PART I

An Overview of Applied Social Statistics

25,094

This section discusses the contexts in which the tools and techniques of social statistics are applied. The several topics covered here are meant to answer in advance many of the questions that are likely to arise as the details of descriptive and inductive statistics are introduced: What is statistics? Why is the field increasingly viewed as a necessary part of an undergraduate education? What is its relationship to the scientific method? What are the basic concepts in the field?

In addressing these concerns, the material is divided into two parts. The first, to which Chapter 2 is devoted, is the more philosophically and methodologically oriented. For many readers, this will be more of a review of principles presented in other courses, including courses on the philosophy of science and social research methods, than the presentation of entirely new ideas. Yet, even for those of you who are familiar with concepts such as *theory*, *sample*, *hypothesis*, and *variable*, this is likely to be your first encounter with the specific roles that these play in statistical applications.

Chapter 3 focuses on the fundamentals of scientific observation. Beginning with a closer look at variables (the topic with which Chapter 2 ends), the discussion is guided by three main sets of learning objectives: (1) to become familiar with the notion of data and the related concept of information, (2) to understand how measurement is applied to observations and data, and (3) to learn how data are collected. With these goals achieved, the connections among the other concepts and techniques presented throughout this book—up to and including those in the last chapter—should be clear and easy to grasp.

A course in statistics, like a course in language or music, is highly cumulative. Every idea builds on every other. It is true that if one falls behind in such a course, it is difficult to catch up. Yet it is also true that if one does attain a good grasp of earlier material, there are few "surprises" later on. Every new concept or principle is related to something else learned earlier. It is with this in mind that the chapters in this section are meant to be truly a foundation for what is to follow.

2

Statistics

What It Is and Why Sociologists Use It

You are about to embark on your first course in social statistics, and most likely your first course in any kind of statistics. Based on that assumption, this chapter is designed to provide you with an overview of the portion of the field covered in this book. As mentioned in the introductory chapter, our primary interest is in application: how and why the principles and techniques we discuss are used in research, in planning, in program evaluation, and in many other contexts. This means that we will not be especially concerned about methods for deriving the equations and formulas we use or with certain refinements of the techniques we introduce. Such an emphasis should not suggest that these more technical matters are unimportant, for they are very important in the larger scheme of things. But because a one-semester course cannot do everything, we have decided to stress those aspects that will be immediately relevant to your future studies in sociology, political science, and related fields and, most important, to everyday life.

The applied spirit is reflected in this chapter, which consists of three main sections, beginning with the more general topics and ending with the more specific ones. The first section discusses the term *statistics* itself. Here we see that the term (and the field itself) extends surprisingly far back into history. We also note that, even today, it has not just one meaning but several.

The next section focuses on the role played by statistics in the scientific method, with an emphasis on understanding the method as a type of inductive thinking. Induction, the orderly process of going beyond what is immediately observed, is one of the two major functions of statistics in science. It is sometimes referred to as "inferential" or "decision-making" statistics.

We then conclude with a look at the other main function of statistics: description —communicating to others what we observe. In this context, the key elements of descriptive statistics are introduced and defined: units, samples, and variables. These elements determine the specific fields and subject areas of statistical application, such as engineering statistics, biostatistics, and—as in our case—social statistics. Although every field employs the same basic procedures and rules, some use them to understand physical forces, some are more concerned with life processes, and others with human relationships. The examination of variables that begins in this latter section continues in greater detail in Chapter 3.

What Is/Are Statistics?

Perhaps you have read or heard the word *statistics*, been reminded of the word *status* or *state*, and wondered whether there is a connection. The answer is, yes; all three words have the same Latin root, *status*, meaning "how affairs stand." The name of the field that today is known as *statistics* (which has the Greek endings *-ist-* and *-ic*) still refers to that meaning as the science that informs us about the status of situations, phenomena, or events of interest. According to Webster's *International Dictionary*, 2nd ed., statistics is "(1) a science dealing with the collection, analysis, presentation, and interpretation of masses of numerical data." That, I think, is a fair catalog of the things you will be learning about this semester: collection, analysis, presentation, and interpretation.

Of course, understanding the word *state* in its political sense, the connection might also suggest that statisticians are especially concerned with government data.[1] This is certainly true to an extent, because the kinds of data collected by governments, such as census findings, are especially appropriate for statistical treatment. In fact, we will collect and analyze data from the United States and other national censuses at several points in the following chapters. However, this connection is after the fact, so to speak; it is not part of the original meaning of statistics, which is far broader and could apply to any kind of data.

