CHAPTER 1

Introduction

1.1 INTRODUCTION TO STATISTICAL METHODOLOGY
1.2 DESCRIPTIVE STATISTICS AND INFERENTIAL STATISTICS
1.3 THE ROLE OF COMPUTERS IN STATISTICS
1.4 CHAPTER SUMMARY

1.1 INTRODUCTION TO STATISTICAL METHODOLOGY

The past quarter-century has seen a dramatic increase in the use of statistical methods in the social sciences. There are many reasons for this. More research in the social sciences has taken on a quantitative orientation. Like research in other sciences, research in the social sciences often studies questions of interest by analyzing evidence provided by empirical data. The growth of the Internet has resulted in an increase in the amount of readily available quantitative information. Finally, with the evolution of evermore powerful computers, software, and statistical methodology, new methods are available that can more realistically address the questions that arise in social science research.

Why Study Statistics?

The increased use of statistics is evident in the changes in the content of articles published in social science research journals and reports prepared in government and private industry. A quick glance through recent issues of journals such as American Political Science Review and American Sociological Review reveals the fundamental role of statistics in research. For example, to learn about which factors have the greatest impact on student performance in school or to investigate which factors affect people’s political beliefs or the quality of their health care or their decision about when to retire, researchers collect information and process it using statistical analyses. Because of the role of statistics in many research studies, more and more academic departments require that their majors take statistics courses.

These days, social scientists work in a wide variety of areas that use statistical methods, such as governmental agencies, business organizations, and health care facilities. For example, social scientists in government agencies dealing with human welfare or environmental issues or public health policy invariably need to use statistical methods or at least read reports that contain statistics. Medical sociologists often must evaluate recommendations from studies that contain quantitative investigations of new therapies or new ways of caring for the elderly. Some social scientists help managers to evaluate employee performance using quantitative benchmarks and to determine factors that help predict sales of products. In fact, increasingly many jobs for social scientists expect a knowledge of statistical methods as a basic work tool. As the joke goes, “What did the sociologist who passed statistics say to the sociologist who failed it? ‘I’ll have a Big Mac, fries, and a Coke.’”
But an understanding of statistics is important even if you never use statistical methods in your career. Every day you are exposed to an explosion of information from advertising, news reporting, political campaigning, surveys about opinion on controversial issues, and other communications containing statistical arguments. Statistics helps you make sense of this information and better understand the world. You will find concepts from this text helpful in judging the information you encounter in your everyday life.

We realize you are not reading this book in hopes of becoming a statistician, but you may suffer from math phobia and feel fear at what lies ahead. Just be assured that you can read this book and learn the primary concepts and mathematics of statistics with little knowledge of mathematics. Just because you have difficulty in math courses before does not mean you will be at a disadvantage. To understand this book, logical thinking and perseverance are more important than mathematics. In our experience, the most important factor is how well you do in a statistics course is how much time you spend on the course—attending class, doing homework, reading and re-reading this text, studying your class notes, working with your fellow students, getting help from your professor or teaching assistant—not your mathematical knowledge or your gender or your race or whether you feel fear of statistics at the beginning.

Don’t be frustrated if learning comes slowly and you need to read a chapter several times before it starts to make sense. Just as you would not expect to take a course in a foreign language and be able to speak that language fluently the first time you read text, however, you will better understand how to make sense of statistical information.

Data

Information gathering is at the heart of all sciences, providing the raw data in statistical analyses. The observations gathered on the characteristics of interest are collectively called data.

For example, a study might conduct a survey of 1000 people to observe characteristics such as opinion about the legalization of marijuana, political party affiliation, religious ideology, how often attend religious services, number of years of education, monthly income, marital status, race, and gender. The data for a particular person would consist of observations such as (opinion = do not favor legalization, party = Republican, ideology = conservative, religiosity = once a week, education = 14 years, monthly income in range 40–60 thousand dollars, marital status = married, race = white, gender = female). Looking at the data in the right way helps us learn about how characteristics are related. We can then answer questions such as, “Do people who attend church more often tend to be more politically conservative?”

To generate data, the social sciences use a wide variety of methods, including surveys, experiments, and direct observation of behavior in natural settings. In addition, social scientists often analyze data already recorded for other purposes, such as records, census materials, and hospital files. Existing archived collections of data are called databases. Many databases are now available on the Internet. A very important database for social scientists contains results since 1972 of the General Social Survey (GSS).

**EXAMPLE 1.1 The General Social Survey (GSS)**

Every other year, the National Opinion Research Center at the University of Chicago conducts the General Social Survey (GSS). This survey of about 2000 adults provides data about opinions and behaviors of the American public. Social scientists use the survey to investigate how adult Americans answer a wide diversity of questions, such as “Do you believe in life after death?,” “Would you be willing to pay higher prices if...
to protect the environment?,” and “Do you think a preschool child is likely to suffer if his or her mother works?” Similar surveys occur in other countries, such as the General Social Survey administered by Statistics Canada, the British Social Attitudes Survey, and the Eurobarometer survey and European Social Survey for nations in the European Union.

It is easy to get summaries of data from the GSS database. We’ll demonstrate, using a question it asked in one survey, “About how many good friends do you have?”

- Go to the Web site sda.berkeley.edu/GSS/ at the Survey Documentation and Analysis site at the University of California, Berkeley.
- Click on New SDA.
- The GSS name for the question about number of good friends is NUMFREND. Type NUMFREND as the Row variable name. Click on Run the table.

Now you’ll see a table that shows the possible values for ‘number of good friends’ and the number of people and the percentage who made each possible response. The most common responses were 2 and 3 (about 16% made each of these responses).

What Is Statistics?

In this text, we use the term “statistics” in the broad sense to refer to methods for obtaining and analyzing data.

<table>
<thead>
<tr>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics consists of a body of methods for obtaining and analyzing data.</td>
</tr>
</tbody>
</table>

Specifically, statistics provides methods for

1. **Design**: Planning how to gather data for research studies
2. **Description**: Summarizing the data
3. **Inference**: Making predictions based on the data

**Design** refers to planning how to obtain the data. For a survey, for example, the design aspects would specify how to select the people to interview and would construct the questionnaire to administer.

