CHAPTER 1

Introduction

The goal of this book is to improve and systematize the use of Benford's law in the social sciences in assessing the validity of self-reported data. We do this by first introducing a new measure of conformity to the Benford distribution that is created through permutation statistical methods and that employs the concept of statistical agreement. More specifically, we introduce a chance-corrected measure of effect size that measures the level of agreement that a researcher's data have with the Benford distribution. In a switch from a typical Benford application, we move from using Benford's law to test whether the data conform to the Benford distribution to using the Benford distribution to draw conclusions about the validity of the data. But before describing this new technique let's discuss reliability and validity, as these are the primary concerns that, we argue, lie behind using the Benford distribution strategy that we suggest.

Virtually all social science research methods textbooks dedicate a chapter to the study of reliability and validity. This, of course, makes sense, given how central these concepts are to undertaking rigorous social science research. Indeed, reliability and validity are necessary conditions for social science research. Simply put, without reliability and validity the results produced by social science research cannot be trusted. Reliability is the ability of measurement instruments to produce consistent results over time and in different contexts. Validity refers to how accurately the measurement instrument captures the concept that the researcher purports to measure with it.

A commonsense example is often used to illustrate the concepts of reliability and validity. Consider a bathroom scale that measures weight in pounds (lbs.). That scale can be considered reliable if it produces the same reading for weight when a person gets on and off the scale repeatedly under the same conditions (i.e., if the person doesn't change clothes, eat or drink, or use the restroom). In this hypothetical example, suppose that a person weighs 200 lbs. according to the scale. That person gets off the

https://doi.org/10.1017/9781009127950.002 Published online by Cambridge University Press

scale, changing nothing else, and lets it recalibrate before getting back on again. The scale again reads 200 lbs. If, after several iterations of getting on and off, the scale continues to produce a reading of 200 lbs., that scale can be considered a reliable instrument for measuring a person's weight. Social scientists would like research instruments to be reliable because reliable instruments produce measurements that are the same under the same set of conditions, and there is a degree of trust that can be placed in those measurements. When research does produce different results, this should be due to changes in the phenomena that are measured and not a result of the measurement itself. For instance, if the person in the example above started a diet or an exercise program while consuming the same calories, we might expect the reading on the scale to decrease over time. This is because the conditions surrounding the observation (i.e., the person's weight) have changed. This idea that measurements should give different readings when conditions change leads us directly into our discussion of validity.

Consider, again, the bathroom scale. For the scale to be a valid measure of weight, it must accurately report, in an acceptable measure of weight, the weight of the person on the scale. One common method of assessing the validity of a measure is to assess it based on some other measure of the same concept that is trusted and validated. For example, let's consider the scale for measuring weight in a hypothetical medical center to be a valid measure of weight in pounds. In this case, let's say that the person in our example just returned from a medical appointment. When they checked in, the nurse weighed them and recorded a weight of 200 lbs. Equipped with the knowledge that the scale in the medical office showed a weight of 200 lbs., the person returns home and proceeds to step on their bathroom scale. Their bathroom scale also reads 200 lbs., exactly like the scale at the medical center. Given that the bathroom scale is reporting the same weight as the scale at the medical center (a scale that is known to be a valid measure of weight), we can consider the bathroom scale to also be a valid measure of weight. The bathroom scale is indeed accurately measuring what it purports to measure – weight in pounds.

It is possible to have a reliable instrument that is not valid. In the example above, if the bathroom scale reads 225 lbs. every time the person gets on the scale, it is reliable. However, given that we know (from the medical center) the weight of the person to be 200 lbs., the bathroom scale cannot be producing a valid measure of weight. Rather it reliably measures weight in pounds, but at a weight 25 lbs. greater than the valid weight of 200 lbs. The converse is not true. In other words, a measure cannot be valid if it is not reliable. Common sense should be enough to understand

that a measure cannot have even face validity – that is, capture the concept it purports to capture – while not producing the same results repeatedly uses – that is, while lacking reliability.

