CHAPTER

ONE

Statistics and Communication Science

1.1 Welcome

Let me be the first to welcome you to the exciting world of statistics and data analysis. Statistics is a way of organizing, describing, and making inferences from data, and statistical methods are used throughout the physical, natural, and social sciences, including the field of communication. Statistics is a way of thinking, a language, and a means of making an argument based on data available (Abelson, 1995). Most importantly, statistics is fundamental to the scientific process. It may seem a strange way of thinking at first, but with enough perseverance and practice, thinking statistically will eventually become second nature to you. Once you have developed the ability to understand and apply statistical principles and concepts to your scientific investigations, you will find that your everyday thinking has changed as well. You will find yourself more analytical, more rational, and you may even possess a new and healthy skepticism when it comes to interpreting information and evaluating claims people make. Furthermore, you will be able to participate in the exciting world of knowledge generation that is the field of communication. So again, welcome.

The field of communication has not always been recognized as a leader in the use of statistical methods, but any observer of the communication literature would recognize statistics as fundamental to the discipline by spending only 10 minutes or so looking through the major journals in the field. The majority of the articles published in such places as *Communication Research*, *Human Communication Research*, the *Journal of Communication*, the *Journal of Broadcasting and Electronic Media*, and *Media Psychology* read much like the research articles found in other social sciences, with detailed method and results sections where the research design and data analysis procedures the researcher used to collect and analyze the data are described. Several communication programs, such as mine at The Ohio State University, are appropriately located in the same administrative colleges as other disciplines that use statistical methods reside, such as psychology, sociology, and political science. And it is common, more so now than in the past, for graduate programs in communication to require students to take a few and often several data analysis courses.

There are at least two mutually interdependent components of the scientific process. The first component is research design—the various approaches you might take to collecting the data to answer your question of interest. You may have taken a research methods course at some point in your education. If so, the course probably focused primarily on how to collect data and the various categories of research designs, such as surveys, experiments, content analysis, archival research, observational studies, interviewing, and so forth. The second component is the process of data analysis—how to evaluate the data with respect to the original question that motivated the research in the first place. As you work through this book, you will come to appreciate the interdependence of design and analysis. In the field of statistics there is a saying: "garbage in, garbage out." This means that even the best data analyst can't take data from a poorly designed study and produce something useful or meaningful. What you put in is what you get out. It is important before embarking on any study to carefully think through what questions you want your study to help you answer. Before collecting the data, you have to constantly keep sight of why you are doing the study and how to best conduct the study to attain those objectives. The questions you are attempting to answer will determine how your study is designed. During the design phase, you need to think about how others might criticize your study and how you can preempt such criticisms with a better design. If your study isn't designed well, no amount of data analysis can turn badly collected data into something meaningful and interpretable with respect to your original question.

Good design is extremely important to the research process. However, the focus of this book is squarely on the data analysis aspect of science. This is a book about statistics, and as such, it focuses on statistical methods for describing and making inferences from data. Along the way, there will be the occasional discussion of design issues, as design and analysis are intertwined. Chapter 3, for instance, discusses sampling strategies—ways of recruiting "units" for your study, whether those units are people, newspaper columns, advertisements, or whatever. This is a design issue, but it is also a statistical one in the sense that our method of recruiting research units affects the kinds of inferences we can make with statistics. Although I cannot help but talk about research design, I focus primarily on data analysis in this book.

If you approach statistics as an exercise in mathematics rather than as an integrated part of the scientific process, you probably will ever understand just why a background in statistics is not only helpful but essential to success as a communication scientist. But before diving head first into the topic of statistics, it is worthwhile to first do a brief overview of just what science is, how communication scientists use the scientific method, and to make explicit the assumptions that communication scientists make when going about their business.

1.2 Why Do Science?

There are as many definitions of science as there are scientists. When I use the term *science* I am referring to something very explicit, but at the same time a bit nebulous. When you think of science, your mind probably conjures up images of people wandering around laboratories in white coats, looking through microscopes, mixing chemicals, filling test tubes, or peering through a telescope at the stars. This is how the media portrays science. Indeed, if you were to read the science section of any newspaper, most likely it would be filled with stories of biology, medicine, chemistry, geology, and astronomy. But to the scientist, science is more a way of thinking about and answering questions rather than a specific discipline, like chemistry or physics or biology. Science is more a *process* than a thing. Someone who is engaging in science is using a particular set of methods, based on a common philosophy about knowledge and discovery, to answer questions of interest to him or her or the scientific community.

But not all disciplines are scientific ones, even if the discipline does focus on the discovery of knowledge and the answering of questions. For instance, philosophy and history are two fields of study not traditionally known as relying on the scientific method even though understanding and discovery are a part of these fields. Other disciplines do share this common method called science (or the "scientific method"). Stereotypically, we think of only the hard or natural sciences as scientific—chemistry, biology, physics, and astronomy, for instance. But there are also the *social sciences*—sociology, psychology, economics, and communication, for example—all of which use the scientific method as a means of discovery and exploration.

I've already described science as a process rather than a thing. I've also described it as a way of thinking and as a collection of methods for answering questions. These are all true. Others define science differently. Looking through the books I have conveniently available in my office, science has been defined as "a process of disciplined inquiry" (Liebert & Liebert, 1995, p. 3) or as "the development of knowledge through a combination of rationalism and empiricism" (Stewart, 2001, p. 4) where rationalism is the use of logic and empiricism is the use of observation and the senses. Both definitions suggest that science is a method of discovery guided by rules. But what are we trying to discover? What is the purpose of doing science?

Science as Problem Solving. One purpose of science is to try to solve some problem of relevance to the world. That world need not be the entire world, in the sense of the planet Earth. That "world" may be a very confined one—a business, for example, or it may be larger, such as a city, state, or country. Communication scientists often do scientific research as a means to solving some problem of relevance to our daily lives and the world in which we live. For example, AIDS can be thought of as a problem of global proportions, and it is clearly a problem with behavioral origins. A large proportion of infections occur through unprotected sex. If we want to reduce or eliminate human suffering resulting from the proliferation of AIDS, what needs to be done? Scientific research can be helpful here by helping us to discover the sources of the behavior (e.g., lack of knowledge, feelings of invulnerability) and evaluate methods to change that behavior (e.g., making condoms conveniently available, educating people on the causes of the disease, informing them how to protect themselves, and so forth). Scientific research that focuses on some kind of real world problem is sometimes called *applied research.*

Science as Information Acquisition. We often engage in scientific discovery for the purpose of information acquisition. Information is important in all aspects of life, including the quest to understand and explain communication processes and to influence behavior. Information is also very useful or even necessary before deciding on a course of action. Suppose, for example, that you are running for a political office and are developing a campaign strategy. What kind of information would be helpful in designing your campaign? You could just campaign on the issues that you find important, but this strategy probably won't get you elected. You need to know what issues are important to your potential constituents. Science could help you figure out what your platform should be so potential voters will pay attention to your message. Or imagine you work for a public relations firm trying to help a fast food chain capture greater market share. Why do some people prefer the competition, and what can you do to get people through the doors of your client's chain and enjoy the experience, thus enhancing the likelihood they will return? You might want to know if people have had bad experiences at your client's chain, whether the prices are perceived to be lower at the competition, or whether the physical environment of your client's chain is unappealing or unwelcoming. The scientific method could help you answer these questions, thus arming you with information about what your client needs to change.

Science as the Development and Testing of Theory. Most disciplines have a large body of theories that researchers in the discipline study. Just like there are many definitions of science, there are many definitions of theory. A *theory* can be thought of as an explanation of a process or phenomenon. There are many scientific theories, some of which you have undoubtedly heard about and equate with "science," such as Einstein's theory of relativity or Darwin's theory of natural selection. Communication has many theories too, such as *communication accommodation theory* (e.g., Giles, Mulac, Bradac, & Johnson, 1987), *spiral of silence theory* (Noelle-Neumann, 1993), and the *elaboration likelihood model of persuasion* (Petty & Cacioppo, 1986). Entire books are devoted to nothing but a description and discussion of communication theories (e.g., Littlejohn, 2001; McFleur & Ball-Rokeach, 1989; Severin & Tankard, 2001), illustrating that communication scientists are busy trying to understand and explain the communication process. These theories may be based on lots of prior research, or they may be based on the theorist's intuition, personal experiences, or the application of rules of logic. Regardless, theories are only attempts at explaining. They may not adequately explain the phenomenon or process. We use science to test the validity of those theories, to see if predictions the theory makes actually work out when they are put to the test. And we do science to determine how theories should be modified to better describe and explain the process under investigation. Research motivated by the testing and development of theory is sometimes called *basic research*. However, this does not mean that basic research is never guided by applied concerns, nor does it mean that applied research is never theoretical. People who conduct basic research often are very focused on eventual application but may feel that their understanding of the process is not sufficiently developed for them to apply the theory in a particular context of interest.

Consider the *elaboration likelihood model of persuasion*, which predicts that "peripheral cues" to persuasion such as a messenger being a well-liked celebrity should have a greater effect on people who are uninvolved in or don't care much about the topic compared with those who are more involved. If a researcher found that the same message when delivered by a celebrity was more persuasive than when it was presented by someone who was not a celebrity, but only for people who didn't care about the topic, then this gives support to the theory—it has passed at least one test. If the data are not consistent with the theory's prediction, and assuming the study was well designed, this is a strike against the theory. If the theory fails repeatedly when put to the test, this suggests that theory is invalid. A theory can't be an adequate explanation of the process if predictions it makes about what a researcher should find in a study designed to test it are rarely or never right. Indeed, it has been said that falsification through disconfirming evidence is the only way that theories can be tested, because a confirmation of a theory's prediction is only suggestive of the accuracy of the theory. There could be (and perhaps is) some other theory that would make the same prediction. However, if data aren't consistent with a theory, this is strong evidence against the validity of the theory.

Chaffee and Berger (1987) discuss in considerable detail the important role that theory development and testing plays in the life of the communication scientist. In the end, we can all create theories, and we routinely develop our own intuitive theories as we manage our day-to-day social affairs, but that doesn't mean that these theories are good. Researchers in a discipline are in charge of evaluating the many theories that attempt to explain some process or phenomenon of interest to the discipline, and they do this using the scientific method. If you are a graduate student, at least some of your time as a student will likely be spent learning about, testing, and evaluating theories through the methods of science.

Science as the Satiation of Curiosity. We also do science simply because we are curious. Curious people are naturally drawn to science because science provides people with a systematic method for answering their own questions. If you are someone who likes to wonder or invent your own explanations or theories for something, then you will enjoy science. Few things are more exciting than answering your own questions through the scientific method. But such pleasures usually are short lived because curious people have many curiosities and there is always some new curiosity to be satiated, some new question to answer.

I've presented these uses of science as if they are nonoverlapping and unrelated, but a researcher may use science for any of these reasons in combination. For example, a researcher may have a natural curiosity in persuasion and is perhaps interested in applying his or her study of persuasion to the development of information campaigns focused on improving childhood health. But just where are the health problems? Are children getting sufficient vitamins at critical stages of development? Are vaccines available and being used? What deters a parent from vaccinating his or her child? Answers to these questions are important before an information campaign can be developed. There are many theories of persuasion, and the researcher may not know which would be most helpful or maximally effective in a particular situation. So the researcher may conduct various tests of the different theories by focusing on issues relating to childhood health. If a theory doesn't make accurate predictions in this context, it probably shouldn't be used when developing an information campaign in that context.

1.3 Assumptions and Philosophies of Scientific Investigation

There are many disciplines that use the scientific method as a means of answering questions, and each discipline tailors the scientific method to suit the special issues and problems conducting research in that discipline. However, users of the scientific method, including communication scientists, make a number of common assumptions when they apply it to their field of study.

The World is Orderly. Scientists are in search of the logic and orderliness of the world. That is, we assume that there are rules or laws governing human behavior and thought. If there were no laws of human behavior, studying human behavior scientifically would be pointless because every situation would lead people to respond, think, or act differently. There would be nothing about the "human condition" to explain, discover, or understand. And there would be no way of systematically attempting to influence or change human behavior because each person's thinking and behavior would be guided not by some common processes or principles but instead by their own whims and idiosyncrasies. Social scientists, such as communication scholars, believe that there are laws of human behavior to be discovered even if they don't always apply to everyone in every situation. These laws are propensities to think, feel, or act in certain ways.

For example, it is well accepted that similarity breeds attraction (Byrne, 1971). We tend to like and be attracted to people that are similar to us. This relationship can be thought of as a behavioral law because it is such a consistent finding in the literature on human attraction. To be sure, laws may have boundary conditions, meaning that they may apply only to certain people in certain situations. We know that in some circumstances dissimilarity breeds attraction, but such circumstances are rare. The important point is that if we didn't believe that there are such laws to be discovered, then there would be no point in doing research in communication. Communication scientists generally believe that such laws exist and can be discovered and described with the methods of science.

Empiricism. Researchers who use the scientific method believe that research should be based on observation that is objective and replicable. Science largely rejects subjective data that are not visible or replicable, such as anecdotes, rumors, or other data that can't be publicly observed and verified. This does not mean that what we study must be directly observable. Communication scientists study things that you can't actually see. Communication anxiety, for instance, is a widely studied construct that you can't actually see directly. But you can indirectly study it by examining observable characteristics of communication anxiety such as how a person responds to a question about his or her worries about speaking in public, or how that person acts when communicating. Furthermore, the methods we use must be replicable. Someone else should be able to conduct the study exactly as we have. To replicate a study, it is important that we clearly describe how the study was conducted and the data analyzed so that others can attempt to replicate not only our methods but hopefully our findings too.

