Statistical Science
1986, Vol. 1, No. 1, 54–77

Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy

B. Efron and R. Tibshirani

Abstract. This is a review of bootstrap methods, concentrating on basic ideas and applications rather than theoretical considerations. It begins with an exposition of the bootstrap estimate of standard error for one-sample situations. Several examples, some involving quite complicated statistical procedures, are given. The bootstrap is then extended to other measures of statistical accuracy such as bias and prediction error, and to complicated data structures such as time series, censored data, and regression models. Several more examples are presented illustrating these ideas. The last third of the paper deals mainly with bootstrap confidence intervals.

Key words: Bootstrap method, estimated standard errors, approximate confidence intervals, nonparametric methods.

1. INTRODUCTION

A typical problem in applied statistics involves the estimation of an unknown parameter θ. The two main questions asked are (1) what estimator $\hat{\theta}$ should be used? (2) Having chosen to use a particular $\hat{\theta}$, how accurate is it as an estimator of θ? The bootstrap is a general methodology for answering the second question. It is a computer-based method, which substitutes considerable amounts of computation in place of theoretical analysis. As we shall see, the bootstrap can routinely answer questions which are far too complicated for traditional statistical analysis. Even for relatively simple problems computer-intensive methods like the bootstrap are an increasingly good data analytic bargain in an era of exponentially declining computational costs.

This paper describes the basis of the bootstrap theory, which is very simple, and gives several examples of its use. Related ideas like the jackknife, the delta method, and Fisher's information bound are also discussed. Most of the proofs and technical details are omitted. These can be found in the references given, particularly Efron (1982a). Some of the discussion here is abridged from Efron and Gong (1983) and also from Efron (1984).

Before beginning the main exposition, we will describe how the bootstrap works in terms of a problem where it is not needed, assessing the accuracy of the sample mean. Suppose that our data consists of a random sample from an unknown probability distribution F on the real line,

$$(1.1) \qquad X_1, X_2, \cdots, X_n \sim F.$$

Having observed $X_1 = x_1$, $X_2 = x_2$, $\cdots$, $X_n = x_n$, we compute the sample mean $\bar{x} = \sum_1^n x_n/n$, and wonder how accurate it is as an estimate of the true mean $\theta = E_F\{X\}$.

If the second central moment of F is $\mu_2(F) \equiv E_F X^2 - (E_F X)^2$, then the standard error $\sigma(F; n, \bar{x})$, that is the standard deviation of $\bar{x}$ for a sample of size n from distribution F, is

$$(1.2) \qquad \sigma(F) = [\mu_2(F)/n]^{1/2}.$$

The shortened notation $\sigma(F) \equiv \sigma(F; n, \bar{x})$ is allowable because the sample size n and statistic of interest $\bar{x}$ are known, only F being unknown. The standard error is the traditional measure of $\bar{x}$'s accuracy. Unfortunately, we cannot actually use (1.2) to assess the accuracy of $\bar{x}$, since we do not know $\mu_2(F)$, but we can use the *estimated standard error*

$$(1.3) \qquad \bar{\sigma} = [\bar{\mu}_2/n]^{1/2}.$$

where $\bar{\mu}_2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$, the unbiased estimate of $\mu_2(F)$.

There is a more obvious way to estimate $\sigma(F)$. Let

B. Efron is Professor of Statistics and Biostatistics, and Chairman of the Program in Mathematical and Computational Science at Stanford University. His mailing address is Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305. R. Tibshirani is a Postdoctoral Fellow in the Department of Preventive Medicine and Biostatistics, Faculty of Medicine, University of Toronto, McMurrick Building, Toronto, Ontario, M5S 1A8, Canada.

$\hat{F}$ indicate the empirical probability distribution,

(1.4) $\hat{F}$: probability mass $1/n$ on $x_1, x_2, \cdots, x_n$.

Then we can simply replace F by $\hat{F}$ in (1.2), obtaining

$$\hat{\sigma} \equiv \sigma(\hat{F}) = [\mu_2(\hat{F})/n]^{1/2}, \tag{1.5}$$

as the estimated standard error for $\bar{x}$. This is the *bootstrap estimate*. The reason for the name "bootstrap" will be apparent in Section 2, when we evaluate $\sigma(\hat{F})$ for statistics more complicated than $\bar{x}$. Since

$$\hat{\mu}_2 \equiv \mu_2(\hat{F}) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}, \tag{1.6}$$

$\hat{\sigma}$ is not quite the same as $\bar{\sigma}$, but the difference is too small to be important in most applications.

Of course we do not really need an alternative formula to (1.3) in this case. The trouble begins when we want a standard error for estimators more complicated than $\bar{x}$, for example, a median or a correlation or a slope coefficient from a robust regression. In most cases there is no equivalent to formula (1.2), which expresses the standard error $\sigma(F)$ as a simple function of the sampling distribution F. As a result, formulas like (1.3) do not exist for most statistics.

This is where the computer comes in. It turns out that we can always numerically evaluate the bootstrap estimate $\hat{\sigma} = \sigma(\hat{F})$, without knowing a simple expression for $\sigma(F)$. The evaluation of $\hat{\sigma}$ is a straightforward Monte Carlo exercise described in the next section. In a good computing environment, as described in the remarks in Section 2, the bootstrap effectively gives the statistician a simple formula like (1.3) for any statistic, no matter how complicated.

Standard errors are crude but useful measures of statistical accuracy. They are frequently used to give approximate confidence intervals for an unknown parameter θ

$$\theta \in \hat{\theta} \pm \hat{\sigma} z^{(\alpha)}, \tag{1.7}$$

where $z^{(\alpha)}$ is the $100 \cdot \alpha$ percentile point of a standard normal variate, e.g., $z^{(.95)} = 1.645$. Interval (1.7) is sometimes good, and sometimes not so good. Sections 7 and 8 discuss a more sophisticated use of the bootstrap, which gives better approximate confidence intervals than (1.7).

The standard interval (1.7) is based on taking literally the large sample normal approximation $(\hat{\theta} - \theta)/\hat{\sigma} \sim N(0, 1)$. Applied statisticians use a variety of tricks to improve this approximation. For instance if θ is the correlation coefficient and $\hat{\theta}$ the sample correlation, then the transformation $\phi = \tanh^{-1}(\theta)$, $\hat{\phi} = \tanh^{-1}(\hat{\theta})$ greatly improves the normal approximation, at least in those cases where the underlying sampling distribution is bivariate normal. The correct tactic then is to transform, compute the interval (1.7) for ϕ, and transform this interval back to the θ scale.

We will see that bootstrap confidence intervals can automatically incorporate tricks like this, without requiring the data analyst to produce special techniques, like the $\tanh^{-1}$ transformation, for each new situation. An important theme of what follows is the substitution of raw computing power for theoretical analysis. This is not an argument against theory, of course, only against unnecessary theory. Most common statistical methods were developed in the 1920s and 1930s, when computation was slow and expensive. Now that computation is fast and cheap we can hope for and expect changes in statistical methodology. This paper discusses one such potential change, Efron (1979b) discusses several others.

2. THE BOOTSTRAP ESTIMATE OF STANDARD ERROR

This section presents a more careful description of the bootstrap estimate of standard error. For now we will assume that the observed data $\mathbf{y} = (x_1, x_2, \cdots, x_n)$ consists of independent and identically distributed (iid) observations $X_1, X_2, \cdots, X_n \sim_{\text{iid}} F$, as in (1.1). Here F represents an unknown probability distribution on $\mathscr{X}$, the common sample space of the observations. We have a statistic of interest, say $\hat{\theta}(\mathbf{y})$, to which we wish to assign an estimated standard error.

Fig. 1 shows an example. The sample space $\mathscr{X}$ is $\mathbb{R}^{2+}$, the positive quadrant of the plane. We have observed $n = 15$ bivariate data points, each corresponding to an American law school. Each point x_i consists of two summary statistics for the 1973 entering class at law school i

$$x_i = (\text{LSAT}_i, \text{GPA}_i); \tag{2.1}$$

LSAT$_i$ is the class' average score on a nationwide exam called "LSAT"; GPA$_i$ is the class' average undergraduate grades. The observed Pearson correlation

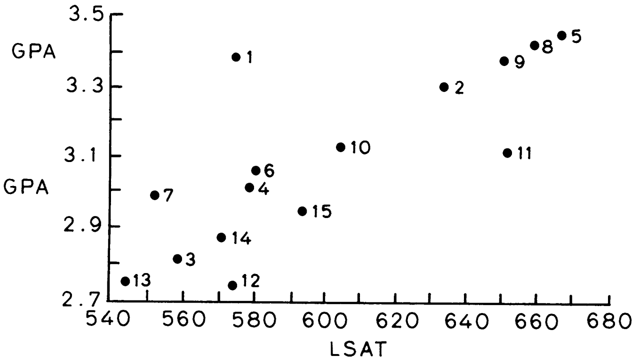

FIG. 1. *The law school data (Efron, 1979b). The data points, beginning with School 1, are* (576, 3.39), (635, 3.30), (558, 2.81), (578, 3.03), (666, 3.44), (580, 3.07), (555, 3.00), (661, 3.43), (651, 3.36), (605, 3.13), (653, 3.12), (575, 2.74), (545, 2.76), (572, 2.88), (594, 2.96).

coefficient for these 15 points is $\hat{\theta} = .776$. We wish to assign a standard error to this estimate.

Let $\sigma(F)$ indicate the standard error of $\hat{\theta}$, as a function of the unknown sampling distribution F,

$$\sigma(F) = [\mathrm{Var}_F\{\hat{\theta}(\mathbf{y})\}]^{1/2}. \tag{2.2}$$

Of course $\sigma(F)$ is also a function of the sample size n and the form of the statistic $\hat{\theta}(\mathbf{y})$, but since both of these are known they need not be indicated in the notation. The bootstrap estimate of standard error is

$$\hat{\sigma} = \sigma(\hat{F}), \tag{2.3}$$

where $\hat{F}$ is the empirical distribution (1.4), putting probability $1/n$ on each observed data point x_i. In the law school example, $\hat{F}$ is the distribution putting mass $1/15$ on each point in Fig. 1, and $\hat{\sigma}$ is the standard deviation of the correlation coefficient for 15 iid points drawn from $\hat{F}$.

In most cases, including that of the correlation coefficient, there is no simple expression for the function $\sigma(F)$ in (2.2). Nevertheless, it is easy to numerically evaluate $\hat{\sigma} = \sigma(\hat{F})$ by means of a Monte Carlo algorithm, which depends on the following notation: $\mathbf{y}^* = (x_1^*, x_2^*, \cdots, x_n^*)$ indicates n independent draws from $\hat{F}$, called a *bootstrap sample*. Because $\hat{F}$ is the empirical distribution of the data, a bootstrap sample turns out to be the same as a random sample of size n drawn *with replacement* from the actual sample $\{x_1, x_2, \cdots, x_n\}$.

The Monte Carlo algorithm proceeds in three steps: (i) using a random number generator, independently draw a large number of bootstrap samples, say $\mathbf{y}^*(1)$, $\mathbf{y}^*(2), \cdots, \mathbf{y}^*(B)$; (ii) for each bootstrap sample $\mathbf{y}^*(b)$, evaluate the statistic of interest, say $\hat{\theta}^*(b) = \hat{\theta}(\mathbf{y}^*(b))$, $b = 1, 2, \cdots, B$; and (iii) calculate the sample standard deviation of the $\hat{\theta}^*(b)$ values

$$\begin{aligned} \hat{\sigma}_B &= \left(\frac{\sum_{b=1}^B \{\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)\}^2}{B-1}\right)^{1/2}, \\ \hat{\theta}^*(\cdot) &= \frac{\sum_{b=1}^B \hat{\theta}^*(b)}{B}. \end{aligned} \tag{2.4}$$

It is easy to see that as $B \to \infty$, $\hat{\sigma}_B$ will approach $\hat{\sigma} = \sigma(\hat{F})$, the bootstrap estimate of standard error. All we are doing is evaluating a standard deviation by Monte Carlo sampling. Later, in Section 9, we will discuss how large B need be taken. For most situations B in the range 50 to 200 is quite adequate. In what follows we will usually ignore the difference between $\hat{\sigma}_B$ and $\hat{\sigma}$, calling both simply "$\hat{\sigma}$."

Why is each bootstrap sample taken with the same sample size n as the original data set? Remember that $\sigma(F)$ is actually $\sigma(F, n, \hat{\theta})$, the standard error for the statistic $\hat{\theta}(\ )$ based on a random sample of size n from the unknown distribution F. The bootstrap estimate $\hat{\sigma}$ is actually $\sigma(F, n, \hat{\theta})$ evaluated at $F = \hat{F}$. The Monte Carlo algorithm will not converge to $\hat{\sigma}$ if the bootstrap sample size differs from the true n. Bickel and Freedman (1981) show how to correct the algorithm to give $\hat{\sigma}$ if in fact the bootstrap sample size is taken different than n, but so far there does not seem to be any practical advantage to be gained in this way.

Fig. 2 shows the histogram of $B = 1000$ bootstrap replications of the correlation coefficient from the law school data. For convenient reference the abscissa is plotted in terms of $\hat{\theta}^* - \hat{\theta} = \hat{\theta}^* - .776$. Formula (2.4) gives $\hat{\sigma} = .127$ as the bootstrap estimate of standard error. This can be compared with the usual normal theory estimate of standard error for $\hat{\theta}$,

$$\hat{\sigma}_{\mathrm{NORM}} = (1 - \hat{\theta}^2)/(n-3)^{1/2} = .115, \tag{2.5}$$

[Johnson and Kotz (1970, p. 229)].

REMARK. The Monte Carlo algorithm leading to $\hat{\sigma}_B$ (2.4) is simple to program. On the Stanford version of the statistical computing language S, Professor Arthur Owen has introduced a single command which bootstraps any statistic in the S catalog. For instance the bootstrap results in Fig. 2 are obtained simply by typing

```
tboot(lawdata, correlation, B = 1000).
```

The execution time is about a factor of B greater than that for the original computation.

There is another way to describe the bootstrap standard error: $\hat{F}$ is the nonparametric maximum likelihood estimate (MLE) of the unknown distribution F (Kiefer and Wolfowitz, 1956). This means that the bootstrap estimate $\hat{\sigma} = \sigma(\hat{F})$ is the nonparametric MLE of $\sigma(F)$, the true standard error.

In fact there is nothing which says that the bootstrap must be carried out nonparametrically. Suppose for instance that in the law school example we believe the true sampling distribution F must be bivariate normal. Then we could estimate F with its *parametric* MLE $\hat{F}_{\mathrm{NORM}}$, the bivariate normal distribution having the same mean vector and covariance matrix as the

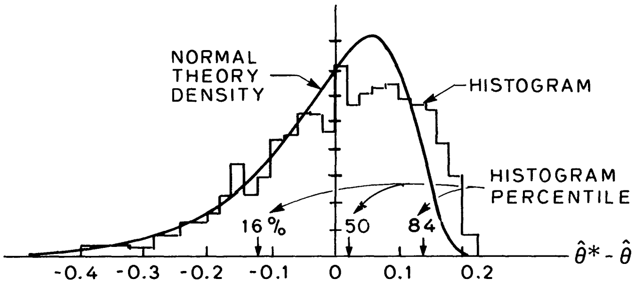

FIG. 2. *Histogram of B = 1000 bootstrap replications of $\hat{\theta}^*$ for the law school data. The normal theory density curve has a similar shape, but falls off more quickly at the upper tail.*

data. The bootstrap samples at step (i) of the algorithm could then be drawn from $\hat{F}_{\text{NORM}}$ instead of $\hat{F}$, and steps (ii) and (iii) carried out as before.

