illustration used in the previous section, the subject of the hypothesis is a specific city, designated by the term "City D."

The predicate of a scientific hypothesis contains one or more *variables*. A variable—one of the most important concepts of scientific research—is an observed characteristic or aspect of a unit that can vary from one unit to another and/or change over the course of time. In our illustration, the variable is "crime rate," which of course might differ from city to city or increase or decrease from year to year. Every variable has at least two *values*—also called "attributes"—which may be qualitative or quantitative. The values are the actual properties between/among which a variable can vary or change. In the crime rates illustration, the pair of attributes is *high/low*. Because there are only two values when crime rate is measured in this way, we refer to such a variable as a *dichotomy*, which is the simplest kind.

If, in a specific context, an observed characteristic does not vary (it has only one value), we refer to it as a *constant*. For example, if I were to collect the ages of a group of 10 students, it would be understood that age is a variable. However, these individuals also have at least one characteristic in common that is not a variable but rather a constant. This is the characteristic "status," with the possible values *student* and *nonstudent*. We stress that this depends on context because one might collect data about status from a group of 10 people of the same age (in years). Suppose that in the process we discovered that 5 were students and 5 were not. In this context, age would be the constant and status would be the variable.

The following two sections explore these matters in greater detail. We begin with the subjects of research, the units; we then return to a further discussion of variables, their values, and how they are measured.

Units and Samples in Social Research

Social scientists study and hypothesize about various aspects of (human) social life. Two types of units used in this kind of research have already been introduced: cities in the crime rate illustration and individual students in the case of finding average age. In addition to these, several other kinds of units can be found in the literature of sociology and related fields. A partial list is shown in table 2.1.

To summarize, social scientists study people (singly and in aggregates), formal and informal organizations, political units, and products of human activity. When we consider the fact that for every type of unit there are numerous (perhaps hundreds of) variables, we can appreciate what a complex task sociologists have taken on. In addition, unlike the objects of study in other fields, those in social science are often quite aware they are being studied, they sometimes object, and they can even dispute or try to alter the findings. With these factors in mind, it is easy to understand why the field has been termed "the impossible science" (Turner and Turner, 1990).

TABLE 2.1 Units of observation in sociology

Typical units
• Individuals with various attributes: students, women, veterans, senior citizens, etc.
• Families, both nuclear and extended
• Households
• Sports teams and work groups
• City blocks, census tracts, and other statistical divisions
• States, provinces, and regions of countries
• Small businesses, corporations, public sector bureaucracies
• Colleges and universities
• Nations and groups of nations (e.g., developed and developing countries)
• Documents such as novels and newspapers
• Works of art, TV programs, music videos

In sociology and in other sciences as well, researchers generally do not observe just one unit at a time, although this is often done with case studies. This fact points to another important and related concept: *sample.* A sample is a set of (usually more than one) units upon which scientific observation focuses. In other words, we observe only the units that form our samples and none other. When we define descriptive statistics as the process of organizing our observations and turning data into information, we are implying that we are describing the sample. Recall that *a* statistic is an indicator that conveys information about a sample. So, although it is somewhat redundant, it would be entirely appropriate to refer to descriptive statistics as *sample* statistics.

One such statistic is very familiar to you; in fact, we have mentioned it already. This is the average—or, to be more exact, the sample mean of a variable. But an even more important statistic—so important that we tend to forget that it is a statistic—is the *size* of a sample. This is nearly always designated by the symbol "n" (lowercase), which stands for the *n*umber of units. So when we say, for example, "$n = 10$," we are indicating that a particular sample has 10 units—people, households, and so on.

It is generally the case that a sample size is smaller than the total number of units to which our hypotheses refer. That is, we might form a hypothesis to this effect: "The average age of the students in my college is 21.5 years." But in order to test this hypothesis, we ordinarily do not attempt to track down all of the students—of whom there may be thousands—to determine their ages. Instead, we select a sample of, say, $n = 30$ and determine the age of each member of the sample. Thus, as the 30 students in this illustration are part of the entire student body, a sample is always part of a whole.