Related to the assumption of empiricism is the importance of *measurement* or *measurability.* To be able to study something, we have to be able to measure it, usually in some quantitative form (although there is some debate about whether it must be quantified). If something can't be measured or, at minimum, categorized, then we can't study it with the scientific method. However, nearly anything of interest can be measured, and so nearly everything can be studied through the scientific method.

This empiricist philosophy when combined with the assumption that the world is orderly is sometimes known as *positivism.* Positivism is akin to the belief that "the truth is out there," and that anyone can discover that truth if they approach it objectively, using a systematic set of replicable methods that can be communicated to other researchers. Nevertheless, not everyone believes that there is a single truth to be known. Some researchers believe that truth is subjective and in the eye of the perceiver (rather than some reality to be observed objectively). According to this philosophy, the conclusions one reaches after observation depend on who is doing the observing. Two people observing the same phenomenon may have very different interpretations of that "reality," and to really understand what makes a person think and act as he or she does, we must understand that person's social environment and how he or she interprets the world. As such, reality is subjective and cannot be known in the same way that a positivist believes is possible. Researchers that reject empiricism or positivism tend to use a different set of methods than do strict positivists and approach knowledge and discovery very differently. But a strong positivist would argue that any approach to knowledge acquisition that doesn't allow the object of study to be perceived objectively and the "truth" to be discovered is pseudoscientific at best. As you can imagine, for researchers in one camp or the other this can be a very emotional issue.

Parsimony. Also known as *Occam's Razor*, the rule of parsimony states that scientists should not evoke explanations for a phenomenon that are more complicated than need be to adequately describe it. In other words, simpler is better. When two competing explanations are used to describe the same thing and can do so equally effectively, the one that is simpler and that makes fewer assumptions is the one that should be preferred. That does not mean that the simpler explanation is always the correct one. It merely means that when two explanations that differ in complexity are equally consistent with the data, the simpler one is to be preferred until new data are available to contradict the simpler explanation and that favor the more complicated one.

Progression in Small Steps. A researcher who believes in the importance of the scientific method is also a modest one. No single research study is ever the definitive one. The development of knowledge through the scientific method is a slow process that progresses in small steps. If you were working on a puzzle, each piece that you fit into its proper place gets you closer to the picture. The result of a single study is like a single piece of the puzzle. However, unlike is true when working on a puzzle, each study often raises new questions, making the puzzle even bigger. Imagine what it would be like if each piece you correctly fit into a puzzle increased the size and complexity of the picture. But that is a little what science is like!

The Nonexistence of Proof. If you ever hear someone say "a study was done that proved" such and such, you know that person isn't familiar with the doctrines of science because no scientist would ever use that word to describe the state of any knowledge we have (at least not in the company of other scientists). No single study is itself especially revealing. More important is what the collection of studies on a similar topic says because a single study is open to multiple interpretations and always has some limitations. Our knowledge is cumulative and evolving. What we know today may not be true tomorrow because new data are always coming in, and there are always alternative explanations for something that can't be categorically ruled out. And just because we can't think of an alternative explanation doesn't mean some other explanation doesn't exist. A theory that is widely accepted today may not be accepted tomorrow. So nothing is ever proven. The things we believe to be true we believe only because the data seem to compel us to accept them as true, but we shouldn't get too comfortable with that truth because it may change as more data become available. So we talk about theories or explanations being *supported* by the data or that the data *suggest* the correctness of such and such explanation. Our beliefs are always tentative ones that we hold until the day that some data lead us to reject those beliefs in favor of something else. That day may never come in our own life time, but that doesn't mean the day will never come.

The fact that proof doesn't exist in science is important to keep in mind when analyzing data and interpreting the results of an analysis. It is all too easy to assume that the statistics tell the story objectively and, because statistics cannot lie (although users of statistics can), the proof of one's claim is to be found in the numbers. Such an assumption places far more importance on statistics than is justified. Any study result, the statistics included, has only limited meaning in the context of a single investigation. There usually are many different ways of quantifying the results of a study, some of which may be better than others. Someone who analyzed your data differently might come away with a very different interpretation or conclusion. Statistics is an area that, like any area of science, is controversial. Professional statisticians disagree on such seemingly uncontroversial matters such as how to best quantify "average" and "variability," and yet we use these concepts routinely in communication science. There are many different ways of analyzing data, and your interpretation of the results of your research must be done in light of this fact. Indeed, throughout this book you will see there are many different statistical tests that have been proposed to answer the same statistical question, and it isn't always clear which test is best in a particular circumstance. What is important in the scheme of science is consistency. Do many studies, conducted using different methods and using different approaches to data analysis all converge on a similar result or interpretation? To the extent that the answer to this question is yes, we can be more confident in our interpretation of the corpus of studies on a topic. But we can never say with 100% confidence that anything has been proven.

Falsifiability. An explanation for some phenomenon must be falsifiable, which means, therefore, that it must be testable and it must be possible for evidence inconsistent with the explanation to exist. Not all theories or claims are falsifiable. One example in communication is McCluhan's phrase "the medium is the message." McCluhan appreciated that the phrase is ambiguous and could have multiple interpretations across people and over time (Sparks, 2002). As such, it is very difficult to falsify and so it is very difficult to study whether there is any "truth" to this phrase using the scientific method. Similarly, could we ever falsify the well accepted claim that "We cannot not communicate"? As alluring and obvious this "fact" might seem, I'm not sure how a study to discredit such a claim could even be undertaken. If it cannot be discredited, if no data inconsistent with the claim could ever be produced, then it is not falsifiable and not in the realm of things amenable to scientific investigation.

1.4 Building Your Statistical Vocabulary

Statistics is an important part of the scientific process because it is through statistics that we extract information from our application of the scientific method to a research problem. As we will discuss later, we use statistics both to describe what we found when conducting a study and to make inferences from the data. Unfortunately, statistics is a topic that often scares the burgeoning scientist. When someone takes his or her first statistics course, sometimes anxiety about mathematics interferes with the ultimate goal of conceptual understanding. This is based on the belief that statistics is about numbers and math. But it is not. Statistics is more like a language; it is a means of communicating ideas and evidence. To understand how to use statistics in science, you need to grasp the concepts as well as the vocabulary used to discuss those concepts. Computers are used to do much of the computational work, so one's initial exposure to statistics is best focused on understanding how statistical procedures are used and getting a handle on the vocabulary. Formulas are always available in printed form, here and elsewhere, and you will rarely need to do computations by hand. But computers are limited in their role as number crunchers. You, the user of statistics, need to understand the concepts and the vocabulary to truly master statistics and interpret what a computer is telling you. In this section, I begin to define some of the more widely used and important terms that are used in this book and throughout the scientific community.

We typically conduct research to answer some kind of question. As you read the communication literature, you will frequently come across the term *research question*, often denoted symbolically as *RQ*. A research question is simply a question that a researcher poses in an investigation. These questions are often vague and abstract, like theories are, but they are much more limited in scope and rarely if ever attempt to explain a process in the same way that a theory does. Someone studying persuasion might ask the following research questions:

> RQ_1: To what extent does the depth of one's thinking while being exposed to a message influence whether attitude change is short or long-lasting?
>
> RQ_2: Does distraction during presentation affect the persuasiveness of a health-related message?

In contrast to a research question, which is usually a vague statement, a *hypothesis* (sometimes denoted in the literature with the letter H) is a prediction about what the researcher expects to find in a study. Two hypotheses corresponding to the research questions posed above might be

> H_1: The more time a person spends thinking about a message, the longer any attitude change resulting from that message will persist.
>
> H_2: The more tasks the person is given when being exposed to a persuasive message, the less attitude change the message will produce.

Notice that these are much more specific than the research questions because they explicitly state what is being predicted. Hypotheses can take various forms. A *one-tailed* or *directional* hypothesis makes a prediction about the direction of the result expected. The following hypotheses are one-sided or directional:

> H_1: Males will spend more time reading the sports section than females.
>
> H_2: People who read the print form of the news will learn more current events than those who read the online version of the same paper.

Observe that these hypotheses make a prediction about how the groups should differ. The hypotheses predict not only the groups will differ, but that one group will do more of something (such as learn more or read more). The corresponding hypotheses presented in *two-tailed* or *nondirectional* form don't specify the precise direction of the result expected. For example:

> H_1: Males will differ from females in how much time they read the sports section
>
> H_2: There will be a difference between readers of print and online newspapers with respect to how much they know about current events.

Hypothesis testing is an important part of the scientific process because theories often lead to predictions about what an investigator should find if the theory is a good representation of the process. In this book I devote considerable space to how to test a hypothesis using statistics.

When you conduct a research study, at some point you will be collecting *data*. Data refers to some kind of representation, quantitative or otherwise, of a variable or variables in a research study. For instance, if we asked 10 students their grade point average, their answers (such as 3.4, 2.4, 3.93, etc.) constitute our data. If we administered the *Personal Report of Communication Apprehension* (see Rubin, Palmgreen, & Sypher, 2004) to a classroom of tenth graders, their scores on this measure in addition to perhaps information about their age and sex would be our data. If we wanted to know how much violent television a person watches during the typical week, we might ask the person which shows they regularly watch and then count up the number of shows watched that could be classified as violent. Our data would be the number of violent shows each person reported watching. Data need not be quantitative. For example, the sex of the 10 students is qualitative rather than quantitative information. The term can also be used to refer to a collection of statistics or more generically to refer to any kind of evidence. Someone might ask if you have any data to support your claim. You might ask yourself what your data tell you about the process you are studying. A *data set* is simply a collection of data. After data collection, the researcher will typically enter the data into a computer prior to analysis, and this file constitutes the data set.

The term *case* is often used to refer to each unit that provides data to the researcher. The unit may be a person, or it may be a single advertisement in a collection of advertisements being analyzed, or it may be one of several letters to the editor of major newspapers that are being content analyzed. Each of the students in the example above is a case in the data. If you had 100 letters to the editor or a collection of advertisements that you were content analyzing as part of a study, each letter or advertisement would be a case in the data set.

Each case in the data set is typically measured on one or more *variables.* A variable is anything that the units providing data in your study vary or differ from each other on. Take a group of newspapers, for example. On any given Sunday, the front section of the *New York Times* has a certain number of pages. *The Washington Post* may have a different number of pages in the front section that day, and the *Los Angeles Times* may have still a different number. The number of pages varies across newspapers, and thus it can be thought of as a variable. Other variables might be the number of square inches of total space devoted to advertising or how many stories about crime are printed on a given day. The *New York Times* may have had 12 stories last Sunday, whereas the *Washington Post* had only 10. So the number of stories about crime that different newspapers print in a day could be considered a variable.

People are also the unit of study quite frequently in communication research. People vary from each other in a lot of ways, such as sex, religion, or years of formal education. People may have different reactions to an advertisement—some may like the advertisement, others may dislike it, and others may not have even noticed. People may differ with respect to how much they think others support their own opinion on some social topic, such as affirmative action or the death penalty. Some may think they are in a minority and others may think they are in a majority. So perceived support for one's opinion is a variable. Other things can vary between individuals, such as how long a person reads the newspaper each day. Some people may regularly spend two hours with the paper. Others may not read the paper at all. "Variable" is an extremely important term, and we will use it a lot. We will often talk about the relationship between variables in this book. Theories make predictions about how variables should be related to each other. We test hypotheses and theories by looking at whether the relationship we find in a study between two or more variables is in the direction a theory or hypothesis predicts it should be.

In research, we often distinguish between two types of variables. The definitions you see people use vary, so I'll do my best to capture the flavor of all of them. When we conduct research or theorize, we often think of one variable as coming before another in time or sequence or that one variable is some how causally prior to another. That variable, the one that affects something else, that is presumed to be the cause of, is used to predict something, or that in some way "explains" variation in something else is referred to as the *independent variable.* It is also called the *explanatory* or *predictor variable* or sometimes the *exogenous variable.* In contrast, the *dependent variable* is on the other side of this chain. The dependent variable is thought to be the result of some process or affected by something else in the model or theory. Usually, the outcome of some process we are interested in studying is the dependent variable. The dependent variable is also sometimes called the *criterion, outcome,* or *endogeneous* variable.

This sounds like a confusing set of definitions, so I'll make it more concrete with some examples. Arguably, the more you are exposed to news through various media outlets, the more informed you will become. In this case, exposure to the news is the independent variable and knowledge is the dependent variable. We presume, as phrased above, that exposure to news causes an increase in knowledge. And it has been said that managers who give workers the freedom to determine how they do their job tend to be liked more by their employers. If we were studying this process, we would think of how much freedom the manager gives the employees as the independent variable, and the employees' liking of the manager as the dependent variable. But sometimes it doesn't matter which variable is the dependent variable and which is the independent variable. For example, you may be interested in determining whether people who fear public presentations tend to be anxious in the course of their daily lives. You may have no particular interest in claiming that being anxious causes fear of public presentations or that fear of giving public speeches leads to general anxiety, nor may you be trying to make the case that there is some causal relationship between them. Therefore, it doesn't matter which is conceptualized as the independent variable and which is defined as the dependent variable.

In a single research study, a variable can serve as both an independent and a dependent variable. For example, the *knowledge gap* refers to the tendency for people with low socioeconomic status to be less informed about things going on in the community or world than those higher in socioeconomic status (Tichenor, Donohue, & Olien, 1970). It could be argued that relatively poor people tend not to get exposed to the media, and because the media is a major source of information, poorer people tend not to get as much information about the community or current events in the world as people with more money. In this process, media exposure is both an independent and a dependent variable. It is a dependent variable in the relationship between socioeconomic status and media exposure (being poor leads to less exposure), and an independent variable in the relationship between exposure and knowledge (knowledge results from exposure to media). This example also illustrates what we will call a *mediating variable* in Chapter 15. Low socioeconomic status causes people to be less exposed to media, and as a result of that low exposure, people low in socioeconomic status tend to know less. Therefore, media exposure is a mediating variable in the relationship between socioeconomic status and knowledge.

