1

THE BIG PICTURE

The research process is a fascinating method for increasing the understanding of our world. The research objectives we may pursue using the research process are virtually limitless. Those who understand and apply this process have the power to determine truths by collecting and evaluating empirical evidence. It is our purpose in this chapter to introduce you to this process at a foundational level.

Let us begin by considering a few questions fundamental to the research process. Why do we collect data in the first place? What research questions do we pursue? How do those research questions inspire us to identify variables to measure as we seek to find out more and more about our questions of interest? How do research questions grow into research programs?

We, as researchers, often identify our research questions by starting with one interesting variable. That variable may be "depression" or "aggression" or "self-esteem," but whatever it is, it is something that we decide we want to know more about. And what we want to know about the variable often relates to a few questions:

What affects this interesting variable?

What is related to this interesting variable?

What does this interesting variable affect?

The first question allows researchers to identify and test other variables as potential antecedents, precursors, or causes of the variable of interest. The second question allows researchers to identify and test potential correlates to the variable of interest. The third question allows researchers to identify and test variables as outcomes of the variable of interest. For each of these questions, researchers collect data and then assess the evidence for the existence of these relationships. As relationships emerge, the research questions often get more complicated, including additional variables as antecedents, correlates, and/or outcomes related to our variable of interest. The models representing, and the analyses consequently used to test, the various relationships become more sophisticated.

But the underlying approach remains the same. In each case, the researchers investigate the extent to which their variables carry information about each other.

Arguably, the objective of all research is to collect and explain information about interesting variables. What this task amounts to, practically, is accounting for variance in our variables. Variables, by definition, take on multiple values, and through our research, we try to determine why these values are not all the same. Researchers attempt to identify other variables that provide some explanation for why these values are not all the same, then test the degree to which the variables share variance. That is, researchers assess how much the variables show some systematic deviation in their values by evaluating the direction and magnitude of relationships among the variables, and researchers build predictive models to summarize these relationships.

This process relies on collecting data to answer these questions; therefore, let's consider the data collection methods of a study investigating the effects of cyberbullying. The question this study attempted to answer was *How do participants feel after being cybervictimized?* (note these are partial data from Barlett, 2015). Participants came into the lab individually and were told that they would be interacting with a "partner" on a group task. Participants were told that they would be completing simple puzzle tasks with their partner. Specifically, participants were instructed to complete a Sudoku puzzle as quickly as they could while their partner waited. Once finished, the partner had to complete the next Sudoku puzzle, and, if it was completed, the participants tried to complete the third, and so forth. All participants were told that they and their partner had a combined 10 minutes to complete as many puzzles as they could. They were also informed that the number of puzzles completed would equal the number of raffle tickets they would receive. Each raffle ticket was an entry to a drawing to win an expensive prize. In actuality, there was no partner. Participants completed these tasks in their own room, having been told that their partner was in the next room. Also, the Sudoku puzzle chosen was unsolvable, and all participants failed. Once the 10-minute time frame had elapsed, participants were allowed to chat online with their partner for a few minutes while the researcher "got the next part of the study ready." Participants were then randomly assigned to receive an insulting ("You Suck") or nice ("It's OK") message from their partner. Finally, participants completed a measure of state hostility. (The instrument measured hostility using 34 items on a 1 = *not at all* to 5 = *extremely* rating scale; scores could range from 34 to 170, with higher scores indicating more hostility.)

The first research question we can ask is whether there was a difference in hostility between those who received the insulting and those who received the nice online messages. We can attempt to answer that question by comparing the hostility scores from the insulted group and the nice group, and we will do that in upcoming chapters. Prior to doing so, though, a good rule of thumb is this: *Look at your data before you do anything else*. We will describe a number of useful techniques for exploring your data in Chapter 2. Statistical techniques aid us in the interpretation and evaluation of our data, and they help us draw conclusions, based on our data, about the variables we study. The purpose of this book is to examine statistical theory and its relationship to research, particularly in the behavioral and social sciences where the use of statistics plays an important role.

An important point to keep in mind throughout this book is that statistical analysis is not an automatic method for interpreting data and drawing conclusions about our research questions and hypotheses. A great deal of subjective evaluation goes into using statistics and interpreting the results of statistical analyses, and the unwary researcher can stumble and be misled in numerous ways.

For example, a statistical formula or test can be applied to any set of numbers regardless of their origin. You can employ a t-test statistic to analyze two sets of numbers, derive a value for t, and compare that value to the critical values in a t-distribution table. How that value should be interpreted depends on the source of the numbers. Your interpretation of the results of a t-test on data from two groups of individuals who are either independent samples from different populations or were randomly assigned to conditions should be very different from your interpretation of the results of the same t-test comparing the performance of a single participant across a number of repeated observations. In this latter case, the scores in each treatment condition are not independent of each other, and the use of a t-statistic here for other than descriptive purposes would be inappropriate. Therefore, it is important to know and understand the assumptions involved in the use of various statistics and how these assumptions affect our interpretations of our data. These assumptions arise from the statistical models we employ to represent the sources of our data. Therefore, we will begin with a discussion of models in general and in statistics.