**Description** refers to summarizing data, to help understand the information they provide. For example, an analysis of the number of good friends based on the GSS data might start with a list of the number reported for each of the people who responded to that question that year. The raw data are a complete listing of observations, person by person. These are not easy to comprehend, however. We get bogged down in numbers. For presentation of results, instead of listing all observations, we could summarize the data with a graph or table showing the percentages reporting 1 good friend, 2 good friends, 3, . . . , and so on. Or we could report the average number of good friends, which was 6, or the most common response, which was 2. Graphs, tables and numerical summaries are called **descriptive statistics**.

**Inference** refers to making predictions based on data. For instance, for the GSS data on reported number of good friends, 6.2% reported having only 1 good friend. Can we use this information to predict the percentage of the more than 200 million adults in the U.S. at that time who had only 1 good friend? A method presented in this book allows us to predict that that percentage is no greater than 8%. Predictions made using data are called **statistical inferences**.

**Description** and **inference** are the two types of **statistical analysis**—ways of analyzing the data. Social scientists use descriptive and inferential statistics to answer questions about social phenomena. For instance, “Is having the death penalty
available for punishment associated with a reduction in violent crime?" “Does student performance in schools depend on the amount of money spent per student the size of the classes, or the teachers’ salaries?"

1.2 DESCRIPTIVE STATISTICS AND INFERENTIAL STATISTICS

Section 1.1 explained that statistics consists of methods for designing studies and analyzing data collected in the studies. Methods for analyzing data include descriptive methods for summarizing the data and inferential methods for making predictions. A statistical analysis is classified as descriptive or inferential, according to whether its main purpose is to describe the data or to make predictions. To explain this distinction further, we next define the population and the sample.

Populations and Samples

The entities that a study observes are called the subjects for the study. Usually the subjects are people, such as in the GSS, but they might instead be families, schools, cities, or companies, for instance.

<table>
<thead>
<tr>
<th>Population and Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>The population is the total set of subjects of interest in a study. A sample is the subset of the population on which the study collects data.</td>
</tr>
</tbody>
</table>

In the 2004 GSS, the sample was the 2813 adult Americans who participated in the survey. The population was all adult Americans at that time — more than 200 million people.

The ultimate goal of any study is to learn about populations. But it is almost always necessary, and more practical, to observe only samples from those populations. For example, the GSS and polling organizations such as the Gallup poll usually select samples of about 1000–3000 Americans to collect information about opinions and beliefs of the population of all Americans.

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive statistics summarize the information in a collection of data.</td>
</tr>
</tbody>
</table>

Descriptive statistics consist of graphs, tables, and numbers such as averages and percentages. The main purpose of descriptive statistics is to reduce the data to simpler and more understandable forms without distorting or losing much information.

Although data are usually available only for a sample, descriptive statistics are also useful when data are available for the entire population, such as in a census. By contrast, inferential statistics apply when data are available only for a sample but we want to make a prediction about the entire population.

<table>
<thead>
<tr>
<th>Inferential Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inferential statistics provide predictions about a population, based on data from a sample of that population.</td>
</tr>
</tbody>
</table>

EXAMPLE 1.2 Belief in Heaven

In two of its surveys, the GSS asked, “Do you believe in heaven?” The population of interest was the collection of all adults in the United States. In the most recent survey
in which this was asked, 86% of the 1158 sampled subjects answered yes. We would be interested, however, not only in those 1158 people but in the entire population of all adults in the U.S.

Inferential statistics provide a prediction about the larger population using the sample data. An inferential method presented in Chapter 5 predicts that the population percentage that believe in heaven falls between 84% and 88%. That is, the sample value of 86% has a “margin of error” of 2%. Even though the sample size was tiny compared to the population size, we can conclude that a large percentage of the population believed in heaven.

Inferential statistical analyses can predict characteristics of entire populations quite well by selecting samples that are small relative to the population size. That’s why many polls sample only about a thousand people, even if the population has millions of people. In this book, we’ll see why this works.

In the past quarter-century, social scientists have increasingly recognized the power of inferential statistical methods. Presentation of these methods occupies a large portion of this textbook, beginning in Chapter 5.

### Parameters and Statistics

<table>
<thead>
<tr>
<th>Parameters and Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A parameter</strong> is a numerical summary of the population. A <strong>statistic</strong> is a numerical summary of the sample data.</td>
</tr>
</tbody>
</table>

Example 1.2 estimated the percentage of Americans who believe in heaven. The parameter was the population percentage who believed in heaven. Its value was unknown. The inference about this parameter was based on a statistic—the percentage of the 1158 subjects interviewed in the survey who answered yes, namely, 86%. Since this number describes a characteristic of the sample, it is a descriptive statistic.

In practice, the main interest is in the values of the parameters, not the values of the statistics for the particular sample selected. For example, in viewing results of a poll before an election, we’re more interested in the population percentages favoring the various candidates than in the sample percentages for the people interviewed. The sample and statistics describing it are important only insofar as they help us make inferences about unknown population parameters.

An important aspect of statistical inference involves reporting the likely precision of the sample statistic that estimates the population parameter. For Example 1.2 on belief in heaven, an inferential statistical method predicted how close the sample value of 86% was likely to be to the unknown percentage of the population believing in heaven. The reported margin of error was 2%.

When data exist for an entire population, such as in a census, it’s possible to find the actual values of the parameters of interest. Then there is no need to use inferential statistical methods.

### Defining Populations: Actual and Conceptual

Usually the population to which inferences apply is an actual set of subjects. In Example 1.2, it was adult residents of the U.S. Sometimes, though, the generalizations refer to a conceptual population—one that does not actually exist but is hypothetical.
For example, suppose a consumer organization evaluates gas mileage for a model of an automobile by observing the average number of miles per gallon for five sample autos driven on a standardized 100-mile course. Their inferences refer to the performance on this course for the conceptual population of all autos of this model that will be or could hypothetically be manufactured.

### 1.3 THE ROLE OF COMPUTERS IN STATISTICS

Over time, ever more powerful computers reach the market, and powerful and easy-to-use software is further developed for statistical methods. This software provides an enormous boon to the use of statistics.

**Statistical Software**

SPSS (Statistical Package for the Social Sciences), SAS, MINITAB, and Stata are the most popular statistical software on college campuses. It is much easier to apply statistical methods using these software than using hand calculation. Moreover, many methods presented in this text are too complex to do by hand or with hand calculators.

Most chapters of this text, including all those that present methods requiring considerable computation, show examples of the output of statistical software. The purpose of this textbook is to teach you what to look for in output and how to interpret it. Knowledge of computer programming is not necessary for using statistical software or for reading this book.

