CHAPTER 1

Introduction

1.1 Toward a new political methodology

Since the emergence of the discipline, political scientists have focused their methodological efforts on creating new data sources. Scholars have polled millions of citizens in personal interviews, recorded exhaustive lists of cooperative and conflictual international events, collected enormous quantities of aggregate data, and preserved the formal and informal decisions of numerous elected officials. These new data sources have enabled political scientists to consider myriad new empirical questions and unexplored theories.

In data collection, political science outdistances most other social sciences. However, in developing and adapting statistical methods to accommodate theoretical requirements and the special data that is collected, political methodology lags far behind the methodological subfields of virtually every other discipline. *This imbalance in political science research between state-of-the-art data collection and weak methods means that future statistical innovations are likely to have disproportionate impact on scholarly research.*[1]

This is not to say that political scientists have shied away from quantitative analyses. To the contrary, quantitative methods have gained widespread acceptance over the past decade. However, the overwhelming majority of statistical methods currently in use have been imported into political science directly from almost every other scientific discipline. Since these tools are generally imported intact and without adaptation, many are ill-suited to the particular needs of political science data and theory. In addition, most political scientists learn only a few of these methods and methodological perspectives, often with unfortunate consequences. For example, many observed political variables are fundamentally discrete; voter preferences, international event counts, and political party identifications are cases in point. The problem is that the application of techniques such as linear regression analysis forces researchers to define naturally discrete concepts as continuous or to try

[1] Since data in political science do not permit a standard set of statistical assumptions, some argue that we should use only very simple quantitative methods. But this is precisely incorrect. Indeed, sophisticated methods are required only when data are problematic; extremely reliable data with sensational relationships require little if any statistical analysis.

to elicit a continuous survey response about a discrete concept. The result is measurement error, bias, and the loss of a significant amount of information.

Given this practice of borrowing from diverse disciplines, political scientists have amassed considerable expertise in using methods created by others. Unfortunately, with a few notable exceptions, expertise in evaluating and adapting existing methods, and in creating new methodologies, is severely lacking. Paradoxically, political scientists make excellent applied econometricans, sociometricans, and psychometricians but often poor political methodologists.

The absence of a coherent field of political methodology has at least three important consequences: the lack of methodological standards, communication, and cumulation.

First, political science has few, if any, established methodological standards. Some researchers abide by statisticians' standards; some have adopted those of economists; others adhere to the statistical norms of sociological research. In short, the discipline of political science has nearly as many methodological standards as there are statistical methods. Unfortunately, this mosaic of largely incompatible standards and traditions has not coalesced into a consistent whole. If progress is to be made in substantive research, we must work toward some uniform, discipline-wide methodological standards. Pluralism and diversity in methods are welcome features; however, diversity in standards to evaluate the merits of these different methods is equivalent to the absence of standards. The frequency of statistical fads is a consequence of this undesirable state of affairs. The scholarly journals are littered with the surge, decline, and, in too few cases, eventual reasoned incorporation of numerous new techniques. Factor analysis, log-linear models for contingency tables, Box-Jenkins time series analysis, and vector autoregression are but a few examples of what were once statistical fads.

Second, institutional mechanisms to disseminate methodological developments, when they occur, are rudimentary at best. This compounds the discipline's methodological shortcomings in several ways. The absence of communication leads to frequent replication of the same statistical mistakes and repeated attempts to address methodological problems that have already been solved (King, 1986a). Moreover, scholars are slow to recognize that methodological developments in one field of political science may be applicable to similar problems in other fields.

Finally, the absence of cumulative progress in political methodology reduces the potential for methodology to influence substantive research in political science in positive ways. "Whatever else may be uncertain in methodology, there is ample evidence that the shaky foundation it provides has weakened intellectual structures across the discipline" (Achen, 1983: 71).

What political science needs, then, is a more highly developed and unified

field of political methodology. From this should come a set of standards for evaluating quantitative research, more flexible and powerful statistical techniques able to extract significant amounts of new information from our wealth of quantitative data, new methodological perspectives that provide novel ways of looking at the political world, and the potential for cumulative progress in political science research.

This book is intended to take political methodology some way down this path. I have sought to unify the diverse statistical methods existing in political science under a general theory of scientific inference. Indeed, since "a theory is the creation of unity in what is diverse by the discovery of unexpected likenesses" (Bronowski, 1958: 8), unifying political methodology and articulating a general theory of inference are compatible, if not equivalent, goals. With the likelihood theory of statistical inference, one can go a very long way toward extracting the essential features of these methods while leaving behind the peculiar conventions and practices more appropriate in other disciplines. Likelihood is neither the only theory of statistical inference nor the best theory for every political science problem, but it generates, by far, the most widely applicable methods for data analysis.

