Are you taking statistics because it is required in your major—not because you find it interesting? In this book, we will show you that statistics can be a lot more interesting and easy to understand than you may have been led to believe. As we draw on your previous knowledge and experience and relate materials to interesting and important social issues, you’ll begin to see that statistics is not just a course you have to take but a useful tool as well.

There are two major reasons why learning statistics may be of value to you. First, you are constantly exposed to statistics every day of your life. Marketing surveys, voting polls, and the findings of social research appear daily in newspapers and popular magazines. By learning statistics, you will become a sharper consumer of statistical material. Second, as a major in the social sciences, even if conducting research is not a part of your job, you may still be expected to understand and learn from other people’s research or to be able to write reports based on statistical analyses.

Just what is statistics anyway? You may associate the word with numbers that indicate birthrates, conviction rates, per capita income, marriage and divorce rates, and so on. But the word statistics also refers to a set of procedures used by social scientists. They use these procedures to organize, summarize, and communicate information. Only information represented by numbers can be the subject of statistical analysis. Such information is called data; researchers use statistical procedures to analyze data to answer research questions and test theories. It is the latter usage—answering research questions and testing theories—that this textbook explores.
Statistics  A set of procedures used by social scientists to organize, summarize, and communicate information.

Data  Information represented by numbers, which can be the subject of statistical analysis.

THE RESEARCH PROCESS

To give you a better idea of the role of statistics in social research, let's start by looking at the research process. We can think of the research process as a set of activities in which social scientists engage so that they can answer questions, examine ideas, or test theories.

As illustrated in Figure 1.1, the research process consists of five stages:

1. Asking the research question
2. Formulating the hypotheses
3. Collecting data
4. Analyzing data
5. Evaluating the hypotheses

Figure 1.1  The Research Process
Each stage affects the theory and is affected by it as well. Statistics is most closely tied to the data analysis stage of the research process. As we will see in later chapters, statistical analysis of the data helps researchers test the validity and accuracy of their hypotheses.

---

**Research process**  A set of activities in which social scientists engage to answer questions, examine ideas, or test theories.

---

## ASKING RESEARCH QUESTIONS

The starting point for most research is asking a *research question*. Consider the following research questions taken from a number of social science journals:

- Does managed health care influence the quality of health care?
- Does social class influence voting behavior?
- What factors influence the economic mobility of female workers?

These are all questions that can be answered by conducting *empirical research*—research based on information that can be verified by using our direct experience. To answer research questions, we cannot rely on reasoning, speculation, moral judgment, or subjective preference. For example, the questions, “Is racial equality good for society?” and “Is an urban lifestyle better than a rural lifestyle?” cannot be answered empirically because the terms *good* and *better* are concerned with values, beliefs, or subjective preference and, therefore, cannot be independently verified. One way to study these questions is by defining *good* and *better* in terms that can be verified empirically. For example, we can define *good* in terms of economic growth and *better* in terms of psychological well-being. These questions could then be answered by conducting empirical research.

---

**Empirical research**  Research based on evidence that can be verified by using our direct experience.

---

You may wonder how to come up with a research question. The first step is to pick a question that interests you. Ideas for research problems are all around you, from media sources to personal experience or your own intuition. Talk to other people, write down your own observations and ideas, or learn what other social scientists have written about.

Take, for instance, the issue of gender and work. In 2008, women who were employed full-time earned about $682 per week on average; men who were employed full-time earned $876 per week on average.¹ Women’s and men’s work are also very different. Women continue to be the minority in many of the higher-ranking and higher-salaried positions in professional and managerial occupations. For
example, in 2008 women made up 10.4% of civil engineers, 30.5% of physicians, 27.2% of dentists, and 24.8% of architects. In comparison, among all those employed as preschool and kindergarten teachers, 97.6% were women. Among all administrative assistants in 2008, 96.1% were women. Another noteworthy development in the history of labor in the United States took place in January 2010: Women outnumbered men for the first time by holding 50.3% of the nonfarm payroll jobs. These observations may prompt us to ask research questions such as the following: Are women paid, on an average, less than men for the same type of work? How much change has there been in women’s work over time?

Identify one or two social science questions amenable to empirical research.

THE ROLE OF THEORY

You may have noticed that each preceding research question was expressed in terms of a relationship. This relationship may be between two or more attributes of individuals or groups, such as gender segregation in the workplace and income disparity. The relationship between attributes or characteristics of individuals and groups lies at the heart of social scientific inquiry.