Another kind of variable is the *moderating* or *moderator variable*, discussed in more detail in Chapter 16. A moderating variable is one that determines or is related to the size, strength, or direction of the relationship between two variables. For example, suppose men who play sports more often than most other men also spend more time than those other men reading the sports section of the newspaper on the weekend. But suppose there is no relationship between how often women play sports and how much time they spend reading the sports section. So the relationship between sports activity and sports reading is different in men compared to women. As such, we say that gender *moderates* the relationship between sports activity and how frequently a person reads the sports section, and so gender is a moderator variable. If variable Z moderates the relationship between X and Y, we also say that X and Z *interact* in explaining variation in Y.

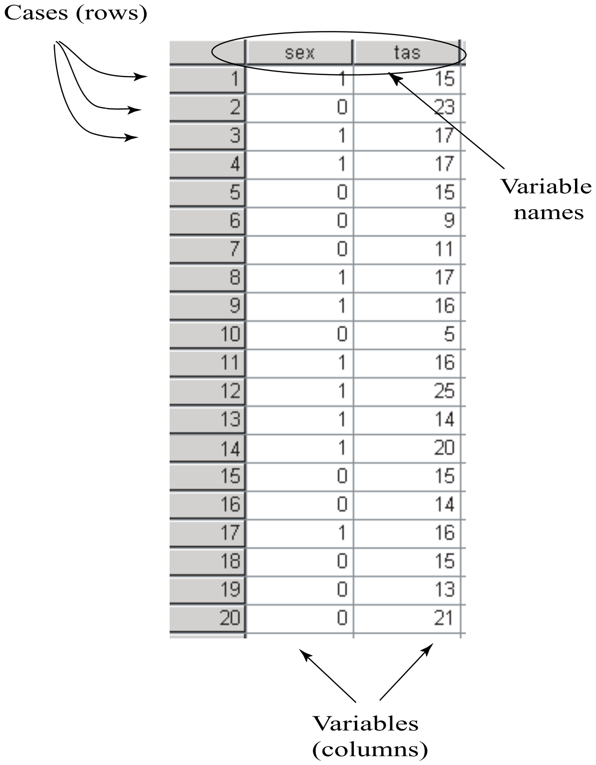
Of course, there are many other important terms that you need to be comfortable with eventually, and they will be introduced at the appropriate times throughout this book.

1.5 The Role of Statistics in Scientific Investigation

As discussed at the beginning of this chapter, science consists of at least two interdependent stages—research design and data analysis. Research design involves such issues as how to measure the variables of interest, how to approach the data collection procedure, who to include as participants in the study, how to find them, and other matters pertaining to procedure. Data analysis involves how to describe the results of the study and how those results relate to or reflect upon the research question or hypothesis being tested. Relatedly, statistics focuses on the kinds of inferences that can be made about the people or process being studied given the data available. The branch of statistics known as *descriptive statistics* focuses on graphical and numerical summaries of data. We use descriptive statistics to summarize a set of measurements taken from a sample from a population as well as when we have measurements taken from every member of the population. *Inferential statistics*, on the other hand, focuses on the kinds of inferences that can be made from the data collected from a sample to either some broader population from which the sample was derived or to the *process* under investigation. Both categories of statistics are important.

When we collect data, we often assume that the units providing data to the researcher are only a small subset of all the possible units that could have provided data but simply did not. For example, you may be interested in studying gender differences in the importance that people place on the television in the course of their daily lives. So you administer the *Television Affinity Scale* (see Rubin et al., 2004) to 10 men and 10 women. These 20 men and women who provide data to you constitute your *sample.* In contrast, you may have some specific *population* of interest from which your sample was derived. If you are interested in sex differences in general, your population may be defined as all males and females. Your population, however, may be more specific, depending on your research interests and objectives. Perhaps you are interested not in men and women in general but *adult* males and females. Thus, your population is defined as all men and women over 18. Or you may be focusing only on college men and women, so your population is all men and women enrolled in college. If your unit of analysis is not people but something else, such as newspaper advertisements, your sample may be a small subset of the newspaper advertisements that appeared in major newspapers that week, and your population is newspaper advertisements published in major newspapers. Ideally, you should specify your population in advance before collecting your sample from that population, although this is not often done. It goes without saying that your sample should be derived from your population of interest. If you were interested in college men and women, you would want to make sure that the data that you analyze come only from men and women enrolled in college.

After you have administered the *Television Affinity Scale* to your sample of men and women, you will probably want to examine the data in some way and describe what you found. A possible data set as it would be set up in *SPSS*, a data analysis program widely used by communication scientists, can be found in Figure 1.1. This figure illustrates the way that a data set is typically entered in a computer program and shows how some of the concepts described in the previous section relate to the data set. Each case gets a row in the data matrix, and the variables measured are



	sex	tas
1	1	15
2	0	23
3	1	17
4	1	17
5	0	15
6	0	9
7	0	11
8	1	17
9	1	16
10	0	5
11	1	16
12	1	25
13	1	14
14	1	20
15	0	15
16	0	14
17	1	16
18	0	15
19	0	13
20	0	21

Figure 1.1 An example data set as it might appear as an SPSS data file.

represented in each column. So if the variable “sex” codes whether the participant in the study is a male (1) or female (0), then case 12 is a male that scored 25 on the *Television Affinity Scale* (named “tas” in the data file). This is a relatively small data set. In reality, the data set from a typical research study would be much larger than this example, with many rows (cases) and many columns (variables).

Making sense of the data in this format is difficult. Researchers usually condense or transform their data into a format that is more digestible and easier to interpret using various numerical and graphical approaches to describing data. To forecast where we are headed, you might compute a measure of central tendency such as the *arithmetic mean*, as well as a measure of variation such as the *standard deviation*. Such numerical summaries of the data would tell you that the 10 men scored higher on the *Television Affinity Scale* than 10 women *on average* but that the women differ from each other more than the men. Such descriptions are an important part of the scientific process for they are a simple way of representing a study’s results.

One thing is for certain, and that is when you conduct a study, you will get some result. For instance, if you are comparing two groups of people in their response to some question, most certainly there will be some difference between the groups in your data. The question that needs to be answered is how much faith can be placed in the result you found as a good description of what is true in the population under investigation, or what it reflects about the process producing the pattern of results observed. Had you used 10 different women and 10 different men in the study, it is entirely possible that there would be no difference between men and women on average in their responses to the *Television Affinity Scale.* Or perhaps in a different set of 10 men and women, the women would actually be *higher* on average than the men. Researchers must acknowledge that any result found may just be the luck of the draw—the fact that the data were collected on these 20 men and women rather than some other group of 20 that could have provided data but that did not. *Chance* can be used to explain any study result, and we will talk about several models of chance throughout this book. Inferential statistical procedures are often used to gauge the role of "chance" in producing a study result and give the researcher a means of testing whether any result could be reasonably attributed to this thing we call "chance."

Suppose a theory predicts that men would feel television is more important in their lives than would women. The data in Figure 1.1 support this prediction. But as I just described, if a different group of 10 men and women participated in the study, this could have produced a different result. Before making some grand claim that men and women in the population of interest differ, or whether the theory is supported by the data, it is necessary to contend with the possibility that chance is the best and sole explanation for our obtained result. Chance is generally regarded as the simplest explanation for a study result, and in accordance with the rule of parsimony, scientists assume chance is the mechanism until we have evidence to the contrary. Inferential statistical procedures are used to assess whether such evidence exists. Again, to forecast where we are headed, you may use an inferential statistical procedure to estimate that if you assume that men and women in the population of interest don't actually differ on average in their television affinity, then chances of getting a difference favoring men as large or larger than you found is only about 1 in 50. When a research finding is determined to be unlikely to occur "just by chance," then chance is discounted as a good explanation for the result and this leads us to infer that the result reflects a real difference between men and women in the population from which the sample was derived.

Before closing this first chapter, I want to make one point regarding inferential statistics. The model I just described, where one uses the results from a small group of units in the study to some larger population is only one way of thinking about statistical inference, the so called *population model* or *random sampling model* of inference. This model of inference is used to make *population inferences*, which are statements about some broader population using information obtained from a subset of that population. Other models of inference are very appealing and arguably more appropriate for many of the research problems that communication scientists confront, such as the *random assignment model.* Alternative models of inference are not widely known or discussed in the communication literature, and I feel obligated to bring them to your attention when appropriate. Having said this, the population model of inference is so pervasive that it is important that you understand it thoroughly. So I will emphasize this model in much of my discussion of statistical inference in the chapters that follow.

1.6 Summary

Communication is a scientific discipline. Communication researchers use the scientific method to test theory, to help solve real world problems, to acquire information, and to satiate their own curiosities about the world. Researchers who subscribe to the canons of science acknowledge the importance of being objective and systematic in their approach by using only observable data to answer research questions and test hypotheses that are falsifiable. They recognize that science is a cumulative process and that the results of a single study are only one small part of a bigger puzzle being put together. They live their professional lives knowing that nothing can be proven and that “facts” exist only to the extent that there is no compelling evidence to the contrary. Researchers recognize that evidence disconfirming what they believe to be true could surface at any time. Communication scientists are active users of both descriptive and inferential statistics. They use statistics to reduce and summarize a complex data set into simpler and more interpretable forms and to assess the extent to which chance is the more parsimonious explanation for a study result. They also use statistics as a means of acknowledging the uncertainty attached to any findings resulting from collecting data from only a sample of a larger population.

My hope is that after working through this book you will appreciate that statistics is not simply a form of mathematics, where you plug numbers into one end of an equation and get another number out the other end. Statistics is a field of inquiry, not just a branch of mathematics. As in any field of inquiry, when a bunch of smart people work independently to solve a problem without a correct answer, they are bound to disagree while at the same time believing that their own solution is the best or most correct. In that way, statistics is just like the field of communication in that it is filled with controversies, differences of opinion, strong personalities, and people who are trying to find the “truth” in a world where truth is often elusive.

CHAPTER
TWO

Fundamentals of Measurement

Statistical procedures are used in communication science to both describe study results and to make inferences from a study result to some population or process that the investigator is studying. The use of statistics ultimately involves some kind of mathematical operation on numbers and then a translation of those numbers into something interpretable in the context of the study design and its purpose. The numbers in the data set being analyzed are almost always the result of some kind of measurement of the units on one or more constructs. A *construct* is simply a concept—typically discussed with some kind of verbal label—that we use to describe a psychological, cognitive, biological, or behavioral process or attribute of the object being studied. Media exposure, shyness, attitudes toward censorship, interest in politics, aggression, news learning, gender, nonverbal sensitivity, and family communication patterns are among the many constructs that communication scientists study. It is through the measurement process that we are able to do such things as quantify the association between constructs, compare groups in their behavior, thoughts, or feelings, or otherwise test hypotheses that our intuitions, curiosities, or theories suggest. Because measurement is important to statistics and the scientific process, it is important to understand some of the fundamental concepts in measurement, the topic of this chapter.

2.1 Measurement Concepts

In Chapter 1, I described the philosophy of empiricism—that scientists use observable data and that to be studied empirically the constructs must be measurable. But just what is measurement, and how is measurement done? There are many definitions of measurement, such as "a scheme for the assignment of numbers or symbols to specify different characteristics of a variable" (Williams & Monge, 2001, p. 11) or "careful, deliberative observations for the purpose of describing objects and events in terms of the attributes composing a variable" (Baxter & Babbie, 2004, p. 107). I think of measurement as *the process of meaningful quantification.* When we measure something, we are assigning numbers to the units on some construct, as Williams and Monge's definition states. But it is not enough, in my judgment, merely to assign numbers to objects to qualify as measurement because those numbers may not be meaningful. To be considered meaningful quantification, those numbers must actually correspond to some kind of quantitative attribute of the object being measured. This important property of measurement will be described shortly. But first, what methods of measurement do communication scientists typically use? How do we obtain the measurements that we subject to data analysis?

2.1.1 Methods of Measurement

Do portrayals of violence on television or other media forms facilitate or promote aggressive, violent behavior in society? Although for experts in the field this question is largely resolved (e.g., Anderson, Berkowitz, Donnerstein et al., 2003; Anderson & Bushman, 2002), the popular media still discuss this debate, and communication researchers continue to conduct research on the specific processes linking exposure to violent media and aggression (see e.g., Slater, Henry, Swaim, & Anderson, 2003). Suppose you wanted to determine whether exposure to violent media is related to a person's tendency to act aggressively or violently. Where would you get the data you would need to answer this question?

One possibility is to develop a questionnaire that asks people to list all the television shows they watch regularly. From their responses, you could count up the number of shows the person watches that could be construed as violent. This could be your measurement of each person's exposure to violent TV. Those data could then be compared to ratings that the participants in the study receive from family members, friends, or acquaintances about how violent or aggressive they tend to be. This hypothetical study illustrates one simple method of measurement: the *survey.* Self-report methods such as surveys are a common source of data that communication scientists use. Indeed, they are perhaps the most common method of measurement in communication research. Surveys are conducted over the telephone, in person, by mail, or over the Internet, and can be administered in either written, oral, or electronic format. Surveys have obvious problems and disadvantages, including such worries as whether people can accurately report the information that you are requesting and the possibility that respondents will not be truthful. Nevertheless, they remain an important source of data to communication researchers and will continue to be used widely.

