1 A Brief Orientation to Research Methods and Statistics for the Social and Behavioral Sciences

John E. Edlund

Austin Lee Nichols

> *Science is a struggle for truth against methodological, psychological, and sociological obstacles.*
>
> (Fanelli & Ioannidis, 2013)

When teaching about research methods in our classes, the first thing we discuss is the nature of science itself and what makes the scientific method the best way to learn about the world. In 1959 Karl Popper wrote the seminal treatise on the nature of the scientific method. Central to Popper's treatise is the falsifiability of a theory. Popper argues that we start with observations of the world; from there he suggests that scientists start building theories about the world. We then collect additional observations that can test the veracity of the theory. If the theory holds up, we have greater trust in the theory. If the theory fails, we should abandon it in light of a better theory that can account for the additional observations. The falsifiability of a theory is a critical component of science (and what distinguishes science from pseudoscience).

Ultimately, the scientific method (as initially suggested by Popper) is the best way we can learn about the world. As such, his approach underpins all of the methods used in the social and behavioral sciences. Before we get into each of these topics in detail, we thought it best to overview some concepts that are important on their own, but especially necessary to understand before reading the chapters that follow. To do so, we first discuss what good science is. Next, we review some of the basic background material that we expect readers to be familiar with throughout the book. From there, we turn to the nature of data. Finally, we orient the reader to the basic statistics that are common in the social and behavioral sciences.

1.1 What Is Good Science?

Good science is not only characterized by being falsifiable – good science should be universal, communal, disinterested, and skeptical (Merton, 1973). Merton suggested that good science would share all of these traits and would advance the scientific endeavor. The first trait science should have is universality. This norm suggests that science should exist

independently of the scientists doing the research. As such, anyone investigating a particular problem should be able to obtain the same results. Of course, science has not always lived up to this norm (e.g., Eagly & Carli, 1981).

Merton's next trait of good science is communality. Communality refers to the idea that scientific evidence does not belong to the individual scientists – rather it belongs to everyone. As such, secrecy and failing to share your material and/or data violates this norm. This suggests that when scientists investigate a particular finding, they should share those results regardless of what they are. Many governments have embraced communality as part of their grant-awarding protocol and require recipients to share their data in a publicly accessible forum (European Research Council, 2018; National Institutes of Health, 2003).

The third trait of good science is disinterestedness. This suggests that the scientific enterprise should act for the benefit of the science rather than for the personal enrichment of the scientists. That is, science should work for the betterment of others rather than to protect one's own interest. One logical consequence of this (which is challenging to implement in practice) is that reviewers of work critical of their own research should set aside their views and potential impacts on themselves. Additionally, this suggests that research should not be incentivized for obtaining certain results. However, in practice, this has proven to be a challenge (e.g., Oreskes & Conway, 2010; Redding, 2001; Stevens et al., 2017).

The final trait that Merton suggests is a component of good science is organized skepticism. This suggests that any scientific claim should be subjected to scrutiny prior to being accepted. This norm was taken further by Carl Sagan, who suggested that "Extraordinary claims require extraordinary evidence" (Sagan, 1980). As such, any scientific report should face critical review in light of the accumulated scientific wisdom. Sadly, this norm has also failed in many instances (e.g., power posing: Cesario, Jonas, & Carney, 2017).

Using Popper's (1959) falsifiability criterion and Merton's (1973) norms, science can teach us about our world. Although incorrect findings do occur and get published, the self-correcting nature of science allows us to learn more and to benefit from that increased knowledge. In moving beyond the philosophy of science, we need to start defining key terms, practices, and approaches that are used in the social sciences.

1.2 Research Methods

This next section will lay out a basic primer on terminology. Certainly, many of these terms will be familiar to you, but our goal is to bring everyone up to speed for the subsequent chapters.

1.2.1 Quantitative and Qualitative Research

One of the first key distinctions to make relates to the basic approach to scientific research. Roughly speaking, research can be broken down into two classes: quantitative and qualitative. Qualitative research is typically more in depth – these studies will explore the totality of the situation; the who, what, where, when, why, and how of a particular question. One common example of qualitative research is a case study. In a case study,

the particular target is fully explored and contextualized. Other examples of qualitative research include literature reviews, ethnographic studies, and analyses of historical documents. Qualitative methods may also make use of interviews (ranging from unstructured participant reports, to fully structured and guided interviews) and detailed coding of events, and prescribed action plans. Innumerable fields use qualitative methods and most commonly employ this method when first starting an investigation of a certain research question. Qualitative research is then often used to generate theories and hypotheses about the world.

