PART I

INTRODUCTION

Chapter 1
The Nature of Statistics

Chapter 2
Learning about MINITAB
Chapter 1

The Nature of Statistics

Looking Ahead

After reading this chapter, you will know what you can expect to learn from this text. Among other things, you will:

1. come to know statistics as a field of study that develops and uses techniques for the careful collection, effective presentation, and proper analysis of numerical information (a fact that is reflected in the titles of Parts II to VI of this text),
2. get a chapter-by-chapter preview of the kinds of business and economic problems you will learn to solve and, in the process, come to see why statistics is said to facilitate wise decision making in the face of uncertainty and is viewed as a “universal guide to the unknown,”
3. meet a number of basic statistical concepts, including statistical populations, samples, variables, data, and more, and
4. learn to distinguish different types of data (such as nominal, ordinal, interval, and ratio data) and, thus, develop an awareness of different data qualities, which, in turn, determines what types of arithmetic operations can be performed with these data.

And Here is a Typical Problem You Will Be Able to Solve:

You just received Fortune magazine’s latest Global 500 report, which provides information on the world’s 500 largest corporations. Besides each company’s name, country, industry code, and number of employees, the report includes dollar amounts and rankings of each firm’s revenues, profits, assets, and stockholders’ equity.

a. Identify the elementary units.
b. How many variables can you find in this report? What are they?
c. Identify the variables as quantitative or qualitative.
d. Identify variables as discrete/continuous or binomial/multinomial.
e. Can you find examples of nominal, ordinal, interval, and ratio data in this report? Why would you care?
You are a film producer and your studio has just spent millions of dollars creating a new soap opera that seems destined to be shown on television. Naturally, you want to make as much money as possible. Thus, it is time to devise a marketing strategy. Several possibilities come to mind:

- First, all rights to the new series could be sold to a distributor who is willing to pay $125 million right now, and that could be the end of the story as far as you are concerned.

- Second, the program could be offered to a TV network for review with these possible results (in your judgment): a 60 percent chance of rejection (which ruins all further chances of a sale to anyone and spells a $30 million loss) or a 40 percent chance of getting a contract (which means a $300 million profit).

- Third, you could hire a consulting firm, which is willing to offer advice on the network’s likely reaction for a $1 million fee. The consulting firm’s “track record” is given in Table 1.1. During the past decade, the consulting firm has issued numerous reports in similar situations. An ultimate rejection—an event here designated as \( E_1 \)—was preceded by a report predicting rejection, \( R_1 \), 80 percent of the time, and by a report predicting a contract offer, \( R_2 \), 20 percent of the time. On the other hand, an eventual contract offer—an event here designated as \( E_2 \)—was preceded by a report predicting rejection 30 percent of the time and by a report predicting a contract offer 70 percent of the time.

What then is your optimal strategy? Should you grasp the sure thing and pocket $125 million now by selling the rights? Should you take the $30 million versus $300 million gamble by showing the pilot to network executives? Should you buy the advice and then take the action that maximizes your likely revenue (sell the rights if report \( R_1 \) is received; offer the film to the network if report \( R_2 \) is received)? Or should you buy the advice now and, after having received your report, rethink the whole matter in light of Table 1.1?

Surely, you can picture yourself now, with report \( R_1 \) in hand, about to make that $125 million deal with the distributor, yet thinking about the lost chance of making $300 million if the report is wrong. Surely, you can see yourself fretting all night, with report \( R_2 \) in hand, about to contact the network and collect that $300 million prize, while thinking about the very real chance, if the report is wrong, of ending up $30 million in the red instead of pocketing $125 million for sure. . . .

Business executives face decision-making problems like this every day. By the time you have studied the last chapter of this text, you will be able to solve this particular problem in no time. And you will acquire similar skills in every chapter in between.

### TABLE 1.1 | Film Consultant’s Track Record

<table>
<thead>
<tr>
<th>Prior Advice Given</th>
<th>( E_1 = \text{Rejection} )</th>
<th>( E_2 = \text{Contract Offer} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_1 ) = network will reject</td>
<td>80%</td>
<td>30%</td>
</tr>
<tr>
<td>( R_2 ) = network will offer contract</td>
<td>20%</td>
<td>70%</td>
</tr>
</tbody>
</table>
1.1 Introduction

Ask anyone to define the nature of *statistics* and, just as in the dictionary, you are likely to get one of three answers. Some, like you perhaps, are about to take a course in the subject. They will naturally think of statistics as a *field of study* that somehow deals with the collection, presentation, and interpretation of numerical data. There will be others, the vast majority of people no doubt, who will instantly think of *masses of data*, seemingly infinite in number, that constantly bombard us in our daily lives. Just think of all those numbers ceaselessly spewing forth from television sets, radios, newspapers, and sites on the World Wide Web: data about the weather and sports events; election results and opinion polls; prices of bonds, stocks, foreign monies, and commodity futures; rates of inflation, unemployment, and economic growth.... Finally, a few other people, already trained in the discipline, will conjure up a highly technical meaning of the term that we, too, will meet in later chapters. The term *statistics* can refer to *summary measures*, such as sample averages and sample proportions, that have been computed from relatively few data gathered by *sampling* a much larger collection of data called a *population*.

In fact, these three definitions are linked. Statistics, viewed as a scientific discipline, inevitably uses as raw material those very masses of data that most people associate with the term. Indeed, statistics courses used to have an ugly reputation precisely because they involved endless, boring hours of manipulating masses of data. But such number crunching is no more. Sophisticated computer software, such as Microsoft’s EXCEL or the MINITAB program packaged with one version of this text, can perform powerful magic. As Chapter 2 illustrates, upon starting either program, you encounter a screen that is nicely divided into a series of columns and rows and invites you to enter data, masses of them, if you want. Indeed, in the case of EXCEL, a single Workbook contains 16 blank spreadsheets. Each of these measures 256 columns by 65,536 rows. That comes to 16,777,216 cells, which could be printed on a sheet of paper 19 feet wide and 1,300 feet long! In the case of MINITAB’s Professional Version, an eager user can fill in 150 million of those pretty cells (provided there is sufficient computer memory), but even the lower limit of a mere 5,000 entries in the Student Version can keep you busy for quite some time. Relax! You won’t have to do that in this course.

Having entered your data into either program, you must specify an appropriate statistical technique. A few well-chosen keystrokes will do, and wham! In a fraction of a second, you have your result. Times have certainly changed. Not so many years ago, some of these calculations might have taken weeks and even months or years of work.

1.2 Collecting Data

Any practical statistical work requires data, data, and more data. A first branch of the discipline of statistics, therefore, focuses on the careful collection of this crucial type of raw material. Such collection can proceed in one of three ways: (1) An investigator can look for data that already exist because others have gathered them in the past. (2) Brand-new data can be generated with the help of so-called observational studies that involve census taking or sampling. (3) Brand-new data can be generated by conducting carefully controlled experiments. These three approaches are explained in Chapters 3, 4, and 5 of this text, respectively.

**FINDING EXISTING DATA**

When relevant information already exists somewhere, an investigator need only find it. A business administrator, for example, might simply search the firm’s internal records for material that quietly resides in filing cabinets or computer memories. Thus, customer records would provide...
names, addresses, telephone numbers, data on amounts purchased, credit limits, and more. Employee records would provide names, addresses, job titles, years of service, salaries, Social Security numbers, and even numbers of sick days used. Production records would contain lists of products, part numbers, and quantities produced, along with associated labor costs, raw material consumption, and equipment usage. A government economist would, similarly, have access to vast databases held by the Bureau of the Census, the Department of Labor, the Federal Reserve Board, and the Office of Management and Budget, to name but a few. From the point of view of the business administrator or the government economist, respectively, all of the sources just mentioned are \textit{internal} sources.

In addition to scouring internal sources of information, our business administrator or government economist could also look for \textit{external} depositories of already existing data and persuade their owners to share information. Indeed, all kinds of organizations, ranging from Dow Jones and Company to the Dun and Bradstreet Corporation to the Medical Economics Company, routinely gather data and sell them to would-be users in the private sector and in government agencies alike. As a student, you are already familiar with the use of preexisting data. When writing your next paper, you are unlikely to generate brand-new information (although you will learn how to do so in this text). More likely, you will turn to some source of data gathered by someone else. You might check out the \textit{Statistical Abstract of the United States} or one of the numerous other sources listed in its Appendix I, Guide to Sources of Statistics, State Statistical Abstracts, and Foreign Statistical Abstracts, a perusal of which we highly recommend. Better yet, you might check out the Internet’s FedStats site (introduced in Chapter 3) that leads you to data collected by over 70 agencies of the U.S. government. Or you might click on the link to any one of almost 200 foreign government sites or visit the sites of any \textit{Fortune 500} company, every one of which is listed on the CD-ROM accompanying this text. Yet, from the point of view of the professional statistician, the matter of collecting data, especially \textit{current} data, is usually much more complicated than checking out internal or external sources of existing data. More often than not, this question arises: How can trustworthy \textit{new} data be generated?

\section*{Generating New Data}

New data about persons or objects possessing characteristics that interest a statistician can be generated either by conducting a complete or sample \textit{survey} or by performing a \textit{controlled experiment}. These two approaches are presented in Chapters 4 and 5, respectively. Here we introduce the general idea, which is conveyed by these definitions:

\begin{quote}
\textbf{Definition 1.1} The collection of data about persons or objects by merely recording information about selected characteristics of interest (such as A or B), while paying no attention to possibly widely diverging other characteristics (such as C or D) that may affect the chosen characteristics, is called an \textit{observational study} or \textit{survey}.

\textbf{Definition 1.2} The collection of data about persons or objects by deliberately exposing them to some kind of change, while leaving all else unchanged, and subsequently recording how identical persons or objects respond to different types of change, or how different types of persons or objects respond to identical change, is called a \textit{controlled experiment}.
\end{quote}

\section*{Generating Survey Data}

In a \textit{survey}, a characteristic such as the annual salary of workers, for example, may simply be observed and recorded for different workers without regard to factors, like length of service, education, and work experience, that make workers different from one another. These factors may, in fact, be responsible for the observed differences in their salaries. If, by pure accident, a firm employs lots of recently hired women with little education
and next to no work experience, while also employing lots of men hired decades ago who are highly educated and have plenty of work experience, you can guess what will happen: A survey of salaries that ignores length of service, education, and work experience can easily create the impression that women are being paid less than men merely because of their sex.

