COMPUTER INTENSIVE METHODS IN STATISTICS

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 150

MAY 1980

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS77-16974

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
COMPUTER INTENSIVE METHODS IN STATISTICS

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 150

MAY 1980

PREPARED UNDER THE AUSPICES OF
NATIONAL SCIENCE FOUNDATION GRANT

MPS 77-16974

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
COMPUTER INTENSIVE METHODS IN STATISTICS

Bradley Efron

Dedicated to the Academy of Sciences of Lisbon, on the Occasion of their Bicentenary

Abstract

Four nonparametric methods, the jackknife, the delta method, the bootstrap, and the smoothed bootstrap, are compared for estimating the standard deviation of the correlation coefficient. The numerical results motivate some speculation on the future of statistical theory.
COMPUTER INTENSIVE METHODS IN STATISTICS

Bradley Efron

Dedicated to the Academy of Sciences of Lisbon, on the Occasion of their Bicentenary

1. Introduction

The Academy has invited us to speak on the theme "Recent Advances in Statistics." The time is ripe for such a discussion. After a decade of drift, following the postwar heyday of Neyman-Wald decision theory, statistics seems to be gathering momentum for another move forward. The line of advance involves the electronic computer, which is now not only very fast, but very cheap.

Most of our commonly used statistical methods were developed during a period when computation was slow and expensive. I am thinking now not only of such obvious examples as analysis of variance and chi-square goodness of fit, but of the entire parametric theory developed in this century, running say from Fisher's original papers of the 1920's to Lehmann's influential book on hypothesis testing (1959). The honest answer to "Why assume normality?" is most often "Because then we can compute the answer." One would expect, and hope, that the ability to compute more ambitious answers would lead to more ambitious theories.

Consider, for example, a simple analysis of variance situation. We wish to test the null hypothesis that a data vector \( \mathbf{X} = (X_1, X_2, \ldots, X_n) \) has mean vector \( \mathbf{\mu} = 0 \), versus the alternative that \( \mathbf{\mu} \) lies in \( \mathcal{F} \), a
predetermined subspace of Euclidean n-space, containing \((1, 1, \ldots, 1)\), the main diagonal. The standard approach compares the angle \(A\) between \(X\) and \(L\) with a standard number \(A_0\), and rejects the null hypothesis if \(A\) is smaller than \(A_0\). The crucial constant \(A_0\) is determined by normal theory, which is the only distribution for the components \(X_i\) which allows its theoretical calculation. (See Efron, 1969.)

Suppose though that the scientist's preferred null hypothesis is the one-sided exponential \(f(x) = e^{-(x-1)}, x > -1\), rather than \(f(x) = e^{-x^2/2}/\sqrt{2\pi}\). These days there is nothing to stop the statistician from going to the computer and calculating, by Monte Carlo methods, the constant \(A_0\) appropriate for this situation. Essentially the statistician would be replacing a universal table of constants, that of the "F distribution", with a special table constructed to fit the situation at hand. Most of us don't use printed tables of the exponential or logarithm any more because it is easier for our calculators to compute the function in each particular case. I am suggesting the same fate for the F tables, at a more profound level of replacement.