Quantitative research, the focus of this book, typically will not look at a given participant or phenomenon in as much depth as qualitative research does. Quantitative research is typically driven by theory (in many cases, by suggestions from qualitative research). Importantly, quantitative research will make mathematical predictions about the relationship between variables. Quantitative research may also explore the who, what, where, when, why, and how (and will generally make predictions whereas qualitative research generally does not).

In many cases, researchers will make use of both qualitative and quantitative methods to explore a question. For example, Zimbardo used both methods in the seminal Stanford prison study (Haney, Banks, & Zimbardo, 1973). In this volume, we will focus on quantitative methods; certainly, many concepts overlap and readers will end up better prepared for learning about qualitative methods after reading this volume. However, our primary focus is on quantitative methods and concepts.

1.2.2 Theories and Hypotheses

Simply put, a scientific theory is an idea about how the world works. We have already discussed that a scientific theory must be falsifiable. This is an immutable requirement of theory. Theories about the world can be driven by qualitative or quantitative observations. Importantly, a theory will allow for the generation of testable hypotheses. A hypothesis is a specific prediction of how two variables will relate (or fail to relate).

These fairly simple definitions hide a great deal of complexity. A "good" theory is much more than simply falsifiable. Good theories and hypotheses share a number of traits when compared to poor theories and hypotheses. Simply put, a good theory will allow for unique predictions about how the world is, and demonstrate that it can predict relationships better than other theories can. Mackonis (2013) specified four traits that characterize good theory:

Simplicity. A good theory will be parsimonious – convoluted theories are generally lacking and would benefit from refinement or the development of a better theory. This is sometimes termed Occam's Razor.

Breadth. Good theories will generally be larger in scope than rival theories (in terms of situations in which the theory can make predictions). Certainly, there are few theories which are truly expansive (perhaps the generalized theory of evolution being the largest in scope). However, better theories will be more expansive than more limited theories (all things being equal).

Depth. Good theories will be less variant across levels of analysis. For instance, a theory that specifies testable mechanisms

for how the theory works will be preferred to theories that fail to specify mechanisms (and instead have the action occurring in a "black box").

Coherence. Good theories will also conform to generally accepted laws of the universe (although these laws are subject to change themselves based on the development of new theories about the universe).

To illustrate these concepts (and the concepts that follow), we would like to first offer a theory and some hypotheses. Our theory is that readers of this volume are great people who will know more about research methods in the social sciences as a result of reading this volume. This theory is lacking in scope (only applying to a limited number of people), but it is a parsimonious theory that offers concrete predictions that no other theory does. It is also certainly a theory that could be falsified (if you aren't great!). Given the specificity of this theory, two hypotheses naturally flow from the theory: a) compared to some other group, readers of this volume are "great" (we will discuss defining this term in the next section) and b) readers will know more about research methods after reading this volume than they did before.

1.2.3 Operationalization and Constructs

Good theories will suggest ways in which two phenomena are related to one another. These phenomena are often referred to as constructs. We might specify that two concepts are related to one another in some way (at this moment, in the abstract). This is termed conceptualization. These concepts are typically non-observable in their essence, although we can measure aspects of a particular construct with a good operationalization. Operationalization is the specific definition of a construct for a particular study. Operationalization can vary between researchers – reasonable people can design very different ways of operationalizing the same variable. You might ask, isn't this a bad thing? No! In fact, this will allow us to have more confidence in the results we obtain (we will return to this concept later in this chapter as well as in Wagner & Skowronski, Chapter 2).