The whole from which a sample is selected (or "drawn") is referred to as a sampling *universe* or, more commonly, *population.* Both terms convey the idea of being

all-inclusive, and both are relative—which is why they are qualified here with the word "sampling." For in this context, a universe is not the entirety of objects and events on the earth and beyond. Nor is a population the sum total of all persons living in a part of the world or in the entire world. Rather, these terms are defined as the entire collection of units to which a hypothesis refers and from which a sample is drawn. In a circular but useful definition, then, a sample is a proper subset of a population. Every unit contained in a sample is also included in the population, but some units in the population are not in the sample.

The procedure that allows us to go from observations of samples to conclusions about populations that we do not observe—especially the conclusion that a hypothesis should be accepted or rejected—is at the heart of inductive statistics. We return to it briefly below and then at length in later chapters. Before doing so, however, let us consider the other main part of hypotheses, the predicate: the part that contains the variable(s).

Variables in Social Research

We have noted that a variable refers to a characteristic of a unit that can change over time or differ from unit to unit. In addition, we have seen that every variable must have at least two values or attributes between which it can vary or differ. The variety of variables used in social research is as great as or greater than the types of units that are studied. A partial listing of the categories of variables one is likely to encounter in this context is shown in table 2.2.

It would be difficult to condense such a list much further, in part because there is a great diversity of units and research interests found in the social sciences. However, survey researchers use an acronym to remind them of the kinds of variables that are

TABLE 2.2 Sociological variables

Typical variables

- Biological characteristics: gender and age
- Stratification variables: wealth, political power, prestige, income
- Group characteristics: ethnic membership, marital status, sexual orientation, occupation, nationality, religious affiliation
- Geographic factors: place of birth, residence, migration history
- Knowledge factors: degree to which one is informed about current issues, history, local affairs, and so on
- Attitudinal features: political orientation, attitudes on issues and events
- Practices: frequency of voting, church attendance, TV watching

essential in their work. It is "DKAP." It stands for (1) *D*emographic characteristics, such as age; (2) *K*nowledge; (3) *A*ttitudes; and (4) *P*ractices. This is certainly a useful and easy to remember classification, but of course it refers to only a small part of the enormous number of potentially interesting characteristics of individuals, groups, and products of human activity.

Summary

By this point, you should have a fairly clear idea of what statistics is and why we use it—as phrased in the title of this chapter. According to the concise dictionary definition quoted in the first section, it is the field that focuses on the collection, presentation, analysis, and interpretation of large amounts of numerical data. Social statistics, in particular, is that branch of the field specializing in numerical data that pertain to human units of observation and to variables that refer to the individual characteristics and social relationships among these units.

We use statistics for two broad purposes. One of these, featured in the section on scientific method, is to assist in formulating and testing hypotheses. In the context of testing hypotheses, the procedure is widely known as induction (or inference), and thus the application is referred to as *inductive statistics.* The other purpose to which statistics is put, and the more basic of the two, is description: conveying the meaning of what we observe to others. Several of the key principles of *descriptive statistics* were discussed in the section on units, samples, and variables. An examination of the last of the three, variables, begins our next chapter on measurement and data.

KEY TERMS

Constant: A characteristic or feature of a unit that has only one attribute within a sample and thus does not change, vary, or differ.

Deduction: The logical process in which truth is derived by following set rules, independent of any observations that might be made.

DKAP: Demographics, Knowledge, Attitudes, and Practices. The principal types of variables used in social survey research.

Falsifiability: The property of a statement that makes it possible for it to be shown false.

Hypothesis: A sentence that states what is believed to be true based on prior knowledge.

Induction: The logical process of generalizing from what is immediately observed.

Law-like generalization: A general statement, usually part of a theory, that has been tested and never (yet) shown to be false. Formerly referred to as a "law."

n (lowercase): The *n*umber of units in a sample.

Null hypothesis: A hypothesis that states the opposite of what the researcher believes to be true, usually indicating "no difference," "no relationship," "no connection," etc.

Parameter: A measure of a characteristic of a population.

Research hypothesis: A hypothesis that states what the researcher believes to be true.

Sample: The collection of units actually observed in science, assumed to be drawn from a larger set referred to as a population (or universe).

Scientific theory: A logically organized set of statements that indicates what is known about a certain subject.

Statistic: A measure of a characteristic of a sample.

Unit (of analysis or observation): the individual item, person, group, etc., two or more of which make up a sample or population.

