CHAPTER 1

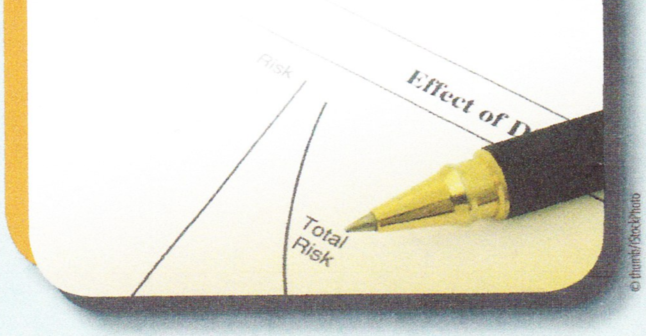
What Is Statistics?

objectives

1-1 Key Statistical Concepts

1-2 Large Real Data Sets

1-3 Statistics and the Computer



© thumb/iStockPhoto

INTRODUCTION

Statistics is a way to get information from data. That's it! Most of this textbook is devoted to describing how, when, and why managers and statistics practitioners[1] conduct statistical procedures. You may ask, "If that's all there is to statistics, then why is this book (and most other statistics books) so large?" The answer is that there are different kinds of information and data to which students of applied statistics should be exposed. We demonstrate some of these with three examples here, one of which is featured later in this book.

[1] The term *statistician* is used to describe so many different kinds of occupations that it no longer has any meaning. It is used, for example, to describe both a person who calculates baseball statistics and an individual educated in statistical principles. We will describe the former as a *statistics practitioner* and the latter as a *statistician*. A statistics practitioner is a person who uses statistical techniques properly. Examples of statistics practitioners include the following:

1. a financial analyst who develops stock portfolios based on historical rates of return;
2. an economist who uses statistical models to help explain and predict variables such as inflation rate, unemployment rate, and changes in the gross domestic product; and
3. a market researcher who surveys consumers and converts the responses into useful information.

Our goal in this book is to convert you into one such capable individual.

The term *statistician* refers to an individual who works with the mathematics of statistics. His or her work involves research that develops techniques and concepts that in the future may help the statistics practitioner. Statisticians are also statistics practitioners, frequently conducting empirical research and consulting. If you're taking a statistics course, your instructor is probably a statistician.

EXAMPLE 1.1

Business Statistics Marks

A student who is enrolled in a business program is attending his first class of the required statistics course. The student is somewhat apprehensive because he believes the myth that the course is difficult. To alleviate his anxiety, the student asks the professor about last year's marks. Because, like all other statistics professors, this one is friendly and helpful, he obliges the student and provides a list of the final marks, which are composed of term work plus the final exam. What information can the student obtain from the list?

This is a typical statistics problem. The student has the data (marks) and needs to apply statistical techniques to get the information he requires. This is a function of **descriptive statistics**.

Peter Dazeley/Photographer's Choice/Getty Images

[illegible]/Shutterstock.com

Descriptive Statistics

Descriptive statistics deals with methods of organizing, summarizing, and presenting data in a convenient and informative way. One form of descriptive statistics uses graphical techniques that allow statistics practitioners to present data in ways that make it easy for the reader to extract useful information. In Chapter 2, we will present a variety of graphical methods.

Another form of descriptive statistics uses numerical techniques to summarize data. One such method that you have already used frequently calculates the average or mean. We can compute the mean mark of last year's statistics course by summing the marks and dividing by the number of marks. Chapter 3 introduces several numerical statistical measures that describe different features of the data.

The actual technique we use depends on what specific information we would like to extract. In Example 1.1, we can see at least three important pieces of information. The first is the "typical" mark. We call this a *measure of central location*. The average is this type of measure. Another measure of central location is the median which is the mark that divides the top half from the bottom half. Suppose the student was told that the average mark last year was 67. Is this enough information to reduce his anxiety? The student would likely respond "no" because he would like to know whether most of the marks were close to 67 or were scattered far below and above the average. He needs a *measure of variability*. The simplest such measure is the *range*, which is calculated by subtracting the smallest number from the largest. Suppose the largest mark is 96 and the smallest is 24. Unfortunately, this provides little information. We need other measures, and these will be introduced in Chapter 3. Moreover, the student must determine more about the marks. In particular, he needs to know how the marks are distributed between 24 and 96. The best way to do this is to use a graphical technique, the histogram, which will be introduced in Chapter 2.