MODELS

The statistical models we choose to represent our data determine how we analyze and interpret those data. The assumptions we make about our data are based on the models we choose.

A **model** is an *attempt to represent or summarize what we believe to be the true state of affairs*. Accordingly, a model is inspired by the questions we pursue and the interesting variables we measure as part of our application of the research process. Since a model is a *summary* of what we *believe* to be true, it is by definition not a perfect representation of what is being modeled. Nevertheless, models are more than mere guesses. In science and in statistics, they are interrelated sets of concepts that are used to describe data and make predictions, and they are based on systematic observations.

A simple example of a model is a road map that we use to represent the routes and landmarks (rivers, cities, connecting roads, and so forth) between various locations. The road map is useful because it presents the important information in a compact form with many details omitted. If the road map were a perfect representation of the geography of the route, it would be as big as the terrain it represents and too cumbersome to use. Like a road map, a model is an imperfect representation and thus will always be incorrect; however, it can be useful if it helps us to understand the phenomenon we are attempting to model and to communicate that understanding to others.

In science, we construct models to help us to understand and predict. These models can be relatively simple (for example, Newton's laws of motion and Fechner's law in perception) or complex (for example, Freud's psychoanalytic theory of behavior). In either case, because they are models, they are imperfect representations of nature, and while they may be consistent with a large number of observations, we can find situations that they do not represent accurately.

Sometimes there are competing models of the same phenomena (for example, Guthrie's, Hull's, Skinner's, and Tolman's models of learning). Each of these models contributes something to our understanding of learning, and consequently we use them at various times to make predictions or provide explanations while recognizing that each is inadequate as a universal model that accurately represents all data.

The fact that models are not exact representations of reality raises the question of why we bother to use them. There are a number of reasons. First, models help us to organize our data in a meaningful way because they are less complex than the phenomena being modeled. Second, they help us to interpret our data and to derive what appear to be general principles or rules. Finally, a well-specified model can be manipulated mathematically or logically to derive predictions or new ways of viewing our data. We choose among alternative models based on how well they perform with respect to each of the reasons given above. A model is of value when it adds to our understanding and leads to useful predictions, even if it is not completely accurate.

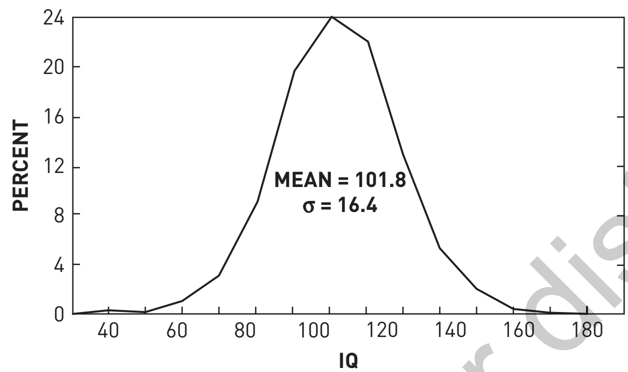
Just as there are models of physical objects and their interactions, and models of the behavior of animals and humans, there are models of numbers and data. In general, there are two classes of statistical models, descriptive models and inferential models.

Descriptive Models

Descriptive models *are used to represent and summarize our data*. Thus, when we say that intelligence test scores have a normal distribution, we are using the normal distribution model to represent those scores. An example of this model is presented in Figure 1.1, which presents the IQ test scores of 3,000 children, ages 2–18, who comprised the standardization group for the 1937 version of the Stanford-Binet.

Clearly, the scores in Figure 1.1 do not create a perfect normal distribution (nor would they even if the sample were larger). When we say that intelligence test scores

Figure 1.1 ■ The distribution of IQ test scores of the standardization group on the 1937 revision of the Stanford-Binet intelligence test.



Source: From Terman, L. M., & Merrill, M. A. (1973). *Stanford-Binet Intelligence Scale: Manual for the Third Revision, Form L-M*. Boston, MA: Houghton Mifflin.

have a normal distribution, we mean that these data approximate a normal distribution, and we can use the normal distribution model to describe the data. While these scores do not perfectly fit the normal distribution, the fit is good enough for us to use this model as a convenient shorthand way to describe the data and to use the normal distribution tables to calculate percentages of scores within various IQ test score ranges.

Descriptive statistics are as important as inferential statistics for the analysis of data, even if we are not interested in trying to fit a particular descriptive model (such as a normal distribution) to our data. Many researchers overlook the importance of examining their data prior to performing further analyses. Failing to look at how well the data fit the descriptive model that informs the assumptions underlying further analyses can result in misanalysis of our data. New and interesting techniques for describing sets of data, which are useful for the social and behavioral sciences, are not widely used by psychologists and other social scientists because they are not widely known; therefore, we will devote some attention to these techniques in Chapter 2.

Inferential Models

We use **inferential models** *to draw inferences from our data*. By making certain assumptions about the behavior of the researcher with respect to the collection of the data and about certain aspects of the data, these models provide methods for testing statistical hypotheses and drawing conclusions. Analysis of variance, *t*-tests, and other "parametric tests" are based on what we refer to here as the *classical statistical model*. Most of this book will be devoted to the classical statistical model and the implications of that model for data analysis.