We can carry the analysis of variance example still further away from parametric theory. Consider the residual vector \(R = (R_1, R_2, \ldots, R_n)\) (which is the projection of \(X\) into the space orthogonal to \(L\)) and the empirical distribution \(\hat{F}\) putting mass \(1/n\) at each value of \(R_i, i=1, 2, \ldots, n\). The statistician can ascertain the constant \(A_0\) by Monte Carlo sampling from \(\hat{F}\). In other words, if \(X^* = (X_1^*, X_2^*, \ldots, X_n^*)\) has \(X_i^* iid \sim \hat{F}, i=1, \ldots, n\), \(A_0\) is that value such that \(A^*\), the angle between \(X^*\) and \(L\), is less than \(A_0\) 5\% (or whatever significance level is desired) of the time.
If this suggestion seems bizarre, it is worth noting how close it remains to standard normal theory. Instead of taking $F$ to be a nonparametric estimate of distribution for the components $X_i$, normal theory takes $\hat{F} \sim N(0, \hat{\sigma}^2)$, where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} R_i^2}{n}$ (or, equivalently for the problem at hand, $\hat{\sigma}^2 = \frac{\bar{R}_1^2}{(n-p)}$, $p$ the dimension of $\mathcal{L}$). The method for determining $A_0$, from this point on, is exactly the same as that described in the previous paragraph. Of course normal theory allows $A_0$ to be determined mathematically rather than by Monte Carlo, which is both its virtue and limitation. Freedman (1975) uses a close cousin of the nonparametric method to successfully analyze a complicated data set.

The remarks so far seem to indicate a statistical world in which mathematical theory is replaced by brute computation. In fact, computational advances precipitate new theory. An example from the last century is matrix theory and multidimensional geometry, which led, 50 years later, to Fisher's development of the analysis of variance. Efron (1979b) discusses several current areas of theoretical development which were impractical before the modern computer era: robust regression, the EM algorithm for maximum likelihood estimation from incomplete data sets, Cox's partial likelihood for censored data, cross-validation, the jackknife, and the bootstrap. Here we take up the latter two topics again. The discussion, which is limited to a very special situation, is intended to illustrate how a more computer-based theory might function in practice, and what new theoretical problems of its own it might raise.

2. A Simple Problem

We will consider the following very special problem. Fourteen bivariate points $x_1, x_2, x_3, \ldots, x_{14}$ independently drawn from the same
bivariate distribution $F$ are observed, and the Pearson sample correlation coefficient

$$
\hat{\rho} = \frac{\frac{14}{\sqrt{\sum_{i=1}^{14} (x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{14} (x_{i2} - \bar{x}_2)^2}}}{\sum_{i=1}^{14} (x_{i1} - \bar{x}_1) (x_{i2} - \bar{x}_2)}
$$

computed, where $\bar{x}_1 = \frac{1}{14} \sum_{i=1}^{14} x_{i1}$, $\bar{x}_2 = \frac{1}{14} \sum_{i=1}^{14} x_{i2}$. It is desired to estimate the standard deviation $SD(\hat{\rho})$. Five different standard deviation estimates will be investigated. A brief description of the five follows, listed in order of increasing computational difficulty.

(i) The normal theory estimate

$$
\hat{SD}_1 = (1 - \hat{\rho}^2)^{1/2}
$$

(2.1)

See Kendall and Stuart (1958), page 236.

(ii) Tukey’s jackknife estimate. Let $\hat{\rho}_{(i)}$ be the value of $\hat{\rho}$ calculated from the 13 data points $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{14}$, the full data set with $x_i$ removed, and define $\hat{\rho}_{(*)} = \sum_{i=1}^{14} \hat{\rho}_{(i)} / 14$. Then

$$
\hat{SD}_2 = \sqrt{\frac{1}{14} \sum_{i=1}^{14} (\hat{\rho}_{(i)} - \hat{\rho}_{(*)})^2}
$$

(2.2)

See Miller (1974).

(iii) The delta method, or infinitesimal jackknife. Let $\mu_{gh} = E_F(X_1 - E_F X_1)^g (X_2 - E_F X_2)^h$, the $gh$th mixed central moment of the random point $X = (X_1, X_2)$ under the true distribution $F$. By using Taylor series expansions, $\hat{\rho}$ can be approximated by a polynomial in the
sample moments of \( x_1, x_2, \ldots, x_{14} \). The standard deviation of this polynomial is expressed in terms of the \( \mu_{gh} \), see Cramer (1946), p. 359, and then estimated by substitution of the sample central moments \( \hat{\mu}_{gh} \) to give the standard deviation estimate

\[
\hat{SD}_3 = \sqrt{\frac{2}{56} \left( \frac{\hat{\mu}_{40}}{\hat{\mu}_{20}^2} + \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} + \frac{2\hat{\mu}_{22}}{\hat{\mu}_{02}^2} + \frac{4\hat{\mu}_{22}}{\hat{\mu}_{11}^2} + \frac{4\hat{\mu}_{31}}{\hat{\mu}_{11} \hat{\mu}_{20}} + \frac{4\hat{\mu}_{13}}{\hat{\mu}_{11} \hat{\mu}_{02}} \right)}.
\] (2.3)