You may also have wondered why the word *statistics* is always in plural form, like the names of other sciences such as economics, physics, and mechanics. In all of these cases, the plural is a carryover from Greek, in which physics (*ta physika*) originally meant "those things that pertain to material nature" (*physis*). Then, you might have wondered, is there such a thing as *a* statistic? Here, too, the answer is, yes. The word *statistic* in the singular (formed from the plural in what linguists call a "back formation") refers to a single piece of information, one that represents an observed characteristic. As we discuss below, we call the data that we observe collectively a *sample*. Thus, a statistic is a measure of a characteristic of a sample. In contrast, we also speak of a *parameter*, which is a characteristic of a larger whole (which we generally do not observe) from which a sample is drawn: a population or universe. Our dictionary makes this point as follows: A statistic is "(3b) a quantity that describes a sample and is thus an estimate of a parameter of a population."

One last point on word origins: Despite the opinions that are commonly circulated on university campuses, there is absolutely no connection between the words *statistics* and *sadistics*. Hopefully, you will be convinced of this by the time you finish this book.

The Place of Statistics in the Scientific Method

We have all heard of the scientific method, and most of us have used it, for instance in courses such as biology and physics. From such knowledge and experience we are aware that it is a systematic way of establishing facts based on observation, and it can be applied in many different contexts in a wide variety of disciplines. What may not be as obvious is that, in applying the scientific method, statistical techniques and reasoning play a major role, descriptively and inductively.

The scientific method has been the subject of much discussion among scientists, philosophers, other researchers, and members of the general public. Numerous books have been written about it (see, for example, Hoover and Donovan, 2004), and most science textbooks feature an overview of it in earlier chapters. Yet, despite the large volume of information that is available, the essence of the scientific method can be easily stated in two words: *testing hypotheses*. Now, all we need to know is what "testing" and "hypotheses" mean. Unfortunately this will require more than just two words.

Hypotheses and Theories

Nearly everyone who has ever used the term knows that a hypothesis[2] is an educated guess. Thus, when we hypothe*size* we are guessing about some event or phenomenon; we do not know whether our statement is actually true. On the other hand, it is not a wild guess that is made with absolutely no knowledge about the situation of interest. It is "educated" in the sense that we believe we are right, based upon previously acquired information.

Suppose, for example, that you and a friend are playing a simple game in which your friend puts a coin in a pocket, either the left or the right, in such a way that you cannot see where it has been hidden. Then you try to state the outcome, with a reward given if you turn out to be correct. If you have no idea about which pocket the friend will use, you will select one of the possible outcomes—let's suppose the left pocket—for no particular reason. This is not so much a hypothesis as it is a "stab in the dark." Now, suppose instead that you have played this game before, and you have noted that your friend, who is right handed, puts the coin in the right pocket far more often than in the left one. Knowing this, you state the outcome that would seem to be part of the pattern: that is, the right pocket. *This* is a hypothesis. It is still a guess, and you could be wrong; but under the circumstances it is the most reasonable guess.

Although the example is highly simplified, it is exactly how hypotheses are formulated in scientific research. They are statements scientists use to anticipate what they believe would be the case in a particular situation, if things operate as they have in the past. The previously acquired information that makes their guesses educated could come from any number of sources. Ideally, however, such information comes from a carefully accumulated and well-organized body of knowledge known as a scientific *theory*. Now it may seem strange to refer to a kind of knowledge as a theory, because in ordinary language we often think of knowledge and theory as opposite concepts. We might say to people who are expressing what they think is true something like "Well, that's just theory; what are the facts?" In scientific usage, however, theories contain facts.

To put the matter somewhat formally, a theory is a linguistic entity; it is in the world of language and not "out there" in the world of objects and events. A theory *refers to* objects and events, but it exists only in the minds of people and/or on paper—as books, articles, and so on—in the form of words and sentences. It was indicated above that a theory is "well organized." To be more specific, the sentences that make up a theory are logically organized. Rather than being a collection of statements piled up together, so to speak, they are logically related to one another by the five basic logical operators: "not," "and/but," "and/or," "implication (if . . . then)," and "identity (is the same as)."[3]

Some of the sentences of a theory refer purely to how words are and are not to be employed. Their truth is based on the proper use of deductive logic, not on any characteristics of the object world. An important type of such sentences is a definition. For example, a theory about urbanization might stipulate, "A city plus its suburbs is identical to a metropolitan region." This definition makes no claim about what is the case in reality; it merely indicates that the terms "metropolitan region" and "city and its suburbs" may be used interchangeably without affecting the truth or falsity of sentences in which they are used. Similarly, the phrase "a city is not a suburb" indicates that the terms may *not* be used interchangeably.