This being said, you can, perhaps, anticipate an answer to the following survey problem:

**CHAPTER 4 | Generating New Data: Census Taking and Sampling**

Domino’s Pizza once was sued by Amstar, maker of Domino sugar, on the grounds that use of this name confused people. Indeed, Amstar had interviewed women shopping in supermarkets, had shown them a Domino’s Pizza box, and had asked whether they thought the pizza makers produced any other product. Some 71 percent said “sugar.” If you were a statistician hired by Domino’s Pizza, how could you help your client?

**GENERATING EXPERIMENTAL DATA** Experimental data, in contrast to survey data, are generated more carefully. Thus, a firm may divide its 40 new employees into two groups of equal size (with the help of some random device you will learn about in Chapter 4). It may then administer a special training program to one of the groups only. If the 20 employees who went through the program exhibited superior productivity later on, the training program might justifiably be credited with those improvements. After all, other factors that could account for this result, such as group differences in age, motivation, or prior work experience, were effectively equalized by random division of the original group of 40 workers.

Can you anticipate an answer to the following question that you will soon encounter?

**CHAPTER 5 | Generating New Data: Controlled Experiments**

Pharmaceutical companies can increase their sales by billions of dollars per year if they are lucky enough to come up with a new best-selling drug. No wonder they are always experimenting. One such company selected 100 adults at random and managed to persuade 80 of them to take calcium supplements, which were suspected of lowering blood pressure. A comparison of the subjects’ before-and-after blood pressure readings confirmed the suspicion. As a statistician hired by the Food and Drug Administration, evaluate this experiment.

Despite the fact that experimental data tend to be more reliable or “stronger” than survey data, most new data in business, economics, and many other fields, are not generated by controlled experiments. More likely than not, they are generated by (complete or sample) surveys. This happens because it is often impossible, or extremely costly, to carry out experiments. We could not easily divide the country’s labor force (or even a segment of it) at random into three groups and then subject each group to different tax rates in order to study the effect of taxes on labor supply. Nor could we simply divide the country’s newly born (or even a segment of them) at random into two groups and then subject only one of the groups to a lifetime of smoking in order to study the effect of smoking on health. Nevertheless, as you will learn later in this text, with proper statistical techniques, we can learn a great deal about such matters even from surveys.

Our discussion so far, however, must not be misunderstood. The modern scientific discipline of statistics is not very well described by the popular image of statistics as a field of study that is preoccupied with the acquisition and publication of masses of data. Much more so, statistics is
about the subsequent description and analysis of data. Indeed, most textbooks, including this one, emphasize the latter two branches of the discipline, that is, descriptive statistics and inferential statistics, over methods for collecting data.

1.3 Describing Data

Once we have collected data, they become the raw material for those laboring in a second branch of the statistics discipline. The descriptive statistician focuses on the postcollection task of effectively presenting data, of organizing and condensing them, usually with the help of tables and graphs (Chapter 6) or with the help of numerical summary measures (Chapter 7). As we will see, unlike mere listings of masses of raw data, such as those that the Internet provides, this type of presentation alone makes data understandable and often reveals patterns otherwise hidden in unprocessed data.

**DEFINITION 1.3** A branch of the discipline that is concerned with developing and using techniques for effectively presenting numerical information so as to highlight patterns otherwise hidden in data sets is called descriptive statistics.

Here are two examples of working with descriptive statistics:

**CHAPTER 6 | Presenting Data: Tables and Graphs**

Imagine you were working for a chemicals firm that is interested in expanding its fertilizer sales in California and Florida. You are supposed to provide information about current fertilizer usage, and you instantly think of oranges. Create a relevant cross tabulation for your next staff meeting. (Hint: To find current data, visit [http://www.fedstats.gov](http://www.fedstats.gov), a site maintained by the federal government’s Interagency Council on Statistical Policy.)

**CHAPTER 7 | Presenting Data: Summary Measures**

Your boss wants to acquire a major software company. You are to gather relevant information. Visit [http://www.fortune.com](http://www.fortune.com), a site maintained by Fortune magazine. Identify the revenues, profits, and numbers of employees of each of the Fortune 500 companies in the computer software industry. Using EXCEL or MINITAB, compute and print out descriptive summary statistics on the three types of data series just noted.

The importance of descriptive statistics is illustrated vividly by numerous other examples found later in this text. As we will show, the effective presentation of data can lead a production engineer to discover the secret behind the recent breakdowns of motors produced by a firm, can help a historian unravel the mystery of a disputed authorship, can aid spies in cracking a secret code, and can enable managers to monitor quality in an ongoing production process.

As useful as it may be, however, descriptive statistics too comprises only a small part of the modern discipline of statistics. Contrary to another common view that identifies data collection and descriptive statistics with the entire field of study, another branch of the discipline is nowadays considerably more important.
1.4 Analyzing Data

Modern statisticians direct most of their effort not toward collecting and presenting numerical information but toward analyzing it. They are laboring in a third branch yet of the statistics discipline. They focus on applying reason to data in order to draw sensible conclusions from them. Their chief concern is making reasonable inferences, from the limited information that is available, about matters that are not known. Accordingly, 16 chapters of this book (Chapters 8–23) deal exclusively with analytical or inferential statistics.

**DEFINITION 1.4** A branch of the discipline of statistics that is concerned with developing and using techniques for properly analyzing (or drawing inferences from) numerical information is called analytical statistics or inferential statistics.

Sometimes a general truth is inferred from particular instances. At other times, statisticians reverse the process and draw conclusions about the particular from their knowledge of the general.

**DRAWING INFERENCES BY INDUCTIVE REASONING**

Drawing inferences about an unknown whole from a known part is called inductive reasoning. Inductive reasoning is at work, for example, when a statistician concludes that between 1 and 3 percent of a firm’s total output is defective because 2 percent of an output sample did, in fact, not meet quality standards. (As we will learn, even this result may be clouded by uncertainty, because the method used to derive this conclusion may provide correct results only 95 percent of the time and incorrect results in the remaining cases.)

**DRAWING INFERENCES BY DEDUCTIVE REASONING**

On the other hand, drawing inferences about an unknown part from a known whole is deductive reasoning. Deductive reasoning is at work, for example, when a statistician concludes that a particular unit of a firm’s output (namely, a unit produced at plant 7, during the night shift, and with the help of components supplied by firm X) has 5 chances in 100 of being included in a quality test. This conclusion may reflect the facts that (a) 5 percent of the firm’s total output meets the above criteria and (b) the portion of output to be tested is being selected from total output by a random process. Thus, it is likely to reflect the characteristics of total output. As a matter of fact, deductive reasoning and inductive reasoning complement one another. As we will see, before statisticians can safely generalize from the part to the whole, they must study how the part has been generated from the whole.

**LOOKING AHEAD**

The importance of inferential statistics is illustrated by the rich array of examples found throughout this book and its supplements. Statistical techniques help firms screen job applicants, budget research and development expenditures, determine the quality of raw materials received or of output produced, and decide whether sales personnel are better motivated by salary or commission. Statistical techniques can, similarly, help firms choose the best one among several product designs, leasing arrangements, oil-drilling sites, fertilizer types, or advertising media. And inferential statistics can tell firms precisely how the quantity of their product that consumers demand relates to the product’s price, to the prices of substitutes and complements, to consumer income, and, perhaps, even to the consumer’s sex.
Government officials are equally avid users of what this text has to teach. Inferential statistics plays an important role in ensuring the reliability of space missions and of more mundane airport lighting systems. And statistical techniques help answer questions such as these: Do motorcycle helmets really reduce accident fatalities? Do nursing homes discriminate against Medicaid recipients? Do the boxes of raisins marketed by this firm truly contain 15 ounces as claimed? Do the firms in this state meet antipollution standards? When is this recession likely to end? What is next year’s probable rate of inflation? This list, too, can be expanded at will.

Here, chapter by chapter, are some specific problems you will encounter and learn to solve. Naturally, the list contains terms you haven’t yet met. Don’t fret. You will learn about them in due course.

CHAPTER 8 | The Theory of Probability

A wine producer has designed a distinctive bottle in the hope of increasing sales. The manager views the probability of success as 50 percent, based on experience, but also orders a survey of customers. The manager knows that when consumers are enthusiastic and sales are about to rise, the type of survey about to be taken will confirm that positive market climate 90 percent of the time. But in 10 percent of the cases, the survey will say the opposite. When consumers are unimpressed and sales prospects look dim, the survey will so indicate 60 percent of the time. But in 40 percent of the cases, it will then say the opposite. The survey is taken and shows great consumer enthusiasm about the bottle. What is the manager’s new assessment of the probability of success?

CHAPTER 9 | Discrete Probability Distributions

An automobile manufacturer has fitted all of the firm’s cars with an identical pollution-control device, designed to meet government standards. Yet experience shows that 5 percent of cars tested perform below these pollution standards. Assume that 20 cars coming off the assembly line during a given month are selected at random. What is the probability that a government inspector who tests 20 cars a month in the above fashion will unjustly accuse the manufacturer of producing more than 5 percent of all cars below standard?

CHAPTER 10 | Continuous Probability Distributions

The city miles-per-gallon (mpg) rating of cars is a normally distributed random variable with a mean of 25.9 and a standard deviation of 2.45. If an automobile manufacturer wants to build a car with an mpg rating that improves upon 99 percent of existing cars, what must the new car’s mpg rating be?

CHAPTER 11 | Sampling Distributions

The manufacturer of batteries for aircraft emergency-locator transmitters claims that the lifetime of these batteries is normally distributed with a mean of 30 months and a standard deviation of 3 months. An aircraft manufacturer checks out 50 batteries and discovers a sample mean of only 29 months. What is the probability that the battery manufacturer’s claim is true?
CHAPTER 12 | Estimation

United Parcel Service wants to determine the gasoline savings if all of its trucks were switched from regular to radial tires. Some 150 trucks get new tires. Half of them are regular tires; the others are radial tires. One truck in each group, furthermore, is matched with one in the other group by make, age, region of the world, and other aspects that might affect gasoline consumption. After 3 months, the mileage on trucks with radial tires is found, on average, to be 5 miles per gallon higher than on trucks with regular tires. The sample standard deviation of the differences is 3 mpg. Construct a 98 percent confidence interval for the potential mileage gain if a similar switchover were made on all of the firm’s trucks worldwide.

CHAPTER 13 | Hypothesis Testing: The Classical Technique

The Environmental Protection Agency is allowing a plant to dump its waste into a river—as long as the effluent averages no more than 4 parts per million (ppm) of a certain toxic substance. During the course of one week, the EPA randomly samples the effluent and finds, in 64 samples, an average of 4.2 ppm of toxic substance, with a standard deviation of 1 ppm. Is the plant violating EPA standards?