It turns out that delta method standard deviation estimates can always be calculated in a simpler manner, almost identical to that of the jackknife. (A proof will appear in a forthcoming paper by the author.)

If \( \hat{F} \) represents the empirical distribution putting mass 1/14 at each point \( x_1, x_2, \ldots, x_{14} \), let \( \hat{F}_{(i)}^E \) be the distribution putting mass \((1-\epsilon)/14\) at \( x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{14} \) and mass \((1+13\epsilon)/14\) at \( x_i \); also let \( \hat{\rho}_{(i)}^E \) be the Pearson correlation coefficient corresponding to \( \hat{F}_{(i)}^E \). Then

\[
\hat{SD}_3 = \lim_{\epsilon \to 0} \sqrt{\sum_{i=1}^{14} \left( \frac{\hat{F}_{(i)}^E - \hat{\rho}}{\epsilon} \right)^2} \quad / \quad 14 \quad .
\] (2.3A)

This last expression is Jaeckel's "infinitesimal jackknife" estimate of standard deviation, see Miller (1974), page 10. Notice that taking \( \epsilon = -1/13 \) gives, almost, the jackknife estimate \( \hat{SD}_2 \). In practice \( \hat{SD}_3 \) can be calculated on the computer by using some sufficiently small value of \( \epsilon \). (The value \( \epsilon = .001 \) was used in the simulation study of Section 3.) The advantage of expression (2.3A) over (2.3) is that the sometimes tedious algebra of the delta method is avoided. Expression (2.3A) also
appears in the literature as the root mean square of the empirical influence curve, see Huber (1977), page 25.

(iv) The bootstrap. The sample correlation coefficient \( \hat{\rho} \) can be thought of as \( \hat{\rho} = \rho(\hat{F}) \), where \( \rho(\cdot) \) is the functional mapping a bivariate distribution function \( F \) into its correlation coefficient \( \rho \). Another functional of interest is \( \sigma(F) \), the standard deviation of \( \hat{\rho} \) obtained from data \( X_1, X_2, \ldots, X_{14} \overset{\text{iid}}{\sim} F \). The bootstrap estimate of standard deviation, Efron (1979a), is simply

\[
\hat{SD}_4 = \sigma(\hat{F}).
\]  

(2.4)

In other words \( \hat{SD}_4 \) is the nonparametric maximum likelihood estimate of \( \sigma(F) \). It is related to \( \sigma(F) \) exactly as \( \hat{\rho} = \rho(\hat{F}) \) is related to \( \rho = \rho(F) \).

Except for unusually favorable situations, not including the present case, \( \hat{SD}_4 \) must be computed by a Monte Carlo algorithm:

(1) Construct the empirical distribution \( \hat{F} \) putting mass 1/14 at each point \( x_1, x_2, \ldots, x_{14} \).

(2) With \( \hat{F} \) fixed, draw a random sample of size 14 from \( \hat{F} \), say \( X_1^*, X_2^*, \ldots, X_{14}^* \overset{\text{iid}}{\sim} F \), (i.e. \( X_1^*, \ldots, X_{14}^* \) is a simple random sample drawn with replacement from \( \{x_1, x_2, \ldots, x_{14}\} \)).