EXAMPLE 1.2

Pepsi's Exclusivity Agreement with a University

In the last few years, colleges and universities have signed exclusivity agreements with a variety of private companies. These agreements bind the university to sell that company's products exclusively on

The Power of Forever Photography/iStockphoto.com

the campus. Many of the agreements involve food and beverage firms.

A large university with a total enrollment of about 50,000 students has offered Pepsi-Cola an exclusivity agreement that would give Pepsi exclusive rights to sell its products at all university facilities for the next year with an option for future years. In return, the university would receive 35% of the on-campus revenues and an additional lump sum of $200,000 per year. Pepsi has been given 2 weeks to respond.

The management at Pepsi quickly reviews what it knows. The market for soft drinks is measured in terms of 12-ounce cans. Pepsi currently sells an average of 22,000 cans per week (over the 40 weeks of the year that the university operates). The cans sell for an average of one dollar each. The costs including labor amount to 30 cents per can. Pepsi is unsure of its market share but suspects it is considerably less than 50%. A quick analysis reveals that if its current market share were 25%, then, with an exclusivity agreement, Pepsi would sell 88,000 (22,000 is 25% of 88,000) cans per week or 3,520,000 cans per year. The gross revenue would be computed as follows:[2]

[2] We have created an Excel spreadsheet that does the calculations for this case. To access it, click **Excel Workbooks** and **Example 1.2**. The only cell you may alter is cell C3, which contains the average number of soft drinks sold per week per student, assuming a total of 88,000 drinks sold per year.

$$\text{Gross revenue} = 3{,}520{,}000 \times \$1.00/\text{can} = \$3{,}520{,}000$$

This figure must be multiplied by 65% because the university would rake in 35% of the gross. Thus,

$$\text{Gross revenue after deducting 35\% university take} = 65\% \times \$3{,}520{,}000 = \$2{,}288{,}000$$

The total cost of 30 cents per can (or $1,056,000) and the annual payment to the university of $200,000 are subtracted to obtain the net profit:

$$\text{Net profit} = \$2{,}288{,}000 - \$1{,}056{,}000 - \$200{,}000 = \$1{,}032{,}000$$

Pepsi's current annual profit is

$$40 \text{ weeks} \times 22{,}000 \text{ cans/week} \times \$.70 = \$616{,}000$$

If the current market share is 25%, the potential gain from the agreement is

$$\$1{,}032{,}000 - \$616{,}000 = \$416{,}000$$

The only problem with this analysis is that Pepsi does not know how many soft drinks are sold weekly at the university. Coke is not likely to supply Pepsi with information about its sales, which together with Pepsi's line of products constitute virtually the entire market.

Pepsi assigned a recent university graduate to survey the university's students to supply the missing information. Accordingly, she organizes a survey that asks 500 students to keep track of the number of soft drinks they purchase in the next 7 days. The responses are stored in a file named Xm01-02 available from our Web site.

Inferential Statistics

The information we would like to acquire in Example 1.2 is an estimate of annual profits from the exclusivity agreement. The data are the numbers of cans of soft drinks consumed in 7 days by the 500 students in the sample. We can use descriptive techniques to learn more about the data. In this case, however, we are not so much interested in what the 500 students are reporting as we are in knowing the mean number of soft drinks consumed by all 50,000 students on campus. To accomplish this goal, we need another branch of statistics—**inferential statistics.**

Inferential statistics is a body of methods used to draw conclusions or inferences about characteristics of populations based on sample data. The population in question in this example is the soft drink consumption of the university's 50,000 students. The cost of interviewing each student would be *prohibitive* and extremely time consuming. Statistical techniques make such endeavors unnecessary. Instead, we can sample a much smaller number of students (the sample size is 500) and infer from the data the number of soft drinks consumed by all 50,000 students. We can then estimate annual profits for Pepsi.

