 ticket sales data. Comment on any interesting features of the dotplot.

c. In what ways are the distributions of the 2007 and 2008 ticket sales observations similar? In what ways are they different?

1.21 ● About 38,000 students attend Grant MacEwan College in Edmonton, Canada. In 2004, the college surveyed non-returning students to find out why they did not complete their degree **(Grant MacEwan College Early Leaver Survey Report, 2004).** Sixty-three students gave a personal (rather than an academic) reason for leaving. The accompanying frequency distribution summarizes the primary reason for leaving for these 63 students.

Primary Reason for Leaving	Frequency
Financial	19
Health	12
Employment	8
Family issues	6
Wanted to take a break	4
Moving	2
Travel	2
Other personal reasons	10

a. Summarize the reason for leaving data using a bar chart.

b. Write a few sentences commenting on the most common reasons for leaving.

1.22 Figure EX-1.22 is a graph similar to one that appeared in ***USA Today*** **(June 29, 2009).** This graph is meant to be a bar graph of responses to the question shown in the graph.

a. Is response to the question a categorical or numerical variable?

b. Explain why a bar chart rather than a dotplot was used to display the response data.

c. There must have been an error made in constructing this graph. How can you tell that the graph is not a correct representation of the response data?

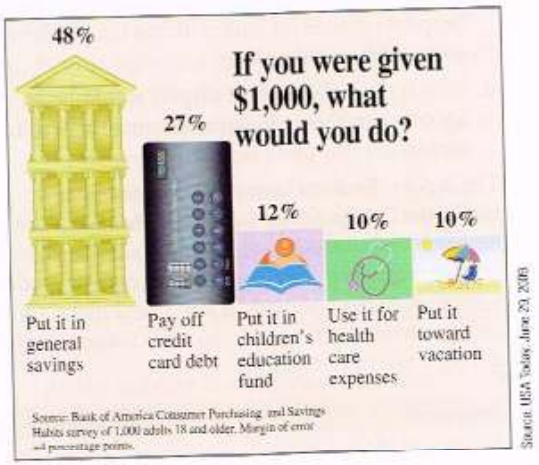


FIGURE EX-1.22

1.23 ● The online article **"Social Networks: Facebook Takes Over Top Spot, Twitter Climbs" (Compete.com, February 9, 2009)** included the accompanying data on number of unique visitors and total number of visits for January 2009 for the top 25 online social network sites. The data on total visits and unique visitors were used to compute the values in the final column of the data table, in which

$$\textit{visits per unique visitor} = \frac{\textit{total visits}}{\textit{number of unique visitors}}$$

Site	Unique Visitors	Total Visits	Visits per Unique Visitor
facebook.com	68,557,534	1,191,373,339	17.3777
myspace.com	58,555,800	810,153,536	13.8356
twitter.com	5,979,052	54,218,731	9.0681
fixter.com	7,645,423	53,389,974	6.9833
linkedin.com	11,274,160	42,744,438	3.7914
tagged.com	4,448,915	39,630,927	8.9080
classmates.com	17,296,524	35,219,210	2.0362
myyearbook.com	3,312,898	33,121,821	9.9978
livejournal.com	4,720,720	25,221,354	5.3427
imeem.com	9,047,491	22,993,608	2.5414
reunion.com	13,704,990	20,278,100	1.4796
ning.com	5,673,549	19,511,682	3.4391
blackplanet.com	1,530,329	10,173,342	6.6478
bebo.com	2,997,929	9,849,137	3.2853
hi5.com	2,398,323	9,416,265	3.9262
yuku.com	1,317,551	9,358,966	7.1033
cafemom.com	1,647,336	8,586,261	5.2122
friendster.com	1,568,439	7,279,050	4.6410
xanga.com	1,831,376	7,009,577	3.8275
360.yahoo.com	1,499,057	5,199,702	3.4686
orkut.com	494,464	5,081,235	10.2762
urbanchat.com	329,041	2,961,250	8.9996
fubar.com	452,090	2,170,315	4.8006
asiantown.net	81,245	1,118,245	13.7639
tickle.com	96,155	109,492	1.1387

continued

a. A dotplot of the total visits data is shown in Figure EX-1.23a. What are the most obvious features of the dotplot? What does it tell you about the online social networking sites?

b. A dotplot for the number of unique visitors is shown in Figure EX-1.23b. In what way is this dotplot different from the dotplot for total visits in Part (a)? What does this tell you about the online social networking sites?

c. A dotplot for the visits per unique visitor data is shown in Figure EX-1.23c. What new information about the online social networks is provided by this dotplot?

