11

STATISTICAL INFERENCE FOR APPARENT POPULATIONS

*Richard A. Berk**
Bruce Western†
*Robert E. Weiss**

In this paper we consider statistical inference for datasets that are not replicable. We call these datasets, which are common in sociology, apparent populations. We review how such data are usually analyzed by sociologists and then suggest that perhaps a Bayesian approach has merit as an alternative. We illustrate our views with an empirical example.

1. INTRODUCTION

It is common in sociological publications to find statistical inference applied to datasets that are not samples in the usual sense. For the substantive issues being addressed, the data on hand are all the data there are. No additional data could be collected, even in principle. In this paper, we call the complete set of all units comprising such datasets an "apparent population." Consider the following examples.

Thanks go to Jan de Leeuw, David Freedman, Edward Leamer, Roderick Little, William Mason, and Donald Rubin for the many discussions that helped us think about the issues raised in this paper. Additional and special thanks go to David Freedman for his detailed written comments. Finally, we are indebted to several reviewers who held our feet to the fire. If we still do not have it right, it is not for the lack of smart, knowledgeable, and helpful colleagues.

*University of California, Los Angeles
†Princeton University

Cross-national Research. Perhaps the most common illustrations of apparent populations are found in cross-national research when a set of countries is studied at one point in time. An early example is Lipset's (1959) classic study of the relationship between democrary and economic growth in 48 industrialized and developing countries. Since then, many scholars have explored the links between democracy, development, income inequality, and political violence. Reviews of this material can be found in Bollen and Jackman (1985), Rueschmeyer and colleagues (1992), and Jenkins and Schock (1992).

OECD Countries. A second generation of quantitative comparative research has focused on a much smaller group of industrialized capitalist democracies. Analyses of approximately 18 OECD countries have explored the causes of social welfare spending, economic performance, the development of political institutions, and unionization (Stephens 1979; Wilensky 1985; Lange and Garret 1985; Griffin et al. 1990; Western 1991).

American Cities. Urban sociologists also commonly rely on data from apparent populations. In a recent paper on the effects of American trade unions on racial and ethnic conflict, for instance, Olzak (1989) analyzed data on the 81 largest American cities. Other popular datasets contain information on a collection of SMSAs. For example, Ross's 1987 study of corporate networks in metropolitan centers uses data from all 157 SMSAs in the United States. Lincoln (1978) and Abrahamson and DuBick (1977) used SMSAs in their earlier studies of urban dominance.

Miscellaneous Data. Data from apparent populations are common in political sociology (e.g., Stepan-Norris and Zeitlin 1989), criminology (e.g., Ehrlich 1977), evaluation research (e.g., Wright et al. 1979) and economic sociology (e.g., Wallace and Kalleberg 1981).

Data from apparent populations create conceptual problems for statistical inference. First, the data are not generated by probability sampling or random assignment. Second, and far more important for this paper, an apparent population is the result of a data generation mechanism that produces only a single batch of data. In effect, the machinery is turned off after a single batch is produced; the data generation mechanism cannot be expected to produce another dataset. Extending common illustrations, a coin is only flipped once and then thrown away, or a roulette wheel is spun a single time, and then dismantled, or one hand of blackjack is dealt and then the game is over.

There have been two kinds of reasoned responses to such data-generating mechanisms. On the one hand, the data are treated as "fixed," and the usual concerns about sampling error are discarded. That is, the data are treated as if they could not have been different as a result of sampling error. Then, conventional statistical inference becomes irrelevant. An apparent population has become a true population, and descriptive statistics are employed.

On the other hand, the data are treated as a "realization" of some set of social process that could have in principle produced a very large number of other realizations. These realizations, in turn, constitute a super population. That is, the data could have been different as a result of random sampling from the "super population." Then, conventional statistical inference is applied as usual. An apparent population has now become a random sample.

In this paper, we explore the tradeoffs between these two approaches and then argue that Bayesian reformulation might be a better alternative. We argue that since at least part of the problem is the usual frequentist interpretation of probability, it may well make sense to replace it. Our arguments are meant to reach a broad audience of sociologists; we have tried to pitch the paper down the middle.

2. A BRIEF REVIEW OF FREQUENTIST STATISTICAL INFERENCE

Frequentist statistical inference is meant to address uncertainty in data (and therefore in any computed sample statistics) introduced by the *sampling* mechanisms that produced the data. In particular, the data are typically treated as a random sample from a specified parametric distribution (Barnett 1982 pp. 121–23). Other sources of uncertainty are not addressed. For example, in sample surveys, such problems as nonresponse, response errors, and interviewer error may be terribly important but are not formally included (Groves 1989, ch. 1).

One of two generic sampling strategies is typically assumed (Berk and Brewer 1978; Noreen 1989): probability sampling (e.g., in surveys) and random assignment (e.g., in experiments). Since for both probability sampling and random assignment the data are generated through *known* probability procedures, it is possible to formally represent the uncertainty that results. In effect, we have an accurate model of the data generation process. By capitalizing on what would happen in principle if numerous samples were generated independently by the same process, it then becomes possible to represent the uncertainty in our parameter estimates by such devices as a sampling distribution for parameter estimates in a very large number of identical, hypothetical, "trials." Confidence intervals and significance tests follow.

It is important to stress, however, that the entire enterprise as usually undertaken depends on at least three premises. The first is that we know how the data were generated. That is, probability sampling and/or random assignment were used in known ways. Often, however, it is quite difficult to know how the data were sampled or assigned because the research design breaks down in practice. For example, some prospective respondents may refuse to be interviewed.

The second premise is that we are satisfied with an interpretation of uncertainty for parameter estimates as a distribution of outcomes over a very large number of hypothetical, identical, and independent trials. We share the concerns of many observers who are uncomfortable with a long-run interpretation of uncertainty when, as a practical research matter, all the data are on hand. Also in practice, researchers cannot wait for the long run to arrive (Barnett 1982, ch. 3). Thus we might be able in principle to draw 500 independent samples of 1,000 from the population of all adults living in a single metropolitan area, but what does that say about sampling error for the single random sample actually on hand? For other criticisms of frequentist inference, see Leamer (1978) and Barnett (1982).

The third premise is that the thought experiment of a large number of identical and independent trials is consistent with how in fact the data were generated. That is, we do not need to invent a science fiction world to make instructive the idea of a large number of identical and independent trials. How easy it is to do this is controversial (Leamer 1978, ch. 2). What kind of world would one have to have, for instance, to be able to actually take 500 independent simple random samples of all adults in a single metropolitan area? And how realistic is it to think about someone actually doing it?

In short, statistical inference based on imagery of the long run is an imperfect tool even in theory and even when applied in the field as well as scientists know how. When we do not have the data generated by random sampling or random assignment, a number of new problems are introduced. We turn to these now.

3. STATISTICAL INFERENCE OF APPARENT POPULATIONS

Imagine a longitudinal dataset for a single developing nation in which a number of economic indicators and measures of social unrest are included. The substantive question might be whether in this country for the historical period represented, there is a relationship between economic growth and social unrest. Clearly, the data do not derive either from random sampling from some defined population or from random assignment to experimental and control conditions. The data generation mechanism is, on its face, not ignorable. More important for our purposes, both the country and the particular historical period are not exchangeable with other countries or historical periods. Consequently, the data are a single trial. What then might be done about sampling error? Or is there any need to worry about it at all?

3.1 *Treating an Apparent Population as a True Population*

Under one conception, the data may be treated as a (fixed) population. There is, after all, only one such country during the historical period of interest. Then, statistical inference is moot. Descriptive summaries of the population can be used. In effect, we assume the data could not have been different because of sampling error. And the data from the apparent population become data from a true population.

It cannot be overemphasized that a completely deterministic world is necessarily assumed, at least with respect to sampling error. If the data could not have been different because of sampling error, the data-generation process is effectively deterministic by default. Disparities found in attempts to replicate the results are caused by other kinds of error or substantive differences across sites or historical periods, not sampling error. In our example, this implies that if the historical process for the particular period and for the particular country in question could be started again, the data would turn out exactly the same. Put another way, given the conditions extant in the beginning of the period, one (and only one) set of events could have unfolded. Should such a conception of historical processes be disconcerting, the alternative is to introduce the possibility of stochastic historical processes, which we shall address shortly.

The same issues arise with any apparent population treated as a true population. We might have, for example, data on all of the NATO countries for 1980, or ten years of data on each of the ten largest cities in South America, or 20 years of the data from the turn of the century on all of the counties in Louisiana. If we are prepared to accept a deterministic view of the social process that generated the data, we may dispense with statistical inference. Perhaps the major message is that there is a price to be paid for treating the data as a population; we are not simply discarding a technical embellishment but making a positive commitment to a particular view of how the data were produced.