The smooth curve in Fig. 2 shows the results of carrying out this "normal theory bootstrap" on the law school data. Actually there is no need to do the bootstrap sampling in this case, because of Fisher's formula for the sampling density of a correlation coefficient in the bivariate normal situation (see Chapter 32 of Johnson and Kotz, 1970). This density can be thought of as the bootstrap distribution for $B = \infty$. Expression (2.5) is a close approximation to $\hat{\sigma}_{\text{NORM}} = \sigma(\hat{F}_{\text{NORM}})$, the parametric bootstrap estimate of standard error.

In considering the merits or demerits of the bootstrap, it is worth remembering that all of the usual formulas for estimating standard errors, like $\hat{\mathscr{I}}^{-1/2}$ where $\hat{\mathscr{I}}$ is the observed Fisher information, are essentially bootstrap estimates carried out in a parametric framework. This point is carefully explained in Section 5 of Efron (1982c). The straightforward nonparametric algorithm (i)–(iii) has the virtues of avoiding all parametric assumptions, all approximations (such as those involved with the Fisher information expression for the standard error of an MLE), and in fact all analytic difficulties of any kind. The data analyst is free to obtain standard errors for enormously complicated estimators, subject only to the constraints of computer time. Sections 3 and 6 discuss some interesting applied problems which are far too complicated for standard analyses.

TABLE 1

A sampling experiment comparing the bootstrap and jackknife estimates of standard error for the 25% trimmed mean, sample size $n = 15$

	F standard normal			F negative exponential		
	Ave	SD	CV	Ave	SD	CV
Bootstrap $\hat{\sigma}$ ($B = 200$)	.287	.071	.25	.242	.078	.32
Jackknife $\hat{\sigma}_J$	.280	.084	.30	.224	.085	.38
True (minimum CV)	.286		(.19)	.232		(.27)

How well does the bootstrap work? Table 1 shows the answer in one situation. Here $\mathscr{X}$ is the real line, $n = 15$, and the statistic $\hat{\theta}$ of interest is the 25% trimmed mean. If the true sampling distribution F is $N(0, 1)$, then the true standard error is $\sigma(F) = .286$. The bootstrap estimate $\hat{\sigma}$ is nearly unbiased, averaging .287 in a large sampling experiment. The standard deviation of the bootstrap estimate $\hat{\sigma}$ is itself .071 in this case, with coefficient of variation .071/.287 = .25. (Notice that there are two levels of Monte Carlo involved in Table 1: first drawing the actual samples $\mathbf{y} = (x_1, x_2, \cdots, x_{15})$ from F, and then drawing bootstrap samples $(x_1^*, x_2^*, \cdots, x_{15}^*)$ with $\mathbf{y}$ held fixed. The bootstrap samples evaluate $\hat{\sigma}$ for a fixed value of $\mathbf{y}$. The standard deviation .071 refers to the variability of $\hat{\sigma}$ due to the random choice of $\mathbf{y}$.)

The jackknife, another common method of assigning nonparametric standard errors, is discussed in Section 10. The jackknife estimate $\hat{\sigma}_J$ is also nearly unbiased for $\sigma(F)$, but has higher coefficient of variation (CV). The minimum possible CV for a scale-invariant estimate of $\sigma(F)$, assuming full knowledge of the parametric model, is shown in brackets. The nonparametric bootstrap is seen to be moderately efficient in both cases considered in Table 1.

Table 2 returns to the case of $\hat{\theta}$ the correlation coefficient. Instead of real data we have a sampling experiment in which the true F is bivariate normal, true correlation $\theta = .50$, sample size $n = 14$. Table 2 is abstracted from a larger table in Efron (1981b), in

TABLE 2

Estimates of standard error for the correlation coefficient $\hat{\theta}$ and for $\hat{\phi} = \tanh^{-1}\hat{\theta}$; sample size $n = 14$, distribution F bivariate normal with true correlation $\rho = .5$ (from a larger table in Efron, 1981b)

	Summary statistics for 200 trials							
	Standard error estimates for $\hat{\theta}$				Standard error estimates for $\hat{\phi}$			
	Ave	SD	CV	$\sqrt{\text{MSE}}$	Ave	SD	CV	$\sqrt{\text{MSE}}$
1. Bootstrap $B = 128$	.206	.066	.32	.067	.301	.065	.22	.065
2. Bootstrap $B = 512$	.206	.063	.31	.064	.301	.062	.21	.062
3. Normal smoothed bootstrap $B = 128$	.200	.060	.30	.063	.296	.041	.14	.041
4. Uniform smoothed bootstrap $B = 128$	.205	.061	.30	.062	.298	.058	.19	.058
5. Uniform smoothed bootstrap $B = 512$	.205	.059	.29	.060	.296	.052	.18	.052
6. Jackknife	.223	.085	.38	.085	.314	.090	.29	.091
7. Delta method (Infinitesimal jackknife)	.175	.058	.33	.072	.244	.052	.21	.076
8. Normal theory	.217	.056	.26	.056	.302	0	0	.003
True standard error	.218				.299			

which some of the methods for estimating a standard error required the sample size to be even.

The left side of Table 2 refers to $\hat{\theta}$, while the right side refers to $\hat{\phi} = \tanh^{-1}(\hat{\theta}) = .5 \log(1 + \hat{\theta})/(1 - \hat{\theta})$. For each estimator of standard error, the root mean squared error of estimation $[E(\hat{\sigma} - \sigma)^2]^{1/2}$ is given in the column headed $\sqrt{\text{MSE}}$.

The bootstrap was run with $B = 128$ and also with $B = 512$, the latter value yielding only slightly better estimates in accordance with the results of Section 9. Further increasing B would be pointless. It can be shown that $B = \infty$ gives $\sqrt{\text{MSE}} = .063$ for $\hat{\theta}$, only .001 less than $B = 512$. The normal theory estimate (2.5), which we know to be ideal for this sampling experiment, has $\sqrt{\text{MSE}} = .056$.

We can compromise between the totally nonparametric bootstrap estimate $\hat{\sigma}$ and the totally parametric bootstrap estimate $\hat{\sigma}_{\text{NORM}}$. This is done in lines 3, 4, and 5 of Table 2. Let $\hat{\Sigma} = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'/n$ be the sample covariance matrix of the observed data. The *normal smoothed bootstrap* draws the bootstrap sample from $\hat{F} \oplus N_2(0, .25\hat{\Sigma})$, $\oplus$ indicating convolution. This amounts to estimating F by an equal mixture of the n distributions $N_2(x_i, .25\hat{\Sigma})$, that is by a normal window estimate. Each point x_i^* in a smoothed bootstrap sample is the sum of a randomly selected original data point x_j, plus an independent bivariate normal point $z_j \sim N_2(0, .25\hat{\Sigma})$. Smoothing makes little difference on the left side of the table, but is spectacularly effective in the $\hat{\phi}$ case. The latter result is suspect since the true sampling distribution is bivariate normal, and the function $\hat{\phi} = \tanh^{-1}\hat{\theta}$ is specifically chosen to have nearly constant standard error in the bivariate normal family. The *uniform smoothed bootstrap* samples from $\hat{F} \oplus \mathfrak{U}(0, .25\hat{\Sigma})$, where $\mathfrak{U}(0, .25\hat{\Sigma})$ is the uniform distribution on a rhombus selected so $\mathfrak{U}$ has mean vector 0 and covariance matrix $.25\hat{\Sigma}$. It yields moderate reductions in $\sqrt{\text{MSE}}$ for both sides of the table.

Line 6 of Table 2 refers to the *delta method*, which is the most common method of assigning nonparametric standard error. Surprisingly enough, it is badly biased downward on both sides of the table. The delta method, also known as the method of statistical differentials, the Taylor series method, and the infinitesimal jackknife, is discussed in Section 10.

3. EXAMPLES

Example 1. Cox's Proportional Hazards Model

In this section we apply bootstrap standard error estimation to some complicated statistics.

The data for this example come from a study of leukemia remission times in mice, taken from Cox (1972). They consist of measurements of remission time (y) in weeks for two groups, treatment ($x = 0$) and control ($x = 1$), and a 0–1 variable (δ_i) indicating whether or not the remission time is censored (0) or complete (1). There are 21 mice in each group.

The standard regression model for censored data is Cox's proportional hazards model (Cox, 1972). It assumes that the hazard function $h(t \mid x)$, the probability of going into remission in next instant given no remission up to time t for a mouse with covariate x, is of the form

$$h(t \mid x) = h_0(t)e^{\beta x}. \tag{3.1}$$

Here $h_0(t)$ is an arbitrary unspecified function. Since x here is a group indicator, this means simply that the hazard for the control group is e^{β} times the hazard for the treatment group. The regression parameter β is estimated independently of $h_0(t)$ through maximization of the so called "partial likelihood"

$$\text{PL} = \prod_{i \in D} \frac{e^{\beta x_i}}{\sum_{j \in R_i} e^{\beta x_j}}, \tag{3.2}$$

where D is the set of indices of the failure times and R_i is the set of indices of those at risk at time y_i. This maximization requires an iterative computer search.

The estimate $\hat{\beta}$ for these data turns out to be 1.51. Taken literally, this says that the hazard rate is $e^{1.51} = 4.33$ times higher in the control group than in the treatment group, so the treatment is very effective. What is the standard error of $\hat{\beta}$? The usual asymptotic maximum likelihood theory, one over the square root of the observed Fisher information, gives an estimate of .41. Despite the complicated nature of the estimation procedure, we can also estimate the standard error using the bootstrap. We sample with replacement from the triples $\{(y_1, x_1, \delta_1), \cdots, (y_{42}, x_{42}, \delta_{42})\}$. For each bootstrap sample $\{(y_1^*, x_1^*, \delta_1^*), \cdots, (y_{42}^*, x_{42}^*, \delta_{42}^*)\}$ we form the partial likelihood and numerically maximize it to produce the bootstrap estimate $\hat{\beta}^*$. A histogram of 1000 bootstrap values is shown in Fig. 3.

The bootstrap estimate of the standard error of $\hat{\beta}$ based on these 1000 numbers is .42. Although the bootstrap and standard estimates agree, it is interesting to note that the bootstrap distribution is skewed to the right. This leads us to ask: is there other information that we can extract from the bootstrap distribution other than a standard error estimate? The answer is yes—in particular, the bootstrap distribution can be used to form a confidence interval for β, as we will see in Section 9. The shape of the bootstrap distributiion will help determine the shape of the confidence interval.

In this example our resampling unit was the triple (y_i, x_i, δ_i), and we ignored the unique elements of the problem, i.e., the censoring, and the particular model being used. In fact, there are other ways to bootstrap

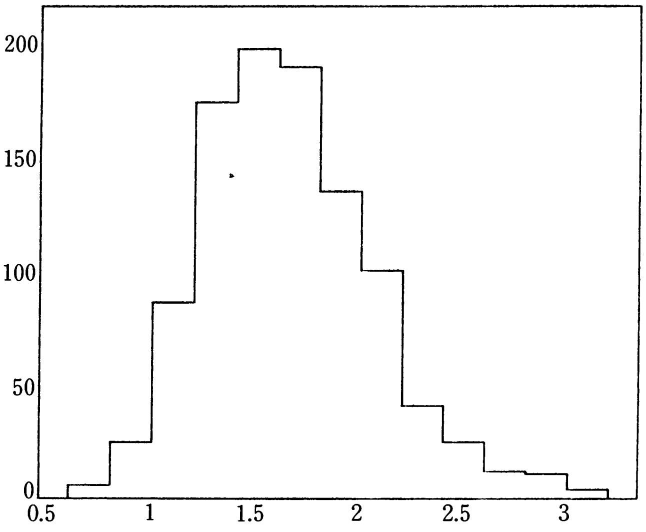

FIG. 3. *Histogram of 1000 bootstrap replications for the mouse leukemia data.*

this problem. We will see this when we discuss bootstrapping censored data in Section 5.

Example 2: Linear and Projection Pursuit Regression

We illustrate an application of the bootstrap to standard linear least squares regression as well as to a nonparametric regression technique.

Consider the standard regression setup. We have n observations on a response Y and covariates $(X_1, X_2, \cdots, X_p)$. Denote the ith observed vector of covariates by $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$. The usual linear regression model assumes

$$(3.3) \qquad E(Y_i) = \alpha + \sum_{j=1}^{p} \beta_j x_{ij}.$$

Friedman and Stuetzle (1981) introduced a more general model, the *projection pursuit* regression model

$$(3.4) \qquad E(Y_i) = \sum_{j=1}^{m} s_j(a_j \cdot x_i).$$

The p vectors a_j are unit vectors ("directions"), and the functions $s_j(\cdot)$ are unspecified.

Estimation of $\{a_1, s_1(\cdot)\}, \cdots, \{a_m, s_m(\cdot)\}$ is performed in a forward stepwise manner as follows. Consider $\{a_1, s_1(\cdot)\}$. Given a direction a_1, $s_1(\cdot)$ is estimated by a nonparametric smoother (e.g., running mean) of y on $a_1 \cdot x$. The projection pursuit regression algorithm searches over all unit directions to find the direction $\hat{a}_1$ and associated function $\hat{s}_1(\cdot)$ that minimize $\sum_1^n (y_i - \hat{s}_1(\hat{a} \cdot x_i))^2$. Then residuals are taken and the next direction and function are determined. This process is continued until no additional term significantly reduces the residual sum of squares.

Notice the relation of the projection pursuit regression model to the standard linear regression model. When the function $s_1(\cdot)$ is forced to be linear and is estimated by the usual least squares method, a one-term projection pursuit model is exactly the same as the standard linear regression model. That is to say, the fitted model $\hat{s}_1(\hat{a}_1 \cdot x_i)$ exactly equals the least squares fit $\hat{\alpha} + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$. This is because the least squares fit, by definition, finds the best direction and the best linear function of that direction. Note also that adding another linear term $\hat{s}_2(\hat{a}_2 \cdot x_2)$ would not change the fitted model since the sum of two linear functions is another linear function.

