1

Introduction

PURPOSE

This text is written for a first course on statistics and quantitative methods for Ph.D. students in social science and allied fields. Anyone undertaking to write such a book must sooner or later confront the question of whether the world really needs another introductory statistics textbook. In my surveys of the market for my own classes on this subject in two social science Ph.D. programs, I clearly decided that it does.

Students in social science Ph.D. programs outside of economics have widely divergent levels of previous exposure to statistical methods, as well as comfort with mathematical expression of concepts. The typical Ph.D. program does not have the luxury of multiple "tracks" to suit different backgrounds, so one course must accommodate all of them. That course must be accessible to students with divergent levels of preparation but must also prepare them technically for future quantitative methods coursework ahead of them.

More important, I have found that students of whatever background will plunge relatively enthusiastically into methods training once they understand why it is essential for the purely substantive elements of their research. Simply put, many students, particularly those without much prior exposure to statistics, do not understand what it is or how it can help them as social scientists. Without this understanding they lack the buy-in necessary to make the technical rigors of the course seem worthwhile.

I never found a textbook that was pitched at the right technical level for introductory students and also focused on what social scientists in particular can do with statistics. Textbooks on mathematical statistics, although useful for the technically ambitious students (and entirely suitable for use in conjunction with this text), leave most students behind at a formal level. They also tend to focus, at least in their core, on univariate inference. This framework allows for

clarity of statistical theory – but it obscures what social scientists can do with statistics and leaves many students wondering why they should bother when they have social science research to do. Statistics textbooks at an intermediate level may be technically suitable for most Ph.D. students in social science but are not much better at explicating the integrative link between social science theory and statistics. These problems with the textbook market are widely enough recognized that instructors often assemble reading lists culled from various textbooks and research papers published in home-discipline journals, but this approach is inherently piecemeal and obscures the overall integration of the material from the students.

This text is my attempt to redress these limitations of current offerings. Mathematically, I assume students have had exposure to and a solid grasp of algebra for scalars. Prior exposure to calculus is helpful (especially for some proofs) but not strictly necessary; when I use calculus notation and concepts in the main text, I explain them on the fly. I essentially do not use matrix algebra in this text except for occasional use of vector notation as a shorthand. Other than this, the text is mostly self-contained, conceptually and mathematically. That is not to say it will be technically easy for most students; it is not for most of mine. But prior formal training in higher-level mathematics is not the limiting factor. What students really need to be able to do is integrate formal definitions and notations into their thinking right away when they are introduced.

This in itself is not a substitute for deep training of future methodologists or theoreticians of statistical modeling but should suffice for students who need solid intuition about the tools they are using more than anything else. The text does not skimp on formal proofs and derivations of core results, for example, of the law of iterated expectations, the expected value of ordinary least squares regression coefficients, canonical formulas for standard errors, and the like. My conviction is that any intelligent user of statistical methods should understand the connection between assumptions they make about their data and inference problem and the properties of the resulting statistical output; formal proofs are simply the arguments that we use to establish those connections. To keep the length and formal demands manageable, I also often present an important theorem or approach verbally (e.g., the Cramér-Rao theorem, derivation of the F distribution, properties of parametric bootstrapped standard errors) to alert students to the key issue and indicate important topics for work in more advanced quantitative methods courses.

This approach implies that mathematical demands of the presentation are not uniform across chapters and topics. For example, the formal presentation of the sampling distribution for least squares regression coefficients (Chapter 7) is much more intense than my presentation of the concept of maximum likelihood estimation (Chapter 9), which is rather more intuitive. Because the purpose of the book is to develop solid understanding of core ideas in statistical inference and modeling rather than to present a formally complete and precise treatment of every topic for its own sake, this seems to be a useful compromise. In

any case, like many instructors, I have always found that many students in more advanced courses are not clear enough on the *ideas* – irrespective of the mathematical technicalities – that are to be covered in a course such as this. So my treatment is an attempt to stress those ideas. I hope thereby to reduce the incidence of students proceeding to later courses or self-study without a clear idea of the meaning of a regression, sampling distribution, p value, estimator, likelihood function, statistical model, and the like.