Another possibility would be to recruit a group of children to a laboratory. You show some of them a series of violent TV shows over a week long period, whereas you show others a series of nonviolent shows. After this exposure period, the children in the study could be placed into a situation that provides an opportunity to respond in an aggressive manner such as in a playroom with a group of other children. The interaction could be videotaped and each child's aggression in that situation quantified in some manner later (such as the number of times a child pushes or hits another child). This is an example of an *experiment.* In this study, exposure to violent television is *manipulated,* whereas aggression is measured through *observation.* Communication scientists often conduct experiments through the manipulation of one or variables because experiments are one of the best ways of assessing cause-effect relationships. Observation, although a bit more complicated and not as widely used by communication scientists, has some obvious advantages over other methods of measurement such as self-report surveys. Observational measurement does not require the study participants to reflect on or recall their own behavior, and if behavior is what you are ultimately interested in measuring, it is hard to argue that the data you obtained do not reflect on a person's actual behavior, unlike self-reports collected through survey methods.

Still another possibility is to examine if the publication of news stories about violence is related to later crime statistics, such as the number of homicides in the community. To conduct such a study, you might select a number of cities throughout the country and examine how many column inches the major newspaper for each city published during a month-long period that contained references to some violent act (such as homicides, assaults). This information could then be compared to crime statistics in those cities a month later compiled by various government agencies after adjusting in some fashion for the amount of crime in the previous month. This study illustrates measurement through *content analysis* and *archival data sources*, both methods of measurement that enjoy frequent use by communication scientists. Content analysis involves the quantification and analysis of information contained in a stream of communication, be it in print, visual, or auditory form. Archival data are data a researcher uses that someone else collected, perhaps some time ago but for a different purpose. Archival data can also be publicly available information that a researcher compiles from public records. Archival data are the primary source of measurement used in many empirical articles published in the communication literature. The *National Election Study*, for example, is a continuously updated, expanding, and publicly-available data set maintained by the University of Michigan and used by many political scientists and political communication scholars interested in political processes and the role of communication in political activity. Data files from the National Election Study are available free on the Internet (http://www.umich.edu/˜nes/), as are the data from many other surveys (for example, from polls regularly conducted by the Pew Center for People and the Press, available at http://www.people-press.org, or the Inter-University Consortium for Political and Social Research, at http://www.icpsr.umich.edu/).

Although this is not an exhaustive list of methods of measurement, and the specific studies described above are in many ways ill-designed to answer the research question unequivocally, the methods described above do capture perhaps 75% or more of the methods that communication researchers use to collect data. This is book is about data analysis more than research design, so I refer you to specialized books on the topic of design and data collection for more details on the methods of measurement communication researchers use (e.g., Baxter & Babbie, 2004; Shadish, Cook, & Campbell, 2002; Frey, Kreps, & Botan, 1999; Fowler, 2001; Krippendorff, 2003; Neuendorf, 2001).

2.1.2 Operationalization

Before you can measure using one or more of the procedures described above, you first need to decide how to operationalize the constructs you intend to measure. *Operationalization* is a term used in two different ways in science—as a noun and as a verb. As a noun, an operationalization is the specific manner in which a construct is measured in a research study. But used as a verb, operationalization is the process of translating a construct into a measureable form. Let's consider this concept of operationalization by continuing with the earlier example of examining the relationship between exposure to violent media and aggression.

Although it may seem on the surface that exposure to violent media is a simple enough construct, the fact is that how to best measure it is far from simple, and it is certainly controversial. A number of possible operationalizations of exposure to violent media exist. Using survey methods you could ask people to rate their exposure to violent television with a simple question asking them to rate on a 1 (not at all) to

10 (very frequently) scale how much television they watch that they would consider violent. Or you might impose a definition of violent television on them by asking them to indicate which television shows in a list of violent shows you provide to them they watched in the last month. Or you might choose a method of data collection that doesn't require people to self-report. For example, if you had the technical know-how or access to such technology, you might provide participants in your study with a device to attach to their TV that records the shows that are on the channel the TV is tuned to when it is on. Over a course of a month, this would provide a relatively objective measure of at least what programs were being projected into the person's living room (although it may not adequately capture what a specific person in the household is actually watching). Or you might control how much violent TV the participants in your study are exposed to by controlling, temporarily, which shows your participants are allowed to watch. This might be easy to do if your participants were children, as the parents could regulate the shows the children are allowed to watch under your instruction. Rather than controlling the content children are exposed to, it might be easier to ask the child's parents to indicate which shows the child watches and then quantifying how much violent content those shows contain in order to quantify each child's exposure to violent television images. Each child's "exposure score" might be the average violence content of the shows that his or her parents say that the child watches regularly.

It is through the process of operationalization that social science is possible. Many of the constructs that communication scientists study, such as communication apprehension, nonverbal sensitivity, media exposure, political knowledge, and the like are abstract, nebulous, and somewhat difficult to define. To study these things, we need to be able to measure them concretely and specifically. A good source of information for how to operationalize various communication-related constructs is Rubin, Palmgreen, and Sypher's (2004) *Communication Research Measures*. Although it focuses almost exclusively on self-report survey-based measures, I believe that every communication researcher should have a copy of this useful resource in their personal library. Other good resources on operationalization include *Measures of Personality and Social Psychological Attitudes* (Robinson, Shaver, & Wrightsman, 1990), *Measures of Political Attitudes* (Robinson, Shaver, & Wrightsman, 1998), the *Handbook of Marketing Scales* (Bearden, 1998), and the *Sourcebook of Nonverbal Measures* (Manusov, 2005). A more thorough overview of the process of operationalization than I have provided here can be found in Baxter and Babbie (2004).

Choosing an operationalization. As the example above illustrates, the same construct can be measured in different ways, and none of the operationalizations of exposure to TV violence described above are perfect. So how do you choose how you should operationalize your constructs of interest in your specific study? There are a lot of things to consider.

First, you should use an operationalization that is going to be convincing to potential critics and anyone who might read your research. Faulty operationalization is a reason many papers are rejected when they are submitted for publication in the communication field. The research community will not hear much about your research if you don't measure well, using a sound operationalization of the constructs you intend to be measuring. For this reason, operationalizations that others have used in published research, particularly in research journals that have a peer-review system, are a relatively safe choice because those studies have gone through a screening process focused on the quality of the design of the study prior to publication. And it is easier to defend yourself by citing precedent rather than just saying "I made it up!" Having said this, there are some obvious dangers to blindly following the lead of others. For example, you may end up making the same mistakes that others have made. So keep up with the scientific literature in your area of interest because that literature often informs you not only how to operationalize and measure what interests you, but the literature is also peppered with criticisms of operationalizations others have used and that perhaps you should avoid.

Second, if possible use more than one operationalization. Imagine you found that people who watch lots of violent television tend to be perceived by their friends and neighbors as aggressive. This finding could be used as evidence supporting your claim that exposure to violent media is related to a person's tendency to be aggressive or violent. But wouldn't it be even more convincing if you could also report that those same people were more likely to have been arrested for some kind of violence-related crime, were more likely to endorse statements such as "sometimes a punch in the mouth is the only way to resolve our differences," and were evaluated by a clinical psychologist as prone to sociopathic behavior? When you get the same finding when measuring something different ways, you can say you have *converging evidence* for your claim. A finding based on a single operationalization of the construct may be interesting and even important, but the results can be highly vulnerable to criticism. A set of consistent findings obtained with several different operationalizations of the construct is more difficult to criticize and considerably easier to defend.

Third, you should use operationalizations that are consistent with the resources you have available. Some operationalizations produce data quite easily and nearly anyone can use them. For example, finding people to respond to a set of questions about their attitudes, beliefs, or behaviors (unless you are studying young children or people who have some kind of disability that makes communication difficult) is fairly easy. Other operationalizations may be difficult or next to impossible for you to use. If you were studying anxiety reactions to fear-arousing advertisements, you may not have access to sophisticated equipment to measure physiological markers of anxiety such as heart rate or skin conductance. Or if you wanted to use something such as crime statistics, the community you are studying simply may not have the data you need, or it may be too costly to obtain.

2.1.3 Levels of Measurement

Stevens (1958) popularized a scheme for conceptualizing *levels of measurement* on a continuum from low to high. According to this scheme, the highest level of measurement is the *ratio level*. Measurement at the ratio level has the following properties:

1. A measurement of "zero" implies a complete absence of what is being measured.

2. A one-unit increase on the measurement scale corresponds to the same increase in what is being measured regardless of where you start on the scale.

Exposure to violent TV programs, defined as the number of violent TV shows you reported watching last week, is an example of ratio-level measurement. The difference between 4 shows and 3 shows is the same as the difference between 2 shows and 1 show. And zero implies an absence of what is being measured. Zero truly means zero—no TV shows were watched. Other examples include minutes spent reading the newspaper today and how many children you have.

The existence of an absolute zero point that corresponds to an absence of what is being measured means that ratios of measurements can be interpreted to mean that two research units who differ by a certain ratio on the measurement scale can be said to differ from each other by the same ratio on the construct being measured. For instance, someone who reports that he watches 6 hours of television a week can be said to watch television twice as frequently as someone who reports watching television only 3 hours a week.

Measurement at the *interval* level is said to be of a "lower" level of measurement than the ratio level. Interval-level measurement also has the property that equal differences on the scale correspond to equal differences in what is actually measured. However, with interval-level measurement, the value of zero does not imply the absence of what is being measured. Temperature on the Fahrenheit or Celsius scale is often used as a generic example. The difference between 20 degrees and 30 degrees corresponds physically to the same actual difference in temperature as the difference between 10 and 20 degrees. But zero does not imply an absence of temperature. So you cannot say that 80 degrees is twice as hot as 40 degrees. As I soon discuss, true interval-level measurement is fairly rare in communication research.

The next lowest level of measurement, the *ordinal* level, does not have the property that interval and ratio level measurement share—equal differences in the scale corresponding to equal differences in the construct being measured. Ordinal measurement quantifies only with respect to a *relative* but not *absolute* amount of what is being measured. For example, a customer might provide an evaluation of an Internet service provider's service as excellent, good, not so good, poor, or very poor. We can assign numbers to these ratings in a meaningful way: 5 = excellent, 4 = good, 3 = not so good, 2 = poor, 1 = very poor. The numbers themselves are arbitrary, but at least they seem to scale the quality of the service along an ordinal continuum where the lowest value on the scale (a value of "1") indicates the worst evaluation, and the highest value ("5") indicates the best evaluation. Furthermore, increasing steps on the arbitrary numerical scale seem, on the surface at least, to indicate increasing quality of the service. An alternative may be a simple rating scale where you ask someone to judge something on a 1 (worst) to 10 (best) scale. Another common ordinal-level measurement procedure is the Likert-type scale, where you ask a person to indicate the extent to which he or she agrees with a statement.

Unlike ratio and interval level measurement, ordinal level measurement doesn't provide precise information about *how much* two measurements differ from each other on the construct being measured. Even though the difference between 1 (very poor) and 2 (poor) is the same as the difference between 2 (poor) and 3 (not so good) in the numerical representation, we don't know that the difference in quality is the same in the minds of those interpreting the scale. The same one point difference between 1 and 2 and between 2 and 3 may not correspond to the same difference in what is actually being measured.

Ordinal measurement is very common in communication research, whereas interval-level measurement is rather rare. But many researchers treat ordinal-level measurement as interval level. Indeed, the practice of conceptualizing ordinal-level data as if were measured at the interval level is so common that rating scales such as those described above are sometimes used as examples of interval measurement in some textbooks (e.g., Frey, Kreps, & Botan, 1999). With interval-level data, you can legitimately apply arithmetic operations to the measurements, like adding and subtracting, knowing that the result means the same thing regardless of where you start with respect to the amount of what is being measured. So it is sensible to say that someone who reads the newspaper twice a week reads it one day more than someone who reads it only once a week and that this difference of one day is the same as the difference between 2 people who read the paper 4 and 3 days a week. But it is difficult at least in principle to justify doing arithmetic operations on ordinal data. Suppose, for example, we ask 4 people to respond to the statement "There is too much violence on television" with one of 5 options: strongly disagree, disagree, neither agree nor disagree, agree, or strongly agree. It would be common to assign numbers to these responses, such as 1 through 5, as a way of quantifying level of agreement. Imagine Jerry responds agree (4), Carroll responds strongly agree (5), Chip replies neither (3), and Matt says disagree (2). We cannot say that the difference in agreement between Jerry and Carroll is the same as the difference between Matt and Chip. Nor can we say that Jerry agrees with the statement twice as much as Matt. The numbers we assign to their responses belie the fact that they are arbitrary and carry only relative information, not absolute information, about what is being measured (level of agreement).

Because many statistics used in research require such mathematical operations, many people treat ordinal measures as if they were interval data and don't worry about it. This is very controversial from a measurement perspective, but it also very common. For example, grade point average (GPA) is based on arithmetic operations on a set of ordinal level measurements (where A = 4, B = 3, C = 2, D = 1, F = 0). But GPA is rarely questioned as a legitimate measure of academic performance. Communication researchers routinely compare people based on ordinal evaluations they provide of something, and they often arithmetically combine a set of ordinal measurements. The statistical methods literature is replete with research on the legitimacy of treating ordinal-level data as if it were interval. The research suggests probably no serious harm is done in doing so under many of the conditions that communication researchers confront (e.g., Baker, Hardyck, & Petrinovich, 1966). Some have called ordinal measures treated like interval as "quasi-interval," a term that I think deserves recognition, reflecting the fact that many of the measurement procedures communication researchers use produce data that, though ordinal, can be treated like interval without doing too much damage.

In Steven's (1958) measurement scheme, the lowest level of measurement is known as the *nominal* level. A variable measured at the nominal level does nothing other than categorize into groups. A person either does or does not speak their opinion when asked in class. The leadership style of a CEO may be classified as either authoritarian, laissez-faire, or democratic. A person's primary source of news might be either the newspaper, the television, the radio, or the Internet. So nominal "measurements" are those that place people in categories. Although widely discussed as such, I do not consider this measurement because measurement requires meaningful quantification according to my definition. Nominal measures do not meet my definition of measurement because they do not reflect the assignment of a number that indicates the *amount* or quantity of something. Consider a construct such as religious affiliation. Recalling William and Monge's (2001) definition of measurement, we can come up with a scheme that assigns numbers (a symbol) to objects (people) depending on whether they are Protestant (religion = 1), Catholic (religion = 2), Jewish (religion = 3), Muslim (religion = 4), Nondenominational (religion = 5) or "Other" (religion = 6). But the numbers themselves have no quantitative interpretation, and thus this is not meaningful measurement. A Jewish person isn't somehow more religious or more affiliated with a religion than a Protestant. Categorization is not really measurement by my definition. It is simply categorization.