Validity is, arguably, more difficult to attain than reliability in social science research. For the sake of simplicity and clarity, we began with an example using a bathroom scale. Of course, social science concepts are generally much more complicated and multifaceted than weight. As we are sociologists and criminologists, let's take as an example a core concept from a foundational thinker in sociology: Karl Marx's notion of alienation. According to Marx (1844) [1959], in the process of wage labor members of the working class become alienated from (1) the act of production, (2) their product, and (3) their species being (i.e., human nature); and this can be considered a defining characteristic of capitalism. In the middle of the nineteenth century, Marx observed that the move to capitalism and wage labor brought about a fundamental transformation in the relationship between workers and their jobs. First, capitalism reduced jobs to a series of routine, simple, repetitive tasks that required little brain power and in which labor was treated as an exchange value. Consequently workers no longer increased the element of overall satisfaction in their lives through their work; work simply became a means of exchanging labor for wages and earning money, and this represented alienation from the act of production. Second, the act of production no longer allowed for the worker's input, knowledge, and experience; in other words the products were designed by the capitalists, not by the workers – and this, too, generated alienation from production. Third, this process also alienated the worker from the product: the worker no longer controlled the product, having sold the capitalist the labor for producing a commodity that the capitalist owns and controls. Fourth, according to Marx, one of the core identities of human beings more generally is as workers or producers. As he observed, humans begin "to distinguish themselves from animals the moment as soon as they begin to produce their means of subsistence" (Marx & Engels, 1845 [1968], p. 6). Work also gives humans a purpose. But that purpose can be lost when work is reduced to capitalist wage labor, in which it has lost its intellectual and creative aspects and has been brought down to repetitious, boring tasks. In sum, for the purposes of this illustration, we can consider alienation to be, in a simplistic definition, the state of feeling estranged from one's work along the dimensions highlighted here.

Now, imagine a contemporary study in which a researcher wants to examine the alienation of workers in 2023 by using a survey distributed to

a random sample of wage laborers. Clearly alienation is a multifaceted concept, and one with which many wage laborers would not be familiar. So asking a question such as "Are you alienated at work?" or "Do you suffer from alienation at work?" is meaningless. Rather the researcher is going to have to design a suite of questions that get at different aspects of alienation, and then perhaps to combine them into a scale. Now, in order to assess the validity of the alienation scale before the survey is administered, the researcher should consult the opinions of other researchers who work on alienation. Additionally, the researcher should do a comprehensive review of the alienation literature.

Doing these two things should provide researchers with some understanding of the validity of their measure by identifying all the different aspects of alienation. Then, once the survey data have been collected, the alienation measure data can be compared to other data on alienation that use a more accepted scale. At this point, the researchers should have a good idea as to whether the measures and the data are capturing what they purport to capture – that is, whether they possess validity. Do the measures and the data actually measure alienation, or do they measure, say, how much a survey respondent dislikes the boss? It should be obvious, at this point, that operationalizing complicated social science concepts is difficult and requires that the validity of the concepts be assessed. While the methods for assessing the validity of scales and other social science composite indicators are imperfect, at least they exist. The same cannot be said for all types of social science data.

A substantial literature discusses measures and processes of establishing reliability and validity. Perhaps the most common example of a quantitative indicator is Cronbach's coefficient α, a measure of the internal consistency of the component parts of indices that is often reported as a measure of scale reliability (Cronbach, 1951; Cronbach & Meehl, 1955). However, there is no recognized similar quantitative measure of validity that is regularly used in social science research. Rather researchers assess the validity of indicators by qualitative means (e.g., through face validity) or by correlating the indicators with existing ones that are known to be valid (i.e., through criterion validity). Those correlated indicators are then used to predict the indicators suggested by theory.

In social sciences, self-reported data are quite common. These data present some advantages because they are often widely reported and easy to obtain. Self-reported data are simply data that rely on an individual's descriptions of things that social scientists are interested in studying, such as characteristics, beliefs, attitudes, and behaviors. Individuals may

self-report their own characteristics, beliefs, attitudes, and behaviors or report those things for groups and organizations, in an official or unofficial capacity. For example, police departments often provide the public with data on the annual number of crimes that occurred in their jurisdictions. When used for research, self-reported data are often referred to as "secondary data." That is, police officials are simply self-reporting crimes that took place in their jurisdiction and that are known to their department; and researchers take these reported data and use them to test research hypotheses that are usually not considered when the police collected and reported the data in question.

There are any number of well-established reasons to challenge or criticize secondary data. For instance, in the example above, the definition of crime is determined by the state and not by the researcher. The researcher's hypotheses on crime are limited to what the state views as crime (for an extended discussion, see Lynch, Stretesky, & Long, 2015). This definition may lack face validity, as it ignores some important behaviors that should be included among crimes (e.g., crimes committed by corporations). However, secondary data are usually viewed as beneficial by the research community, because they allow researchers to examine a variety of research questions using data that would otherwise be difficult, if not impossible, to collect. Thus self-reported data, especially in the form of secondary data, are often seen as beneficial to social sciences.