Values (also called "attributes"): Specific categories, rankings, or numbers between which a variable can vary, change, or differ.

Variable: A characteristic or feature of a unit that has more than one value and can thus vary, change, or differ within a sample.

WEB SITES TO BOOKMARK

At the end of each chapter, the URLs (addresses) of several web sites are listed along with a brief description of their contents. You will find visiting these sites to be a useful and interesting way to expand your knowledge about social statistics. Although these addresses were valid at the time of publication, we know that some will be discontinued and others will have changed by the time you read this section. In that case, please consider these URLs to be suggestions, using the list to begin your own exploration. One of the best search engines for such an exploration is www.google.com. Try it.

1. www.emory.edu/EMORY_CLASS/PSYCH230/psych230.html
   This site has annotated lecture materials and descriptions of statistical methods in psychology by Professor J. J. McDowell of Emory University in Atlanta.
2. math.uc.edu/~brycw/classes/147/blue/tools.htm
   Here is a very comprehensive guide to online resources for elementary statistics, including textbooks, web-based software, data sets, exercises and tutorials, and simulations on the Web.
3. www.mste.uiuc.edu/hill/dstat/dstat.html
   This is an online introduction to descriptive statistics maintained by the College of Education at the University of Illinois at Champaign-Urbana.
4. www.statsoftinc.com/textbook/stathome.html
   This Electronic Statistical Textbook (Statistics Homepage) offers training in the understanding and application of statistics.
5. www.yale.edu/ynhti/curriculum/guides/1985/8/85.08.02.x.html
   Yale University maintains this Introduction to Elementary Statistics by Lauretta J. Fox.
6. www.statsoftinc.com/textbook/esc1.html
   Here is another electronic statistics textbook that allows the user to search for terms and general statistical concepts.

SOLUTION-CENTERED APPLICATIONS

In each chapter you will find a set of exercises intended to help you apply some of the key concepts and techniques featured in the preceding text. Some of these will take you to the library; others—in the later chapters—will ask you to solve statistical problems using a computer, and others will ask you simply to use your imagination. It is also possible that some of these applications will in themselves not seem especially interesting; however, they might suggest to you or your instructor other statistics-related activities that would be more appropriate. Making up your own problems in this way is not only a good idea, but it expands the usefulness of the "Solution-Centered Applications" section, which for reasons of space can provide only a sample of relevant, learning-oriented applications.

1. Social research often focuses on institutions of higher education (see Dentler, 2002: Chapter 6). In fact, many colleges and universities now maintain their own department or office of institutional research. Does your school have such a department? This kind of research examines various aspects of a school: its size, its personnel, its financial condition, its curriculum, attitudes of faculty and students, etc. The results of such studies are used for several purposes, including recruitment of future students and strategic planning. The

following exercise asks you to begin an applied institutional research project. In later chapters we will expand on the work you have done here.

Suppose that you have been asked to contribute to a booklet that is to be sent to prospective students and their parents to help them decide whether your school is right for them. In a brief written report, state a reasonable research hypothesis about a relevant aspect of your college or university. For example, "University X has increased its out-of-state enrollment during the past five years." Then state the associated null hypothesis. Next, indicate the "theory" you used to produce your hypothesis; that is, what information do you have that would suggest that the hypothesis is a good guess? Finally, briefly discuss how you would go about testing the hypothesis.

2. Once a client or a community presents a problem to an applied social researcher, often the first step taken is to examine the scholarly literature to see (1) whether, when, and by whom the problem has been encountered in the past and (2) how the problem was approached by others. One essential place to begin this literature search is in the periodical section of a university or college library. The following exercise will help to acquaint you with this crucial aspect of applied social research by asking you not only to visit the periodical section but also to examine some of the material there with regard to units of observation, sampling, and variables.

   Go to the periodical section of your school library and select recent issues of two leading research journals, one in a social science and one in a nonsocial science. Ask the librarian for help in finding these journals, if necessary. From each journal select one article to read that reports a research project. Can you identify, in each article, (1) the units of observation, (2) the sample size, and (3) one or more variables? Write a brief report on the results of your search, and be sure to include the correct citation to the articles you used.

