Chapter 1
Introduction

Statistics plays a crucial role in scientific research, especially social science research based on observational (nonexperimental, survey) data. In addition, statistics plays an increasingly important role in our daily lives. We are constantly inundated with statistics in the news in the form of political polls and general information about one or another aspect of life, and these statistics routinely find their way into politics—often as sound bites—as well as into policy.

At the same time that the use of statistics permeates our lives, statistics used as a evidence for a position or argument are much maligned, and statistics as a course of study is dreaded. Few college students look forward to taking a statistics course, few college graduates seem to recall their introductory statistics course fondly, and in the general population, people never seem to have anything good to say about the subject. Indeed, a number of colloquialisms have emerged about statistics, including: "there are lies, damned lies, and statistics," "you can prove anything with statistics," and "87 % of statistics are made up on the spot" (this percentage, of course, can be replaced with any number you like!).

In fact, "statistics" don't prove anything by themselves; they must be interpreted. For example, a 10 % unemployment rate in the US could reflect any number of things. It could mean that the country has a lot of lazy people. It could mean that the economy is in a recession, with businesses cutting payroll in order to maintain a stable level of profit. It could mean that there is no recession, but simply that the working-age population is growing faster than demand for products and services. In short, such a statistic in the absence of other data could mean just about anything. And using such a statistic as "proof" of some point is no more valid than using poor logical reasoning with words is. Put another way, one could just as easily—and falsely—say that "you can prove anything with words." The difference is simply that statistics, because their production involves mathematics—another area of study often dreaded—have an air of mysteriousness and sophistication about them that seems to lend greater legitimacy to arguments that involve them than those that do not.

This is not to say that statistics are useless or that learning about statistics is a waste of time. There is a right way (or right ways) to use statistics, and the

S.M. Lynch, *Using Statistics in Social Research: A Concise Approach*,
DOI 10.1007/978-1-4614-8573-5_1,

main goal of this book is to show how to use statistics properly in social science research. A secondary goal is to make you aware of how statistics can be misused, misinterpreted, or faultily constructed in a variety of settings, especially in the media and in social and political debate.

As you will learn in the book, there are three main purposes of statistics: summarizing data, making inference from samples to populations, and predicting (or "forecasting") future—or at least unobserved—events based on extant data. Most people are aware that statistics are used to summarize data. For example, when a news agency reports that 15 % of the US population does not have health insurance, this is a single number summarizing the status of a population of 300 million people.

Fewer people are aware that this type of percentage is not usually based on a complete enumeration and canvassing of the population, but rather a small sample drawn from it. Drawing conclusions about the entire population from a small sample is the purview of "inferential statistics," and this aspect of statistics is less well understood. Indeed, when I read viewer comments on news websites reporting the results of polls measuring political attitudes, I often see comments questioning the validity of the results based on several recurring arguments. Some question how a poll of 600 persons in the US could possibly accurately reflect the attitudes of 300 million people. Some suggest that poll results are inherently flawed because "I've never been polled, and my views (and those of everyone I know) disagree with the results." As you will learn in this book, extremely precise and accurate results—"inferences" to the larger population (in fact, an infinitely large population)—can be obtained from very small samples. Furthermore, given the small samples that are usually taken in polls, it is not surprising that most persons will never be involved in a "scientific" poll. Indeed, a very rough calculation suggests that at least a half million polls would have to be taken before any given individual in the U.S. population could expect to be polled. While polls have become increasingly common over the last few decades, far fewer than 500,000 polls have been conducted.

In terms of prediction, we often hear statements like: by 2037, the Social Security system will be bankrupt, or, by 2020, white non-Hispanics will no longer be the majority of the US population. These sorts of claims also fall within the purview of statistics. Such claims are based on projecting the current state of the world forward in time based on knowledge of observed patterns of fertility, mortality, migration, tax rates, etc. We often refer to such predictions as "forecasts," because their predictive validity—that is, their ability to accurately predict future events—rests on the assumption that future conditions and patterns will remain as they have been, or at least as they are expected to be. If there is a change from expectation, then all bets are off. For example, imagine the consequence for solvency if Social Security tax rates were to increase (or decrease).

Another type of prediction in statistics regards the ability to make statements about unobserved situations or individuals based on observed situations or individuals. For example, suppose I observe in a sample that men, on average, earn more than women, on average. I might then predict that, in a future random sample, the men will earn more than the women on average. Many people have trouble with

this type of statistical argument. Some might say that they know plenty of women who earn more than men that they know—and this may be true. For example, in the 2008 General Social Survey (data from which will be used throughout this book, and so it will be abbreviated as GSS; see Smith et al. 2011), the average income for men in the sample was just over $48,000, while the average for women was just over $32,000. However, there were certainly many women in the sample who earned more than some men (and vice versa). In particular, 20 % of women in the sample earned more than the average for men, and 36 % of men earned less than the average for women.