CHAPTER 14 | Hypothesis Testing: The Chi-Square Technique

An advertising agency wants to know whether consumer preferences for three brands of coffee are independent of the person’s gender. The answer will determine whether different ads must be created for men’s and women’s magazines. A simple random sample of 100 persons yields Table 1.2. So, can we say that people’s coffee preferences do not depend on their sex?

TABLE 1.2 | Preferences Indicated by Coffee Drinkers

<table>
<thead>
<tr>
<th>(A) Sex</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>18</td>
<td>25</td>
<td>17</td>
<td>60</td>
</tr>
<tr>
<td>Female</td>
<td>32</td>
<td>5</td>
<td>3</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>30</td>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>

CHAPTER 15 | Analysis of Variance

An analyst wants to test whether the average price per share differs among three stock exchanges, A–C. Independent random samples of eight stocks from each market yield the following (in dollars per share):

A: 45, 56, 82, 49, 53, 61, 48, 51
B: 17, 19, 27, 22, 31, 41, 15, 16
C: 30, 19, 82, 49, 31, 19, 16, 51

Perform the desired test.
A marketing manager wants to establish the relationship between the number of cereal boxes sold, \( Y \), and the shelf space devoted to them, \( X \). Given the data of Table 1.3, determine an appropriate equation relating \( Y \) to \( X \).

### TABLE 1.3 | Cereal Boxes Sold and Shelf Space

<table>
<thead>
<tr>
<th>Number of Boxes Sold, ( Y )</th>
<th>Feet of Shelf Space, ( X )</th>
<th>Number of Boxes Sold, ( Y )</th>
<th>Feet of Shelf Space, ( X )</th>
</tr>
</thead>
<tbody>
<tr>
<td>145</td>
<td>3</td>
<td>125</td>
<td>7</td>
</tr>
<tr>
<td>151</td>
<td>6</td>
<td>190</td>
<td>5</td>
</tr>
<tr>
<td>235</td>
<td>9</td>
<td>210</td>
<td>10</td>
</tr>
<tr>
<td>120</td>
<td>5</td>
<td>118</td>
<td>8</td>
</tr>
<tr>
<td>272</td>
<td>13</td>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>300</td>
<td>15</td>
<td>390</td>
<td>12</td>
</tr>
<tr>
<td>110</td>
<td>2</td>
<td>210</td>
<td>6</td>
</tr>
</tbody>
</table>

An executive of a shoe manufacturing company wants to assess the relationship between average daily sales at the firm’s factory outlet stores, \( Y \), and a number of possible determinants. These include the number of competitors within a 3-mile radius, \( X_1 \), per capita annual income in the county, \( X_2 \), and the average price per pair of shoes, \( X_3 \). Given the data of Table 1.4, compute a multiple regression equation.

A textbook publisher has collected the data of Table 1.5. Create an economic model that relates copies sold to the review ratio and the Web site availability.

The following time series represents U.S. corporate tax liabilities from 1959 to 1997 (in billions of dollars): 23.6; 22.7; 22.8; 24.0; 26.2; 28.0; 30.9; 33.7; 32.7; 39.4; 39.7; 34.4; 37.7; 41.9; 49.3; 51.8; 50.9; 64.2; 73.0; 83.5; 88.0; 84.8; 81.1; 63.1; 77.2; 94.0; 96.5; 106.5; 127.1; 137.0; 141.3; 140.5; 133.4; 143.0; 165.2; 186.6; 211.0; 226.1; 246.1.

a. Estimate the trend in the form of a 7-year moving-averages series.

b. On the basis of the slope of the trend line between the last two numbers estimated in (a), forecast corporate taxes in 2004.
The manager of an orchard has collected the data of Table 1.6. Compute a 1995 *Laspeyres quantity index* for the orchard’s output, based on 1985. Can you see any problem with interpreting the index number?

### TABLE 1.4 | Factory Outlet Stores Data

<table>
<thead>
<tr>
<th>Average Daily Sales ($1,000s), $Y$</th>
<th>Competitors within 3 Miles (number), $X_1$</th>
<th>Per Capita Income in County ($1,000/year), $X_2$</th>
<th>Average Price of Shoes ($/pair), $X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.8</td>
<td>0</td>
<td>3.1</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>0</td>
<td>4.0</td>
<td>31</td>
</tr>
<tr>
<td>.6</td>
<td>4</td>
<td>2.9</td>
<td>67</td>
</tr>
<tr>
<td>1.9</td>
<td>1</td>
<td>5.9</td>
<td>44</td>
</tr>
<tr>
<td>1.9</td>
<td>1</td>
<td>6.9</td>
<td>47</td>
</tr>
<tr>
<td>1.4</td>
<td>1</td>
<td>4.9</td>
<td>43</td>
</tr>
<tr>
<td>.6</td>
<td>6</td>
<td>2.7</td>
<td>78</td>
</tr>
<tr>
<td>2.5</td>
<td>3</td>
<td>4.7</td>
<td>34</td>
</tr>
<tr>
<td>1.7</td>
<td>1</td>
<td>4.2</td>
<td>23</td>
</tr>
<tr>
<td>2.9</td>
<td>2</td>
<td>6.0</td>
<td>33</td>
</tr>
<tr>
<td>4.8</td>
<td>2</td>
<td>7.3</td>
<td>55</td>
</tr>
<tr>
<td>3.9</td>
<td>3</td>
<td>7.2</td>
<td>66</td>
</tr>
<tr>
<td>5.1</td>
<td>0</td>
<td>8.0</td>
<td>43</td>
</tr>
<tr>
<td>8.6</td>
<td>0</td>
<td>14.7</td>
<td>83</td>
</tr>
</tbody>
</table>

### TABLE 1.5 | Sample Data on Textbook Sales

<table>
<thead>
<tr>
<th>Copies Sold (1,000s/year), $Y$</th>
<th>Favorable-to-Unfavorable Review Ratio, $X_1$</th>
<th>Web Site ($\equiv 0$ if unavailable; $= 1$ if available), $D_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>5.1</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>4.6</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>6.0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>2.0</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>3.1</td>
<td>1</td>
</tr>
<tr>
<td>39</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>52</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>10.3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5.0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>12.0</td>
<td>0</td>
</tr>
</tbody>
</table>
A business school admissions board stands accused of manipulating admissions according to a secret daily quota system based on applicants’ sex. The board denies the charge and claims that sex is never even considered during the admission process. Therefore, the order of male/female admissions must be random. Investigators acquire the information given in Table 1.7 for 60 successive admissions. (M = male, F = female, and data are to be read in successive rows.) Make an appropriate hypothesis test and decide whether the board discriminates on the basis of sex.

<table>
<thead>
<tr>
<th>TABLE 1.7</th>
<th>Business School Admissions Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Order of Admission</strong></td>
<td><strong>Sex of Admission</strong></td>
</tr>
<tr>
<td>1–10</td>
<td>M M M M M F F M M M</td>
</tr>
<tr>
<td>11–20</td>
<td>M M F F F M M M M M</td>
</tr>
<tr>
<td>21–30</td>
<td>F F M F M M M M M F</td>
</tr>
<tr>
<td>31–40</td>
<td>F F M M M M M M M F</td>
</tr>
<tr>
<td>41–50</td>
<td>F F M F M M M M M M</td>
</tr>
<tr>
<td>51–60</td>
<td>M M F F M M F F M M</td>
</tr>
</tbody>
</table>

A production process is designed to fill bottles with 16.5 ounces of liquid detergent on the average. The population of filling weights is normally distributed and has a standard deviation of .8 ounce. Inspectors take periodic samples of 35 bottles. One sample yields a mean filling weight of 16.2 ounces, the next two yield 15.8 ounces and 17.3 ounces. Is the production process running properly?

A U.S. firm plans to enter a new market in China. The firm’s executives are considering four alternatives: (1) building a plant in China, (2) hiring a Chinese sales force to sell U.S.-made products exported to China, (3) sending mail-order catalogues to Chinese consumers, and (4) teaming
up with Chinese firms that would act as sales agents. The executives believe that the profit consequences of each of these approaches will differ, depending on whether demand turns out to be low, moderate, or huge. In millions of dollars, the next year profit predictions are 2, 4, and 10 for (1); 3, 3, and 3 for (2); −5, −1, and +20 for (3); and −2, 0, and 30 for (4). Assuming they want to maximize profits, what should the executives do?

1.5 Statistics—A Universal Guide to the Unknown

All of the previous examples have at least one thing in common. They illustrate dramatically how inferential statistics can facilitate decision making in the face of uncertainty. Indeed, one can argue that such is the main purpose of the entire discipline.

DEFINITION 1.5 The field of study known as statistics is a branch of mathematics that is concerned with facilitating wise decision making in the face of uncertainty and that, therefore, develops and uses techniques for the careful collection, effective presentation, and proper analysis of numerical information.

This definition clearly incorporates all of the branches of statistics discussed so far. In addition, by failing to specify who the decision makers are, the above definition quite correctly suggests the universal applicability of what statistics has to offer. As the earlier examples have shown, modern statistical techniques routinely guide business executives, as well as government economic-policy makers, in making reasonable decisions in the face of uncertainty. In addition, the same techniques are just as useful, and are just as frequently applied, outside the fields of business and economics. Like mathematics in general, statistics is a universal type of language that all sciences use regularly. Drawing valid inferences from limited information is just as important to historians and psychologists, to geneticists and medical researchers, to astronomers and engineers as it is to business executives and economists. We shall see examples of this universal use of statistics as well. Among others, questions such as these will be posed and answered: How can we decipher a secret code? How can we resolve the issue of a disputed authorship? Were Mendel’s genetics data fudged? Does smoking cause heart disease? Is toothpaste A better than toothpaste B? Is ESP real?

1.6 Basic Statistical Concepts

As is true of practitioners in all scientific disciplines, statisticians have a language all of their own. In this section, we meet some of their favorite terms. This will prove helpful in later chapters when we discuss issues of collecting, describing, and analyzing data more fully.

ELEMENTARY UNITS AND THE FRAME

A statistical investigation invariably focuses on people or things with characteristics in which someone is interested. The persons or objects that have characteristics of interest to statisticians are called elementary units. Thus, someone who wanted to learn about the racial composition of a firm’s labor force would quickly identify the individual employees of that firm as the elementary units. But someone concerned about the amount of credit extended by that firm might view individual credit accounts as the elementary units to be investigated. Even the flashcubes
produced by the firm, the lightbulbs installed in its plants, or the boxes of cereal shipped by one of its divisions could be regarded as elementary units—provided someone was interested in discovering, respectively, the percentage of defective flashcubes produced, the lifetimes of lightbulbs used, or the content weights of cereal boxes sold. A complete listing of all elementary units relevant to a statistical investigation is called a frame.