3. The following is a purely logical exercise that focuses on the issue of falsifiability. It is designed to help you understand why applied social researchers insist that the hypotheses they are testing can reasonably expected to be true but that they might prove false, depending on the observations made in testing them. On a sheet of paper, write two sentences that involve social scientific units and variables. Structure one of the sentences so that it is falsifiable and the other so that it is not falsifiable. Discuss why you categorized the sentences in this way.

4. Survey research studies that employ self-administered questionnaires are among the most important tools in applied sociology and related fields. For this exercise, you are to search the Internet for a questionnaire used in an applied social research project. Study the questionnaire and the accompanying explanation. Write a brief report (including the correct citation to the questionnaire) identifying (1) the unit(s) of analysis, (2) the key variable(s) in the study, (3) the hypothesis that the questionnaire appears designed to be testing, (4) the researcher(s) conducting the study, and (5) the client for whom the research is being conducted.

CHAPTER 3

Measurement in Sociology

Quantity, Quality, and Social Aggregates

Now that we have introduced the uses and the main principles of applied social statistics, our overview continues with a discussion of how researchers formally record, or measure, what they observe. The discussion is divided into three main parts. The first consists of a definition and some illustrations of the important concept of *level of measurement*. This refers to the different types of variables employed in social research, ranging from simple categories to "absolute" quantities. Next we consider the distinction between independent and dependent variables, which is the cornerstone of bivariate and multivariate applications. The third and last section of the chapter focuses on sources of social scientific data. These include both data for which the researcher makes the actual observations employed in descriptive and inductive contexts and those that the researcher uses but that have already been collected by another individual or group.

Levels of Measurement

At the end of Chapter 2, we discussed the classification of variables according to substantive categories: that is, the kinds of characteristics to which they refer, such as the survey researcher's DKAP types. However, a far more manageable and, from the statistical point of view, meaningful way of classifying variables is in terms of their types of attributes. The different types are often called *levels of measurement* because (1) each type of attribute can be measured with different scales and (2) there is a hierarchy among these types, with some considered to be at higher levels than others. Another reason why this is an important way of classifying variables is that some statistical procedures and techniques apply to one or two levels but not to the others. This is true of both descriptive and inductive applications.

BOX 3.1 Statistics for Sociologists

Quantity or Quality?

As noted in Chapter 2, a debate has been going on among sociologists for several decades concerning the uses and abuses of quantitative research methods, in general, and statistical approaches, in particular. In a highly controversial book entitled *Fads and Fables in Sociology* (Sorokin, 1956), Pitirim Sorokin, the founder of Harvard University's Department of Sociology, parodied what he considered to be the excessive use of statistics with terms such as "testomania" and "quantophrenia." Some years later, in *The Sociological Imagination* (Mills, 1959), a widely cited book by C. Wright Mills, the terms of the argument were established for decades to come. Indeed, Mills' ideas still resonate among sociologists today. In the book, Mills coined the term "abstracted empiricism" to refer to a style of research in which quantitative facts are gathered for their own sake. He contrasts this with a style that he finds equally troublesome, "Grand Theory": speculation with no reference to facts. The approach that he considers to be the correct combination of the two is "intellectual craftsmanship."

Frequently, the abstracted empiricism style is referred to—with negative connotations—as "positivism," with the suggestion that Henri Saint-Simon and Auguste Comte, who coined the term, believed that statistical analysis in itself can lead to truths about social relations. However, as the contemporary theorist Jonathan Turner (1993: Ch. 1) has clearly demonstrated, the real positivism of Comte assumes that statistical analysis is a necessary part of sociology but not a sufficient one. Qualitative and deductive analysis, as well as a commitment to application, are equally important parts of the scientific enterprise.

In fact, the problem had been discussed centuries before even Saint-Simon took it up by the English philosopher Francis Bacon. In his famous book *The New Science* (Bacon, 2004 [1620]), Bacon speaks of different approaches to science using the metaphors of the ant, the spider, and the bee. The ant, says Bacon, collects sticks and ends up with a pile of sticks. This is the abstracted empiricism of the sociologist who collects facts in the belief that they speak for themselves. The spider, for Bacon, spins elaborate webs out of its own substance, as the grand theorist theorizes for theory's sake. But Bacon views the bee as representing the ideal approach, that of intellectual craftsmanship. For the bee collects nectar (facts) and then combines the nectar with its own substance (qualitative and deductive analysis) to produce honey (scientific truth).