As far as my treatment of what statistics does for social science, I focus early and throughout on theoretical arguments about relationships between variables and statistics as a set of tools to assess those relationships in data. Early and sustained focus on this is one of the principal distinguishing features of this book. In our published research that uses quantitative methods, this is almost always what we social scientists are doing with statistics and statistical models. Understanding how theories are couched in terms of relationships and that statistical models are good ways to measure relationships, makes for a good motivation to learn statistical modeling. I have sought to write a book that introduces students to this way of thinking as soon as possible. I have found this to be unusual in a textbook, but it is probably the most important thread running through this one. I understand "theory" in social science as an argument that two or more concepts are related and the reason why; it is therefore not an accident that the substantive core of social science is intimately linked to statistics.

This focus on relationships in both social science theory and statistical modeling is reflected at several points that are unusual in first treatments but are formally no more difficult than what is usually covered. First, canonical families of probability distributions are introduced not as abstract and mysterious entities that students must suffer through today for some unknown future good but rather as convenient models for various types of social processes. I introduce them immediately with common functional forms linking the parameters of these distributions to covariates of theoretical interest to a social scientist as such and with derivation of marginal effects of these covariates. The student therefore encounters formal probability models as a way to structure and express their theorizing about some dependent variable. This is, by design, entirely separate from statistical issues involved in estimating the parameters of these models. When a student wonders what a logit model or a negative binomial model is all about, the first thing he or she needs to understand is how it expresses dependencies of what is explained on what does the explaining, not on how the magnitude of that dependency is estimated or on its numerical stability.

Second, the core statistical concepts of sampling, estimation, and hypothesis testing are all covered with an early and sustained focus on inference for parameters in regression models, not just univariate inference. The idea of a sample regression coefficient as random in repeated sampling is no more difficult than the idea of a sample mean as a random variable (which is not to say

either concept is simple or intuitive for newcomers to the field; they are not), so there is no reason not to introduce these concepts at the same time. Yet doing so saves endless heartache as students learn regression in earnest, both through their own research and in further coursework. To be sure, in preference for a "clean" analytical approach, I typically focus on plain-vanilla regression and simple extensions in these discussions but also attempt to explicate how the concepts translate to parameter estimates from more intricate estimators.

Upon finishing this book, students will know how to use and critique the application of several workhorse techniques for establishing and interpreting relationships in social science research and will have developed a framework for evaluating and understanding the contribution made by more advanced techniques (the details of which they may grasp only loosely). Concretely, through the theory and applications discussed subsequently, students should finish the book with the ability to understand and critique journal articles and books in their field that make use of linear models, generalized linear models, maximum likelihood estimation, and classical hypothesis testing, among other techniques. (In my experience, this is not the case after completing most introductory treatments.) In addition, upon finishing, students will have built a solid foundation in the theory of statistics for future quantitative methods course work – and more importantly, for teaching themselves new methods as necessary for their own research. The last result is one of the most important for any scholar doing quantitative empirical work; the field of quantitative methods in social science is huge and expanding much more rapidly than any person has the ability to grasp. Therefore, the need to master a technique for a research project that one has not learned in formal coursework is a normal state of affairs.

SCOPE

In basic and abstract terms, a *positive theory* in social science asserts that two or more concepts or events are related to each other and specifies the reason why.[1] To maintain abstraction and for lack of better names, we might as well call these concepts x and y. Often a theory is useful and important because it postulates a process or channel or "mechanism" through which x causes y. So when x changes, it will cause y to change in a fashion that is explicated by the theory. It is possible that the theory says that x and y are only related conditional on some third concept z. That does not pose any conceptual problems for what follows.

For example, in political science, a theory of the "democratic peace" asserts that two nation-states are less likely to go to war with each other if both are

[1] By "positive" theories I mean ones that attempt to specify how social or political processes actually work, as distinct from"normative" theories. Positive theory in my usage does not imply derivation from any particular type of reasoning, such as formal or mathematical modeling. Nor does my usage imply that one developing or exploring positive theories must adhere to positivism as a philosophy of science.

democracies than if both are not. So here y indicates the concept that two states are "at war" with each other, and x indicates that both are democracies. The theory asserts that these concepts are related to each other in a specific way: the chance that a dyad of nation-states go to war is greater when they are not both democracies than when they are.

To show that a theory deserves a place in the conversation and debate of some field of research, a researcher advancing it should show that it organizes and renders intelligible actual events in the world. This requires confronting the theory with data to see if the former explains the latter. To wit, as a matter of fact, does x cause changes in y as the theory asserts? If it is, then at the very least, x and y should be associated with each other or should "move together" in some specific way spelled out by the theory. A major reason social scientists pursue formal statistical analysis is to answer that question of how things play out in fact and whether they comport with the theory in question. This empirical analysis of theory constitutes the stock-in-trade of a great many social scientists and comprises a large portion of the research published in disciplinary journals.