3.2. *Treating an Apparent Population as a Random Sample*

An alternative to treating an apparent population as a true population is treating an apparent population as a random sample. But if an apparent population is to be a random sample, the population from which it is drawn must be defined. This population is sometimes called a "super population," with samples from the population called "realizations." Then we can rely on the thought experiment of what would happen in a very large number of random and independent draws (or realizations) from this super population and apply conventional statistical inference as usual.

It is likely that in the sociological literature whenever statistical inference has been applied to apparent populations, the super population framework is being implicitly assumed. It seems useful, therefore, to consider the issues in a bit more depth. We suspect that even researchers who are thoroughly familiar with the mathematics of frequentist inference sometimes lose sight of its applicability.

Consider first a model of radioactive decay (Eisberg and Lerner 1981, p. 907). Over a short period of time, counts of the number of particles per time period are well described by a Poisson distribution (Johnson and Kotz 1969, pp. 88–89), and the assumptions necessary for that distribution are consistent with what is known about the physics. In effect, therefore, there is an unobserved population distribution that is well described by some member of the Poisson family of distributions, and each count can be viewed as a random and independent observation from that distribution. A set of such observations can be treated as a random sample from which summary statistics, such as the mean, can be computed. In short, the physical world is modeled as if a simple random sample from an unknown but parameterized population were the data-generating mechanism. Finally, we can easily envision a very large number of realizations of the same process and sampling distributions of the summary statistics.

In contrast to a random sample from a known population, the population of particle counts does not already exist. In the case of radioactive decay, however, we could in principle generate a very close approximation of the population by monitoring the radioactive decay over a long period. Likewise, we could in principle collect data from a very large number of realizations and construct the sampling distribution of a parameter estimate. Furthermore, what is known about the data-generating mechanism indicates that the process by which particles are produced is a very good approximation of simple random sampling, well described by a tractable distribution. Consequently, it may be instructive to apply the same mathematics to both processes; random sampling done by a scientist and radioactive decay.

Consider now a study of the number of suicides occurring in Manhattan per day. We might begin by assuming on theoretical grounds that the data-generating mechanism is a Poisson process, implying that the count on any given day is a Poisson random variable (Larson 1969, pp. 146–47). That is, there is an underlying *social* process producing suicides whose outcomes on each day may be described by the density for a Poisson random variable. In principle, any given day may be associated with one of several different counts. As a result, the count associated with any given day can be conceptualized as a realization of a Poisson process and as a random sample of one from a distribution of all possible counts. Then, a set of days becomes a realization of the process from which summary statistics can be computed. And as before, we can easily imagine the empirical world generating a very large number of such realizations and sampling distributions for the summary statistics.

In effect, it can be assumed that the *social world* generates the equivalent of a random sample much like the samples generated by radioactive decay; the social world replaces the physical world as a sampler. In this context, there are three key assertions. First, the data on hand could have been different. Second, the number of suicides on any given day (or some other time period) is a simple random sample of one from an underlying Poisson distribution characterizing the population.[1] Third, the social process producing suicides is stable enough so that it is not fatuous to imagine a very large number of realizations behaving like a very large number of independent, identical trials. Put another way, with sufficient patience we could actually produce a close approximation to the sampling distributions in question. For us, both the second and third assumptions require a very long reach. (See, for example, Phillips' work on suicide [1974]).

3.3. *Randomization Tests*

An analogous imagery can be applied to random assignment. We might be interested in, for example, whether there is any association between voter turnout and how close an election is (Noreen 1989, pp. 19–22). In particular, for the presidential election of 1844, data might be collected for each state on the percentage of eligible voters who voted and the absolute difference in the percentage of the total obtained by each of the two candidates (i.e., Polk and Clay). One empirical possibility is that the two variables are really independent. To pursue this, we might imagine how the variables would be related if they were indeed independent; which suggests considering what would happen if the data for one of the variables were thoroughly shuffled. (Shuffling one variable is enough, and equivalent to shuffling both.) That is, the order of the observations for the first variable are random with respect to the order of the second. Put another way, the values of the first variables are assigned to states at random.

[1] This requires that the population be very large, that the probabilty of an individual committing suicide is constant over time and very small (or equivalently, that the mean number of suicides in the population is constant), and that individuals are independent.

What would the association between the two variables be then? And if a very large number of such shuffles were undertaken, and each time a measure of association computed, a sampling distribution under the usual null hypothesis of the measure would result. It would then be possible to formally address the variability in our measure of association in the null situation of no association by comparing the actual value computed from the data in the absence of shuffling to the null value, through such devices as tests of statistical significance. We embed the actual event observed in an artificially constructed population of presumably similar events.

Such randomization tests capitalize on the formal mathematics of random assignment (Edgington 1987), but it is hypothesized that the social world, not the researcher, is doing the assignment at random. How reasonable this is typically depends upon the three premises much like those discussed earlier. First, the empirical relationship between the two variables is stochastic; the observed association could have been different. Second, given the null hypothesis, the data on hand are a random sample from a null population of all possible random shuffles of the data. Third, it is not fatuous to think of the social world (under the null hypothesis) as producing a very large number of such shuffles, at least in principle. In this instance, the three premises add up to rather strong statements about the historical processes associated with the election of 1844. For example, it seems implausible to us that the social world could manage to shuffle the data so that the bivariate statistics would change but the univariate statistics would not.

Freedman and Lane (1983) provide an interesting variation on this theme. They operationalize a surprise faced by researchers with apparent populations as the degree to which the data represent in some general way an artifact or fluke. If in all possible shuffles, summary statistics equal to or greater than the given summary statistic are very rare, that statistic is to be taken seriously. Note that nothing about the *social world* is being asserted. In effect, Freedman and Lane operationalize an "artifact" as an outcome that would have been uncommon under a shuffling scheme implemented by the *scientist.* Unfortunately, it is not altogether clear why we should care. The original problem was uncertainty produced by the social world. Uncertainty introduced by the scientist once the data are in hand is another matter entirely. Therefore, justification for the Freedman and Lane approach must be found elsewhere. We might try to make the case, for example, that the test is somehow *useful* despite the absence of any explicit links to the empirical world from which the data came.

However, approaching conventional statistical inference through its "usefulness" is very different from approaching statistical inference as if we were modeling the data-generation mechanism. When we have random sampling or random assignment, frequentist inference rests (at least in principle) on a reasonable approximation of how the data were *actually* sampled. And there are typically empirical clues (e.g., the response rate in sample surveys) with which one can explore how well the model corresponds to the ways in which the data were produced. For apparent populations, there is no sampling and, therefore, no clue about whether the data were generated according to some sampling plan. To proceed as if the data were generated by random sampling or random assignment is to embrace a fiction. Then, to justify the application of conventional statistical inference by claiming that it is nevertheless "useful" is unsatisfying unless it is possible to clearly explain what is meant by "useful." We suspect that this is very difficult to do, and we have yet to find an instructive discussion in the scholarly literature (see, for example, Morrison and Henkel 1970; Barnett 1982; Oaks 1986).

In summary, there are two ways to treat an apparent population as a sample: the first employs the concept of a realization from a super population and the second the concept of random shuffling. But in both cases, we proceed as if the social world is capable of producing a very large number of independent trials. It is, of course, easy to assume that a very large number of independent trials could be produced, but for the kinds of apparent populations likely to be of interest to sociologists, the implied data-generation mechanism would likely be at variance with the relevant social theory. More examples follow shortly.

3.4. *Extensions to Regression Problems*

For most work in sociology, the associations of interest are typically conditioned on "control" variables. Thus we might examine the relationships between turnout and the competitiveness of the presidential election of 1844, controlling for such things as the proportion of voters registered as Democratic. This requires moving to a slightly more complicated framework. However, there are really no new issues involving statistical inference for apparent populations, so we will be very brief.

Multiple regression is ubiquitous in sociological research and will illustrate nicely how the influential approaches described above may easily be extended to conventional multivariate problems. We will assume, consistent with most expositions, that the explanatory variables are fixed. That is, we condition on the observed values of the explanatory values. Only the response variable is subject to stochastic forces.

If the data being used do not derive from probability sampling, one can, once again, tell a story in which the social world produces data *as if* by repeated random sampling. For each configuration of values for the explanatory variables, an observed value of the response variable is viewed as but one of a very large number of values that could have been observed in principle.