This chapter introduces the principles of measurement of sociological variables. You will note that the first, and clearly a major, type of variable consists of qualitative variables, those at the nominal level: gender, ethnicity, geographic location, political orientation, and so on. These form an essential part of the human experience, yet in themselves they cannot be quantified. Therefore, the statistical techniques that can be applied to

FIGURE 3.1 Sir Francis Bacon (1526–1651), the first modern philosopher of science, who argued that good science must include both theory and empirical observation.

them are limited. But this by no means suggests that they can be ignored.

Similarly, we do not claim that statistical measurement and related techniques can substitute for nonquantitative approaches. Field work (participant observation), interpretation of texts and language, deductive theory building, and the other tools of qualitative analysis are essential aspects of the sociological enterprise. More often than not, sociological research and practice are substantially improved when several types of techniques are incorporated: statistical and qualitative analysis, observation and survey research, sociological generalization and the search for "laws" (also called the "nomothetic" method), as well as the exploration of uniqueness in social situations (the "idiographic" method). Hopefully, you will find in this and related chapters encouragement to avoid the path of the ant (or the spider for that matter) and to strive to emulate the bee.

Nominal Level

Some variables, such as resident status (in a particular state), have attributes that *simply describe a condition*: "resident" and "nonresident." The same would apply to marital status ("never married," "married," "divorced or widowed") and to student status ("current student," "past student," and "nonstudent"). Because the attributes are names for the condition, they are referred to as *nominal* (from the Latin word for "name"). Nominal-level variables are the simplest type, and only a limited range of techniques can be employed with them. Nominal-level data in our Colleges data set include *control* (public/private) and *state*.

Ordinal Level

Some variables have attributes that not only name conditions but also include an obvious ranking among the set of all attributes. The word *ordinal*, from the same Latin

word from which we get "order," means just that: rank-ordered. One familiar example of an ordinal-level variable is class in school: freshman, sophomore, junior, senior, and graduate student. You can see that "freshman" and "sophomore," for example, are different names for different classes; but they also can be clearly ranked: sophomore above freshman, and so on. Other examples include socioeconomic class (high, medium, and low) and results of a competition (first, second, third, . . . , last). All of the statistical techniques that apply at the nominal level also apply at the ordinal, and several additional ones apply at the ordinal but not the nominal. An ordinal-level variable in our Colleges data set is *competitiveness rank.*

Numerical (Quantitative) Variables

Many variables have numbers as attributes. Here we are not referring to the kinds of numbers that indicate rank (1st, 2nd, etc.), because those do not really designate amounts but rather refer to relative positions in a ranking system and are therefore ordinal attributes. The kinds of variables known as *numerical* really do refer to a quantity, such as years of education (e.g., 12), household income (e.g., $45,000), GPA (e.g., 3.5), percentage of something (e.g., 32.6%), and number of siblings (e.g., 3). One attribute for each of the preceding variables is shown in parentheses. These numbers are known as "counting" numbers or "cardinal" numbers—suggesting that they are of basic importance. All of the statistical techniques that apply at the nominal and the ordinal levels also apply to numerical variables, and several additional ones apply to numerical variables but not to the others.

Interval versus Ratio Levels

Often a distinction is made between two types of numerical variables, depending on how the zero point (0.0) is derived. If the zero is merely an arbitrary starting point that is set for the sake of convenience, but does not indicate a complete absence of the variable, the variable is designated *interval level.* If a value of 0 on a variable does mean that no amount has been observed, the variable is understood to be *ratio level.* The ratio level is viewed as the higher of the two, and some operations can be performed with variables at that level that do not apply at the interval level.

One example of an interval-level variable is calendar date. We speak of the year in which Barack Obama was inaugurated as president as 2009 because it was the 2,009th year in the Common (or Christian) Era, which began with the year when, according to medieval scholars, Jesus Christ was born.[1] But we know that there are many other calendars—Muslim, Chinese, Mayan, etc.—and each has its own starting point. For example, the Christian year 2009 is equivalent to the year 5769–5770 in the Jewish calendar, because ancient Jewish scholars set their starting year as what they believed was the year the world was created, and the Common Era began 3,760 years later. So, we might ask,