Consider a statistician who is hired to evaluate charges of racial and sex discrimination allegedly occurring in one of those small firms that can be found at any one of thousands of private airports across the United States. (Believe it or not, in 1999 there were over 18,000 operating airports in the United States, and over 13,000 of them were private.) Our airport operator’s personnel records might provide the information listed in Table 1.8. In this case, any one entry in column 1 is an elementary unit. All the entries in the shaded portion of that column jointly represent the frame.

VARIABLES AND DATA

In general, any one elementary unit may possess one or more characteristics that interest a statistician. In Table 1.8, five characteristics, ranging from race and sex to annual salary, are listed in the headings of columns 2–6. Such characteristics of elementary units are called variables, presumably because observations about these characteristics will likely vary from one elementary unit to the next.

Any single observation about a specified characteristic of interest is called a datum. It is the basic unit of the statistician’s raw material. Any collection of observations about one or more characteristics of interest, for one or more elementary units, is called a data set. A data set is univariate, bivariate, or multivariate depending on whether it contains information on one variable only, on two variables, or on more than two. The 45 entries in columns 2–6 of Table 1.8 (namely, 5 data for each of 9 elementary units), thus, constitute a multivariate data set.

QUALITATIVE AND QUANTITATIVE VARIABLES

Table 1.8 teaches us something else: Any given characteristic of interest to the statistician can differ in kind or in degree among various elementary units. A variable that is normally described in words rather than numerically (because it differs in kind rather than degree among elementary units) is called a qualitative variable. Table 1.8 contains three qualitative variables: race, sex, and job title. Qualitative variables can, in turn, be binomial or multinomial. Observations about a binomial qualitative variable can be made in only two categories: for example, male or female, employed or unemployed, correct or incorrect, defective or satisfactory, elected or defeated, absent or present. Observations about a multinomial qualitative variable, in contrast, can be made in more than two categories; consider job titles, colors, languages, religions, or types of businesses.

On the other hand, a variable that is normally expressed numerically (because it differs in degree rather than kind among the elementary units under study) is called a quantitative variable. Table 1.8 contains two of them: years of service and annual salary. Quantitative variables can, in turn, be discrete or continuous. Observations about a discrete quantitative variable can assume values only at specific points on a scale of values, with inevitable gaps between them. Such data differ from each other by clearly defined steps. Consider observing the number of children in families, of employees in firms, of students in classes, of rooms in houses, of cars in stock, of cows in pastures. Invariably, the individual data will be disconnected from each other by gaps on the scale of values. In the above instances, they will look like 1, 2, 3, . . . and 49; never like 3.28 or 20.13. It is impossible to have 3.28 children in a family or to observe 20.13 cows in a pasture because these items come in whole units only. But note that the gaps representing impossible values need not span the entire space between whole numbers. Stock prices, for example, are reported in
TABLE 1.8  Selected Characteristics of All Full-Time Employees of Mountain Aviation, Inc.; December 31, 2000

This table illustrates a number of basic statistical concepts. Thus, column 1 lists nine elementary units that jointly constitute the frame (shaded). The headings of columns 2–6 show various characteristics of the elementary units that are called variables. They can be qualitative (race, sex, job title) or quantitative (years of service, annual salary). Any single observation about a given elementary unit is a datum (the plural is data). This particular table contains a multivariate data set because it records observations about several variables for each elementary unit. All possible observations about a given variable, such as the shaded entries in column 3 or 6, constitute a statistical population. Any subset of a population or frame, such as the boxed data in column 6, is a sample.

<table>
<thead>
<tr>
<th>List of Employees (1)</th>
<th>Race (2)</th>
<th>Sex (3)</th>
<th>Job Title (4)</th>
<th>Years of Service (5)</th>
<th>Annual Salary (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abel</td>
<td>White</td>
<td>Male</td>
<td>Pilot</td>
<td>2</td>
<td>$39,000</td>
</tr>
<tr>
<td>Cruz</td>
<td>White</td>
<td>Male</td>
<td>Chief Mechanic</td>
<td>10</td>
<td>70,000</td>
</tr>
<tr>
<td>Dunn</td>
<td>Black</td>
<td>Male</td>
<td>Chief Pilot</td>
<td>23</td>
<td>85,000</td>
</tr>
<tr>
<td>Hill</td>
<td>Black</td>
<td>Female</td>
<td>Secretary</td>
<td>5</td>
<td>17,000</td>
</tr>
<tr>
<td>King</td>
<td>White</td>
<td>Male</td>
<td>Janitor</td>
<td>8</td>
<td>21,000</td>
</tr>
<tr>
<td>Otis</td>
<td>White</td>
<td>Male</td>
<td>Grounds Keeper</td>
<td>10</td>
<td>21,000</td>
</tr>
<tr>
<td>West</td>
<td>Black</td>
<td>Male</td>
<td>Mechanic</td>
<td>2</td>
<td>36,000</td>
</tr>
<tr>
<td>Wolf</td>
<td>White</td>
<td>Female</td>
<td>Pilot</td>
<td>7</td>
<td>36,000</td>
</tr>
<tr>
<td>Zorn</td>
<td>White</td>
<td>Female</td>
<td>Mechanic</td>
<td>7</td>
<td>40,000</td>
</tr>
</tbody>
</table>

Frame (containing 9 elementary units) | Population of employee sexes (containing 9 data) | Sample of employee salaries (containing 3 data)

NOTE: As you will learn in Chapter 4, many types of samples exist. Not all of them are equally likely to reflect the makeup of the sampled population. For example, the group of three salaries found in the column 6 box here might be a convenience sample that was selected merely for the ease of illustration. It might also be a simple random sample that was selected by some procedure such as writing the nine salaries on slips of paper, mixing the slips in a bowl, and pulling out three. In the latter case, as you will learn in Chapter 8, it is possible to select 84 different samples of 3 out of 9. That would give us 1 chance in 84 of selecting the particular sample shown here.

eighths of dollars (or to the nearest $0.125). These discrete figures can take on values of 67 1/8, 67 2/8, 67 3/8 (or equivalent dollar decimals per share) but cannot take on values between these. (The quoting of prices by eighths is a throwback to the old pirate days and the Spanish gold “pieces of eight.”)
Observations about a **continuous quantitative variable** can, in contrast, assume values at all points on a scale of values, with no breaks between possible values. Consider height, temperature, time, volume, or weight. Weight, for instance, might be reported as 7 pounds or 8 pounds but also as 7.3 pounds or even 7.3425 pounds, depending entirely on the sensitivity of the measuring instrument involved. No matter how close two values are to each other, it is always possible for a more precise device to find another value between them.

**CAUTION**

The distinction between qualitative and quantitative variables is visually obvious in Table 1.8. The observations about one type of variable are recorded in words; those about the other type in numbers. Yet that distinction can easily be blurred. Quantitative variables can be converted into seemingly qualitative variables, and the opposite is also true. Thus, a statistician who could replace the column 5 and 6 numerical data by words, such as *low*, *intermediate*, or *high*, although probably nobody would wish to give up the more precise information recorded in Table 1.8. On the other hand, it is common practice to code observations about qualitative variables with the help of numbers. Thus, a statistician might turn the verbal entries of Table 1.8 into numbers by recording, say, “white” as 1 and “black” as 2 in column 2, by recording “male” as 0 and “female” as 1 in column 3, and by assigning numbers between 0 and 6 to the seven job titles in column 4. Nevertheless, the distinction between qualitative and quantitative variables, although then hidden, would remain.

Being aware of the distinction is important for this reason: Even when qualitative data are encoded numerically, we cannot perform meaningful arithmetic operations with them, whereas we can do so with quantitative data. Thus, it would make no sense to report the “sum of races” in our firm as 12 (using the code just noted), but it would be valid to report the sum of annual salaries as $365,000.

**POPULATION VERSUS SAMPLE**

We must finally consider two other concepts of particular importance.

**DEFINITION 1.6** The set of all possible observations about a specified characteristic of interest is called a statistical **population**.

**DEFINITION 1.7** A subset of a statistical population, or of the frame from which it is derived, is called a **sample**.

As Table 1.8 illustrates, it is possible to draw several statistical populations from a given frame. We have one frame (the shaded list of elementary units in column 1), but five variables (the headings of columns 2–6). Hence our table contains five populations. The shaded entries in column 3, for example, make up the population of employee sexes; those in column 6 make up the population of employee salaries; and so on for columns 2, 4, and 5.

Note that a statistical population consists of all possible observations about a variable. Because they correspond to all the employees of our firm, the entries in column 6 make up the population of employee salaries in our hypothetical study. In a firm of only nine employees, it is easy to observe the entire salary population. But imagine the difficulty of such an undertaking if we attempted to carry out a similar study for the entire aviation industry or even for the entire labor force of the United States!

Under such circumstances, statisticians often make observations concerning selected elementary units only. They observe only \( n \) such units out of the larger number \( N \) that exist. Naturally, they end up with a subset of all the possible observations about the characteristic of interest, which is then called a **sample**. The boxed entries in column 6 of Table 1.8, for instance, make up one of many possible samples of employee salaries—namely, the sample...
based on observing the salary characteristics of only Otis, West, and Wolf. These three names themselves can, in turn, be viewed as a sample of the frame.

**CAUTION**

What constitutes a population or a sample of that population depends entirely on the context in which the question arises. If the goal were to study salaries only at Mountain Aviation, Inc., the data in shaded column 6 of Table 1.8 would, as a group, make up the relevant population. If the goal were to study salaries in the entire aviation industry, however, the identical column 6 data, even as a group, would constitute only a (pitifully small and probably not very representative) sample of the much larger population of salaries in the industry as a whole.

In addition, as you will learn in Chapter 4, it is important to note that several types of samples exist and not all of them provide information of equal quality. The particular sample illustrated in Table 1.8 may well be a *convenience sample*, selected because the three salaries in question just happened to be located next to one another. In fact, however, it would be possible to select 84 different samples of size $n = 3$ from among the $N = 9$ salaries listed in Table 1.8. Thus, if we had used a random process to select a *random sample* of three salaries in Table 1.8, we would have had only 1 chance in 84 to select the particular sample shown there. (More about *that* in Chapter 8!)