The text appendix explains how to use SPSS and SAS, organized by chapter. You can refer to this appendix as you read each chapter to learn how to use them perform the analyses of that chapter.

**Data Files**

Figure 1.1 shows an example of data organized in a data file for analysis by statistical software. A data file has the form of a spreadsheet:

- Any one row contains the observations for a particular subject in the sample.
- Any one column contains the observations for a particular characteristic.

Figure 1.1 is a window for editing data in SPSS. It shows data for the first subjects in a sample, for the characteristics sex, racial group, marital status, age, annual income (in thousands of dollars). Some of the data are numerical, and some consist of labels. Chapter 2 introduces the types of data for data files.

**Uses and Misuses of Statistical Software**

A note of caution: The easy access to statistical methods using software has drawbacks as well as benefits. It is simple to apply inappropriate methods. A computer performs the analysis requested whether or not the assumptions required for its proper use are satisfied.

Incorrect analyses result when researchers take insufficient time to understand the statistical method, the assumptions for its use, or its appropriateness for a specific problem. It is vital to understand the method before using it. Just knowing how to use statistical software does not guarantee a proper analysis. You'll need a good background in statistics to understand which method to select, which of the options to choose in that method, and how to make valid conclusions from the output of the method.
1.4 CHAPTER SUMMARY

The field of statistics includes methods for
- designing research studies,
- describing the data, and
- making inferences (predictions) using the data.

Statistical methods normally are applied to observations in a sample taken from the population of interest. Statistics summarize sample data, while parameters summarize entire populations. There are two types of statistical analyses:

- Descriptive statistics summarize sample or population data with numbers, tables, and graphs.
- Inferential statistics make predictions about population parameters, based on sample data.

A data file has a separate row of data for each subject and a separate column for each characteristic. Statistical methods are easy to apply to data files using software. This relieves us of computational drudgery and helps us focus on the proper application and interpretation of the methods.

PROBLEMS

Practicing the Basics

1.1. The Environmental Protection Agency (EPA) uses a few new automobiles of each brand every year to collect data on pollution emission and gasoline mileage performance. For the Toyota Prius brand, identify the (a) subject, (b) sample, (c) population.

1.2. In the 2006 gubernatorial election in California, an exit poll sampled 2705 of the 7 million people who voted. The poll stated that 56.5%
Chapter 2

Sampling and Measurement

2.1 VARIABLES AND THEIR MEASUREMENT
2.2 RANDOMIZATION
2.3 SAMPLING VARIABILITY AND POTENTIAL BIAS
2.4 OTHER PROBABILITY SAMPLING METHODS*
2.5 CHAPTER SUMMARY

To analyze social phenomena with a statistical analysis, descriptive methods summarize the data and inferential methods use sample data to make predictions about populations. In gathering data, we must decide which subjects to sample. Selecting a sample that is representative of the population is a primary topic of this chapter.

Given a sample, we must convert our ideas about social phenomena into data through deciding what to measure and how to measure it. Developing ways to measure abstract concepts such as achievement, intelligence, and prejudice is one of the most challenging aspects of social research. A measure should have validity, describing what it is intended to measure and accurately reflecting the concept. It should also have reliability, being consistent in the sense that a subject will give the same response when asked again. Invalid or unreliable data-gathering instruments render statistical manipulations of the data meaningless.

The first section of this chapter introduces definitions pertaining to measurement, such as types of data. The other sections discuss ways, good and bad, of selecting the sample.

2.1 VARIABLES AND THEIR MEASUREMENT

Statistical methods help us determine the factors that explain variability among subjects. For instance, variation occurs from student to student in their college grade point average (GPA). What is responsible for that variability? The way those students vary in how much they study per week? in how much they watch TV per day? in their IQ? in their college board score? in their high school GPA?

Variables

Any characteristic we can measure for each subject is called a variable. The name reflects that values of the characteristic vary among subjects.

<table>
<thead>
<tr>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>A variable is a characteristic that can vary in value among subjects in a sample or population.</td>
</tr>
</tbody>
</table>

Different subjects may have different values of a variable. Examples of variables are income last year, number of siblings, whether employed, and gender. The values the variable can take form the measurement scale. For gender, for instance, the
measurement scale consists of the two labels, female and male. For number of siblings it is 0, 1, 2, 3, ... .

The valid statistical methods for a variable depend on its measurement scale. We treat a numerical-valued variable such as annual income differently than a variable with a measurement scale consisting of categories, such as (yes, no) for whether employed. We next present ways to classify variables. The first type refers to whether the measurement scale consists of categories or numbers. Another type refers to the number of levels in that scale.

**Quantitative and Categorical Variables**

A variable is called *quantitative* when the measurement scale has numerical values. The values represent different magnitudes of the variable. Examples of quantitative variables are a subject’s annual income, number of siblings, age, and number of years of education completed.

A variable is called *categorical* when the measurement scale is a set of categories. For example, marital status, with categories (single, married, divorced, widowed), is categorical. For Canadians, the province of residence is categorical, with the categories Alberta, British Columbia, and so on. Other categorical variables are whether employed (yes, no), primary shopping destination (local mall, local downtown, Internet, other), favorite type of music (classical, country, folk, jazz, rock), religion (Protestant, Catholic, Jewish, Muslim, other, none), and political party preference.

For categorical variables, distinct categories differ in quality, not in numerical magnitude. Categorical variables are often called *qualitative*. We distinguish between categorical and quantitative variables because different statistical methods apply to each type. Some methods apply to categorical variables and others apply to quantitative variables. For example, the *average* is a statistical summary for a quantitative variable, because it uses numerical values. It’s possible to find the average for a quantitative variable such as income, but not for a categorical variable such as religious affiliation or favorite type of music.

**Nominal, Ordinal, and Interval Scales of Measurement**

For a quantitative variable, the possible numerical values are said to form an *interval* scale. Interval scales have a specific numerical distance or *interval* between each pair of levels. Annual income is usually measured on an interval scale. The interval between $40,000 and $30,000, for instance, equals $10,000. We can compare outcomes in terms of how much larger or how much smaller one is than the other.

Categorical variables have two types of scales. For the categorical variable mentioned in the previous subsection, the categories are unordered. The scale does not have a "high" or "low" end. The categories are then said to form a *nominal scale*. For another example, a variable measuring primary mode of transportation to work might use the nominal scale with categories (automobile, bus, subway, bicycle, walk).