Self-reported data may also be problematic; for example, they may be intentionally or unintentionally misreported. Because researchers do not gather the self-reported data themselves, they rely on those who self-report to provide data that are valid and reliable. Researchers are often unaware of any intentional or unintentional misreporting that may have taken place. For instance, it has long been recognized that police departments may misreport crimes to the public in order to make the department look more effective in fighting crime or in order to make the community in which the department operates look safer than it is (or both) (Seidman & Couzens, 1974). This practice has been so widespread in the United States, for example, that a special phrase, "going down on crime," has been created to describe it (Scoville, 2013).

Unfortunately, in social science research many self-reported data are not assessed for validity (i.e., for whether they are valid or invalid). For example, while the US government works hard to ensure that it identifies instances of police departments misreporting crime, this is but one instance of misreporting in the vast array of social science data. At this juncture it is important to point out that the failure to assess misreporting

in self-reported data is problematic because it is not possible to know whether the results of studies are based on valid data. If data are not valid, the study itself may not draw valid conclusions and those kinds of outcomes can undermine the scientific process. This is why the goal of the present book is to provide one quantitative measure of validity, along with a workflow that can be used to comprehensively assess self-reported data for validity when it comes to potentially intentional or unintentional misreporting. One way to achieve this goal, we suggest, is to employ Benford's law.

Benford's Law

Our measure and method of validity analysis for misreporting is based on Benford's law – a probability distribution for the leading digit in a set of numbers (Benford, 1938; Newcomb, 1881; see also Frunza, 2015; Nigrini, 2012). The nine leading digits of the numbers in a dataset (i.e., digits 1–9) are not equally likely; rather, as Benford's law states, leading digits should be distributed in the manner reported in Table 1.1 (assuming that the data meet several conditions that will be discussed in future chapters, e.g., $n \geq 500$, spans several orders of magnitude, etc.). Further, Benford's law also provides expected probabilities for the successive digits of numbers in a dataset (see Table 1.1), although the range of expected probability values is much smaller for the second digit than for the first digit, and smaller still for the third and fourth digits. To illustrate, the range of probability values for first digits is $0.3010 - 0.0458 = 0.2863$, the range of probability values

Table 1.1. *Benford's law: probability values for first, second, third, and fourth significant digits for* $d = 0, \ldots, 9$.

d	First digit	Second digit	Third digit	Fourth digit
0	-	0.1197	0.1018	0.1002
1	0.3010	0.1139	0.1014	0.1001
2	0.1761	0.1088	0.1010	0.1001
3	0.1249	0.1043	0.1007	0.1001
4	0.0969	0.1003	0.1002	0.1000
5	0.0792	0.0967	0.0998	0.1000
6	0.0669	0.0934	0.0994	0.0999
7	0.0580	0.0904	0.0990	0.0999
8	0.0512	0.0876	0.0986	0.0998
9	0.0458	0.0850	0.0983	0.0998

https://doi.org/10.1017/9781009127950.002 Published online by Cambridge University Press

for second digits is 0.1197−0.0850 = 0.0347, the range of probability values for third digits is 0.1018−0.0983 = 0.0035, and the range of probability values for fourth digits is only 0.1002−0.0998 = 0.0004.

Over time, researchers discovered that they could use Benford's law as a method of checking the accuracy of data and, more recently, of detecting accounting and financial fraud (Nigrini, 2012). Specifically, comparing the observed distribution of a dataset to the expected Benford distribution and (somehow) quantifying the difference provides a measure of whether the data can be considered accurate. In other words, a high level of conformity with the Benford distribution suggests that the data lack measurement error, be it intentional or unintentional. While the original tests of conformity to the Benford distribution focused almost exclusively on the first digits of the data, some later applications provide a test for the second digit or for the first two digits at the same time. We focus only on first digit tests in this book. The reason is that successive digits approach uniformity, as demonstrated in Table 1.1 and as noted by Newcomb: "In the case of the third figure [digit,] the probability will be nearly the same for each digit, and for the fourth and following ones the difference will be inappreciable" (Newcomb, 1881, p. 40). It should be noted, however, that there are some advocates for examining second digits instead of first digits (Diekmann, 2007).

Benford's law has occasionally been used to assess data accuracy in the social sciences (see, e.g., Badal-Valero, Alvarez-Jareño, & Pavía, 2018; Beiglou et al., 2017; Breunig & Goerres, 2011; Brown, 2005; Cole, Maddison, & Zhang, 2020; Coracioni, 2020; de Marchi & Hamilton, 2006; de Vries & Murk, 2013; Deckert, Myagkov, & Ordeshook, 2011; Hickman & Rice, 2010; Judge & Schechter, 2009; Koch & Okamura, 2020; Mir, 2012, 2014; Tam Cho & Gaines, 2007), using a number of different approaches and measures of conformity. However, in our view, the social sciences would benefit from a more systematic use of the Benford distribution as a means of assessing data validity.