Most of us use the term theory quite casually to explain events and experiences in our daily life. We may have a “theory” about why our boss has been so nice to us lately or why we didn’t do so well on our last history test. In a somewhat similar manner, social scientists attempt to explain the nature of social reality. To the social scientist, a theory is a more precise explanation that is frequently tested by conducting research.

A theory is an explanation of the relationship between two or more observable attributes of individuals or groups. The theory attempts to establish a link between what we observe (the data) and our conceptual understanding of why certain phenomena are related to each other in a particular way. Suppose we wanted to understand the reasons for the income disparity between men and women; we may wonder whether the types of jobs men and women have and the organizations in which they work have something to do with their wages.

Theory An elaborate explanation of the relationship between two or more observable attributes of individuals or groups.

One explanation for gender inequality in wages is gender segregation in the workplace—the fact that American men and women are concentrated in different kinds of jobs and occupations. For example, in 2008, of the approximately 69 million women in the labor force, more than one third (34%) worked in only 1 of the 14 industries listed by the census. The jobs in which women and men are segregated are not only different but also unequal. Although the proportion of women in the labor force has markedly increased, women are still concentrated in occupations with low pay, low prestige, and few opportunities for promotion. In particular, women’s
segregation into different jobs and occupations from those of men is the most immediate cause of the pay gap. Women receive lower pay than men do even when they have the same level of education, skills, and experience as men in comparable occupations.

FORMULATING THE HYPOTHESES

So far, we have come up with a number of research questions about the income disparity between men and women in the workplace. We have also discussed a possible explanation—a theory—that helps us make sense of gender inequality in wages. Is that enough? Where do we go from here?

Our next step is to test some of the ideas suggested by the gender segregation theory. Theories suggest specific concrete predictions about the way that observable attributes of people or groups are interrelated in real life. These predictions, called hypotheses, are tentative answers to research problems. Hypotheses are tentative because they can be verified only after they have been tested empirically. For example, one hypothesis we can derive from the gender segregation theory is that wages in occupations in which the majority of workers are female are lower than the wages in occupations in which the majority of workers are male.

Hypothesis  A tentative answer to a research problem.

Not all hypotheses are derived directly from theories. We can generate hypotheses in many ways—from theories, directly from observations, or from intuition. Probably, the greatest source of hypotheses is the professional literature. A critical review of the professional literature will familiarize you with the current state of knowledge and with hypotheses that others have studied.

Let’s restate our hypothesis:

Wages in occupations in which the majority of workers are female are lower than the wages in occupations in which the majority of workers are male.

Note that this hypothesis is a statement of a relationship between two characteristics that vary: wages and gender composition of occupations. Such characteristics are called variables. A variable is a property of people or objects that takes on two or more values. For example, people can be classified into a number of social class categories, such as upper class, middle class, or working class. Similarly, people have different levels of education; therefore, education is a variable. Family income is a variable; it can take on values from zero to hundreds of thousands of dollars or more. Wages is a variable, with values from zero to thousands of dollars or more. Similarly, gender composition is a variable. The percentage of females (or males) in an occupation can vary from 0 to 100. (See Figure 1.2 for examples of some variables and their possible values.)

Variable  A property of people or objects that takes on two or more values.
Each variable must include categories that are both exhaustive and mutually exclusive. Exhaustiveness means that there should be enough categories composing the variables to classify every observation. For example, the common classification of the variable marital status into the categories “married,” “single,” “divorced,” and “widowed” violates the requirement of exhaustiveness. As defined, it does not allow us to classify same-sex couples or heterosexual couples who are not legally married. (We can make every variable exhaustive by adding the category “other” to the list of categories. However, this practice is not recommended if it leads to the exclusion of categories that have theoretical significance or a substantial number of observations.)

Mutual exclusiveness means that there is only one category suitable for each observation. For example, we need to define religion in such a way that no one would be classified into more than one category. For instance, the categories “Protestant” and “Methodist” are not mutually exclusive because Methodists are also considered Protestant and, therefore, could be classified into both categories.