Hastie and Tibshirani (1984) applied the bootstrap to the linear and projection pursuit regression models to assess the variability of the coefficients in each. The data they considered are taken from Breiman and Friedman (1985). The response Y is Upland atmospheric ozone concentration (ppm); the covariates X_1 = Sandburg Air Force base temperature (C°), X_2 = inversion base height (ft), X_3 = Daggot pressure gradient (mm Hg), X_4 = visibility (miles), and X_5 = day of the year. There are 330 observations. The number of terms (m) in the model (3.4) is taken to be two. The projection pursuit algorithm chose directions $\hat{a}_1 = (.80, -.38, .37, -.24, -.14)'$ and $\hat{a}_2 = (.07, .16, .04, -.05, -.98)'$. These directions consist mostly of Sandburg Air Force temperature and day of the year, respec-

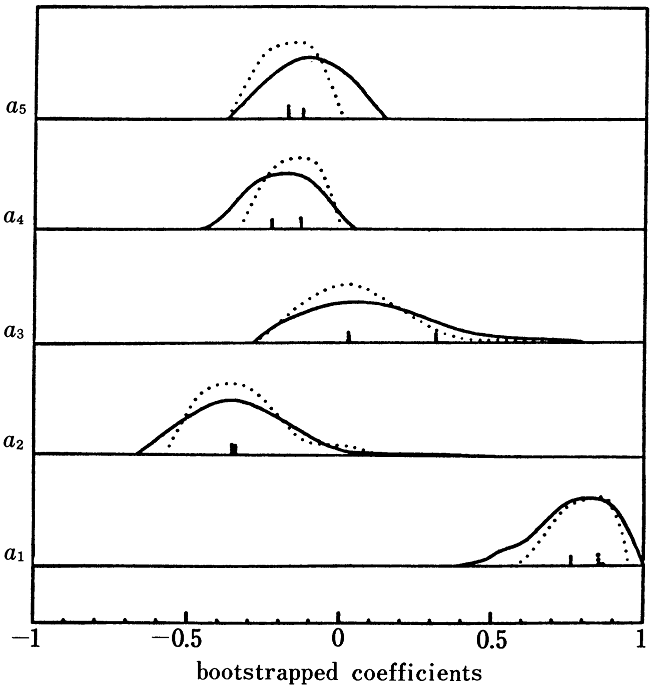

FIG. 4. *Smoothed histograms of the bootstrapped coefficients for the first term in the projection pursuit regression model. Solid histograms are for the usual projection pursuit model; the dotted histograms are for linear $\hat{s}(\cdot)$.*

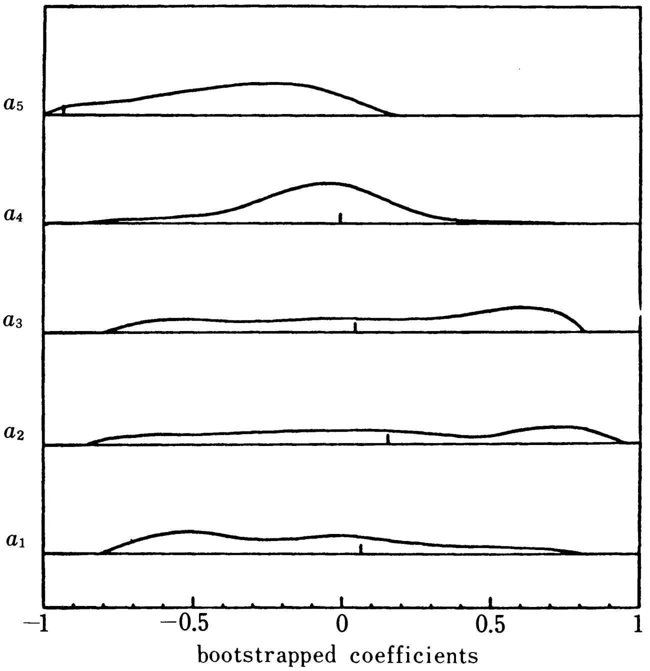

FIG. 5. *Smoothed histograms of the bootstrapped coefficients for the second term in the projection pursuit model.*

tively. (We do not show graphs of the estimated functions $\hat{s}_1(\cdot)$ and $\hat{s}_2(\cdot)$ although in a full analysis of the data they would also be of interest.) Forcing $\hat{s}_1(\cdot)$ to be linear results in the direction $\hat{a}_1 = (.90, -.37, .03, -.14, -.19)'$. These are just the usual least squares estimates $\hat{\beta}_1, \cdots, \hat{\beta}_p$ scaled so that $\sum_1^p \beta_j^2 = 1$.

To assess the variability of the directions, a bootstrap sample is drawn with replacement from $(y_1, x_{11}, \cdots, x_{15}), \cdots, (y_{330}, x_{3301}, \cdots, x_{3305})$ and the projection pursuit algorithm is applied. Figs. 4 and 5 show histograms of the directions $\hat{a}_1^*$ and $\hat{a}_2^*$ for 200 bootstrap replications. Also shown in Fig. 4 (broken histogram) are the bootstrap replications of $\hat{a}_1$ with $\hat{s}_1(\cdot)$ forced to be linear.

The first direction of the projection pursuit model is quite stable and only slightly more variable than the corresponding linear regression direction. But the second direction is extremely unstable! It is clearly unwise to put any faith in the second direction of the original projection pursuit model.

Example 3: Cox's Model and Local Likelihood Estimation

In this example, we return to Cox's proportional hazards model described in Example 1, but with a few added twists.

The data that we will discuss come from the Stanford heart transplant program and are given in Miller and Halpern (1982). The response y is survival time in weeks after a heart transplant, the covariate x is age at transplant, and the 0–1 variable δ indicates whether the survival time is censored (0) or complete (1). There are measurements on 157 patients. A proportional hazards model was fit to these data, with a quadratic term, i.e, $h(t \mid x) = h_0(t)e^{\beta_1 x + \beta_2 x^2}$. Both $\hat{\beta}_1$ and $\hat{\beta}_2$ are highly significant; the broken curve in Fig. 6 is $\hat{\beta}_1 x + \hat{\beta}_2 x^2$ as a function of x.

For comparison, Fig. 6 shows (solid line) another estimate. This was computed using *local likelihood estimation* (Tibshirani and Hastie, 1984). Given a general proportional hazards model of the form $h(t \mid x) = h_0(t)e^{s(x)}$, the local likelihood technique assumes nothing about the parametric form of $s(x)$; instead it estimates $s(x)$ nonparametrically using a kind of local averaging. The algorithm is very computationally intensive, and standard maximum likelihood theory cannot be applied.

A comparison of the two functions reveals an important qualitative difference: the parametric estimate suggests that the hazard decreases sharply up to age 34, then rises; the local likelihood estimate stays approximately constant up to age 45 then rises. Has the forced fitting of a quadratic function produced a misleading result? To answer this question, we can bootstrap the local likelihood estimate. We sample with replacement from the triples $\{(y_1, x_1, \delta_1) \cdots (y_{157}, x_{157}, \delta_{157})\}$ and apply the local likelihood algorithm to each bootstrap sample. Fig. 7 shows estimated curves from 20 bootstrap samples.

Some of the curves are flat up to age 45, others are decreasing. Hence the original local likelihood estimate is highly variable in this region and on the basis of these data we cannot determine the true behavior of the function there. A look back at the original data shows that while half of the patients were under 45, only 13% of the patients were under 30. Fig. 7 also shows that the estimate is stable near the middle ages but unstable for the older patients.

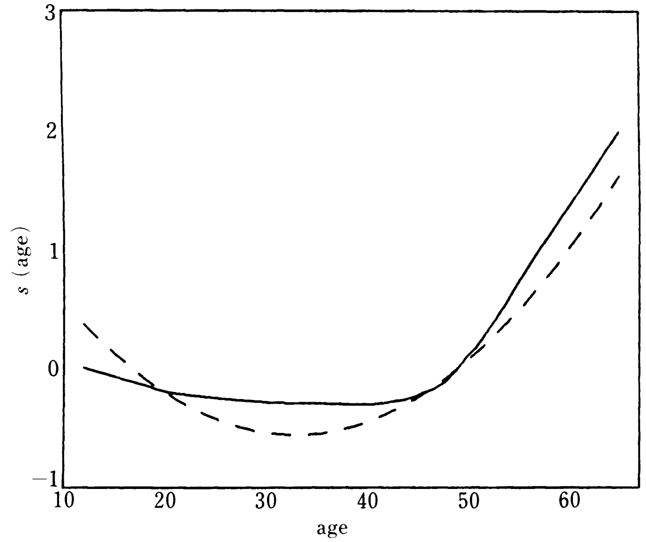

FIG. 6. *Estimates of log relative risk for the Stanford heart transplant data. Broken curve: parametric estimate. Solid curve: local likelihood estimate.*

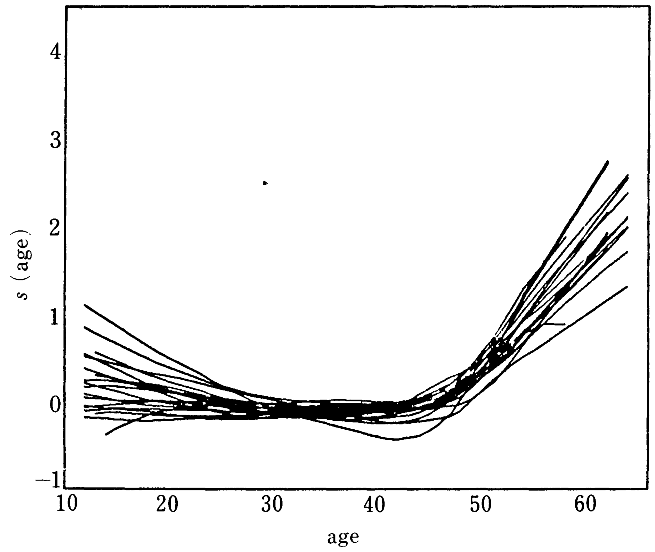

FIG. 7. *20 bootstraps of the local likelihood estimate for the Stanford heart transplant data.*

4. OTHER MEASURES OF STATISTICAL ERROR

So far we have discussed statistical error, or accuracy, in terms of the standard error. It is easy to assess other measures of statistical error, such as bias or prediction error, using the bootstrap.

Consider the estimation of bias. For a given statistic $\hat{\theta}(\mathbf{y})$, and a given parameter $\mu(F)$, let

$$R(\mathbf{y}, F) = \hat{\theta}(\mathbf{y}) - \mu(F). \tag{4.1}$$

(It will help keep our notation clear to call the parameter of interest μ rather than θ.) For example, μ might be the mean of the distribution F, assuming the sample space X is the real line, and $\hat{\theta}$ the 25% trimmed mean. The bias of $\hat{\theta}$ for estimating μ is

$$\beta(F) = E_F R(\mathbf{y}, F) = E_F\{\hat{\theta}(\mathbf{y})\} - \mu(F). \tag{4.2}$$

The notation E_F indicates expectation with respect to the probability mechanism appropriate to F, in this case $\mathbf{y} = (x_1, x_2, \cdots, x_n)$ a random sample from F.

The bootstrap estimate of bias is

$$\hat{\beta} = \beta(\hat{F}) = E_{\hat{F}} R(\mathbf{y}^*, \hat{F}) = E_{\hat{F}}\{\hat{\theta}(\mathbf{y}^*)\} - \mu(\hat{F}). \tag{4.3}$$

As in Section 2, $\mathbf{y}^*$ denotes a random sample $(x_1^*, x_2^*, \cdots, x_n^*)$ from $\hat{F}$, i.e., a bootstrap sample. To numerically evaluate $\hat{\beta}$, all we do is change step (iii) of the bootstrap algorithm in Section 2 to

$$\begin{aligned} \hat{\beta}_B &= \frac{1}{B} \sum_{b=1}^{B} R(\mathbf{y}^*(b), \hat{F}). \\ &= \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B} - \mu(\hat{F}). \\ &= \hat{\theta}^*(\cdot) - \hat{\mu}(F). \end{aligned} \tag{4.4}$$

As $B \rightarrow \infty$, $\hat{\beta}_B$ goes to $\hat{\beta}$ (4.3).

TABLE 3
BHCG blood serum levels for 54 patients having metasticized breast cancer in ascending order

0.1, 0.1, 0.2, 0.4, 0.4, 0.6, 0.8, 0.8, 0.9, 0.9, 1.3, 1.3, 1.4, 1.5, 1.6, 1.6, 1.7, 1.7, 1.7, 1.8, 2.0, 2.0, 2.2, 2.2, 2.2, 2.3, 2.3, 2.4, 2.4, 2.4, 2.4, 2.4, 2.4, 2.5, 2.5, 2.5, 2.7, 2.7, 2.8, 2.9, 2.9, 2.9, 3.0, 3.1, 3.1, 3.2, 3.2, 3.3, 3.3, 3.5, 4.4, 4.5, 6.4, 9.4

As an example consider the blood serum data of Table 3. Suppose we wish to estimate the true mean $\mu = E_F\{X\}$ of this population using $\hat{\theta}$, the 25% trimmed mean. We calculate $\hat{\mu} = \mu(\hat{F}) = 2.32$, the sample mean of the 54 observations, and $\hat{\theta} = 2.24$, the trimmed mean. The trimmed mean is lower because it discounts the effect of the large observations 6.4 and 9.4. It looks like the trimmed mean might be more robust for this type of data, and as a matter of fact a bootstrap analysis, $B = 1000$, gave estimated standard error $\hat{\sigma} = .16$ for $\hat{\theta}$, compared to .21 for the sample mean. But what about bias?

The same 1000 bootstrap replications which gave $\hat{\sigma} = .16$ also gave $\hat{\theta}^*(\cdot) = 2.29$, so

$$\hat{\beta} = 2.29 - 2.32 = -0.03. \tag{4.5}$$

according to (4.4). (The estimated standard deviation of $\hat{\beta}_B - \hat{\beta}$ due to the limitations of having $B = 1000$ bootstraps is only 0.005 in this case, so we can ignore the difference between $\hat{\beta}_B$ and $\hat{\beta}$.) Whether or not a bias of magnitude −0.03 is too large depends on the context of the problem. If we attempt to remove the bias by subtraction, we get $\hat{\theta} - \hat{\beta} = 2.24 - (-0.03) = 2.27$. Removing bias in this way is frequently a bad idea (see Hinkley, 1978), but at least the bootstrap analysis has given us a reasonable picture of the bias and standard error of $\hat{\theta}$.

Here is another measure of statistical accuracy, different from either bias or standard error. Let $\hat{\theta}(\mathbf{y})$ be the 25% trimmed mean and $\mu(F)$ be the mean of F, as in the serum example, and also let $\hat{i}(\mathbf{y})$ be the interquartile range, the distance between the 25th and 75th percentiles of the sample $\mathbf{y} = (x_1, x_2, \cdots, x_n)$. Define

$$R(\mathbf{y}, F) = \frac{\hat{\theta}(\mathbf{y}) - \mu(F)}{\hat{i}(\mathbf{y})}. \tag{4.6}$$

R is like a Student's t statistic, except that we have substituted the 25% trimmed mean for the sample mean and the interquartile range for the standard deviation.