2.1.4 Measurement Precision

Another way of conceptualizing measurement is in terms of its fineness versus its courseness. A fine measure has very many possible values. A course measure, in contrast, has few possible measurements. The finest possible measurement is called *continuous*. A continuously measured variable has an infinite (or at least a very large) number of possible values. The number of seconds it takes you to answer a question correctly is continuous, in that there is no limit to the number of possible values a measurement could take because time can be measured at an infinitely fine level (e.g., one hundredths of a second), and often there is no upper bound imposed on measures of time. The number of children a woman has, however, is not continuous. Human biology imposes a limit on how many children a woman can conceive, and it is not possible for a person to have 1.2 children, 2.6 children, or any number of children that is not an integer.

By contrast, a *discrete* measure has a limited number of possible values. Examples of discrete measurement include the number of days a week you read the newspaper at least once (only 8 values possible), your rating of the quality of care your doctor provides (1 = very poor, 2 = poor, 3 = adequate, 4 = good, 5 = very good), how many of your grandparents are still alive (only 5 values possible—0 to 4), and how many children a person has.

Whether a measurement procedure produces continuous or discrete measurements is not always obvious, and it depends in part on how you use the measuring instrument. Consider you allowed a person up to 60 seconds to view a web page, and you recorded the number of seconds the person chose to view it in one-second intervals. Technically, this is discrete measurement. Had you recorded it in milliseconds, while technically there is a limit to the number of possible values, the measurement is so fine (there are 60,000 possible measurements) that it would be sensible to think of this as essentially continuous. Differences of less than one millisecond would have no importance to any communication-related research using this fine of a measure, although in other fields such a difference might be meaningful. You can always record time even more precisely than the millisecond, so theoretically, time could be construed as continuous when measured at a fine level. However, it is clear that the number of children a person has could never be thought of as continuous. It is clearly discrete.

It is also sensible to conceptualize measurement precision not only in terms of possible values but plausible values. Though it might be possible for a person to have 60,000 children, and thus 59,999 is possible, as is 59,998, and so on, you could think of such a measure as continuous (using the same logic as above). But it is implausible that you'd ever get a measurement greater than perhaps 10 or so in everyday use of such a measure. It is therefore sensible to think of this as discrete measurement. Although the number of possible values is perhaps quite large, the number of plausible values is relatively small.

Analyzing data using the highest level of measurement precision available to you is generally best. It is far too common for a researcher to reduce measurement precision by making a continuous or nearly continuous set of measurements discrete. For example, participants in telephone surveys are often asked their year of birth to derive a measure of their age in years by subtracting the reported birth year from the year of data collection. In that case, age is a ratio-level variable that can be treated as practically continuous in analyses involving age. But to their misfortune, researchers often first turn age into an ordinal variable by classifying people into age groups (e.g., under 18, 19 to 35, 36 to 50, 50 or older) prior to analysis. Another common example is the *median* or *mean split*, where an investigator creates "high" and "low" groups by classifying respondents' measurements on some dimension as either above or below an arbitrary cutoff. For instance, a study participant might be classified as either high or low in self-esteem based on whether or not his or her score on the Rosenberg self-esteem index exceeds the sample median or mean. This practice has little value and should be avoided unless there are justifiable reasons for treating people lumped into ordinal or discrete groups as if they are the same on the variable being measured. For a discussions of the many reasons for avoiding this practice, see Cohen (1983), MacCallum, Zhang, Preacher, & Rucker (2002), Irwin & McClelland (2003), Maxwell & Delaney (1993), and Streiner (2002).

2.1.5 Qualitative Data versus Quantitative Measurement

A final distinction pertinent to measurement is the distinction between qualitative and quantitative data. Qualitative data describe or code the object being measured in qualitative rather than quantitative form. You will recognize the nominal level of measurement as qualitative. For the same reasons I gave when discussing the nominal level of "measurement," qualitative measurement cannot be considered true measurement by my definition. Quantitative measurements, in contrast, represent the object of measurement in quantitative terms on the dimension being measured. Ordinal-, interval-, and ratio-level measurement all qualify as quantitative measurement, although an argument could be made that ordinal-level measurement possesses both qualitative and quantitative features depending on what is being measured.

2.2 Measurement Quality

Science requires measurement, but good science requires good measurement. Anybody can measure something, but it takes effort to measure well. The quality of one's measurement of a construct can vary along a continuum from poor to perfect (although if you ever succeed in measuring something perfectly you deserve an award, as perfect measurement is next to impossible). No single numerical index of measurement quality exists. However, measurement quality can be judged on two important dimensions: *reliability* and *validity.*

2.2.1 Reliability of Measurement

Reliability assesses how much numerical error there is in the measurements. The quantification of reliability of measurement is too advanced a topic at this stage of the book because a complete understanding requires some familiarity with material not yet introduced. The theory of reliability and ways of quantifying reliability are discussed in Chapter 6 after some of the relevant prerequisite statistical concepts are introduced. But at this stage we can discuss some of the basic ideas.

A method of measuring some construct, such as a self-report questionnaire, produces an *observed measurement* of that construct. This observed measurement is the empirical quantification of the unit on the construct being measured. But this observed measurement is unlikely to be equal to that unit's *true score* on what is being measured. The true score is the actual amount of the construct possessed by that unit. The observed and true scores are not necessarily the same thing. The observed measurement is linked to the true score by the equation

$$\text{Observed Measurement} = \text{True Score} + \text{Measurement Error}$$

So the observed measurement contains both the true score of the unit on that construct as well as some error in measurement. The smaller the error, the higher reliability. Various reliability indices that exist are means of quantifying how large these errors in measurement tend to be.

Reliability is often conceptualized as the repeatability of a measurement or set of measurements. If you measured your height with a ruler, recorded your height in inches on a piece of paper, and then measured your height again, you'd be surprised if your measurement changed. Your surprise would stem from that fact that your intuition tells you that a ruler is a reliable measure of height. By the same reasoning, you expect your weight not to vary much from measurement to measurement. To be sure, at different times of the day your weight may fluctuate (depending how much you had eaten recently, what you are wearing, etc), but you'd be surprised if it varied by more than a few pounds from measurement to measurement. This is because most measures of weight, such as a bathroom scale, are very reliable measures of weight. These same ideas can be applied to measures of communication-related constructs, such as media exposure, communication apprehension, and the like. If a measuring instrument produces measurements that contain little error, you would expect that if you repeatedly measured something on the same unit in your study and the unit has not actually changed on what you are measuring, the resulting measurement should be the same or at least very similar time after time. As will be discussed in Chapter 6, rarely would you expect two measurements of the same unit over time to be exactly the same even if the unit has not changed on the construct being measured, but if the measurement instrument is a good one (that is, if the measurements contain relatively little error) then those repeated measurements should be very similar.

But as will be discussed later, repeatability is an *outcome* of high reliability, not the definition of reliability. To be sure, if a set of units is measured on some construct X with relatively little measurement error, then repeated measurement of those same units should be similar over time. But reliability refers to the amount of measurement error that pervades a set of measurements or the measurement process, not just the repeatability of those measurements over time.

2.2.2 Validity of Measurement

Validity of measurement speaks to whether the obtained measurements can be thought of as sensible and high quality quantifications of what the researcher *intends* to be measuring. Whereas reliability quantifies how much *numerical error* exists in the measurement, validity assesses *conceptual error.* A bathroom scale is a reliable measure, but it is not a valid measure of height. Even though a scale will provide consistent measurements, it would be silly to use someone's weight as a measure of his or her height. Indeed, tall people tend to be heavier, but a person's weight is not a sensible or valid measure of his or her height.

Although this example illustrates the point, determining whether or not a method of measurement is valid is typically not at all clear cut. There is no way to prove that a measuring instrument or set of measurements are valid. Validity is assessed through logical argumentation and research. Unfortunately, many researchers just

Box 2.1: The *Television Affinity Scale*

For each statement, please indicate the extent to which you strongly disagree, disagree, agree and disagree, agree, or strongly agree as it applies to you. Circle a number from 1 to 5, where 1 = strongly disagree, 2 = disagree, 3 = both agree and disagree, 4 = agree, 5 = strongly agree.

(1) Watching television is one of the more important things I do each day. 1 2 3 4 5

(2) If the television set wasn't working I would really miss it. 1 2 3 4 5

(3) Watching television is very important in my life 1 2 3 4 5

(4) I could easily do without television for several days 1 2 3 4 5

(5) I would feel lost without television to watch 1 2 3 4 5

TO COMPUTE YOUR TELEVISION AFFINITY, FIRST SUBTRACT YOUR RESPONSE TO THE FOURTH QUESTION FROM 6. THEN ADD TO THE RESULT THE SUM OF YOUR RESPONSES TO QUESTIONS 1, 2, 3, and 5. YOUR SCORE SHOULD BE BETWEEN 5 and 25.

assume that a measurement procedure they are using is producing valid measurements of the construct they want to measure without providing any kind of argument or research to support that claim. A good measurement instrument should meet many different criteria of validity if it is to be accepted as valid. Not all measures will meet all criteria, and not all criteria are relevant for all measures.

Face Validity. We talk about a measuring instrument's *face validity* as a kind of intuitive feeling that the measure seems or feels right. When you look at a measuring instrument and compare it conceptually to the construct you intend to measure and find yourself thinking "yeah, it seems about right," then you are saying the measure has face validity. Consider for instance the *Television Affinity Scale* (see Box 2.1), which reportedly measures the importance one attaches to television in the course of day to day life (see Rubin et al., 2004). This is known as a *summated rating scale*, and it yields a single score for each person who responds to the questions. This single score is constructed by adding up the person's responses to each of the questions, as described in Box 2.1, such that a higher score reflects a greater sense of the importance of television to the person. For example, someone who responded agree or strongly agree to questions 1, 2, 3, and 5 and disagree or strongly disagree to question 4 would have a relatively high score on the measure—in the 20 to 25 range. The opposite pattern of responses would yield a low score—in the 5 to 10 range. If you look at this measure and you agree that it seems like a reasonable and sensible measure of this construct, then you are agreeing that the measure has face validity. This is a very subjective judgment, and if all you can say is that a measure has face validity, you haven't really said that much. Surprisingly, however, some of the published research in the communication literature includes measures that at best satisfy only the face validity criterion. Researchers often construct measures ad hoc, based on their needs for the particular study, and never give any kind of evidence, logical or otherwise, that the measurement procedure they are using is actually measuring what they claim. This is a problem not in just the field of communication; it permeates the social sciences and reflects the fact that demonstrating that a measurement procedure meets the more rigorous criteria of validity described below is a lot more difficult.

Content Validity. A closely related kind of validity is *content validity.* A measurement instrument is said to be content valid if the items in the measure adequately represent the universe of relevant behaviors or indicators of that construct. Consider a test of your mastery of the material presented in this book up to this point. If this test had 90% of its questions on levels of measurement and nothing on the philosophies and assumptions of science (in the previous chapter), you could reasonably criticize such a test on the grounds that it is not a content valid measure of your mastery of the material because it does not adequately cover all the material or even a decent representation of it. Similarly, people often criticize many measures of "intelligence" as low in validity because they fail to measure abilities that could be argued are intelligence, such as creativity or the ability to interact appropriately with people in various situations. Or if you were measuring "aggressive tendencies" by asking a person if he or she has ever punched someone in the face, that clearly would not be a content valid measure of "aggressive tendencies." Aggression can take many physical forms, as well as many nonphysical forms. Consider the *Television Affinity Scale* again. To assess its content validity, you need to ask whether something important may have been left out. Think about this for a minute. Might there be an item or two missing that would improve this measure?

I leave this question unanswered to illustrate that both content validity and face validity are somewhat difficult to assess objectively. You could say that a measurement instrument you are using has content and face validity, but someone else may disagree. All you can do is argue about it, duke it out verbally, and no one can really be "proven" right or wrong. So what can you do? Probably the best way to assess face and content validity is to show the instrument you are using to a group of experts on the topic of interest and see if they feel like the measure is missing something important. Does an expert or a group of experts feel like you are adequately capturing all the relevant domains, dimensions, behaviors, and so forth, that are relevant to the construct?

Criterion-related Validity. If the *Television Affinity Scale* is a valid measure of the importance that one places on television from day to day, then you'd expect that people who score low on the *Television Affinity Scale* would also watch relatively little television compared to those who score relatively high on the index. If that was the case, it would be reasonable to say that the *Television Affinity Scale* has *criterion-related validity.* Criterion-related validity is assessed by seeing if scores on the measurement instrument are related to other things that you'd expect them to be related to if the measure is measuring what you claim it is measuring.

There are a few special forms of criterion-related validity, but in practice the distinction between them is not important, and the terms I describe next are often used interchangeably. *Concurrent validity* refers to the extent to which a measurement instrument produces measurements that are similar to the measurements that a different measure of the same construct yields. For example, if you were developing a measure of shyness, you would hope that people who score relatively high in shyness on your measure also score as relatively shy on other measures of shyness. *Predictive validity,* also a form of criterion-related validity, assesses the extent to which scores on a measure accurately predict something that they should be able to predict. For example, you could argue that your measure of shyness is valid if it accurately predicts performance on a job in which shyness should inhibit good performance, such as a door-to-door sales.