Although the different categories are often called the *levels* of the scale, for a nominal variable no level is greater than or smaller than any other level. Names or labels such as “automobile” and “bus” for mode of transportation identify the categories but do not represent different magnitudes. By contrast, each possible value of a quantitative variable is greater than or less than any other possible value.

A third type of scale falls, in a sense, between nominal and interval. It consists of categorical scales having a natural *ordering* of values. The levels form an *ordinal scale*. Examples are social class (upper, middle, lower), political philosophy (very liberal, slightly liberal, moderate, slightly conservative, very conservative),
government spending on the environment (too little, about right, too much), and frequency of religious activity (never, less than once a month, about 1–3 times a month, every week, more than once a week). These scales are not nominal, because the categories are ordered. They are not interval, because there is no defined distance between levels. For example, a person categorized as very conservative is more conservative than a person categorized as slightly conservative, but there is no numerical value for how much more conservative that person is.

In summary, for ordinal variables the categories have a natural ordering, whereas for nominal variables the categories are unordered. The scales refer to the actual measurement and not to the phenomena themselves. Place of residence may indicate a geographic place name such as a county (nominal), the distance of that place from a point on the globe (interval), the size of the place (interval or ordinal), or other kinds of variables.

Quantitative Aspects of Ordinal Data

As we’ve discussed, levels of nominal scales are qualitative, varying in quality, not in quantity. Levels of interval scales are quantitative, varying in magnitude. The position of ordinal scales on the quantitative–qualitative classification is fuzzy. Because their scale is a set of categories, they are often analyzed using the same methods as nominal scales. But in many respects, ordinal scales more closely resemble interval scales. They possess an important quantitative feature: Each level has a greater or smaller magnitude than another level.

Some statistical methods apply specifically to ordinal variables. Often, though, it’s helpful to analyze ordinal scales by assigning numerical scores to categories. By treating ordinal variables as interval rather than nominal, we can use the more powerful methods available for quantitative variables.

For example, course grades (such as A, B, C, D, E) are ordinal. But we treat them as interval when we assign numbers to the grades (such as 4, 3, 2, 1, 0) to compute a grade point average. Treating ordinal variables as interval requires good judgment in assigning scores. In doing this, you can conduct a “sensitivity analysis” by checking whether conclusions would differ in any significant way for other choices of the scores.

Discrete and Continuous Variables

One other way to classify a variable also helps determine which statistical methods are appropriate for it. This classification refers to the number of values in the measurement scale.

<table>
<thead>
<tr>
<th>Discrete and Continuous Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>A variable is discrete if its possible values form a set of separate numbers, such as 0, 1, 2, 3, . . . . It is continuous if it can take an infinite continuum of possible real number values.</td>
</tr>
</tbody>
</table>

Examples of discrete variables are the number of siblings and the number of visits to a physician last year. Any variable phrased as “the number of . . .” is discrete, because it is possible to list its possible values {0,1,2,3,4, . . .}.

Examples of continuous variables are height, weight, and the amount of time it takes to read a passage of a book. It is impossible to write down all the distinct potential values, since they form an interval of infinitely many values. The amount of time needed to read a book, for example, could take on the value 8.6294473 . . . hours.

Discrete variables have a basic unit of measurement that cannot be subdivided. For example, 2 and 3 are possible values for the number of siblings, but 2.5716 is
not. For a continuous variable, by contrast, between any two possible values there always another possible value. For example, age is continuous in the sense that individual does not age in discrete jumps. At some well-defined point during the year in which you age from 21 to 22, you are 21.3851 years old, and similarly for every other real number between 21 and 22. A continuous, infinite collection of age values occurs between 21 and 22 alone.

Any variable with a finite number of possible values is discrete. All categorical variables, nominal or ordinal, are discrete, having a finite set of categories. Quantitative variables can be discrete or continuous; age is continuous, and number of siblings is discrete.

For quantitative variables the distinction between discrete and continuous variables can be blurry, because of how variables are actually measured. In practice, we round continuous variables when measuring them, so the measurement is actually discrete. We say that an individual is 21 years old whenever that person’s age is somewhere between 21 and 22. On the other hand, some variables, although discrete, have a very large number of possible values. In measuring annual family income in dollars, the potential values are 0, 1, 2, 3, …, up to some very large value in millions.

What’s the implication of this? Statistical methods for discrete variables are mainly used for quantitative variables that take relatively few values, such as the number of times a person has been married. Statistical methods for continuous variables are used for quantitative variables that can take lots of values, regardless of whether they are theoretically continuous or discrete. For example, statisticians treat variables such as age, income, and IQ as continuous.

In summary,

- Variables are either quantitative (numerical valued) or categorical. Quantitative variables are measured on an interval scale. Categorical variables with unordered categories have a nominal scale, and categorical variables with ordered categories have an ordinal scale.
- Categorical variables (nominal or ordinal) are discrete. Quantitative variables can be either discrete or continuous. In practice, quantitative variables that can take lots of values are treated as continuous.

Figure 2.1 summarizes the types of variables, in terms of the (quantitative, categorical, nominal, ordinal, interval), and (continuous, discrete) classifications.

![Diagram showing the classification of variables]

Note: Ordinal data are treated sometimes as categorical and sometimes as quantitative.

**FIGURE 2.1:** Summary of Quantitative—Categorical, Nominal—Ordinal—Interval, Continuous—Discrete Classifications


2.2 RANDOMIZATION

Inferential statistical methods use sample statistics to make predictions about population parameters. The quality of the inferences depends on how well the sample represents the population. This section introduces an important sampling method that incorporates randomization, the mechanism for achieving good sample representation.

Simple Random Sampling

Subjects of a population to be sampled could be individuals, families, schools, cities, hospitals, records of reported crimes, and so on. Simple random sampling is a method of sampling for which every possible sample has equal chance of selection.

Let \( n \) denote the number of subjects in the sample, called the sample size.