In addition to such logical truths, a theory also must contain some "empirical" sentences. These *do* refer to the object world, and they might be true or false depending on what is the case in reality. For instance, one might say, "Boston has fewer suburbs than Chicago." Now whether or not this is a fact is determined not by logical criteria but by an actual count of the suburbs of the two cities. The philosopher of science Karl Popper (2002 [1959]) and others have referred to this criterion as "falsifiability." By this it is meant that the possibility always exists that such a sentence can be proved false; note that this cannot apply to a nonempirical sentence such as "A city plus its suburbs is identical to a metropolitan region." There is one important stipulation in this regard: to be a proper part of a theory an empirical sentence must be *true as far as we know*. That is, it surely might be false, but so far it has proved to be otherwise. It is for this reason that we speak of theories as bodies of *knowledge*—they reflect what is known (at a given moment).

One final, and essential, aspect of a theory is that some of its empirical sentences must refer to a general category of phenomena or events in the real world. This property can be identified by the use of a qualifying adjective such as "all," "every," "any," or "none." Such a sentence is known by several names, including "categorical," "general principle," "law-like generalization," and, rarely today, "law." The reason that the last name is no longer used is that "law" conveys the sense of an absolute truth that can never be false under any circumstances. Of course, from what has just been said about falsifiability, such a sentence could not be part of a theory. Like all empirical sentences in a theory, categoricals must be true as far as we know, but they must also be capable of being proved false under some conceivable circumstances.

This is one of the most important and most widely misunderstood characteristics of the scientific method. The only place where absolute certainty is found in theories is in their logical statements such as definitions. Any other statement—one that would refer to the object world, including even our most firmly established "laws"—is subject to doubt. As we will see at several points in the following chapters, where truth is sought but cannot be absolute, statistics has a major role to play.

Having noted this about theories, it might be clearer why they are viewed as the ideal source for producing hypotheses. When researchers want to learn more about some aspects of urban life, for example, they are wise to turn to the body of urban theory that summarizes what is known, provides logical connections between these facts, and offers generalizations drawn from previous research. In practice, this is easier said than done. Ordinarily, especially in the social sciences, theories of urbanization or of anything else are not to be found in just one book or research report. Instead, relevant theories and parts of theories may be scattered among several different sources (often in several languages). It is for this reason that scientific research typically begins with what is called a *literature search*. This is the activity of locating and examining potentially fruitful sources to discover what is already known about the object of study. This can be quite time-consuming and often frustrating. But because theory is such a powerful part of the research act, it is usually worth the effort. One of the Solution-Centered Applications at the end of this chapter illustrates the procedure of a literature search.

The close ties between theories and hypotheses also underscore the fact that hypotheses, too, are part of the world of language. In fact they are sentences, declarative sentences. You will recall that there are three types of sentences in this respect: declarative, interrogative, and exclamatory. The first type, which includes hypotheses, makes a statement: This is the case. Something is true, etc. The second type, the interrogative, poses a question: Is this the case? Is something true? The exclamatory type of sentence expresses a strong observation, usually with an exclamation mark to stress the point: This book is easy to read! I have found it (*Eureka*)! In indicating that a hypothesis is an educated guess expressed as a declarative sentence, we are saying that it states what is believed to be true: "My boat will float." It does not pose a question: "Will my boat float?" Thus, we speak of a hypothesis as being *accepted* if true and *rejected* if false, rather than saying "yes" or "no" in answer to a question. The final point in connection with seeing hypotheses as sentences is that, like all declarative sentences, they have a subject (noun part) and a predicate (verb part). We return to this feature below.