Social scientists can choose which level of social life to focus their research on. They can focus on individuals or on groups of people such as families, organizations, and nations. These distinctions are referred to as units of analysis. A variable is a property of whatever the unit of analysis is for the study. Variables can be properties of individuals, of groups (e.g., the family or a social group), of organizations (e.g., a hospital or university), or of societies (e.g., a country or a nation). For example, in a study that looks at the relationship between individuals’ level of education and their income, the variable income...
refers to the income level of an individual. On the other hand, a study that compares how differences in corporations’ revenues relate to differences in the fringe benefits they provide to their employees uses the variable revenue as a characteristic of an organization (the corporation). The variables wages and gender composition in our example are characteristics of occupations. Figure 1.3 illustrates different units of analysis frequently employed by social scientists.

Unit of analysis  The level of social life on which social scientists focus. Examples of different levels are individuals and groups.

Independent and Dependent Variables: Causality

Hypotheses are usually stated in terms of a relationship between an independent and a dependent variable. The distinction between an independent and a dependent variable is important in the language of research. Social theories often intend to provide an explanation for social patterns or causal relations between variables. For example, according to the gender segregation theory, gender segregation in the workplace is the primary explanation (although certainly not the only one) of the male-female earning gap. Why should jobs where the majority of workers are women pay less than jobs that employ mostly men? One explanation is that

societies undervalue the work women do, regardless of what those tasks are, because women do them... For example, our culture tends to devalue caring or nurturant work at least partly because it is done by women. This tendency accounts for child care workers’ low rank in the pay hierarchy.

In the language of research, the variable the researcher wants to explain (the “effect”) is called the dependent variable. The variable that is expected to “cause” or account for the dependent variable is called the independent variable. Therefore, in our example, gender composition of occupations is the independent variable, and wages is the dependent variable.

Dependent variable  The variable to be explained (the “effect”).

Independent variable  The variable expected to account for (the “cause” of) the dependent variable.

Cause-and-effect relationships between variables are not easy to infer in the social sciences. To establish that two variables are causally related, you need to meet three conditions: (1) The cause has
to precede the effect in time, (2) there has to be an empirical relationship between the cause and the effect, and (3) this relationship cannot be explained by other factors.

**Independent and Dependent Variables: Guidelines**

Because of the limitations in inferring cause-and-effect relationships in the social sciences, be cautious about using the terms *cause* and *effect* when examining relationships between variables. However, using the terms *independent variable* and *dependent variable* is still appropriate even when this relationship is not articulated in terms of direct cause and effect. Here are a few guidelines that may help you identify the independent and dependent variables:
1. The dependent variable is always the property that you are trying to explain; it is always the object of the research.

2. The independent variable usually occurs earlier in time than the dependent variable.

3. The independent variable is often seen as influencing, directly or indirectly, the dependent variable.

The purpose of the research should help determine which is the independent variable and which is the dependent variable. In the real world, variables are neither dependent nor independent; they can be switched around depending on the research problem. A variable defined as independent in one research investigation may be a dependent variable in another. For instance, *educational attainment* may be an independent variable in a study attempting to explain how education influences political attitudes. However, in an investigation of whether a person’s level of education is influenced by the social status of his or her family of origin, *educational attainment* is the dependent variable. Some variables, such as race, age, and ethnicity, because they are primordial characteristics that cannot be explained by social scientists, are never considered dependent variables in a social science analysis.

Identify the independent and dependent variables in the following hypotheses:

- Younger Americans are more likely to support health care reform than older Americans.
- People who attend church regularly are more likely to oppose abortion than people who do not attend church regularly.
- Elderly women are more likely to live alone than elderly men.
- Individuals with postgraduate education are likely to have fewer children than those with less education.

What are the independent and dependent variables in your hypothesis?

**COLLECTING DATA**

Once we have decided on the research question, the hypothesis, and the variables to be included in the study, we proceed to the next stage in the research cycle. This step includes measuring our variables and collecting the data. As researchers, we must decide how to measure the variables of interest to us, how to select the cases for our research, and what kind of data collection techniques we will be using. A wide variety of data collection techniques are available to us, from direct observations to survey research, experiments, or secondary sources. Similarly, we can construct numerous measuring instruments. These instruments can be as simple as a single question included in a questionnaire or as complex as a composite measure constructed through the combination of two or more questionnaire items. The choice of a particular data collection method or instrument to measure our variables depends on the study objective. For instance, suppose we decide to study how social class position is related to attitudes about abortion. Since attitudes
about abortion are not directly observable, we need to collect data by asking a group of people questions about their attitudes and opinions. A suitable method of data collection for this project would be a survey that uses some kind of questionnaire or interview guide to elicit verbal reports from respondents. The questionnaire could include numerous questions designed to measure attitudes toward abortion, social class, and other variables relevant to the study.