<table>
<thead>
<tr>
<th>Simple Random Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>A simple random sample of ( n ) subjects from a population is one in which each possible sample of that size has the same probability (chance) of being selected.</td>
</tr>
</tbody>
</table>

For instance, suppose a researcher administers a questionnaire to one randomly selected adult in each of several households. A particular household contains four adults—mother, father, aunt, and uncle—identified as M, F, A, and U. For a simple random sample of \( n = 1 \) adult, each of the four adults is equally likely to be interviewed. You could select one by placing the four names on four identical ballots and selecting one blindly from a hat. For a simple random sample of \( n = 2 \) adults, each possible sample of size two is equally likely. The six potential samples are \((M, F), (M, A), (M, U), (F, A), (F, U),\) and \((A, U)\). To select the sample, you blindly select two ballots from the hat.

A simple random sample is often just called a random sample. The simple adjective is used to distinguish this type of sampling from more complex sampling schemes presented in Section 2.4 that also have elements of randomization.

Why is it a good idea to use random sampling? Because everyone has the same chance of inclusion in the sample, so it provides fairness. This reduces the chance that the sample is seriously biased in some way, leading to inaccurate inferences about the population. Most inferential statistical methods assume randomization of the sort provided by random sampling.

How to Select a Simple Random Sample

To select a random sample, we need a list of all subjects in the population. This list is called the sampling frame. Suppose you plan to sample students at your school. The population is all students at the school. One possible sampling frame is the student directory.

The most common method for selecting a random sample is to (1) number the subjects in the sampling frame, (2) generate a set of these numbers randomly, and (3) sample the subjects whose numbers were generated. Using random numbers to select the sample ensures that each subject has an equal chance of selection.

<table>
<thead>
<tr>
<th>Random Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random numbers are numbers that are computer generated according to a scheme whereby each digit is equally likely to be any of the integers (0, 1, 2, \ldots, 9) and does not depend on the other digits generated.</td>
</tr>
</tbody>
</table>
Table 2.1 shows a table containing random numbers. The numbers fluctuate according to no set pattern. Any particular number has the same chance of being a 0, 1, 2, ..., or 9. The numbers are chosen independently, so any one digit chosen has no influence on any other selection. If the first digit in a row of the table is 9, for instance, the next digit is still just as likely to be a 9 as a 0 or 1 or any other number. Random numbers are available in published tables and can be generated with software and many statistical calculators.

Suppose you want to select a simple random sample of \( n = 100 \) students from a university student body of size 30,000. The sampling frame is a directory of the students. Select the students by using five-digit sequences to identify them, as follows:

1. Assign the numbers 00001 to 30000 to the students in the directory, using 0001 for the first student in the list, 00002 for the second student, and so on.
2. Starting at any point in the random number table, or by generating random numbers using software or a calculator, choose successive five-digit numbers until you obtain 100 distinct numbers between 00001 and 30000.
3. Include in the sample the students with assigned numbers equal to the random numbers selected.

For example, for the first column of five-digit numbers in Table 2.1, the first three random numbers are 10480, 22368, and 24130. The first three students selected are those numbered 10480, 22368, and 24130.

In selecting the 100 five-digit numbers, skip numbers greater than 30000, such as the next three five-digit numbers in Table 2.1, since no student in the directory has an assigned number that large. After using the first column of five-digit numbers, move to the next column of numbers and continue. If the population size were between 1000 and 9999, you would use four digits at a time. The column (or row) from which you begin selecting the numbers does not matter, since the numbers have no pattern. Most statistical software can do this all for you.

**Collecting Data with Sample Surveys**

Many studies select a sample of people from a population and interview them to collect data. This method of data collection is called a *sample survey*. The interview can be a personal interview, telephone interview, or self-administered questionnaire.

The General Social Survey (GSS) is an example of a sample survey. It gathers information using personal interviews of a random sample of subjects from the U.S. adult population to provide a snapshot of that population. (The survey does not use simple random sampling but rather a method discussed later in the chapter.)
incorporates multiple stages and clustering but is designed to give each family the same chance of inclusion.) National polls such as the Gallup Poll are also sample surveys. They usually use telephone interviews. Since it is often difficult to obtain a sampling frame, many telephone interviews obtain the sample with random digit dialing.

A variety of problems can cause responses from a sample survey to tend to favor some parts of the population over others. Then results from the sample are not representative of the population. We’ll discuss some potential problems in Section 2.3.

Collecting Data with an Experiment

In some studies, data result from a planned experiment. The purpose of most experiments is to compare responses of subjects on some outcome measure, under different conditions. Those conditions are levels of a variable that can influence the outcome. The scientist has the experimental control of being able to assign subjects to the conditions.

For instance, the conditions might be different drugs for treating some illness. The conditions are called treatments. To conduct the experiment, the researcher needs a plan for assigning subjects to the treatments. These plans are called experimental designs. Good experimental designs use randomization to determine which treatment a subject receives.

In the late 1980s, the Physicians’ Health Study Research Group at Harvard Medical School designed an experiment to analyze whether regular intake of aspirin reduces mortality from heart disease. Of about 22,000 male physicians, half were randomly chosen to take an aspirin every other day. The remaining half took a placebo, which had no active agent. After five years, rates of heart attack were compared. By using randomization to determine who received which treatment, the researchers knew the groups would roughly balance on factors that could affect heart attack rates, such as age and quality of health. If the physicians could decide on their own which treatment to take, the groups might have been out of balance on some important factor. Suppose, for instance, younger physicians were more likely to select aspirin. Then, a lower heart attack rate among the aspirin group could occur merely because younger subjects are less likely to suffer heart attacks.

Collecting Data with an Observational Study

In social research, it is rarely possible to conduct experiments. It’s not possible to randomly assign subjects to the groups we want to compare, such as levels of gender or race or educational level or annual income. Many studies merely observe the outcomes for available subjects on the variables without any experimental manipulation of the subjects. Such studies are called observational studies. The researcher measures subjects’ responses on the variables of interest but has no experimental control over the subjects.

With observational studies, comparing groups is difficult because the groups may be imbalanced on variables that affect the outcome. This is true even with random sampling. For instance, suppose we plan to compare black students, Hispanic students, and white students on some standardized test. If white students have a higher average score, a variety of variables might account for that difference, such as parents’ education or parents’ income or quality of school attended. This makes it difficult to compare groups with observational studies, especially when some key variables may not have been measured in the study.
Establishing cause and effect is central to science. But it’s not possible to establish cause and effect definitively with a nonexperimental study, whether it be an observational study with an available sample or a sample survey using random sampling. With an observational study, there’s the strong possibility that the sample does well reflect the population. With an observational study or a sample survey, the always the possibility that some unmeasured variable could be responsible for factors observed in the data. With an experiment that randomly assigns subjects treatments, those treatments should roughly balance on any unmeasured variable. For example, in the heart attack study mentioned above, the doctors taking aspirin would not tend to be younger or of better health than the doctors taking placebo. Because a randomized experiment balances the groups being compared on all factors, it’s possible to study cause and effect more accurately with an experiment than with an observational study.