1.24 Heal the Bay is an environmental organization that releases an annual beach report card based on water quality **(Heal the Bay Beach Report Card, May 2009)**. The 2009 ratings for 14 beaches in San Francisco County during wet weather were:

A+ C B A A+ A+ A A+ B D C D F F

a. Would it be appropriate to display the ratings data using a dotplot? Explain why or why not.

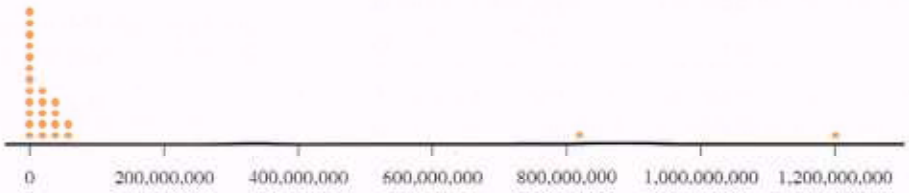


FIGURE EX-1.23a

FIGURE EX-1.23b

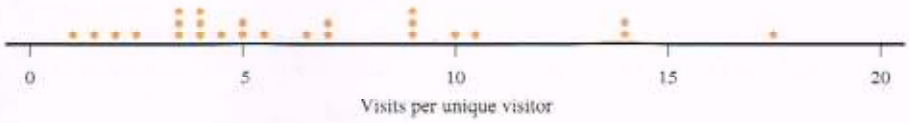


FIGURE EX-1.23c

b. Summarize the wet weather ratings by constructing a relative frequency distribution and a bar chart.

c. The dry weather ratings for these same beaches were:

A B B A+ A F A A A A A A B A

Construct a bar graph for the dry weather ratings.

d. Do the bar graphs from Parts (b) and (c) support the statement that beach water quality tends to be better in dry weather conditions? Explain.

1.25 ● The article **"Going Wireless"** (***AARP Bulletin*, June 2009**) reported the estimated percentage of households with only wireless phone service (no landline) for the 50 states and the District of Columbia. In the accompanying data table, each state was also classified into one of three geographical regions—West (W), Middle states (M), and East (E).

Wireless %	Region	State
13.9	M	AL
11.7	W	AK
18.9	W	AZ
22.6	M	AR
9.0	W	CA
16.7	W	CO
5.6	E	CN
5.7	E	DE
20.0	E	DC
16.8	E	FL
16.5	E	GA
8.0	W	HI
22.1	W	ID
16.5	M	IL
13.8	M	IN
22.2	M	IA
16.8	M	KA
21.4	M	KY
15.0	M	LA
13.4	E	ME
10.8	E	MD
9.3	E	MA
16.3	M	MI
17.4	M	MN
19.1	M	MS
9.9	M	MO
9.2	W	MT
23.2	M	NE
10.8	W	NV
16.9	M	ND
11.6	E	NH
8.0	E	NJ
21.1	W	NM
11.4	E	NY
16.3	E	NC
14.0	E	OH
23.2	M	OK
17.7	W	OR
10.8	E	PA
7.9	E	RI
20.6	E	SC
6.4	M	SD
20.3	M	TN
20.9	M	TX
25.5	W	UT
10.8	E	VA
5.1	E	VT
16.3	W	WA
11.6	E	WV
15.2	M	WI
11.4	W	WY

a. Display the data graphically in a way that makes it possible to compare wireless percent for the three geographical regions.

b. Does the graphical display in Part (a) reveal any striking differences in wireless percent for the three geographical regions or are the distributions of wireless percent observations similar for the three regions?

1.26 ● Example 1.5 gave the accompanying data on violent crime on college campuses in Florida during 2012 (from the FBI web site):

b. Write a short paragraph that could be used as part of a newspaper article on flight delays that could accompany the dotplot of the *rate per 10,000 flights* data.

1.28 The report **"Trends in Education 2010: Community Colleges" (www.collegeboard.com/trends)** included the accompanying information on student debt for students graduating with an AA degree from a public community college in 2008.