© Mark Richards/PhotoEdit

EXAMPLE 10.1

Exit Polls (see Chapter 10)

When an election for political office takes place, the television networks cancel regular programming and provide election coverage instead. When the ballots are counted, the results are reported. However, for important offices such as president or senator in large states, the networks actively compete to see which will be the first to predict a winner. This is done through *exit polls* in which a random sample of voters who exit polling booths are asked who they voted for. From the data, the sample proportion of voters supporting the candidates is computed. A statistical technique is applied to determine whether there is enough evidence to infer that the leading candidate will garner enough votes to win. Suppose that the exit poll results from the state of Florida during the 2000 year elections were recorded. Although there were a number of candidates running for president, the exit pollsters recorded only the votes of the two candidates who had any chance of winning: Republican candidate George W. Bush and Democrat Albert Gore. The results (900 people who voted for either Bush or Gore) were stored on a file (Xm10-01), which can be downloaded from our Web site. The network analysts would like to know whether they can conclude that George W. Bush will win the state of Florida.

Example 10.1 describes a very common application of statistical inference. The population the television networks wanted to make inferences about is the approximately 5 million Floridians who voted for Bush or Gore for president. The sample consisted of the 900 people randomly selected by the polling company who voted for either of the two main candidates. The characteristic of the population that we would like to know is the proportion of the Florida total electorate that voted for Bush. Specifically, we would like to know whether more than 50% of the electorate voted for Bush (counting only those who voted for either the Republican or Democratic candidate). Because we will not ask every one of the 5 million actual voters who they voted for, it must be made clear that we cannot predict the outcome with 100% certainty. This is a fact that statistics practitioners and even students of statistics must understand. A sample that is only a small fraction of the size of the population can lead to correct inferences only a certain percentage of the time. You will find that statistics practitioners can control that fraction and usually set it between 90% and 99%.

Incidentally, on the night of the United States election in November 2000, the networks goofed badly. Using exit polls as well as the results of previous elections, all four networks concluded at about 8 P.M. that Al Gore would win the state of Florida. Shortly after 10 P.M., with a large percentage of the actual vote having been counted, the networks reversed course and declared that George W. Bush would win Florida. By 2 A.M., another verdict was declared: the result was too close to call. This experience is often used by statistics instructors when they teach how *not* to use statistics.

Contrary to what you probably believed, notice that data are not necessarily numbers. The marks in Example 1.1 and the number of soft drinks consumed in a week in Example 1.2, of course, are numbers, but the votes in Example 10.1 are not. In Chapter 2, we will discuss the different types of data you will encounter in statistical applications and how to deal with them.

1-1 Key Statistical Concepts

© Glowimages

Statistical inference problems involve three key concepts: the population, the sample, and the statistical inference. We now discuss each concept in more detail.

1-1a Population

A **population** is the group of all items of interest to a statistics practitioner. It is frequently very large and may, in fact, be infinitely large. In the language of statistics, *population* does not necessarily refer to a group of people. It may, for example, refer to the population of diameters of ball bearings produced at a large plant. In Example 1.2, the population of interest consists of the 50,000 students on campus. In Example 10.1 the population consists of the Floridians who voted for Bush or Gore.

A descriptive measure of a population is called a **parameter**. The parameter of interest in Example 1.2 is the mean number of soft drinks consumed by all the students at the university. The parameter in Example 10.1 is the proportion of the 5 million Florida voters who voted for Bush. In most applications of inferential statistics, the parameter represents the information we need.

1-1b Sample

A **sample** is a set of data drawn from the population. A descriptive measure of a sample is called a **statistic**. We use statistics to make inferences about parameters. In Example 1.2, the statistic we would compute is the mean number of soft drinks consumed in the last week by the 500 students in the sample. We would then use the sample mean to infer the value of the population mean, from which we can estimate the profit. In Example 10.1,

© Anatoliy Samara/Shutterstock.com

we compute the proportion of the sample of 900 Floridians who voted for Bush. The sample statistic is then used to make inferences about the population of all 5 million votes. In other words, we predict the election results even before the actual count.

1-1c Statistical Inference

Statistical inference is the process of making an estimate, prediction, or decision about a population based on sample data. Because populations are almost always very large, investigating each member of the population would be impractical and expensive. It is far easier and cheaper to take a sample from the population of interest and draw conclusions or make estimates about the population on the basis of information provided by the sample. However, such conclusions and estimates will not always be correct. For this reason, we build into the statistical inference a measure of reliability. There are two such measures: the **confidence level** and the **significance level**. The *confidence level* is the proportion of times that an estimating procedure will be correct. In Example 1.2, we could produce an estimate of the average number of soft drinks to be consumed by all 50,000 students that has a confidence level of 95%. In other words, in the long run, estimates based on this form of statistical inference will be correct 95% of the time. When the purpose of the statistical inference is to draw a conclusion about a population, the *significance level* measures how frequently the conclusion will be wrong in the long run. For example, suppose that as a result of the analysis in Example 10.1, we conclude that more than 50% of the electorate will vote for George W. Bush and thus he will win the state of Florida. A 5% significance level means that, in the long run, this type of conclusion will be wrong 5% of the time.