Benford Agreement Analysis and Benford Validity

Most studies that employ the Benford distribution to assess data do so to find fraud or misreporting. Recall that, in our case, we flip from using Benford's law to test whether the data conform to the Benford distribution to using the Benford distribution to draw conclusions concerning the validity of the data. We advocate the systematic use of Benford's law in social sciences to assess the validity of self-reported data by introducing our

https://doi.org/10.1017/9781009127950.002 Published online by Cambridge University Press

new measure of conformity, which employs permutation statistical methods and statistical agreement. Specifically, we introduce a chance-corrected measure of effect size that measures the level of agreement a researcher's data have with the Benford distribution.

Once we establish the new measure of agreement with the Benford distribution (symbolized by the Fraktur letter $\mathfrak{R}$), we turn our attention to our recommended workflow for Benford agreement analysis. In other words, we describe the steps that researchers should take to comprehensively assess their data for validity using the Benford agreement analysis procedure and establish whether these data have what we call "Benford validity." Moreover, *we argue that establishing that self-reported data possess Benford validity should be a regular step in the social science research process.*

Plan for the Book

The remainder of the book is organized as follows. Chapter 2 begins with a brief discussion of validity in the social sciences. We note that, while some types of quantitative social science data (e.g., scales and latent variables) have methods for assessing whether the measures and the data are valid, self-reported data do not have the same options for assessing validity. This is particularly true for secondary self-reported data; therefore we provide illustrations of self-reported data that are emblematic of the types of data social scientists use. These discussions are based on data on the US prison population, COVID-19 new cases, toxic releases in the United States, and global fish catches. We then provide a list of social science studies that employ Benford's law to assess data accuracy. We conclude the chapter with a brief introduction to the concept of Benford validity.

Chapter 3 presents Benford's law in detail. We then turn to a discussion of the existing methods for measuring conformity with the Benford distribution. These methods oftentimes employ probability values generated from chi-squared or likelihood-ratio goodness-of-fit tests, which, we argue, are of limited utility for making meaningful decisions on whether data conform to the Benford distribution. We next introduce permutation statistical methods, which we employ in the construction of our measure of conformity to the Benford distribution. The benefits of permutation methods include that they do not require random sampling, make no assumption of normality, are completely data-dependent, and yield exact probability values (Berry et al., 2016, p. 15). Next, we argue that a measure of effect size that allows the researcher to assess practical or substantive significance is the most useful approach to assessing conformity with the

Benford distribution. After reviewing the different classes of effect size, we settle on a measure of effect size that is chance-corrected and based on the statistical concept of agreement – a measure symbolized by $\mathfrak{R}$. We conclude the chapter by changing the language of interpretation of Benford results from "conformity with the Benford distribution" to "agreement with the Benford distribution."

Chapter 4 begins with a discussion of the characteristics that data must possess to be suitable for what we call "Benford agreement analysis" (i.e., $n \geq 500$ four significant digits and spans several orders of magnitude). We then provide a program written in the statistical package R that calculates $\mathfrak{R}$ and provides additional information, for example a histogram of the results. Next we move to a description of the workflow of Benford agreement analysis using state-level US prison population data for the years 1972–2002. This example

1. highlights the use of Benford agreement analysis for the full dataset,
2. highlights the use of an analysis of subgroups, and
3. ends with an illustration of how to identify where the misreporting is clustered.

The chapter concludes with a discussion on how to decide whether the data that are examined have Benford validity and can be used in social science research.

Chapters 5 and 6 provide detailed examples of using our measure of chance-corrected agreement with the Benford distribution and the associated workflow on publicly available, self-reported secondary data. In Chapter 5 we analyze reported and unreported fish-landings data (measured in 2010 US dollars [USD] or US$) from the Sea Around Us project. In Chapter 6 we analyze new COVID-19 case data from US Centers for Disease Control and Prevention (CDC), for US states, over time. We then turn our attention to new COVID-19 case data at the country level, over time.

In Chapter 7 we return to the Sea Around Us data and describe a process by which data with unacceptable Benford validity can be further examined. Here we explore the extent to which deviation from Benford validity meaningfully affects the results of the desired statistical model. This diagnostic process uses values of $\mathfrak{R}$ as both independent and dependent variables in regression models. In this chapter we highlight that, in some cases, data with problematic Benford validity may still be usable for certain analyses.

The book ends with a short conclusion in Chapter 8 that begins by summarizing the work presented in Chapters 2 through 7. We then discuss the conditions under which a Benford agreement analysis should be used. Finally, we conclude with a few thoughts on the validity of self-reported social science data.