How would we go about collecting data to test the hypothesis relating the gender composition of occupations to wages? We want to gather information on the proportion of men and women in different occupations and the average earnings for these occupations. This kind of information is routinely collected by the government and published in sources such as bulletins distributed by the U.S. Department of Labor’s Bureau of Labor Statistics and the *Statistical Abstract of the United States*. The data obtained from these sources could then be analyzed and used to test our hypothesis.

**Levels of Measurement**

The statistical analysis of data involves many mathematical operations, from simple counting to addition and multiplication. However, not every operation can be used with every variable. The type of statistical operations we employ depends on how our variables are measured. For example, for the variable *gender*, we can use the number 1 to represent females and the number 2 to represent males. Similarly, 1 can also be used as a numerical code for the category “one child” in the variable *number of children*. Clearly, in the first example, the number is an arbitrary symbol that does not correspond to the property “female,” whereas in the second example the number 1 has a distinct numerical meaning that does correspond to the property “one child.” The correspondence between the properties we measure and the numbers representing these properties determines the type of statistical operations we can use. The degree of correspondence also leads to different ways of measuring—that is, to distinct *levels of measurement*. In this section, we will discuss three levels of measurement: nominal, ordinal, and interval ratio.

**Nominal Level of Measurement**

At the nominal level of measurement, numbers or other symbols are assigned a set of categories for the purpose of naming, labeling, or classifying the observations. *Gender* is an example of a nominal-level variable. Using the numbers 1 and 2, for instance, we can classify our observations into the categories “females” and “males,” with 1 representing females and 2 representing males. We could use any of a variety of symbols to represent the different categories of a nominal variable; however, when numbers are used to represent the different categories, we do not imply anything about the magnitude or quantitative difference between the categories. Because the different categories (e.g., males vs. females) vary in the quality inherent in each but not in quantity, nominal variables are often called qualitative. Other examples of nominal-level variables are political party, religion, and race.
**Ordinal Level of Measurement**

Whenever we assign numbers to rank-ordered categories ranging from low to high, we have an **ordinal** level of measurement. *Social class* is an example of an ordinal variable. We might classify individuals with respect to their social class status as “upper class,” “middle class,” or “working class.” We can say that a person in the category “upper class” has a higher class position than a person in a “middle-class” category (or that a “middle-class” position is higher than a “working-class” position), but we do not know how much higher “upper class” is compared with the “middle class.”

Many attitudes that we measure in the social sciences are ordinal-level variables. Take, for instance, the following statement used to measure attitudes toward same-sex marriages: “Same-sex partners should have the right to marry each other.” Respondents are asked to mark the number representing their degree of agreement or disagreement with this statement. One form in which a number might be made to correspond with the answers can be seen in Table 1.1. Although the differences between these numbers represent higher or lower degrees of agreement with same-sex marriage, the distance between any two of those numbers does not have a precise numerical meaning.

**Ordinal measurement**

Numbers are assigned to rank-ordered categories ranging from low to high.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Strongly agree</td>
</tr>
<tr>
<td>2</td>
<td>Agree</td>
</tr>
<tr>
<td>3</td>
<td>Neither agree nor disagree</td>
</tr>
<tr>
<td>4</td>
<td>Disagree</td>
</tr>
<tr>
<td>5</td>
<td>Strongly disagree</td>
</tr>
</tbody>
</table>

**Interval-Ratio Level of Measurement**

If the categories (or values) of a variable can be rank ordered, and if the measurements for all the cases are expressed in the same units, then an **interval-ratio** level of measurement has been achieved. Examples of variables measured at the interval-ratio level are *age*, *income*, and *SAT scores*. With all these variables, we can compare values not only in terms of which is larger or smaller but also in terms of *how much* larger or smaller one is compared with another. In some discussions of levels of measurement, you will see a distinction made between interval-ratio variables that have a natural zero point (where zero means the absence of the property) and those variables that have zero as an arbitrary point. For example, weight and length have a natural zero point, whereas temperature has an arbitrary zero point. Variables with a natural zero point are also called *ratio variables*. In statistical practice, however,
ratio variables are subjected to operations that treat them as interval and ignore their ratio properties. Therefore, no distinction between these two types is made in this text.