A second inferential model, the *randomization model*, will be presented in some detail in Chapters 9, 10, and 14. That model makes different assumptions about the behavior of the experimenter and the data. Interestingly enough, although the assumptions of the randomization model are more realistic with respect to the way experimental research is conducted in psychology, statistical techniques based on this model have not been widely used in psychology for a number of reasons. As mentioned, the most important reason is that most psychologists are also not familiar with this approach—but more about that later.[1]

Given the widespread use of the classical statistical model, we will describe it in some detail now.

THE CLASSICAL STATISTICAL MODEL

The classical statistical model posits that our data are randomly sampled from populations that have certain parameters, some of which are unknown to us. Our task is to use the information in our random samples to estimate the values of the unknown parameters.

The inference model most widely used in social and behavioral science research is the classical statistical model. To use the classical statistical model, one has to make two assumptions, one about the data and one about the behavior of the experimenter.

Assumption about the data:

Populations with certain attributes (called parameters*) exist.*

The elements of these populations can be conceptualized in two ways. We tend to think of the elements of a population as individuals (people, rats, and so forth), and we assume that our study sample is composed of a subset of these individuals. However, in the behavioral sciences, the object of our inquiry is not the individual but the behavior of the individual on the measurement scale we choose to study (e.g., test score, number of correct responses, response rate, and so forth). Therefore, it is more appropriate to think of our study sample as being a subset of the scores or numbers generated by the individuals in our population of interest.

The **parameters** (attributes) of these populations *describe the populations*. These parameters typically include the population mean, population variance, and form or shape of the population (normal distribution, exponential distribution, and so forth).

[1] These are not the only two statistical classes of inferential statistical models. Another is the Bayesian model, which is not covered in this book.

Some of these parameters may be assumed to be known to us,[2] and some parameters are assumed to be unknown. The researcher's job is to infer the unknown parameters from the data, using the help of another assumption.

Assumption about the behavior of the experimenter:

Random samples are drawn from the population(s) under study.

In a **random sample**, *each element in the population has an equal chance of being sampled*. The fact that some elements are sampled more frequently than others occurs because more of these elements exist in the population. For example, in a normal distribution, there are more elements in the center of the distribution than at the tails; hence, random samples from normal distributions will tend to contain more elements from the center.

The **purpose of random sampling** is *to obtain a set of scores (a sample) that is representative of the population*, meaning that the sample represents the distribution of the scores in the population. Unless we take the entire population as our sample, we would not expect our sample to match the population exactly. Nevertheless, we hope that by using a random sampling procedure, our sample will adequately reflect the characteristics of the population. We must emphasize that while random sampling does not guarantee that our sample will be representative of our population, a random sample will usually contain enough information for us to make useful inferences about the characteristics of that population.

Despite the importance of the assumption of random sampling for using the classical statistical model, random sampling is a rare event in much of the research in the social and behavioral sciences. Nevertheless, use of the classical statistical model is based on the assumption that random sampling has occurred.

Sometimes, we add the additional assumption that we have *independent* random samples. The assumption that these samples are composed of independent events is of great importance for the correct use of this model in social science research. We have an **independent event** *when the occurrence of one event does not affect the probability of the occurrence of another event*. In terms of sampling, the selection of any element in a sample does not affect the selection of any other element; in other words, the selection of one element does not change the probability that another element might be selected.

Non-independence of events occurs in a number of situations: When sampling is done without replacement, sampling of one element of the population changes the probability that the others will be sampled. Players of card games such as draw poker and blackjack know that the appearance of a given card on the table affects the probability that they will draw another card of the same face value. For example, if no aces are showing (and assuming no aces are

[2]For example, Student's *t*-test statistic (see Chapter 7) and analysis of variance (see Chapter 14) are based on assumptions that our data are independent random samples drawn from normal distributions that have equal variances.

already face down on the table), your probability of getting an ace is 4 divided by the number of cards not showing, but if there are already 2 aces face up on the table (and again assuming no other aces are already face down on the table), the probability of drawing another ace is only 2 divided by the number of cards not showing. For the most part, the populations under study in the social and behavioral sciences are (or are assumed to be) so large that sampling without replacement is really not an issue. In the behavioral sciences, non-independent events occur, for example, when we measure the behavior of the same individual more than once. Clearly, there is a relationship between an individual's behavior at one point in time and his or her behavior at a later time. We will examine how to analyze data from non-independent observations later when we discuss the randomization model in Chapter 9.

Having introduced the classical statistical model, let's examine how statistical theory and practice impact how we analyze our data. We will use as an example the experiment we described earlier in this chapter on the effects of online message content on hostility.

DESIGNING EXPERIMENTS AND ANALYZING DATA

In this section, we will examine the steps a researcher would take to design and execute an experiment and analyze the data to try to answer the research question. The researcher must generate a set of experimental hypotheses that specify possible outcomes of the experiment as well as a set of statistical hypotheses to test.