The main ingredient of that analysis, of course, is data. Needless to say, in social science there are often no uncontested, cut-and-dried measures of the concepts to which a theory pertains. With respect to the democratic peace, a political scientist interested in empirical analysis would first have to consider an operational definition of "war" as well as "democratic nation-state." Careful and sensible measurement are obviously crucially important for meaningful quantitative analysis, but these issues are mostly beyond the scope of the coverage in this book, save for some cursory discussion in Chapter 2 and sprinkled in a few other places. For the most part, I take as given a set of observations of variables that are tied to theoretically important concepts in a substantively meaningful way but do not offer much advice about how to obtain them.

With data in hand, an analyst can explore whether x and y are related as the theory asserts. That is what this book is about. Statistics provides a set of tools for assessing relationships in data. That is the main reason social scientists use it. Statistics provides a battery of formulas and techniques, simple in their basics, for assessing relationships among variables. Social scientists also often (and increasingly so) want to know if a relationship is causal and, if so, what is causing what (i.e., x causes y, vice versa, both, or neither). Statistics provides a useful conceptual framework for assessing what causal inference means and what it requires, although it provides no simple formulas that ensure valid causal inferences are possible in a given context.

A significant complication in assessing relationships is introduced because social scientists often deal with data that has some uncertainty behind it. You only get to see one dataset, but the data in it might be generated by some type of random or *stochastic* process. It might be that the observable data represents only a part of a larger, unknown universe of possible observations; it might be that the social processes and behind the data involve behavior that can only be described as (at least partly) random. Assessing relationships in the presence of

data with randomness behind it, and expressing the degree of uncertainty about those assessments, is difficult. That assessment is what a lot of the theory of statistics, especially statistical inference and modeling, is about. The difficulty explains why we have to spend so much time on it.

This text explains (*i*) tools for summarizing data and assessing relationships in data, (*ii*) sources of uncertainty in the data social scientists can see, (*iii*) tools for formalizing and analyzing uncertainty and randomness, and (*iv*) tools for assessing relationships in the face of randomness. It also briefly discusses (*v*) concepts for assessing the extent to which an observed relationship is a *causal* relationship.

This book's coverage of these topics is almost entirely based on "classical" or"frequentist" statistics. It almost entirely excludes nonparametric approaches in the frequentist tradition. It also pays little attention to Bayesian statistics, which is achieving a status approaching standard fare for some types of estimation and in some fields. For a variety of reasons, some good and some less good, most quantitative research in social science still builds on the classical, parametric approach. There will be more than enough balls in the air with the restriction to that topic, even for students who have some solid statistics training, so that introducing whole new schools of thought as well seems ill advised.

OUTLINE

More specifically, the plan of the book is as follows. Chapter 2 introduces descriptive statistics. One of the most important sources of insight about political and social processes, and in practice a necessary precursor to formal statistical analysis, is simply "playing around with the data." This exercise gives researchers a feel for the behavior of variables, which can serve as a guide to specific modeling choices, and can suggest relationships for further analysis. Students and scholars that dive into modeling exercises without understanding their data are not uncommonly exposed to embarrassingly simple questions about specific observations that they cannot answer. This chapter develops several important benchmark techniques to structure such investigations. It introduces the concept of a distribution and variable, and presents tools for creating information from mere data. These tools are designed to distill a mass of numbers into a few useful summaries that are readily interpretable and comparable across distributions. In the case of distributions of several variables, these tools include summary measures of *association* among variables – that is, relationships –which are the bread and butter of most efforts to test theories in empirical social research.

Chapter 3 discusses the need for inference about unobservable quantities based on observable ones – not just when using statistics but when performing almost any kind of theoretically guided research based on observed data. It makes the distinction between observable data, from which summary measures

and descriptions can be computed, and the crucial concept of a "data-generating process," which is itself unobservable but which gives rise to the observable data. It is this data-generating process that, in virtually all conceivable cases (even when it doesn't seem like it at first), is the ultimate object of theoretical interest to a social scientist. Furthermore, most data-generating processes that might interest anyone involve uncertainty and randomness (or at least what looks like randomness to an analyst or researcher): that is, they give rise to particular sets of observations not with certainty but only with some probability.