---

### Interval-ratio measurement

Measurements for all cases are expressed in the same units.

---

### Cumulative Property of Levels of Measurement

Variables that can be measured at the interval-ratio level of measurement can also be measured at the ordinal and nominal levels. As a rule, properties that can be measured at a higher level (interval-ratio is the highest) can also be measured at lower levels, but not vice versa. Let’s take, for example, *gender composition of occupations*, the independent variable in our research example. Table 1.2 shows the percentage of women in four major occupational groups as reported in the 2010 *Statistical Abstract of the United States*.

The variable *gender composition* (measured as the percentage of women in the occupational group) is an interval-ratio variable and, therefore, has the properties of nominal, ordinal, and interval-ratio measures. For example, we can say that the management group differs from the natural resources group (a nominal comparison), that service occupations have more women than the other occupational categories (an ordinal comparison), and that service occupations have 34.8 percentage points more women (57.2 – 22.4) than production occupations (an interval-ratio comparison).

The types of comparisons possible at each level of measurement are summarized in Table 1.3 and Figure 1.4. Note that differences can be established at each of the three levels, but only at the interval-ratio level can we establish the magnitude of the difference.

---

### Levels of Measurement of Dichotomous Variables

A variable that has only two values is called a **dichotomous variable**. Several key social factors, such as gender, employment status, and marital status, are dichotomies—that is, you are male or female, employed or unemployed, married or not married. Such variables may seem to be measured at the nominal level: You fit in either one category or the other. No category is naturally higher or lower than the other, so they can’t be ordered.

---

### Table 1.2  Gender Composition of Four Major Occupational Groups

<table>
<thead>
<tr>
<th>Occupational Group</th>
<th>Women in Occupation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management, professional, and related occupations</td>
<td>50.8</td>
</tr>
<tr>
<td>Service occupations</td>
<td>57.2</td>
</tr>
<tr>
<td>Production, transportation, and materials occupations</td>
<td>22.4</td>
</tr>
<tr>
<td>Natural resources, construction, and maintenance occupations</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Table 1.3  Levels of Measurement and Possible Comparisons

<table>
<thead>
<tr>
<th>Level</th>
<th>Different or Equivalent</th>
<th>Higher or Lower</th>
<th>How Much Higher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Interval-ratio</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Figure 1.4  Levels of Measurement and Possible Comparisons: Education Measured on Nominal, Ordinal, and Interval-Ratio Levels

**Nominal Measurement**
- Graduated from public high school
- Graduated from private high school
- Graduated from military academy

**Ordinal Measurement**
- Holds a high school diploma
- Holds a college diploma
- Holds a PhD

**Interval-Ratio Measurement**
- Has 8 years of education
- Has 12 years of education
- Has 16 years of education

---

**Dichotomous variable**  A variable that has only two values.
However, because there are only two possible values for a dichotomy, we can measure it at the ordinal or the interval-ratio level. For example, we can think of “femaleness” as the ordering principle for gender, so that “female” is higher and “male” is lower. Using “maleness” as the ordering principle, “female” is lower and “male” is higher. In either case, with only two classes, there is no way to get them out of order; therefore, gender could be considered at the ordinal level.

Dichotomous variables can also be considered to be interval-ratio level. Why is this? In measuring interval-ratio data, the size of the interval between the categories is meaningful: The distance between 4 and 7, for example, is the same as the distance between 11 and 14. But with a dichotomy, there is only one interval. Therefore, there is really no other distance to which we can compare it.

Mathematically, this gives the dichotomy more power than other nominal-level variables (as you will notice later in the text).

For this reason, researchers often dichotomize some of their variables, turning a multicategory nominal variable into a dichotomy. For example, you may see race (originally divided into many categories) dichotomized into “white” and “nonwhite.” Though this is substantively suspect, it may be the most logical statistical step to take.

When you dichotomize a variable, be sure that the two categories capture a distinction that is important to your research question (e.g., a comparison of the number of white vs. nonwhite U.S. senators).

Discrete and Continuous Variables

The statistical operations we can perform are also determined by whether the variables are continuous or discrete. Discrete variables have a minimum-sized unit of measurement, which cannot be subdivided. The number of children per family is an example of a discrete variable because the minimum unit is one child. A family may have two or three children, but not 2.5 children. The variable wages in our research example is a discrete variable because currency has a minimum unit (1 cent), which
cannot be subdivided. One can have $101.21 or $101.22 but not $101.21843. Wages cannot differ by less than 1 cent—the minimum-sized unit.