With the concept of criterion-related validity now introduced, it is worth making the distinction between validating a measurement instrument and validating its *use*. We use measurement instruments to accomplish certain research objectives, such as examining the relationship between exposure to violent TV and aggression. If the measure allows us to accomplish those research objectives because our measures are measuring the constructs we want, then they are valid for that use. But there are other ways that a measure can be valid, even if we don't accept the validity of the measure as a measure of something in particular. You may not believe that such college admission tests as the SAT or ACT are valid measures of scholastic aptitude because they focus so much on verbal and mathematical skills. But they are valid predictors of college grade point average, and so it is perfectly legitimate from the perspective of making good decisions to use SAT or ACT scores of applicants for admission as a criterion in college admission decisions.

Here is another admittedly absurd example that makes the same point. The U.S. government employs thousands of people in the Transportation Security Administration (TSA) as airport security screeners. How does the TSA know if an applicant for a position as a security screener is going to be good? Suppose for argument's sake that research found that good screeners, when given the choice, prefer bananas to oranges whereas bad screeners prefer oranges over bananas (I doubt this is true, but let's pretend it is true). Fruit preference clearly is not a valid measure of the skills required to be a good screener using the simple definition of validity I gave earlier. But it is a valid predictor of performance in this example. As such, on the surface at least, it doesn't seem unreasonable to use an applicant's response to such a question to determine who should be and should not be hired if it is a valid predictor of performance as a screener.

Construct Validity. The final form of validity is *construct validity*. In my judgment, construct validity is really what most people are interested in when they ask whether a method of measurement can be said to be validly measuring the desired construct. We can say a measure has construct validity if it meets most of the criteria I've been describing. Does it "feel" right. Do others agree that the measure is representing the relevant dimensions or indicators of the construct the instrument reportedly measures? And is the pattern of associations between scores on the instrument and measures of other constructs consistent with the claim that the instrument is measuring the desired construct?

This latter criterion is probably the most difficult to grasp, so let me use an example from my own research on a construct I developed and named *Willingness to Self-Censor* (Hayes, Glynn, & Shanahan, in press a). Willingness to Self-Censor refers to a person's general, cross-situational willingness to censor their opinion expression around others thought to disagree with that opinion. I developed the *Willingness to Self-Censor Scale*, an 8–item self-report instrument to tap this individual difference in people (see Box 6.1). To establish that the *Willingness to Self-Censor Scale* is a construct-valid measure of this individual difference, I embarked on a series of studies to examine its correlates with other individual differences you would expect people who are relatively high versus relatively low on this construct to differ on. For instance, expressing an opinion that others don't agree with can result in an argument or some kind of interpersonal conflict. Therefore, you would expect people who are relatively more willing to censor their own opinion expression would be relatively averse to argumentation, a construct measured with the *Argumentativeness Scale* (Infante & Rancer, 1992). Indeed, I found that people who scored relatively high on the *Willingness to Self-Censor Scale* (compared to those relatively low) did indeed score relatively low on the *Argumentativeness Scale.* Additionally, I argued that the expression of a dissenting opinion is a risky act that you would expect people to do only if they were relatively confident in their own goodness, because our opinions are reflections in part of the things we value and find important and define who we are as people. To test this prediction, I administered the *Willingness to Self-Censor Scale* to a group of people and also had them respond to a series of questions that are known to measure self-esteem. I found that, as expected, people who scored relatively low on the *Willingness to Self-Censor Scale* were relatively higher in self-esteem. Furthermore, in an experimental context, I showed that people who scored high on the *Willingness to Self-Censor Scale* were more sensitive to the distribution of others' opinions when deciding whether or not to voice their opinion publicly than were people who scored relatively low on the scale (Hayes, Glynn, & Shanahan, in press b). These findings, combined with several others described in the validation paper and elsewhere, all can be used as evidence that the *Willingness to Self-Censor Scale* is a construct valid measure of this individual difference. It is measuring this construct rather than something else.

Validation of a measurement instrument is a complex and long process. If you are being innovative and creating a new instrument or method of measurement, it is very reasonable for a reader of your research to expect you to provide some evidence or argument that your measurement method is producing measurements of what you claim it is producing. If you are using a measurement instrument that others have used before and that others accept as valid when used in the way you used it in your research, then most are willing to give you the benefit of the doubt that your method of measurement is valid for that purpose. For an advanced discussion of measurement validation, see Cronbach and Meehl (1955).

Before closing this chapter, it is important to make the distinction between a set of observed measurements and the measuring instrument or procedure. A measurement instrument or procedure is what yields the observed measurements. A self-report measure such as the *Interpersonal Attraction Scale* (Rubin et al., 2004) is a measuring instrument, and if it is a valid measure of interpersonal attraction, it will produce observed measurements for a person's attraction to another (on three dimensions—social, physical, and task). We often talk about measurement instruments as being reliable or valid and that is a sensible thing to do. Some measures of communication constructs are certainly more reliable and valid than others. But it is not necessarily true that a reliable and valid measure will yield reliable and valid measurements in all circumstances. For example, a measure of self-esteem may have been developed and validated based on the responses of people 18 years old and older living in the United States. And it may produce reliable and valid measurements when administered to adult residents of, for example, the state of Ohio. Whether it produces reliable and valid measurements in adolescents, or residents of Sydney, Australia, or Barcelona, Spain is open to question. You can assume that the measure would be reliable and valid if used on adolescents or Australians or Spaniards, but that doesn't mean it is. Therefore, it is important whenever possible to report at least the reliability of your measurements when describing your research. A reader will want to know if your method of measurement yielded reliable measurements. Whether the instrument is reliable for other investigators who have used it in their research is only indirectly relevant. Of course, the chances of the observed measurements being reliable are much higher if you use a measurement instrument that has lots of evidence supporting its reliability in many circumstances or many populations. In contrast, there is almost no way of knowing whether a measurement instrument is valid in all circumstances or situations in which it is likely to be used. You can only assume it is valid, make the argument that it is valid, or provide data suggesting that it is valid.

2.3 Summary

Communication researchers typically quantify the constructs they are interested in measuring. The process of operationalization results in a means of measuring the constructs of interest to the researcher. The outcome of the measurement process is a set of observed measurements of the research units on the constructs being measured. These measurements, and the methods of measurement being used, can vary in precision and in quality. Communication researchers should strive to measure their constructs with as little error as possible (reliability) and ensure that they are measuring what they intend to be measuring (validity). Although reliability can be quantified (see Chapter 6), establishing the validity of measurement is considerably more complex and as much a logical and argumentative process as a statistical one, ideally buttressed by data when possible.

CHAPTER THREE

Sampling

Before any data analysis can begin, you have to have some data to analyze. Where do your data come from, and how do you go about obtaining the data you need to test your hypotheses and answer your research questions? Data can take many forms, it can come from many different sources, and there are various approaches to finding the data you want. As discussed in Chapters 1 and 2, data are the outcome of the measurement or categorization process. But before you can measure, you have to have someone or something to measure. In other words, you need to have recruited participants for your study or otherwise obtained the research units to be measured on the variables relevant to your study. In this chapter, I give a broad overview of the various approaches to recruiting or otherwise obtaining the research units that provide data to the researcher.

3.1 Population Inference

When we conduct research, we are often interested in some kind of inference. In statistics, the most common conceptualization of inference is *population inference*—the practice of making a statistical statement about a collection of objects from a subset of that collection. For example, if 60% of 200 people you ask approve how the president is doing his or her job, you might make the claim that around 60% of *all* people (rather than just those 200 you asked) approve. In so doing, your claim that 60% of all people approve is a population inference. To make such an inference, you must be very careful in how you go about recruiting or finding research units to be measured (e.g., people who respond to a question or participate in the research in some fashion) because most statistical methods make some kind of assumption about how the units were obtained in order to apply those methods to making population inferences.

When conducting research, we usually collect data only from a *sample*—a small subset of the population. The *population* is the universe of objects you are trying to make some kind of inference about, whereas the sample consists of a collection of members of the population (see Figure 3.1). The sample will almost always be smaller in size than the population and typically *much* smaller. We often (but not always) are interested in the population that a sample represents and not the individuals in

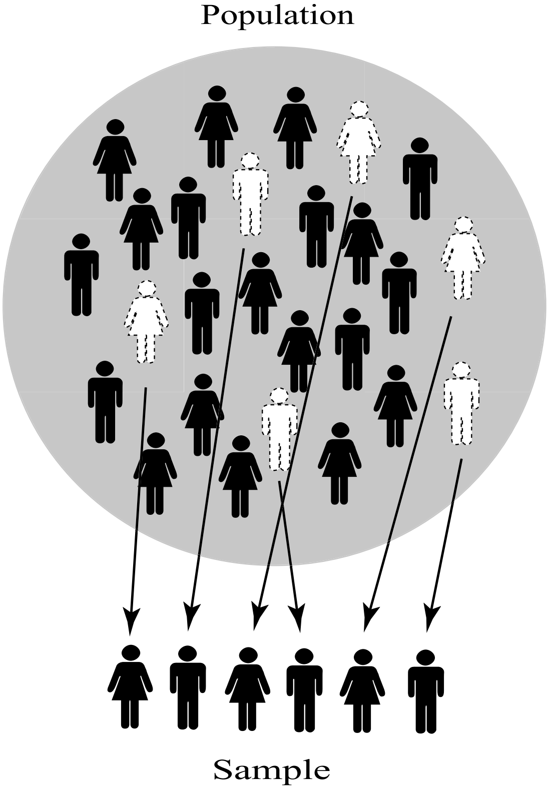


Figure 3.1 Population versus sample.

the sample itself. Public opinion polls are a good example of this. When a pollster calls 1,000 people on the telephone and asks their opinion on some issue, the pollster is interested in understanding the population of people from which the sample was derived. So the focus of the inference is on the entire population, not just the 1,000 people telephoned. Of course, it would be far too burdensome (and typically impossible) to obtain data from the entire population. But if you were somehow able to obtain data from every member of the population, then you would have what is technically called a *census*. Fortunately, when population inference is the goal of the research, it isn't necessary to obtain a census of the population because the sample can provide a window through which the population can be viewed if proper sampling procedures are followed.

Similarly, a media researcher may be interested in how news broadcasts (the population) portray minorities, but the researcher will probably examine this by studying only a small subset of news broadcasts, such as all 10 o'clock news broadcasts from four local networks in Los Angeles, Dallas, Chicago, and New York for a week (the sample). The researcher will have data about how this small subset portrays minorities, and if proper sample procedures are followed, he or she can use this small subset to make statements about the broader population of interest. Just how population inference is accomplished constitutes the theory of estimation, aspects of which are discussed in Chapter 7 and elsewhere in this book.

When we make a population inference, we are making a numerical statement about a population by generalizing a statistical description of the data in the sample to the population from which the sample was derived. But there is another kind of inference that is equally or even more important than population inference—*process inference*. Any research finding is constrained in some respects by who provided the data and how the units were recruited to participate in the study. When a researcher's primary interest is a process inference, there is less focus on specific numerical descriptions of the population (such as 60% of the U.S. population approves of the president, or people watch 2.5 hours of TV per week on average). Furthermore, who provided the data and how the participants were obtained typically take a back seat to other concerns when interpreting study results. Instead, the focus is on whether the theory or hypothesis is being adequately tested given the design of the study and whether the study helps to illuminate the process that is producing the effect observed. Process inference is usually the primary interest of communication researchers. This important distinction between a process and a population inference will be made clearer toward the end of this chapter and elsewhere throughout this book. For now, we will focus on sampling methods as a means of making good population inferences.

3.1.1 The *Literary Digest* Poll: Population Inference Gone Awry

Most of the statistical methods that communication researchers use are based on the *population model* of inference. The concept of population inference is probably best illustrated by first discussing one of the most well known population inference failures, the infamous *Literary Digest* poll of 1936. The *Literary Digest* was a magazine published in the early half of the 20^{th} century, and the publishers had been conducting a poll to predict U.S. Presidential election results since 1916. Until 1936, it had correctly predicted the winner of the presidential election every single time. But in 1936, it got it wrong by predicting that Republican Alf Landon would beat Democrat Franklin Delano Roosevelt, with Landon garnering roughly 57% of the votes. This prediction was a population inference, reflecting their estimate (57%) of the percent of the U.S. voting population that would vote for Landon, as well as a qualitative inference that Landon would win the election. But history tells us Roosevelt won the Presidency, not Landon, and by a huge margin (62% vs. 37%).

To understand how the *Literary Digest* got it wrong, it is necessary to understand the methodology of the poll in some detail. To conduct this poll, they mailed ballots to roughly 10 million people. The respondents were asked to indicate on the ballot who they were going to vote for in the upcoming election and then return the ballot by mail. The 10 million people were obtained largely from automobile registration lists and telephone books. Roughly 23% of those that received the ballot did return it; a sample size of over 2 million, which is a very large sample by any standard. But sample size is not everything because from responses returned, the *Literary Digest* incorrectly forecasted that Landon would win the election.

Just what went wrong has been a source of debate for some time, and although we will never know for certain just what happened, Squire (1988) published a rather convincing paper in *Public Opinion Quarterly* that attributed the failure to both data collection methods and nonresponse bias. His data came from an opinion poll conducted in 1937, just after the election. This poll contained questions about how the respondent voted in the 1936 election as well as whether he or she owned a car or telephone, received the *Literary Digest* ballot, and if so, whether he or she filled it out and returned it to the *Digest*.