Whether or not a study is experimental, it’s important to incorporate randomization in any study that plans to make inferences. This randomization could take the form of randomly selecting a sample for a survey, or randomly allocating subjects to different treatments for an experimental study.

### 2.3 Sampling Variability and Potential Bias

Even if a study wisely uses randomization, the results of the study still depend on which subjects are sampled. Two researchers who separately select random samples from the same population may have little overlap, if any, between the sample memberships. Therefore, the values of sample statistics will differ for the samples, and the results of analyses based on these samples may differ.

#### Sampling Error

Suppose the Gallup, Harris, Zogby, and Pew polling organizations each randomly sample 1000 adult Canadians, in order to estimate the percentage of Canadians who give the prime minister’s performance in office a favorable rating. Based on their different samples, perhaps Gallup reports an approval rating of 63%, Harris reports 68%, Zogby 65%, and Pew 64%. These differences could reflect slight differences in question wording. But even if the questions were worded exactly the same, the percentages would probably differ somewhat because the samples are different.

For conclusions based on statistical inference to be worthwhile, we should know the potential sampling error—how much the statistic differs from the parameter it predicts because of the way results naturally exhibit variation from sample to sample.

**Sampling Error**

The sampling error of a statistic equals the error that occurs when we use a statistic based on a sample to predict the value of a population parameter.

Suppose that the actual percentage of the population of adult Canadians who give the prime minister a favorable rating is 66%. Then the Gallup organization, which sampled 63%, had a sampling error of $63\% - 66\% = -3\%$. The Harris organization, which predicted 68%, had a sampling error of $68\% - 66\% = 2\%$. In practice, sampling error is unknown, because the values of population parameters are unknown.

Random sampling protects against bias, in the sense that the sampling error tends to fluctuate about 0, sometimes being positive (as in the Harris poll) and sometimes being negative (as in the Gallup poll). Random sampling also allows us to predict the likely size of the sampling error. For sample sizes of about 1000, we’ll see
the sampling error for estimating percentages is usually no greater than plus or minus 3%. This bound is the margin of error. Variability also occurs in the values of sample statistics with nonrandom sampling, but the extent of the sampling error is not predictable as it is with random samples.

**Sampling Bias: Nonprobability Sampling**

Other factors besides sampling error can cause results to vary from sample to sample. These factors can also possibly cause bias. We next discuss three types of bias. The first is called sampling bias.

For simple random sampling, each possible sample of n subjects has the same probability of selection. This is a type of probability sampling method, meaning that the probability any particular sample will be selected is known. Inferential statistical methods assume probability sampling. Nonprobability sampling methods are ones for which it is not possible to determine the probabilities of the possible samples. Inferences using such samples have unknown reliability and result in sampling bias.

The most common nonprobability sampling method is volunteer sampling. As the name implies, subjects volunteer to be in the sample. But the sample may poorly represent the population and yield misleading conclusions. For instance, a mail-in questionnaire published in TV Guide posed the question, “Should the President have the Line Item Veto to eliminate waste?” Of those who responded, 97% said yes. For the same question posed to a random sample, 71% said yes.¹

Examples of volunteer sampling are visible any day on many Internet sites and television news programs. Viewers register their opinions on an issue by voting over the Internet. The viewers who respond are unlikely to be a representative cross section, but will be those who can easily access the Internet and who feel strongly enough to respond. Individuals having a particular opinion might be much more likely to respond than individuals having a different opinion. For example, one night the ABC program Nightline asked viewers whether the United Nations should continue to be located in the United States. Of more than 186,000 respondents, 67% wanted the United Nations out of the United States. At the same time, a poll using a random sample of about 500 respondents estimated the population percentage to be about 28%. Even though the random sample had a much smaller size, it is far more trustworthy.

A large sample does not help with volunteer sampling—the bias remains. In 1936, the newsweekly Literary Digest sent over 10 million questionnaires in the mail to predict the outcome of the presidential election. The questionnaires went to a relatively wealthy segment of society (those having autos or telephones), and fewer than 25% were returned. The journal used these to predict an overwhelming victory by Alfred Landon over Franklin Roosevelt. The opposite result was predicted by George Gallup with a much smaller sample in the first scientific poll taken for this purpose. In fact, Roosevelt won in a landslide.

Unfortunately, volunteer sampling is sometimes necessary. This is often true in medical studies. Suppose a study plans to investigate how well a new drug performs compared to a standard drug, for subjects who suffer from high blood pressure. The researchers are not going to be able to find a sampling frame of all who suffer from high blood pressure and take a simple random sample of them. They may, however, be able to sample such subjects at certain medical centers or using volunteers. Even then, randomization should be used wherever possible. For the study patients, they can randomly select who receives the new drug and who receives the standard one.

Even with random sampling, sampling bias can occur. One case is when sampling frame suffers from **undercoverage**: It lacks representation from specific groups in the population. A telephone survey will not reach prison inmates, homeless people, or people too poor to afford a telephone, whereas families with many phones will tend to be over-represented. Responses by those not having a telephone might tend to be quite different from those actually sampled, leading to biased results.

**Response Bias**

In a survey, the way a question is worded or asked can have a large impact on results. For example, when a *New York Times/CBS News* poll in 2006 asked the interviewee whether they would be in favor of a new gasoline tax, only 12% said yes. When the tax was presented as reducing U.S. dependence on foreign oil, 55% said yes, even when asked about a gas tax that would help reduce global warming, 59% said yes.

Poorly worded or confusing questions result in **response bias**. Even the order in which questions are asked can influence the results dramatically. During the Vietnam War, a study asked, “Do you think the U.S. should let Russian newspaper reporters come here and send back whatever they want?” and “Do you think Russia should let American newspaper reporters come in and send back whatever they want?” with percentage of yes responses to the first question was 36% when it was asked first, 73% when it was asked second.