### 1.7 Major Types of Data

No matter how hard they try to do a good job, data gatherers will always come up with data of varying quality. This is so because different data sets are *inherently* different, as another look at Table 1.8 can quickly show: Observations about *qualitative* variables (columns 2–4) are typically made in words but are possibly coded into numbers later on for purposes of data processing. Observations about *quantitative* variables, in contrast, are numerical at the outset (columns 5–6). Anyone who works with numbers, therefore, must be very clear about their precise meaning.

Consider the numbers in the series 1, 2, 3, ... 10, 11, 12. They could be house numbers found along a street. They could be numbers on the Beaufort scale, measuring wind velocity. They could be numbers on the Fahrenheit scale, indicating temperature. They could be hourly wages paid different workers in a factory. Believe it or not, despite the fact that we are looking at the same numbers, 1 through 12, these four examples point to radically different *types* of data.

We must discuss these data types before we turn, in Chapter 2, to the task of entering data into a computer. Not every data type is suitable for the arithmetic operations that computers can perform so rapidly and well.

In fact, the assignment of numbers to characteristics that are being observed—which is *measurement*—can yield any one of four types of data. In order of increasing sophistication, it can produce *nominal*, *ordinal*, *interval*, or *ratio* data, and different statistical concepts and techniques are appropriately applied to each type.

**Nominal Data**

Suppose you were working, as we will later in this text, with an alphabetic list of the 110 largest multinational firms that maintained headquarters in the United States. (Table 4.1 on page 110 contains such a list.) Continually referring to the actual company names, such as Goodyear Tire & Rubber or Minnesota Mining & Manufacturing, may soon become awkward and unwieldy. So you decide to substitute *numbers* for those company names, ranging from 00 for Abbott Laboratories to 99 for Xerox. These numbers are *nominal data*. They merely *name* or label differences in kind. Thus, they serve the purpose of classifying observations about qualitative variables into mutually exclusive groups where the numbers in each group can then be counted. (Numbers
between 00 and 99, for example, might refer to multinational companies, numbers between 100 and 159 to other types of firms, and so on.)

In fact, we meet nominal data every day. House numbers provide a good example. The green house at the corner might be assigned the number 1, the yellow house across the street a 2, the white house in the middle of the block a 6, and so on, until the brick house at the end is labeled with a 12. Similarly, a statistician working with Table 1.8 presented earlier might code “male” as 0 and “female” as 1 for the sake of mere convenience, but alternative labels of “male” = 100 and “female” = 50 would serve as well.

Invariably, nominal data provide the weakest level of measurement in the sense that they contain only the tiniest amount of useful information. More importantly, as the slightest bit of thought about these examples can confirm, it never makes sense to add, subtract, multiply, divide, rank, average, or otherwise manipulate nominal data arithmetically. We can merely count them. The presence of 12 numbers on a street denotes the existence of 12 houses. Five 1’s, according to one of the above codes, indicates the presence of five females. And that is all.

Consider how adding all the house numbers on our street would yield a meaningless number 78. Summing six 0’s and three 1’s to a total of 3 (because, say, six men and three women are working in a firm) would be equally silly. Ordering nominal numbers by size, or ranking them, would be senseless as well. Although 2 is smaller than 6, in what sense is the yellow house numbered 2 smaller than the white house numbered 6? Although 1 is greater than 0, in what sense is “female” greater than “male”? Nor could we assume that equal differences or intervals between nominal data carry any meaning at all: Just because 12 – 10 = 2 and 10 – 8 = 2 as well, could we assume that the distance between house #12 and house #10 is the same as that between house #10 and house #8? Hardly. And dividing one house number by another would be pointless, too. True enough, the ratio of 12/6 is 2, but can we say that house #12 is somehow twice as large or otherwise more important than house #6 down the street?

**ORDINAL DATA**

The next level of measurement produces ordinal data. These are numbers that label differences in kind, as nominal data do, but that, in addition, by their very size also order or rank observations on the basis of importance. Consider another list of those 100 multinational companies, but this time let it not be alphabetical. Let the companies be ranked from the one with the smallest profit, labeled 00, to the one with the highest profit, labeled 99. We can compare such ordinal numbers meaningfully as greater than, smaller than, or equal to one another. But they contain no information about how much greater or smaller one labeled item is compared to the other. Thus, company 99 has a larger profit than company 63, and company 63 has a smaller profit than company 69, but that is all we can say. Differences between ordinal numbers or ratios of such numbers remain meaningless. Just because 69 – 63 = 6 and 17 – 11 = 6 as well, can we conclude that the profit difference between companies 69 and 63 comes to the same dollar figure as that between companies 17 and 11? Certainly not. Just because (80/20) = 4, can we conclude that the profit of company 80 is four times as large as that of company 20? We cannot.

Just like nominal data, we encounter ordinal data often in our daily lives. The Beaufort wind scale used by weather forecasters provides a good example. It codes a wind velocity of less than 1 mile per hour (mph) as 0 or “calm,” a velocity of 1–3 mph as 1 or “light air,” a velocity of 4–7 mph as 2 or “light breeze,” and so on, until velocities above 75 mph are reached and labeled 12 or “hurricane.” Clearly, the order of numbers matters here. The larger the number, the stronger the wind. Yet these make no statement about how much stronger or weaker the assessment becomes as we move along the scale in one direction or another. It may look as if moving along the Beaufort scale from 0 to 12 in equal steps of 1 represents equal increases in velocity, but such is not the case. Codes 1 and 2 represent average wind speeds of 2 and 5.5 mph, respectively. Thus the difference between “light air = 1” and “light breeze = 2” comes to 3.5 mph. Yet codes 10
and 11 represent average wind speeds of 59 and 69.5 mph, respectively. Thus the difference between “whole gale = 10” and “storm = 11” is 10.5 mph.

Similar examples abound. Assessments of a product as superb, average, or poor might be recorded as 2, 1, 0, as 250, 10, 2, or even as 10, 9, 4.5—the important thing is that larger ordinal numbers denote a more favorable assessment, or a higher ranking, while smaller ones do the opposite. Yet, in such an assessment, a 2 is deemed better than a 1 but not necessarily twice as good. A 250 is deemed better than a 10 but not necessarily 25 times as good. A 4.5 is deemed worse than a 9 but not necessarily half as good. And that is all.

Once again, meaningful arithmetic operations with ordinal data, as with nominal data, are not possible.

**INTERVAL DATA**

Somewhat more information is contained in interval data. These are numbers that possess all the characteristics of ordinal data and, in addition, relate to one another by meaningful intervals or distances. This is so because all numbers are referenced to a common (although admittedly arbitrary) zero point. As a result, addition and subtraction are permissible, but multiplication and division continue to make no sense.

Certain scales of calendar time, clock time, and temperatures provide good examples of measurements that start from an arbitrarily located zero point and then use an equally arbitrary but consistent distance unit for expressing intervals between numbers. Consider how the Celsius scale places zero at the water-freezing point, whereas the Fahrenheit scale places it far below the freezing point. Within the context of either scale, the distance unit (degree of temperature) has a consistent meaning. Each degree Celsius equals 1/100 of the distance between water’s freezing and boiling points. Each degree Fahrenheit equals 1/180 of that distance. However, the zero point, being arbitrarily located, does not denote the absence of the characteristic being measured. Unlike 0°F on the absolute (Kelvin) temperature scale that is familiar to scientists, neither 0°F nor 0°C indicates a complete absence of heat. As a result, any ratio of Fahrenheit or Celsius data fails to convey meaningful information. For example, 90°F is not twice as hot as 45°F. Indeed, the ratio of the corresponding Celsius figures (32.2° and 7.2°) does not equal 2:1 but well over 4:1.

**RATIO DATA**

The highest level of measurement, producing the most useful information, yields ratio data. These are numbers that possess all the characteristics of interval data and, in addition, have meaningful ratios because they are referenced to an absolute or natural zero point that denotes the complete absence of the characteristic being measured. All types of arithmetic operations, even multiplication and division, can be performed with such data. Unlike in the Fahrenheit/Celsius example, the ratio of any two such numbers is independent of the unit of measurement because each number is a distance measure from the same zero point. For example, the measurement of hourly wages, monthly salaries, age, area, distance, height, volume, or weight produces ratio data. So does a measurement of temperature on the Kelvin scale, the zero point of which is tied to zero molecular speed. Consider hourly wages as a case in point.

Clearly, it makes sense to say that an hourly wage of $12 is larger than one of $9, which is larger than one of $6. (In contrast, when working with those nominal data taking the place of alphabetically ordered company names, it made no sense to say that company 12 is larger than company 9, which is larger than company 6.) Thus, hourly wage data give the kind of information provided by ordinal data.

In addition, it makes sense to compare intervals between hourly wage data and to say that the distance between $12 and $9 equals the distance between $9 and $6. (In contrast, the differ-
A perfect example of the birth of *interval data* is provided by the worldwide effort to replace the familiar 24-hour clock with a new type of *universal* time. Such adjustments of our way of measuring time are nothing new. In the United States, for example, different cities used to set their clocks by the sun until the late 1800s. Then the railroads came along and introduced the now-familiar time zones in order to coordinate their schedules and avoid collisions. An 1874 treaty established Greenwich mean time, which later served lots of people very well, including ships and airlines. In 1960, scientists in Paris set up an atomic clock, based on vibrations of the cesium atom. And now there is the Swatch Group, the world’s largest watchmaker, which wants us to live on *Internet time*.

Figure 1.A illustrates what is involved.

The new time starts at an arbitrary zero point and beats 000 at midnight over the Swatch building in Biel, Switzerland. It also divides the day into 1,000 *swatch beats*, each equivalent to 86.4 good old seconds. So, if it’s 3 P.M. local time, or 15:00 hours, by the old clock, you are at 625 universal time, as the display shows.

As you might have guessed, Swatch is selling an Internet watch around the world ($70 at the time of this writing) and you can even download software at its site to teach your computer a trick or two about the meaning of Internet time.