The experiment described at the beginning of this chapter was designed to attempt to answer the following research question: Does being a victim of cyberbullying influence hostility levels? The research design used here is rather simple, and it is not the only possible design or even the most optimal design for this type of study; however, it does allow us to see the issues relevant to designing experiments and analyzing their data clearly.

The first step in any experiment is to decide *specifically* what you want to investigate—in other words, *specifically* what questions you want to ask and hopefully answer. Here the experimenter was interested in whether cybervictimization affects later state hostility. To study this variable, the experimenter had to devise a way to manipulate the participants' behavior through the messages they received and a way to measure the hostility that might result from that experience. Respectively, these components are the independent and dependent variables in this experiment; that is, **independent variables** *are the aspects of the situation the researcher manipulated*, and **dependent variables** *are the participant behaviors the researcher measured.* While these **operational definitions** (the specific ways in which variables are represented or measured in research studies) are properly matters of experimental design, one cannot ask a meaningful question about the effects of cybervictimization on participants' consequent hostility without also considering the context in which the independent variables are presented.

Because an experiment requires at least two levels of the independent variable (e.g., experimental vs. control condition), the experimenter had to select at least two types of electronic communication that were similar in terms of length and number of messages but were very different in terms of content. The idea behind the messages was simple: To possibly elicit hostility from the participant in an ecologically valid and ethical way, the participant had to be cyberbullied. Therefore, the experimenter went to great lengths to convince the participants that they were interacting with another partner they could not see, they were working together to solve the puzzles so that they could win a prize, there would be a raffle and prizes awarded for completing these puzzles, and their partner sent the message they received online.

The experimenter could have used a number of dependent measures to assess the effects of these messages. As described earlier, the experimenter assessed the effects of the "partner's" messages by measuring state levels of hostility after participants received their message from their "partner." Hostility was measured using the State Hostility Scale (SHS; Anderson, Deuser, & DeNeve, 1995), with higher scores being indicative of feeling more hostile. Accordingly, participants were asked to indicate how they felt at that moment by reporting their levels of agreement with several statements, such as *I feel furious*.

Experimental Hypotheses

Experimental hypotheses are possible answers to your research question. They should be mutually exclusive (only one can be true) and exhaustive (these are the only possibilities), and they are stated in terms of the independent and dependent variables in your experiment.

Having defined the task, the next step is to generate a set of **experimental hypotheses**: *Statements about what may or may not be the actual state of affairs in terms of the relationships between the variables you are examining.* Therefore, the experimental hypotheses are phrased in terms of the independent and dependent variables in the experiment and in such a way as to be **mutually exclusive** (they do not overlap such that *only one of them can be true*) and **exhaustive** (*they are the only possibilities*). There are two possibilities for the effects of cybervictimization (message content) on state hostility in this experiment:

1. Cybervictimization (message content) *does not* affect state hostility.
2. Cybervictimization (message content) *does* affect state hostility.

Note that these experimental hypotheses are mutually exclusive and exhaustive because *only one of them can be true* and *they are the only possibilities*. The researcher's task is to discover which of these statements is true. If the researcher knew which experimental hypothesis was true, there would be no reason to conduct the experiment, but because the researcher does not know which statement is true, he or she collects and analyzes data to decide which one is true. However, keep in mind that the number of messages and severity of the derogatory comments in the messages may influence the strength of the effect of cybervictimization on hostility. Thus, the design of the experiment and the operationalization of the variables may affect the outcome and the conclusions about which of the experimental hypotheses is true.

Experimental Design

While many of the details regarding how you conduct an experiment are not statistical issues, they can affect the interpretation of your data, especially if the experiment is not designed properly.

In the experimental design phase of the research process, the researcher decides how to conduct the study. How will the participants be obtained? How many will there be? How many conditions will there be in the experiment? How will the participants be treated in each condition? Exactly what will be measured? With the exception of the questions about how the participants are obtained and the number of participants, the rest of these questions do not involve statistical issues; they are matters of proper experimental procedure.

How we design an experiment that examines the effects of cybervictimization on hostility depends on what we wish to learn about these variables. Because we stated our experimental hypotheses in terms of whether or not cybervictimization has an effect on hostility, a two-group (experimental group and control group) design should be sufficient. In the case of the Barlett (2015) study, one group was designated the experimental group and received insulting messages. The other group was designated the control group and received nice messages. Both groups were treated the same in all other respects; the only difference in their treatment was *in terms of the variable under investigation* (i.e., the type of messages received). That is, participants in both the experimental and control groups received the same instructions, interacted with the same experimenters, used the same equipment, and so forth. If care is not taken to eliminate as many potentially confounding variables as possible from the experimental design, the researcher will not be able to determine whether any of the observed differences between the performance of the participants in these two groups are due to the treatments (such as types of messages) or to other factors (such as the participants' knowledge of the treatment they received or expectations about how they should respond). Again, these points are not statistical issues, but considerations regarding proper experimental design procedures that affect the internal validity of the study and, thus, can impact the interpretation of the study's results.