In an interview, characteristics of the interviewer may result in response bias. Respondents might lie if they think their belief is socially unacceptable. They may be more likely to give the answer that they think the interviewer prefers. An example is provided by a study on the effect of the interviewer’s race. Following a phone interview, respondents were asked whether they thought the interviewer was black or white (all were actually white). Perceiving a white interviewer resulted in more conservative opinions. For example, 14% agreed that “American society is fair to everyone” when they thought the interviewer was black, but 31% agreed to this when they thought the interviewer was white.

**Nonresponse Bias: Missing Data**

Some subjects who are supposed to be in the sample may refuse to participate; it may not be possible to reach them. This results in the problem of **nonresponse bias**. If only half the intended sample was actually observed, we should worry about whether the half not observed differ from those observed in a way that causes bias in the results. Even if we select the sample randomly, the results are questionable if there is substantial nonresponse, say, over 20%.

For her book *Women in Love*, author Shere Hite surveyed women in the United States. One of her conclusions was that 70% of women who had been married at least five years have extramarital affairs. She based this conclusion on responses to questionnaires returned by 4500 women. This sounds like an impressively large sample. However, the questionnaire was mailed to about 100,000 women. We can know whether the 4.5% of the women who responded were representative of the 100,000 who received the questionnaire, much less the entire population of American women. This makes it dangerous to make an inference to the larger population.

---

Missing data is a problem in almost all large studies. Some subjects do not provide responses for some of the variables measured. Even in censuses, which are designed to observe everyone in a country, some people are not observed or fail to cooperate. Most software ignores cases for which observations are missing for at least one variable used in an analysis. This results in wasted information and possible bias. Statisticians have recently developed methods that replace missing observations by predicted values based on patterns in the data. See Allison (2002) for an introduction to ways of dealing with missing data.

Summary of Types of Bias

In summary, sample surveys have potential sources of bias:

- **Sampling bias** occurs from using nonprobability samples or having undercoverage.
- **Response bias** occurs when the subject gives an incorrect response (perhaps lying), or the question wording or the way the interviewer asks the questions is confusing or misleading.
- **Nonresponse bias** occurs when some sampled subjects cannot be reached or refuse to participate or fail to answer some questions.

In any study, carefully assess the scope of conclusions. Evaluate critically the conclusions by noting the makeup of the sample. How was the sample selected? How large was it? How were the questions worded? Who sponsored and conducted the research? The less information that is available, the less you should trust it.

Finally, be wary of any study that makes inferences to a broader population than is justified by the sample chosen. Suppose a psychologist performs an experiment using a random sample of students from an introductory psychology course. With statistical inference, the sample results generalize to the population of all students in the class. For the results to be of wider interest, the psychologist might claim that the conclusions extend to all college students, to all young adults, or even to all adults. These generalizations may well be wrong, because the sample may differ from those populations in fundamental ways, such as in average age or socioeconomic status.

2.4 OTHER PROBABILITY SAMPLING METHODS*

Section 2.2 introduced simple random sampling and explained its importance to statistical inference. In practice, other probability sampling methods that have elements of randomness are sometimes preferable to simple random sampling or are simpler to obtain.

**Systematic Random Sampling**

Systematic random sampling selects a subject near the beginning of the sampling frame list, skips several names and selects another subject, skips several more names and selects the next subject, and so forth. The number of names skipped at each stage depends on the desired sample size. Here’s how it’s done:

<table>
<thead>
<tr>
<th>Systematic Random Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denote the sample size by ( n ) and the population size by ( N ). Let ( k = N/n ), the population size divided by the sample size. A systematic random sample (1) selects a subject at random from the first ( k ) names in the sampling frame, and (2) selects every ( k )th subject listed after the first. This method is called the skip number.</td>
</tr>
</tbody>
</table>
Suppose you want a systematic random sample of 100 students from a pool of 30,000 students listed in a campus directory. Then, \( n = 100 \) and \( N = 30 \), so \( k = \frac{30,000}{100} = 300 \). The population size is 300 times the sample size you need to select one of every 300 students. You select one student at random using random numbers, from the first 300 students in the directory. Then select every 300th student after the one selected randomly. This produces a sample of size 100. The first three digits in Table 2.1 are 104, which falls between 1 and 300, so you first select the student numbered 104. The numbers of the students selected are 104 + 300 = 404, 404 + 300 = 704, 704 + 300 = 1004, 1004 + 300 = 1304, and so on. The 100th student selected is listed in the 300 names in the directory.

In sampling from a sampling frame, it’s simpler to select a systematic random sample than a simple random sample because it uses only one random number. This method typically provides as good a representation of the population, because the alphabetic listings such as directories of names, values of most variables fluctuate randomly through the list. With this method, statistical formulas based on simple random sampling are usually valid.

A systematic random sample is not a simple random sample, because all samples of size \( n \) are not equally likely. For instance, unlike in a simple random sample, subjects listed next to each other on the list cannot both appear in the sample.

### Stratified Random Sampling

Another probability sampling method, useful in social science research for comparing groups, is **stratified sampling**.

#### Stratified Random Sample

A **stratified random sample** divides the population into separate groups, called **strata**, and then selects a simple random sample from each stratum.

Suppose a study in Cambridge, Massachusetts plans to compare the opinion of registered Democrats and registered Republicans about whether government should guarantee health care to all citizens. Stratifying according to political party registration, the study selects a random sample of Democrats and another random sample of Republicans.

Stratified random sampling is called **proportional** if the sampled strata proportions are the same as those in the entire population. For example, if 90% of the population interest is Democrat and 10% is Republican, then the sampling is proportional if the sample size for Democrats is nine times the sample size for Republicans.

Stratified random sampling is called **disproportional** if the sampled strata proportions differ from the population proportions. This is useful when the population size for a stratum is relatively small. A group that comprises a small part of the population may not have enough representation in a simple random sample to allow precise inferences. It is not possible to compare accurately Democrats to Republicans, for example, if only 10 people in a sample size of 100 are Republican. By contrast, disproportional stratified sample size of 100 might randomly sample 50 Democrats and 50 Republicans.

To implement stratification, we must know the stratum into which each subject in the sampling frame belongs. This usually restricts the variables that can be used for forming the strata. The variables must have strata that are easily identifiable. For example, it would be easy to select a stratified sample of a school population...
using grade level as the stratification variable, but it would be difficult to prepare an adequate sampling frame of city households stratified by household income.