I omit several aspects of ongoing research in political methodology from this book. First, with a few exceptions, I do not talk about how to derive statistical models from economic theory. Although several excellent examples exist in the discipline (see Enelow and Hinich, 1984), most areas of political science do not have fully developed formal theories. Second, political science contributions to other disciplines are not discussed except tangentially (e.g., Brady, 1985; Hildebrand, Laing, and Rosenthal, 1977; Rivers and Vuong, 1988). Finally, neither political statistics and data collection (see Tufte, 1977) nor descriptive statistics (Alker, 1975) and methods of graphical presentation (Tufte, 1983) are discussed in any detail. Each of these topics is of considerable importance, but I have tried to keep my focus on the theory of inference and creating statistical models for research in political science.

The remainder of this chapter defines terms, symbols, and the elementary ideas to be used throughout the book. Part I (Chapters 1–4) contains a comprehensive analysis of *uncertainty* and *inference,* the two key concepts underlying empirical research in political science. The basic concepts are introduced in Chapter 2. Then, the probability model of uncertainty is presented in Chapter 3 and the likelihood model of scientific inference in Chapter 4. The likelihood model is based on the probability model, but it is far more useful as a theory of inference, providing a more fundamental basis for statistical analysis. The chapters comprising Part II draw on the theory of inference developed in Part I. In each chapter, I present a different set of statistical models applicable to a different type of data and theory. A variety of discrete regression models is presented in Chapter 5. Chapter 6 reviews models based

on cross-classified tabular data, and Chapter 7 demonstrates how to model time series processes. An introduction to multiple equation models appears in Chapter 8. Models for nonrandom selection are in Chapter 9, and general classes of multiple equation models are derived in Chapter 10.

Thus, this book explores a very broad set of statistical models. However, the subject of this book is a new perspective on the unity inherent in what are and should be the methods of political methodology, not a comprehensive guide to every possible method or even an exhaustive analysis of any one. My subject is the trunk of the tree and some of its branches, not the detail in the leaves. In addition, in many places, I have gone beyond the list of techniques commonly used in political science research and have described models that could profitably be used but have not been. Whereas some of these statistical methods are new just to political science, others were derived here for the first time.

1.2 A language of inference

In order to begin from a common frame of reference, I briefly develop a language to describe key features of the enterprise. The terms, phrases, and mathematical symbols in this language are not sacrosanct, but they do differ from more standard treatments for specific theoretical reasons (for a discussion and criticism of statistical language, see Bross, 1971). This is meant only as a brief general overview; each of the concepts mentioned here is more fully explained in later chapters.

Social system: The largest component of social scientific inference is the social system – the object under study. A social system may be a neighborhood, a government, an election, a war, a stock market, or anything else a researcher wants to know about. The social system usually contains certain *features* that are either known or estimable, but important elements remain unobserved and indeed unobservable.

Explanatory variables: Explanatory variables are measures of the features of a social system, symbolized as X. Unless otherwise stated, I assume that X is *known* a priori. X contains measures of each of these features for all n observations $(x_1, \ldots, x_i, \ldots, x_n)$. If only one feature is known and relevant, x_i is a scalar variable. If several known features are relevant, x_i is a vector of variables.

Output (dependent variable): Every interesting social system generates some type of output. Outputs are consequences of the social system that can be

observed and measured. For example, an election produces a victorious candidate, a coup d'etat in a developing nation could produce changes in policies toward the industrialized world, and a stock market may generate competitive prices for existing firms.

When parts of the social system are under control of the researcher, the process of producing output is called an *experiment.* In experiments, the researcher can manipulate the explanatory variables (the known features of the social system) and make it produce output as needed. Although true experiments are only infrequently possible in the social sciences, I will use the term generally where an experiment could be conducted, if only in theory. For example, a presidential election cannot be controlled by political scientists, but one can imagine the same election being "run" multiple times under essentially identical conditions. Let Y stand for a vector of n outputs of the social system. Y_i, one of the n outputs, is generated in theory by setting the known features of the social system (x_i) to particular values and running the experiment. Even under identical conditions (i.e., values of the explanatory variables), the social system outputs different results for each experiment; this is because output is produced in a probabilistic instead of deterministic manner.