Chapter 3 discusses reasons for randomness in social and political data and some important implications of that randomness for the observable data from which descriptive statistics are derived. The most important such implication for the practice of empirical research is that any summary of observed data – quantitative or otherwise – itself has some randomness bound up in it. This chapter is somewhat "metaphysical," but it covers important (although often unstated) philosophical commitments and disagreements behind different approaches to empirical research.

Given the importance of uncertainty and randomness in the data social scientists can observe, a language for discussing and analyzing randomness precisely is necessary. Such a language allows for analysis and evaluation of the information provided in observable data about its underlying data generating process. That language is provided by the theory of probability, which is introduced in Chapter 4. The discussion starts from the primitive idea of an experiment and the set-theoretic ideas of a sample space and event. Probability measures are defined in terms of these foundations, and some important properties of probabilities of events are presented. The discussion then turns to the crucial concept of a random variable and the cumulative distribution function and mass or density functions associated with random variables. These functions are the central tools for modeling data-generating processes in statistical research. This chapter also introduces multiple random variables and multivariate probability distributions.

Probability distributions convey much information about the behavior of random variables that can be compressed into more compact summaries of their behavior. These summaries are addressed in Chapter 5. In particular, this material focuses on expectation and variance for univariate distributions, and conditional expectation and covariance for multivariate distributions, as important characteristics and summaries of data-generating processes. In many ways, these summaries of probability distributions are analogous to the ones that descriptive statistics offers for data distributions, as covered in Chapter 2.

Theories in social science often suggest or imply certain features of a data-generating process, such as the behavior of its summary characteristics as some variable changes. For example, many theories that assert that one variable is associated with another imply something about the behavior of a conditional mean. Chapter 6 develops the link between positive theory in political and

social science and the form and characteristics of a data-generating process. This chapter introduces several important families of probability distributions that are useful as models of specific kinds of social and political processes. It explores the relationship between theoretical assertions that variables are related and the statistical comparison of expectations, variances, or other features of distributions. It also notes that many theoretical assertions of association imply that the conditional mean of one variable is a function of another, thereby introducing regression modeling as a central tool of theory evaluation. This chapter also explores the expression of additive as well as multiplicative effects of one variable on another. The main point of this chapter is that many positive theories in social science are essentially assertions about specific aspects of data-generating processes. Therefore, it establishes a crucial link between statistical modeling and the theory development and assessment that are the ultimate purpose of most empirical research in social science.

The chapters up to this point explore (*i*) the description of information in observable data, (*ii*) the distinction between observable data and the data-generating process, (*iii*) the data-generating process as a probability model and properties of these models, and (*iv*) the link between theory and data-generating process. But any empirical conclusion about theory can (by definition) only be based on observable data. Therefore, the use of observable data as a basis for theory assessment in social science requires an understanding of the connection between this observable data and the data-generating process. That connection was hinted at in Chapter 3. The rest of the book from Chapter 7 to the end focus on formally analyzing it. Collectively these chapters comprise the discussion in this text of inferential statistics, as opposed to the descriptive statistics and probability theory that occupied attention before Chapter 7.

The first step is to explore how the process of sampling or essentially taking "draws" of observable data from a data-generating process affect the connection between the data and the generating process. This is discussed in Chapter 7. The discussion draws heavily on the canonical concept of a "random sample." With such a sample, the summary statistics introduced in Chapter 2 have particularly appealing connections to the underlying data-generating process. This chapter formally defines a statistic and discusses the foundational statistical concepts of a sampling distribution and the central limit theorem. The chapter then explores how the linkages between summaries of observed data and characteristics of the data-generating process are modified by various violations of the random sampling assumption. Because these violations are commonplace in actual data in social science, these modifications cannot be treated as mere footnotes to the main point. Instead, whenever possible and necessary, they must occupy a central place in empirical investigation because they strongly affect the link between observable data and data-generating process, and therefore between observable data and theoretical question.

Chapter 6 argues that positive theories in social science often imply a comparison of conditional means or other features of data-generating processes.