Unlike discrete variables, continuous variables do not have a minimum-sized unit of measurement; their range of values can be subdivided into increasingly smaller fractional values. Length is an example of a continuous variable because there is no minimum unit of length. A particular object may be 12 in. long, it may be 12.5 in. long, or it may be 12.532011 in. long. Although we cannot always measure all possible length values with absolute accuracy, it is possible for objects to exist at an infinite number of lengths. In principle, we can speak of a tenth of an inch, a ten thousandth of an inch, or a ten trillionth of an inch. The variable gender composition of occupations is a continuous variable because it is measured in proportions or percentages (e.g., the percentage of women in medicine), which can be subdivided into smaller and smaller fractions.

This attribute of variables—whether they are continuous or discrete—affects subsequent research operations, particularly measurement procedures, data analysis, and methods of inference and generalization. However, keep in mind that, in practice, some discrete variables can be treated as if they were continuous, and vice versa.

Name three continuous and three discrete variables. Determine whether each of the variables in your hypothesis is continuous or discrete.

A CAUTIONARY NOTE: MEASUREMENT ERROR

Social scientists attempt to ensure that the research process is as error free as possible, beginning with how we construct our measurements. We pay attention to two characteristics of measurement: reliability and validity.

Reliability means that the measurement yields consistent results each time it is used. For example, asking a sample of individuals “Do you approve or disapprove of President Obama’s job performance?” is more reliable than asking “What do you think of President Obama’s job performance?” While responses to the second question are meaningful, the answers might be vague and could be subject to different interpretation. Researchers look for the consistency of measurement over time, in relationship with other related measures, or in measurements or observations made by two or more researchers. Reliability is a prerequisite for validity: We cannot measure a phenomenon if the measure we are using gives us inconsistent results.

Validity refers to the extent to which measures indicate what they are intended to measure. While standardized IQ tests are reliable, it is still debated whether such tests measure intelligence or one’s test-taking ability. A measure may not be valid due to individual error (individuals may want to provide socially desirable responses) or method error (questions may be unclear or poorly written).

Specific techniques and practices for determining and improving measurement reliability and validity are the subject of research methods courses.
Following the data collection stage, researchers analyze their data and evaluate the hypotheses of the study. The data consist of codes and numbers used to represent our observations. In our example, each occupational group would be represented by two scores: (1) the percentage of women and (2) the average wage. If we had collected information on 100 occupations, we would end up with 200 scores, 2 per occupational group. However, the typical research project includes more variables; therefore, the amount of data the researcher confronts is considerably larger. We now must find a systematic way to organize these data, analyze them, and use some set of procedures to decide what they mean. These last steps make up the statistical analysis stage, which is the main topic of this textbook. It is also at this point in the research cycle that statistical procedures will help us evaluate our research hypothesis and assess the theory from which the hypothesis was derived.

Descriptive and Inferential Statistics

Statistical procedures can be divided into two major categories: descriptive statistics and inferential statistics. Before we can discuss the difference between these two types of statistics, we need to understand the terms population and sample. A population is the total set of individuals, objects, groups, or events in which the researcher is interested. For example, if we were interested in looking at voting behavior in the last presidential election, we would probably define our population as all citizens who voted in the election. If we wanted to understand the employment patterns of Latinas in our state, we would include in our population all Latinas in our state who are in the labor force.

**Population**  The total set of individuals, objects, groups, or events in which the researcher is interested.

Although we are usually interested in a population, quite often, because of limited time and resources, it is impossible to study the entire population. Imagine interviewing all the citizens of the United States who voted in the last election or even all the Latinas who are in the labor force in our state. Not only would that be very expensive and time-consuming, but we would also probably have a very hard time locating everyone! Fortunately, we can learn a lot about a population if we carefully select a subset from that population. A subset selected from a population is called a sample. Researchers usually collect their data from a sample and then generalize their observations to the larger population.