(3) Let \( \hat{F}^* \) be the empirical distribution of \( X_1^*, X_2^*, \ldots, X_{14}^* \), and \( \hat{\rho}^* = \rho(\hat{F}^*) \). Repeat step (2) independently \( N \) times, obtaining bootstrap replications \( \hat{\rho}^1, \hat{\rho}^2, \ldots, \hat{\rho}^N \). Approximate \( \hat{SD}_4 \) by

\[
\hat{SD}_4 = \sqrt{\frac{1}{\sum_{j=1}^{N} [\hat{\rho}^* - \hat{\rho}^j]^2 / (N-1)}},
\]  

(2.4A)

where \( \hat{\rho}^* = \frac{1}{N} \sum_{j=1}^{N} \hat{\rho}^j / N \). The values \( N = 128 \) and \( N = 512 \) were used in the simulation study of Section 3.
(v) Smoothed bootstrap. The bootstrap algorithm defined above starts from \( \hat{F} \), the nonparametric maximum likelihood estimate for the true \( F \). The statistician may be willing to assume that \( F \), whatever it is, is reasonably smooth. In this case the bootstrap algorithm might better be started from a smoothed version of \( \hat{F} \), say the convolution of \( \hat{F} \) with some smooth distribution \( \hat{W} \). In the simulation study of Section 3, \( \hat{W} \) was taken to be a uniform distribution over a rhombus \( \tilde{\mathcal{R}} \), where \( \tilde{\mathcal{R}} \) was selected so that \( \hat{W} \) had covariance matrix .5 times \( \hat{\rho} \), the sample covariance matrix of \( x_1, x_2, \ldots, x_{14} \). The points \( x_1^*, x_2^*, \ldots, x_{14}^* \) were then selected independently from \( \hat{F} \star \hat{W} \), and \( \hat{SD}_5 \) calculated as in (2.4A).

It would be more natural to take \( \hat{W} \) bivariate normal. However this would be self-serving in the context of Section 3, since the simulation study selected the original points \( x_1, x_2, \ldots, x_{14} \) themselves from a bivariate normal distribution.

Notice that \( \hat{F} \star \hat{W} \) has the same correlation coefficient as \( \hat{\rho} \), namely \( \hat{\rho} \). If the covariance matrix of \( \hat{W} \) were not taken proportional to \( \hat{\rho} \) this would not be true, and a bias would be introduced into the smoothed bootstrap procedure.

3. A Simulation Study

The study involved 200 trials\( ^\dagger \) of the following situations

\[
X_1, X_2, \ldots, X_{14} \overset{iid}{\sim} h_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}\right).
\]

---

\( ^\dagger \) In what follows, "trial" refers to a new selection of the data points \( X_1 = x_1, X_2 = x_2, \ldots, X_{14} = x_{14} \); "replication" refers to the bootstrap selection of \( X_1^*, X_2^*, \ldots, X_{14}^* \), with \( x_1, x_2, \ldots, x_{14} \) held fixed, as described in Section 2.
In this case it can be shown that \( \hat{\rho} \) has expected value \(.486\), standard deviation \(.221\),

\[
\hat{\rho} \sim (0.486, 0.221^2)
\]  \(3.2\)

see Johnson and Kotz (1970), page 225.

### Summary Statistics, 200 trials

<table>
<thead>
<tr>
<th>Method</th>
<th>AVE</th>
<th>SD</th>
<th>CV</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Jackknife</td>
<td>.223</td>
<td>.085</td>
<td>.38</td>
<td>.085</td>
</tr>
<tr>
<td>3. Delta Method</td>
<td>.175</td>
<td>.058</td>
<td>.33</td>
<td>.074</td>
</tr>
<tr>
<td>(Infinitesimal Jackknife)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Bootstrap</td>
<td>.206</td>
<td>.066</td>
<td>.32</td>
<td>.068</td>
</tr>
<tr>
<td>(N = 128)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N = 512)</td>
<td>.206</td>
<td>.063</td>
<td>.31</td>
<td>.065</td>
</tr>
<tr>
<td>5. Smoothed Bootstrap</td>
<td>.205</td>
<td>.061</td>
<td>.30</td>
<td>.063</td>
</tr>
<tr>
<td>(Rhomboid window, (N = 128))</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. A comparison of different methods for estimating \( \text{SD}(\hat{\rho}) \), where \( \hat{\rho} \) is the sample correlation coefficient for \( x_1, x_2, \ldots, x_{14} \) generated from \((3.1)\). The true value is \( \text{SD}(\hat{\rho}) = .221 \). *Relative Bias \( \geq .10 \), **Relative Bias \( \geq .20 \).