---

**APPLICATION 1.1**

**TIME IN CYBERSPACE**

---

Qualitative variables are usually described verbally. When coded, these verbal descriptions turn into numbers that are nominal or ordinal data. Arithmetic operations with such data—including adding, subtracting, multiplying, dividing, or averaging—yield pure nonsense. In contrast, quantitative variables are always described numerically, either by interval or ratio data. Interval data allow some types of arithmetic operations; ratio data allow all types.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Associated Data Type</th>
<th>Description of Data</th>
<th>Permitted Operations</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative</td>
<td>Nominal</td>
<td>Numbers that merely name or label differences in kind; their order has no particular meaning, nor do their differences or ratios</td>
<td>Placing numbers into mutually exclusive groups; counting numbers in each group</td>
<td>The two-digit codes from 00 to 99 for an alphabetical list of 100 most admired firms; highway or telephone numbers</td>
</tr>
<tr>
<td>Qualitative</td>
<td>Ordinal</td>
<td>Numbers that not only label differences in kind, but that by their size also order or rank observations on the basis of importance, while differences between numbers or ratios of such numbers are meaningless</td>
<td>As above plus comparing numbers as greater than, smaller than, or equal to one another</td>
<td>The ranking of college courses as 3 for excellent, 2 for good, and 1 for poor; student class ranks; the ranking of runners at the finish line</td>
</tr>
<tr>
<td>Quantitative</td>
<td>Interval</td>
<td>Numbers that not only label differences in kind and by their size rank observations, but that are also referenced to a common (but arbitrary) zero point, which makes their intervals or differences comparable, while their ratios are meaningless due to the arbitrariness of the zero point</td>
<td>As above plus adding and subtracting</td>
<td>Fahrenheit and Celsius temperature scales; clock time; calendar time</td>
</tr>
<tr>
<td>Quantitative</td>
<td>Ratio</td>
<td>Numbers that not only label differences in kind, by their size rank observations, and are separated by meaningful intervals, but that are also referenced to an absolute or natural zero point, which makes their ratios meaningful as well</td>
<td>As above plus multiplying and dividing</td>
<td>The salary data listed in Table 1.8; measures of length, volume, weight</td>
</tr>
</tbody>
</table>
The fact that ratios of numbers convey meaningful information is the advantage of ratio data over interval data. No wonder that statisticians, when they have a choice, prefer ratio data to interval data, interval data to ordinal data, and ordinal data to nominal data. Table 1.9 summarizes our discussion of data types for those who seek a quick review.

1. The term *statistics* has at least three different meanings to people. Some think of it as a *field of study* that somehow deals with the collection, presentation, and interpretation of numerical data. For others, the term conjures up images of *masses of data*, seemingly infinite in number. Still others attribute a highly technical meaning to the term, thinking of *summary measures*—such as sample averages and sample proportions—that have been computed from relatively few data gathered by sampling a much larger collection of data.

2. Masses of data are, indeed, the statistician’s raw material, and a first branch of the discipline of statistics focuses on the careful collection of data. Such collection can proceed in one of three ways:
   a. An investigator can look for data that already exist because others have gathered them in the past.
   b. Brand-new data can be generated with the help of *observational studies* that involve census taking or sampling.
   c. Brand-new data can be generated by conducting carefully controlled experiments.

3. A second branch of the discipline of statistics, known as *descriptive statistics*, is concerned with developing and using techniques for the effective presentation of numerical information so as to highlight patterns otherwise hidden in a data set.

4. A third branch of the discipline of statistics, undoubtedly the most important one, is concerned with developing and using techniques for properly analyzing (or drawing inferences from) numerical information and is therefore called *analytical statistics* or *inferential statistics*.

5. In the end, the discipline of statistics is, perhaps, best viewed as a branch of mathematics that develops and uses techniques for the careful collection, effective presentation, and proper analysis of numerical information. As such it facilitates wise decision making in the face of uncertainty and becomes a universal guide to the unknown.

6. The process of data collection, a crucial prerequisite for subsequent descriptive and analytical work, employs a number of basic concepts. Thus, a statistical investigation focuses on persons or objects, which are called *elementary units*. These elementary units possess characteristics of interest, called *variables*. Observations about them, which can be qualitative or quantitative, are called *data*. The set of all possible observations about a specified characteristic of interest is called a *population*. A subset of it (or of the *frame* from which the population is derived) is referred to as a *sample*.

7. The assignment of numbers to characteristics that are being observed, which is *measurement*, can yield any one of four types of data. In order of increasing sophistication, it can produce *nominal*, *ordinal*, *interval*, or *ratio* data. Different statistical concepts and techniques are appropriately applied to each type. Arithmetic operations with nominal and ordinal data, for example, are out of the question.

---

**Key Terms**

- analytical statistics
- binomial qualitative variable
- bivariate data set
- continuous quantitative variable
- controlled experiment
- data set
data set
- datum
-deductive reasoning
-descriptive statistics
-discrete quantitative variable
-elementary units
-frame
-inductive reasoning
-inferential statistics
-interval data
-measurement
-multinomial qualitative variable
-multivariate data set
-nominal data
-observational study
-ordinal data
-population
-qualitative variable
-quantitative variable
-ratio data
-sample
-statistics
-survey
-univariate data set
-variables
Practice Problems

Section 1.1 Introduction

1. This is a fun question that elaborates on this chapter’s Preview. It challenges you to do some serious thinking, but you can also learn much by merely looking up the answers on the Student CD-ROM. First, study the solution that is given here to this chapter’s Preview problem; then consider the questions that follow.

If you forgo buying the consulting firm’s advice, the optimal action is to sell the rights for $125 million, which is illustrated in Figure 1.1.

If you do buy the consulting firm’s advice, the optimal action, now illustrated in Figure 1.2, is this: If $R_1$ is received, sell the rights and take the $125 million minus the $1 million fee. If $R_2$ is received, offer the film to the network and earn an expected $200 million.

a. Can you guess the meaning of the $102 million number at point $b$ in Figure 1.1?

b. Can you guess the meaning of the $154.4 million number at point $b$ in Figure 1.2?

c. Can you guess what strategy the filmmaker would be well advised to follow: forgoing the advice or buying the advice?

FIGURE 1.1 | The Soap Opera Decision without Advice

Note: The 60 percent chance of rejection and the 40 percent chance of a contract offer are indicated as probability of event $E_1$ being .6 and probability of event $E_2$ being .4.
CHAPTER 1 ■ THE NATURE OF STATISTICS

FIGURE 1.2 | The Soap Opera Decision with Advice

Note: You need not fully understand all of the entries in Figure 1.2 at this point.

Payoff (millions)

- $125 – 1 = $124
- $124
- $154.4
- $200
- $35

Get advice
Sell rights
Offer film to network

R1 = Rejection predicted
p(R1) = .6

E1 = Rejection
p(E1|R1) = .8

E1 = Rejection
p(E1|R2) = .3

R2 = Contract predicted
p(R2) = .4

E2 = Contract offer
p(E2|R1) = .2

E2 = Contract offer
p(E2|R2) = .7

$124
$35
$200
$200

$125 – 1 = $124
$300 – 1 = $299
$300 – 1 = $299

$125 – 1 = $124
$300 – 1 = $299
$300 – 1 = $299

$124
$35
$200
$200

$30 – 1 = $31
$30 – 1 = $31
$30 – 1 = $31

- $31

d. What do you think is the maximum amount the filmmaker could be made to pay for the (admittedly imperfect) advice?

SECTION 1.2 ■ THE COLLECTION OF DATA

2. This problem provides a preview of the type of material to be discussed at length in Chapter 3. If you are connected to the Internet, visit http://www.fedstats.gov, a site maintained by the Interagency Council on Statistical Policy. Click on Agencies and explore the manifold sources of U.S. federal government statistics. List five agencies that supply data to this site.

3. This problem provides a preview of the type of material to be discussed at length in Chapter 3. If you are connected to the Internet, visit http://www.statcan.ca, a site maintained by Statistics Canada. Click on English (unless you prefer French) > Canadian Statistics > The Economy: The Latest Indicators. Then find the latest monthly Canadian merchandise export figure.
4. This problem provides a preview of the type of material to be discussed at length in Chapter 3. If you are connected to the Internet, visit http://www.inegi.gob.mx, a site maintained by Mexico’s National Institute of Statistics, Geography, and Informatics.
   Click on English (unless you prefer Spanish) > Economy > Short Term Economic Indicators > Financial, Stock Market, and Monetary Indicators. Then find the latest monthly yield figure for commercial paper.

5. This problem provides a preview of the type of material to be discussed at length in Chapter 4. If you are connected to the Internet, visit http://www.gallup.com, a site maintained by the Gallup Organization. Check it out; then write an essay on what Gallup tells you about its sampling techniques.

6. This problem provides a preview of the type of material to be discussed at length in Chapter 4. If you are connected to the Internet, visit http://www.roper.com, a site maintained by the Roper Organization. Check it out; then write an essay on what Roper tells you about its sampling techniques.

7. This problem provides a preview of the type of material to be discussed at length in Chapter 4. If you are connected to the Internet, visit http://www.louisharris.com, a site maintained by the Harris Organization. Check it out; then write an essay on the latest monthly Harris poll.

8. Section 1.2 briefly anticipates issues that will be discussed at length in later chapters. One of these issues is the difference between surveys and experiments and the significance of exercising control over elementary units whose characteristics are being scrutinized. Imagine annual salaries of workers in a firm to equal $10,000 for everyone, plus $1,000 for every year of work experience. Salaries are, thus, totally unrelated to race. Then imagine that most of a firm’s black workers are young (and, therefore, have had little work experience), while most of its white workers are older (and have had many years of experience on the job). Someone merely surveying salaries might find an average salary of $15,000 a year among blacks and of $28,000 a year among whites and might conclude, quite incorrectly, that the firm’s management is discriminating based on race.
   In contrast, a controlled study would divide the firm’s workers into groups according to work experience and would compare salaries within each group. Such a study would find identical salaries between (1) the few white and (2) the many black workers in the younger and less experienced group. And it would find identical but higher salaries between (1) the many white and (2) the few black workers in the older and more experienced group. Thus, the controlled study would avoid the false racial discrimination charge.
   Make up a detailed numerical example to corroborate the story told by the numbers given here.

Section 1.6 Basic Statistical Concepts

9. Consider Table 1.10, which contains selected data found in Fortune magazine’s 1999 Global 500 report.
   a. Identify the elementary units.
   b. How many variables can you find in this table? Which are they?
   c. Identify the variables as quantitative or qualitative.
   d. Identify variables as discrete/continuous or binomial/multinomial.
   e. What kind of data set does this table contain?

10. In 1999, Fortune magazine surveyed more than 10,000 executives, directors, and securities analysts to rank corporate reputations on the basis of eight criteria, including innovativeness, quality of management, employee talent, quality of products/services, long-term investment value, financial soundness, social responsibility, and use of corporate assets. On the basis of their answers, it created something like a report card that ranked U.S. corporations from “most admired” = 1 to “least admired” = 469. Table 1.11 provides information about the top ten.
   a. Identify the elementary units.
   b. How many variables can you find in this table? Which are they?
   c. Identify the variables as quantitative or qualitative.
   d. Identify variables as discrete/continuous or binomial/multinomial.
   e. What kind of data set does this table contain?