When such cases have to be made, it is often easier to formulate the regression model so that each observed value for the response variable is comprised of two parts: a deterministic part and a stochastic part. The deterministic part is the linear combination of values of the explanatory variables. The stochastic part is a disturbance term that behaves as if each value were independently drawn at random from some specified distribution, usually the normal distribution. More formally we can write

$$Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + u_i,\ u_i \sim N(0, \sigma^2). \tag{1}$$

The disturbance term is interpreted as a random perturbation that introduces stochastic variation into Y_i. Then, a plausible argument must be made that the social world can repeatedly generate these perturbations in a manner that meets certain distributional assumptions; in a very large number of opportunities, the social world must produce disturbances that behave as if drawn over and over independently from a particular normal distribution (although the assumption of normality is often not critical).

We are hard-pressed to find any examples in the sociological literature in which the relevant theory and/or empirical literature directly justifies such formulations. Why should the disturbances associated with labor unrest, for example, behave as if drawn independently and randomly from a particular normal distribution? In contrast, consider illustrations for the physical and biological sciences where the sources of stochastic variation are a key part of any theoretical formulations and from which particular probability distributions necessarily follow. In addition to the illustration of radioactive decay described earlier, consider Mendelian genetics, the dispersion of pollutants in stream channels (Young and Lee 1993) or variation in the size of fish populations (Speed 1993).

Therefore the usual justification for the disturbance distribution, when one is provided, is that if there are a large number of omitted variables, each uncorrelated with the explanatory variables included in the model and with each other, and each with a small impact on the response variable, they will produce in the aggregate a normal distribution of IID disturbances consistent with the assumptions of the regression model. In effect, potential explanatory variables are of two rather different kinds: a few with "big" effects and many with "small" effects. Then, if the former are explicitly included in one's regression model, the latter can be swept into the disturbance term. Why the social world should be so conveniently divided into distinctly big and distinctly small effects is, of course, unaddressed. And why should these effects be uncorrelated? Equally important is whether it is sensible, once again, to imagine a social world capable of generating a very large number of independent draws from the same underlying disturbance distribution.

To bring these statistical ideas down to concrete applications, consider a factory in which a machine produces over the course of an hour a number of discrete products such as small ball bearings. Suppose one measure of quality is the smoothness of each ball bearing and that despite efforts to ensure homogeneity in raw materials and production processes, some ball bearings are smoother than others. Viewed from the perspective of a regression model, smoothness could be the response variable, measures of the inputs (e.g., purity) and the production processes (e.g., temperature) could be the explanatory variables, and the minor variation in smoothness from other sources could be represented by a disturbance term.

The key point is that for all practical purposes, one has in the ball bearing machine a replicable process. If the machine is running properly and the mix of inputs is stable, a very large number of ball bearings could be produced that would be more or less alike. And it is arguably instructive to apply a regression formulation with the disturbances approximated by random, independent draws from some distribution.

For the kinds of macro datasets common in sociology, it is difficult to believe that the social processes possess the same kind of replicability as our ball bearing machine. We would have to postulate a social world that, much like the ball bearing machine, is capable of producing some outcome over and over again in more or less the same fashion with more or less the same result. In effect, *we would have to assume that history can be rerun, not just once or twice, but over and over.*

In summary, the usual regression formulations for apparent populations rests on a particular and, in our view, implausible view of the social world. In order to tell a sensible story about the sources of stochastic variation, a hypothetical world of replicability is constructed. In effect, the social world replaces the scientist and is compliant enough to generate just the kinds of data the scientist needs. Arguably, this is a formulation driven more by convenience than reality.

4. BAYESIAN INFERENCE

We have argued that for apparent populations, a fundamental problem with conventional imagery of statistical inference is the requirement of a very large number of independent trials. But this derives in part from the frequentist interpretations of probability. Recall that for a frequentist, the probability of an event is commonly interpreted as the limiting value of the relative frequency of that event as the number of independent trials grows without limit. Subjective conceptions of probability have very different foundations (Barnett 1982, ch. 3). In brief, subjective probability is an expression of uncertainty experienced by an individual. Thus, a statement that the probability of rain on a given day is .90 denotes how certain the speaker is that it will rain. It is not a claim about the limiting proportion of days, like the one in question, when it will rain.

Modern formulations of subjective probability rest on a few axioms that many people would agree are sensible ways to conceptualize uncertainty assessments (Lee 1989:2–10). With a few simple and reasonable axioms, it is possible to define probability as a degree of belief that, in turn, obeys the usual probability theorems. That is, the usual mathematics apply by and large, but they are given a different interpretation. Many scientists and statisticians argue that the subjective conceptions of probability fit more naturally with the way science is done, especially when combined with an apparatus for expressing and revising one's degree of belief (e.g., Howson and Urbach; Press 1989). That apparatus in called Bayesian inference.

Bayesian inference takes many forms that imply different computations and different inferential philosophies. But its most common form begins with a *prior* probability density expressing one's beliefs about some empirical fact. Then those prior beliefs are revised in light of some data, with the result being a *posterior* probability distribution expressing one's new beliefs about the empirical fact.

We intend to just touch a few highlights and then to illustrate the use of Bayesian inference for the kinds of problems common in macrosociology. Textbook treatments are easily found (Lee 1989; Press 1989; Pollard 1986; Broemeling 1985; Box and Tiao 1973; Zellner 1971; O'Hagan 1994; Bernardo and Smith 1994). Readers might also keep in mind that *empirical* Bayes methods, such as those advocated by Mason and his colleagues (Mason et al. 1983) are very different from the Bayesian methods we will consider.

4.1. *An Introductory Illustration*

For many applications, the numerical conclusions from Bayesian analysis will be identical, or nearly identical, to the conclusions from frequentist analysis. Nevertheless, comparing the two approaches can be very instructive.

Suppose the parameter of interest for some data analysis is the mean μ of a population from which we draw a large sample of n observations. A mean $\bar{X}$ is computed and by the central limit theorem, $\bar{X}$ is known to be approximately normally distributed with a mean of μ and a variance (σ^2/n). This knowledge is usually referred to as a *model,* and can be written as

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{2}$$

Often n is large enough so that the sample variance s^2 can be substituted as an estimate for σ^2. Confidence intervals may be constructed for the unknown population mean μ. The 95 percent confidence interval for μ is

$$\left(\bar{X} - 1.96\frac{s}{\sqrt{n}}, \bar{X} + 1.96\frac{s}{\sqrt{n}}\right). \tag{3}$$

From the frequentist point of view, the 95 percent confidence interval will in *repeated samples* cover the true mean μ 95 percent of the time. Recall that this is a statement about the confidence interval construction procedure and not a statement about the particular interval constructed from the particular data set. Scientists typically are interested in the particular data set, conclusions about the data set, and the population it was collected from, not in statisticians' claims about their statistical procedures.

A Bayesian approach may initially proceed in a similar fashion. However, having written down the model in equation (2), we can proceed further and write

$$\mu \sim N(\bar{X}, \frac{\sigma^2}{n}). \tag{4}$$

That is, μ has a normal distribution centered on $\bar{X}$ with a variance of σ^2/n. It follows that a 95 percent confidence interval is exactly the same, but the interpretation is different.[2] The Bayesian interpretation is that the μ for this problem is within the computed interval with a probability of .95. The probability .95, in turn, is a statement about the degree of belief of the researcher, who is 95 percent certain that μ falls in the particular interval computed from the data. There is no reference to hypothetical repeated sampling, which means that there is no need to postulate or conceptualize the existence of an arbitrarily large number of additional random samples.

Given a null hypothesis, equation (2) can also be used by frequentists to implement tests of statistical significance. With mean μ assumed equal to some value, the frequentist can easily compute

[2] Bayesian intervals are often called credible intervals to distinguish them from confidence intervals. However, calling Bayesian intervals confidence intervals is not uncommon, and it is a practice that we follow here. Under suitable conditions, Bayesian intervals are asymptotically confidence intervals, so it seems reasonable to use the familiar term confidence intervals.

the probability that in repeated samples drawn from the population characterized by the null hypothesis, means more extreme than the observed mean $\bar{X}$ could have occurred. Once again, the inference is about what might have happened in a large number of additional random samples. Additionally, however, the null hypothesis can be viewed as information external to the data brought by the researcher to the data analysis. Before looking at the data, the researcher makes an assertion about a likely value of μ. Bayesians develop and use such external information rather differently.