To continue our earlier example, the constructs are defined in the hypothesis. We are exploring the "greatness" of the reader along with the reader's knowledge about research methods. These are our theoretical constructs. We must first have a personal understanding, or conceptualization, of what each of these constructs is. What exactly is "greatness"? As authors, we went back and forth in terms of discussing this concept as we wrote in this chapter (i.e., is the concept even definable?). To us, greatness is an idea that we think people will agree on that ties in aspects of likeability, talent, and skill. Next, we need to decide how to operationalize these variables. Obviously, given our conceptualization, there are many ways we could operationalize greatness. One way we could operationalize this construct would be to measure friendliness (likeability). Certainly, friendliness has been measured (Reisman, 1983) and would capture an aspect of greatness. Perhaps we might want to operationalize the variable as intelligence (given our interest in how a textbook might change levels of greatness). There are innumerable intelligence measures available (e.g., Flanagan, Genshaft, & Harrison, 1997). We could also design our own greatness measure for this study (see Stassen & Carmack, Chapter 12 for a detailed

exposition on the construction of a scale). Any of these operationalizations might be acceptable for testing whether using this text contributes to greatness. We are also interested in testing whether the reader knows more about research methods as a result of reading this text. In this case, the conceptualization of content knowledge is likely understood and shared by many. Similarly, many would suggest an operationalization of a quiz or test that would assess factual knowledge of research methods. Of course, there are hundreds (or thousands) of possible questions that could be assessed; some may be excellent and some may not. For the purposes of this chapter, let us say that we decided to operationalize greatness with a new measure that we created using items representative of likeability, talent, and skill and that we chose last year's research methods final exam to operationalize content knowledge.

1.2.4 Variables

Once we have operationalized our constructs, we need to start using two key terms associated with our research (which will become key when we turn to statistics). The **independent variable (IV)** is the variable that we believe causes another variable when we do inferential statistics. IVs can be experimentally manipulated (two or more conditions are manipulated and assigned to the participants), quasi-experimental (two or more non-randomly assigned conditions; biological sex is a commonly reported quasi-experimental variable), or simply be measured variables that are hypothesized to affect another variable. In our example, we could experimentally compare students who learn research methods due to this volume compared to students who learn research methods via another text (and as long as students were randomly assigned to textbooks, we could say that any difference was caused by the different text). As such, each book (this volume or the other volume) would represent different levels of the IV. The **dependent variable(s) (DV)** is the variable that the IV is proposed to act upon. This is the variable that many researchers would expect to show a difference as a result of the IV. In our example, the DV would be content knowledge (i.e., the test/quiz score).

1.2.5 Reliability and Validity

Reliability and validity are key concepts in the social sciences. In fact, they are so key that the next chapter in this book is dedicated to exploring these issues more fully than can be covered here. For simplicity's sake, we will say that **reliability** is the consistency of measurement. Are you getting the same (or very similar) value across multiple modes and administrations? **Validity** is the meaning of your measurement. Validity asks whether you are getting values that correspond to your underlying construct and theoretical variables (see Wagner & Skowronski, Chapter 2 for a detailed exposition on these issues).

1.3 Data

Before we can cover the kinds of statistics encountered in the social sciences, we need to define the kinds of data generated in these studies. The first kind of data is **nominal** data (sometimes called categorical data). Nominal data consists of discrete groups that possess no rankings. Biological sex is one commonly used example of nominal data. There is no intrinsic ordering to sex; you could just as easily say females come before males or vice

versa. Additionally, when you assign quantitative values to sex, any number could reasonably be used (females could = 1, males could = −1, or any other mutually exclusive categorizations). Another example of nominal data is the reading of this volume (readers of the volume = 1, and non-readers of this volume = 0).

The next kind of data that we can work with is called **ordinal** data. Ordinal data possesses all of the traits of nominal, but ordinal data adds a specific ordering to the categories. Military rank is a classic example of ordinal data. For instance, the lowest category of Naval Officer rank is Ensign, followed by Lieutenant Junior Grade, Lieutenant, etc. The categories are discrete and ordered, but, beyond that, simply knowing the rank will not provide any additional information.

The third kind of data encountered in the social and behavioral sciences is **interval** data. Interval data possesses all of the traits of nominal data, but adds in equal intervals between the data points. A common example of interval data is temperature in Fahrenheit. For instance, the difference between 22°F and 23°F is the same as the difference between 85°F and 86°F. Importantly, zero simply represents a point on the continuum with no specific characteristics (e.g., in an interval scale you can have negative values, such as −5°F). This is the first kind of data that can be referred to as continuous.