11. In 1999, Fortune magazine surveyed more than 10,000 executives, directors, and securities analysts to rank corporate reputations. On the basis of their answers, it created something like a report card
TABLE 1.10 | Selected Characteristics of the World’s Largest Corporations in 1999

<table>
<thead>
<tr>
<th>Company</th>
<th>Country</th>
<th>Industry</th>
<th>Revenues ($ billion)</th>
<th>Profits ($ million)</th>
<th>Employees (thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. General Motors</td>
<td>U.S.</td>
<td>Autos</td>
<td>161.3</td>
<td>2,956</td>
<td>594</td>
</tr>
<tr>
<td>2. DaimlerChrysler</td>
<td>Germany</td>
<td>Autos</td>
<td>154.6</td>
<td>5,656</td>
<td>442</td>
</tr>
<tr>
<td>3. Ford Motor</td>
<td>U.S.</td>
<td>Autos</td>
<td>144.4</td>
<td>22,071</td>
<td>345</td>
</tr>
<tr>
<td>4. Wal-Mart Stores</td>
<td>U.S.</td>
<td>Retail</td>
<td>139.2</td>
<td>4,430</td>
<td>910</td>
</tr>
<tr>
<td>5. Mitsui</td>
<td>Japan</td>
<td>Trading</td>
<td>109.4</td>
<td>233</td>
<td>33</td>
</tr>
<tr>
<td>6. Itochu</td>
<td>Japan</td>
<td>Trading</td>
<td>108.7</td>
<td>−267</td>
<td>6</td>
</tr>
<tr>
<td>7. Mitsubishi</td>
<td>Japan</td>
<td>Trading</td>
<td>107.2</td>
<td>244</td>
<td>36</td>
</tr>
<tr>
<td>8. Exxon</td>
<td>U.S.</td>
<td>Oil</td>
<td>100.7</td>
<td>6,370</td>
<td>79</td>
</tr>
<tr>
<td>9. General Electric</td>
<td>U.S.</td>
<td>Electrical</td>
<td>100.5</td>
<td>9,296</td>
<td>293</td>
</tr>
<tr>
<td>10. Toyota Motor</td>
<td>Japan</td>
<td>Autos</td>
<td>99.7</td>
<td>2,787</td>
<td>184</td>
</tr>
<tr>
<td>11. Royal Dutch Shell</td>
<td>Brit./Neth.</td>
<td>Oil</td>
<td>93.7</td>
<td>350</td>
<td>102</td>
</tr>
<tr>
<td>12. Marubeni</td>
<td>Japan</td>
<td>Trading</td>
<td>93.6</td>
<td>−921</td>
<td>65</td>
</tr>
<tr>
<td>13. Sumitomo</td>
<td>Japan</td>
<td>Trading</td>
<td>89.0</td>
<td>−102</td>
<td>31</td>
</tr>
<tr>
<td>14. IBM</td>
<td>U.S.</td>
<td>Computers</td>
<td>81.7</td>
<td>6,328</td>
<td>291</td>
</tr>
<tr>
<td>15. AXA</td>
<td>France</td>
<td>Insurance</td>
<td>78.7</td>
<td>1,702</td>
<td>88</td>
</tr>
</tbody>
</table>


TABLE 1.11 | Fortune’s Most Admired U.S. Companies, 1998

<table>
<thead>
<tr>
<th>Company Name</th>
<th>Company Rank</th>
<th>Chief Executive</th>
<th>1998 Total Return</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Electric</td>
<td>1</td>
<td>Jack Welch</td>
<td>41.0%</td>
</tr>
<tr>
<td>Coca-Cola</td>
<td>2</td>
<td>Doug Ivester</td>
<td>1.3%</td>
</tr>
<tr>
<td>Microsoft</td>
<td>3</td>
<td>Bill Gates</td>
<td>114.6%</td>
</tr>
<tr>
<td>Dell Computer</td>
<td>4</td>
<td>Michael Dell</td>
<td>248.5%</td>
</tr>
<tr>
<td>Berkshire Hathaway</td>
<td>5</td>
<td>Warren Buffett</td>
<td>52.2%</td>
</tr>
<tr>
<td>Wal-Mart Stores</td>
<td>6</td>
<td>David Glass</td>
<td>107.6%</td>
</tr>
<tr>
<td>Southwest Airlines</td>
<td>7</td>
<td>Herb Kelleher</td>
<td>38.4%</td>
</tr>
<tr>
<td>Intel</td>
<td>8</td>
<td>Craig Barrett</td>
<td>69.0%</td>
</tr>
<tr>
<td>Merck</td>
<td>9</td>
<td>Raymond Gilmartin</td>
<td>41.3%</td>
</tr>
<tr>
<td>Walt Disney</td>
<td>10</td>
<td>Michael Eisner</td>
<td>−8.5%</td>
</tr>
</tbody>
</table>

SOURCE: Author’s table based on Fortune, March 1, 1999, pp. 68ff.
that ranked U.S. corporations from “most admired” = 1 to “least admired” = 469. Table 1.12 provides information about the bottom ten.

a. Identify the elementary units.

b. How many variables can you find in this table? Which are they?

c. Identify the variables as quantitative or qualitative.

d. Identify variables as discrete/continuous or binomial/multinomial.

e. What kind of data set does this table contain?

### TABLE 1.12 | Fortune’s Least Admired U.S. Companies, 1998

<table>
<thead>
<tr>
<th>Company Name</th>
<th>Company Rank</th>
<th>1998 Total Return</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foundation Health</td>
<td>460</td>
<td>−46.6%</td>
</tr>
<tr>
<td>Fruit of the Loom</td>
<td>461</td>
<td>−46.1%</td>
</tr>
<tr>
<td>Viad</td>
<td>462</td>
<td>+59.4%</td>
</tr>
<tr>
<td>Olston</td>
<td>463</td>
<td>−49.9%</td>
</tr>
<tr>
<td>U.S. Industries</td>
<td>464</td>
<td>−38.4%</td>
</tr>
<tr>
<td>Stone Container</td>
<td>465</td>
<td>+36.2%</td>
</tr>
<tr>
<td>Oxford Health Plans</td>
<td>466</td>
<td>−4.4%</td>
</tr>
<tr>
<td>MedPartners</td>
<td>467</td>
<td>−76.5%</td>
</tr>
<tr>
<td>Shoney’s</td>
<td>468</td>
<td>−57.7%</td>
</tr>
<tr>
<td>Trump Hotels &amp; Casinos</td>
<td>469</td>
<td>−43.9%</td>
</tr>
</tbody>
</table>

SOURCE: Author’s table based on Fortune, March 1, 1999, pp. 68ff.

12. Consider the data of Table 1.13.

a. Identify the elementary units.

b. How many variables can you find in this table? Which are they?

### TABLE 1.13 | Best Picture Nominees for the 1998 Academy Awards

<table>
<thead>
<tr>
<th>Title</th>
<th>Distributor</th>
<th>Gross Revenue by February 1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saving Private Ryan</td>
<td>Dreamworks</td>
<td>$196.0 million</td>
</tr>
<tr>
<td>Shakespeare in Love</td>
<td>Miramax</td>
<td>37.6 million</td>
</tr>
<tr>
<td>Thin Red Line</td>
<td>Fox</td>
<td>31.2 million</td>
</tr>
<tr>
<td>Elizabeth</td>
<td>Gramercy</td>
<td>21.6 million</td>
</tr>
<tr>
<td>Life Is Beautiful</td>
<td>Miramax</td>
<td>18.8 million</td>
</tr>
</tbody>
</table>

SOURCE: Author’s table, based on Variety, various issues.
c. Identify the variables as quantitative or qualitative.

d. Identify variables as discrete/continuous or binomial/multinomial.

e. What kind of data set does this table contain?

13. Consider the data of Table 1.14.

a. Identify the elementary units.

b. How many variables can you find in this table? Which are they?

c. Identify the variables as quantitative or qualitative.

d. Identify variables as discrete/continuous or binomial/multinomial.

e. What kind of data set does this table contain?

**TABLE 1.14 | Best Picture Winners at 1993 to 1997 Academy Awards**

<table>
<thead>
<tr>
<th>Title</th>
<th>Distributor</th>
<th>Gross Revenue by February 1999</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schindler’s List, 1993</td>
<td>Universal</td>
<td>$96.1 million</td>
</tr>
<tr>
<td>Forrest Gump, 1994</td>
<td>Paramount</td>
<td>329.7 million</td>
</tr>
<tr>
<td>Braveheart, 1995</td>
<td>Paramount</td>
<td>75.6 million</td>
</tr>
<tr>
<td>The English Patient, 1996</td>
<td>Miramax</td>
<td>78.7 million</td>
</tr>
<tr>
<td>Titanic, 1997</td>
<td>Paramount</td>
<td>600.8 million</td>
</tr>
</tbody>
</table>

SOURCE: Author’s table, based on Variety, various issues.

14. Consider the data of Table 1.15.

a. Identify the elementary units.

b. How many variables can you find in this table? Which are they?

c. Identify the variables as quantitative or qualitative.

d. Identify variables as discrete/continuous or binomial/multinomial.

e. What kind of data set does this table contain?

**TABLE 1.15 | Best U.S. Video Sales, February 1–7, 1999**

<table>
<thead>
<tr>
<th>Title</th>
<th>Sales Rank</th>
<th>Distributor</th>
<th>Retail Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mulan</td>
<td>1</td>
<td>Disney</td>
<td>$26.99</td>
</tr>
<tr>
<td>City of Angels</td>
<td>2</td>
<td>Warner</td>
<td>19.95</td>
</tr>
<tr>
<td>Tae-Bo Workout</td>
<td>3</td>
<td>Ventura</td>
<td>39.95</td>
</tr>
<tr>
<td>U.S. Marshals</td>
<td>4</td>
<td>Warner</td>
<td>19.98</td>
</tr>
<tr>
<td>Parent Trap</td>
<td>5</td>
<td>Disney</td>
<td>22.99</td>
</tr>
</tbody>
</table>

SOURCE: Author’s table, based on Videoscan.
15. Consider the data of Table 1.16.
   a. Identify the elementary units.
   b. How many variables can you find in this table? Which are they?
   c. Identify the variables as quantitative or qualitative.
   d. Identify variables as discrete/continuous or binomial/multinomial.
   e. What kind of data set does this table contain?