4.2. *Combining Information*

In our discussion of confidence intervals, both frequentists and Bayesians proceeded as if they had no information about μ before looking at the data. This is often a reasonable way to proceed, and we favor such approaches as a part of a larger data analysis. However, there is sometimes information available about plausible values for μ before the data are examined. For example, there may be theoretical reasons to believe that μ falls within a particular range, or there may be estimates of μ from earlier studies. When such prior information exists, Bayesian methods provide a principled, general apparatus for combining the prior information with information from the current study.

The use of prior information within the Bayesian framework is often controversial and a full discussion of the issues is well beyond the scope of this paper. But in brief, much of the concern arises because the prior information may be internal to the data analyst. The prior information is not generated by a process to which others necessarily have access. The process therefore, is not necessarily replicable and as a result, prior information may be sometimes deemed unscientific. Another concern is that the manner in which Bayesians must present prior information (see below) may lead to formulations of the prior information driven more by mathematical convenience than by the prior information itself.

In response, Bayesians note that frequentists also employ prior information. Furthermore, the prior information is used in a manner that is neither well rationalized nor well documented. How, for example, is a particular null hypothesis generated? How is a particular model or estimator selected for one data set and not another? In addition, Bayesians observe that if prior information is unacceptable, critics are not just free, but encouraged, to employ prior information of their own, or proceed as if no prior information existed. Also, if the number of observations in a dataset is large and the analysis is not too complex, the prior information will usually be swamped; the prior information will have very little impact on conclusions of the data analysis. Finally, in complex studies where the prior information is based on clearly explained previous studies, the prior may find greater support among skeptical researchers than the model itself.

To illustrate, we can extend our previous example. We begin with equation (2) as before: $\bar{X}$ is distributed normally with a mean of μ and a variance of σ^2/n. We then introduce prior information about the value of μ. In one of the mathematically simplest possible situations, the prior information can be represented as the following prior distribution for μ:

$$\mu \sim N(\mu_0, \sigma_o^2). \tag{5}$$

The subscript "0" is used to indicate "time 0," before the data have been examined. Equation (5) implies that μ is near μ_0, but how near is uncertain. That uncertainty, in turn, is represented as a probability: The probability that μ is more than $z_{1-\alpha}\ \sigma_0$ greater than (or less than) μ_0 is α, where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of a standard normal variable. In effect, the $(1 - 2\alpha)$ percent confidence interval is constructed to characterize the researcher's information about μ before looking at the data.

By Bayes theorem, equations (2) and (5) can be combined to yield

$$\mu|X \sim N(\mu_1, \sigma_1^2), \tag{6}$$

where the subscript 1 is used to indicate time 1, after the data have been examined, and

$$\mu_1 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_o^2}\right)^{-1} \left(\frac{n}{\sigma^2}\bar{X} + \frac{1}{\sigma_0^2}\mu_0\right), \tag{7}$$

and

$$\sigma_1^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_o^2}\right)^{-1}. \tag{8}$$

Equation (6) is called the posterior distribution of μ and represents what the researcher believes after the prior information about μ is combined with information from the data. The posterior mean of μ is a weighted average of μ_0 and $\bar{X}$, with the weights determined by the relative sizes of their variances. A larger variance implies less weight. Thus, if the researcher feels that virtually nothing is known about the likely value of μ, σ_0^2 will be a very large number, which makes $1/\sigma_o^2$ a very small number. As a result, μ_1 will be very close to $\bar{X}$. In this case, prior information does not affect the conclusions greatly. In contrast, if the researcher feels that the value of μ is almost certainly μ_0, then σ_o^2 will be very small, which makes $1/\sigma_o^2$ a very large number. Then μ_0 and μ_1 will be almost the same. Prior information dominates the story.

The variance σ_1^2 of μ_1 is determined by the variances of $\bar{X}$ and μ_0, and it is smaller than either. That is, the spread of the posterior distribution will be smaller than the spread of the prior or the standard error of the sample mean. In this example, the information contained in the prior distribution leads to improved precision (i.e., smaller variance).

For the Bayesian, all of the information required for statistical inference is contained in the posterior distribution. That information may be summarized in several ways. Confidence intervals are popular, but it is also common to see plots of the entire posterior distribution. In any case, however the posterior is summarized, there is no reference to replications of the data generation process because the distribution refers to the researcher's degree of belief about a parameter after examining the data and conditional on the given model.

4.3. *Multiparameter Applications*

Now we consider the same issue more abstractly. Drawing on Lee's (1989) discussion, suppose for the substantive phenomenon in question, there are k unknown quantities of interest,

$$\theta = (\theta_1, \theta_2, \ldots, \theta_k). \tag{9}$$

These may be means, variances, correlations, regression coefficients or other kinds of parameters, or transformations of them. Also suppose that the researcher is prepared to express some prior beliefs about the values of these k unknowns in the form of a joint prior density function $p(\theta)$. In practice, the expression for the prior density of the parameters θ often includes a convenient model for the prior distribution (e.g., a multivariate normal density) and statements about the mean and variance of that multivariate normal density. In our illustration to follow shortly, there are three parameters for a model with an intercept in which union density is regressed on two predictors.

Data are collected that are thought to contain information about the k unknown quantities. For n observations we have

$$X = (X_1, X_2, \ldots, X_i, \ldots, X_n). \tag{10}$$

In our illustration, X_i for country i is the value for union density. Union density can be conceptualized as having a probability density that depends on the two other explanatory variables and on θ. Often, we write the conditional density of the data more generally as

$$p(X|\theta) = \prod_{i=1}^{n} p(X_i|\theta). \tag{11}$$

This says that the individual countries provide independent information about the parameters, and the joint density is the product of the conditional densities of the X_i given the parameters.

There are at least two ways in which the conditional density function for the data may be interpreted. For the "classical" Bayesian (Diaconis and Freedman 1986), the conditional density function is part of a "true" model for the data, and there is a "true" parameter that is to be estimated from the data. It is then possible to resurrect the imagery of a very large number of independent trials; the conditional density is the limiting distribution of data, given θ, as the number of independent trials grows without limit. For the "subjective" Bayesians, in contrast, the conditional density function for the data is a statement of belief about the likely values that the data will take given θ (Jeffreys 1961, pp. 28–29). That is, uncertainty about the parameters, *and* uncertainty about the data given the parameters, are expressed in subjective terms. The implications of these different interpretations of the conditional density function will be considered later. (See Leamer 1978, ch. 2, for an extended discussion of these differences.)

A key step in any Bayesian data analysis is to revise the prior density function based on information contained in the likelihood. Bayes theorem for random variables provides the apparatus, and the general method for it is

$$p(\theta|X) \propto p(\theta)p(X|\theta). \tag{12}$$

Stated in words, the posterior density of θ, given data X, is proportional to the prior density of θ times the density of X given θ.[3] The last term is the likelihood function, and it is sometimes written

$$l(\theta|X) = p(X|\theta). \tag{13}$$

Writing $l(\theta|X)$ in this way, we can emphasize that the data are treated as fixed, and the likelihood is considered to be a function of θ.

4.4. *Interpreting Multiparameter Results*

The natural product of Bayesian analysis is a joint posterior density for all the unknowns. However, the story that the researcher wants to tell is usually in terms of the *marginal* density for each unknown. That is, we want to *average over* $k - 1$ unknowns to obtain a density for the k^{th} unknown alone. Then, we repeat the process for the other unknown θ's in turn. If there are six θ's, for example, we usually want to extract some summary of the marginal density for each θ individually (e.g., a confidence interval for each θ), not some summary of a single joint density for all six θ's as in a joint confidence interval for all of the θ's).

In practice, obtaining the marginal density for each of the unknowns from the single joint density can be extremely difficult, especially if exact results are desired. However, there a number of recent developments that lead to useful approximations in practice (e.g., Tierney and Kadane 1986; Gelfand and Smith 1990) and that are increasingly available in statistical software (e.g., Xlispstat, S-Plus). For a variety of applications, therefore, it is relatively easy to obtain a useful approximation of the desired marginal densities.

[3] The proportionality relationship can be replaced with an equality relationship by including the proper constant in the denominator on the right-hand side. This is unnecessary for most applications and is commonly ignored (Lee 1989 pp. 34–35).

4.5. *Bayesian Sensitivity Analysis and Diagnostics*

Since our version of Bayesian inference depends on a subjective notion of probability, it is reasonable to ask how one researcher can rely on another researcher's results. Bayesians have at least two responses. The first is to note that the Bayesian analysis frequently gives similar results to a likelihood analysis. Consequently, if we expect research consumers to depend on the results of a likelihood analysis, we can expect them to depend on the Bayesian analysis as well. In reality, this argument rests on the false assumption that a frequentist analysis should be automatically usable by a research consumer. For example, Begg (1990) presents several analyses of a single data set, and the wide variety of results that can arise from different sets of starting assumptions.