Random variables: Operationally, a random variable is the assignment of numbers to events (Democratic President = 1, Republican President = 0; income = number of dollars, etc.) with a probability assigned to each number. One should not confuse "random" with "haphazard." Random processes adhere to very specific probabilistic rules defined below. Y_i is a *random* variable since the actual data are randomly produced from the social system's outputs according to each event's probability. The key idea behind a random variable is that Y_i varies across an infinite number of hypothetical experiments. If the actual experiment were "run" again under essentially the same conditions, the observed values of the dependent variable (the "realizations" of the random variables) would differ but the nature of the experiment and the social system would remain constant.

The data: The data, y, are n observed realizations of the *random variables* Y. They are a set of numbers, each y_i being a random draw from a corresponding random dependent variable, Y_i. The process by which portions of output are chosen and measured is called *sampling,* and the data themselves are sometimes called "the sample." Survey sampling is one general category of examples, but many others exist. The specific sampling method used is central to both modeling and estimation. (Only occasionally will I use "the data" to refer to both y and X.)

Models: A model is a mathematical simplification of, and approximation to, a more complex concept or social system. Models are never literally "true," although one often proceeds as if they were. Indeed, the idea that a model *is* a social system or that parameters exist in nature is certainly odd. Within this definition, many types of models exist. The next paragraph introduces a general *statistical model,* from which many even more specific statistical models are derived. In the next chapter, I also discuss *probability* as a general model of uncertainty and *likelihood* as a general model for statistical inference. Indeed, modeling is the core concept of this volume: Models provide precise views to the world; they contribute new ways of approaching old problems; and they provide a means of inferring from observed data to unobserved phenomena of interest (see Bender, 1978).

Statistical models: A statistical model is a formal representation of the *process* by which a social system produces output. The essential goal is to learn about the underlying process that generates output and hence the observed data. Since no interesting social systems generate outcomes deterministically, statistical models are assumed to have both *systematic* and *stochastic* components.

The most common way of writing a complete linear-Normal regression model, for example, is to express the random dependent variable (Y_i) as the sum of a systematic ($x_i\beta$) and a stochastic (ϵ_i) component:

$$\begin{aligned} Y_i &= x_i\beta + \epsilon_i, \\ \epsilon_i &\sim f_n(e_i|0,\sigma^2), \end{aligned} \tag{1.1}$$

where f_n refers to the Normal distribution with mean zero and constant variance σ^2. This representation is very convenient for linear-Normal models, but quite burdensome in the more general case. Fortunately, these equations can be equivalently expressed as follows:

$$\begin{aligned} Y_i &\sim f_n(y_i|\mu_i,\sigma^2). \\ \mu_i &= x_i\beta. \end{aligned} \tag{1.2}$$

In this format, the randomness in the dependent variable Y_i is modeled directly, and the systematic component models variation in one of its parameters (the mean, μ_i) over the observations. Since Y_i itself is random, this formulation requires no artificial analytical construct for the random error (ϵ_i).

In Equation (1.1), we generally assume that x_i and ϵ_i are independent. We have become used to thinking in these terms, but conceptualizing ϵ_i and what it correlates with is not as easy to explain in terms close to one's data and theory. Fortunately, an equivalent assumption in Equation (1.2) is considerably easier to theorize about: All one needs to assume is that Y_i depends on x_i only through its mean, $E(Y_i) \equiv \mu_i$. In other words, x_i and Y_i are *parametrically*

related but, conditional on this parametric relationship, are not *stochastically* related.[2] This is much easier to understand in more complicated cases because this parametric relationship is expressed explicitly in the second line of Equation (1.2).

The general form of the complete statistical model, for the linear-Normal or almost any other model, may then be written in two equations as follows:

$$Y \sim f(y|\theta,\alpha), \tag{1.3}$$

$$\theta = g(X,\beta). \tag{1.4}$$

I now proceed to more precisely define each of these equations and the symbols contained therein.

Stochastic component: The stochastic component is not a technical annoyance, as it is sometimes treated, but is instead a critical part of the theoretical model: "The fundamental intellectual breakthrough that has accompanied the development of the modern science of statistical inference is the recognition that the random component has its own tenuous regularities that may be regarded as part of the underlying structure of the phenomenon" (Pollock, 1979: 1).[3] The stochastic component may be written, in general, as Equation (1.3) and, for random variable i, as $Y_i \sim f_i(y_i|\theta_i,\alpha_i)$. Equation (1.3) is read: "Y is distributed as f of y given θ and α." f is a probability distribution, an explicit model of the form of uncertainty in the random variable Y across repeated experiments (see Chapter 3). θ and α are both parameter vectors. The distinction between the two is conceptual rather than mathematical or statistical. In general, θ is of more interest. In the linear regression special case, θ is $E(Y_i) \equiv \mu_i$, the expected value of the dependent random variable, Y. In another case, θ might be π, the probability of observing one of the two

[2] Still another way to think about this assumption is that the variation across hypothetical experiments in Y_i and over observations in μ_i ($i = 1, \ldots, n$) is orthogonal.