HOW WILL THE PARTICIPANTS BE OBTAINED? In most experimental research in psychology, participants are recruited as "volunteers" through a system whereby they are enticed to participate by offers of credit toward course requirements, money, or some other compensation for their services. Of course, there are procedures to remove coercion from this process by allowing for alternative methods of earning credit or money, by informing participants of the risks and possible benefits of their participation, and by allowing participants to withdraw at any time without loss of credit or money. Having volunteered, and if they are accepted into the experiment, participants are then assigned to the different conditions of the experiment.

As noted above, one of the assumptions of the classical statistical model is that our data are obtained by random sampling from populations. However, it is obvious that there is nothing remotely random about the manner in which participants are obtained for our experiments. Those who volunteer to be participants do so for various reasons (experimental credit, money, curiosity, and so forth) and are certainly not representative of college students or any other group from which volunteers are solicited. Here is the first instance of a deviation from the assumptions of the classical statistical model. However, this deviation may be somewhat addressed by studying psychological processes that are reasonably assumed to operate similarly in the larger population and by using random assignment to conditions.

RANDOM ASSIGNMENT TO CONDITIONS Random assignment to groups helps us to make up for the fact that we do not use random sampling in experimental research. **Random assignment** is *a procedure for assigning study participants to the various conditions of an experiment, regardless of how the participants are obtained*.

The purpose of the random assignment procedure is to attempt to remove any systematic effects that might produce the observed differences in performance between the groups. For example, we could use the participants' middle initials, flip a coin for each participant (when there are two groups, as in this example), or use a table of random numbers. In this way, the participants in each condition are not systematically different from each other prior to their participation in the study. That is, any differences between the participants in the various conditions prior to the study's manipulations would be random. If, following random assignment, both groups are treated the same—that is, given the same treatment and then tested—the sample statistics (e.g., means, variances, medians) for both groups would tend to be similar (but not necessarily identical), and any differences between the sample statistics for the two groups would presumably not be due to any systematic effects (such as assigning all females to one group, all participants with high IQs to one group, or all older participants to one group). Of course, it is possible for random assignment to produce such a "nonrandom" outcome, but such a possibility is rare; therefore, it should be noted that random assignment (and random sampling) refers to procedures and not to the results of those procedures. Thus, random assignment does not necessarily create groups that are equivalent on the variable of interest. Rather, random assignment attempts to create groups in which the differences on the variable of interest are random, and such randomness (or error) can then be accounted for by the statistical techniques we use to analyze our data.

By using random assignment and then applying different treatments to the two groups, we should feel confident that any large differences in the sample statistics are due to the treatments and not to other factors. Of course, it is possible for there to be another systematic factor operating (that is, a confound), but, as noted above, we can rightly assume such an occurrence is unlikely if we have been careful in how we treat the participants and in how we randomly assigned them to conditions. On the other hand, if we do not use random assignment and we apply different treatments to our two groups, then we will not know whether any large differences in performance are due to the treatments or to any other systematic difference between the groups.

EXPERIMENTAL TREATMENTS Following random assignment to groups, the next steps are the administration of the experimental treatments and the collection of data. The experimental group in our representative experiment received the “insulting” message and then was given the State Hostility Scale to complete.

To assess the effects of cybervictimization on hostility, the researcher needed also to measure the performance of a group that received the “nice” message; this group is the control group. The term *control* is used to indicate that all relevant variables other than the treatment under consideration were controlled; that is, with the exception of the experimental treatment, the two groups were treated identically. If the only difference between the two groups is the presence or absence of the experimental treatment, then the control group provides a baseline against which to compare the data obtained from the experimental group. Any large differences in behavior between these groups would then be attributed to the experimental treatment. For this reason, the control group is sometimes referred to as a *baseline condition or group* (because the results for the control group are what you would expect to observe in a default situation in which the experimental manipulation, in this case the “insulting” message, did not occur) or as a *comparison condition or group*.

Failure to adequately control for other systematic variables compromises the interpretation of the results of the experiment. As in the case of nonrandom assignment, the presence of other systematic differences between groups (such as differential instructions, testing at different times of the day or year, use of different experimenters, and so forth) produces a confounded design. Such a design makes it impossible to determine whether the treatment produced the differences between the groups or the systematic differences that accompanied but did not comprise the experimental manipulation produced the differences between the groups. These are matters of good experimental design procedures and not statistical issues (although nonrandom assignment is a statistical issue);

nevertheless, they should not be overlooked. Statistical methods cannot untangle a confounded design unless the confound is purposely part of the original design.

Data Analysis

> *Inspection of your data should influence which statistical model you select to analyze your data. If you select the classical statistical model, then your statistical hypotheses will be about the parameters of the populations from which your data were presumably sampled.*

After the data are collected, the process shifts to one of analysis. *The first step in any data analysis is to inspect the data*. Inspection of the data is usually facilitated by calculating summary statistics of your data (such as the mean, standard deviation, and median), by organizing the data into tables, and by visually presenting the data using graphs and plots. Inspection of the data in these ways will inform decisions about which inferential statistics to use to analyze the data, and these tables, plots, and graphs may then be used to reinforce the conclusions we will draw from our inferential statistics. We will discuss this step in detail in Chapter 2.