Some Bayesians (as well as some frequentists) offer a second response. Since others (consumers) may bring different prior information to an analysis, Bayesians use sensitivity analyses to assess the stability of conclusions to this prior information. If the conclusions are stable over a range of prior information, then anyone whose prior information is at least close to the set of priors considered should be able to depend on the research results. If the sensitivity analysis shows that the results change dramatically with varying prior information, the research consumer can simply employ his or her preferred priors.

Prior information has traditionally meant the prior distribution $p(\theta)$, but it is being realized increasingly that prior information also includes the model $p(X|\theta)$. In this context, it cannot be overemphasized that any inferences that might follow from the posterior density depend on how well the model captures either the essential features of the data-generating processes or, for a subjective Bayesian, how well the model captures one's opinion about the relation between the data and the unknown parameters. Bayesians and frequentists may face similar issues in model specification. Even for randomized experiments or probability sample surveys, there will be important features of the structure that cannot be known with certainty, and matters are typically far worse for observational studies (Berk 1988; Freedman 1991; Draper 1993). In short, for both Bayesians and frequentists, statistical inference may be moot well before any computations are undertaken.

There are a variety of diagnostics for Bayesian inference, many related to diagnostics used in frequentist inference (e.g., Pettit 1986; Allenby 1990; Draper 1993). For example, it is often useful to consider how the posterior density of key parameters changes when potentially influential observations are deleted one at a time (Weiss and Cook 1992). It is sometimes instructive to examine the predictive distribution $p(X)$ of a Bayesian analysis either for diagnostic or substantive reasons (Box 1980).

In our earlier notation:

$$p(X) = \int p(x|\theta)p(\theta)d\theta. \qquad (14)$$

The predictive distribution is the distribution of X implied by the model taking into account the uncertainty in the models parameters (Lee 1989 p. 36). As such, it represents a prediction of the likely value of X. Just like the conditional distribution $p(X|\theta)$, the predictive distribution may be interpreted from either a frequentist or subjective perspective. Under either interpretation, changes in the predictive distribution under various alterations in a proposed model can be used to gauge the importance of those alterations.

4.6. *Reference Priors*

To this point we have been assuming that the researcher can construct a prior density to represent a priori knowledge. What if the researcher is really approaching the data with a blank slate? What if there is little or no preexisting information on which to base a prior density?

As noted in Section 4.2, the larger the variance of the prior density, the less certain the researcher is a priori about the likely values of the unknowns. One could imagine, therefore, simply increasing the variances substantially so that for all practical purposes, the density is "flat," at least in the region of interest. Prior densities of this sort are called "flat priors," "ignorance priors," "uninformative priors," "Jeffrey's priors," (Box and Tiao 1973) or "reference priors."

Although there is a long and controversial literature on how best to represent prior ignorance (e.g., Barnett 1982, pp. 206–17), in many circumstances it may not matter very much. In practice, we can simply drop the prior from our computations, and let the posterior density depend solely on the likelihood function. Alternatively, we can try several different reference priors and see whether the posterior densities are meaningfully different. Except possibly in small samples, they will not be.

4.7. *Bayesian Regression*

At the introductory level at which we are proceeding, it is relatively easy to make the jump to Bayesian regression. To begin, we formally express our beliefs about the regression coefficients before the data are examined. The regression coefficients are treated as uncertain in the sense that we do not know their values.[4] This uncertainty, in turn, is expressed as a joint probability density of these parameters. For example, if θ_0 is a vector of prior means, and Σ_0 is a prior covariance matrix, and the joint density is taken to be normal, the prior for joint density of the regression coefficients may be summarized as

$$\beta \sim N(\theta_0, \Sigma_0). \tag{15}$$

The vector of means locates the prior probability density; large values imply that the researcher believes that the coefficients are large. The main diagonal of the covariance matrix contains the variances; as measures of spread, larger variances imply greater uncertainty. The off-diagonal elements contain the covariances, which indicate the degree of dependence between the regression coefficients.

Sometimes, a researcher may have very little information a priori about the joint prior probability density of the regression parameters. In this case, the prior joint density will contain very large elements down the main diagonal of Σ. Effectively, the joint prior density is "flat." In practice, however, we think "total ignorance" is very rare. Although it may seem contrived to introduce prior information when in fact the researcher has very little, null hypotheses in conventional frequentist inference are essentially beliefs about the regression coefficients articulated before the data are analyzed, and hypothesis testing is endemic in sociology.

With the prior density specified, information in the data is

[4] For this discussion, we will assume that the sampling variance of the distribution is known. No essential points will be lost. See Box and Tiao (1973) for a more general treatment.

introduced into the analysis through the likelihood function. In the regression case, this requires that a joint probability density for the data given the parameters, has to be previously specified. The normal is certainly convenient, but it is only an option. The data are taken as fixed, and relative likelihood values are associated with different possible values for parameters. Then, the joint and marginal posterior densities may be computed as described earlier.

4.8. *Bayesian Inference for Apparent Populations*

Recall once again that apparent populations are characterized by single batches of data; the data-generation mechanism produces only the data on hand. Consequently, it is difficult to employ the imagery of a limiting distribution as the number of trials increases without limit. And this difficulty applies to both frequentists and classical Bayesians when the sampling density of the data given the parameters is considered. However, subjective Bayesians, who treat the conditional density as a representation of *their* uncertainty, have no inferential problem with a single batch of data. A sample space may be identified as usual, but the probabilities attached to each event in that sample space are subjective and do not depend on a large number of trials (Leamer 1978, ch. 2).

Consider the following example. Suppose a sociologist is interested in the possible impact of the North American Free Trade Agreement (NAFTA) on jobs in the southwestern United States. Clearly, NAFTA is a very particular treaty set in very particular geographic historical circumstances. Whatever its impact, the imagery of a large number of independent trials is a very long reach.

Suppose that one measure of impact will be the total number of jobs lost or gained within three years after the treaty's ratification. Suppose also that before the data are collected, the sociologist chooses to represent the structure of her *beliefs* about the likely *value of the total* with a normal distribution, conditional on parameters for the mean and variance. In other words, her model for the conditional density of the total number of jobs lost or gained is a particular (but as yet unspecified) normal distribution. This implies that she can attach subjective probabilities to different possible ranges of the total, consistent with the density of a normal curve. Note that the model captures the structure of what the sociologist believes. It is not a model of the data-generation mechanism. So, why should anyone care? Why should representation of the beliefs of a single sociologist be of interest to the larger scientific community? Scientists will care insofar as the model is consistent with the existing empirical and theoretical knowledge.

For example, suppose at one extreme, the sociologist is unable to justify her use of the normal with any kind of external information. She appeals solely to some form of introspection inaccessible to anyone else. Then, there is indeed little reason to pay attention to her model.

Suppose instead that the sociologist explains that as far as the approximation, she is proceeding as if the count of the number of jobs lost or gained is affected independently by a large number of variables, with no small subset dominating. It is affected, for instance, by the sum total of many decisions made independently by employers across the nation. Her use of the normal is then justified by an appeal to the central limit theorem. This is clearly an improvement because the assertion about the role of a large number of causal variables is in principle subject to empirical scrutiny. Perhaps now a few other scientists will take notice.

Finally, at the other extreme, suppose there are several other studies of the impact of other trade agreements, perhaps based on the recent European experience, and some relevant theory that has credibility in the scientific community. With this information, she is now able to specify, in addition, how the mean of her normal model varies conditional upon the values of several explanatory variables. She has made the distinction between the role of a few variables that alter the mean and the role of the variables excluded from her model that affect the dispersion of the count around conditional means. Perhaps now she will get some respect. Moreover, she can specify a prior density for the model parameters, collect longitudinal data before and after NAFTA, and then generate posterior densities for both the model's parameters and the predictive density of the number of jobs gained or lost.

To summarize, if we are prepared to interpret a statistical model in subjective terms, we can do Bayesian inference for apparent populations. However, we would be among the first to admit that the models interpreted in subjective terms raise difficult issues. For example, should we equate our sociologist's subjective model with a scientific theory? Are not both statements of belief about how the empirical world works? While a discussion of these issues is well beyond the scope of this paper (and a thorough discussion well beyond our capabilities), it is clear that we would need to consider nothing less than the nature of scientific theory and the scientific enterprise itself. Provocative discussion of Bayesian inference in science can be found in books by Howson and Urbach (1989) and Earman (1992).