The final kind of data that we can encounter is **ratio** data. Ratio data possesses all of the traits of interval data but incorporates an absolute zero value. A common example of ratio data is reaction time to a stimulus. In this setting, time cannot be negative and a person's reaction time to a particular stimulus might be measured at 252 milliseconds.

1.4 Statistics

The final section of this volume explores new statistical trends in the social and behavioral sciences. Chapter 18 explores Bayesian statistics, Chapter 19 explores item response theory, Chapter 20 explores social network analysis, and Chapter 21 introduces meta-analysis. However, before we introduce you to these advanced forms of statistics, we want to ensure that you have a grounding in other common statistical procedures that are often encountered in social and behavioral science research. In this volume, we will focus on linear relationships (and statistics) as these are the most common statistics used in these fields.

Before discussing these statistics, we need to briefly discuss the common assumptions of the data. Many of the statistics discussed are based on the **normal distribution**. This is the idea that, for a particular variable with a large enough sample drawn from a population, the values of that variable will concentrate around the mean and disperse from the mean in a proportional manner. This distribution is sometimes referred to as a bell curve (see Figure 1.1). Importantly, only interval and ratio data can potentially obtain a bell curve. As such, unless otherwise noted, the statistics we discuss in this chapter are only appropriate for these kinds of data as DVs (although it is worth explicitly noting that nominal and ordinal data can be used as IVs in these analyses).

There are a number of commonly reported descriptive statistics in the social and behavioral sciences. The following list is certainly not exhaustive, but it represents the most commonly used descriptive statistics. Additionally, for brevity's sake, we will not discuss the intricacies of the calculation of these statistics

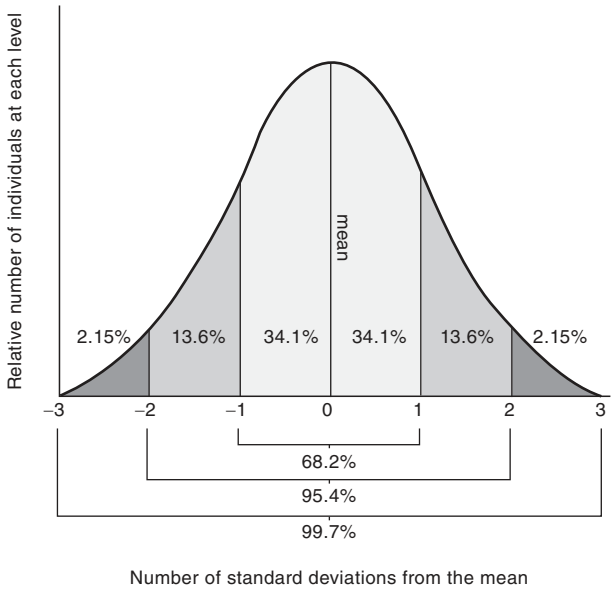

FIGURE 1.1 The normal curve

(for further details, we encourage you to explore our suggested readings); we will instead conceptually discuss these statistics.

1.4.1 Measures of Central Tendency

One of the kinds of information that we will want to know about our data is where the center of that data is. There are three common measures of central tendency in the social and behavioral sciences. The **mean** is the mathematical average of all of the values. However, it is important to know that the mean can potentially be impacted by a few outlier variables (we will discuss this when we discuss standard deviation), and as such, should be carefully inspected. The **median** is the central number in a distribution (the middle number). The median is often less impacted by outliers. The **mode** is the most common value in a distribution. The mode is not be impacted by outlier values. Of particular note, in the typical normal distribution (when the sample is large enough), the mean, median, and mode will be the same value (approximately).

1.4.2 Measures of Distribution

The second kind of information that we typically seek in our data are measures of the distribution of values (around the central value). The simplest value commonly reported is the range. The **range** is simply the distance from the smallest value to the largest value. One of the appeals of the range is that it is easy to calculate and understand. However, this value does not tell us how the values between the extremes are distributed and can also be affected by outliers. The most common measure of distribution reported in the social sciences is standard deviation. **Standard deviation** is the measure of how much divergence there is, on average, in values across the distribution. Mathematically, standard deviation is the square root of its variance (how far a value is from the mean). As such, it is a measure of how far away from the mean the "typical" value is in the distribution. Finally, skew and kurtosis describe how far from normality (i.e., the bell curve pictured in Figure 1.1) the observed data is. Specifically, **skew** is a measure of how asymmetrical the