The use of observable data to compare data-generating processes is the subject of hypothesis testing, covered in Chapter 8. Hypothesis tests assess the degree of support in observed data for specific conjectures pertaining to underlying data-generating processes. They are inherently linked to the sampling distributions of statistics used to estimate the features of the data-generating process that are of ultimate theoretical interest. The chapter discusses the conceptual aspects of hypothesis testing and develops methods for several important and commonly used tests (e.g., for difference in means, proportions, and variances; for the sign of regression coefficients; for categorical and tabular data).

The linkage reviewed in Chapter 7 between summaries of observable data and the data-generating process can be used another way: if we don't know some important characteristics of the data-generating process, we can use summary measures from observable data to estimate them. Estimating characteristics of the underlying data-generating process in this way is "point estimation" and is the subject of Chapter 9. This chapter presents several criteria for evaluating point estimators and uses them to evaluate the summary measures familiar from previous discussions. In so doing, it presents a version of the law of large numbers, an important theoretical result, and extends the discussion of sampling distributions for several important estimators. The chapter then discusses important techniques for developing estimators, specifically the methods of maximum likelihood, least squares, and (briefly) Bayesian estimation.

The book then briefly considers causation. Social science needs statistics because our theories typically concern relationships among variables, and statistics has a variety of tools to estimate relationships among observable quantities. But most of these statistical tools used in social science focus on *association* among variables, whereas most of our positive theories assert *causal* relationships among variables. "Correlation does not imply causation" is a mantra in empirical research and deservedly so. Chapter 10 discusses the distinction between association and causation and how and when one can bridge that gap in empirical research. This chapter covers the meaning of causation, in terms of the potential outcomes causal model and what causation looks like at an empirical level. It discusses specific conditions under which one can interpret a demonstration of mere association as an assertion of causation, and how the source of the data (e.g., experimental data vs. observational data) affects how strong a case can be made that these conditions hold. The chapter demonstrates implicit assumptions in important empirical techniques that are necessary for assertions of causation based on observable data. This material is essential and rapidly growing in importance for many social scientists, given what we hope to get out of empirical analysis. Although the treatment here is terse, I believe it is useful to present the themes and simple technicalities in a style of notation and exposition to which students are already accustomed. I hope this will make for a smoother transition to further coursework covering these issues in depth.

Finally, a brief afterword puts what is covered in this book in the larger context of empirical social science research, research that in many or most

cases attempts to find what events or variables in social and political processes are related to each other, how strongly, and why.

NOTES ON COURSE USE

I have taught the sequence of topics around which this book is written numerous times in the political science departments at U.C. Berkeley and Northwestern University. In each department, the course was the first in a sequence on quantitative methods. In a semester-length class at Berkeley, I typically make it through most of the text, with time for at least an overview on Chapter 10. In a quarter-length class, a few more choices must be made. The content to de-emphasize depends somewhat on the course that comes after the one for which this book is used.

This book includes a relatively large amount of material on regression for a course on foundations of probability and inference. I have found this essential to motivate the technical details for the many students for whom an interest in methods follows an interest in social science. That said, if this course is followed by a full course dedicated to regression or statistical modeling, a few (lengthy) sections of this book can be covered briefly or skipped to save time: Sections 6.7, 6.9, 7.5, 8.5, and 9.5 all deal with modeling and inference for regression. Most of Chapter 6 is written around the concept of a statistical model, but most of this is interwoven with canonical families of probability distributions. It is inadvisable to skip that material for introductory students, although it can be downplayed. In addition, Chapter 10 is self-contained, and the core theory of probability and inference is completed before this chapter. So that chapter can also be skipped or left for self-study if the course runs short on time.

It is tempting to skip Chapters 2 and 3 because they provide background material but not the core theory of probability or inference, but I have always found it important to cover these chapters. It is simply not that unusual for introductory Ph.D. students to have a hazy idea or less about the basic toolkit of descriptive statistics or the sorts of questions one can ask (and maybe answer) with these tools, so Chapter 2 is important for making the core theoretical ideas concrete. As for Chapter 3, most social science disciplines continue to have ongoing, alive-and-well research traditions drawing on methods one might call "qualitative," or at any rate far outside the ambit of statistics. This chapter is important because it provides a framework for thinking about what kinds of assumptions one must make so that a statistical approach is coherent and that is an essential part of helping Ph.D. students understand what they are doing with research methods and why. Discussions of the sort engaged in this chapter might even have the ancillary benefit of reducing the amount of incoherent quantitative research done.

I have taught from this text to students with widely divergent levels of math preparation, including some without a "math camp" or "prefresher" before