In other words, both statements are true; one does not contradict the other. To argue that the claim about average income is false because of one's own experience is to ignore the fact that one's own experience may not be representative of all of reality, while the average is a summary of all of reality. Put another way: individual experience does not invalidate a general pattern, so long as that general pattern is properly established. We call such appeals to one's personal experience "anecdotes," and it does not matter how many anecdotes one can provide, they do not establish or reject patterns established by actual data.

A more substantial argument may be that my claim about average income perpetuates an invalid "stereotype" about women. This argument is somewhat more difficult to counter, and this is one dilemma that anyone using statistics inevitably will come across. By definition, a stereotype is a generalization about a "type" of individual that has a "kernel" of truth. In the example just discussed, the fact is that there are more women who make less than the average income for men than there are above the average for men. So, my claim is not a stereotype: it is more than a "kernel" of truth. Think about it this way. If I selected a man and woman at random from the population, would you be willing to place a significant bet that the woman earns more than the man? Given that the average tells us something about where the distribution of incomes is centered for men and women, it would probably not be a smart bet.

Importantly, one thing that differentiates a statistical argument from a stereotype is that a statistical argument generally places limits on its claim. For example, suppose it's true that unemployed persons are more likely to engage in robbery than employed persons, such that 2 % of the unemployed commit robberies, while 1 % of the employed commit roberies. While it's true that the unemployed commit robberies at twice the rate of the employed (the "kernel of truth" here), it would be fallacious (and ridiculous) to fear the unemployed but not fear the employed. In neither group is robbery a predominant behavior, and so to claim that unemployed persons are robbers would be to unjustly "stereotype" the unemployed as criminal.

This type of stereotypical reasoning may be obvious, but there are certainly less obvious instances of it. For example, in reading viewers' comments on news websites, I have often seen the argument that all unemployed persons are simply lazy, bolstered by the argment that the commenter knows one or more lazy unemployed persons. The claim that ALL unemployed folks are therefore lazy is fallacious, because it is a "hasty generalization." The generalization is "hasty," because a single person's view of his/her own experience (observing, usually,

only a couple of persons at most) is not representative of the population. The opposite argument—that, because "I know one unemployed persons who is not lazy, therefore no unemployed persons are lazy"—is equally fallacious. Therefore, debates involving this type of faulty reasoning can go nowhere. Instead, properly used statistics may help arbitrate the issue.

In short, we cannot generalize—in any direction—from our personal experience. This is one of the most difficult concepts to learn in statistics, and yet it is one of the most important. To be blunt, be wary of any claim that says that "all X are Y." Sound statistical reasoning involves making generalizable claims, but equally importantly, it places qualifiers and limits on those claims.

1.1 Goals of This Book

In this book, we will focus on the role of statistical reasoning and methods in social science research, paying particular attention to the roles of summarization and inference. As such, the important starting point is discussing what scientific research is about—as well as what it is not about. The next chapter discusses the process of scientific research, from the stage of asking a general question through the specification of hypotheses that are to be tested using statistics. Chapter 3 discusses how to find and/or collect data to address a hypothesis, with a particular focus on issues of sampling design and construction of quantitative survey instruments. Although this book is not geared toward showing you how to develop your own instrument and collect data—indeed, the focus of the book is on using extant ("secondary") data—it is extremely important to understand how one's data was collected, and how survey questions were asked, in order to produce meaningful statistical analyses.

Chapter 4 shows how to take a data set and produce meaningful summaries of it using descriptive statistical methods. The chapter begins by discussing how to summarize univariate data, that is, how to summarize single variables that may be of interest, like years of schooling or earnings, before showing how to summarize bivariate data using "cross-tabulations."

Chapter 5 discusses probability theory in considerable depth in order to lay the foundation for understanding statistical inference. Chapter 6 begins by introducing one of the most difficult ideas in introductory statistics that follows from probability theory—the Central Limit Theorem. The chapter then shows how statistical inference can be made using the theorem to conduct "hypothesis tests" and to construct "confidence intervals" for univariate data, that is, data consisting of a single variable. The chapter continues this discussion in extending inference about one variable to inference about group differences in a continuous variable.

Chapters 7 through 10 extend the basic concepts of inferential statistical reasoning for single, continuous variables, to different types of quantitative data and more sophisticated types of research questions.

Finally, Chapter 11 shows how to summarize the results of statistical analyses in reader-friendly tables and figures, as well as how to summarize the results and write the discussion and conclusion sections of a scientific research paper.