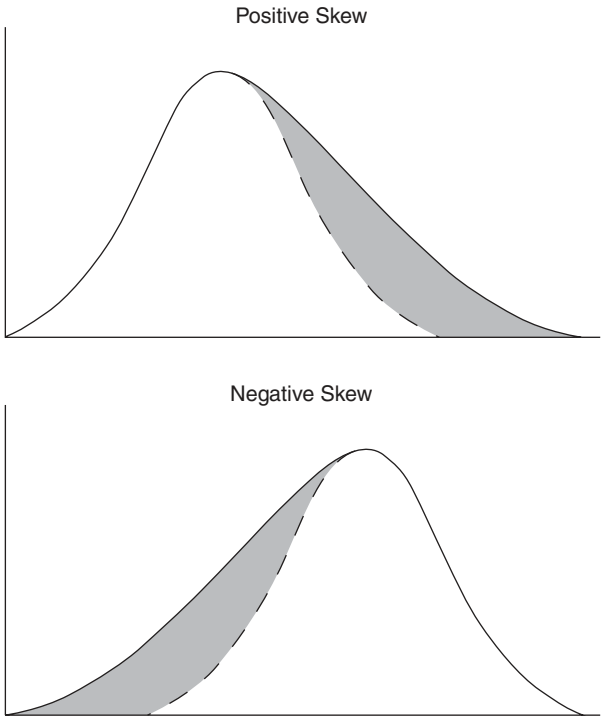

FIGURE 1.2 Positive and negative skew

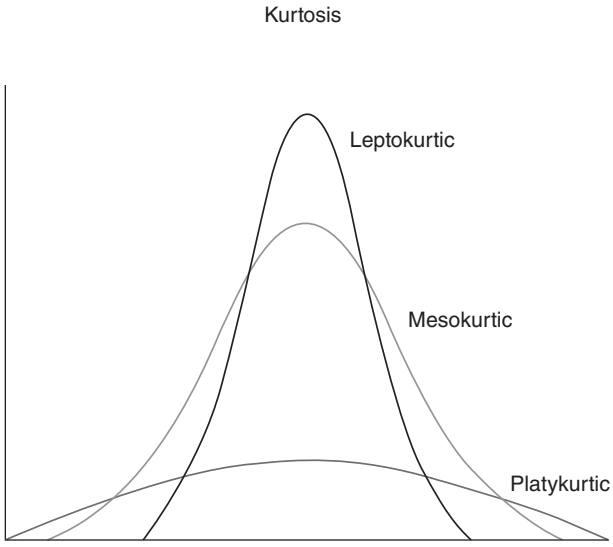

FIGURE 1.3 Types of kurtosis

distribution is around the mean (see Figure 1.2). Skew values range from positive numbers that describe a distribution skewed to the right (i.e., the right tail is longer than the left tail) to negative values that describe a distribution skewed to the left (i.e., the left tail is longer than the right tail). **Kurtosis** is a measure of the central peak of the data (see Figure 1.3). Higher

kurtosis values indicate a higher central peak in the data (termed "Leptokurtic") whereas lower values indicate a relatively flat peak (termed "Platykurtic"). A normal kurtoic curve is termed "Mesokurtic."

1.5 Statistical Analyses

As we move on to discussing correlation and inferential statistics, we need to briefly discuss statistical significance and the assumptions of analyses. In null hypothesis testing, we *actually* test the likelihood that the observed distribution would occur if the null hypothesis (that there is no difference between groups) was true. In the social and behavioral sciences, we typically set our level of confidence for a type 1 error (i.e., that there is an effect when, in reality, there isn't $-\alpha$) as $p < .05$. Of course, there is nothing magical about that likelihood as it is simply an arbitrarily chosen (but accepted) value (Rosnow & Rosenthal, 1989). The other kind of error that we are concerned about is a type 2 error (saying there is no difference between the groups when, in reality, there is $-\beta$). Researchers are concerned with obtaining sufficient power to detect effects to avoid the possibility of type 2 errors. See Figure 1.4 for a humorous illustration of the two types of errors.