Table 1 shows summary statistics for the 200 trials, for each of the methods described in Section 2. For example the 200 values of

\[
\hat{\text{SD}}_1 = (1-\hat{\rho})^2/\sqrt{14}
\]

averaged \(.192\), with sample standard deviation \(.050\).

The sample coefficient of variation was \(.050/\cdot192 = .26\), and the root mean square error, from the true value \(.221\), equaled 

\[
[(.221 - .192)^2 + (199/200)(.050)^2]^{1/2} = .058.
\]

One asterisk indicates a relative bias \( \geq .10 \), e.g. \(|(.192 - .221)/.221| = .13\); two asterisks indicate relative bias \( \geq .20 \). This notation
reflects the author's opinion that biases smaller than 10% are negligible in this context. In any case, it should be noticed that the bias situation would look considerably different if the summary statistics were given in terms of variances rather than standard deviations. The true variance is \( \text{VAR}(\hat{\rho}) = .221^2 = .0488 \). The averages, over 200 trials, of the estimated variances were .0394 (normal parametric), .0569 (jackknife), .0340 (delta method), .0468 (bootstrap 128), .0464 (bootstrap 512), and .0457 (smoothed bootstrap). The reader is referred to Efron and Stein (1978) for a proof that the jackknife estimate of variance is always biased upward in a certain specific sense. Among the nonparametric estimates of standard deviation, there is noticeably improved performance as we move in the direction of increasing computation, from the top toward the bottom of the table. The smoothed bootstrap has root mean square error less than 75% of that of the jackknife. Of course the normal parametric method is better still, but that is no surprise since the simulation involved only normal distributions.

<table>
<thead>
<tr>
<th>Summary Statistics, 200 trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVE</td>
</tr>
<tr>
<td>--------------------------------</td>
</tr>
<tr>
<td>1. Normal Parametric (= 1/\sqrt{11} = .302)</td>
</tr>
<tr>
<td>2. Jackknife</td>
</tr>
<tr>
<td>3. Delta Method (Infinitesimal Jackknife)</td>
</tr>
<tr>
<td>4. Bootstrap N = 128</td>
</tr>
<tr>
<td>N = 512</td>
</tr>
<tr>
<td>5. Smoothed Bootstrap (Rhomboid window, N = 128)</td>
</tr>
</tbody>
</table>

Table 2. A comparison of different methods for estimating \( \text{SD}(\tanh^{-1} \hat{\rho}) \), where \( \hat{\rho} \) is the sample correlation coefficient for \( x_1, x_2, \ldots, x_{14} \) generated from (3.1). The true value is \( \text{SD}(\tanh^{-1} \hat{\rho}) = .299. * \) Relative bias \( \geq .10. \)
Table 2 is the same as Table 1, except that the quantity of interest is \( SD(\tanh^{-1} \hat{\rho}) \) rather than \( SD(\hat{\rho}) \). Fisher's transformation \( \tanh^{-1} \hat{\rho} = (1/2) \log(1 + \hat{\rho})/(1 - \hat{\rho}) \) is variance stabilizing, that is it gives a (nearly) constant standard deviation, \( 1/\sqrt{11} = .302 \), when sampling from a bivariate normal distribution. Once again the nonparametric estimates of standard deviation improve as we go from the jackknife to the smoothed bootstrap.