Should it be necessary to employ inferential statistics to decide whether the differences we observe in our data are due to the experimental treatments or to chance, a set of statistical hypotheses have to be developed. These statistical hypotheses parallel our experimental hypotheses. If we decide to use the classical statistical model, *our **statistical hypotheses** will be about the parameters of the populations from which we presumably sampled*.

For our example experiment, there are two experimental hypotheses:

1. Message content (cybervictimization) does not affect hostility.
2. Message content (cybervictimization) does affect hostility.

The researcher's task is to attempt to determine from the data which of these hypotheses is true. As stated, the experimental hypotheses cannot both be true, yet one must be true (because they are both mutually exclusive and exhaustive). For this situation, a corresponding set of statistical hypotheses using the classical statistical model are as follows:

1. $\mu_{\text{insulting message}} = \mu_{\text{nice message.}}$ The mean score on the state hostility measure of the population from which the participants who received the insulting message were sampled equals the mean score of the population from which the participants who received the nice message were sampled.
2. $\mu_{\text{insulting message}} \neq \mu_{\text{nice message.}}$ The mean score on the state hostility measure of the population from which the participants who received the insulting message were sampled does not equal the mean score of the population from which the participants who received the nice message were sampled.

Note that the statistical hypotheses are phrased in terms of population parameters, *not* in terms of sample statistics.[3] The statement $\overline{X}_{\text{insulting message}} = \overline{X}_{\text{nice message}}$ can be tested directly by inspecting the data: Either the sample means are equal, or they are not. Note that the statement about the difference between the sample means is *not* a correct mathematical formulation of our statistical hypotheses, because hypotheses are about things we cannot directly observe. We can observe whether the sample means are equal or not, but we cannot observe the population means. If we could observe the population means, we would have no reason to collect sample data!

It is on the basis of the results of our statistical hypothesis tests that we use our data to infer which of our various statistical hypotheses is likely to be correct. Using the classical statistical model for our representative experiment, a *t*-test seems most appropriate. We will discuss the rationale for using the *t*-test in this situation in Chapter 7, after we have mastered some basic facts and concepts pertaining to the classical statistical model.

WHERE ARE THE POPULATIONS? With the classical statistical model, we assume that the participants in each condition of our experiment represent a random sample from various populations. While we acknowledge that we rarely use random sampling in experimental research, when we use this model, we identify our participants as being sampled from populations and therefore serving as representatives of those populations. Thus, the hypotheses we test are about the parameters of these populations.

Can we identify the populations from which our participants might be randomly sampled? Where, for example, is the population from which we would draw a sample of participants who are in our experimental group? The answer to these questions is that *these populations do not exist outside of our experiment*. We create both the experimental participants (and the control participants) by giving them the treatments appropriate to their respective conditions. If we want more participants for either condition, we can create them with the appropriate treatment. We could theoretically have 1,000 people receive the insulting message and then randomly sample a small number of them to be in the experimental group. The parameters in our statistical hypotheses would be the attributes of these "created" populations. We call these populations **potential populations** because *although they do not exist outside of our experiments, they could exist if we chose to create them. In other words, they could potentially exist.*

Obviously, we need the concept of potential populations in order to use the classical statistical model. The fact that these populations do not really exist should cause us little concern. The purpose of inferential statistics is not only to test statistical hypotheses but

[3]The convention for using Greek letters to represent parameters of a population and Roman letters to represent estimates of those parameters and sample statistics was adopted in the 1920s and 1930s, although there are a few exceptions (some of which will appear in this book).

to make inferences about our experimental hypotheses. In this sense, these populations (and their parameters) are like the hypothetical constructs and intervening variables of many psychological theories; that is, they are useful abstractions that aid us in our work.

From Statistical Hypotheses to Experimental Hypotheses

The goal of research is to answer our research questions. Therefore, an important step is to make correct inferences about our experimental hypotheses from the results of statistical analyses of our statistical hypotheses. We can use good experimental design and perform all of the statistical hypothesis tests correctly yet still not produce a correct answer to our research question.

The results of our statistical hypothesis tests allow us to make inferences from our data to our statistical hypotheses, meaning that we can infer which of our hypotheses we believe to be correct. This, however, is not the end of the story. We conduct our research to learn about the effects of our independent variables on behavior, not to learn about population parameters. *The goal of experimental research is to make inferences about experimental hypotheses*, in this case, the effects of cybervictimization on hostility.

What could the researcher conclude in this experiment if he or she were to accept the statistical hypothesis (on the basis of the statistical hypothesis test on the data) that, for example, $\mu_{\text{insulting message}} \neq \mu_{\text{nice message}}$ (the mean of the population from which the participants who received the insulting message were sampled does not equal the mean of the population from which the participants who received the nice message were sampled)? If, for example, the sample mean for the insulting message group were higher than the sample mean for the nice message group, should we conclude that insulting messages increase hostility? The answer is clearly no! We can only justifiably say that receipt of the insulting message used in this experiment by this group of participants (who may be distinct from other populations in terms of age, gender, social class, educational level, and so forth) affected hostility as measured in this experiment. It would take more data from more experiments to claim more than that. The matter of inference from the results of statistical hypothesis tests to experimental hypotheses is more a matter of logic than statistics; nevertheless, it is an aspect of the research enterprise in which mistakes are frequently made. For an excellent discussion of this topic in the context of statistical hypothesis testing, you should read David Bakan's 1966 paper.