**Sample**  A relatively small subset selected from a population.

**Descriptive statistics** includes procedures that help us organize and describe data collected from either a sample or a population. Occasionally data are collected on an entire population, as in a census.
Inferential statistics, on the other hand, is concerned with making predictions or inferences about a population from observations and analyses of a sample. For instance, the General Social Survey (GSS), from which numerous examples presented in this book are drawn, is conducted every other year by the National Opinion Research Center (NORC) on a representative sample of several thousands of respondents (e.g., the sample size in 2006 was 4,510, and in 2008 it was 2,023). The survey, which includes several hundred questions, is designed to provide social science researchers with a readily accessible database of socially relevant attitudes, behaviors, and attributes of a cross section of the U.S. adult population. NORC has verified that the composition of the GSS samples closely resembles census data. But because the data are based on a sample rather than on the entire population, the average of the sample does not equal the average of the population as a whole. The tools of statistical inference help determine the accuracy of the sample average obtained by the researchers.

**Descriptive statistics**  Procedures that help us organize and describe data collected from either a sample or a population.

**Inferential statistics**  The logic and procedures concerned with making predictions or inferences about a population from observations and analyses of a sample.

**Evaluating the Hypotheses**

At the completion of these descriptive and inferential procedures, we can move to the next stage of the research process: the assessment and evaluation of our hypotheses and theories in light of the analyzed data. At this next stage, new questions might be raised about unexpected trends in the data and about other variables that may have to be considered in addition to our original variables. For example, we may have found that the relationship between gender composition of occupations and earnings can be observed with respect to some groups of occupations but not others. Similarly, the relationship between these variables may apply for some racial/ethnic groups but not for others.

These findings provide evidence to help us decide how our data relate to the theoretical framework that guided our research. We may decide to revise our theory and hypothesis to take account of these later findings. Recent studies are modifying what we know about gender segregation in the workplace. These studies suggest that race as well as gender shapes the occupational structure in the United States and helps explain disparities in income. This reformulation of the theory calls for a modified hypothesis and new research, which starts the circular process of research all over again.

Statistics provides an important link between theory and research. As our example on gender segregation demonstrates, the application of statistical techniques is an indispensable part of the research process. The results of statistical analyses help us evaluate our hypotheses and theories, discover unanticipated patterns and trends, and provide the impetus for shaping and reformulating our theories. Nevertheless, the importance of statistics should not diminish the significance of the preceding phases of the research process. Nor does the use of statistics lessen the importance of our own judgment in the entire process. Statistical analysis is a relatively small part of the research process, and even the most rigorous statistical
procedures cannot speak for themselves. If our research questions are poorly conceived or our data are flawed due to errors in our design and measurement procedures, our results will be useless.

**LOOKING AT SOCIAL DIFFERENCES**

By the middle of this century, if current trends continue unchanged, the United States will no longer be a predominantly European society. Due mostly to renewed immigration and higher birthrates, the United States is being transformed into a “global society” in which nearly half the population will be of African, Asian, Latino, or Native American ancestry.

Is the increasing diversity of American society relevant to social scientists? What impact will such diversity have on the research methodologies we employ?

In a diverse society stratified by race, ethnicity, class, and gender, less partial and distorted explanations of social relations tend to result when researchers, research participants, and the research process itself reflect that diversity. Such diversity shapes the research questions we ask, how we observe and interpret our findings, and the conclusions we draw.

A statistical approach that focuses on social differences uses statistical tools to examine how variables such as race, class, and gender as well as other demographic categories such as age, religion, and sexual orientation shape our social world and explain our social behavior. Numerous statistical procedures can be applied to describe these processes, and we will begin to look at some of those options in the next chapter.

Whichever model of social research you use—whether you follow a traditional one or integrate your analysis with qualitative data, whether you focus on social differences or any other aspect of social behavior—remember that any application of statistical procedures requires a basic understanding of the statistical concepts and techniques. This introductory text is intended to familiarize you with the range of descriptive and inferential statistics widely applied in the social sciences. Our emphasis on statistical techniques should not diminish the importance of human judgment and your awareness of the person-made quality of statistics. Only with this awareness can statistics become a useful tool for viewing social life.

**MAIN POINTS**

- Statistics are procedures used by social scientists to organize, summarize, and communicate information. Only information represented by numbers can be the subject of statistical analysis.
- The research process is a set of activities in which social scientists engage to answer questions, examine ideas, or test theories. It consists of the following stages: asking the research question, formulating the hypotheses, collecting data, analyzing data, and evaluating the hypotheses.
- A theory is an elaborate explanation of the relationship between two or more observable attributes of individuals or groups.
- Theories offer specific concrete predictions about the way observable attributes of people or groups would be interrelated in real life. These predictions, called hypotheses, are tentative answers to research problems.
- A variable is a property of people or objects that takes on two or more values. The variable that the researcher wants to explain (the “effect”) is
called the dependent variable. The variable that is expected to “cause” or account for the dependent variable is called the independent variable.