To summarize this introduction to hypotheses, let us continue with the example of urbanization research and illustrate the main point of our discussion. Suppose that you are interested in the relationship between crime and urbanization. In the course of your studies you find some information about a city (let's call it City D) which apparently has not been included in investigations of the type you are conducting. You learn that it is a large (one million plus) city in the northern industrial, "rustbelt" region of the United States. You are interested in the crime situation in D, and with this in view you refer to existing theories on crime in cities in D's region. You search the literature and find included in the theoretical material a generalization to this effect: "All large cities in the rustbelt have crime rates higher than the national average." With this knowledge in hand, we derive a hypothesis by deductive logic, as follows.

- All large rustbelt cities have higher than average crime rates.
- D is a large rustbelt city.
- ∴ D has a higher than average crime rate.

In the third statement is the hypothesis; the "∴" symbol means "therefore," indicating that you have deduced the hypothesis.

The claim that D has a high crime rate is just that, a claim or guess. You do not at this point *know* that it is true. But you do know that it *ought to be* true if the theory from which you took the categorical statement is sound and if your observation that D is large and in the rustbelt corresponds to the actual situation. The next step is to decide whether to accept the hypothesis or reject it. This procedure is what we referred to in our two-word characterization of the scientific method as "testing," the subject to which we now turn.

Testing Our Expectations

In the formal sense in which we have introduced the concept, a hypothesis is neither true nor false. Rather, it is a statement of a likelihood that is awaiting confirmation. In principle, there are several possible procedures that might be used to help in deciding whether or not a hypothesis is true. For example, the question might be referred to recognized authorities in the field of interest: urban studies, in our example. Those people might then declare that, based upon their knowledge, D does have a high crime rate. Another possibility is to seek to establish what ordinary people have always understood to be the situation: that is, traditional knowledge. If, for example, residents and nonresidents alike have consistently learned from their parents and grandparents that D does *not* have a serious crime problem, then this information would be used to support a decision to reject our hypothesis.

One or both of these bases for determining the truth value of hypotheses, authority and tradition, have been used throughout human history, and in many instances they are still used to this day. However, neither is considered acceptable in contemporary scientific work. Instead, as an aspect of scientific method, the only legitimate means of testing hypotheses is to compare the stated, expected outcome with observable data. In our illustration, this would mean counting the number of crimes committed in D, dividing that number by the population size, and determining whether the quotient is above or below the national average. If it is above, then the hypothesis would be accepted; otherwise, it would be rejected. A result such as this, based upon observable, empirical data, would override both the opinions of authorities and traditional knowledge, if either differed—assuming, of course, our observations and calculations were accurate. Thus, *to test a hypothesis one compares the expectation it states to relevant observable data.*

The rule that only observable data have the "power" to establish the truth or falsity of hypotheses is of primary importance in science generally and in statistical applications in sociology in particular. As noted earlier, when hypotheses are derived from theories, we expect them to be true because of the way theories are constructed. However, the fact that a hypothesis follows logically from a well-established base of knowledge cannot guarantee that it will be accepted when it is tested. If, on the basis of testing, a hypothesis is accepted, then we can conclude that the theory remains sound. But if a hypothesis is rejected—which is often a more interesting outcome of research—then a series of questions must be asked and answered.

- Have the observations, calculations, and other aspects of the testing situation been conducted correctly? If not, then the associated procedures need to be improved and repeated. If so, then
- Were the logical steps taken in deducing the hypothesis followed properly? If not, then the procedure should be corrected and repeated. If so, then
- Were the theoretical principles used in the deductive operation correctly understood and interpreted? If not, then the principles need to be more carefully examined. If so, then the theory has been weakened. New principles need to be established to cover the new, disconfirming evidence.

It is obvious from the foregoing procedure that the act of comparing observations with expectations is far from simple. There are many points at which errors can be made or when our fallible human judgment can lead us astray. One important tool in this process involves the organization of the data we observe in such a way that it becomes informative: clear, unambiguous, and capable of assisting in the process of deciding whether or not to accept a hypothesis. This tool or set of tools for turning observed data into usable information is the concern of descriptive statistics, featured in Part II of this book: Chapters 4 through 6.

Assuming that our observations have been described effectively, we might still wonder how one could be sure that the decision the researcher makes to accept or to reject a hypothesis is the correct one. This very important question is easy to answer: one never can be *sure*. The scientific method is the most effective means for establishing the truth about many things. But the truths it reveals cannot be absolute; they are always subject to doubt because of the complexity of the hypothesis-testing situation. The best that can be said about the results of scientific research is that we are "fairly confident," "very confident," or even "extremely confident" that we are right—but we know that we always could be mistaken. This is reflected in the statistician's use of the term *confidence interval*, as discussed in Chapter 9.