There is also the possibility that the limited conclusion will not stand the test of replication; that is, it may be that results similar to those obtained in one experiment will not occur again when the experiment is repeated. We will discuss the matter of errors in hypothesis testing and the failure to replicate the results of an experiment later in Chapter 6.

Summary

In this chapter, we introduced the purpose of the research process and a general overview of the steps comprising this process. To illustrate the discussion, we used as an example a study examining the effects of cyberbullying.

We noted that all research starts with a research question or questions. For each question, we frame possible answers as experimental hypotheses, statements about the effects of an independent variable (or variables) on dependent variables of interest to us. These hypotheses guide the design of our experiment as we decide how to operationalize our independent and dependent variables. Eliminating possible confounds, while not necessarily a statistical issue, is important because confounds can compromise the effectiveness of our experiments in providing answers to our research questions.

After we collect our data, we need to look at those data to decide what statistical model to use for analysis. The model we choose will determine how we frame our statistical hypotheses. When we use the classical statistical model, which is based on the assumptions that populations of scores exist and we can learn the values of unknown parameters by taking random samples from those populations, our statistical hypotheses are about the parameters of those populations.

Although random sampling from populations is an assumption of the classical statistical model, we do not take random samples when we conduct experiments; rather, we assemble a number of participants and randomly assign them to the conditions in our experiment. We invoke the concept of potential populations, populations that do not exist outside our experiment (but could be created by giving more individuals our experimental treatments) to allow us to use the classical statistical model with experiments.

In Box 1.1, you will find a flowchart that you can use as a guide to the structure of the research process.

BOX 1.1

FLOWCHART FOR EXPERIMENTS

Step 1: Generate your research question

Example: How do participants feel after being cybervictimized?

Step 2: Generate experimental hypotheses as possible answers to your research question

These are statements about what may or may not be the actual state of affairs. They should be mutually exclusive and exhaustive.

Example: Cybervictimization does not produce feelings of hostility.

Cybervictimization produces feelings of hostility.

Step 3: Design your experiment

Choose independent and dependent variables that define the elements of your experimental hypotheses.

Examples: Cybervictimization is defined by the content of the message someone receives. Hostility is measured by responses to the State Hostility Scale (SHS).

Decide what the participants will experience in the different conditions in your experiment.

(Continued)

BOX 1.1 (Continued)

FLOWCHART FOR EXPERIMENTS

Decide who participants will be, how you will recruit or obtain them, how many will be in your experiment, and how they will be assigned to the different conditions.

Step 4: Collect your data

Step 5: Decide what statistical model to use to analyze your data

Step 6: Generate your statistical hypotheses

For the classical statistical model, your statistical hypotheses will be statements about the parameters of the population from which you presumably obtained your samples.

Example: H_0: $\mu_{\text{insulting message}} = \mu_{\text{nice message.}}$ The mean score on the state hostility measure of the population from which the participants who received the insulting message were sampled equals the mean score of the population from which the participants who received the nice message were sampled.

H_1: $\mu_{\text{insulting message}} \neq \mu_{\text{nice message.}}$ The mean score on the state hostility measure of the population from which the participants who received the insulting message were sampled does not equal the mean score of the population from which the participants who received the nice message were sampled.

Step 7: Perform the statistical analyses and draw the appropriate conclusions about your statistical hypotheses

Step 8: Draw the appropriate conclusions about your experimental hypotheses to answer your research question

Questions Raised by the Use of the Classical Statistical Model

The first section of this book will be devoted to answering the following questions about the proper use of the classical statistical model:

1. To test hypotheses about population parameters with the classical statistical model, we must extract information from our samples to estimate these parameters. The question is, *What information from our samples should we use for these estimates?* It is customary to use the sample mean as the "best" estimate of the population mean, but in the case of the estimate of the population variance, the best estimate is not the sample variance. Why that is the case and how we determine the "best" estimates for these parameters is the topic of Chapter 4.

2. The statistical hypotheses in our example experiment are the following:
   a. $\mu_{\text{insulting message}} = \mu_{\text{nice message.}}$ The mean of the population from which the participants who received the insulting message were sampled equals the mean of the population from which the participants who received the nice message were sampled.

b. $\mu_{\text{insulting message}} \neq \mu_{\text{nice message.}}$ The mean of the population from which the participants who received the insulting message were sampled is not equal to the mean of the population from which the participants who received the nice message were sampled.

Hypothesis (a) is usually referred to as the "null hypothesis" and hypothesis (b) as the "alternative hypothesis." Why are they stated as they are and given those names? Questions 2–8 will be addressed in Chapter 6.