4. Discussion

Tables 1 and 2 raise many theoretical questions, most of which have not yet been solved.

(i) How much computation should one invest in a given situation? For estimating either \( SD(\hat{\rho}) \) or \( SD(\tanh^{-1} \hat{\rho}) \), it was definitely worthwhile to go from the jackknife, which requires 14 recalculations of \( \hat{\rho} \), to the bootstrap with \( N = 128 \) replications. Multiplying \( N \) by 4, to \( N = 512 \), gave little additional improvement. In the present case it can be shown that further increase of \( N \) would be pointless, at least for the estimation of standard deviations. (\( N = \infty \) would give \( SD = .062 \) for the bootstrap in Table 1, and \( SD = .061 \) for the bootstrap in Table 2.) In a real situation, as opposed to a Monte Carlo simulation, it is more difficult to choose \( N \), though standard sampling calculations give a rough idea of when to stop.

(ii) What is the best way to invest computational effort? In the author's experience the bootstrap nearly always outperforms the jackknife. Sometimes though the improvement is slight, and not worth the additional computational expense. There are many other resampling plans which have been suggested, see Remark I of Efron (1979a). All of these are closely
related to the bootstrap. (For example the jackknife can be explained as a delta method approximation to the bootstrap, as in Section 5 of Efron (1979a).) However they may have different, perhaps better, properties in a given situation. The bootstrap, because of its maximum likelihood derivation, enjoys some theoretical advantages, which so far have been verified in the author's modest comparative studies, but the issue is by no means settled.

(iii) What is the relation with parametric methods? The parametric method described in Section 2 is itself a version of the bootstrap. The only conceptual difference is that at Step 1 of the bootstrap algorithm, \( \hat{F} \) is taken to be the parametric maximum likelihood estimation of \( F \), i.e. the bivariate normal with first and second moments matching those of the sample, rather than the nonparametric MLE. Of course the computation of the standard deviation estimate can then be carried out theoretically, rather than by Monte Carlo.

What would we do in a practical situation, where typically we would have no strong grounds for either believing or disbelieving the bivariate normal model? One approach is to use the smoothed bootstrap, sampling from \( \hat{F} \ast \hat{W} \) where \( \hat{W} \sim \mathcal{N}_2(0, \hat{C}) \), trying different values of \( c \). The choice \( c=0 \) gives the unsmoothed bootstrap, while \( c=\infty \) gives the normal parametric estimate of standard deviation. Once again, there is no theory to rely on. See Sections 3 of Efron (1979 a,b).

(iv) How should confidence intervals be determined? Computing a standard deviation is a rough substitute for the more ambitious problem of assigning a two-sided confidence interval. Tukey's original suggestion was to use the jackknife and the standard t-table to assign confidence
intervals of the form $\hat{\rho} \pm t_{\alpha} \cdot \hat{SD}_2$. However there is no theoretical support for this approach, at least not for small enough samples sizes to make the $t$ intervals much different than the normal intervals $\hat{\rho} \pm z_{\alpha} \cdot \hat{SD}_2$. Efron (1979c) has suggested a method of constructing confidence intervals using the entire bootstrap distribution $\hat{\rho}^*1, \hat{\rho}^*2, \ldots, \hat{\rho}^*N$. This method requires $N$ to be larger than that for estimating the standard deviation, on the order of $N = 1000$. A theoretical argument involving transformations supports this method, but also suggests other approaches.

5. Conclusions

The next twenty years should be exciting ones for statisticians. My prediction is for a partial replacement of parametric models, and the accompanying mathematical calculations we have gotten used to, by computer intensive methods. These methods will replace "theory from a book", typified by t-tables and F-tables, by "theory from scratch", generated anew by the computer for each new data analysis problem. We theoreticians won't be put out of work though. Statistical computer methods themselves are perfectly worthy objects of mathematical attention. Their analysis, comparison, and refinement is a formidable prospect for the statistical theorists of the late 20th century.
References