Suppose we know the 5th and 95th percentiles of $R(\mathbf{y}, F)$, say $\rho^{(.05)}(F)$ and $\rho^{(.95)}(F)$, where the definition of $\rho^{(.05)}(F)$ is

$$\text{Prob}_F\{R(\mathbf{y}, F) < \rho^{(.05)}(F)\} = .05, \tag{4.7}$$

and similarly for $\rho^{(.95)}(F)$. The relationship $\text{Prob}_F\{\rho^{(.05)} \le R < \rho^{(.95)}\} = .90$ combines with definition (4.6) to give a central 90% "t interval" for the mean $\mu(F)$,

$$\mu \in [\hat\theta - \hat\imath\rho^{(.95)}, \hat\theta - \hat\imath\rho^{(.05)}]. \tag{4.8}$$

Of course we do not know $\rho^{(.05)}(F)$ and $\rho^{(.95)}(F)$, but we can approximate them by their bootstrap estimates $\rho^{(.05)}(\hat F)$ and $\rho^{(.95)}(\hat F)$. A bootstrap sample $\mathbf{y}^*$ gives a bootstrap value of (4.6), $R(\mathbf{y}^*, \hat F) = (\hat\theta(\mathbf{y}^*) - \mu(\hat F))/\hat\imath(\mathbf{y}^*)$, where $\hat\imath(\mathbf{y}^*)$ is the interquartile range of the bootstrap data $x_1^*, x_2^*, \cdots, x_n^*$. For any fixed number ρ, the bootstrap estimate of $\mathrm{Prob}_F\{R < \rho\}$ based on B bootstrap samples is

$$\#\{R(\mathbf{y}^*(b), \hat F) < \rho\}/B. \tag{4.9}$$

By keeping track of the empirical distribution of $R(\mathbf{y}^*(b), \hat F)$, we can pick off the values of ρ which make (4.9) equal .05 and .95. These approach $\rho^{(.05)}(\hat F)$ and $\rho^{(.95)}(\hat F)$ as $B \to \infty$.

For the serum data, $B = 1000$ bootstrap replications gave $\rho^{(.05)}(\hat F) = -.303$ and $\rho^{(.95)}(\hat F) = .078$. Substituting these values into (4.9), and using the observed estimates $\hat\theta = 2.24$, $\hat\imath = 1.40$, gives

$$\mu \in [2.13, 2.66] \tag{4.10}$$

as a central 90% "bootstrap t interval" for the true mean $\mu(F)$. This is considerably shorter than the standard t interval for μ based on 53 degrees of freedom, $\bar x \pm 1.67\bar\sigma = [1.97, 2.67]$. Here $\bar\sigma = .21$ is the usual estimate of standard error (1.3).

Bootstrap confidence intervals are discussed further in Sections 7 and 8. They require more bootstrap replications than do bootstrap standard errors, on the order of $B = 1000$ rather than $B = 50$ or 100. This point is discussed briefly in Section 9.

By now it should be clear that we can use any random variable $R(\mathbf{y}, F)$ to measure accuracy, not just (4.1) or (4.6), and then estimate $E_F\{R(\mathbf{y}, F)\}$ by its bootstrap value $E_{\hat F}\{R(\mathbf{y}^*, \hat F)\} \doteq \sum_{b=1}^{B} R(\mathbf{y}^*(b), \hat F)/B$. Similarly we can estimate $E_F R(y, F)^2$ by $E_{\hat F} R(\mathbf{y}^*, \hat F)^2$, etc. Efron (1983) considers the prediction problem, in which a training set of data is used to construct a prediction rule. A naive estimate of the prediction rule's accuracy is the proportion of correct guesses it makes on its own training set, but this can be greatly over optimistic since the prediction rule is explicitly constructed to minimize errors on the training set. In this case, a natural choice of $R(\mathbf{y}, F)$ is the over optimism, the difference between the naive estimate and the actual success rate of the prediction rule for new data. Efron (1983) gives the bootstrap estimate of over optimism, and shows that it is closely related to cross-validation, the usual method of estimating over optimism. The paper goes on to show that some modifications of the bootstrap estimate greatly out perform both cross-validation and the bootstrap.

5. MORE COMPLICATED DATA SETS

The bootstrap is not restricted to situations where the data is a simple random sample from a single distribution. Suppose for instance that the data consists of two independent random samples,

$$U_1, U_2, \cdots, U_m \sim F \text{ and } V_1, V_2, \cdots, V_n \sim G, \tag{5.1}$$

where F and G are possibly different distributions on the real line. Suppose also that the statistic of interest is the Hodges–Lehmann shift estimate

$$\hat\theta = \text{median}\{V_j - U_i, i = 1, 2, \cdots, m, j = 1, 2, \cdots, n\}. \tag{5.2}$$

Having observed $U_1 = u_1$, $U_2 = u_2$, $\cdots$, $V_n = v_n$, we desire an estimate for $\sigma(F, G)$, the standard error of $\hat\theta$.

The bootstrap estimate of $\sigma(F, G)$ is $\hat\sigma = \sigma(\hat F, \hat G)$, where $\hat F$ is the empirical distribution of $u_1, u_2, \cdots, u_m$, and $\hat G$ is the empirical distribution of $v_1, v_2, \cdots, v_n$. It is easy to modify the Monte Carlo algorithm of Section 2 to numerically evaluate $\hat\sigma$. Let $\mathbf{y} = (u_1, u_2, \cdots, v_n)$ be the observed data vector. A bootstrap sample $\mathbf{y}^* = (u_1^*, u_2^*, \cdots, u_m^*, v_1^*, v_2^*, \cdots, v_n^*)$ consists of a random sample $U_1^*, \cdots, U_m^*$ from $\hat F$ and an independent random sample $V_1^*, \cdots, V_n^*$ from $\hat G$. With only this modification, steps (i) through (iii) of the Monte Carlo algorithm produce $\hat\sigma_B$, (2.4), approaching $\hat\sigma$ as $B \to \infty$.

Table 4 reports on a simulation experiment investigating how well the bootstrap works on this problem. 100 trials of situation (5.1) were run, with $m = 6$, $n = 9$, F and G both Uniform [0, 1]. For each trial, both $B = 100$ and $B = 200$ bootstrap replications were generated. The bootstrap estimate $\hat\sigma_B$ was nearly unbiased for the true standard error $\sigma(F, G) = .167$ for either $B = 100$ or $B = 200$, with a quite small standard deviation from trial to trial. The improvement in going from $B = 100$ to $B = 200$ is too small to show up in this experiment.

In practice, statisticians must often consider quite complicated data structures: time series models, mul-

TABLE 4

Bootstrap estimate of standard error for the Hodges–Lehmann two-sample shift estimate; 100 trials

	Summary statistics for $\hat\sigma_B$		
	Ave	SD	CV
$B = 100$	.165	.030	.18
$B = 200$	.166	.031	.19
True σ	.167		

Note: $m = 6$, $n = 9$; true distributions F and G both uniform [0, 1].

tifactor layouts, sequential sampling, censored and missing data, etc. Fig. 8 illustrates how the bootstrap estimation process proceeds in a general situation. The actual probability mechanism P which generates the observed data $\mathbf{y}$ belongs to some family P of possible probability mechanism. In the Hodges–Lehmann example, $P = (F, G)$, a pair of distributions on the real line, P equals the family of all such pairs, and $\mathbf{y} = (u_1, u_2, \cdots, u_m, v_1, v_2, \cdots, v_n)$ is generated by random sampling m times from F and n times from G.

We have a random variable of interest $R(\mathbf{y}, P)$, which depends on both $\mathbf{y}$ and the unknown model P, and we wish to estimate some aspect of the distribution of R. In the Hodges–Lehmann example, $R(\mathbf{y}, P) = \hat{\theta}(\mathbf{y}) - E_P\{\hat{\theta}\}$, and we estimated $\sigma(P) = \{E_P R(\mathbf{y}, P)^2\}^{1/2}$, the standard error of $\hat{\theta}$. As before, the notation E_P indicates expectation when $\mathbf{y}$ is generated according to mechanism P.

We assume that we have some way of estimating the entire probability model P from the data $\mathbf{y}$, producing the estimate called $\hat{P}$ in Fig. 8. (In the two-sample problem, $\hat{P} = (\hat{F}, \hat{G})$, the pair of empirical distributions.) *This is the crucial step for the bootstrap.* It can be carried out either parametrically or nonparametrically, by maximum likelihood or by some other estimation technique.

Once we have $\hat{P}$, we can use Monte Carlo methods to generate bootstrap data sets $\mathbf{y}^*$, according to the same rules by which $\mathbf{y}$ is generated from P. The bootstrap random variable $R(\mathbf{y}^*, \hat{P})$ is observable, since we know $\hat{P}$ as well as $\mathbf{y}^*$, so the distribution of $R(\mathbf{y}^*, \hat{P})$ can be found by Monte Carlo sampling. The bootstrap estimate of $E_P R(\mathbf{y}, P)$ is then $E_{\hat{P}} R(\mathbf{y}^*, \hat{P})$, and likewise for estimating any other aspect of $R(\mathbf{y}, P)$'s distribution.

A regression model is a familiar example of a complicated data structure. We observe $\mathbf{y} = (y_1, y_2, \cdots, y_n)$, where

$$y_i = g(\beta, t_i) + \varepsilon_i \quad i = 1, 2, \cdots, n. \tag{5.3}$$

Here β is a vector of unknown parameters we wish to estimate; for each i, t_i is an observed vector of covariates; and g is a known function of β and t_i, for instance $e^{\beta' t_i}$. The ε_i are an iid sample from some unknown distribution F on the real line,

$$\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \sim F, \tag{5.4}$$

where F is usually assumed to be centered at 0 in some sense, perhaps $E\{\varepsilon\} = 0$ or $\text{Prob}\{\varepsilon < 0\} = .5$. The probability model is $P = (\beta, F)$; (5.3) and (5.4) describe the step $P \longrightarrow \mathbf{y}$ in Fig. 8. The covariates $t_1, t_2, \cdots, t_n$, like the sample size n in the simple problem (1.1), are considered fixed at their observed values.

For every choice of β we have a vector $\mathbf{g}(\beta) = (g(\beta, t_1), g(\beta, t_2), \cdots, g(\beta, t_n))$ of predicted values for $\mathbf{y}$. Having observed $\mathbf{y}$, we estimate β by minimizing some measure of distance between $\mathbf{g}(\beta)$ and $\mathbf{y}$,

$$\hat{\beta}: \min_{\beta} D(\mathbf{y}, \mathbf{g}(\beta)). \tag{5.5}$$

The most common choice of D is $D(\mathbf{y}, \mathbf{g}) = \sum_{i=1}^{n} \{y_i - g(\beta, t_i)\}^2$.

How accurate is $\hat{\beta}$ as an estimate of β? Let $R(\mathbf{y}, P)$ equal the vector $\hat{\beta} - \beta$. A familiar measure of accuracy is the mean square error matrix

$$\begin{aligned} \Sigma(P) &= E_P(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \\ &= E_P R(\mathbf{y}, P) R(\mathbf{y}, P)'. \end{aligned} \tag{5.6}$$

The bootstrap estimate of accuracy $\hat{\Sigma} = \Sigma(\hat{P})$ is obtained by following through Fig. 8.

There is an obvious choice for $\hat{P} = (\hat{\beta}, \hat{F})$ in this case. The estimate $\hat{\beta}$ is obtained from (5.5). Then $\hat{F}$ is the empirical distribution of the residuals,

$$\hat{F}: \text{mass}(1/n) \quad \text{on} \quad \hat{\varepsilon}_i \equiv y_i - g(\hat{\beta}, t_i), \qquad i = 1, \cdots, n. \tag{5.7}$$

A bootstrap sample $\mathbf{y}^*$ is obtained by following rules (5.3) and (5.4),

$$y_i^* = g(\hat{\beta}, t_i) + \varepsilon_i^*, \quad i = 1, 2, \cdots, n, \tag{5.8}$$

where $\varepsilon_1^*, \varepsilon_2^*, \cdots, \varepsilon_n^*$ is an iid sample from $\hat{F}$. Notice that the ε_n^* are independent bootstrap variates, even though the $\hat{\varepsilon}_i$ are not independent variates in the usual sense.

Each bootstrap sample $\mathbf{y}^*(b)$ gives a bootstrap value $\hat{\beta}^*(b)$,

$$\hat{\beta}^*(b): \min_{\beta} D(\mathbf{y}^*(b), \mathbf{g}(\beta)), \tag{5.9}$$

as in (5.5). The estimate

$$\hat{\Sigma}_B = \frac{\sum_{b=1}^{B} \{\hat{\beta}^*(b) - \hat{\beta}^*(\cdot)\}\{\hat{\beta}^*(b) - \hat{\beta}^*(\cdot)\}'}{B} \tag{5.10}$$

approaches the bootstrap estimate $\hat{\Sigma}$ as $B \to \infty$. (We could just as well divide by $B - 1$ in (5.10).)

In the case of ordinary least squares regression, where $g(\beta, t_i) = \beta' t_i$ and $D(\mathbf{y}, \mathbf{g}) = \sum_{i=1}^{n} (y_i - g_i)^2$,

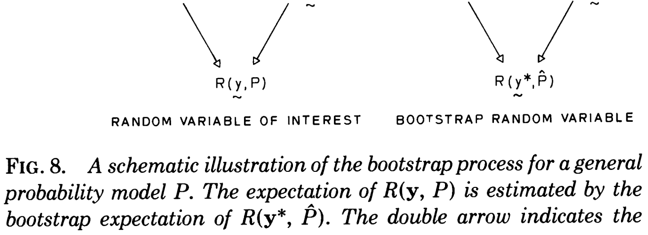
FAMILY OF POSSIBLE PROBABILITY MODELS
ACTUAL PROBABILITY MODEL
OBSERVED DATA
ESTIMATED PROBABILITY MODEL
BOOTSTRAP DATA
$\mathcal{P}$
P
$\underset{\sim}{y}$
$\hat{P}$
$\underset{\sim}{y}^*$
$\underset{\sim}{R}(y,P)$
$\underset{\sim}{R}(y^*,\hat{P})$
RANDOM VARIABLE OF INTEREST
BOOTSTRAP RANDOM VARIABLE

FIG. 8. *A schematic illustration of the bootstrap process for a general probability model P. The expectation of* $R(\mathbf{y}, P)$ *is estimated by the bootstrap expectation of* $R(\mathbf{y}^*, \hat{P})$. *The double arrow indicates the crucial step in applying the bootstrap.*

Section 7 of Efron (1979a) shows that the bootstrap estimate, $B = \infty$, can be calculated without Monte Carlo sampling, and is

$$(5.11) \qquad \hat{\Sigma} = \hat{\sigma}^2 \left(\sum_{i=1}^{n} t_i t_i' \right)^{-1} \quad \left[\hat{\sigma}^2 \equiv \sum_{1}^{n} \frac{\hat{\varepsilon}_n^2}{n} \right].$$

This is the usual Gauss–Markov answer, except for the divisor n in the definition of $\hat{\sigma}^2$.

There is another, simpler way to bootstrap a regression problem. We can consider each covariate-response pair $x_i = (t_i, y_i)$ to be a single data point obtained by simple random sampling from a distribution F. If the covariate vector t_i is p-dimensional, F is a distribution on $p + 1$ dimensions. Then we apply the bootstrap as described originally in Section 2 to the data set $x_1, x_2, \cdots, x_n \sim_{\text{iid}} F$.

The two bootstrap methods for the regression problem are asymptotically equivalent, but can perform quite differently in small sample situations. The class of possible probability models P is different for the two methods. The simple method, described last, takes less advantage of the special structure of the regression problem. It does *not* give answer (5.11) in the case of ordinary least squares. On the other hand the simple method gives a trustworthy estimate of $\hat{\beta}$'s variability *even if the regression model* (5.3) *is not correct*. The bootstrap, as outlined in Fig. 5, is very general, but because of this generality there will often be more than one bootstrap solution for a given problem.