Debt	Relative Frequency
None	0.62
Less than $10,000	0.23
Between $10,000 and $20,000	0.10
More than $20,000	0.05

a. Use the given information to construct a bar chart.

b. Write a few sentences commenting on student debt for public community college graduates.

1.29 The article **"Where College Students Buy Textbooks" (*USA Today*, October 14, 2010)** gave data on where students purchased books. The accompanying frequency table summarizes data from a sample of 1152 full-time college students.

Where Books Purchased	Frequency
Campus bookstore	576
Campus bookstore web site	48
Online bookstore other than campus bookstore	240
Off-campus bookstore	168
Rented textbooks	36
Purchased mostly eBooks	12
Didn't buy any textbooks	72

a. Construct a bar chart to summarize the data distribution.

b. Write a few sentences commenting on where students are buying textbooks.

1.30 ▼ The article **"Americans Drowsy on the Job and the Road" (Associated Press, March 28, 2001)** summarized data from the 2001 Sleep in America poll. Each individual in a sample of 1004 adults was asked questions about his or her sleep habits. The article states that

> "40 percent of those surveyed say they get sleepy on the job and their work suffers at least a few days each month, while 22 percent said the problems occur a few days each week. And 7 percent say sleepiness on the job is a daily occurrence."

Assuming that everyone else reported that sleepiness on the job was not a problem, summarize the given information by constructing a relative frequency bar chart.

1.31 **"Ozzie and Harriet Don't Live Here Anymore" (*San Luis Obispo Tribune*, February 26, 2002)** is the title of an article that looked at the changing makeup of America's suburbs. The article states that nonfamily households (for example, homes headed by a single professional or an elderly widow) now outnumber married couples with children in suburbs of the nation's largest metropolitan areas. The article goes on to state:

> In the nation's 102 largest metropolitan areas, "nonfamilies" comprised 29 percent of households in 2000, up from 27 percent in 1990. While the number of married-with-children homes grew too, the share did not keep pace. It declined from 28 percent to 27 percent. Married couples without children at home live in another 29 percent of suburban households. The remaining 15 percent are single-parent homes.

Use the given information on type of household in 2000 to construct a frequency distribution and a bar chart. (Be careful to extract the 2000 percentages from the given information).

Bold exercises answered in back • Data set available online ▼ Video Solution available

ACTIVITY 1.1 Head Sizes: Understanding Variability

Materials needed: Each team will need a measuring tape.

For this activity, you will work in teams of 6 to 10 people.

1. Designate a team leader for your team by choosing the person on your team who celebrated his or her last birthday most recently.
2. The team leader should measure and record the head size (measured as the circumference at the widest part of the forehead) of each of the other members of his or her team.
3. Record the head sizes for the individuals on your team as measured by the team leader.
4. Next, each individual on the team should measure the head size of the team leader. Do not share your measurement with the other team members until

all team members have measured the team leader's head size.

5. After all team members have measured the team leader's head, record the different team leader head size measurements obtained by the individuals on your team.
6. Using the data from Step 3, construct a dotplot of the team leader's measurements of team head sizes. Then, using the same scale, construct a separate dotplot of the different measurements of the team leader's head size (from Step 5).

Now use the available information to answer the following questions:

7. Do you think the team leader's head size changed in between measurements? If not, explain why the measurements of the team leader's head size are not all the same.
8. Which data set was more variable—head size measurements of the different individuals on your team or the different measurements of the team leader's head size? Explain the basis for your choice.
9. Consider the following scheme (you don't actually have to carry this out): Suppose that a group of 10 people measured head sizes by first assigning each person in the group a number between 1 and 10. Then person 1 measured person 2's head size, person 2 measured person 3's head size, and so on, with person 10 finally measuring person 1's head size. Do you think that the resulting head size measurements would be more variable, less variable, or show about the same amount of variability as a set of 10 measurements resulting from a single individual measuring the head size of all 10 people in the group? Explain.

ACTIVITY 1.2 Estimating Sizes

1. Construct an activity sheet that consists of a table that has 6 columns and 10 rows. Label the columns of the table with the following six headings: (1) Shape, (2) Estimated Size, (3) Actual Size, (4) Difference (Estimated − Actual), (5) Absolute Difference, and (6) Squared Difference. Enter the numbers from 1 to 10 in the "Shape" column.
2. Next you will be visually estimating the sizes of the shapes in Figure 1.9. Size will be described as the number of squares of this size

that would fit in the shape. For example, the shape

would be size 3, as illustrated by

You should now quickly *visually* estimate the sizes of the shapes in Figure 1.9. *Do not* draw on the figure—these are to be quick visual estimates. Record your estimates in the "Estimated Size" column of the activity sheet.

3. Your instructor will provide the actual sizes for the 10 shapes, which should be entered into the "Actual Size" column of the activity sheet. Now complete the "Difference" column by subtracting the actual value from your estimate for each of the 10 shapes.
4. What would cause a difference to be negative? What would cause a difference to be positive?
5. Would the sum of the differences tell you if the estimates and actual values were in close agreement? Does a sum of 0 for the differences indicate that all the estimates were equal to the actual value? Explain.
6. Compare your estimates with those of another person in the class by comparing the sum of the absolute values of the differences between estimates and corresponding actual values. Who was better at estimating shape sizes? How can you tell?
7. Use the last column of the activity sheet to record the squared differences (for example, if the difference for shape 1 was -3, the squared difference would be $(-3)^2 = 9$. Explain why the sum of the squared differences can also be used to assess how accurate your shape estimates were.
8. For this step, work with three or four other students from your class. For each of the 10 shapes, form a new size estimate by computing the average of the size estimates for that shape made by the individuals in your group. Is this new set of estimates more accurate than your own individual estimates were? How can you tell?
9. Does your answer from Step 8 surprise you? Explain why or why not.

SUMMARY Key Concepts and Formulas

TERM OR FORMULA	COMMENT
Population	The entire collection of individuals or measurements about which information is desired.
Sample	A part of the population selected for study.
Descriptive statistics	Numerical, graphical, and tabular methods for organizing and summarizing data.
Inferential statistics	Methods for generalizing from a sample to a population.
Categorical data	Individual observations are categorical responses (nonnumerical).
Numerical data	Individual observations are numerical (quantitative) in nature.
Discrete numerical data	Possible values are isolated points along the number line.
Continuous numerical data	Possible values form an entire interval along then umber line.
Univariate, bivariate, and multivariate data	Each observation consists of one (univariate), two (bivariate), or two or more (multivariate) responses or values.
Frequency distribution for categorical data	A table that displays frequencies, and sometimes relative frequencies, for each of the possible values of a categorical variable.
Bar chart	A graph of a frequency distribution for a categorical data set. Each category is represented by a bar, and the area of the bar is proportional to the corresponding frequency or relative frequency.
Dotplot	A graph of numerical data in which each observation is represented by a dot on or above a horizontal measurement scale.

CHAPTER REVIEW Exercises 1.32 - 1.37

1.32 ● The report "Testing the Waters 2009" (www.nrdc.org) included information on the water quality at the 82 most popular swimming beaches in California. Thirty-eight of these beaches are in Los Angeles County. For each beach, water quality was tested weekly and the data below are the percent of the tests in 2008 that failed to meet water quality standards.

Los Angeles County

32	4	6	4	4	7	4	27	19	23
19	13	11	19	9	11	16	23	19	16
33	12	29	3	11	6	22	18	31	43
17	26	17	20	10	6	14	11		

Other Counties

0	0	0	2	3	7	5	11	5	7
15	8	1	5	0	5	4	1	0	1
1	0	2	7	0	2	2	3	5	3
0	8	8	8	0	0	17	4	3	7
10	40	3							

a. Construct a dotplot of the percent of tests failing to meet water quality standards for the Los Angeles County beaches. Write a few sentences describing any interesting features of the dotplot.

b. Construct a dotplot of the percent of tests failing to meet water quality standards for the beaches in other counties. Write a few sentences describing any interesting features of the dotplot.

c. Based on the two dotplots from Parts (a) and (b), describe how the percent of tests that fail to meet water quality standards for beaches in Los Angeles County differs from those of other counties.

1.33 The U.S. Department of Education reported that 14% of adults were classified as being below a basic literacy level, 29% were classified as being at a basic literacy level, 44% were classified as being at an intermediate literacy level, and 13% were classified as