[3] At one time, scientists had as their goal the complete systematic modeling of Y, with no stochastic component. Indeed, "18th century French mathematician Pierre Simon de LaPlace once boasted that given the position and velocity of every particle in the universe, he could predict the future for the rest of time. Although there are several practical difficulties to achieving LaPlace's goal, for more than 100 years there seemed to be no reason for his not being right at least in principle" (Crunchfield et al., 1986: 47). An example demonstrating the modern understanding of this phenomenon is "a game of billiards, somewhat idealized so that the balls move across the table and cross with a negligible loss of energy." In predicting the consequences of one shot, "if the player ignored an effect even as minuscule as the gravitational attraction of an electron at the edge of the galaxy, the prediction would become wrong after one minute!" (Crunchfield et al., 1986: 49). The combined effects of many small random processes in social science research would seem considerably more substantial than in this simple physical example. See also Suppes (1984).

outcomes of a dichotomous variable. α is a vector of parameters that is present in only some models. α contains subsidiary, but potentially important information about the process that drives the observed data. In the linear regression model, for example, σ^2 is often considered ancillary.

Systematic component: The systematic component of the statistical model is a statement of how θ_i varies over observations as a function of a vector of explanatory variables. This may be written generally as Equation (1.4) for the entire parameter vector and $\theta_i = g(x_i, \beta)$ for the ith element of θ. We have already seen that in linear regression, the systematic component is a linear function of the expected value of the output of the social system Y:

$$E(Y) \equiv \mu = X\beta$$

for all n random observations, and $\mu_i = x_i\beta$ for observation i. Chapter 5 introduces the logit specification which lets the parameter $\pi \equiv \Pr(Y = 1)$ vary over observations in a particular way:

$$\pi = \frac{1}{1 + \exp(-X\beta)}.$$

In a different specialized model, we might not want σ^2 to be constant over observations, as in the homoscedastic linear regression model. Instead, σ^2 might vary as a particular function of a set of explanatory variables,

$$\sigma^2 = \exp(X\beta),$$

so that the variance of some observations would be predicted to be larger than others. Subsequent chapters describe how to formulate and interpret these alternative *functional forms,* written more generally as $g(\cdot, \cdot)$. Functional forms are precise statements of how a particular characteristic (θ) of the random dependent variable (Y) is generated by certain features of the social system (X). Although Equation (1.4) need not be a causal model, the elements of the vector β are called the *effect parameters,* representing the degree and direction of dependence of θ on the explanatory variables X. The specific interpretation of β depends on the functional form (see Section 5.2). In general, leaving β as a symbol rather than assigning it a number enables one to estimate it by empirical analysis and thus to effectively choose a particular model from the family of models in Equation (1.4).

The systematic component of the statistical model is a statement of how different known features of the social system (the explanatory variables) generate various characteristics of the random output (the parameter values), such as the average output μ. The particular function of the known features involves the functional form and the effect parameters, each of which may or may not be known in advance of any data analysis. In general, since g and β

are restricted in only very limited ways, Equation (1.4) is an extremely flexible form for the systematic component of a model.

Statistical models exist that do not fit into the framework of Equations (1.3) and (1.4), but nearly all of those used in actual research situations are either formally special cases or simple generalizations. This formulation at least conceptually encompasses linear regression, Poisson regression, probit and logit, models for tabular data, time series, nonrandom selection, factor analysis, structural equations, switching regressions, variance functions, and hundreds of others. The list includes the vast majority of statistical techniques used in political science and those of other social sciences as well.

One form that does not *appear* to fit into the general model of Equations (1.3) and (1.4) is when y is "transformed" (such as by taking the logarithm) and a linear regression is run on the transformed variable. However, transformation is merely a way of "tricking" a linear regression computer program into running a nonlinear regression. If we concentrate on what process drives our data, represented in its most substantively natural form, we are likely to retain an emphasis on theory rather than computing and technical issues.

For example, most people find it more natural to theorize about income than the log of income. If the functional form of the relationship between income and the explanatory variables is log-linear, then g can be appropriately specified [rather than trying to conceptualize and model $\ln(y)$], and we can continue to theorize directly about the variable in its most interesting form – in this case, dollars.