As the final example of this section, we discuss *censored data*. The ages of 97 men at a California retirement center, Channing House, were observed either at death (an uncensored observation) or at the time the study ended (a censored observation). The data set $\mathbf{y} = \{(x_1, d_1), (x_2, d_2), \cdots, (x_{97}, d_{97})\}$, where x_i was the age of the ith man observed, and

$$d_i = \begin{cases} 1 & \text{if } x_i \text{ uncensored} \\ 0 & \text{if } x_i \text{ censored.} \end{cases}$$

Thus (777, 1) represents a Channing House man observed to die at age 777 months, while (843, 0) represents a man 843 months old when the study ended. His observation could be written as "843+," and in fact d_i is just an indicator for the absence or presence of "+." A full description of the Channing House data appears in Hyde (1980).

A typical data point (X_i, D_i) can be thought of as generated in the following way: a real lifetime X_i^0 is selected randomly according to a survival curve

$$(5.12) \qquad S^0(t) \equiv \text{Prob}\{X_i^0 > t\}, \quad (0 \le t < \infty)$$

and a censoring time W_i is independently selected according to another survival curve

$$(5.13) \qquad R(t) \equiv \text{Prob}\{W_i > t\}, \quad (0 \le t < \infty).$$

The statistician gets to observe

$$(5.14) \qquad X_i = \min\{X_i^0, W_i\}$$

and

$$(5.15) \qquad D_i = \begin{cases} 1 & \text{if } X_i = X_i^0 \\ 0 & \text{if } X_i = W_i. \end{cases}$$

Note: $1 - S^0(t)$ and $1 - R(t)$ are the cumulative distribution functions (cdf) for X_i^0 and W_i, respectively; with censored data it is more convenient to consider survival curves than cdf.

Under assumptions (5.12)–(5.15) there is a simple formula for the nonparametric MLE of $S^0(t)$, called the *Kaplan–Meier estimator* (Kaplan and Meier, 1958). For convenience suppose $x_1 < x_2 < x_3 < \cdots < x_n$, $n = 97$. Then the Kaplan–Meier estimate is

$$(5.16) \qquad \hat{S}^0(t) = \prod_{j=1}^{k_t} \left(\frac{n-i}{n-i+1} \right)^{d_i},$$

where k_t is the value of k such that $t \in [x_k, x_{k+1})$. In the case of no censoring, $\hat{S}^0(t)$ is equivalent to the observed empirical distribution of $x_1, x_2, \cdots, x_n$, but otherwise (5.16) corrects the empirical distribution to account for censoring. Likewise

$$(5.17) \qquad \hat{R}(t) = \prod_{j=1}^{k_t} \left(\frac{n-i}{n-i+1} \right)^{1-d_i}$$

is the Kaplan–Meier estimate of the censoring curve $R(t)$.

Fig. 9 shows $\hat{S}^0(t)$ for the Channing House men. It crosses the 50% survival level at $\hat{\theta} = 1044$ months. Call this value the observed median lifetime. We can use the bootstrap to assign a standard error to the observed median.

The probability mechanism is $P = (S^0, R)$; P produces (X_i^0, D_i) according to (5.12)–(5.15), and $\mathbf{y} = \{(x_1, d_1), \cdots, (x_n, d_n)\}$ by $n = 97$ independent repetitions of this process. An obvious choice of the estimate $\hat{P}$ in Fig. 8 is $(\hat{S}^0, \hat{R})$, (5.14), (5.15). The rest of

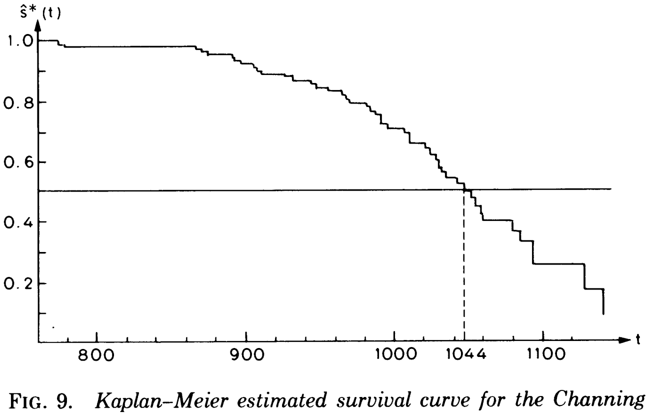

FIG. 9. *Kaplan–Meier estimated survival curve for the Channing House men; t = age in months. The median survival age is estimated to be* 1044 *months* (87 *years*).

bootstrap process is automatic: $\hat{S}^0$ and $\hat{R}$ replace S^0 and R in (5.12) and (5.13); n pairs (X_i^*, D_i^*) are independently generated according to rules (5.12)–(5.15), giving the bootstrap data set $\mathbf{y}^* = \{x_1^*, d_1^*, \cdots, (x_n^*, d_n^*)\}$; and finally the bootstrap Kaplan-Meier curve $\hat{S}^{0*}$ is constructed according to formula (5.16), and the bootstrap observed median $\hat{\theta}^*$ calculated. For the Channing House data, $B = 1600$ bootstrap replications of $\hat{\theta}^*$ gave estimated standard error $\hat{\sigma} = 14.0$ months for $\hat{\theta}$. An estimated bias of 4.1 months was calculated as at (4.4). Efron (1981b) gives a fuller description.

Once again there is a simpler way to apply to bootstrap. Consider each pair $y_i = (x_i, d_i)$ as an observed point obtained by simple random sampling from a bivariate distribution F, and apply the bootstrap as described in Section 2 to the data set $y_1, y_2, \cdots, y_n \sim_{\text{iid}} F$. This method makes no use of the special structure (5.12)–(5.15). Surprisingly, it gives *exactly the same answers* as the more complicated bootstrap method described earlier (Efron, 1981a). This leads to a surprising conclusion: bootstrap estimates of variability for the Kaplan-Meier curve give correct standard errors even when the usual assumptions about the censoring mechanism, (5.12)–(5.15), fail.

6. EXAMPLES WITH MORE COMPLICATED DATA STRUCTURES

Example 1: Autoregressive Time Series Model

This example illustrates an application of the bootstrap to a famous time series.

The data are the Wolfer annual sunspot numbers for the years 1770–1889 (taken from Anderson, 1975). Let the count for the ith year be z_i. After centering the data (replacing z_i by $z_i - \bar{z}_i$), we fit a first-order autoregressive model

(6.1) $$z_i = \phi z_{i-1} + \varepsilon_i$$

where $\varepsilon_i \sim$ iid $N(0, \sigma^2)$. The estimate $\hat{\phi}$ turned out to be .815 with an estimated standard error, one over the square root of the Fisher information, of .053.

A bootstrap estimate of the standard error of $\hat{\phi}$ can be obtained as follows. Define the residuals $\hat{\varepsilon}_i = z_i - \hat{\phi} z_{i-1}$ for $i = 2, 3, \cdots, 120$. A bootstrap sample $z_1^*, z_2^*, \cdots, z_{120}^*$ is created by sampling $\hat{\varepsilon}_2^*, \hat{\varepsilon}_3^*, \cdots, \hat{\varepsilon}_{120}^*$ with replacement from the residuals, then letting $z_1^* = z_1$, and $z_i^* = \hat{\phi} z_{i-1}^* + \hat{\varepsilon}_i^*$, $i = 2, \cdots, 120$. Finally, after centering the time series $z_1^*, z_2^*, \cdots, z_{120}^*$, $\hat{\phi}^*$ is the estimate of the autoregressive parameter for this new time series. (We could, if we wished, sample the $\hat{\varepsilon}_i^*$ from a fitted normal distribution.)

A histogram of 1000 such bootstrap values $\hat{\phi}_1^*, \hat{\phi}_2^*, \cdots, \hat{\phi}_{1000}^*$ is shown in Fig. 10.

The bootstrap estimate of standard error was .055, agreeing nicely with the usual formula. Note however that the distribution is skewed to the left, so a confidence interval for ϕ might be asymmetric about $\hat{\phi}$ as discussed in Sections 8 and 9.

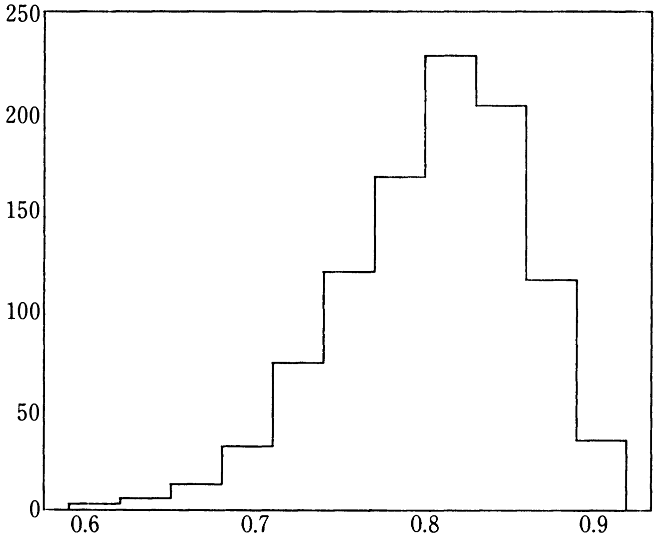

FIG. 10. *Bootstrap histogram of* $\hat{\phi}_1^*, \cdots, \hat{\phi}_{1000}^*$ *for the Wolfer sunspot data, model* (6.1).

In bootstrapping the residuals, we have assumed that the first-order autoregressive model is correct. (Recall the discussion of regression models in Section 5.) In fact, the first-order autoregressive model is far from adequate for this data. A fit of second-order autoregressive model

(6.2) $$z_i = \alpha z_{i-1} + \theta z_{i-2} + \varepsilon_i$$

gave estimates $\hat{\alpha} = 1.37$, $\hat{\theta} = -.677$, both with an estimated standard error of .067, based on Fisher information calculations. We applied the bootstrap to this model, producing the histograms for $\alpha_1^*, \cdots, \alpha_{1000}^*$ and $\theta_1^*, \cdots, \theta_{1000}^*$ shown in Figs. 11 and 12, respectively.

The bootstrap standard errors were .070 and .068, respectively, both close to the usual value. Note that the additional term has reduced the skewness of the first coefficient.

Example 2: Estimating a Response Transformation in Regression

Box and Cox (1964) introduced a parametric family for estimating a transformation of the response in a regression. Given regression data $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, their model takes the form

(6.3) $$z_i(\lambda) = x_i \cdot \beta + \varepsilon_i$$

where $z_i(\lambda) = (y_i^\lambda - 1)/\lambda$ for $\lambda \neq 0$ and $\log y_i$ for $\lambda = 0$, and $\varepsilon_i \sim$ iid $N(0, \sigma^2)$. Estimates of λ and β are found by minimizing $\sum_1^n (z_i - x_i \cdot \beta)^2$.

Breiman and Friedman (1985) proposed a nonparametric solution for this problem. Their so called ACE

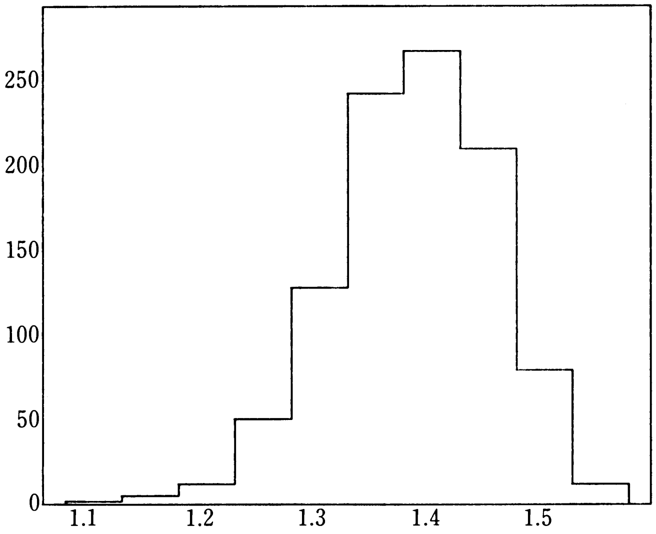

FIG. 11. *Bootstrap histogram of $\hat{\alpha}^*, \cdots, \hat{\alpha}^*_{1000}$ for the Wolfer sunspot data, model* (6.2).

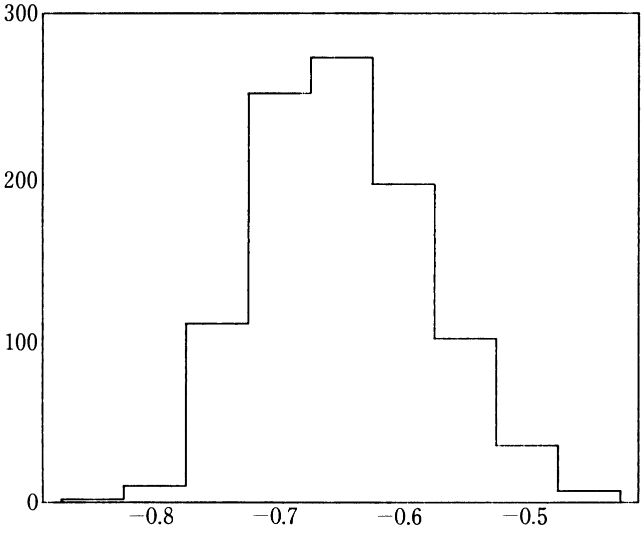

FIG. 12. *Bootstrap histogram of $\hat{\theta}^*, \cdots, \hat{\theta}^*_{1000}$ for the Wolfer sunspot data, model* (6.2).

(alternating conditional expectation) model generalizes (6.3) to

$$s(y_i) = x_i \cdot \beta + \varepsilon_i, \tag{6.4}$$

where $s(\cdot)$ is an unspecified smooth function. (In its most general form, ACE allows for transformations of the covariates as well.) The function $s(\cdot)$ and parameter β are estimated in an alternating fashion, utilizing a nonparametric smoother to estimate $s(\cdot)$.

In the following example, taken from Friedman and Tibshirani (1984), we compare the Box and Cox procedure to ACE and use the bootstrap to assess the variability of ACE.

The data from Box and Cox (1964) consist of a 3 × 3 × 3 experiment on the strength of yarns, the response Y being number of cycles to failure, and the factors length of test specimen (X_1) (250, 300, and 350 mm), amplitude of loading cycle (X_2) (8, 9, or 10 mm), and load (X_3) (40, 45, or 50 g). As in Box and Cox, we treat the factors as quantitive and allow only a linear term for each. Box and Cox found that a logarithmic transformation was appropriate, with their procedure producing a value of −.06 for $\hat{\lambda}$ with an estimated 95% confidence interval of (−.18, .06).

Fig. 13 shows the transformation selected by the ACE algorithm. For comparison, the log function is plotted (normalized) on the same figure.