3. Even though we might collect data that lead us to reject the null hypothesis ($\mu_{\text{insulting message}} = \mu_{\text{nice message}}$) and accept the alternative hypothesis ($\mu_{\text{insulting message}} \neq \mu_{\text{nice message}}$), can we conclude anything about the *direction* of the effect (whether insulting messages increase or decrease hostility)? And when we expect a result in one direction but not the other, can we state our alternative hypothesis in that direction and accept it only when the difference is in that direction?
4. Our statistical hypotheses, like our experimental hypotheses, are stated so that they are mutually exclusive (only one can be true in a given situation) and exhaustive (there are no other possibilities such that one of these must be true). Because of these properties, if we can reject the null hypothesis on the basis of our data, we can accept the alternative hypothesis. While it is logically possible to accept an alternative hypothesis by testing and rejecting the null hypothesis, why can we not logically test the alternative hypothesis directly?
5. On the basis of what criterion do we reject the null hypothesis?
6. While it is possible to reject the null hypothesis, can we ever accept it? Is non-rejection of the null hypothesis the same as acceptance of it? What conclusion can we draw about our experimental hypotheses when we cannot reject the null hypothesis?
7. When we reject the null hypothesis and accept an alternative, we say that we have found a "significant difference" or that the difference is "statistically significant." What do we mean by this statement?
8. The decision about the number of participants to be included in an experiment is an important one. While it certainly is true that it is easier to demonstrate that the observed difference between two sets of numbers is not due to chance with a large number of participants, the use of large numbers of participants can make small differences "statistically significant." Furthermore, the use of too many participants can be wasteful of time and money when the same inferences could be drawn from a smaller number of participants. How can we decide how many participants to use in an experiment in order to maximize our chances of finding an effect if there is one, while at the same time not finding spurious effects or wasting resources?
9. What are the effects of violating the assumptions of the classical statistical model? We have already briefly looked at the fact that we usually do not draw random samples from populations and how we deal with the violation of that assumption. For the statistical tests based on this

model, assumptions have to be made about the forms of the (potential) populations from which "random samples" are taken. For example, in the case of the *t*-test suggested for our example experiment, it is assumed that the populations have normal distributions with equal variances. But suppose this is not the case? How would we be able to test these assumptions when our populations really do not exist outside of our experiment? We provide some answers in Chapter 2.

10. What is the appropriate test statistic for testing various statistical hypotheses in different situations? We provide answers in Chapters 6 through 15.

Before we answer these questions, however, we need to step back and learn about some interesting techniques for looking at our data and testing assumptions about our populations. We will accomplish that goal in the next chapter.

Conceptual Exercises

1. Why do we state our statistical hypotheses in terms of population parameters and not in terms of sample statistics?
2. Write the following statement in an English sentence:

$$\mu_1 \neq \mu_2$$

3. What is a potential population? Why do we need this concept for the classical statistical model?
4. What are the two assumptions of the classical statistical model? With respect to these two assumptions, how do we actually do things in psychology?
5. What is the difference between experimental hypotheses and statistical hypotheses? Why do we need both sets of hypotheses when we use the classical statistical model to analyze our data?
6. What is wrong with the following statement?

$$H_0 = \overline{X}_E = \overline{X}_C$$

How should it be written? Write the correct statement in an English sentence.

7. What is the difference between random sampling and random assignment? What is the purpose of each procedure? When do we use each?
8. Self-fulfilling prophecies involve having strong beliefs that something will happen and then unknowingly acting in a way to make it happen. We tried to see if what we told students in an experimental methods class about the intelligence of their rats would affect how long it took students to shape a rat to press a lever. We gave each student a rat and taught them how to shape their rat to press a lever to obtain food. Students in the class were randomly assigned to two groups: One group was told that their rat was bred for superior performance in a maze, and the other group was told that their rat was bred for poor performance in a maze. (In actuality, all rats were the same breed.) Then

all students attempted to shape their rat to press a lever for food. We recorded how many minutes it took them to shape their rat.

One way to analyze the data in this experiment is to use a *t*-test, which is based on the classical statistical model.

a. To use the classical statistical model, we have to make two general assumptions. What are they? How realistic are these two assumptions for this situation? Why?
b. What are the *statistical* hypotheses being tested here? (State them in symbols and in words.)
c. Do we need the concept of a potential population here? Why or why not?

9. In a study of the effects of suggestion on the taste of drinking water, 12 students were recruited from general psychology and randomly assigned to one of two groups. One group drank tap water and rated its taste on a scale from 1 (*very bad*) to 10 (*very good*). The other group also drank tap water but rated it believing it to be a well-advertised, expensive spring water.

a. What are the experimental hypotheses for this experiment?
b. What are the statistical hypotheses being tested here? (State them in symbols and in words.)
c. What is the purpose of random sampling, and how is random sampling related to this situation? Would you consider the participants to be a random sample from a population? Why or why not?
d. Why were the participants in this experiment randomly assigned to conditions? What is the purpose of random assignment, and how is that purpose relevant here?
e. Is there a need to invoke the concept of potential population here? Why or why not?

Student Study Site

Visit the Student Study Site at **https://study.sagepub.com/friemanstats** for a variety of useful tools including data sets, additional exercises, and web resources.