One explanation that had been offered for the failure of the *Literary Digest* poll is that the list of 10 million people who were sent the ballot did not adequately represent the people of the United States. Of concern was the possibility that the use of automobile registrations and phone lists produced a collection of respondents who were more well-to-do than the U.S. population as a whole and therefore more likely to vote Republican. Remember that this was 1936, when the richer segment of the population would have been much more likely to own a car or telephone and therefore receive a ballot. However, Squire's data suggests that this explanation doesn't completely explain the failure of the *Literary Digest* poll. To be sure, those with neither a car nor a telephone were much more likely to vote for the Democrat (Roosevelt) than the Republican (Landon) by about a 4 to 1 margin (79% vs. 19%). But even among those who owned both a car and a telephone, Roosevelt received more votes than Landon (55% to 45%).

Squire also examined whether people who received the ballot were more likely to vote for Landon than those who did not receive it. Indeed, those that reported that they did not receive the *Literary Digest* poll were much more likely to vote for Roosevelt (71%) than Landon (27%). However, even those that reported that they did receive the poll were more likely to vote for Roosevelt (55% vs. 45%, respectively). Combined, these findings suggest that there were differences between those who did and did not receive the poll with respect to how they voted, and people of different income levels probably did vote differently. But in none of the subgroups of this sample discussed thus far was there a preference for Landon over Roosevelt.

A better explanation for the source of the failure of the poll to correctly forecast the election is *nonresponse* bias. Squire tabulated how the respondents in the 1937 poll voted as a function of whether or not they received and *returned* the ballot. The results were striking. Of those who received but did not return the ballot, 69% reported they voted for Roosevelt. Of those who both received *and* returned the ballot, the majority voted for Landon (51% vs. 48% for Roosevelt). So the failure of the *Literary Digest* poll can be attributed, at least in part, to the fact that those who participated in the poll by returning the ballot did not represent the views of the American public as a whole. Probably less important, although its influence cannot be ruled out, was the tendency for wealthier people to be more likely to receive a ballot, who were more likely to vote for Landon than the less wealthy. In the end, the results from their poll were not generalizable away from their data (the 2+ million respondents to the ballot) to the larger group the publishers of the *Digest* were interested in understanding—the U.S. voting population as a whole. The population inference was wrong.

It is tempting to attribute the failure of the *Literary Digest* poll to early 20^{th}-century methodological ignorance. Certainly, such a huge mistake could never happen again by expert, modern-day researchers. To be sure, there were lessons learned from the *Literary Digest* failure that changed the way that pollsters collect and interpret data. But there is evidence that the same processes that lead the *Literary Digest* to overestimate support for Landon in 1936 lead Edison Media Research and Mitofsky International, the companies that managed the November 3, 2004 exit polls in the U.S. Presidential election between Senator John Kerry and President George W. Bush, to overestimate how many people were voting for Kerry on election day. Exit polls are conducted just after voters cast their ballots, the results of which are given to the media so a winner can be forecasted after the polling stations close but before votes are counted. Historically, exit polls have been very accurate, but in this case Bush ended up winning many of states the exit poll data said should have been won by Kerry. The discrepancies between exit poll data and voter returns resulted in many conspiracy theories of election day fraud and vote rigging. But when the exit poll data were closely analyzed, it turned out that in many voting precincts, exit poll staff were more successful at recruiting Kerry supporters to participate in the exit poll than they were at recruiting Bush supporters (Edison Media Research & Mitofsky International, 2005). As a result, Kerry supporters were more likely to have their opinions registered in exit poll data, and the result was an inaccurate inference from sample to population. So even expert researchers well-versed in sampling theory and methodology are not immune to errors in inference.

3.1.2 Population Inference Through Representativeness

If the goal of the research is to generalize from the data to some population of interest (which was the goal for the *Literary Digest* publishers), we want to do our best to make sure that the sample is *representative* of the population. A sample is representative if it is *similar to the population in all important aspects relevant to the research.* To clarify, suppose you are interested in the political attitudes of the people of the city you live in. So you conduct a study by getting 500 residents of the city to fill out a questionnaire in which you ask various questions about political orientation, voting habits, and so forth. To the extent that the 500 people in your study are similar to the population as a whole on aspects relevant to your research, you can say that your sample is representative, and you can probably generalize your results from your sample to the population.

But on what important aspects relevant to the research would you want to see high correspondence between the sample and the population? For starters, you'd probably want to make sure that the sample contains men and women in about the same proportion as the population. We know that political attitudes of men and women are different, and if you want to generalize from sample to population, you'd want to make sure that your sample didn't substantially overrepresent males or females, lest your findings may be slanted in the direction of the overrepresented group. Similarly, income is also related to many political attitudes. If it was known that 20% of the city is lower class, 50% middle class, and 30% upper class (how these categories are defined is not important for our purposes), it would be important to make sure that your sample is similar in this respect. It would probably be all right, however, if your sample did not represent the population with respect to something irrelevant to the research, like pet ownership. Whether or not someone owns a pet is probably unrelated to their political attitudes, so you probably wouldn't care too much if 90% of your sample respondents had a pet even though only, say, 30% of the residents of the city own a pet. Of course, a difference on a seemingly irrelevant characteristic such as pet ownership might reflect other differences between the sample and the population that affect representativeness (e.g., perhaps pet owners are more likely to own a house and thus are more likely to have a higher income).

Although it is important that the sample be representative of the population when the goal is to make a population inference, we never really know whether or not it is representative. Why? Because you won't usually know in advance (a) all the aspects of the population that are relevant to the research and (b) just how the population differs from the sample on those characteristics. There are some things, however, that you can check after sampling, such as demographic similarity. For instance, the U.S.

Census Bureau keeps lots of statistics pertaining to the demographic make up of cities all over the United States. You could see if your sample is similar to the population by comparing the demographic characteristics of the sample to the statistics compiled by the Census Bureau. Or you could attempt to make a logical argument that even though your sample may not be representative of the population, you are safe in making a population inference from your data. Of course, you can argue all you want, and that doesn't mean you arguments are valid and the resulting inference will be correct. Therefore, it is best that you not end up in a situation having to defend your population inferences on logical grounds. Instead, maximize representativeness through the use of a good sampling strategy.

3.2 Sampling Methods

If your goal is population inference, you should focus your energy on obtaining research units that make it likely that your sample represents the population by the careful selection of sampling method. Although there is no way to guarantee a representative sample, there are things that you can do to enhance the likelihood that it will be representative. But first it is worth overviewing methods of sampling that tend to produce poor population inferences.

3.2.1 Nonprobability Sampling

A *nonprobability sampling method* is any method of recruiting or obtaining units for analysis in which inclusion in the sample is not determined by a random process. Just what is meant by a “random process” will be clearer in section 3.2.2 when I discuss probability sampling. For now, I describe two common methods of nonprobability sampling.

Convenience Sampling. If the investigator selects who (if the unit is people) or what (if the unit is something else, such as newspaper advertisements) provides data merely because those units are conveniently available to the researcher, then that investigator is conducting a *convenience sample*. Perhaps the people live in the same neighborhood as the investigator, or they are students in the researcher's college class, or they happen to shop at a local shopping mall. Or suppose a media researcher examines newspapers that he or she happens to subscribe to or that are available at the local newsstand or library. Although the use of convenience samples is very common in social science research, including communication research (see, e.g., Potter, Cooper, & Dupagne, 1993; Rossiter, 1976; or Sears, 1986), it is very difficult to make accurate population inferences from a sample of convenience. This is not to say that other forms of inference are not possible, and often (or even typically) population inference is not the goal of the researcher. But when it is, population inference rests on very shaky ground when a sample is based on convenience of the research units.

Why? The major problem is that there is no way of knowing that units conveniently available to the researcher are representative of the population to which the population inference is being made. For example, if I were interested in numerically estimating gender differences in communication apprehension by walking around the campus of The Ohio State University and asking men and women to fill out McCroskey's *Personal Report of Communication Apprehension* (see Rubin et al., 2004), all kinds of processes could render any difference (or lack of difference) I find ungeneralizable to people in general, college students in general, or even college students at The Ohio State

University. For instance, I might hesitate to approach someone if he or she appears angry, unhappy, or unsociable. Or I may have some other conscious or unconscious bias that leads me to approach only certain types of people, such as tall people or physically attractive people. To the extent that attributes of a person that lead me not to approach that person are related to what I am measuring (or things related to what I am measuring), my sample will be not be representative of any population, except perhaps the trivial population of people that I find approachable. I think it is easy to see how such people might differ in important ways related to communication apprehension. They may be especially outgoing, gregarious, and not at all anxious about communication. As a result, I may find no sex differences only because I only approached the more sociable and outgoing people I ran across wandering around campus. By the same reasoning, newspapers that a researcher happens to subscribe to or that are locally available may not represent newspapers in general. The community in which the researcher lives may be especially conservative or liberal, with newspapers in the area reflecting the political or social environment of the region. It would be difficult to make any kind of population inference (such as the proportion of editorials in major newspapers that are critical of the president's foreign policy) from what such a restricted sample of newspapers yields, unless the population of interest was the regional newspapers or newspapers with similar characteristics to those locally and conveniently available.

Another problem with samples of convenience is that it isn't clear just what the population being sampled is. Consider again my convenience sample of men and women I approach around campus. Just what is the population from which I sampled? Is it all people? Probably not. Is it young adults? Probably not. Students? People who live or work around The Ohio State University? I don't know, and there is no way I could determine what the population I sampled is. So even if the sample was representative, it isn't clear just what population it represents.

Volunteer Sampling. Another nonprobability sampling method that enjoys wide use is *volunteer sampling*. In a volunteer sample, the researcher recruits people to participate in the research, and those who volunteer to participate are included in the sample. In some sense, almost all samples of people are volunteer samples because you can't force a person to participate in a research study. What makes volunteer sampling distinguishable is the method of recruitment. Volunteer sampling procedures include advertising a study in the newspaper, posting a sign-up sheet in a classroom or around campus, or emailing potential participants. Volunteer samples suffer from the same problems as convenience samples, in that it is very difficult to know how representative your sample is of your population of interest, and when you obtain your sample, it usually isn't clear just what population your sample does represent. There is considerable evidence that (a) people interested in the topic of the research are more likely to volunteer to participate in that research, (b) when given the choice of several studies to participate in, volunteers who choose different studies sometimes differ in potentially important ways such as personality, (c) that students who volunteer to participate in research early in the academic term differ from those who volunteer later, (d) and that people who require more coaxing to participate in research may differ from those who more willingly participate (see, e.g., Abeles, Iscoe, & Brown, 1954–1955; Jackson, Procidano, & Cohen, 1989; Rosenthal & Rosnow, 1975; Wang & Jentsch, 1998). So there is no question that volunteers often differ from nonvolunteers in ways that might be relevant to the research. They are not likely to be representative of the population of interest, and population inference is often unwarranted and unjustifiable.

As I will soon discuss, convenience samples or other samples derived from a nonrandom sampling plan can be an efficient way of collecting data, depending on the kinds of inferences one wants to make. But if the goal is population inference, convenience samples should be avoided whenever better alternatives are possible.

3.2.2 Probability Sampling

Probability sampling methods are better suited to research where population inference is the goal. Probability sampling procedures are those where the determination of who or what provides data to the researcher is determined by a random process. In a probability sample, the only influence the investigator has in determining who is included in a sample is in the determination of the population to be sampled. Once the population is identified, the researcher (or his or her assistants or associates) has no say in who ends up in the sample. Random selection of the sample greatly increases the likelihood that the sample will be representative of the population in all aspects relevant to the research.

Simple Random Sampling. The simplest form of random sampling is *simple random sampling.* To conduct a simple random sample, all members of the population are identified and enumerated in the form of a list or database of some kind. Members of the population are then included in the sample through a random selection procedure, such as having a computer randomly select from the list, or even putting the names or identifiers of each member of the population on a slip of paper, putting the slips into a container and drawing out the desired number of members to be included in the sample. The key to conducting a simple random sample is to make sure that each member of the population has an equal chance of being included in the sample. If you are successful at doing so, this means that all possible samples of size n are equally likely samples. For example, if my population contained $N = 25$ members and I wanted a sample of size $n = 5$ from that population, there are 53,130 possible samples of size n $= 5$ (where that number comes from will be discussed in Chapter 7). A simple random sampling procedure guarantees that all 53,130 of these possible samples of size 5 are equally likely samples.

A simple random sample is in some ways the statistical ideal. However, in practice, it is difficult or impossible to collect a truly simple random sample because for most populations of interest to a researcher, there is no list or from which the sample can be derived in a simple random fashion.

Systematic Random Sampling. Another probability sampling method is *sequential random sampling.* In systematic or sequential random sampling, the researcher selects a random start point in the list of the population and then includes every k^{th} member of the population from that point, selecting k so that the desired sample size is obtained once the end of the list is reached. For example, if the population is of size $N = 100$ and a sample of size $n = 20$ is desired, a starting point in the list between 1 and 5 is randomly selected and then every fifth person in the list from that point is included in the sample.

Interestingly, this kind of sampling strategy is possible even if the population list is not available. Suppose a researcher wanted to sample people who visit a particular mall (the population of interest). The researcher could stand at the entrance to the mall over a series of days or weeks and interview or otherwise collect the relevant data from every 20^{th} person that enters the mall. If the researcher followed this procedure, then we know that he or she will not be influencing who is ultimately approached to provide data.

Stratified Random Sampling. Another kind of random sampling is *stratified random sampling.* A stratified random sampling procedure might be used if the researcher wants to make sure that the sample represents the population on one or more especially important dimensions, such as the distribution of males versus females. In a stratified random sample, the researcher first identifies two or more *strata*, defined by values on the stratification variable. A *stratum* (the singular of strata) is defined as a subset of the population the members of which share a common feature. For example, a researcher might want to stratify based on sex. Therefore, sex is the stratification variable, and the two strata are the males in the population and the females in the population. A stratified random sample is obtained by taking a simple random sample from each stratum, ensuring that the proportion of the total sample obtained from each stratum reflects the distribution of the population on the stratification variable.