The similarity is truly remarkable! In order to assess the variability of the ACE curve, we can apply the bootstrap. Since the X matrix in this problem is fixed by design, we resampled from the residuals instead of from the (x_i, y_i) pairs. The bootstrap procedure was the following:

```
Calculate residuals  $\hat{\varepsilon}_i = \hat{s}(y_i) - x_i \cdot \hat{\beta}$,  $i = 1, 2, \cdots, n$.
Repeat $B$ times
  Choose a sample  $\hat{\varepsilon}_1^*, \cdots, \hat{\varepsilon}_n^*$
                   with replacement from  $\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n$
  Calculate  $y_i^* = \hat{s}^{-1}(x_i \cdot \hat{\beta} + \hat{\varepsilon}_i^*)$,  $i = 1, 2, \cdots, n$
  Compute  $\hat{s}^*(\cdot)$ = result of ACE algorithm
                   applied to  $(x_1, y_1^*), \cdots, (x_n, y_n^*)$
End
```

The number of bootstrap replications B was 20. Note that the residuals are computed on the $s(\cdot)$ scale, not the y scale, because it is on the $s(\cdot)$ scale that the true residuals are assumed to be approximately iid. The 20 estimated transformations, $\hat{s}_1^*(\cdot), \cdots, \hat{s}_{20}^*(\cdot)$ are shown in Fig. 14.

The tight clustering of the smooths indicates that the original estimate $\hat{s}(\cdot)$ has low variability, especially for smaller values of Y. This agrees qualitatively with

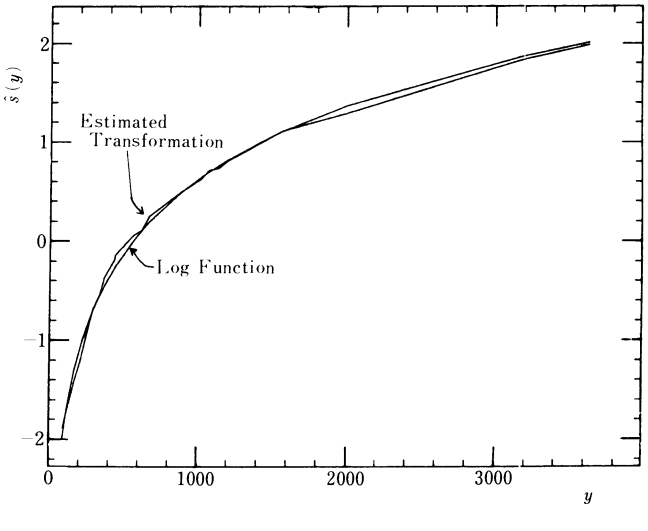

FIG. 13. *Estimated transformation from ACE and the log function for Box and Cox example.*

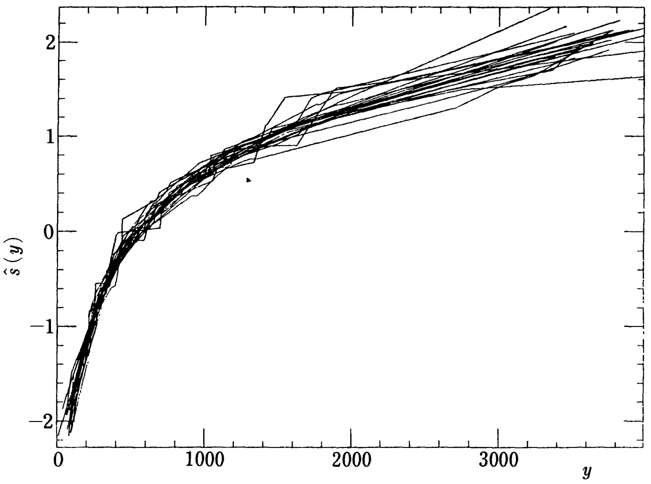

FIG. 14. *Bootstrap replications of ACE transformations for Box and Cox example.*

the short confidence interval for λ in the Box and Cox analysis.

7. BOOTSTRAP CONFIDENCE INTERVALS

This section presents three closely related methods of using the bootstrap to set confidence intervals. The discussion is in terms of simple parametric models, where the logical basis of the bootstrap methods is easiest to see. Section 8 extends the methods to multiparameter and nonparametric models.

We have discussed obtaining $\hat{\sigma}$, the estimated standard error of an estimator $\hat{\theta}$. In practice, $\hat{\theta}$ and $\hat{\sigma}$ are usually used together to form the approximate confidence interval $\theta \in \hat{\theta} \pm \hat{\sigma} z^{(\alpha)}$, (1.7), where $z^{(\alpha)}$ is the $100 \cdot \alpha$ percentile point of a standard normal distribution. The interval (1.7) is claimed to have approximate coverage probability $1 - 2\alpha$. For the law school example of Section 2, the values $\hat{\theta} = .776$, $\hat{\sigma} = .115$, $z^{(.05)} = -1.645$, give $\theta \in [.587, .965]$ as an approximate 90% central interval for the true correlation coefficient.

We will call (1.7) the *standard interval for* θ. When working within parametric families like the bivariate normal, $\hat{\sigma}$ in (1.7) is usually obtained by differentiating the log likelihood function, see Section 5a of Rao (1973), although in the context of this paper we might prefer to use the parametric bootstrap estimate of σ, e.g., $\hat{\sigma}_{\text{NORM}}$ in Section 2.

The standard intervals are an immensely useful statistical tool. They have the great virtue of being automatic: a computer program can be written which produces (1.7) directly from the data **y** and the form of the density function for **y**, with no further input required from the statistician. Nevertheless the standard intervals can be quite inaccurate as Table 5 shows. The standard interval (1.7), using $\hat{\sigma}_{\text{NORM}}$, (2.5), is strikingly different from the exact normal theory interval based on the assumption of a bivariate normal sampling distribution F.

TABLE 5
Exact and approximate central 90% *confidence intervals for* θ, *the true correlation coefficient, from the law school data of Fig.* 1

1. Exact (normal theory)	[.496, .898]	R/L = .44
2. Standard (1.7)	[.587, .965]	R/L = 1.00
3. Transformed standard	[.508, .907]	R/L = .49
4. Parametric bootstrap (BC)	[.488, .900]	R/L = .43
5. Nonparametric bootstrap (BC_a)	[.43, .92]	R/L = .42

Note: R/L = ratio of right side of interval, measured from $\hat{\theta} = .776$, to left side. The exact interval is strikingly asymmetric about $\hat{\theta}$. Section 8 discusses the nonparametric method of line 5.

In this case, it is well known that it is better to make the transformation $\hat{\phi} = \tanh^{-1}(\hat{\theta})$, $\phi = \tanh^{-1}(\theta)$, apply (1.7) on the ϕ scale, and then transform back to the θ scale. The resulting interval, line 3 of Table 5, is moved closer to the exact interval. However, there is nothing automatic about the $\tanh^{-1}$ transformation. For a different statistic from the correlation coefficient or a different distributional family from the bivariate normal, we might very well need other tricks to make (1.7) perform satisfactorily.

The bootstrap can be used to produce approximate confidence intervals in an automatic way. The following discussion is abridged from Efron (1984 and 1985) and Efron (1982a, Chapter 10). Line 4 of Table 5 shows that the parametric bootstrap interval for the correlation coefficient θ is nearly identical with the exact interval. "Parametric" in this case means that the bootstrap algorithm begins from the bivariate normal MLE $\hat{F}_{\text{NORM}}$, as for the normal theory curve of Fig. 2. This good performance is no accident. The bootstrap method used in line 4 in effect transforms $\hat{\theta}$ to the best (most normal) scale, finds the appropriate interval, and transforms this interval back to the θ scale. All of this is done automatically by the bootstrap algorithm, without requiring special intervention from the statistician. The price paid is a large amount of computing, perhaps $B = 1000$ bootstrap replications, as discussed in Section 10.

Define $\hat{G}(s)$ to be the parametric bootstrap cdf of $\hat{\theta}^*$,

$$\hat{G}(s) = \text{Prob}_*\{\hat{\theta}^* < s\}, \tag{7.1}$$

where Prob_* indicates probability computed according to the bootstrap distribution of $\hat{\theta}^*$. In Fig. 2 $\hat{G}(s)$ is obtained by integrating the normal theory curve. We will present three different kinds of bootstrap confidence intervals in order of increasing generality. All three methods use percentiles of $\hat{G}$ to define the confidence interval. They differ in which percentiles are used.

The simplest method is to take $\theta \in [\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)]$ as an approximate $1-2\alpha$ central interval for θ. This is called the *percentile method* in Section 10.4 of Efron (1982a). The percentile method interval is just the interval between the $100 \cdot \alpha$ and $100 \cdot (1-\alpha)$ percentiles of the bootstrap distribution of $\hat{\theta}^*$.

We will use the notation $\theta[\alpha]$ for the α level endpoint of an approximate confidence interval for θ, so $\theta \in [\theta[\alpha], \theta[1-\alpha]]$ is the central $1-2\alpha$ interval. Subscripts will be used to indicate the various different methods. The percentile interval has endpoints

$$\theta_P[\alpha] \equiv \hat{G}^{-1}(\alpha). \tag{7.2}$$

This compares with the standard interval,

$$\theta_S[\alpha] = \hat{\theta} + \hat{\sigma} z^{(\alpha)}. \tag{7.3}$$

Lines 1 and 2 of Table 6 summarize these definitions.

Suppose the bootstrap cdf $\hat{G}$ is perfectly normal, say

$$\hat{G}(s) = \Phi((s-\hat{\theta})/\hat{\sigma}), \tag{7.4}$$

where $\Phi(s) = \int_{-\infty}^{s} (2\pi)^{-1/2} e^{-t^2/2}\, dt$, the standard normal cdf. In other words, suppose that $\hat{\theta}^*$ has bootstrap distribution $N(\hat{\theta}, \hat{\sigma}^2)$. In this case the standard method and the percentile method agree, $\theta_S[\alpha] = \theta_P[\alpha]$. In situations like that of Fig. 2, where $\hat{G}$ is markedly nonnormal, the standard interval is quite different from (7.2). Which is better?

To answer this question, consider the simplest possible situation, where for all θ

$$\hat{\theta} \sim N(\theta, \sigma^2). \tag{7.5}$$

That is, we have a single unknown parameter θ with no nuisance parameters, and a single summary statistic $\hat{\theta}$ normally distributed about θ with constant standard error σ. In this case the parametric bootstrap cdf is given by (7.4), so $\theta_S[\alpha] = \theta_P[\alpha]$. (The bootstrap estimate $\hat{\sigma}$ equals σ.)

Suppose though that instead of (7.5) we have, for all θ,

$$\hat{\phi} \sim N(\phi, \tau^2), \tag{7.6}$$

for some monotone transformation $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$, where τ is a constant. In the correlation coefficient example the function g was $\tanh^{-1}$. The standard limits (7.2) can now be grossly inaccurate. However it is easy to verify that the percentile limits (7.2) are still correct. "Correct" here means that (7.2) is the mapping of the obvious interval for ϕ, $\hat{\phi} \pm \tau z^{(\alpha)}$, back to the θ scale, $\theta_P[\alpha] = g^{-1}(\hat{\phi} + \tau z^{(\alpha)})$. It is also correct in the sense of having exactly the claimed converge probability $1-2\alpha$.

Another way to state things is that the percentile intervals are transformation invariant,

$$\phi_P[\alpha] = g(\theta_P[\alpha]) \tag{7.7}$$

for any monotone transformation g. This implies that if the percentile intervals are correct on some transformed scale $\phi = g(\theta)$, then they must also be correct on the original scale θ. The statistician does not need to know the normalizing transformation g, only that it exists. Definition (7.2) automatically takes care of the bookkeeping involved in the use of normalizing transformations for confidence intervals.

Fisher's theory of maximum likelihood estimation says that we are always in situation (7.5) to a first order of asymptotic approximation. However, we are also in situation (7.6), for any choice of g, to the same order of approximation. Efron (1984 and 1985) uses higher order asymptotic theory to differentiate between the standard and bootstrap intervals. It is the higher order asymptotic terms which often make exact intervals strongly asymmetric about the MLE $\hat{\theta}$ as in Table 5. The bootstrap intervals are effective at capturing this asymmetry.

The percentile method automatically incorporates normalizing transformations, as in going from (7.5)–(7.6). It turns out that there are two other important ways that assumption (7.5) can be misleading, the first of which relates to possible bias in $\hat{\theta}$. For example consider $f_\theta(\hat{\theta})$, the family of densities for the observed correlation coefficient $\hat{\theta}$ when sampling $n = 15$ times from a bivariate normal distribution with true corre-

TABLE 6

Four methods of setting approximate confidence intervals for a real valued parameter θ

Method	Abbreviation	α level endpoint	Correct if	
1. Standard	$\theta_S[\alpha]$	$\hat{\theta} + \hat{\sigma} z^{(\alpha)}$	$\hat{\theta} \sim N(\theta, \sigma^2)$	σ constant
			There exists monotone transformation $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$ such that:	
2. Percentile	$\theta_P[\alpha]$	$\hat{G}^{-1}(\alpha)$	$\hat{\phi} \sim N(\phi, \tau^2)$	τ constant
3. Bias-corrected	$\theta_{BC}[\alpha]$	$\hat{G}^{-1}(\Phi\{2z_0 + z^{(\alpha)}\})$	$\hat{\phi} \sim N(\phi - z_0\tau, \tau^2)$	z_0, τ constant
4. BC$_a$	$\theta_{BC_a}[\alpha]$	$\hat{G}^{-1}\left(\Phi\left\{z_0 + \dfrac{(z_0 + z^{(\alpha)})}{1 - a(z_0 + z^{(\alpha)})}\right\}\right)$	$\hat{\phi} \sim N(\phi - z_0\tau_\phi, \tau_\phi^2)$ where $\tau_\phi = 1 + a\phi$	z_0, a constant

Note: Each method is correct under more general assumptions than its predecessor. Methods 2, 3, and 4 are defined in terms of the percentiles of $\hat{G}$, the bootstrap distribution (7.1).

lation θ. In fact it is easy to see that no monotone mapping $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$ transforms this family to $\hat{\phi} \sim N(\phi, \tau^2)$, as in (7.6). If there were such a g, then $\text{Prob}_\theta\{\hat{\theta} < \theta\} = \text{Prob}_\phi\{\hat{\phi} < \phi\} = .50$, but for $\theta = .776$ integrating the density function $f_{.776}(\hat{\theta})$ gives $\text{Prob}_{\theta=.776}\{\hat{\theta} < \theta\} = .431$.

The *bias-corrected percentile method* (BC method), line 3 of Table 6, makes an adjustment for this type of bias. Let

$$z_0 \equiv \Phi^{-1}\{\hat{G}(\hat{\theta})\}, \tag{7.8}$$

where Φ^{-1} is the inverse function of the standard normal cdf. The BC method has α level endpoint

$$\theta_{\text{BC}}[\alpha] \equiv \hat{G}^{-1}(\Phi\{2z_0 + z^{(\alpha)}\}). \tag{7.9}$$

Note: if $\hat{G}(\hat{\theta}) = .50$, that is if half of the bootstrap distribution of $\hat{\theta}^*$ is less than the observed value $\hat{\theta}$, then $z_0 = 0$ and $\theta_{\text{BC}}[\alpha] = \theta_P[\alpha]$. Otherwise definition (7.9) makes a bias correction.

Section 10.7 of Efron (1982a) shows that the BC interval for θ is exactly correct if

$$\hat{\phi} \sim N(\phi - z_0\tau, \tau^2) \tag{7.10}$$

for some monotone transformation $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$ and some constant z_0. It does not look like (7.10) is much more general than (7.6), but in fact the bias correction is often important.

In the example of Table 5, the percentile method (7.2) gives central 90% interval [.536, .911] compared to the BC interval [.488, .900] and the exact interval [.496, .898]. By definition the endpoints of the exact interval satisfy

$$\begin{aligned} \text{Prob}_{\theta=.496}\{\hat{\theta} > .776\} &= .05 \\ &= \text{Prob}_{\theta=.898}\{\hat{\theta} < .776\}. \end{aligned} \tag{7.11}$$

The corresponding quantities for the BC endpoints are

$$\begin{aligned} \text{Prob}_{\theta=.488}\{\hat{\theta} > .776\} &= .0465, \\ \text{Prob}_{\theta=.900}\{\hat{\theta} < .776\} &= .0475, \end{aligned} \tag{7.12}$$

compared to

$$\begin{aligned} \text{Prob}_{\theta=.536}\{\hat{\theta} > .776\} &= .0725, \\ \text{Prob}_{\theta=.911}\{\hat{\theta} < .776\} &= .0293. \end{aligned} \tag{7.13}$$

for the percentile endpoints. The bias correction is quite important in equalizing the error probabilities at the two endpoints. If z_0 can be approximated accurately (as mentioned in Section 9), then it is preferable to use the BC intervals.

Table 7 shows a simple example where the BC method is less successful. The data consists of the single observation $\hat{\theta} \sim \theta(\chi^2_{19}/19)$, the notation indicating an unknown scale parameter θ times a random variable with distribution $\chi^2_{19}/19$. (This definition makes $\hat{\theta}$ unbiased for θ.) A confidence interval is desired for the scale parameter θ. In this case the BC interval based on $\hat{\theta}$ is a definite improvement over the standard interval (1.7), but goes only about half as far as it should toward achieving the asymmetry of the exact interval.

TABLE 7

Central 90% confidence intervals for θ having observed $\hat{\theta} \sim \theta(\chi^2_{19}/19)$

1. Exact	$[.631 \cdot \hat{\theta}, 1.88 \cdot \hat{\theta}]$	R/L = 2.38
2. Standard (1.7)	$[.466 \cdot \hat{\theta}, 1.53 \cdot \hat{\theta}]$	R/L = 1.00
3. BC (7.9)	$[.580 \cdot \hat{\theta}, 1.69 \cdot \hat{\theta}]$	R/L = 1.64
4. BC_a (7.15)	$[.630 \cdot \hat{\theta}, 1.88 \cdot \hat{\theta}]$	R/L = 2.37
5. Nonparametric BC_a	$[.640 \cdot \hat{\theta}, 1.68 \cdot \hat{\theta}]$	R/L = 1.88

Note: The exact interval is sharply skewed to the right of $\hat{\theta}$. The BC method is only a partial improvement over the standard interval. The BC_a interval, $a = .108$, agrees almost perfectly with the exact interval.

It turns out that the parametric family $\hat{\theta} \sim \theta(\chi^2_{19}/19)$ cannot be transformed into (7.10), not even approximately. The results of Efron (1982b) show that there does exist a monotone transformation g such that $\hat{\phi} = g(\hat{\theta})$, $\phi = g(\theta)$ satisfy to a high degree of approximation

$$\hat{\phi} \sim N(\phi - z_0\tau_\phi, \tau_\phi^2) \quad (\tau_\phi = 1 + a\phi). \tag{7.14}$$

The constants in (7.14) are $z_0 = .1082$, $a = .1077$.

The BC_a method (Efron, 1984), line 4 of Table 6, is a method of assigning bootstrap confidence intervals which are exactly right for problems which can be mapped into form (7.14). This method has α level endpoint

$$\theta_{\text{BC}_a}[\alpha] \equiv \hat{G}^{-1}\left(\Phi\left\{z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}\right\}\right). \tag{7.15}$$

If $a = 0$ then $\theta_{\text{BC}_a}[\alpha] = \theta_{\text{BC}}[\alpha]$, but otherwise the BC_a intervals can be a substantial improvement over the BC method as shown in Table 7.

The constant z_0 in (7.15) is given by $z_0 = \Phi^{-1}\{\hat{G}(\hat{\theta})\}$, (7.8), and so can be computed directly from the bootstrap distribution. How do we know a? It turns out that in one-parameter families $f_\theta(\hat{\theta})$, a good approximation is

$$a \doteq \frac{\text{SKEW}_{\theta=\hat{\theta}}(\dot{l}_\theta(t))}{6}, \tag{7.16}$$

where $\text{SKEW}_{\theta=\hat{\theta}}(\dot{l}_\theta(t))$ is the skewness at parameter value $\theta = \hat{\theta}$ of the score statistic $\dot{l}_\theta(t) = (\partial/\partial\theta)\log f_\theta(t)$. For $\hat{\theta} \sim \theta(\chi^2_{19}/19)$ this gives $a \doteq .1081$, compared to the actual value $a = .1077$ derived in Efron (1984). For the normal theory correlation family of Table 5 $a \doteq 0$ which explains why the BC method, which takes $a = 0$, words so well there.

The advantage of formula (7.18) is that we need not know the transformation g leading to (7.14) in order to approximate a. In fact $\theta_{\mathrm{BC}_a}[\alpha]$, like $\theta_{\mathrm{BC}}[\alpha]$ and $\theta_P[\alpha]$, is transformation invariant, as in (7.7). Like the bootstrap methods, the BC$_a$ intervals are computed directly from the form of the density function $f_\theta(\cdot)$, for θ near $\hat{\theta}$.

Formula (7.16) applies to the case where θ is the only parameter. Section 8 briefly discusses the more challenging problem of setting confidence intervals for a parameter θ in a multiparameter family, and also in nonparametric situations where the number of nuisance parameters is effectively infinite.

To summarize this section, the progression from the standard intervals to the BC$_a$ method is based on a series of increasingly less restrictive assumptions, as shown in Table 6. Each successive method in Table 6 requires the statistician to do a greater amount of computation; first the bootstrap distribution $\hat{G}$, then the bias correction constant z_0, and finally the constant a. However, all of these computations are algorithmic in character, and can be carried out in an automatic fashion.

Chapter 10 of Efron (1982a) discusses several other ways of using the bootstrap to construct approximate confidence intervals, which will not be presented here. One of these methods, the "bootstrap t," was used in the blood serum example of Section 4.

8. NONPARAMETRIC AND MULTIPARAMETER CONFIDENCE INTERVALS

Section 7 focused on the simple case $\hat{\theta} \sim f_\theta$, where we have only a real valued parameter θ and a real valued summary statistic $\hat{\theta}$ from which we are trying to construct a confidence interval for θ. Various favorable properties of the bootstrap confidence intervals were demonstrated in the simple case, but of course the simple case is where we least need a general method like the bootstrap.

Now we will discuss the more common situation where there are nuisance parameters besides the parameter of interest θ; or even more generally the nonparametric case, where the number of nuisance parameters is effectively infinite. The discussion is limited to a few brief examples. Efron (1984 and 1985) develops the theoretical basis of bootstrap approximate confidence intervals for complicated situations, and gives many more examples. The word "approximate" is important here since exact nonparametric confidence intervals do not exist for most parameters (see Bahadur and Savage, 1956).

Example 1. Ratio Estimation

TABLE 8

Central 90% confidence intervals for $\theta = \eta_2/\eta_1$ and for $\phi = 1/\theta$ having observed $(y_1, y_2) = (8, 4)$ from a bivariate normal distribution $\mathbf{y} \sim N_2(\boldsymbol{\eta}, \mathbf{I})$

	For θ	For ϕ
1. Exact (Fieller)	[.29, .76]	[1.32, 3.50]
2. Parametric boot (BC)	[.29, .76]	[1.32, 3.50]
3. Standard (1.7)	[.27, .73]	[1.08, 2.92]
MLE	$\hat{\theta} = .5$	$\hat{\phi} = 2$

Note: The BC intervals, line 2, are based on the parametric bootstrap distribution of $\hat{\theta} = y_2/y_1$.

The data consists of $\mathbf{y} = (y_1, y_2)$, assumed to come from a bivariate normal distribution with unknown mean vector $\boldsymbol{\eta}$ and covariance matrix the identity,

$$\mathbf{y} \sim N_2(\boldsymbol{\eta}, \mathbf{I}). \tag{8.1}$$

The parameter of interest, for which we desire a confidence interval, is the ratio

$$\theta = \eta_2/\eta_1. \tag{8.2}$$

Fieller (1954) provided well known exact intervals for θ in this case. The Fieller intervals are based on a clever trick, which seems very special to situation (8.1), (8.2).

Table 8 shows Fieller's central 90% interval for θ having observed $\mathbf{y} = (8, 4)$. Also shown is the Fieller interval for $\phi = 1/\theta = \eta_1/\eta_2$, which equals $[.76^{-1}, .29^{-1}]$, the obvious transformation of the interval for θ. The standard interval (1.7) is satisfactory for θ, but not for ϕ. Notice that the standard interval does not transform correctly from θ to ϕ.

Line 2 shows the BC intervals based on applying definitions (7.8) and (7.9) to the parametric bootstrap distribution of $\hat{\theta} = y_2/y_1$ (or $\hat{\phi} = y_1/y_2$). This is the distribution of $\hat{\theta}^* = y_2^*/y_1^*$ when sampling $\mathbf{y}^* = (y_1^*, y_2^*)$ from $\hat{F}_{\mathrm{NORM}} \sim N_2((y_1, y_2), \mathbf{I})$. The bootstrap intervals transform correctly, and in this case they agree with the exact interval to three decimal places.

Example 2. Product of Normal Means

For most multiparameter situations, there do not exist exact confidence intervals for a single parameter of interest. Suppose for instance that (8.2) is changed to

$$\theta = \eta_1\eta_2, \tag{8.3}$$

still assuming (8.1). Table 9 shows approximate intervals for θ, and also for $\phi = \theta^2$, having observed $\mathbf{y} = (2, 4)$. The "almost exact" intervals are based on an analog of Fieller's argument (Efron, 1985), which with suitable care can be carried through to a high degree of accuracy. Once again, the parametric BC intervals are a close match to line 1. The fact that the standard intervals do not transform correctly is particularly obvious here.

TABLE 9
Central 90% confidence intervals for $\theta = \eta_1\eta_2$ *and* $\phi = \theta^2$ *having observed* $\mathbf{y} = (2, 4)$, *where* $\mathbf{y} \sim N_2(\boldsymbol{\eta}, \mathbf{I})$

	For θ	For ϕ
1. Almost exact	[1.77, 17.03]	[3.1, 290.0]
2. Parametric boot·(BC)	[1.77, 17.12]	[3.1, 239.1]
3. Standard (1.7)	[0.64, 15.36]	[−53.7, 181.7]
MLE	$\hat\theta \doteq 8$	$\hat\phi = 64$

Note: The almost exact intervals are based on the high order approximation theory of Efron (1985). The BC intervals of line 2 are based on the parametric bootstrap distribution of $\hat\theta = y_1y_2$.

The good performance of the parametric BC intervals is not accidental. The theory developed in Efron (1985) shows that the BC intervals, based on bootstrapping the MLE $\hat\theta$, agree to high order with the almost exact intervals in the following class of problems: the data $\mathbf{y}$ comes from a multiparameter family of densities $f_{\boldsymbol{\eta}}(\mathbf{y})$, both $\mathbf{y}$ and $\boldsymbol{\eta}$ k-dimensional vectors; the real valued parameter of interest θ is a smooth function of $\boldsymbol{\eta}$, $\theta = t(\boldsymbol{\eta})$; and the family $f_{\boldsymbol{\eta}}(\mathbf{y})$ can be transformed to multivariate normality, say

$$g(\mathbf{y}) \sim N_k(h(\boldsymbol{\eta}), \mathbf{I}), \tag{8.4}$$

by some one-to-one transformations g and h.

Just as in Section 7, it is not necessary for the statistician to know the normalizing transformations g and h, only that they exist. The BC intervals are obtained directly from the original densities $f_{\boldsymbol{\eta}}$: we find $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}(\mathbf{y})$, the MLE of $\boldsymbol{\eta}$; sample $\mathbf{y}^* \sim f_{\hat{\boldsymbol{\eta}}}$; compute $\hat\theta^*$, the bootstrap MLE of θ; calculate $\hat G$, the bootstrap cdf of $\hat\theta^*$, usually by Monte Carlo sampling, and finally apply definitions (7.8) and (7.9). This process gives the same interval for θ whether or not the transformation to form (8.4) has been made.

Not all problems can be transformed as in (8.4) to a normal distribution with constant covariance. The case considered in Table 7 is a one-dimensional counter example. As a result the BC intervals do not always work as well as in Tables 8 and 9, although they usually improve on the standard method. However, in order to take advantage of the BC$_a$ method, which is based on more general assumptions, we need to be able to calculate the constant a.

Efron (1984) gives expressions for "a" generalizing (7.16) to multiparameter families, and also to nonparametric situations. If (8.4) holds, then "a" will have value zero, and the BC$_a$ method reduces to the BC case. Otherwise the two intervals differ.

Here we will discuss only the nonparametric situation: the observed data $\mathbf{y} = (x_1, x_2, \cdots, x_n)$ consists of iid observations $X_1, X_2, \cdots, X_n \sim F$, where F can be *any* distribution on the sample space $\mathscr{X}$; we want a confidence interval for $\theta = t(F)$, some real valued functional of F; and the bootstrap interval are based on bootstrapping $\hat\theta = t(\hat F)$, which is the nonparametric MLE of θ. In this case a good approximation to the constant a is given in terms of the empirical influence function U_i^0, defined in Section 10 at (10.11),

$$a \doteq \frac{1}{6}\,\frac{\sum_{i=1}^n (U_i^0)^3}{\{\sum_{i=1}^n (U_i^0)^2\}^{3/2}}. \tag{8.5}$$

This is a convenient formula, since it is easy to numerically evaluate the U_i^0 by simply substituting a small value of θ into (10.11).

Example 3. The Law School Data

For $\hat\theta$ the correlation coefficient, the values of U_i^0 corresponding to the 15 data points shown in Fig. 1 are −1.507, .168, .273, .004, .525, −.049, −.100, .477, .310, .004, −.526, −.091, .434, .125, −.048. (Notice how influential law school 1 is.) Formula (8.5) gives $a \doteq -.0817$. $B = 100{,}000$ bootstrap replications, about 100 times more than was actually necessary (see Section 10), gave $z_Q = -.0927$, and the central 90% interval $\theta \in [.43, .92]$ shown in Table 5. The nonparametric BC$_a$ interval is quite reasonable in this example, particularly considering that there is no guarantee that the true law school distribution F is anywhere near bivariate normal.