5. EMPIRICAL ILLUSTRATION

To illustrate the use of Bayesian inference for apparent populations, we analyze below the data used in a paper by Wallerstein (1989). This study of unionization in 20 advanced industrial democracies typifies the setting frequently encountered by macrosociological researchers. The regression model used is also representative. We stress, however, that we have chosen to work with Wallerstein's data and formulation because it captures well the kind of research found in major sociological journals. We take no stand one way or the other about the validity of the model or the quality of the data. The data are shown in Table 1.

Wallerstein's response variable is union density D, union members as a proportion of potential membership in a particular country. The mean of union density is taken to be a function of two explanatory variables: the representation of left parties in the government (Left) and the size of the civilian labor force (Size). The effect of the governing left parties is expected to be positive because labor and social democratic governments have historically assisted union organizing legislatively and through the sympathetic administration of labor law. On the other hand, the effect of labor force size is expected to be negative because larger labor forces are more expensive for unions to organize. Wallerstein also experiments with several control variables, but these add little to the fit of the data. Our Bayesian illustration is thus limited to a regression of union density on left parties in government and the size of the civilian labor force.

Building on Wallerstein's analysis, the regression model has a log-log form

$$\ln(D_i) = \beta_0 + \beta_1 \ln(\text{Left}_i) + \beta_2 \ln(\text{Size}_i) + u_i, \tag{16}$$

TABLE 1
Data for Union Density, Civilian Labor Force, and Index of Left Government

Country	Union Density (%)	Labor Force (1000's)	Left Government*
Australia	51.4	5436	33.75
Austria	65.6	2469	48.67
Belgium	71.9	3348	43.25
Britain	48.0	25757	43.67
Canada	31.2	10516	0.00
Denmark	69.8	2225	90.24
Finland	73.3	2034	59.33
France	28.2	18846	8.67
Germany	39.6	23003	35.33
Iceland	70.5	81	17.25
Ireland	68.1	886	0.00
Israel	80.0	997	73.17
Italy	50.6	15819	0.00
Japan	31.0	39930	1.92
Netherlands	37.7	4509	31.50
New Zealand	59.4	1050	60.00
Norway	58.9	1657	83.08
Sweden	82.4	3931	111.84
Switzerland	35.4	2460	11.87
United States	24.5	92899	0.00

Source: Wallerstein (1989).

*Left government is measured by Wilensky's cumulative index in which 1 year of left government in the period since World War I scores 3 points. One was added to each left government score before the log transformation.

where the regression coefficients are interpreted as elasticities; a 1 percent change in X_{ij} yields a β_j percent change in Y_i, and u_i is assumed to meet the usual ordinary least squares assumptions. In particular, u_i is taken to be normally distributed.[5]

Equation (16) can be interpreted as a model of the *subjective conditional density of D*. It is, therefore, a statement about what the researcher believes; it is a representation of his uncertainty about D. Whether others choose to take his beliefs seriously depends on the relationship between the model and what the others believe. Since

[5] A normal quantile-quantile plot of the residuals gave no cause for concern once all the variables were logged.

the relevant actors are scientists, all parties presumably will be affected by shared understandings (at least in general terms) about the role to be played by the current data, past research, and existing theory.

5.1. *Specifying the Priors*

Recall that the first step in a proper Bayesian analysis is to consider the information one has before looking at the data. In this instance, it is clear from Wallerstein's theoretical discussions that he had rather specific expectations about the likely range of values for the regression coefficients. Our reading of the major sociological journals suggests that this is more generally true; perhaps fearing the charge of mindless empiricism, most sociological scholars provide rather clear priors (i.e., the range of values in which the regression coefficients are likely to fall) as a part of the "theory section" of their publications.

Using the existing research literature on the determinants of union density (e.g., Stephens 1979; Korpi 1983; Western 1993), we can specify a joint prior density for the three regression parameters of interest (i.e., β_0, β_1, β_2).[6] It is convenient and reasonable to specify that the three have a joint normal distribution. Put another way, the convenience of the normal for expositional purposes is not outweighed by some substantive reason to represent our prior knowledge with another distribution. It is also convenient and reasonable to specify that all three covariances between the coefficients are zero. This implies, for example, that whether the impact of the representation of the Left party is larger or smaller than an initial guess conveys nothing about whether the impact of the size of the civilian labor force is larger or smaller.

[6] We will proceed as if the variance of the disturbances is known. For our purposes, the variance of the disturbances is a nuisance parameter; we have no substantive interest in its value. Moreover, treating the disturbances as known simplifies the exposition substantially. If the variance is not known, then the marginal distribution of each of the regression coefficients would have a t-distribution rather than a normal distribution. However, for a normal disturbance term and a sample size of 20 (df for the error is 17), the t-distribution and the normal are moderately similar. So, nothing important is being lost here, assuming that the disturbance distribution is known. Moreover, as discussed briefly in the appendix to this chapter, the model was also estimated allowing the variance to be unknown, and the results were substantially changed.

Turning first to the mean and variance of the constant β_0, other research provides very little information. Moreover, the constant is essentially irrelevant for Wallerstein's theoretical ideas. We put, therefore, a rather diffuse prior on the constant to reflect a great amount of uncertainty and very little interest; for Wallerstein, the constant is almost a nuisance parameter. Put another way, using a rather diffuse prior implies that the marginal posterior density for the constant will be heavily determined by the data on hand (via the likelihood function).

The first row in Table 2 provides the details. The value of 10 for the prior variance of β_0 is meant to reflect little prior information about the value of the constant. The precise value of 10 was taken to reflect our considerable uncertainty, a priori. It implies we are 95 percent certain that the value of the constant falls between plus and minus 6.3 (i.e., the prior mean of 0 plus or minus two standard deviations).

The mean of the prior density for the regression coefficient for Left is .2. This indicates that we think that *a priori on average,* for a country with twice the social democratic incumbency over another country, we might expect union density to be about 15 percent higher. The second row also shows that before looking at the data, we were 95 percent certain that the value of the regression coefficient for the Left falls between .1 and .3. This translates into a variance of .0025.

Finally, we are more uncertain about the impact of size. As the third row of Table 2 shows, the mean of the prior density for β_2 (the coefficient of labor force size) is −.1, and the standard deviation is .1. This implies that we are 95 percent certain that regression coefficient β_2 for the impact of size falls between −.3 and .1, an interval that contains zero. Thus our prior for β_2 gives moderate

TABLE 2
Prior Parameters for the Regression of Union Density on Left and Size

Parameters	Mean	Two Times Standard Error	Variance
Intercept	0	6.3	10
Left	.2	.1	.0025
Size	−.1	.2	.01

probability to both positive and negative values. On the average, however, we are asserting before looking at the data that a typical country would have a union density that is about 7 percent less than another country with half the labor force size.

5.2. *Results*

Table 3 provides the means, standard deviations, and 95 percent confidence regions for each of the three regression parameters under three conditions: (1) the usual OLS estimates, implying a flat prior with no prior information; (2) a diffuse prior constructed by multiplying prior covariance matrix described earlier by 5; and (3) the informative prior from Table 2. The first column, therefore, represents both computational business as usual as well as a Bayesian baseline in which prior information is effectively ignored. The third column represents a preferred approach drawing on our prior information. The second is an in-between case where prior information plays a modest role. In effect, we are providing a sensitivity analysis to explore how much our posterior distributions depend on the prior distributions.

The computations for the second and third columns are relatively simple. (See the appendix at the end of this chapter.) But software for Bayesian linear regression is now widely available on IBM-compatible computers. Perhaps more importantly, very general procedures for computing joint and marginal posterior densities,

TABLE 3
Posterior Means, Standard Errors, and 95 Percent Confidence Regions for Regressions of Union Density on Left Government and Labor Force Size, Advanced Industrial Democracies, 1980–1985

	OLS	Diffuse Prior	Informative Prior
Constant	4.849 (.371) [4.121,5.576]	4.757 (.360) [4.052,5.462]	4.519 (.326) [3.880,5.157]
Left	.079 (.036) [.009,.150]	.091 (.034) [.025,.158]	.123 (.029) [.066,.180]
Size	−.137 (.038) [−.211,−.064]	−.131 (.037) [−.202,−.059]	−.113 (.034) [−.180,−.046]

even when one does not have a formally appropriate (conjugate) prior, are now available in S-Plus and Xlispstat; software is really not a problem. For example, we have been using Bayesian routines in Xlispstat for logistic and Poisson regression and have written programs of our own for quite complicated models.