For example, in a study I conducted on high school English teachers in California (Barnes & Hayes, 1995), we used a stratified random sampling procedure to select teachers to be interviewed. It was important in this study to make sure teachers at large, medium, and small schools (defined by enrollment) were represented according to their frequency in the population of schools. From a list of schools the state provided, it was possible to determine the proportion of schools that were large, medium, and small defined by enrollment. Knowing that 40% of schools were large, 40% were medium, and 20% were small using our definition of size, 40% of our sample was obtained by simple random sampling from the list of teachers at large schools, 40% was obtained by simple random sampling from the list of teachers at medium sized schools, and the remaining 20% were obtained from a simple random sample of teachers at small schools. As a result, the distribution of school size in the sample exactly mirrored the population distribution of school size.

If a simple random sample is difficult to obtain because of the difficulty of obtaining a list of the entire population from which the sample can be derived, a stratified random sample is often even more difficult because in addition to having a list from which to sample, you must have some kind of information that identifies which stratum in the population each member belongs. If that information is available, all well and good, but often it is not, making stratification impossible.

Cluster Sampling. Another random sampling method is *cluster sampling.* Cluster sampling is similar to stratified sampling in that the members of the population must first be identified and classified based on some characteristic. Whereas the groups sharing the stratification characteristic are called strata in stratified sampling, in cluster sampling those groups are called *clusters.* In cluster sampling, the population of clusters is randomly sampled. For each randomly selected cluster, *all* members of that cluster are included in the sample. This differs from stratified sampling, in that stratified sampling is based on a random sampling of members of each strata.

To illustrate the distinction, let's imagine you wanted to sample all members of the faculty at a particular university. To conduct a cluster sample, you might randomly select departments from the population of departments and then include in your sample all members of the departments that were randomly selected from the population of departments. By contrast, a stratified sample would randomly sample members from *each and every department* in such a way that the proportion of people in the *sample* from each department corresponds to the proportion of the entire faculty that resides in each of the departments.

Cluster sampling can also be useful for sampling a large geographical area. A map of a region can be divided into smaller regions (clusters), and then the population of clusters is randomly sampled by randomly selecting clusters and including every person living in those randomly selected clusters in the sample. Similarly, neighborhoods can be sampled by dividing up the neighborhood into city blocks and then randomly sampling city blocks and including each resident of those blocks that were selected in the sample.

Notice that cluster sampling does not necessarily require a list of all members of the population prior to sampling. For example, it might be tough to get an accurate list of employees at all North American McDonald's restaurant franchises. Indeed, the corporation might be quite reluctant to provide that information even if it had it available. But it would probably happily provide a list of all of its franchises in North America. You could randomly sample from the list of franchises and then approach the manager of each restaurant that was selected randomly and ask if he or she would be willing to let the employees of that restaurant participate in the research.

Random Digit Dialing. With the exception of cluster sampling, the sampling methods described above presume that it is possible to first enumerate the entire population prior to selecting a sample. There is another way of randomly sampling a population that enjoys widespread use in public opinion research and some other areas of communication research as well. The method of *random digit dialing* capitalizes on the fact that although there are few readily available lists of the population of a city, state, or country, most people can be identified with a phone number. When a researcher conducts a random sample using random digit dialing (often abbreviated *RDD*), people are contacted by dialing a random phone number. Whoever answers the phone and meets the desired selection criteria (e.g., someone over 18) is then included in the sample. By focusing the random dialing on a particular area code (the first three digits) or telephone exchange (the next three digits), it is possible to randomly sample small regions, such as states, cities, or even neighborhoods.

The use of the telephone to recruit participants has a number of clear disadvantages however. Not all telephone numbers are residential numbers. Many are businesses, fax machines, or disconnected. So it takes a lot of dialing to obtain the desired sample size because many phone numbers are not linked to a person. Furthermore, not everyone owns a phone, and we know that phone ownership is related to things that may bias the sample in important ways so that certain types of people (e.g., unemployed, homeless, or poor) are less likely to end up in a sample using RDD sampling. Relatedly, people who only use a cell phone will not be included in public opinion polls because of restrictions placed on pollsters and telemarketers on the calling of cell phone numbers for research and marketing purposes and the fact that there is no database of active cell phone numbers.[1] Also, some people have more than one phone number, meaning such people have a greater chance of being included in the sample. To the extent that having multiple phone lines or a cell phone only is related to things relevant to your research, you could end up with a sample that is biased in favor of such people or types of people.

For these and other reasons, it is relatively rare these days for researchers to use pure random digit dialing as a sampling method. More likely, a research firm will sample from a list of working telephone numbers compiled from one of many different databases. Because not all working telephone numbers are likely to be included in such

[1] This may change in the future, as more and more people abandon traditional “landline” phones in favor of cellular phones.

databases, the last few digits of the phone number might be randomly switched in an attempt to capture at least some unlisted numbers.

Sampling methodology is a very large and technical topic, and it is impossible to do it justice in just a few pages. Many books on the topic are available, and you could spend years developing knowledge and expertise in the area. Stuart (1984) provides a good and relatively nontechnical introduction to some of the more complicated issues in sampling. But the basic introduction I presented here should give you a feeling for the options available and used by communication researchers to recruit research units.

3.3 Is Nonprobability Sampling Really So Bad?

The probability sampling methods described above are the best approach to producing a sample that maximizes the ability to make sound population inferences. The random selection of participants, (if well conducted and the sample is not adulterated by response biases of some sort) allows the researcher to be reasonably confident that a sample represents the population from which sample was derived, and this representativeness affords the researcher the ability to make inferences about the population from information obtained from a sample of members of that population.

Given this, it might seem discouraging to acknowledge the fact that true probability sampling is rarely ever done in communication research. More typically, the samples that communication researchers collect are not selected randomly from any population. Indeed, the population from which a sample is derived is rarely explicitly defined either before or after data collection. Given this, how is it possible to make inferences from sample to population? The simple answer is that, technically, it isn't possible to make population inferences. However, in thinking about this problem, it is clear that the question is better framed not as whether or not it is *possible* but instead whether or not the researcher *wants* to make a population inference. If the researcher does not want to make a specific statistical statement about a population (such as females are 2.3 units more shy than males on average), then the question of whether the sample is random or not becomes moot. If the intent of the researcher is not to make a population inference but instead make a *process inference*, then the origin of the sample should loom less large in our evaluation of that research (Mook, 1983).

Just what do I mean by process inference? This concept is best understood by remembering that we often do research to test theory or a hypothesis (whether or not derived from a theory). Theories make predictions about what researchers should find in a research study motivated by the theory. Theory-driven research focuses less on estimating the size of an effect (such as the average difference between men or women on some measure in the population of interest) than it does on determining whether a prediction the theory makes about what should happen in a research study actually does happen (c.f., Frick, 1998; Mook, 1983). The data are collected, and the researcher analyzes the data to see if the data are consistent with the prediction that the theory makes. If so, then this provides some support to the theory. Remember that theories are explanations of a process. So if the theory is supported by the data, it is sensible to say that, at least in the circumstances in which the theory was tested, the process is probably at work and that in similar circumstances or situations, it is probably at work as well.

I need to be more concrete than this, so here is an example of what I mean. The *elaboration likelihood model of persuasion* (Petty & Cacioppo, 1986) postulates that people will deeply process the contents of a message aimed at persuading only if they are motivated and able to do so. How would you be able to determine if people who are motivated to process a message are actually thinking about the contents of the message more deeply than people less motivated? In other words, how would you be able to test this theoretical proposition? One approach taken by persuasion researchers has been to assess whether people's attitudes are affected by a manipulation of the strength of the arguments in the message. If people are paying attention to the message and thinking about the contents of the message at a deep level, then they should be more persuaded by strong arguments than by weak arguments. But if they are not paying attention and processing the message because they aren't motivated or able to do so, then strong arguments should be no more persuasive than weak ones. To test this, you might recruit some people who happen to be conveniently available to participate in a study. Perhaps the participants are college students or people who happened to respond to an advertisement you placed in the local paper in search of participants. First you determine whether the contents of the message will be of interest to them or relevant to each participant's life. For those people for whom the answer is yes, consider those people motivated to process the message. Those who don't care about the topic or who find it irrelevant to their lives you can consider unmotivated to process the message. Then randomly assign these participants to receive a persuasive message containing either mostly strong arguments or mostly weak arguments, and then assess their agreement with the message after they are exposed to it. The *elaboration likelihood model* predicts that the argument strength manipulation should have a bigger effect on the people motivated to process the message. For those who don't care about the topic or find it irrelevant to their lives, they would be less likely to notice whether the arguments are strong or weak because they aren't processing the message deeply when it is presented. So there should be relatively little difference between the strong and weak argument form of the message in terms of their agreement with it. In contrast, those who are motivated to attend to the message and process it more deeply should be more persuaded by strong arguments than by weak ones, as anyone paying attention to and processing the message deeply should be.

Suppose that you conducted this study using a group of university sophomores who were conveniently available to you, and you found exactly what the *elaboration likelihood model* predicts you should have found. From my discussion of sampling methods, it would seem that this tells you next to nothing given how the sample was obtained. Without knowing which population was sampled, who knows if the results are generalizable to any interesting population, such as "people in general" or even college students (a population of trivial interest in general unless you are particularly interested in studying college students specifically). But this criticism of the study is invalid on the grounds that process inference rather than population inference was the goal. The focus of the study was to test a theoretical proposition, which predicted that people motivated to process a message would attend to the contents and think about it more deeply. The fact that those highly motivated were more persuaded by strong arguments than weak ones compared to those less motivated suggests that the process the theory is attempting to explain about persuasion and the processing of messages is accurate. It matters not at all that the participants were not randomly selected from some larger population of interest because the intent of the study was process inference, not population inference.

And if the data are not consistent with the theoretical prediction? This suggests that either the theory is inaccurate, the researcher didn't adequately test the theory, or the theory has boundary conditions (i.e., it may not apply in all circumstances, or to all people, or it is sensitive to the choice of operationalization the investigator uses). As such, it is sensible to then say that the process at work producing the data is different than the process the theory proposes, it doesn't work this way in the circumstances or situation in which the researcher explicitly examined it, or it doesn't apply to the participants that actually participated in the study. As Mook (1983) aptly stated, it is the process at work that we are often interested in making inferences about, not the specific size of the effect in some population that we may (or may not) have sampled from. Importantly, process inference does not require us to randomly sample. The question as to whether the result generalizes beyond the conditions of the study or the people who participated is based not on the random sampling process but on *replication* of the research in different circumstances, using different methodologies, and different units of analysis. You can always criticize a study on the grounds that the people in the study perhaps had some quality that makes them different from everyone else (e.g., they are students, they are young, they are uneducated, and so forth). If that criticism is valid, if the findings are an artifact of the sample used, then future researchers will discover this when they fail to replicate your finding. Generalizability of a research result is an empirical question as much or more so than a statistical one.

But what about research that is not motivated by theory testing? The same argument applies, although the form of the argument is slightly different. A researcher may conduct a study motivated by curiosity using research units that are conveniently available. After collecting data, the researcher will have some finding, and a good researcher will probably at least attempt to speculate on why he or she found what was found. What is the process that produces the result found? And a good researcher will recognize that the generalizability of the finding away from the constraints of the design (both who provided data and how the data were collected) is dangerous. But those speculations or explanations, by advancing them in a research paper, become fuel for future research. Researchers have to take a "leap of faith" in their peers in the scientific community and trust that science is a self-correcting process (Aronson, 1977). If other people find our research interesting, it will motivate further research. It will serve as the source of some future researcher's predictions about what he or she should or might find in a study. If an explanation for a finding turns out to be wrong, future researchers will discover this. If that explanation is correct, the findings will be replicated when the explanation is put to the test in a new study. So the generalizability of our findings and the correctness of our explanations for a phenomenon are determined by replication not by random sampling (c.f., Amir & Sharon, 1991; Frick, 1998; Hendrick, 1991).

Random sampling also never results in a sample that is verifiably and unequivocally representative, so perhaps it shouldn't be put on a pedestal as the ideal method of recruiting research participants. Consider random digit dialing methods used frequently in public opinion polls. Not everyone is going to own a phone, and not everyone who owns a phone is going to be willing to answer your questions when you call them. There is no way of knowing the extent to which those variables (phone ownership and willingness to cooperate) are related to what you are measuring. How do you know for that certain that your sample doesn't under or over represent certain groups of people in important ways? The answer is that you can't be certain. To be sure, you could at least compare, for example, the demographic makeup of those who actually answered the phone and cooperated with you to see if the distribution of demographic variables in the sample matches population statistics derived from some place such as the U.S. Census Bureau.

Finally, it is important to keep in mind that all samples, including random ones, will yield results that do not necessarily generalize over time. When a public opinion pollster measures, for example, approval of the president's job performance, the result (say, 65%) is generalizable only to the population from which the sample was derived at that time. Next week, approval of the president's performance may be different. Populations change physically (people die and people are born continuously), and the social environment that influences that population on things you might be interested in measuring is in constant flux. Even process inferences may apply only to the time in which the study was conducted. As society changes, the social forces at work that communication theories often attempt to explain change as well, so a theory that is good now (i.e., one that is consistent with the data and makes accurate predictions) may not be an adequate description of the same process in the future. Of course, some theories are based on processes that are not likely to change substantially. We wouldn't expect major changes over time in the way people process information or interact face-to-face, for example, so communication theories that attempt to explain such processes probably are safe from substantial threats to validity resulting from population or social change.

3.4 Summary

Researchers are typically interested in some kind of inference. If population inference is the goal, random sampling methods are the most appropriate methods of recruiting research units. However, even nonrandom sampling plans can be useful to researchers more interested in process inferences. As this chapter illustrated, inference is best phrased as a question or set of questions. What form of inference do you want to make, population, process, or both? And does your sampling method and research design allow you to make the desired inference?