This feature of scientific research may at first seem to be a disadvantage. After all, we would prefer to be certain about our observations and conclusions—especially in social science, when human welfare and even human life might be at stake. Certainty *is* possible in human experience, but only in a subjective sense such as one's faith in God or in the sense of valid deductions. In the latter case, we can be certain that all bachelors are unmarried men; that all unicorns have one horn; and that (1) if all fruit is nourishing, and (2) a banana is a fruit, then (3) a banana is nourishing. The reason we can be certain is that the truth of these statements depends not on observable facts but on the rules of logic. They would be true even if there were no such thing as a bachelor, a unicorn, or a banana. However, once we enter the realm of empirical sentences (which all scientific theories must include), and thus the realm of *inductive* logic, even our most well-established beliefs are still beliefs and subject to disproof.

This situation does have important advantages. Because scientific knowledge is always subject to doubt, we can never be content with yesterday's truths. Instead, the scientist is encouraged to question and challenge the taken-for-granted world, to examine critically what "everyone knows." This is why the sociologist Robert K. Merton (1968) referred to science as "organized skepticism." The scientific method organizes our thinking, and it incorporates the "prove it to me" attitude by calling for fresh observations and new interpretations. When this skepticism fails to refute current knowledge, then such knowledge is reinforced and we are more confident to act on it. But when the doubting and questioning is successful, authentic progress has been achieved.

Were it not for the evidence and argument marshaled by Copernicus and Galileo, we would still be content with the "obvious" understanding that sunset and sunrise are the result of the sun's movement around the earth. Were it not for the findings of Einstein and fellow twentieth-century physicists, we would still believe that subatomic particles and galaxies naturally behaved according to Newton's "laws."

Understanding that one can never be absolutely certain that the decisions to accept or reject hypotheses are correct, scientists therefore depend upon a range of statistical techniques that can at least indicate the *probability* or chances that they have made the right decision. This is where the familiar "margin of error" or "plus or minus 5%" comes from in reports of public opinion polls and the results of other kinds of research. One might think of this parallel: Whereas the rules of deductive logic are used to determine deductive truths, the rules of statistics are used to determine scientific truths. Part III of this book, Chapters 7 through 9, focuses specifically on the principles and techniques of statistical induction. There we will explore in detail such issues as the rules of probability, confidence in our findings, and the errors that are made in accepting or rejecting hypotheses. We will also have the opportunity to elaborate about the place of statistics in the scientific method. Until reaching those chapters, however, we will place such concerns in the background in order to present some fundamental, preliminary topics.

Thus, to review what we have said thus far, the heart of the scientific method is testing expectations, which are derived from theories about an aspect of the world. These expectations, or educated guesses, are formally stated as declarative sentences called hypotheses. These hypotheses are compared to observations. The data that are observed are then organized so that they become information. Based on this information it is determined whether the hypotheses should be accepted as true or rejected as false.

One final point before moving on: You will recall that this procedure has been termed organized skepticism, because any observation or any conclusion based on observation is subject to doubt. In order to ensure that this skeptical orientation is scrupulously maintained, researchers ordinarily do not even overtly attempt to prove their expectations; for this might introduce bias in favor of accepting hypotheses when they should be rejected. Instead, scientists state their expectations in a *negative* form. For example, instead of comparing observations to the sentence "D has a high crime rate," scientists would substitute the sentence "D does *not* have a high crime rate." The former is referred to as the *research* hypothesis because that is what the researcher believes, and the latter as the *null* hypothesis because it is *not* the researcher's expectation. Therefore, in practice it is the null hypothesis that is either accepted or rejected. Researchers then exert maximum effort; they "bend over backwards," to accept the null. But, if this proves all but impossible, then they must, like true skeptics, reluctantly conclude that they were right (until shown otherwise). We return to the role of null hypotheses in research in Part III.

Describing What We Observe: Units, Samples, and Variables

Like all sentences, hypotheses have two main parts: (1) the subject, which states what the sentence is about, and (2) the predicate, which states the action or quality associated with the subject. The subject of a scientific hypothesis is the word or words that refer to the unit of analysis or observation (let us say *unit* for the sake of brevity).[4] In the