The story from the marginal posterior densities is relatively straightforward. To begin, all of the confidence regions in Table 3 are given in Bayesian interpretation. Thus, for the OLS results, we are 95 percent certrain that the regression coefficient for Left falls between .009 and .150, and that the regression coefficient for Size falls between −.211 and −.064. Since neither interval includes zero, we are also confident that we at least have the correct sign. Similar interpretations can be made for all of the 95 percent confidence regions.

However, it cannot be overemphasized that these results are *conditional on the model and the data on hand.* If the model is not credible, our intervals are not credible; all of the usual concerns about specification error apply. And there is nothing explicit in the analysis that speaks to what the intervals would be for other data; our results are conditional upon the data.

Moving across the table, we are able to see the impact of prior information. The mean of the marginal posterior density (and our "best guess" as a point estimate) for the coefficient of Left shifts upward from .079 to .123 (nearly doubling). Simultaneously, the 95 percent confidence region shrinks. If the magnitude of the Left coefficient is of substantial theoretical importance, the impact of our prior information is significant. The coefficient of Size, however, is far less affected by our information. The mean of our prior density was quite close to the OLS estimate based on a flat prior, so there was little "pull." In addition, the more dispersed prior implies that the prior, whatever its mean, would have less influence on the posterior density in any case. In summary, the data in this illustration lead to relatively modest revision of the priors we brought to the data.

5.3. *When Are Bayesian and Frequentist Conclusions Different?*

In our empirical illustration, the substantive conclusions from Bayesian and frequentist analyses would not differ dramatically. But sometimes they can. Perhaps the most common examples are found in studies using small samples when prior information can be very important. Different conclusions can also result from how hypothesis testing is undertaken by frequentists and Bayesians. The issues are well beyond the scope of this paper (see Lee 1989 pp. 123–43) and can vary depending on the *kind* of Bayesian or frequentist doing the analysis; but recall that frequentists build prior information into their null hypotheses while Bayesians build prior information into their prior densities. This can lead to rather different approaches to testing and different substantive conclusions. Such was the case in some of our recent work (Berk et al. 1992). A reasonable generalization may be that "slam dunk" empirical results will be recognized by Bayesians or frequentists alike. Difference can arise, however, when the weight of the evidence provided by the data is not compelling. And, it is our sense that much sociological research falls into this category.

6. SUMMARY AND CONCLUSIONS

This paper had two goals: to evaluate conventional analyses of data from apparent populations and to raise for consideration a Bayesian alternative. Our views on these issues are easily summarized.

To begin, statistical inference is difficult even under the best of circumstances. Demanding assumptions must be made that are hard to meet in practice. However, it is one thing to say that statistical inference is imperfect and quite another to say that it is useless. The decision about whether or not to proceed depends on what may be learned compared to the alternatives. Sometimes statistical inference will yield useful information and sometimes it will not. When it does, appropriate caveats must be introduced; when it does not, statistical inference should not be undertaken. For what it is worth, however, our reading of the sociological literature suggests that statistical inference is very commonly used when it obscures far more than it enlightens.

For apparent populations, there have been two common approaches to uncertainty. On the one hand, it is possible to treat an apparent population as a true population and dispense with statistical inference altogether. One proceeds, then, as if the data could not have been different; historical/macro processes are taken to be deterministic. On the other hand, one may treat an apparent population as a realization from some super population. An apparent population becomes a random sample. One then has recourse to all of the existing frequentist apparatus. We have argued at length that both approaches are not fully satisfying.

As a very different inferential approach, we have raised for consideration the use of Bayesian inference when one's data derive from an apparent population. The form of Bayesian inference required, however, treats the distribution of the data, conditional on the parameters, as a subjective statement about what the researcher believes; what is represented is the structure of the scientist's uncertainty. While this would seem to solve the problem of statistical inference for data that are not replicable, it introduces controversial and complex issues about the nature of the scientific enterprise, which we make no pretense of resolving.

All three alternatives have their strengths and weaknesses, but we believe that relying on the procedures that treat apparent populations as either true populations or random samples ultimately will not suffice. We see no prospect of solving the problem of inference for apparent populations. In contrast, we see promise in the Bayesian approach. However, in the solution we are proposing there are a number of new conceptual problems that need further study. These include justification of the model (and implicitly the likelihood function) and justification of the prior density, although model specification is hardly a Bayesian problem alone.

What then are the implications for practice? We believe that the best way for the sociological community to explore the possibilities of Bayesian inference is to begin applying it. There is really no substitute for hands-on experience with real problems and real data. Philosophic discourse certainly has its place, but the true test is whether Bayesian inference improves the science. That can be determined only by experience. The principal danger is that a new Bayesian ritual will simply be substituted for an old frequentist ritual. Bayesian inference must be employed in a careful, self-conscious, and skeptical manner.

APPENDIX A: SOME TECHNICAL DETAILS

The usual linear regression model can be written as:

$$y = X\beta + u,\ u \sim N(0, \sigma^2 \mathrm{I}), \tag{17}$$

where y is an $N \times 1$ vector of observations on the response variable, X is an $N \times k$ matrix of fixed observations on the explanatory variables, β is a $k \times 1$ vector of regression coefficients, and u is an N by 1 vector of disturbances. Within a Bayesian formulation, uncertainty is attached to the regression parameters β and σ^2. We can proceed as if β has a multivariate normal prior distribution with a vector of means b* and covariance matrix V*. That is

$$\beta \sim N(b^*, V^*). \tag{18}$$

If σ^2 is either known or estimated from the sample data and then treated as known, the posterior distribution of β is normal (e.g., see Zellner 1971) so that

$$\beta \sim N(b^{**}, V^{**}), \tag{19}$$

where

$$V^{**} = (V^{*^{-1}} + s^{-2}X^TX)^{-1}, \; b^{**} = V^{**}(V^{*^{-1}}b^* + s^{-2}X^Ty). \tag{20}$$

If uncertainty in σ^2 is taken into account, the posterior distribution of the regression coefficients has a multivariate t distribution and the expressions for the mean and variance are a bit more complicated (e.g., see Zellner 1971). In either case, however, the necessary computations for the posterior distribution's mean vector and covariance matrix are easily programmed with any of several statistical programming languages: Proc Matrix in SAS, GAUSS, and S-Plus. We used S-Plus.

Far more general approaches are also available in S-Plus and Xlispstat. Xlispstat, which may be obtained at no cost, runs well on Apple Macintoshes, UNIX workstations, and a variety of mainframes. There is also a version that runs under DOS on a PC, but it is limited to small datasets. We understand that a more powerful version of Xlispstat will soon be available for DOS machines. In any case, the senior author is delighted with the generality and power of Xlispstat and has now used it for a variety of problems. For example, it was easy to reestimate the models reported in Table 3 with the variance treated as unknown.

REFERENCES

Abrahamson, M., and M. A. DuBick. 1977. "Patterns of Urban Dominance in the U.S. in 1890." *American Sociological Review* 42:756–68.

Allenby, G. M. 1990. "Cross-validation, the Bayes Theorem, and Small-Sample Bias." *Journal of Business and Economic Statistics* 8(2):171–78.

Barnett, V. 1982. *Comparative Statistical Inference.* New York: Wiley.

Begg, C. 1990. "On Inferences from Wei's Biased Coin Design for Clinical Trials (With Discussion)." *Biometrika* 77: 463–84.

Berger, J. O., and R. L. Wolpert. 1984. *The Likelihood Principle,* 2nd ed. Hayward, Calif.: Institute of Mathematical Statistics.

Berk, R. A. 1988. "Causal Inference for Sociological Data." In *Handbook of Sociology,* edited by Neil Smelser. Beverly Hills: Sage Publications.

Berk, R. A., and M. Brewer. 1978. "Feet in Clay in Hobnail Boots: An Assessment of Statistical Inference in Applied Research. In *Evaluation Studies Review Annual,* vol. 3, edited by Thomas Cook. Beverly Hills: Sage Publications.

Berk, R. A., A. Campbell, R. Klap, and B. Western. 1992. "The Differential Deterrent Effects of an Arrest in Incidents of Domestic Violence: A Bayesian Analysis of Four Randomized Field Experiments." *American Sociological Review* 57(5):698–708.

Berk, R. A., and P. H. Rossi. 1990. *Thinking About Program Evaluation.* Newbury Park, Calif.: Sage Publications.

Bernardo, J. M., and A. F. M. Smith. 1994. *Bayesian Inference.* New York: Wiley.