1-2 Large Real Data Sets

© Glowimages

The author believes that you learn statistics by doing statistics. In life after college and university, we expect that our students will have access to large amounts of real data that must be summarized to acquire the information needed to make decisions. To provide practice in this vital skill. We have included on our Web site two sets of large real data sets. Their sources are the General Social Survey (GSS) and the American National Election Survey (ANES).

1-2a General Social Survey

Since 1972, the General Social Survey has been tracking American attitudes on a wide variety of topics. Except for the United States census, the GSS is the most frequently used source of information about American society. The surveys now conducted every second year measure hundreds of variables and thousands of observations. We have included the results of the last six surveys (years 2002, 2004, 2006, 2008, 2010, and 2012) stored as GSS2002, GSS2004, GSS2006, GSS2008, GSS2010, and GSS2012, respectively. The survey sizes are 2,765, 2,812, 4,510, 2,023, 2,044, and 1974, respectively. We have reduced the number of variables to about 60 and have deleted the responses that are known as *missing* data ("Don't know," "Refused," etc.).

We have included some demographic variables such as, age, gender, race, income, and education. Others measure political views, support for various government activities, and work. The full lists of variables for each year are stored on our Web site in Appendixes GSS2002, GSS2004, GSS2006, GSS2008, GSS2010, and GSS2012.

We have scattered throughout this book examples and exercises drawn from these data sets.

1-2b American National Election Survey

The goal of the American National Election Survey is to provide data about why Americans vote as they do. The surveys are conducted in the years of presidential elections. We have included data from the 2004 and 2008 surveys. Like the General Social Survey, the ANES includes demographic variables. It also deals with interest in the presidential election as well as variables that describe political beliefs and affiliations. Web site Appendixes ANES2004 and ANES2008 contain the names and definitions of the variables.

The 2008 surveys overly sampled black and Hispanic voters. We have "adjusted" these data by randomly deleting responses from these two racial groups.

As is the case with the General Social Surveys, we have removed the missing data.

1-3 Statistics and the Computer

© Glowimages

In virtually all applications of statistics, the statistics practitioner must deal with large amounts of data. For instance, Example 1.2 involves 500 observations. To estimate annual profits, the statistics practitioner would have to perform computations on the data; although the calculations do not require any great mathematical skill, the sheer amount of arithmetic makes this aspect of the statistical method time consuming and tedious.

Fortunately, numerous commercially prepared computer programs are available to perform the arithmetic. We have chosen to use Microsoft Excel, which is a spreadsheet program, and Minitab, a statistical software package. We chose Excel because we believe that it is and will continue to be the most popular spreadsheet package. One drawback, however, is that it does not offer a complete set of the statistical techniques that we introduce in this book. Consequently, we have created add-ins that can be loaded onto your computer to enable you to use Excel for all statistical procedures introduced in this book. The add-ins can be downloaded from our Web site. When installed, they will appear as Data Analysis Plus© on Excel's menu. Also available on the Web site are introductions to Excel and Minitab and detailed instructions for both software packages.

A large proportion of the examples and exercises feature large data sets that are also stored on the Web site. These are denoted with the file name. We demonstrate the solution to the statistical examples in three ways: manually, by employing Excel, and by using Minitab. Moreover, we will provide detailed instructions for all techniques.

The files contain the data needed to produce the solution. However, in many real applications of statistics, additional data are collected. For example, the interviewer at exit polls often records the gender and asks the voter for other information, including race, religion, education, and income. Many other data sets are similarly constructed. In later chapters, we will return to these files and require other statistical techniques to extract the needed information. (Files that contain additional data are denoted by an asterisk on the file name.)

The approach we prefer to take is to minimize the time spent on manual computations and to focus instead on selecting the appropriate method for dealing with a problem and on interpreting the output after the computer has performed the necessary computations. In this way, we hope to demonstrate that statistics can be as interesting and as practical as any other subject in your curriculum.