For those schooled in the "tricks" of linear regression programs, the formulation in Equations (1.3) and (1.4) does encompass models that authors of econometrics textbooks call "nonlinear in the variables" (i.e., those models one can trick into a regression program) and "nonlinear in the parameters" (those models one cannot get into a linear regression). Though this distinction is useful for computational purposes, it is irrelevant to substantive theory. It should have no place in theoretical discussions of research methodology or empirical analyses of particular substantive problems.

Statistics: The word "statistics" derives from the meaning (but not etymology) of the phrase "sample characteristics." Statistics have many purposes, but, by definition, they are no more than functions of the data. The general form of a statistic h is

$$h = h(y) = h(y_1, \ldots, y_n).$$

Statistics generally are of "lower dimension" than the data. For example, the sample mean $[\bar{y} = (1/n)\sum_{i=1}^{n} y_i]$ is one number $(\bar{y})$, whereas the data include n observations $(y_1, \ldots, y_n)$. A sample mean and sample variance together are probably a better description than just the mean. However, there is a trade-

off between concise and complete description: Interpreting a statistic is generally easier than the entire data set, but one necessarily loses information by describing a large set of numbers with only a few. "The fundamental idea of statistics is that useful information can be accrued from individual small bits of data" (Efron, 1982: 341). By carefully reducing dimensionality through statistics, useful information can be extracted from these many small bits of data.

Descriptive statistics: Descriptive statistics are statistics used for no other purpose than to describe the observed data. Since the goal of this enterprise is to model the process by which some social system produces output, it is often of use to begin with a simple description of the general features of the data. "Description in economics, as in other sciences, plays a key role in bringing to the fore what is to be explained by theory" (Zellner, 1984: 29). Descriptive statistics make no reference to any part of the statistical model or social system and, as used, cannot be part of a theory of statistical inference.

Estimation: Whereas *inference* is the general process by which one uses observed data to learn about the social system and its outputs, *estimation* is the specific procedure by which one obtains numerical estimates of features (usually parameters) of the statistical model. To calculate estimates, one uses statistics (functions of the data). These statistics may even be the same as are used for descriptive purposes. The use to which they are put, and the theory behind their calculation, distinguish statistics as estimators from statistics as mere descriptive tools.

Whereas *point estimation* produces a single number as a "best guess" for each parameter, *interval estimation* gives a range of plausible values. Both point and interval estimation reduce the dimensionality of the data considerably. By moving from all the data to a number or range of numbers, much interpretability is gained, but some information is lost. In contrast, *summary estimation* reduces the data as much as possible but without losing any information that could be useful for inference or testing alternative hypotheses under a single model. Summary estimators are usually too unwieldy to use directly for inference purposes, but they are often an excellent first step in reducing the data for analysis. Indeed, point and interval estimators can be easily determined once a summary estimator is available. A more precise mathematical definition of these concepts is given in the next chapter.[4]

[4] See Efron (1982: 343–4) for different definitions of similar concepts.

1.3 Concluding remarks

In contrast to the usual presentation of these concepts in introductory statistics and research methods texts, the language presented here focuses on social science theory rather than on technique: For example, introductory texts usually define the previous concepts more narrowly in terms of populations and samples, along with all the apparatus of designing sample surveys. Although surveys are an important part of political science, relatively few researchers actually design their own. Instead, they either use surveys others have conducted or, even more commonly, other types of data generated by someone or something else. Hence, the language presented above is more general than survey sampling, having the latter as a special case.

In addition, even survey researchers are not primarily concerned with "simple random sampling" or other technical methods of drawing samples from populations. Instead, what generates data in the context of individual survey interviews is more critical. For example, consider modeling the process by which an interviewer asks for the respondent's opinion on a subject about which the respondent is only vaguely aware. Will the respondent guess, report not knowing, terminate the interview, or create an opinion on the spot? Problems of nonresponse bias, self-selection, censoring, truncation, nonopinions, and other interesting theoretical problems should be the focus of statistical modeling efforts in many survey research problems.

My general approach is to focus on the underlying process that generates social science data. This is distinct from the much more narrow and usual focus on simple random sampling from fixed populations. Although understanding how to create reliable data can be important, the process of generating most social science data is not influenced by most data analysts. Indeed, in those cases when political scientists generate their own data, understanding how to derive models of these processes leads one to create considerably more interesting and useful data sets. The important question for political science research, then, is the underlying process that gives rise to the observed data. What are the characteristics of the social system that produced these data? What changes in known features of the social system might have produced data with different characteristics? What is the specific stochastic process driving one's results? By posing questions such as these, statistical modeling will be more theoretically relevant and empirically fruitful.