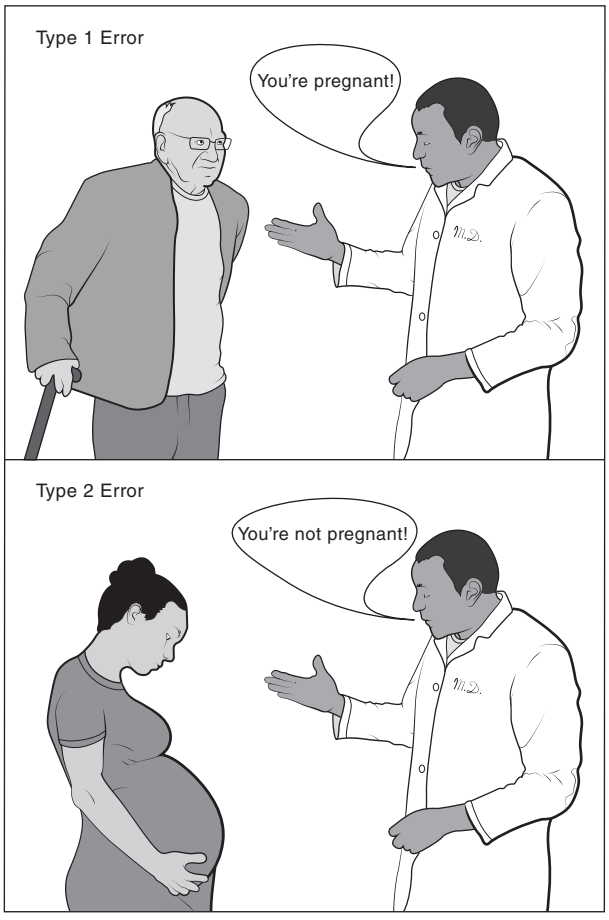

FIGURE 1.4 Type 1 and type 2 errors

However, simply reporting a *p*-value along with a statistical test variable is no longer considered sufficient. Many journals now require the reporting of confidence intervals and effect sizes. **Effect sizes** are measures of how "meaningful" the difference is. There are numerous measures of effect size that can be reported ranging from Cohen's *d*, to Cohen's *g*, to a partial eta squared, to others. **Confidence intervals (CIs)** are a projection of where the true value is likely to occur (typically with a 95% chance). This is often phrased, in practice, as 95% confidence, but we cannot interpret this to mean that there is 95% probability that the interval contains the true value. Instead, what we can simply state is that there is a 95% chance that the interval contains the "true" value. That is, if repeated samples were taken and the 95% confidence interval was computed for each sample, 95% of the intervals would contain the population value. CIs can be calculated to reflect on differences between the means, point values, or other effects sizes.

All statistical tests rest on certain assumptions. Generally speaking, the analyses we will discuss in this chapter will assume interval or ratio data that is normally distributed (low levels of skewness and kurtosis). As such, an inspection of skewness and kurtosis (and a transformation of data, if necessary) should be a part of all analyses (see LaMothe & Bobek, 2018 for a recent discussion of these issues). The good news is that the majority of analyses used by social and behavioral scientists are robust against violations of normality (i.e., even if you run your analyses on data that somewhat violates these assumptions, your chances of a type 1 error will not increase; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010).

1.5.1 Moderation and Mediation

Two terms that describe how your IVs may influence your DV(s) are moderation and mediation. **Moderation** occurs when the level of an IV influences the observed DV. Moderators simply specify under what kinds of circumstances would a particular effect be likely (or unlikely) to occur. An example of moderation might be how great people read this volume. You have two groups (roughly dividing people in the world into great and non-great) and the great people are statistically more likely to read this book when assigned in a class setting. That's moderation!

Mediation, on the other hand, explains how a particular variable leads to an observed change in the DV. That is, a mediating variable is the mechanism through which a particular change can take place. For instance, let's say we measure research methods knowledge in students prior to taking a methods course and after (via a test). Reading this volume may be the mechanism by which students learn and do better (on a similar but different test). If this volume fully explains that relationship, reading this volume would be said to fully mediate the relationship between the changes from pre-test to post-test. However, we like to think that your instructor is valuable as well and is contributing to your learning. In that case, the reading of the volume would be said to partially mediate the relationship between scores on the pre-test and post-test (and that your instructor's contributions would also partially mediate, or explain, that relationship).