Bollen, K. A., and R. W. Jackman. 1985. "Political Democracy and the Size Distribution of Income. *American Sociological Review* 50:438–57.

Box, G. E. P. 1980. "Sampling and Bayes Inference in Scientific Modeling and Robustness (with discussion)." *Journal of the Royal Statistical Society (Series A) 143:383–430.*

Box, G. E. P., and G. C. Tiao. 1973. *Bayesian Inference in Statistical Analysis.* Reading, Mass.: Addison-Wesley.

Broemeling, L. D. 1985. *Bayesian Analysis of Linear Models.* New York: Marcel Dekker.

Cochran, W. G. 1977. *Sampling Techniques.* New York: Wiley.

Dahl, R. A., and E. R. Tufte. 1973. *Size and Democracy.* Stanford: Stanford University Press.

Diaconis, P., and D. Freedman. 1986. "On the Consistency of Bayes Estimates." *The Annals of Statistics* 14(1):1–87.

Draper, D. 1993. "Assessment and Propagation of Model Uncertainty (with discussion)." *Journal of the Royal Statistical Society.* (Series B) 57: 45–89.

Draper, D., J. S. Hodges, C. L. Mallows, and D. Pregibon. 1993. "Exchangeability and Data Analysis" (with discussion). *Journal of the Royal Statistical Society,* Series A, 156, Part 1:9–37.

Earman, J. 1992. *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory.* Cambridge: MIT Press.

Edington, E. S. 1987. Randomization Tests. New York: Marcel Dekker.

Ehrlich, I. 1977. "Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence." *Journal of Political Economy* 85(August):741.

Eisberg, R. M., and I. S. Lerner. 1981. *Physics: Foundations and Applications.* New York: McGraw Hill.

Freedman, D. A. 1991. "Statistical Models and Shoe Leather." In *Sociological Methodology 1991,* edited by Peter V. Marsden, 291–313. Washington: American Sociological Association.

Freedman, D. A., and D. Lane. 1983. "Significance Testing in a Nonstochastic Setting." In *A Festschrift for Eric L. Lehman,* 185–208. Belmont, Calif.: Wadsworth.

Gelfand, A. E., and A. F. M. Smith. 1990. Sampling-Based Approaches to Calculating Marginal Densities." *Journal of American Statistical Association* 85:398–409.

Griffin, L.J., H. J. McCammon, and C. Botsko. 1990. "The Unmaking of a Movement? The Crisis of U.S. Trade Unions in a Comparative Perspective." In *Change in Social Institutions,* edited by M. Hallinan, D. Klein and J. Glass, 169–94. New York: Plenum.

Groves, R. M. 1989. *Survey Errors and Survey Costs.* New York: Wiley.

Howson, C., and P. Urbach. 1989. *Scientific Reasoning: The Bayesian Approach.* LaSalle, Ill.: Open Court.

Jeffreys, H. 1961. *Theory of Probability,* 3rd ed. Oxford: Clarendon.

Jenkins, J. C., and C. Schock. 1992. "Global Structures and Political Processes in the Study of Domestic Conflict." *American Sociological Review* 18:161–85.

Johnson, N. L., and S. Kotz. 1969. *Discrete Distributions.* New York: Wiley.

Korpi, W. 1983. *The Democratic Class Struggle.* London: Routledge.

Lange, P., and G. Garret. 1985. "The Politics of Growth: Strategic Interaction and Economic Performance in Advanced Industrialized Democracies, 1974–1980." *Journal of Politics* 47:792–827.

Larson, H. J. 1969. *Introduction to Probability Theory and Statistical Inference.* New York: Wiley.

Leamer, E. E. 1978. *Specification Searches.* New York: Wiley.

Lee, P. M. 1989. *Bayesian Statistics: An Introduction.* New York: Oxford University Press.

Lincoln, J. 1978. "The Urban Distribution of Headquarters and Branch Plants in Manufacturing: Mechanisms of Metropolitan Dominance. *Demography* 15:213–24.

Lipset, S. M. 1959. "Some Social Requisites of Democracy: Economic Development and Political Legitimacy." *American Political Science Review* 53: 69–105.

Little, R. J. A., and D. B. Rubin. 1987 *Statistical Analysis with Missing Data.* New York: Wiley.

Mason, W. M., G. Y. Wong, and B. Entwisle. 1983. "Contextual Analysis Through the Multilevel Linear Model." In *Sociological Methodology, 1983–1984,* edited by S. Leinhardt. San Francisco: Jossey-Bass.

Morrison, D. E., and R. E. Henkel. 1970. *The Significance Tests Controversy.* London: Butterworth.

Noreen, E. W. 1989. *Computer Intensive Methods for Testing Hypotheses.* New York: Wiley.

Oaks, M. 1986. *Statistical Inference. A Commentary for the Social and Behavioral Sciences.* New York: Wiley.

O'Hagan, A. 1994. *Kendall's Advanced Theory of Statistics, vol 2B, Bayesian Theory.* New York: Halstead Press.

Olzak, S. 1989. "Labor Unrest, Immigration and Ethnic Conflict in Urban America, 1880–1914." *American Journal of Sociology* 14(6):1303–33.

Pettit, L. I. 1986. "Diagnostics in Bayesian Model Choice." *The Statistician* 35:183–90.

Phillips, D. P. 1974. "The Influence of Suggestion on Suicide: Substantive and Theoretical Implications of the Weather Effect." *American Sociological Review* 39:340–54.

Pollard, W. E. 1986. *Bayesian Statistics for Evaluation Research.* Beverly Hills: Sage Publications.

Press, S. J. 1989. *Bayesian Statistics.* New York: Wiley.

Przeworski. A. and J. Sprague. 1986. *Paper Stones: A History of Electoral Socialism.* Chicago: University of Chicago Press.

Ross, C. O. 1987. "Organizational Dimensions of Metropolitan Dominance: Prominence in the Network of Corporate Control, 1955–1975." *American Sociological Review* 52:258–67.

Rossi, P. H., R. A. Berk, and B. K. Edison. 1974. *The Roots of Urban Discontent.* New York: Academic Press.

Rubin, D. B. 1976. "Inference in Missing Data." *Biometrika* 63:593–604.

Rueschmeyer, D., E. H. Stephens, and J. D. Stephens. 1992. *Capitalist Development and Democracy.* Chicago: University of Chicago Press.

Smith, T. M. F., and R. A. Sugden. 1988. "Sampling and Assignment Mechanisms in Experiments, Surveys, and Observational Studies." *International Statistical Review* 56:165–80.

Speed, T. 1993. "Modeling and Managing a Salmon Population." In *Statistics for the Environment,* edited by V. Barnett and K. F. Turkman. New York: Wiley.

Stepan-Norris, J., and M. Zeitlin. 1989. " 'Who Gets the Bird?' or How the Communists Won Power and Trust in America's Unions." *American Sociological Review* 54:503–23.

Stephens, J. D. 1979. *The Transition from Capitalism to Socialism.* London: MacMillan.

Tierney, L., and J. B. Kadane. 1986. "Approximate Marginal Densities for Nonlinear Functions." *Journal of the American Statistical Association* 81:82–86.

Wallace, M., and A. L. Kalleberg. 1981. "Economic Organization of Firms and Labor Market Consequences: Toward a Specification of Dual Labor Market Theory." In *Sociological Perspectives on Labor Markets,* edited by Ivar Berg. New York: Academic Press.

Wallerstein, M. 1989. "Union Organization in Advanced Industrial Democracies." *American Political Science Review* 83:483–501.

Weiss, R. E., and R. D. Cook. 1992. "A Graphical Case Statistic for Assessing Posterior Influence." *Biometrika* 79(1):51–55.

Western, B. A. 1991. "A Comparative Study of Corporatist Development." *American Sociological Review* 56:283–94.

Western, B.A. 1993. "Postwar Unionization in Eighteen Advanced Capitalist Countries." *American Sociological Review* 58:26–82.

Wilensky, H. L. 1985. *The Welfare State and Equality.* Berkeley: University of California Press.

Wonnacott, T. H., and R. J. Wonnacott. 1977. *Introductory Statistics.* New York: Wiley.

Wright, J. D., P. H. Rossi, S. R. Wright, and D. Weber-Budin. 1979. *After the Clean-Up.* Beverly Hills: Sage Publications.

Young, P. C., and M. Lee. 1993. "The Active Mixing Volume: A New Concept in Modelling Environmental Systems." In *Statistics for the Environment,* edited by V. Barnett and K. F. Turkman. New York: Wiley.

Zellner, A. 1971. *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley.