Bootstrap Statistical Inference: Examples and Evaluations for Political Science
Author(s): Christopher Z. Mooney
Source: *American Journal of Political Science*, Vol. 40, No. 2 (May, 1996), pp. 570-602
Published by: Midwest Political Science Association
Stable URL: https://www.jstor.org/stable/2111639
Accessed: 07-01-2020 22:57 UTC

*Bootstrap Statistical Inference: Examples and Evaluations for Political Science**

Christopher Z. Mooney, *West Virginia University*

Theory: Bootstrapping is a nonparametric approach to statistical inference that relies on large amounts of computation rather than mathematical analysis and distributional assumptions of traditional parametric inference. It has been shown to provide asymptotically accurate inferences for a wide variety of statistics.
Hypothesis: Bootstrapping may make more accurate inferences than the parametric approach under two general circumstances: 1) when the assumptions of parametric inference are not tenable, and 2) when no parametric alternative exists for a problem.
Methods: Monte Carlo simulation is used to test the performance of bootstrap and parametric confidence intervals for both types of situations in which bootstrapping is hypothesized to be superior to parametric inference. A single data example is used to illustrate the use of the bootstrap: a seats/votes model of U.S. House elections from 1932 to 1988.
Results: My central conclusions are that in the cases examined: 1) bootstrap confidence intervals are at least as good as the parametric confidence interval and sometimes better, 2) OLS parametric confidence intervals do not perform too badly when the model error is non-normal, especially as sample size increases, and 3) when no parametric alternative exists, the bootstrap provides a reasonable method of making statistical inferences.

I. Introduction

Bootstrapping is a relatively new, nonparametric approach to statistical inference that relies on large amounts of computation rather than on the mathematical analysis and distributional assumptions of traditional parametric inference. Parametric inference requires the *a priori* assumption of a standard functional form for the sampling distribution of a statistic. This is necessary to reduce the complexity of the problem, making the mathematics of the procedure tractable. For example, if a statistic is assumed to be

*I would like to thank Robert D. Duval, Bruce Worton, Hugh Ward, George A. Krause and three reviewers for this journal for their comments on this project, and Burak Saltoglu for research assistance. Financial support was provided by the West Virginia University Faculty Senate, the West Virginia University Water Resources Institute, the University of Essex Research Promotion Fund and the British Academy.[1]

[1]All data and documentation necessary to replicate the analyses contained herein can be obtained from the author. The data analyses were conducted using the author's own GAUSS code, RATS programs incorporating the BOOT command, and the LIMDEP CRMODEL command with the HET subcommand.

American Journal of Political Science, Vol. 40, No. 2, May 1996, Pp. 570–602

normally distributed, only two parameters (the mean and the standard deviation) are needed to characterize completely its sampling distribution. Bootstrapping constructs an estimate of a statistic's sampling distribution empirically, using within-sample variation. This is done using a simple, but computationally intensive, re-sampling procedure. While the effectiveness of this sort of experimental approach has been understood in the statistical community for at least 100 years, it has only recently been developed in detail because it has become computationally feasible due to advances in computer technology.[2]

The promise of the bootstrap is that it may free researchers from the reliance on the limited number of statistical models for which parametric inference works. The familiar *z*- and *t*-tests, exemplars of the parametric approach, were developed in an era of low technology where a 10-key mechanical calculator in the hands of an experienced practitioner typified the state of the art in calculation speed (Mooney and Krause 1996). In this environment, it made sense to settle for sometimes tenuous parametric assumptions and a limited number of models in order to proceed with analysis. But with the high speed PC's and workstations available today, we are due to strike a new compromise between computational effort and the reliance on distributional assumptions (Tukey 1986, 73). The bootstrap has been touted as a major component of this new compromise (Efron 1979b).

The potential advantages of the bootstrap are likely to be greatest for political scientists in two situations: where there exists no or only weak statistical theory about the distribution of a statistic, and where the distributional assumptions necessary for valid parametric inference are violated. If its inferences are accurate, the bootstrap will be exceptionally useful in the former case, as parametric inference offers no alternative. Examples of such estimators of interest to political scientists to which the bootstrap has already been applied include indirect effects of path models (Bollen and Stine 1991), the switch point in a switching regression (Douglas 1987), eigenvalues (Lambert, Wildt, and Durand 1990), and the difference between two sample medians (Mooney and Duval 1993, 50–4).

The other potentially important application of the bootstrap for political scientists is in those situations where parametric inference can be conducted mathematically, but where violations of key distributional assumptions occur. Parametric inference always requires distributional assumptions regarding the tested statistics, which are more or less realistic depending on the specific situation. As a familiar example, parametric inference using

[2]While bootstrapping takes the frequentist approach to inference much further than does parametric inference, the philosophical conception of inference is the same. This conception is quite different, for example, from that of Bayesian inference (Oakes 1986, chap. 6).

OLS regression coefficients requires the assumption of normality for the model's error term if the sample is small, but a non-normal dependent variable can cause this assumption to be violated. For example, Lijphart and Crepaz (1991, 242) face such a situation in modeling the influences on the level of corporatism in 18 industrialized democracies. Likewise, parametric inference for Pearson's correlation coefficient requires a bivariate normality assumption that is often violated (Fisher 1915). This situation arises when Songer and Haire (1992, 974) correlate appeals court judges' decisions with potential influences on them. Many parametric tests are thought to be robust to these sorts of violations, and so researchers often overlook them (Achen 1982, 36–7; Malinvaud 1970, 297–9). Violations of these assumptions, however, can lead to serious and undetectable problems with these tests' error rates (Geary 1947; Pearson 1931; Ghurye 1949).

When confronted with these problematic inferential situations, the reliance on parametric inference has forced political scientists to adopt a variety of strategies. They have transformed their data in various ways; they have resorted to the use of less than optimal estimators whose inferential properties are better known; they have used ad hoc nonparametric tests; and they have even ignored inferential testing altogether, relying solely on point estimation. Each of these strategies has its own idiosyncratic drawbacks, and they all suffer from the general problem of resorting to ad hocery. The bootstrap is a single, general approach to inference that offers a solution to a wide range of statistical problems, and as such may be preferred to the piecemeal approach taken heretofore.

While the bootstrap has begun to receive regular usage in some disciplines, notably economics, biology, and medicine, it has not often been used in political science up until now, probably for at least two reasons. First, this approach to statistical inference is completely foreign to most political scientists. Several generations of researchers have been trained exclusively in z- and t-tests, raising parametric inference and the Gaussian assumptions usually associated with it to the level of an unquestioned paradigm (Tukey 1975, 352). Paradigm changes come with difficulty, even if the benefits of such changes are obvious. One goal of this article is to address this problem by offering a brief and intuitive introduction to bootstrap inference, along with references to more detailed explications.

A second reason why the bootstrap has not been widely used in political science may be that its benefits are not as yet obvious. For example, how and when does the bootstrap outperform traditional parametric inference? When can it be fruitfully applied in political science? Without clear answers to these questions, political scientists will not likely invest the significant time and psychological resources required to incorporate the bootstrap into their methodological repertoires. While a great deal of theoretical explora-

tion of the bootstrap has occurred in the decade and a half since Efron (1979a) developed it, little systematic empirical comparison between bootstrap and parametric inference has been conducted in a context accessible and relevant to political scientists. Therefore, a second goal of this article is to address this problem by conducting such a comparison on two common sorts of political science problems.

At the outset of this article, it is important to specify why a researcher would prefer one inference technique to another, that is, the criteria for inferential performance. Where point estimators can be assessed on the well-known criteria of bias, efficiency and consistency, inference techniques are best judged on 1) how accurately they reproduce the nominal Type I error rate (α) of a problem, and 2) the level of Type II error that they commit in a problem. In this article, I will focus on the accuracy of a technique in the sense of point 1, following the traditional social science emphasis on controlling Type I error. I discuss how this criterion can be assessed in the Appendix describing my Monte Carlo experiments. The reason that this will be the standard of comparison is that when a researcher selects an α-level for a statistical test, he or she accepts a certain probability that a Type I error will be made, trading off for an unknown[3] but inversely related level of Type II error. If the actual α-level of a test is higher than the nominal level, more Type I error than is acceptable will occur. That is, the researcher will reject the true null hypothesis more often than he or she thinks is acceptable. On the other hand, if the actual α-level is lower than the nominal α-level, more Type II errors than necessary will be made, meaning that the researcher will fail to reject more false nulls than required given the nominal α-level. The point is that a test ought to have the α-level that is specified for it at the outset.

For pedagogical purposes, I use a single analytic example that exhibits both of the situations in which the bootstrap may be advantageous—a seats/votes model to test the hypothesis of bias in first-past-the-post electoral systems, which was proposed in the early 20th century as the "cube law" (Kendall and Stuart 1950; Schrodt 1982; Taagepera 1986). This is a good example for my purposes in that it exhibits both an OLS regression with non-normal error (Linehan and Schrodt 1978) and the development of an electoral system bias estimator whose distributional properties are unknown (Jackman 1994). I will evaluate these situations using both Monte Carlo simulation, and analyses of U.S. House of Representatives election data for 1932–88.

[3]The level of Type II error is generally unknown unless a power analysis is carried out.

II. Bootstrap Statistical Inference

At root, bootstrap and parametric inference pursue the same strategy: to estimate the sampling distribution of an estimator, $\hat{\theta}$, in order to make probability statements about the value of the population parameter, θ.[4] The parametric approach is to estimate this sampling distribution by making the assumption, based on statistical theory, that its functional form is characterized by a small number of parameters (e.g., two parameters in the case of the normal distribution). Analytic formulae are then used to estimate the parameters of this distribution from sample data (Mohr 1990, 13–28). Well-known procedures for performing hypothesis tests or constructing confidence intervals are then applied to make inferences to θ using this estimated sampling distribution of $\hat{\theta}$. For some statistics, such as the sample mean, this approach is quite adequate in practical research situations because there is strong statistical theory upon which to base the distributional assumption about $\hat{\theta}$. In many situations such assumptions are untenable, however, and/or it is difficult or impossible to calculate the parameters of the assumed distribution analytically from the data.

Bootstrapping, in contrast, involves making inferences about θ based on the variation of $\hat{\theta}$ within the sample itself. As with parametric inference, there are two steps in this process: 1) estimate the sampling distribution of $\hat{\theta}$, and 2) use this estimated sampling distribution to develop confidence intervals around θ. To undertake the first step, recall that the sampling distribution of $\hat{\theta}$ can be thought of as the relative frequency distribution of the values of $\hat{\theta}$ from an infinite number of samples of size n drawn from a given population.[5] In conducting bootstrap inference, the researcher simulates such a process by randomly drawing a large number of "re-samples" of size n *with replacement* from the original sample, and then constructing a relative frequency distribution of the resulting values of the statistic, $\hat{\theta}^*$, calculated from each of these re-samples (Efron 1979a, 2–3; Hinckley 1988, 322–4; Efron and Tibshirani 1986, 54–5; Mooney and Duval 1993, 10–1). Because resampling is carried out with replacement, each of the re-samples will be slightly and randomly different from the original sample. This is because some of the elements of the original sample may be represented more than once in a re-sample, and some not at all. Therefore, the resulting $\hat{\theta}^*$'s will be slightly and randomly different from one another.

[4]Several primers on the bootstrap exist for those unfamiliar with the technique, e.g., Efron and Tibshirani (1993), Mooney and Duval (1993), and Stine (1990). Here I offer a brief introduction to highlight the key points of the approach. Those who wish to use the bootstrap are advised to consult one of these primers for a more comprehensive treatment.

[5]This again demonstrates the frequentist orientation to inference shared by bootstrapping and the parametric approach, as noted in footnote 2.

The relative frequency distribution of these $\hat{\theta}$*'s is, then, the bootstrap estimate of the sampling distribution of $\hat{\theta}$, denoted as $\hat{f}^*(\hat{\theta}^*)$.

The justification for this procedure rests on the analogies of the sample empirical density function (EDF) with the population density function that generated the data, and the random re-sampling mechanism with the random component of that process (Bickel and Freedman 1981; Singh 1981). The EDF is the nonparametric maximum likelihood estimate of the unknown distribution, $f(X)$, where X is the random component of the model (Rao 1987, 162–6; Rohatgi 1984, 234–6). That is, given that there is no other information about the population, the sample is the single best estimate of the population.[6] Therefore, the sample is treated as the population, with the re-samples being analogous to a series of independent random samples, conditional on $f(X)$. These conditionally independent samples are then used to construct an empirical estimate of $f(\hat{\theta})$, just as Monte Carlo simulation can be used when we have knowledge of the population (Mooney N.d., sec. 1.1).

Two practical questions arise from this description. First: How many re-samples are needed? Typically, the number of re-samples (B) should be 50–200 to estimate the standard error of $\hat{\theta}$, and perhaps 1,000 to construct confidence intervals (Efron and Tibshirani 1986, sec. 9). This is clearly a vast increase in computation over that required for parametric inference. But in practice estimating models of the type most political scientists work with, it usually takes only a matter of minutes to conduct the bootstrap with modern personal computers.[7] The second practical question raised here is: Exactly which quantity is re-sampled? Conceptually, the observed random component of the model should be re-sampled. The determination of which quantity is the random component for a given process requires a thorough understanding of both the model and the data at hand. For example, the random component is the raw variable in the case of a sample mean, the residuals in a classic regression model, and the cases of data for regression with survey data where the independent variables are themselves stochastic (Freedman 1981; Mooney and Duval 1993, 15–6).

Special consideration on this point needs to be given to data which is serially dependent, such as is often the case in a time series. With data of this sort, random re-sampling destroys the integrity of the dependence,

[6]If a researcher has other information about the population, such as empirical or theoretical evidence about its functional form or the value of its parameters, he or she can incorporate them into his or her analysis, as is done in Bayesian and parametric inferential statistics. I am concerned here, however, with cases where no such reliable information exists.

[7]When using the double-bootstrap to get a percentile-*t* confidence interval (described below), the running time can increase to upwards of an hour in some cases (see Table 3). Running time depends heavily on software efficiency and hardware speed.

which is often crucial to understanding the process involved. There are two general strategies to bootstrapping such data. First, one can model the dependence, resulting in random error that can be re-sampled as described above (Efron and Tibshirani 1993, 92–9). For example, an AR(1) corrected regression model could be bootstrapped by re-sampling the residuals if they were found not to be serially dependent (as in my example below). The second approach, known as moving block re-sampling, involves re-sampling groups of cases at a time, and concatenating them into a re-sample (Politis and Romano 1992; Liu and Singh 1992). This approach allows some of the dependence structure to be retained in the data, but still generates random variation in the re-sample. A key issue here is the trade-off between efficiency in creating this random variation and the amount of dependence retained in the data. This trade-off is reflected in the choice of the size of the blocks to be re-sampled.

The second step in making a statistical inference involves using the estimated sampling distribution to make probability statements about θ. In a traditional parametric confidence interval, we do this by selecting the $100*(\alpha/2)$ and $100*(1 - \alpha/2)$ percentile points from a standardized sampling distribution (e.g., using the calculated probabilities in the Student's *t* table), where α is the acceptable probability of Type I error. These standardized scores are then multiplied by the analytically estimated standard error of the sampling distribution and added to the point estimate to arrive at the upper and lower endpoints of an α-level confidence interval.

In bootstrap inference, there are several approaches to developing confidence intervals using $\hat{f}^*(\hat{\theta}^*)$.[8] In this article, I describe and evaluate four of these methods: the normal approximation, the percentile, the bias-corrected accelerated percentile (BCa), and the percentile-*t* methods.[9] I discuss these because the latter two methods appear now to be the consensus choice in the statistical community for the most accurate intervals, while the former two appear to be the most frequently used in practice. As a guide to programming, see the GAUSS code for executing each of these confidence intervals in Mooney (1994), the RATS code in Mooney and Duval (1993, 63–4), and the SAS code in Fan and Jacoby (1995).

The *normal approximation method* of constructing bootstrap confidence intervals is quite analogous to the parametric confidence interval ap-

[8]Most work on bootstrap inference has been in terms of confidence intervals, as opposed to hypothesis testing. Given the greater information contained in the former and the greater simplicity of developing them for the bootstrap, I follow this emphasis. For discussions of bootstrap hypothesis testing, see Fisher and Hall (1990) and Bollen and Stine (1994).

[9]See Mooney and Duval (1993, sec. 2.2), and Efron and Tibshirani (1993) for more thorough descriptions of these techniques.

proach (Noreen 1989, 69). When it is plausible to assume that a statistic is normally distributed, but no analytic formula exists to estimate its standard error, the bootstrapped sampling distribution can be used to estimate that standard error. This procedure is a straightforward application of the notion that the $\hat{\theta}^*$'s are random variables distributed as $\hat{\theta}$:

$$\hat{\sigma}^*_{\hat{\theta}} = \sqrt{\frac{\Sigma(\hat{\theta}^*_b - \hat{\theta}^*_{(\cdot)})}{B - 1}},$$

where

$$\hat{\theta}^*_{(\cdot)} = \frac{\Sigma\hat{\theta}^*_b}{B}, \tag{1}$$

and where B = the number of re-samples, and $\hat{\theta}^*_b = \hat{\theta}^*$ from re-sample b. Efron (1981) shows that as $B \to \infty$, $\hat{\sigma}^*_{\hat{\theta}} \to \hat{\sigma}_{\hat{\theta}}$, but little improvement in the approximation occurs as B exceeds 50–200 (Efron and Tibshirani 1986, sec. 9). To construct a confidence interval, then, the bootstrap estimate of the standard error is simply plugged into the well-known parametric confidence interval formula, using z- or t-scores as appropriate.

The normal approximation method is therefore very similar to parametric inference and, along with its analogous hypothesis test, it appears to be the most frequently used bootstrap method in the social sciences (e.g., King 1991; Green and Krasno 1990; Poole and Rosenthal 1991). This approach still relies on the parametric assumption of $\hat{\theta}$'s normality, however, and so it is seriously flawed. If $\hat{\theta}$ is not normally distributed, we are actually looking up the standardized percentile points in the wrong table. For example, the .025 percentile point of a standardized χ^2 distribution is not the same as that from the t distribution, even if these distributions have the same mean and standard deviation.

The *percentile method* takes literally the notion that $\hat{f}^*(\hat{\theta}^*)$ estimates $f(\hat{\theta})$, and does not require that it has a standard distribution. The basic approach is straightforward: an α-level confidence interval includes all the values of $\hat{\theta}^*$ from the $100^*(\alpha/2)$ to the $100^*(1 - \alpha/2)$ percentiles of the $\hat{f}^*(\hat{\theta}^*)$ distribution (Efron 1982, chap. 10; Stine 1990, 249–50). That is, the endpoints of a 95% confidence interval for θ would be the values of $\hat{\theta}^*$ at the 2.5th and 97.5th percentiles of $\hat{f}^*(\hat{\theta}^*)$. This interval can be developed from a sorted vector of $\hat{\theta}^*$'s in a straightforward manner. This ease of use probably accounts for the percentile method being the most popular bootstrap approach among applied statisticians (Hall 1992, 36). The accuracy of the percentile method has long been seriously questioned, however,

when $f(\hat{\theta})$ is skewed or $\hat{\theta}$ is a biased estimator (Efron 1982, sec. 10.7; Schenker 1985).

The *bias-corrected accelerated (BCa) percentile method* was developed by Efron (1987) to address the problems with the percentile approach arising from skewed $\hat{f}^*(\hat{\theta}^*)$'s and biased $\hat{\theta}$'s, but it is also perfectly applicable for unbiased $\hat{\theta}$'s and symmetric $\hat{f}^*(\hat{\theta}^*)$'s. This interval involves the application of an algebraic adjustment to the percentile points the analyst selects from $\hat{f}^*(\hat{\theta}^*)$ to serve as the confidence interval endpoints. While this procedure is too complex to describe in full here, it is relatively straightforward to program and apply. See Efron and Tibshirani (1993, chap. 14) for a detailed and practical explanation.

An alternative method of accounting for the skew in $f(\hat{\theta})$ which approaches the problem from a completely different angle is the *percentile-t method* (Hall 1992; Bickel and Freedman 1981; Efron 1981, sec. 9). Here we transform each $\hat{\theta}^*$ into a standardized variable, t^*:

$$t_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\hat{\sigma}_{\hat{\theta}b}} \tag{2}$$

where $\hat{\sigma}_{\hat{\theta}b}$ is calculated for each $\hat{\theta}_b^*$ from its corresponding re-sample, and $\hat{\theta}$ is calculated from the full sample. The estimation of $\hat{\sigma}_{\hat{\theta}b}$ may be done analytically if a formula exists for it, or it may be calculated by conducting another round of bootstrapping (Equation 1). This "double-bootstrap" therefore involves another level of re-sampling, so that one is actually re-sampling from a re-sample. This can greatly increase the computations involved, which is the central drawback of the percentile-*t*.

The resulting t_b^*'s are distributed as $\hat{\theta}$, but on a standardized scale. In a way completely analogous to the use of the Student's t distribution in parametric inference, we select the $100*(\alpha/2)$ and $100*(1 - \alpha/2)$ percentile values of t_b^*, and develop a confidence interval around θ as follows:

$$p(\hat{\theta} - t_{\alpha/2}^*\hat{\sigma}_{\hat{\theta}} < \theta < \hat{\theta} - t_{1-\alpha/2}^*\hat{\sigma}_{\hat{\theta}}) = 1 - \alpha \tag{3}$$

using $\hat{\sigma}_{\hat{\theta}}$ calculated from the original sample, whether analytically or through bootstrapping.

Given this variety of bootstrap confidence interval methods, under which conditions do they perform differentially from one another and from a parametric confidence interval? The theoretical literature on this question has been confined by necessity to specific estimators and based largely on the consistency of the raw bootstrap sampling distribution estimate, and its adjustment for the BCa and percentile-*t* intervals. Over the decade and a half since its first development by Efron, the bootstrap sampling distribution

has been found to be consistent for a variety of estimators (e.g., Bickel and Freedman 1981; Freedman 1981; Thombs and Schucany 1990). But of greater concern for inferential practice in political science is the level of Type I and II error an inference approach commits in the long run. Especially we must ask: How closely does the actual Type I error rate of an inference technique approximate the nominal α-level set by the researcher? Without a close approximation, confidence intervals will be too wide or narrow, and the null hypothesis will be incorrectly rejected or not rejected more often than is necessary. Hall (1992) summarizes previous work on this issue and shows explicitly that for some commonly used estimators, the percentile-*t* and the BCa are at least as accurate in this sense as the parametric approach, and often more so, but that the percentile and normal approximation approaches tend to be less accurate.[10]

The problem with these theoretical results is that they rely on asymptotic theory; they hold only as sample size increases to infinity. Their performance in the finite (and even small) samples that many political scientists work with can only be explored through the use of Monte Carlo simulation (Davidson and MacKinnon 1993, 768), and this is the approach taken in this paper (see Appendix). Given a plausible sample size and theoretical model on a data situation of interest to political scientists, how well do the bootstrap techniques perform relative to one another, and relative to the parametric approach? This is the question I turn to now.

III. Inference When Parametric Assumptions Are Violated

Parametric inference involves the use of mathematical analysis and distributional assumptions to make these mathematics tractable. For example, we can make inferences regarding Pearson's correlation coefficient if we are willing to assume bivariate normality for the variables. In deriving this result, Fisher (1915) made this distributional assumption not because it was plausible, but because it would be too difficult mathematically to derive the result without it. While the use of these distributional assumptions allowed early statisticians to proceed with analysis, the question of their plausibility has long been a sticking point (Brewer 1985; Micceri 1989). Bootstrapping allows us to make inferences by substituting statistical theory and computation for these assumptions.

By far the most common type of statistical analysis in political science

[10]The comparison of the Type II error rates of the bootstrap and parametric approaches has been less well-explored in the statistics literature, and following the social science tradition of being primarily concerned with Type I error, I leave evaluation on this criterion to future research.

is ordinary least squares regression (OLS).[11] A variety of well-known assumptions are required to hold true about the error term of a linear regression model in order for the OLS coefficient estimates to be normally distributed, efficient and unbiased (Berry 1993). For example, if the error is distributed normally, the slope estimates will be distributed normally, thus allowing for valid parametric inference. Unlike most of the other OLS assumptions, however, the violation of error normality is not only very difficult to test for with power in small samples (Belsley, Kuh, and Welsch 1980, 16–8; Greene 1993, 309–10), but it is also not easily corrected. Therefore, the plausibility of this distributional assumption for OLS can be questionable. And while it is commonly held that for many statistics, including OLS estimates, parametric inference is robust to "minor" violations of distributional assumptions (Achen 1982, 36–7; Malinvaud 1970, 297–9; Srivastava 1958; Bartlett 1935), there is evidence that a non-normal distribution can adversely affect the error rate of parametric inference, at least in small samples (Geary 1947; Pearson 1931; Ghurye 1949). This suggests that the effect of the violation of this assumption on parametric inference with OLS is an open question.

The typical response to this issue in practice is to invoke the central limit theorem, which holds that as sample size increases, the OLS slopes will be normally distributed regardless of the distribution of the error. A key issue then is just how large a sample size is "large enough" to invoke the central limit theorem and comfortably apply parametric inference. Textbooks advise that a sample of 30–50 is probably large enough in most situations (e.g., Mansfield 1986, 241), but this will vary with the type of error non-normality. For example, the error may be highly skewed, reflecting the skew of the dependent variable (e.g., national GNP or life expectancy), or it could be platykurtic (flatter than a normal distribution) due to some unmodelled bimodality in the process.[12] Further, there are many situations where a political scientist is constrained to conduct analysis on less than 30–50 data points. This is often the case in studies of industrial-

[11] In the 1993 volumes of five top political science journals (*American Journal of Political Science*, *American Political Science Review*, *Journal of Politics*, *British Journal of Political Science*, and *Political Research Quarterly*), fully 35.6% of all articles reported OLS results.

[12] Skewed error is often dealt with by various data transformations, but this is less than desirable for at least two reasons. First, the transformation of variables in regression models can place undue restrictions on the error structure (Linehan and Schrodt 1978). Second, for purposes of interpretation it is better to work with data in its original metric when possible. But more importantly, in small samples it may be difficult to detect non-normalities of error in the first place.

ized countries (Lijphart and Crepaz 1991), presidential primaries in a given year (Norrander 1993), the U.S. states (Ringquist 1993), or limited time series (Costain and Majstorovic 1994), just to name some common examples.

Due to these potential problems with this distributional assumption, we must ask how accurate parametric inferences are in these circumstances, and whether the bootstrap can do better. Given the prevalence in political science of OLS analysis and the potential for the widespread violation of the assumption of normal model error, these are important questions to the discipline.

Therefore, I conducted an extensive series of Monte Carlo experiments to address these questions (see Appendix).[13] The population generating function is a bivariate regression model:

$$Y_i = 1 + 2 * X_i + w * \epsilon_i, \tag{4}$$

where X_i is the set of sequential integers from 1 to n, $\epsilon_i \sim \chi^2_{df}$ centered on 0, and w is an error weight. In order to assess the performance of the various confidence interval methods under a variety of conditions, I systematically varied the shape of the error distribution from being highly skewed to approximating normality by changing the degrees of freedom in the χ^2 distributed error term from 1 to 15 by ones, I varied the R^2 of the model from .05 to .96 via the error weight, w, and I varied the sample size from 10 to 50 by fives.[14] For each unique pseudo-population combination of R^2, error distribution and sample size, a Monte Carlo experiment assessing the performance of each confidence interval technique for the slope coefficient was undertaken. There were 500 trials in each Monte Carlo experiment; 1,000 bootstrap re-samples were used in each trial.

The results of a typical experiment are displayed in Table 1.[15] Here the average R^2 for the 500 trials is .79, the sample size is 30 and the error term is distributed $\chi^2_{df=4}$. This is a marginal case of the violation of the normality assumption. A $\chi^2_{df=4}$ distribution is somewhat skewed to the right, but with

[13]Throughout this paper, the residuals re-sampling approach for bootstrapping regression is used (Mooney and Duval 1993, sec. 1.3).

[14]Sample sizes of 10–50 were used in the experiments as preliminary analysis indicated that this was the relevant range in which variation of the performance variable took place, and the full experiment results bore out this expectation (see Figure 1).

[15]This experiment is typical in as much as the independent variables are all in the middle range of their potential values, but of course with an experimental design, any set of values for the independent variables is equally likely to be observed. This case was chosen as an example especially for its similarity to the House elections data set discussed below.

Table 1. Example Monte Carlo Experiment Comparing Confidence Interval Performance Around a Bivariate Regression Slope

	Median nominal $\alpha/2$ endpoint	Median nominal $1-\alpha/2$ endpoint	α-level[a]
Monte Carlo estimate	1.6310	2.3392	***
Parametric[b]	1.6174	2.3873	.038
	(.21)[c]	(.21)	
Normal Approximation	1.6286	2.3699	.048
	(.21)	(.21)	
Percentile	1.6405	2.3580	.058
	(.21)	(.21)	
BCa	1.6389	2.3542	.050
	(.21)	(.21)	
Percentile-*t*	1.6162	2.3821	.042
	(.21)	(.21)	

Nominal α-level = .05, n = 30, average R^2 = .79.
Pseudo-population model: $Y_i = 1 + 2X_i + 3.25\epsilon_i$, where $\epsilon \sim \chi^2_{df=4}$ and $X = (1, 2, 3, \ldots, 30)$.
Monte Carlo trials = 500, bootstrap re-samples = 1,000, residuals re-sampling.
Monte Carlo estimate of β = 2.00 (s.e. = .19).
Running time = 4.98 hours for a GAUSS 3.0 program on an IBM 486-DX.
[a]The proportion of trials that the true value of $\beta(2.0)$ was excluded from the confidence interval.
[b]Using $t_{.025;df=28} = 2.048$ for the parametric and normal approximation intervals.
[c]The standard error of the confidence interval endpoint estimate.

a sample of 30, one might be willing to invoke the central limit theorem and assume $\hat{\beta}$ was distributed normally regardless of the distribution of the error. Further, it would be difficult in practice to detect this deviation from normality with this sample size. For example, in this experiment a Jarque-Bera test on these residuals could reject the null hypothesis of normality at the .05 α-level in only 202 out of 500 trials (Greene 1993, 309–10).

Given this plausible scenario and tough test for the bootstrap, how does it perform? Table 1 indicates that all of the bootstrap intervals outperformed the parametric interval on the crucial test of approximating the nominal α-level, and two outperformed it by a fair degree. Where the parametric interval yielded almost one-quarter fewer Type I errors than were nominally allowed (thus reducing the power of the test unnecessarily), the BCa method outperformed all the rest by yielding the exact nominal α-level. A closer look at the median values of the endpoints indicates also that while all the intervals' endpoints centered very closely on the Monte Carlo estimates, the parametric interval was a bit too wide, while the bootstrap intervals were all closer to the Monte Carlo standard. The standard errors of

all the endpoints were very similar, indicating that these interval methods cannot be differentiated by their efficiency.

The results of Monte Carlo evaluations of inferential techniques, like those of analytic evaluations, are conditional on the specific statistical situation under which they are carried out. Therefore, in order to generalize from these results to a somewhat broader range of statistical situations, I have conducted response surface analysis on the results of multiple experiments (Hendry 1984). This analysis allows for the penetration of the random component of these processes and the identification of any underlying relationship between interval technique and inferential performance. First, the above described experiment was repeated 177 times, each time independently and systematically varying the error distribution, R^2, and sample size of the pseudo-population. Each confidence interval approach was subjected to the experiment for each pseudo-population, yielding no correlation between any of these factors and the interval method used.[16] For each of these 885 experimental outcomes (five intervals and 177 experiments on each), I developed an α-level performance measure calculated as the negative of the absolute difference between the nominal α-level (α) and that estimated in the experiment ($\hat{\alpha}$):

$$PERFORMANCE = -1 * |\alpha - \hat{\alpha}| \tag{5}$$

The negative is used for interpretation purposes, so that the higher the number, the better the performance. For example, in the experiment displayed in Table 1, the parametric interval has a performance level of $-.012$, and the percentile interval is scored at $-.008$.

Response surface analysis was then carried out using this performance measure as the dependent variable in a multiple regression model, with the independent variables being a set of dummies for the interval methods, the natural log of sample size, and interaction terms.[17] The reference dummy is the parametric interval, since the performance relative to that interval is

[16]Because error structure and sample size have some effect on the R^2, there was some correlation between these factors. I structured the experiments in such a way as to keep this correlation to a minimum, however, with no bivariate correlation between these three factors exceeding .31.

[17]The skew and kurtosis of the estimated bootstrap sampling distribution of the slope estimate, and the pseudo-population's R^2 were also included in the model, but their effects were found not to be statistically significant. I report the simplest model for the sake of clarity. Also, I followed Box and Draper's (1987, 275–8) procedure for determining the polynomial level of this model. The addition of another level of polynomials and interactions did not improve the fit to the data sufficiently to include them, using Box and Draper's adequacy criterion.

Table 2. Response Surface Analysis of Monte Carlo Experiments on the α-Level Performance of Confidence Intervals Around a Bivariate Regression Slope

	Estimated OLS coefficient
Normal Approximation	.0981*
	(.0084)[a]
Percentile	.0011
	(.0091)
BCa	.0082
	(.0098)
Percentile-*t*	.1352*
	(.0069)
Normal Approx. * ln(sample size)	−.0255*
	(.0024)
Percentile * ln(sample size)	.0001
	(.0026)
BCa * ln(sample size)	−.0011
	(.0028)
Percentile-*t* * ln(sample size)	−.0354*
	(.0020)
ln(sample size)	.0342*
	(.0019)
Constant	−.1396*
	(.0065)
n	885
R^2	.70
Breusch-Pagan (df = 9)	196.04

OLS regression analysis: Unit of analysis = a Monte Carlo experiment; Dependent variable = $-1 * |\alpha - \hat{\alpha}|$.
Monte Carlo experiments: $X = (1, 2, 3 \ldots n)$, residuals re-sampling, trials per experiment = 500, bootstrap re-samples = 1,000.
[a]White (1980) estimated standard error to account for heteroskedasticity.
*$p(\beta = 0) < .01$.

of interest. Sample size is included because it will undoubtedly be important in predicting performance. Parametric intervals should become more accurate as sample size increases because the sampling distribution of the OLS slope estimate will approach normality, and the proofs of the accuracy of the bootstrap intervals are also based on asymptotics (e.g., Bickel and Freedman 1981; Efron 1987). This effect should decrease as sample size increases, as performance will improve asymptotically to some limit, and this is reflected by the natural log transformation.

Table 2 displays the results of this analysis. As expected, the natural

log of sample size is positively related to performance to a statistically significant degree. But it is the interval dummies that are of greatest interest here. In these experiments, the general pattern is that all of the bootstrap methods outperform the parametric interval on the basis of approximating the nominal α-level, as indicated by the positive slope coefficients for the bootstrap interval method dummy variables. That is, since all these slope estimates are positive, the predicted value of performance for a case at almost any sample size[18] will be greater for any bootstrap interval than for the reference dummy, the parametric interval. This result corresponds to the findings of Navidi (1989) who uses Edgeworth expansion to show that the bootstrap is at least as good asymptotically as parametric inference for OLS.

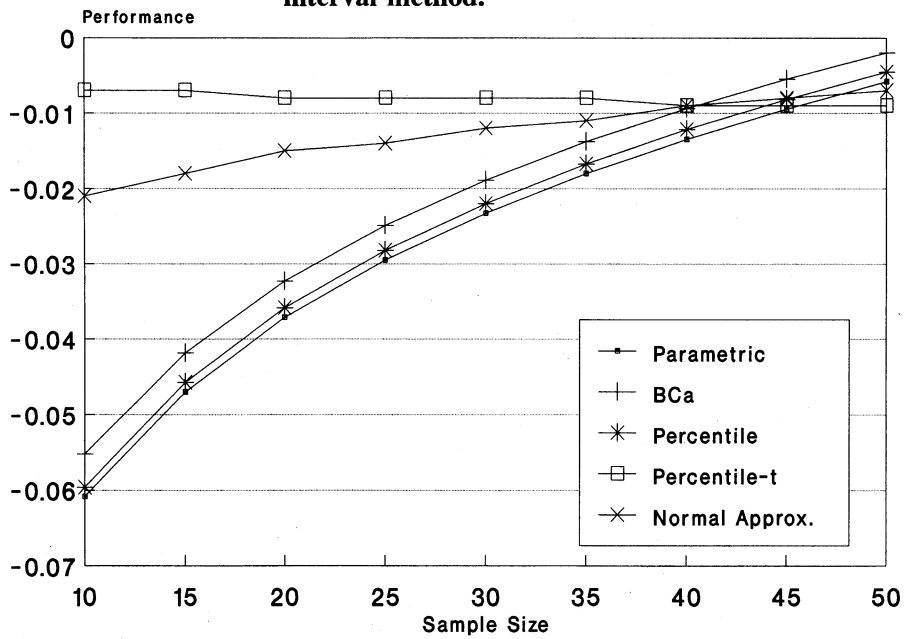
In these experiments, however, only the performance of the percentile-*t* and normal approximation methods are statistically distinct from that of the parametric approach (i.e., the coefficients are statistically distinct from zero).[19] The BCa's poor performance is somewhat surprising in that there is good theoretical reason to believe it is superior to the normal approximation method (Efron 1987). But this theory is based on asymptotics, and therefore does not necessarily hold in small samples. The asymptotics of the percentile-*t* apparently come into play before those of either the BCa or the parametric approach, at least in the case of OLS. The fact that the normal approximation technique outperforms the parametric when both techniques are using the same assumed standardized distribution for $\hat{\beta}$ indicates that the bootstrap estimate of the standard error of $\hat{\beta}$ is superior to that of OLS. Further research is needed to explore this result.

Figure 1 displays the model's predicted performance level for each interval technique at various sample sizes to provide further insight into the results of this analysis. Three points should be noted here. First, the percentile-*t* and normal approximation methods clearly outperform the other intervals in the smallest samples. For example, at $n = 10$ the percentile-*t* is predicted to have an α-level $\pm$ .007 away from the nominal .05

[18]Because of the differences in the curvature of the regression lines for different interval methods (Figure 1), the predicted performance of the parametric interval is higher than that of two other intervals at $n = 50$. This may be an anomaly of the experimental design, in that 50 is the upper limit for sample size in the experiments, making predictions to it less reliable.

[19]It may seem ironic to use parametric inference to test the coefficients of this model in an article describing bootstrap inference. But this is done because the large sample ($n = 885$) provides good reason to believe that the slope coefficients in this model are normally distributed, so there is no reason to expend the extra computation and programming time in bootstrapping this model.

Figure 1. Predicted performance on alpha level by sample size and interval method.



Source: Estimated model in Table 2. Performance = −1 * abs(nominal alpha − observed alpha).

value, whereas the parametric interval is predicted to be off by ± .060. This is the difference between a true α-level of .057 and one of .110, a substantial divergence. Second, note that increasing the sample size affects the performance of the BCa, percentile and parametric intervals to a much greater degree than it affects the percentile-*t* or normal approximation. In fact, the performance of the percentile-*t* is statistically indistinguishable from a flat line in this graph. This is because the percentile-*t* and the normal approximation appear to reach the performance limit at low samples in these experiments, and therefore cannot improve much. The third point to note about this graph is that all of these techniques converge in performance at a sample size of around 40–45. This indicates that any superiority in performance of the percentile-*t* only comes into play for rather small samples, such as for a subset of states or countries, or for a short time series.

As an example of an empirical situation which resembles these simulations, I look at a well-known model that characterizes the translation of

votes to seats in legislative elections. The "cube law" (Kendall and Stuart 1950; March 1957; Brookes 1959) holds that in a first-past-the-post, two-party system, the party with the majority of votes will get more legislative seats than it is entitled to by its proportion of the votes, and that the proportion of seats won will be on the order of the cube of the proportion of the votes received.[20] Using the parameter notation that is conventional in the seats/votes literature, one formulation of this model is:

$$Y_i = \beta X_i^{\rho} \epsilon_i, \tag{6}$$

where Y is the ratio of the minority party's seats to the majority party's seats, and X is the ratio of minority party's votes to the majority party's votes. This model is usually linearized, and so made amenable to OLS analysis, by taking the natural log of both sides, yielding:

$$\ln(Y_i) = \ln(\beta) + \rho \ln(X_i) + \ln(\epsilon_i). \tag{7}$$

The problem with this transformation is that in order for the error term in Equation 7 to be normally distributed, the error term in Equation 6 must be lognormally distributed (Linehan and Schrodt 1978), and there is no reason to believe that this would be the case. In fact, if ϵ is made up of the sum of a series of independent random shocks, as is usually assumed, the central limit theorem would lead us to believe that the error in Equation 6 tends to normality. Therefore, it is quite likely that when estimating Equation 7, the error structure will be non-normal.

To compare the confidence intervals and inferences of the bootstrap and parametric approaches on real data of this sort, I estimated Equation 7 with data from U.S. House of Representatives elections for 1932–88. This was a period of Democratic domination in the House, and the question of whether and to what degree there was some structural advantage for the Democrats is of substantive interest. I run an AR1 correction to eliminate autocorrelation, but otherwise the model is as presented in Equation 7.[21] Both ρ and $\ln(\beta)$ are of theoretical concern in this model. ρ is hypothesized

[20]There has been considerable debate as to the specification of this model (Linehan and Schrodt 1978; Schrodt 1982; Tufte 1973), the value of the exponent (Theil 1969; Sankoff and Mellos 1972) and the possibility of extending it to other types of electoral systems (Taagepera 1986; Casstevens and Morris 1972). I examine only the simplest model in that my principal concern with it is as an example of the use of OLS on a regression model with potentially non-normal error.

[21]An alternative to this model-based approach for bootstrapping a weakly dependent time series is to do moving block re-sampling (Liu and Singh 1992; Politis and Romano 1992).

Table 3. 95% Confidence Intervals Around the Slope and Intercept Parameters of the Seats/Votes Model for U.S. House Elections, 1932–88

Model: $Y_t = \ln(\beta) + \rho X_t + \phi e_{t-1} + \upsilon_t$, where Y = ln(Republican seats/Democratic seats); X = ln(Republican votes/Democratic votes).

1)ln(β): OLS-AR(1) estimate = −.152 (s.e. = .076)

	α/2 endpoint	1-α/2 endpoint
Parametric	−.3085	.0046
Normal Approximation	−.2201	−.0838
Percentile	−.1977	−.0677
BCa	−.2533	−.1076
Percentile-*t*	−.1868	−.0852

2) ρ: OLS-AR(1) estimate = 1.942 (s.e. = .144)

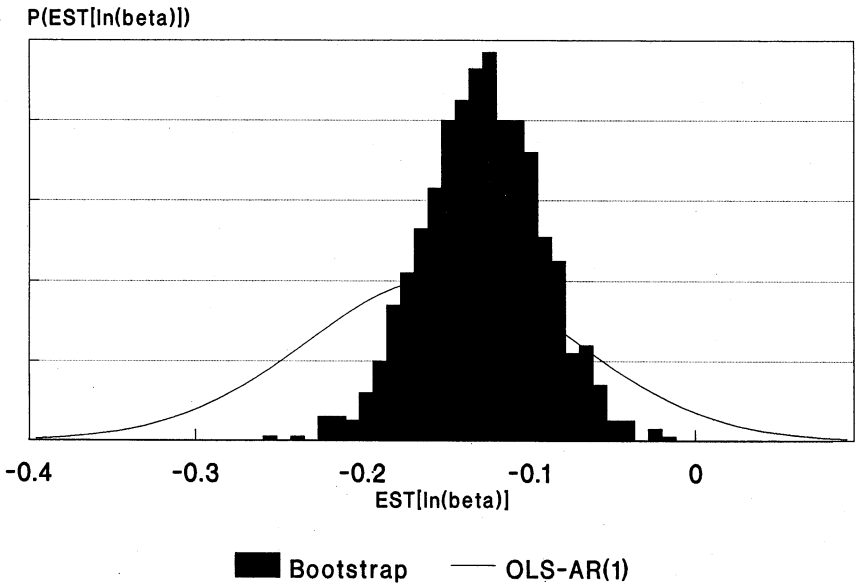
	α/2 endpoint	1-α/2 endpoint
Parametric	1.6468	2.2366
Normal Approximation	1.5677	2.3156
Percentile	1.6764	2.3966
BCa	1.5139	2.2199
Percentile-*t*	1.7023	2.3526

n = 30.
OLS-AR(1) estimate of ϕ = .721 (s.e. = .127).
Durbin-Watson = 1.84.
R_a^2 = .867.
α = .05.
Bootstrap re-samples = 1,000, residuals re-sampling.
Running time = 2.51 hours on a RATS 3.0 program running on an IBM 486-DX. This includes running the calculations for the results displayed in Table 5, which involved the double-bootstrap for *VOTE**.

to be three by the cube law, and ln(β) will be zero if there is no systematic partisan bias in the system (Jackman 1994, 325–7).

Table 3 displays the parametric and bootstrap confidence intervals around both ln(β) and ρ for this model. Several points should be noted about these results. First, the parametric confidence interval around ln(β) is clearly wider than the bootstrap intervals. Note also that the bootstrap intervals are fairly similar to one another in this respect, indicating that the choice between them would probably not be a critical factor here. This leads to the rejection of the substantively important null hypothesis, ln(β)

Figure 2. Bootstrap vs OLS-AR(1) estimates of the sampling distribution of estimated ln(beta) from US House data.



OLS-AR(1) distribution is based on a mean of −.152 and a standard error .076 (see Table 3).

= 0, for the bootstrap, but not for the parametric approach, at the .05 α-level. This is reflective of the significantly narrower bootstrap estimate of the sampling distribution of the estimate of ln(β) from which all the bootstrap confidence intervals have been generated (see Figure 2 and Table 4). Figure 2 shows the effect of the reduction by over half of the standard error of the estimate of ln(β) when comparing the bootstrap to the parametric sampling distribution estimate. And since the critical theoretical value of ln(β) (0) falls into the region where the bootstrap and the parametric estimates are discrepant, the difference is important in this case.

Figure 2 also demonstrates two noteworthy points about the bootstrap sampling distribution generally. First, the distribution will always be irregular and discrete, even with a large number of re-samples. This "messiness," as compared with the smooth parametric sampling distribution estimate, is an aestetic drawback of the bootstrap, but it does not reduce its effectiveness in making inferences. Second, the histogram of this distribution can be used creatively to say more about the variation in the statistic given the

Table 4. Characteristics of the Bootstrap vs. Parametric Sampling Distributions of the Statistics Estimated from the U.S. House Seats/Votes Model

	OLS-AR(1) Sampling Dist.			Bootstrap Sampling Dist.		
	$\hat{\sigma}_{\hat{\theta}}$	Skew	Kurtosis	$\hat{\sigma}_{\hat{\theta}}$	Skew	Kurtosis
EST[ln(β)]	.076	0	0	.033	−.06	.05
$\hat{\rho}$	.144	0	0	.182	.168**	.30*
*VOTE**	NA[a]	NA	NA	.005	.50**	.37**

Values for a normal distribution: skew = 0, kurtosis = 0.
[a]Parametric statistical theory offers no insight into the standard deviation or functional form of *VOTE**.
**$-p(\theta = 0) < .01$.
*$-p(\theta = 0) < .05$.

data than just its standard error or α-level percentile points. For example, one could compare the effect of the skew of a parameter's estimated sampling distribution on a set of null hypotheses. While this is a very under researched area, work continues to develop techniques to use this available information better (Loftus 1993). Note that this sample-based information is made available only with the bootstrap, as the functional form of parametric sampling distribution estimates are imposed by assumption.

In contrast to the results for ln(β), the bootstrap intervals around ρ are wider than the parametric interval. This again reflects the shape of the bootstrap estimate of the sampling distribution, which has a larger standard error and is platykurtic in this case (Table 4). This illustrates that there is no consistent effect on the parametric interval of distributional assumption violation, because different parameters may have different shapes. Only the bootstrap allows the interval to reflect the shape of the sampling distribution that is observed in the data, rather than a shape that is imposed by (in this case, an unsupported) assumption. Also note in this regard that the BCa does a good job of shifting the percentile confidence interval endpoints to the left to correct for the positive skew in the estimated sampling distribution.

Given the results of my response surface analysis (Table 2 and Figure 1) and previous theoretical and empirical work on bootstrapping OLS models (e.g., Hall 1992; Mooney and Duval 1993, 57; Navidi 1989), it is likely that the percentile-*t* is the best confidence interval in this situation. The substantive conclusions are generally the same for each bootstrap method

in this example, however. First, the "cube" aspect of the cube law does not appear to be supported by these data; the disproportionate advantage of the majority party appears to be on the order of the square of the votes ratio. Second, the bootstrap intervals around $\ln(\beta)$ provide evidence that there may have been other structural biases in favor of the Democrats in this period (perhaps gerrymandering or geographic concentration advantage). This is an inference that might not have been made had the parametric interval been relied upon, or at best the evidence for it would have been of a more marginal nature.

IV. Inference When No Parametric Alternative Exists

The above analysis demonstrates that there are situations where parametric inference is mathematically possible, but where it is outperformed by the bootstrap. The above analysis also suggests however, that at least in the case examined, parametric inference does not perform too badly. The parametric error rate in Table 1 is not wildly off the nominal α-level mark, and Figure 1 shows that by a not-too-large sample size, all the methods tend to converge in their performance. Therefore, one may honestly ask, where is the pay-off for all this extra computation and the upfront costs of programming the bootstrap?

I believe that the biggest pay-off from the bootstrap will prove to be in making inferences where no parametric alternative exists at all. As Efron and Tibshirani point out (1993, 11; see also Efron and LePage 1992), virtually the only statistic with an easy to obtain analytic formula for its standard deviation and a strong justification for its distributional theory is the sample mean. Much of the discipline of statistics in the first half of the 20th century was occupied with either trying to figure out ways to use the mean to estimate various characteristics and thereby take advantage of its desirable properties, or working up distributional theory on a few non-mean statistics, usually based on heroic distributional assumptions (for example, the test for a comparison of sample variances requires both variables to be normally distributed; see Daniel 1990, 102–11). This has led political scientists and researchers in other disciplines using statistical models to rely heavily on the few models for which distributional theory is fairly sound. If substantive theory calls for a different sort of model, the researcher who is rightly concerned with statistical inference may well opt for the substantively less appropriate model due to the ability to conduct inference on it. This is clearly undesirable.

What then are these models and estimators that may fit well with substantive theory, but have weak statistical theory associated with them? There is a wide variety of these, and more are being developed continually.

Some of these models and estimators are already frequently used, though the implications of not being able to make parametric inferences with them are usually ignored. For example, there is no analytical way to make inferences using eigenvalues and principal components (Efron and Tibshirani 1993, 61–70), estimates of the parameters of discriminant analysis (Chant and Dalgleish 1992; Dalgleish 1994), or the indirect paths in causal models (Bollen and Stine 1991). The cited authors have applied the bootstrap to the problem of making inferences to these widely used statistics. Examples of other potentially useful statistics with no standard error formulae and/or no known functional form that are currently less widely used in political science include the switch point in a switching regression model (Douglas 1987), loess curve-fitting parameters (Cleveland, Grosse, and Shyu 1992), and Stein and Mundlak's regression slope estimators (Brownstone 1990).

In political science, there are increasingly frequent examples of the use of theoretically appropriate statistics with weak statistical theory, requiring analysts to rely on non-standard strategies to conduct inference. Bartels (1993) uses a ratio of regression parameters to test the distinctiveness of media impact on voters, and employs Monte Carlo simulation to assess that statistic's variability. Krehbiel (1990) examines the difference between medians to assess congressional committee bias, but is forced to conduct his inferential tests on the difference between means. Gelman and King (1994) assess the bias and responsiveness of legislative districting plans using estimators derived from estimated regression coefficients and hypothetical regressor values, and use a Bayesian simulation approach for inference. In each of these cases, the bootstrap could have provided an alternative, perhaps preferable, and certainly more general,[22] inference strategy. The more general and important point is that with the availability of the bootstrap, researchers can focus on selecting the appropriate point estimator for the theory at hand, and still make plausible statistical inferences regardless of the level of distributional theory that has been developed for that statistic.

As an example of such a situation, I return to the seats/votes model of the U.S. House elections of 1932–88. Jackman (1994, 327) suggests a measure of bias in a legislative electoral system that combines the components of the model in Equation 7 in the following way:

$$VOTE^* = \frac{\exp(-\ln[\hat{\beta}]/\hat{\rho})}{1 + \exp(-\ln[\hat{\beta}]/\hat{\rho})}. \tag{8}$$

[22]While Bayesian simulation is certainly a general approach, the need to specify distributional assumptions (as in Gelman and King 1994) restricts its generalizability as compared to the bootstrap.

Table 5. 95% Bootstrap Confidence Intervals Around the "Vote Needed" Parameter for U.S. House Elections, 1932–88

*VOTE** using OLS estimates of $\ln(\beta)$ and ρ from Table 3 = .5196 (bootstrap est. of s.e. = .005)

	$\alpha/2$ endpoint	$1-\alpha/2$ endpoint
Normal Approximation	.5085	.5306
Percentile	.5074	.5289
BCa	.5126	.5349
Percentile-*t*	.5066	.5256

α = .05.
Bootstrap re-samples = 1,000, double-bootstrap re-re-samples = 200, residuals re-sampling.

*VOTE** is the estimate of the proportion of votes the minority party (the Republicans in this case) would need to get in order to win 50% of the seats in a legislative election. If *VOTE** is over .5, then the system is biased against them. For example, using the estimate of $\ln(\beta)$ and $\hat{\rho}$ in Table 3, the best estimate of the proportion of the vote the Republicans needed to get to win 50% of the House seats in this period is .5196. But this point estimate is of course a random variable, and we do not know whether the population parameter being estimated could plausibly be expected to be between .49 and .53 or between .20 and .90. This is where statistical inference comes into play.

While Jackman's combination of these parameter estimates offers a very intuitive and important interpretation of the data, the sampling distribution of this statistic is quite difficult to derive mathematically, and even its standard error cannot be estimated using analytic techniques. Therefore, parametric inference cannot be used to make probability statements about whether the true vote needed is .5, the logical null hypothesis value. This is especially of concern when the point estimate is quite close to the null hypothesis, as it is in this case.

I have developed bootstrap confidence intervals around the proportion of the vote needed for the Republican party to win 50% of seats for the U.S. House data, and these are presented in Table 5. These results indicate that while the point estimate is quite close to the null hypothesis value, its extremely small standard error leads to 95% confidence intervals that do not incorporate that null value for any of the bootstrap interval methods. That is, it seems quite likely that the electoral system for the U.S. House

had a small but statistically significant bias against the Republicans during this period. Had I only the ability to conduct parametric inference, I would have had no idea that the variability of this statistic was so low, and my seat-of-the-pants substantive conclusion may well have been different than the one drawn using the bootstrap.

I was constrained by logistics from conducting a response surface analysis (e.g., Table 2) to evaluate the bootstrap intervals in this case.[23] Therefore, I conducted a single Monte Carlo experiment, attempting to simulate the data analytic situation in the U.S. House data in all relevant respects. The population generating model is:

$$Y_i = -.5 + 2.5 * X_i + .25 * \epsilon_i, \tag{9}$$

where $\epsilon \sim \chi^2_{df=1}$ and X is the set of consecutive integers from one to 30. Sample size is 30, and the average R^2 of the 250 trials is .876. In each trial, I conducted 1,000 re-samples, with 200 "re-re-samples" from each re-sample for the percentile-t interval. Given the population values of $\ln(\beta)$ and ρ (in Equation 9), the population value of the vote needed by the minority party to win 50% of the seats is .55. The question is, how accurate are the bootstrap confidence intervals around this parameter?

Table 6 displays the output from this Monte Carlo experiment. First, note that the Monte Carlo estimate of the parameter (the average *VOTE** for the 250 trials) is .551, and its standard deviation is .010. This indicates both that the Monte Carlo estimate is likely to be unbiased, and that the bootstrap estimate of very low variability for *VOTE** in the U.S. House data is supported by these experimental results.

A comparison of the confidence interval endpoints with those of the Monte Carlo percentile standards indicates that three of the intervals are a bit too narrow in their coverage. The percentile-t interval has very close to the same coverage as the Monte Carlo interval, but it is shifted to the right slightly. These discrepancies from the Monte Carlo standard are reflected in the fact that the α-levels of the bootstrap intervals are all too high, ranging

[23] Since there is no analytic formula for the standard error of *VOTE**, a double-bootstrap procedure was necessary to construct the percentile-t confidence interval here (Mooney and Duval 1993, 40). Therefore, the single experiment undertaken (Table 6) took over a week to run, making it highly impractical to conduct the large number of experiments needed for response surface analysis. But since the main point of the response surface analysis was to compare the bootstrap intervals to the parametric approach, it is less important here because no such comparison can be made.

Table 6. Monte Carlo Experiment Assessing Bootstrap Confidence Intervals Around the "Vote Needed" Parameter

	Median nominal $\alpha/2$ endpoint	Median nominal $1-\alpha/2$ endpoint	α-level[a]
Monte Carlo estimate	.5278	.5673	***
Normal Approximation[b]	.5370	.5646	.084
	(.010)[c]	(.006)	
Percentile	.5370	.5632	.088
	(.010)	(.006)	
BCa	.5383	.5609	.176
	(.010)	(.007)	
Percentile-*t*	.5370	.5707	.076
	(.010)	(.013)	

Nominal α-level = .05, n = 30, average R^2 = .876.
Pseudo-population model: $Y_i = -.5 + 2.5X_i + .25\epsilon_i$, where $\epsilon \sim X^2_{df=1}$ and X = (1, 2, 3 . . . 30); population Vote Needed = .55.
Monte Carlo trials = 250, bootstrap re-samples = 1,000, double-bootstrap re-re-samples = 200, residuals re-sampling.
Monte Carlo estimate of Vote Needed = .551 (s.e. = .010).
Running time = 180.5 hours for a GAUSS 3.0 program on an IBM 486-DX.
[a]The proportion of trials that the true value of Vote Needed (.55) was excluded from the confidence interval.
[b]Using $t_{.025;df=28} = 2.048$ for the normal approximation interval.
[c]The standard error of the confidence interval endpoint estimate.

from a moderately poor .076 for the percentile-*t* to a very weak .176 for the BCa. In other words, the bootstrap intervals have failed to include the true "vote needed" value more times than the nominal α-level would have led us to believe.

While this coverage error should concern us, two points need to be kept in mind. First, these rates of error are on the order of n^{-1} for the percentile-*t* and $n^{-1/2}$ for the BCa. These are standards that can be met by few general inferential techniques. For example, using a *z*-score where a *t*-score is more appropriate in parametric inference yields a coverage error on the order of n^{-1} (Kendall and Stuart 1977, 404). In fact, Hall (1992, 94) argues that it is inherent in almost all inferential techniques to have a coverage error on the order of at least $n^{-3/2}$. And of course as sample size increases, the absolute value of this error shrinks quite quickly.

The second point to keep in mind about these results is that while there may be a bit of perhaps unavoidable coverage error here, the bootstrap still offers the *only feasible method of making probability-based inference for*

this population parameter. This is crucial. Without the bootstrap, there is simply no way to test whether or not the variability in *VOTE** overwhelms the point estimate to the degree that sound statements about substantively interesting values of its parameter can be made.[24] The question then becomes one of whether the researcher decides to be slightly too conservative or too liberal in Type I error rate (depending on the situation), or fails to conduct inference at all.[25]

V. Conclusion

Because of its unique potential to break through the 100-year-old barriers of parametric inference, the bootstrap has captured the interest of the statistical and scientific community. In political science, there are many common data analysis situations where either parametric assumptions are violated, or where there exists no parametric alternative for inference. And as political scientists take advantage of the new and more flexible models that have been developed, the need for the bootstrap will grow. The two data analytic situations discussed in this article are only examples of the broad range of possibilities for better inference and more creative modelling that the bootstrap offers.

At this point, it should be remembered that my examples and evaluations must be considered as suggestive rather than comprehensive or authoritative. For example, the poor performance of the BCa confidence interval in Tables 2 and 6 should not be taken as a categorical condemnation of this approach, but may merely indicate that it requires larger samples than the percentile-*t* to achieve accuracy. The convergence of the predicted performance of all the interval techniques in Figure 1 suggests this may be the case. In fact, Efron (1987) argues that the BCa is indeed the preferred interval method generally, while Hall (1992) argues that the percentile-*t* is preferred in regression-based analyses. It is also the case that my specific evaluation conclusions are only strictly applicable to the particular statistical estimators and situations tested. This is true of all evaluations of infer-

[24]Jackman (1994, 352) conducts inference using *VOTE** with a parametric bootstrap, assuming that the estimates of $\ln(\beta)$ and ρ are bivariate normally distributed. My approach is nonparametric and therefore more general than Jackman's, though when dealing with samples as small as four and seven, as Jackman does, a parametric assumption may be necessary.

[25]It is advisable to conduct a Monte Carlo simulation experiment (such as in Table 6) on the bootstrap's performance for a given statistic the first time it is bootstrapped. In this way, at least an estimate of the performance of the various interval methods can be made, and compensations made in the substantive conclusions for any obvious problems.

ence approaches, however, whether these evaluations be mathematical or simulation-based. But these results and those of many others that have appeared in the statistical literature over the past 15 years lead me to conclude that bootstrap inference provides a better estimate of the variability of a far greater range of statistics than any other general approach.

This paper has shown that the bootstrap performs at least as well as parametric inference in a situation where parametric inference is known to perform very well (OLS estimation of a regression model), and it "outperforms" parametric inference when parametric inference cannot be applied. It is therefore a more general approach, and as such needs to be considered carefully and included in the toolkit of every political scientist doing statistical analysis. It is doubtful that the bootstrap will take the place of parametric inference in the discipline any time soon—more theory and software need to be developed to apply it and know its limitations. Perhaps more importantly, there are generations of political scientists who have been trained exclusively in parametric inference, and re-tooling in such a fundamentally important area is costly. But over time, as new generations of political scientists learn the bootstrap and related methods in graduate school, and as software to perform the technique becomes more readily available, the bootstrap may well supplant the more restrictive and less intuitive parametric approach because of its broader scope. Much like the use of correlation coefficients and partial correlation coefficients was largely supplanted in the discipline by the more general multiple regression approach (perhaps 30 years after the latter's full development in the 1930s), the bootstrap may well become the dominant approach to statistical inference in the 21st century.

Manuscript submitted 16 March 1995.
Final manuscript received 19 June 1995.

APPENDIX
Monte Carlo Methodology

In evaluating the performance of an inference technique, the primary concern is with its error rate in its statements about population parameters. Monte Carlo simulation is the only way to conduct such an evaluation in finite samples, because it is only in these experiments that information about the value of these population parameters can be known (Mooney n.d., sec. 4.3; Davidson and MacKinnon 1993, 768). The Monte Carlo experiments for this paper were conducted as follows. First,

a "pseudo-population" was defined by a random number generating process,[26] and a sample size of n was generated. This was the original sample for the first *trial.* From this sample, the statistic of interest ($\hat{\theta}_t^S$) was calculated and stored, where the superscript "S" denotes it was developed from an original sample, and the subscript "t" indicates the trial number. Confidence intervals (with a nominal $\alpha = .05$ throughout) around θ were then computed using the parametric and four bootstrap methods. This entire procedure, from generating the sample to estimating the confidence intervals, was one Monte Carlo trial, and each Monte Carlo experiment consisted of either 250 or 500 trials, as noted in the text.

These hundreds of confidence intervals for each inference technique in each statistical situation were then used to evaluate the error rate associated with these inference techniques.[27] A basic concern of statistical inference is determining the probability of rejecting a true null hypothesis, that is, the Type I error rate (α).[28] To estimate the Type I error rate for an interval method, I noted if the value of the pseudo-population parameter (known from the random number generating process) was included in that interval for each trial. If it was not included, a Type I error had been committed. That is, the confidence interval failed to cover the value of the pseudo-population parameter, and if this had been the null hypothesis value, the null would have been incorrectly rejected. The proportion of trials with such an error is the observed Type I error rate ($\hat{\alpha}$) for an interval in that specific statistical situation. This is then compared to the nominal α-level (.05) of the test. The main standard of evaluation will be how closely an interval method's observed α-level comes to the nominal α-level.

A secondary standard of evaluation will be the Monte Carlo estimates of the endpoints of the confidence intervals. These are developed from the relative frequency distribution of the $\hat{\theta}_t^S$'s in each experiment. For example, the Monte Carlo estimate of the lower endpoint of a 95% confidence interval around $\hat{\theta}$ is simply the 2.5^{th} percentile value of the sorted $\hat{\theta}_t^S$'s in an experiment. This standard will be compared to the median endpoint generated by each confidence interval method. This comparison will be used to assess the bias in the endpoint estimation of each technique, and the standard deviation of the endpoints' distributions will be used to assess efficiency.

---

[26]I used the pseudo-random number generating process in GAUSS 3.0 to conduct these experiments. This process will not cycle in less than $(2^{31} - 1)$ numbers, which is adequate to ensure that for my purposes these act like random numbers. For brevity, I therefore refer to these as random numbers.

[27]While error rates are typically discussed in the context of hypothesis tests rather than confidence intervals, the parallels are transparent.

[28]Since Type II error rates are typically of less concern to political scientists, I do not compare these confidence interval techniques on this characteristic in this article. In the experiments reported herein, using commonsensical null hypotheses, these techniques differ little in their power. To compare the power of these intervals sensitively, experiments must be designed specifically with this characteristic in mind. I leave this to future research.

REFERENCES

Achen, Christopher H. 1982. *Interpreting and Using Regression.* Sage University Paper series on Quantitative Applications in the Social Sciences, 07-029. Newbury Park, CA: Sage.

Bartels, Larry M. 1993. "Message Received: The Political Impact of Media Exposure." *American Political Science Review* 84:149–63.

Bartlett, M. S. 1935. "The Effect of Non-Normality on the *t* Distribution." *Proceedings of the Cambridge Philosophical Society* 31:223–31.

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics.* New York: John Wiley.

Berry, William D. 1993. *Understanding Regression Assumptions.* Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-092. Newbury Park, CA: Sage.

Bickel, P. J., and D. A. Freedman. 1981. "Some Asymptotics on the Bootstrap." *The Annals of Statistics* 9:1196–1217.

Bollen, Kenneth A., and Robert Stine. 1991. "Direct and Indirect Effects: Classical and Bootstrap Estimates of Variability." In *Sociological Methodology 1991*, ed. Clifford C. Clogg. San Francisco: Joss-Bassey.

Bollen, Kenneth A., and Robert A. Stine. 1994. "Bootstrapping Goodness-of-Fit Measures in Structural Equation Models." *Sociological Methods and Research* 21:205–29.

Box, George E. P., and Norman R. Draper. 1987. *Empirical Model-Building and Response Surfaces.* New York: John Wiley.

Brewer, Ken. 1985. "Behavioral Statistics Textbooks: Sources of Myths and Misconceptions." *Journal of Educational Statistics* 10:252–68.

Brookes, Ralph H. 1959. "Legislative Representation and Party Vote in New Zealand." *Public Opinion Quarterly* 23:287–91.

Brownstone, David. 1990. "Bootstrapping Improved Estimators For Linear Regression Models." *Journal of Econometrics* 44:171–87.

Casstevens, Thomas W., and William D. Morris. 1972. "The Cube Law and the Decomposed System." *Canadian Journal of Political Science* 5:521–31.

Chant, David, and L. I. Dalgleish. 1992. "A SAS Macro for Jackknifing the Results of Discriminant Analysis." *Multivariate Behavioral Research* 27:323–33.

Cleveland, William S., Eric Grosse, and William M. Shyu. 1992. "Local Regression Models." In *Statistical Models in S,* ed. John M. Chambers and Trevor J. Hastie. Pacific Grove, CA: Wadsworth and Brooks.

Costain, Anne N., and Steven Majstorovic. 1994. "Congress, Social Movements and Public Opinion: Multiple Origins of Women's Rights Legislation." *Political Research Quarterly* 47:111–35.

Dalgleish, L. I. 1994. "Discriminant Analysis-Statistical Inference Using Jackknife and Bootstrap Procedures." *Psychological Bulletin* 116:498–508.

Daniel, Wayne W. 1990. *Applied Nonparametric Statistics.* 2nd ed. Boston: PWS-Kent.

Davidson, Russell, and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics.* New York: Oxford University Press.

Douglas, Stratford M. 1987. "Improving the Estimation of a Switching Regressions Model: An Analysis of Problems and Improvements Using the Bootstrap." Ph.D. diss. University of North Carolina, Chapel Hill.

Efron, Bradley. 1979a. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7:1–26.

———. 1979b. "Computers and the Theory of Statistics: Thinking the Unthinkable." *SIAM Review* 21:460–80.
———. 1981. "Nonparametric Standard Errors and Confidence Intervals" (with discussion). *The Canadian Journal of Statistics* 9:139–72.
———. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.
———. 1987. "Better Bootstrap Confidence Intervals" (with discussion). *Journal of the American Statistical Association.* 82:171–200.
Efron, Bradley, and Raoul LePage. 1992. "Introduction to the Bootstrap." In *Exploring the Limits of Bootstrap,* ed. Raoul LePage and Lynne Billard. New York: John Wiley.
Efron, Bradley, and R. Tibshirani. 1986. "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy." *Statistical Science* 1:54–77.
Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap.* London: Chapman & Hall.
Fan, X., and William G. Jacoby. 1995. "BOOTSREG-An SAS Matrix-Language Program for Bootstrapping Linear Regression Models." *Educational and Psychological Measurement* 55:764–8.
Fisher, Nicholas, and Peter Hall. 1990. "On Bootstrap Hypothesis Testing." *Australian Journal of Statistics* 32:177–90.
Fisher, R. A. 1915. "Frequency Distribution of Values of the Correlation Coefficient in Samples From an Indefinitely Large Population." *Biometrika* 10:507–21.
Freedman, David A. 1981. "Bootstrapping Regression Models." *Annals of Statistics* 9: 1218–28.
Geary, R. C. 1947. "Testing For Normality." *Biometrika* 34:209–42.
Gelman, Andrew, and Gary King. 1994. "A Unified Method of Evaluating Electoral Systems and Redistricting Plans." *American Journal of Political Science* 38:514–54.
Ghurye, S. G. 1949. "On the Use of Student's *t*-Test in an Asymmetrical Population." *Biometrika* 36:426–30.
Green, Donald Phillip, and Jonathan S. Krasno. 1990. "Rebuttal to Jacobson's 'New Evidence for Old Arguments'." *American Journal of Political Science* 34:363–72.
Greene, William H. 1993. *Econometric Analysis.* 2nd ed. New York: Macmillan.
Hall, Peter. 1992. *The Bootstrap and the Edgeworth Expansion.* New York: Springer-Verlag.
Hendry, David F. 1984. "Monte Carlo Experimentation in Econometrics." In *Handbook of Econometrics,* vol. II, ed. Z. Griliches and M. D. Intriligator. Amsterdam: Elsevier.
Hinckley, David W. 1988. "Bootstrap Methods." *Journal of the Royal Statistical Society,* series B. 50:321–37.
Jackman, Simon. 1994. "Measuring Electoral Bias: Australia, 1949–1993." *British Journal of Political Science* 24:319–57.
Kendall, M. G., and A. Stuart. 1950. "The Law of Cubic Proportions in Election Results." *British Journal of Sociology* 1:183–97.
———. 1977. *The Advanced Theory of Statistics,* vol. 1, 3rd ed. London: Griffin.
King, Gary. 1991. "Constituency Service and Incumbency Advantage." *British Journal of Political Science* 21:119–28.
Krehbiel, Keith. 1990. "Are Congressional Committees Composed of Preference Outliers?" *American Political Science Review* 84:149–63.
Lambert, Zambert V., Albert R. Wildt, and Richard M. Durand. 1990. "Assessing Sampling Variation Relative to Number of Factors Criteria." *Educational and Psychological Measurement* 50:33–48.

Lijphart, Arend, and Markus M. L. Crepaz. 1991. "Corporatism and Consensus Democracy in Eighteen Countries: Conceptual and Empirical Linkages." *British Journal of Political Science* 21:235–46.

Linehan, William J., and Philip A. Schrodt. 1978. "A New Test of the Cube Law." *Political Methodology* 4:353–67.

Liu, Regina Y., and Kesar Singh. 1992. "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence." In *Exploring the Limits of Bootstrap,* ed. Raoul LePage and Lynne Billard. New York: John Wiley.

Loftus, Geoffrey R. 1993. "A Picture Is Worth a Thousand *p* Values: On the Irrelevance of Hypothesis Testing in the Microcomputer Age." *Behavior Research Methods, Instruments and Computers* 25:250–6.

Malinvaud, E. 1970. *Statistical Methods of Econometrics.* New York: American Elsevier.

Mansfield, Edwin. 1986. *Basic Statistics with Applications.* New York: W. W. Norton.

March, James G. 1957. "Party Legislative Representation as a Function of Election Results." *Public Opinion Quarterly* 21:521–42.

Micceri, T. 1989. "The Unicorn, the Normal Curve, and Other Improbable Creatures." *Psychological Bulletin* 155:155–66.

Mohr, Lawrence B. 1990. *Understanding Significance Testing.* Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-073. Newbury Park, CA: Sage.

Mooney, Christopher Z. 1994. "Constructing Bootstrap Confidence Intervals Using GAUSS." *The Political Methodologist* 6:27–9.

———. N.d. *Monte Carlo Simulation.* Sage University Paper Series on Quantitative Applications in the Social Sciences, forthcoming. Newbury Park, CA: Sage.

Mooney, Christopher Z., and Robert D. Duval. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference.* Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-095. Newbury Park, CA: Sage.

Mooney, Christopher Z., and George Krause. 1996. "Of Silicon and Political Science: Computationally Intensive Techniques of Statistical Estimation and Inference." *British Journal of Political Science.* Forthcoming.

Navidi, W. 1989. "Edgeworth Expansion for Bootstrapping Regression Models." *Annals of Statistics* 17:1472–8.

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

Norrander, Barbara. 1993. "Nomination Choices: Caucus and Primary Outcomes, 1976–1988." *American Journal of Political Science* 37:343–64.

Oakes, Michael. 1986. *Statistical Inference: A Commentary for the Social and Behavioural Sciences.* New York: John Wiley.

Pearson, Egon S. 1931. "The Analysis of Variance in Cases of Non-Normal Variation." *Biometrika* 23:114–33.

Politis, Dimitris N., and Joseph P. Romano. 1992. "A Circular Block-Resampling Procedure for Stationary Data." In *Exploring the Limits of Bootstrap,* ed. Raoul LePage and Lynne Billard. New York: John Wiley.

Poole, Keith T., and Howard Rosenthal. 1991. "Patterns of Congressional Voting." *American Journal of Political Science* 35:228–78.

Rao, B. L. S. Prakasa. 1987. *Asymptotic Theory of Statistical Inference.* New York: Wiley.

Ringquist, Evan J. 1993. "Does Regulation Matter?: Evaluating the Effects of State Air Pollution Control Programs." *Journal of Politics* 55:1022–45.

Rohatgi, Vijay K. 1984. *Statistical Inference.* New York: Wiley.

Sankoff, David, and Koula Mellos. 1972. "The Swing Ratio and Game Theory." *American Political Science Review* 66:551–4.

Schenker, Nathaniel. 1985. "Qualms About Bootstrap Confidence Intervals." *Journal of the American Statistical Association* 80:360–1.

Schrodt, Philip A. 1982. "A Statistical Study of the Cube Law in Five Electoral Systems." *Political Methodology* 7:31–53.

Singh, Kesar. 1981. "On the Asymptotic Accuracy of Efron's Bootstrap." *Annals of Statistics* 9:1187–95.

Songer, Donald R., and Susan Haire. 1992. "Integrating Alternative Approaches to the Study of Judicial Voting: Obscenity Cases in the U.S. Courts of Appeals." *American Journal of Political Science* 36:963–82.

Srivastava, A. B. L. 1958. "The Effects of Non-Normality on the Power Function of the *t*-Test." *Biometrika* 45:421–9.

Stine, Robert A. 1990. "An Introduction to Bootstrap Methods." *Sociological Methods and Research* 18:243–91.

Taagepera, Rein. 1986. "Reformulating the Cube Law for Proportional Representation Elections." *American Political Science Review* 80:489–504.

Theil, Henri. 1969. "The Desired Political Entropy." *American Political Science Review* 63:521–5.

Thombs, L. A., and W. R. Schucany. 1990. "Bootstrap Prediction Intervals for Autoregression." *Journal of the American Statistical Society* 85:486–92.

Tufte, Edward R. 1973. "The Relationship Between Seats and Votes in Two-Party Systems." *American Political Science Review* 67:540–7.

Tukey, John W. 1975. "Instead of Gauss-Markov Least Squares, What?" In *Applied Statistics,* ed. R. P. Gupta. New York: American Elsevier.

———. 1986. "Sunset Salvo." *The American Statistician* 40:72–6.

White, Halbert. 1980. "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity." *Econometrica* 48:817–38.