- Three conditions are required to establish causal relations: (1) The cause has to precede the effect in time, (2) there has to be an empirical relationship between the cause and the effect, and (3) this relationship cannot be explained by other factors.

- At the nominal level of measurement, numbers or other symbols are assigned to a set of categories to name, label, or classify the observations. At the ordinal level of measurement, categories can be rank ordered from low to high (or vice versa). At the interval-ratio level of measurement, measurements for all cases are expressed in the same unit.

- A population is the total set of individuals, objects, groups, or events in which the researcher is interested. A sample is a relatively small subset selected from a population.

- Descriptive statistics includes procedures that help us organize and describe data collected from either a sample or a population. Inferential statistics is concerned with making predictions or inferences about a population from observations and analyses of a sample.

**KEY TERMS**

- data
- dependent variable
- descriptive statistics
- dichotomous variable
- empirical research
- hypothesis
- independent variable
- inferential statistics
- interval-ratio measurement
- nominal measurement
- ordinal measurement
- population
- research process
- sample
- statistics
- theory
- unit of analysis
- variable

**ON YOUR OWN**

Log on to the web-based student study site at [www.sagepub.com/ssdsextentials](http://www.sagepub.com/ssdsextentials) for additional study questions, web quizzes, web resources, flashcards, codebooks and datasets, web exercises, appendices, and links to social science journal articles reflecting the statistics used in this chapter.

**CHAPTER EXERCISES**

1. In your own words, explain the relationship of data (collecting and analyzing) to the research process. (Refer to Figure 1.1.)

2. Construct potential hypotheses or research questions to relate the variables in each of the following examples. Also, write a brief statement explaining why you believe there is a relationship between the variables as specified in your hypotheses.
   a. Gender and educational level
   b. Income and race
   c. The crime rate and the number of police in a city
   d. Life satisfaction and age
e. A nation's military expenditures as a percentage of its gross domestic product (GDP) and that nation's overall level of security
f. Care of elderly parents and ethnicity

3. Determine the level of measurement for each of the following variables:
   a. The number of people in your family
   b. Place of residence classified as urban, suburban, or rural
   c. The percentage of university students who attended public high school
   d. The rating of the overall quality of a textbook, on a scale from “Excellent” to “Poor”
   e. The type of transportation a person takes to work (e.g., bus, walk, car)
   f. Your annual income
   g. The U.S. unemployment rate
   h. The presidential candidate that the respondent voted for in 2008

4. For each of the variables in Exercise 3 that you classified as interval ratio, identify whether it is discrete or continuous.

5. Why do you think men and women, on average, do not earn the same amount of money? Develop your own theory to explain the difference. Use three independent variables in your theory, with annual income as your dependent variable. Construct hypotheses to link each independent variable with your dependent variable.

6. For each of the following examples, indicate whether it involves the use of descriptive or inferential statistics. Justify your answer.
   a. The number of unemployed people in the United States
   b. Determining students' opinion about the quality of food at the cafeteria based on a sample of 100 students
   c. The national incidence of breast cancer among Asian women
   d. Conducting a study to determine the rating of the quality of a new automobile, gathered from 1,000 new buyers
   e. The average GPA of various majors (e.g., sociology, psychology, English) at your university
   f. The change in the number of immigrants coming to the United States from Southeast Asian countries between 2005 and 2010

7. Identify three social problems or issues that can be investigated with statistics. (One example of a social problem is hate crimes.) Which one of the three issues would be the most difficult to study? Which would be the easiest? Why?

8. Construct measures of political participation at the nominal, ordinal, and interval-ratio levels. (Hint: You can use behaviors such as voting frequency or political party membership.) Discuss the advantages and disadvantages of each.

9. Variables can be measured according to more than one level of measurement. For the following variables, identify at least two levels of measurement. Is one level of measurement better than another? Explain.
   a. Individual age
   b. Annual income
   c. Religiosity
   d. Student performance
   e. Social class
   f. Attitude toward affirmative action