AP Stats – Summer Work - Chapter 1 Part 1

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

Contents

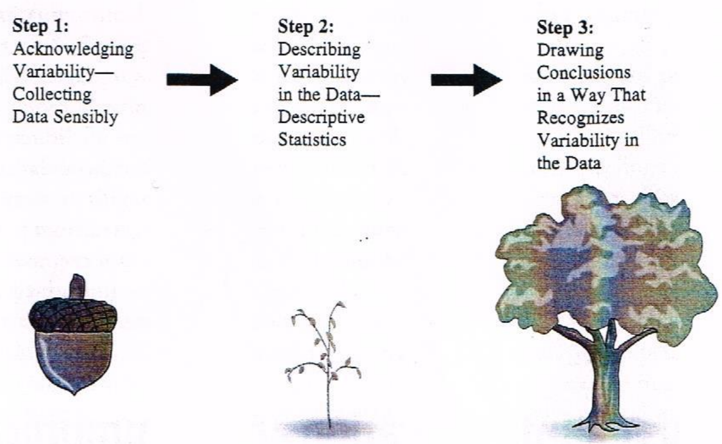
Preface

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.

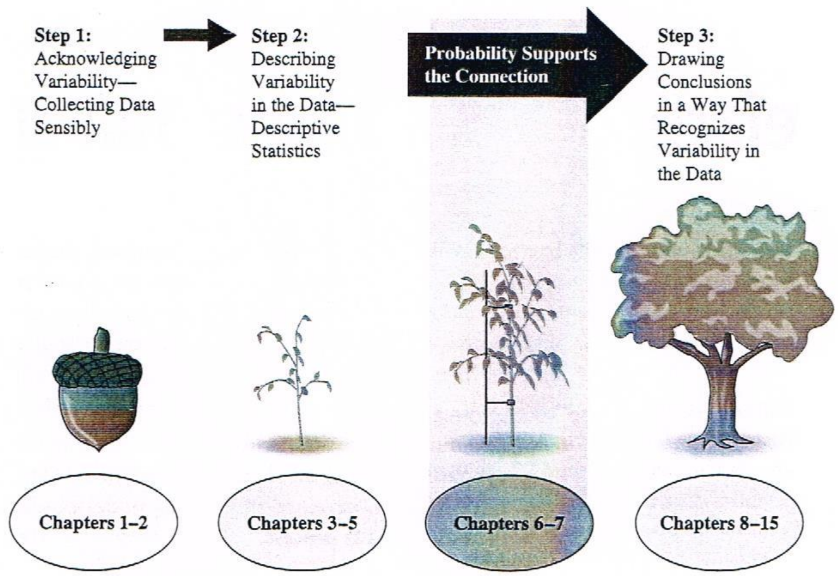
An Organization That Reflects the Data Analysis Process

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 summaries found in journals and other published sources. A subsection titled "A Word to the Wise" reminds readers of things that must be considered in order to ensure that statistical methods are employed in reasonable and appropriate ways.

Consistent with Recommendations for the Introductory Statistics Course Endorsed by the American Statistical Association

In 2005, the American Statistical Association endorsed the report "College Guidelines in Assessment and Instruction for Statistics Education (GAISE Guidelines)," which included the following six recommendations for the introductory statistics course:

1. Emphasize statistical literacy and develop statistical thinking.
2. Use real data.
3. Stress conceptual understanding rather than mere knowledge of procedures.
4. Foster active learning in the classroom.
5. Use technology for developing conceptual understanding and analyzing data.
6. Use assessments to improve and evaluate student learning.

Introduction to Statistics and Data Analysis, Fifth Edition, is consistent with these recommendations and supports the GAISE guidelines in the following ways:

1. **Emphasize statistical literacy and develop statistical thinking.**
   Statistical literacy is promoted throughout the text in the many examples and exercises that are drawn from the popular press. In addition, a focus on the role of variability, consistent use of context, and an emphasis on interpreting and communicating results in context work together to help students develop skills in statistical thinking.
2. **Use real data.**
   The examples and exercises from *Introduction to Statistics and Data Analysis,* Fifth Edition, are context driven, and the reference sources include the popular press as well as journal articles.
3. **Stress conceptual understanding rather than mere knowledge of procedures.**
   Nearly all exercises in *Introduction to Statistics and Data Analysis,* Fifth Edition, are multipart and ask students to go beyond just computation. They focus on interpretation and communication, not just in the chapter sections specifically devoted to this topic, but throughout the text. The examples and explanations are designed to promote conceptual understanding. Hands-on activities in each chapter are also constructed to strengthen conceptual understanding. Which brings us to . . .
4. **Foster active learning in the classroom.**
   While this recommendation speaks more to pedagogy and classroom practice, *Introduction to Statistics and Data Analysis,* Fifth Edition, provides more than 30 hands-on activities in the text and additional activities in the accompanying instructor resources that can be used in class or assigned to be completed outside of class.
5. **Use technology for developing conceptual understanding and analyzing data.**
   The computer brings incredible statistical power to the desktop of every investigator. The wide availability of statistical computer packages such as Minitab,

THIS PAGE

is intentionally

BLANK

CHAPTER 1

The Role of Statistics and the Data Analysis Process

© Andresr/Shutterstock.com

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 ⟩

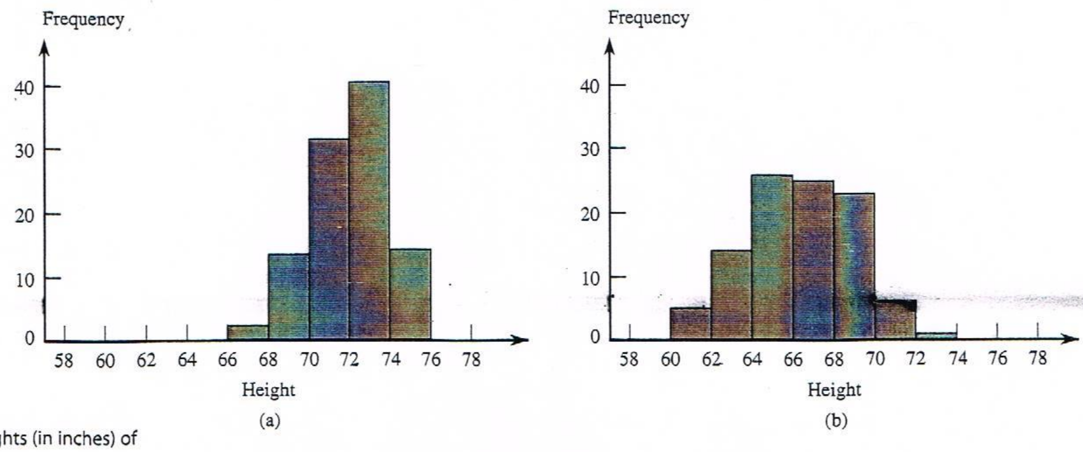


FIGURE 1.1
Histograms of heights (in inches) of female athletes:
(a) 100 basketball players;
(b) 100 gymnasts.

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.

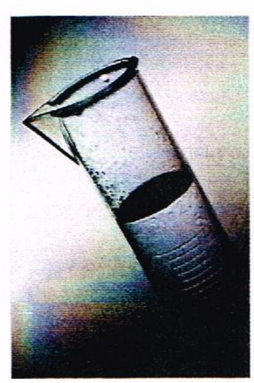
Interpret the results ⟩

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



David Chasey/Photodisc/Getty Images

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?

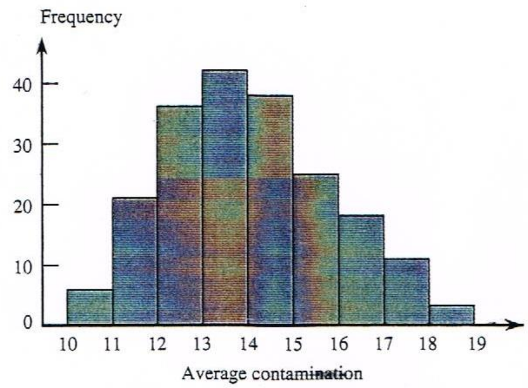


FIGURE 1.2
Average contamination concentration (in parts per million) measured each day for 200 days.

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