**Cluster Sampling**

Simple, systematic, and stratified random sampling are often difficult to implement, because they require a complete sampling frame. Such lists are easy to obtain for sampling cities or hospitals or schools, for example, but more difficult for sampling individuals or families. *Cluster sampling* is useful when a complete listing of the population is not available.

<table>
<thead>
<tr>
<th>Cluster Random Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Divide the population into a large number of clusters, such as city blocks. Select a simple random sample of the clusters. Use the subjects in those clusters as the sample.</td>
</tr>
</tbody>
</table>

For example, a study might plan to sample about 1% of the families in a city, using city blocks as clusters. Using a map to identify city blocks, it could select a simple random sample of 1% of the blocks and then sample every family on each block. A study of patient care in mental hospitals in Ontario could first randomly sample mental hospitals (the clusters) and then collect data for patients within those hospitals.

What's the difference between a stratified sample and a cluster sample? A stratified sample uses *every* stratum. The strata are usually groups we want to compare. By contrast, a cluster sample uses a *sample* of the clusters, rather than all of them. In cluster sampling, clusters are merely ways of easily identifying groups of subjects. The goal is not to compare the clusters but rather to use them to obtain a sample. Most clusters are not represented in the eventual sample.

Figure 2.2 illustrates the distinction among sampling subjects (simple random sample), sampling clusters of subjects (cluster random sample), and sampling subjects from within strata (stratified random sample). The figure depicts ways to survey 40 students at a school, to make comparisons among Freshmen, Sophomores, Juniors, and Seniors.

**Multistage Sampling**

When conducting a survey for predicting elections, the Gallup Organization often identifies election districts as clusters and takes a simple random sample of them. But then it also takes a simple random sample of households within each selected election

---

**FIGURE 2.2:** Ways of Randomly Sampling 40 Students. The figure is a schematic for a simple random sample, a cluster random sample of 8 clusters of students who live together, and a stratified random sample of 10 students from each class (Fr, So, Ju, Sr).
24 Chapter 2 Sampling and Measurement

district. This is more feasible than sampling every household in the chosen district. This is an example of multistage sampling, which uses combinations of sampling methods.

Here’s an example of a multistage sample:

- Treat counties (or census tracts) as clusters and select a random sample of certain number of them.
- Within each county selected, take a cluster random sample of square-mile regions.
- Within each region selected, take a systematic random sample of every third house.
- Within each house selected, select one adult at random for the sample.

Multistage samples are common in social science research. They are simpler to implement than simple random sampling but provide a broader sampling of population than a single method such as cluster sampling.

For statistical inference, stratified samples, cluster samples, and multistage sampling use different formulas from the ones in this book. Cluster sampling requires a larger sample to achieve as much inferential precision as simple random sampling. Observations within clusters tend to be similar, because of the tendency of similar living near one another to have similar values on opinion issues and on economic and demographic variables such as age, income, race, and occupation. So we need more data to obtain a representative cross section. By contrast, the result of stratified sampling may be more precise than those stated in this textbook for simple random sampling. Books specializing in sampling methodology provide further detail (Scheaffer et al., 2005; Thompson, 2002).

2.5 CHAPTER SUMMARY

Statistical methods analyze data on variables, which are characteristics that vary among subjects. The statistical methods used depend on the type of variable:

- Numerically measured variables, such as family income and number of children in a family, are quantitative. They are measured on an interval scale.
- Variables taking value in a set of categories are categorical. Those measured with unordered categories, such as religious affiliation and province of residence, have a nominal scale. Those measured with ordered categories, such as class and political ideology, have an ordinal scale of measurement.
- Variables are also classified as discrete, having possible values that are whole numbers (such as 0, 1, 2, . . .), or continuous, having a continuum of possible values. Categorical variables, whether nominal or ordinal, are discrete. Quantitative variables can be of either type, but in practice treated as continuous if they can take a large number of values.

Much social science research uses observational studies, which use available data to observe variables of interest. One should be cautious in attempting to conduct inferential analyses with data from such studies. Inferential statistical methods require probability samples, which incorporate randomization in some way. Random sampling allows control over the amount of sampling error, which describes how much of the sample can vary from sample to sample. Random samples are much more likely to be representative of the population than are nonprobability samples such as volunteer samples.

- For a simple random sample, every possible sample of size $n$ has the chance of selection.
• Here are other examples of probability sampling: **Systematic** random sampling takes every $k$th subject in the sampling frame list. **Stratified** random sampling divides the population into groups (strata) and takes a random sample from each stratum. **Cluster** random sampling takes a random sample of clusters of subjects (such as city blocks) and uses subjects in those clusters as the sample. **Multistage** sampling uses combinations of these methods.

Chapter 3 introduces statistics for describing samples and corresponding parameters for describing populations. Hence, its focus is on **descriptive statistics**.

### PROBLEMS

#### 2.4. Which scale of measurement is most appropriate for:

(a) Occupation (plumber, teacher, secretary, …)
(b) Occupational status (blue collar, white collar)
(c) Social status (lower, middle, upper class)
(d) Statewide murder rate (number of murders per 1000 population)
(e) County population size (number of people)
(f) Population growth rate (in percentages)
(g) Community size (rural, small town, large town, small city, large city)
(h) Annual income (thousands of dollars per year)
(i) Attitude toward affirmative action (favorable, neutral, unfavorable)
(j) Lifetime number of sexual partners

#### 2.5. Which scale of measurement is most appropriate for “attained education” measured as:

(a) Number of years (0, 1, 2, 3, …)
(b) Grade level (elementary school, middle school, high school, college, graduate school)
(c) School type (public school, private school)

#### 2.6. Give an example of a variable that is (a) categorical, (b) quantitative, (c) ordinal scale, (d) nominal scale, (e) discrete, (f) continuous, (g) quantitative and discrete.

#### 2.7. A poll conducted by YouGov for the British newspaper *The Daily Telegraph* in June 2006 asked a random sample of 1962 British adults several questions about their image of the U.S. One question asked, “How would you rate George W. Bush as a world leader?” The possible choices were (great leader, reasonably satisfactory leader, pretty poor leader, terrible leader).

(a) Is this four-category variable nominal, or ordinal? Why?
(b) Is this variable continuous, or discrete? Why?
(c) Of the 93% of the sample who responded, the percentages in the four categories were 1% (great leader), 16%, 37%, 46% (terrible leader). Are these values statistics, or parameters? Why?

#### 2.8. A survey asks subjects to rate five issues according to their importance in determining voting intention.