<table>
<thead>
<tr>
<th>Title</th>
<th>Rental Rank</th>
<th>Distributor</th>
<th>Rentals per Copy</th>
</tr>
</thead>
<tbody>
<tr>
<td>There’s Something...</td>
<td>1</td>
<td>Fox</td>
<td>3.8</td>
</tr>
<tr>
<td>Rush Hour</td>
<td>2</td>
<td>New Line</td>
<td>4.0</td>
</tr>
<tr>
<td>The Truman Show</td>
<td>3</td>
<td>Paramount</td>
<td>3.7</td>
</tr>
<tr>
<td>Mulan</td>
<td>4</td>
<td>Disney</td>
<td>3.3</td>
</tr>
<tr>
<td>Lethal Weapon 4</td>
<td>5</td>
<td>Warner</td>
<td>2.2</td>
</tr>
</tbody>
</table>

SOURCE: Author’s table, based on Video Business Magazine.

16. Consider the data of Table 1.17.
   a. Identify the elementary units.
   b. How many variables can you find in this table? Which are they?
   c. Identify the variables as quantitative or qualitative.
   d. Identify variables as discrete/continuous or binomial/multinomial.
   e. What kind of data set does this table contain?

<table>
<thead>
<tr>
<th>Title</th>
<th>Sales Rank</th>
<th>Publisher</th>
<th>Retail Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turbo Tax</td>
<td>1</td>
<td>Intuit</td>
<td>$18</td>
</tr>
<tr>
<td>Windows 98 Upgrade</td>
<td>2</td>
<td>Microsoft</td>
<td>88</td>
</tr>
<tr>
<td>Turbo Tax Deluxe</td>
<td>3</td>
<td>Intuit</td>
<td>48</td>
</tr>
<tr>
<td>Quicken</td>
<td>4</td>
<td>Intuit</td>
<td>31</td>
</tr>
<tr>
<td>Quicken Deluxe</td>
<td>5</td>
<td>Intuit</td>
<td>58</td>
</tr>
</tbody>
</table>

SOURCE: Author’s table, based on PC Data.

17. Consider the data of Table 1.18.
   a. Identify the elementary units.
   b. How many variables can you find in this table? Which are they?
### TABLE 1.18 | Best U.S. Business Software Sales (Macintosh), December 1998

<table>
<thead>
<tr>
<th>Title</th>
<th>Sales Rank</th>
<th>Publisher</th>
<th>Retail Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mac OS 8.5</td>
<td>1</td>
<td>Apple</td>
<td>$91</td>
</tr>
<tr>
<td>Norton Utilities 4.0</td>
<td>2</td>
<td>Symantec</td>
<td>94</td>
</tr>
<tr>
<td>Printmaster Gold</td>
<td>3</td>
<td>Learning Co.</td>
<td>25</td>
</tr>
<tr>
<td>Norton Antivirus 5.0</td>
<td>4</td>
<td>Symantec</td>
<td>64</td>
</tr>
<tr>
<td>Adobe Illustrator 8.0</td>
<td>5</td>
<td>Adobe</td>
<td>114</td>
</tr>
</tbody>
</table>

SOURCE: Author’s table, based on PC Data.

c. Identify the variables as quantitative or qualitative.
d. Identify variables as discrete/continuous or binomial/multinomial.
e. What kind of data set does this table contain?

18. In each of the following cases, determine whether the data set is univariate, bivariate, or multivariate:
   a. A table contains income data for 50 consumers.
   b. A table contains data on quantity produced and total cost for 7 factories.
   c. A table contains data on total assets, revenue growth, and management style for 100 firms.
   d. A table contains data on job category, sex, years of experience, and performance indexes for 500 employees.
   e. A table contains interest rate forecasts by 30 so-called experts for numerous financial instruments.

19. In each of the following cases, determine whether the data set is univariate, bivariate, or multivariate:
   a. A table contains data on last year’s dollar purchases, estimated annual income, and geographic location of 1 million customers.
   b. A table contains data on the ask and bid prices for 25 different corporate bonds.
   c. A table contains data on quality ratings (A = best and E = worst) for 10 types of refrigerators.
   d. A table contains data on 100 incoming airline passengers who have been questioned about the reason for their trip (10 categories, ranging from business to honeymoon), the likely length of their stay, their likely expenditures in town, and their type of accommodation (6 categories, ranging from hotel to own home).
   e. A table contains data on 50 shoppers concerning the number of CDs bought in the past 12 months, their age, and their favorite type of music (10 categories, ranging from classical to rock).

20. Classify the following variables, first as qualitative or quantitative, and second as binomial/multinomial or discrete/continuous:
   a. the number of telephone calls made by someone during a day
   b. the dollar figures listed on a sheet of paper
   c. the sexes of corporate executives
   d. the running times of participants in a race
   e. the employment/unemployment status of workers
   f. the types of hair coloring sold in a drugstore
21. Classify the following variables, first as qualitative or quantitative, and second as binomial/multinomial or discrete/continuous:
   a. the weight lost by a dieter
   b. the types of skills found among a firm’s employees
   c. the attendance record of students in a class
   d. the ages of applicants for a marriage license
   e. the types of cars seen in a parking lot
   f. the pressure required to fracture a casting

22. Classify the following variables as qualitative or quantitative:
   a. a firm’s average cost of production
   b. a town’s tax rate
   c. the religious affiliations of a firm’s employees
   d. the national unemployment rate
   e. the brands of gasoline for sale in a city
   f. a listing of the states in which 50 firms achieved their highest sales

23. Classify the following variables as qualitative or quantitative:
   a. a list of foreign exchange rates
   b. the number of black executives in an industry
   c. the depth of tread remaining on aircraft tires after 1,000 landings
   d. the Dow Jones Industrial Average
   e. the political party affiliations of a firm’s employees
   f. the types of sports practiced by a group of people

24. Make a list of 6 variables that are qualitative and binomial.
25. Make a list of 6 variables that are qualitative and multinomial.
26. Make a list of 6 variables that are quantitative and discrete.
27. Make a list of 6 variables that are quantitative and continuous.

Section 1.7 Major Types of Data

28. Identify the data types found in
   a. Table 1.10.
   b. Table 1.11.

29. Identify the data types found in
   a. Table 1.12.
   b. Table 1.13.

30. Identify the data types found in
   a. Table 1.14.
   b. Table 1.15.

31. Identify the data types found in
   a. Table 1.16.
   b. Table 1.17.
   c. Table 1.18.

32. Review each of the cases in Practice Problem 18 and identify the data types involved.

33. Review each of the cases in Practice Problem 19 and identify the data types involved.

34. Review each of the cases in Practice Problem 20 and identify the data types involved.

35. Review each of the cases in Practice Problem 21 and identify the data types involved.

36. Review each of the cases in Practice Problem 22 and identify the data types involved.

37. Review each of the cases in Practice Problem 23 and identify the data types involved.

38. Among numbers describing the following, which are nominal data?
   a. distances traveled
   b. student I.D. numbers
   c. net assets
   d. room numbers
39. Classify numbers describing the following as nominal, ordinal, interval, or ratio data:
   a. the location of voters by district
   b. the ages of employees
   c. the order in which cars finish a race
   d. models of computers
   e. the number of white blood cells found in a cubic centimeter
   f. the colors of new cars

40. Classify numbers describing the following as nominal, ordinal, interval, or ratio data:
   a. ratings of colleges
   b. temperature readings at the airport
   c. the daily receipts of a supermarket
   d. consumer brand preferences concerning types of coffee
   e. army ranks
   f. a corporate hierarchy from president to janitor
   g. calendar years

41. A product is produced in six alternative colors: blue, brown, green, red, yellow, and white.
   a. Code the colors with the help of nominal data.
   b. Show why adding, subtracting, multiplying, dividing, ranking, or averaging such data would be nonsensical.
   c. Review Table 1.8 on page 16 and code the column (2) data as “1” for “white” and “2” for “black.” Enter the nine data into a calculator and sum them. Comment on the result.

42. Find a list of the names of the 50 states of the United States. (Your telephone book area code listing might be a good start.)
   a. Code the names in alphabetical order from Alabama = 1 to Wyoming = 50. What kinds of data do you have?
   b. Add the code numbers and divide by 50 to get the average. Interpret your result.

43. List 6 data types that are clearly nominal in nature.
44. List 6 data types that are clearly interval in nature.
45. List 6 data types that are clearly ordinal in nature.
46. Consider the following situations; identify the types of data involved:
   a. A quality inspector has classified defective units of a product as 1 and satisfactory units as 2.
   b. A hotel manager has labeled rooms on the first, second, or third floors by numbers in the 100’s, 200’s, or 300’s, respectively, while also designating rooms on the north or south side of the building by even or odd last digits. Thus, 102, 104, 106 stand for first-floor rooms to the north; 301, 303, 305 for third-floor rooms facing south.

47. Someone claims that coding Olympic “gold,” “silver,” and “bronze” as 3, 2, and 1 amounts to creating interval data. What do you think?

48. The Fujita or F scale measures the intensity of tornadoes, as follows:
   0 Wind velocity 40–72 mph; damages chimneys, tree limbs, and sign boards.
   1 Wind velocity 73–112 mph; flips cars, mobile homes, peels roofing.
   2 Wind velocity 113–157 mph; tears roofs off houses, splinters mobile homes.
   3 Wind velocity 158–206 mph; tears roofs and walls off houses, uproots trees.
   4 Wind velocity 207–260 mph; levels frame houses, generates missiles.
   5 Wind velocity 261–318 mph; hurls houses and cars long distances.
What kind of data are the intensity numbers 0–5?

49. Identify the types of data created when the following are coded from 1–5:
   b. Restaurant ratings: *, **, ***, ****, *****
   c. Soft-drink sizes: baby, small, medium, large, extra large
   d. Employee salary classes: GS1, GS2, GS3, GS4, GS5
   e. Staff positions: president, vice president, department head, associate department head, secretary

50. Identify the types of data created when the following are coded from 1–5:
   a. Method of payment: cash, check, debit card, Visa card, Mastercard
   b. The largest energy companies on Fortune’s 1999 Global 500 list: Suez Lyonnaise des Eaux, Enron, RAO Gazprom, Dynegy, Transcanada Pipelines
   c. The largest aerospace companies on Fortune’s 1999 Global 500 list: Boeing, Lockheed Martin, United Technologies, Raytheon, AlliedSignal
   d. The largest banks on Fortune’s 1999 Global 500 list: Bank of America Corporation, Credit Suisse, Deutsche Bank, HSBC Holdings, ABN AMRO Holding
   e. The largest chemicals companies on Fortune’s 1999 Global 500 list: E. I. du Pont de Nemours, Bayer, BASF, Hoechst, Dow Chemical