Example 4. Mouse Leukemia Data (the First Example in Section 3)

The standard central 90% interval for β in formula (3.1) is [.835, 2.18]. The bias correction constant $z_0 \doteq .0275$, giving BC interval [1.00, 2.39]. This is shifted far right of the standard interval, reflecting the long right tail of the bootstrap histogram seen in Fig. 3. We can calculate "a" from (8.5), considering each of the $n = 42$ data points to be a triple (y_i, x_i, δ_i): $a \doteq -.152$. Because a is negative, the BC$_a$ interval is shifted back to the left, equaling [.788, 2.10]. This contrasts with the law school example, where a, z_0, and the skewness of the bootstrap distribution added to each other rather than cancelling out, resulting in a BC$_a$ interval much different from the standard interval.

Efron (1984) provides some theoretical support for the nonparametric BC$_a$ method. However the problem of setting approximate nonparametric confidence intervals is still far from well understood, and all methods should be interpreted with some caution. We end this section with a cautionary example.

Example 5. The Variance

Suppose $\mathscr{X}$ is the real line, and $\theta = \mathrm{Var}_F X$, the variance. Line 5 of Table 2 shows the result of applying the nonparametric BC$_a$ method to data sets $x_1, x_2, \cdots, x_{20}$ which were actually iid samples from a $N(0, 1)$ distribution. The number .640 for example is the average of $\theta_{\mathrm{BC}_a}[.05]/\hat{\theta}$ over 40 such data sets, $B = 4000$ bootstrap replications per data set. The upper limit $1.68 \cdot \hat{\theta}$ is noticeably small, as pointed out by Schenker (1985). The reason is simple: the nonparametric bootstrap distribution of $\hat{\theta}^*$ has a short upper tail; compared to the parametric bootstrap distribution which is a scaled χ^2_{19} random variable. The results of Beran (1984), Bickel and Freedman (1981), and Singh (1981) show that the nonparametric bootstrap distribution is highly accurate asymptotically, but of course that is not a guarantee of good small sample behavior. Bootstrapping from a smoothed version of $\hat{F}$, as in lines 3, 4, and 5 of Table 2 alleviates the problem in this particular example.

9. BOOTSTRAP SAMPLE SIZES

How many bootstrap replications must we take? Consider the standard error estimate $\hat{\sigma}_B$ based on B bootstrap replications, (2.4). As $B \to \infty$, $\hat{\sigma}_B$ approaches $\hat{\sigma}$, the bootstrap estimate of standard error as originally defined in (2.3). Because $\hat{F}$ does not estimate F perfectly, $\hat{\sigma} = \sigma(\hat{F})$ will have a non-zero coefficient of variation for estimating the true standard error $\sigma = \sigma(F)$; $\hat{\sigma}_B$ will have a larger CV because of the randomness added by the Monte Carlo bootstrap sampling.

It is easy to derive the following approximation,

$$\text{(9.1)} \qquad \mathrm{CV}(\hat{\sigma}_B) \doteq \left\{ \mathrm{CV}(\hat{\sigma})^2 + \frac{E\{\hat{\delta}\} + 2}{4B} \right\}^{1/2},$$

where $\hat{\delta}$ is the kurtosis of the bootstrap distribution of $\hat{\theta}^*$, given the data $\mathbf{y}$, and $E\{\hat{\delta}\}$ its expected value averaged over $\mathbf{y}$. For typical situations, $\mathrm{CV}(\hat{\sigma})$ lies between .10 and .30. For example, if $\hat{\theta} = \bar{x}$, $n = 20$, $x_i \sim_{\text{iid}} N(0, 1)$, then $\mathrm{CV}(\hat{\sigma}) \doteq .16$.

Table 10 shows $\mathrm{CV}(\hat{\sigma}_B)$ for various values of B and $\mathrm{CV}(\hat{\sigma})$, assuming $E\{\hat{\delta}\} = 0$ in (9.1). For values of $\mathrm{CV}(\hat{\sigma}) > .10$, *there is little improvement past* $B = 100$. In fact B as small as 25 gives reasonable results. Even smaller values of B can be quite informative, as we saw in the Stanford Heart Transplant Data (Fig. 7 of Section 3).

TABLE 10

Coefficient of variation of $\hat{\sigma}_B$, the bootstrap estimate of standard error based on B Monte Carlo replications, as a function of B and $\mathrm{CV}(\hat{\sigma})$*, the limiting CV as* $B \to \infty$

		$B \to$				
		25	50	100	200	∞
$\mathrm{CV}(\hat{\sigma})$	.25	.29	.27	.26	.25	.25
$\downarrow$	.20	.24	.22	.21	.21	.20
	.15	.21	.18	.17	.16	.15
	.10	.17	.14	.12	.11	.10
	.05	.15	.11	.09	.07	.05
	0	.14	.10	.07	.05	0

Note: Based on (9.1), assuming $E\{\hat{\delta}\} = 0$.

The situation is quite different for setting bootstrap confidence intervals. The calculations of Efron (1984), Section 8, show that $B = 1000$ is a rough minimum for the number of Monte Carlo bootstraps necessary to compute the BC or BC_a intervals. Somewhat smaller values, say $B = 250$, can give a useful percentile interval, the difference being that then the constant z_0 need not be computed. Confidence intervals are a fundamentally more ambitious measure of statistical accuracy than standard errors, so it is not surprising that they require more computational effort.

10. THE JACKKNIFE AND THE DELTA METHOD

This section returns to the simple case of assigning a standard error to $\hat{\theta}(y)$, where $\mathbf{y} = (x_1, \cdots, x_n)$ is obtained by random sampling from a single unknown distribution, $X_1, \cdots, X_n \sim_{\text{iid}} F$. We will give another description of the bootstrap estimate $\hat{\sigma}$, which illustrates the bootstrap's relationship to older techniques of assigning standard errors, like the jackknife and the delta method.

For a given bootstrap sample $\mathbf{y}^* = (x_1^*, \cdots, x_n^*)$, as described in step (i) of the algorithm in Section 2, let p_i^* indicate the proportion of the bootstrap sample equal to x_i,

$$\text{(10.1)} \qquad p_i^* = \frac{\#\{x_j^* = x_i\}}{n} \quad i = 1, 2, \cdots, n,$$

$\mathbf{p}^* = (p_1^*, p_2^*, \cdots, p_n^*)$. The vector $\mathbf{p}^*$ has a rescaled multinomial distribution

$$\text{(10.2)} \qquad \begin{gathered} \mathbf{p}^* \sim \mathrm{Mult}_n(n, \mathbf{p}^0)/n \\ (\mathbf{p}^0 = (1/n, 1/n, \cdots, 1/n)), \end{gathered}$$

where the notation indicates the proportions observed from n random draws on n categories, each with probability $1/n$.

For $n = 3$ there are 10 possible bootstrap vectors $\mathbf{p}^*$. These are indicated in Fig. 15 along with their multinomial probabilities from (10.2). For example, $\mathbf{p}^* = (1/3, 0, 2/3)$, corresponding to $\mathbf{x}^* = (x_1, x_3, x_3)$ or any permutation of these values has bootstrap probability $1/9$.

To make our discussion easier suppose that the statistic of interest $\hat{\theta}$ is of functional form: $\hat{\theta} = \theta(\hat{F})$, where $\theta(F)$ is a functional assigning a real number to any distribution F on the sample space X. The mean, the correlation coefficient, and the trimmed mean are all of functional form. Statistics of functional form have the same value as a function of $\hat{F}$, no matter what the sample size n may be, which is convenient for discussing the jackknife and delta method.

For any vector $\mathbf{p} = (p_1, p_2, \cdots, p_n)$ having nonnegative weights summing to 1, define the weighted

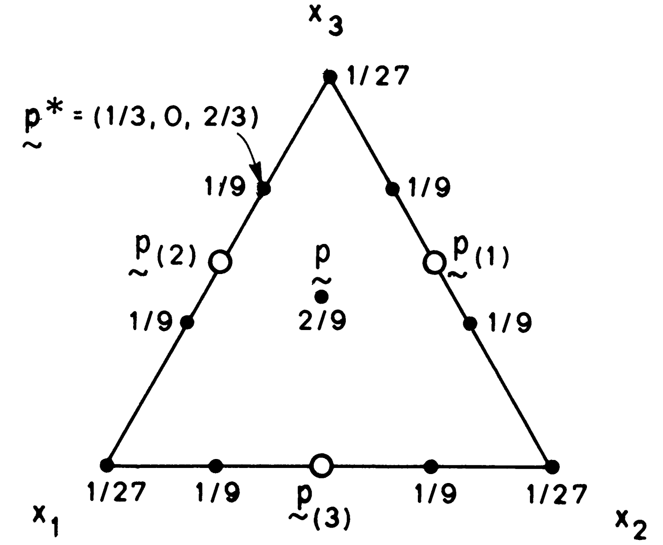
x_3
1/27
$\underset{\sim}{p}^* = (1/3, 0, 2/3)$
1/9
1/9
$\underset{\sim}{p}_{(2)}$
$\underset{\sim}{p}$
$\underset{\sim}{p}_{(1)}$
1/9
2/9
1/9
1/27
1/9
$\underset{\sim}{p}_{(3)}$
1/9
1/27
x_1
x_2

FIG. 15. *The bootstrap and jackknife sampling points in the case* $n = 3$. *The bootstrap points* $(\cdot)$ *are shown with their probabilities.*

empirical distribution

(10.3) $\hat{F}(\mathbf{p})$: probability p_i on x_i $\quad i = 1, \cdots, n.$

For $\mathbf{p} = \mathbf{p}^0 = \mathbf{1}/n$, the weighted empirical distribution equals $\hat{F}$, (1.4).

Corresponding to $\mathbf{p}$ is a resampled value of $\hat{\theta}$,

$$(10.4) \qquad \hat{\theta}(\mathbf{p}) \equiv \theta(\hat{F}(\mathbf{p})).$$

The shortened notation $\hat{\theta}(\mathbf{p})$ assumes that the data $(x_1, x_2, \cdots, x_n)$ is considered fixed. Notice that $\hat{\theta}(\mathbf{p}^0) = \theta(\hat{F})$ is the observed value of the statistic of interest. The bootstrap estimate $\hat{\sigma}$, (2.3), can then be written

$$(10.5) \qquad \hat{\sigma} = [\mathrm{var}_* \hat{\theta}(\mathbf{p}^*)]^{1/2},$$

where var_* indicates variance with respect to distribution (10.2). In terms of Fig. 15, $\hat{\sigma}$ is the standard deviation of the ten possible bootstrap values $\hat{\theta}(\mathbf{p}^*)$ weighted as shown.

It looks like we could always calculate $\hat{\sigma}$ simply by doing a finite sum. Unfortunately, the number of bootstrap points is $\binom{2n-1}{n}$, 77,558,710 for $n = 15$ so straightforward calculation of $\hat{\sigma}$ is usually impractical. That is why we have emphasized Monte Carlo approximations to $\hat{\sigma}$. Therneau (1983) considers the question of methods more efficient than pure Monte Carlo, but at present there is no generally better method available.

However, there is another approach to approximating (10.5). We can replace the usually complicated function $\hat{\theta}(\mathbf{p})$ by an approximation linear in $\mathbf{p}$, and then use the well known formula for the multinomial variance of a linear function. The *jackknife approximation* $\hat{\theta}_J(\mathbf{p})$ is the linear function of $\mathbf{p}$ which matches $\hat{\theta}(\mathbf{p})$, (10.4), at the n points corresponding to the deletion of a single $\mathbf{x}_i$ from the observed data set $x_1, x_2, \cdots, x_n$,

$$(10.6) \qquad \mathbf{p}_{(i)} = \frac{1}{n-1}(1, 1, \cdots, 1, 0, 1, \cdots, 1)$$

$i = 1, 2, \cdots, n$. Fig. 15 indicates the jackknife points for $n = 3$; because $\hat{\theta}$ is the functional form, (10.4), it does not matter that the jackknife points correspond to sample size $n - 1$ rather than n.

The linear function $\hat{\theta}_J(\mathbf{p})$ is calculated to be

$$(10.7) \qquad \hat{\theta}_J(\mathbf{p}) = \hat{\theta}_{(\cdot)} + (\mathbf{p} - \mathbf{p}^0) \cdot \mathbf{U}$$

where, in terms of $\hat{\theta}_{(i)} \equiv \hat{\theta}(\mathbf{p}_{(i)})$, $\hat{\theta}_{(\cdot)} = \sum_{i=1}^n \hat{\theta}_{(i)}/n$, and $\mathbf{U}$ is the vector with ith coordinate

$$(10.8) \qquad U_i = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}).$$

The jackknife estimate of standard error (Tukey, 1958; Miller, 1974) is

$$(10.9) \quad \hat{\sigma}_J \equiv \left[\frac{n-1}{n} \sum_{i=1}^n \{\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\}^2\right]^{1/2} = \left[\frac{\sum_i^n U_i^2}{n(n-1)}\right]^{1/2}.$$

A standard multinomial calculation gives the following theorem (Efron, 1982a),

THEOREM. *The jackknife estimate of standard error equals* $[n/(n-1)]^{1/2}$ *times the bootstrap estimate of standard error for* $\hat{\theta}_J$,

$$(10.10) \qquad \hat{\sigma}_J = \left[\frac{n}{n-1} \mathrm{var}_* \hat{\theta}_J(\mathbf{p}^*)\right]^{1/2}.$$

In other words, the jackknife estimate is itself almost a bootstrap estimate applied to a linear approximation of $\hat{\theta}$. The factor $[n/(n-1)]^{1/2}$ in (10.10) makes $\hat{\sigma}_J^2$ unbiased for σ^2 in the case where $\hat{\theta} = \bar{x}$, the sample mean. We could multiply the bootstrap estimate $\hat{\sigma}$ by this same factor, and achieve the same unbiasedness, but there does not seem to be any consistent advantage to doing so. The jackknife requires n, rather than $B = 50$ to 200 resamples, at the expense of adding a linear approximation to the standard error estimate. Tables 1 and 2 indicate that there is some estimating efficiency lost in making this approximation. For statistics like the sample median which are difficult to approximate linearly, the jackknife is useless (see Section 3.4 of Efron, 1982a).

There is a more obvious linear approximation to $\hat{\theta}(\mathbf{p})$ than $\hat{\theta}_J(\mathbf{p})$. Why not use the first-order Taylor series expansion for $\hat{\theta}(\mathbf{p})$ about the point $\mathbf{p} = \mathbf{p}^0$? This is the idea of Jaeckel's *infinitesimal jackknife* (1972). The Taylor series approximation turns out to be

$$\hat{\theta}_T(\mathbf{p}) = \hat{\theta}(\mathbf{p}^0) + (\mathbf{p} - \mathbf{p}^0)'\mathbf{U}^0$$

where

$$(10.11) \qquad U_i^0 = \lim_{\varepsilon \to 0} \frac{\hat{\theta}((1-\varepsilon)\mathbf{p}^0 + \varepsilon\delta_i) - \hat{\theta}(\mathbf{p}^0)}{\varepsilon},$$

δ_i being the ith coordinate vector. This suggests the infinitesimal jackknife estimate of standard error

$$\text{(10.12)} \quad \hat{\sigma}_{IJ} \equiv [\text{var}_* \hat{\theta}_T(\mathbf{p}^*)]^{1/2} = [\Sigma U_i^{02}/n^2]^{1/2}$$

with var_* still indicating variance under (10.2). The ordinary jackknife can be thought of as taking $\varepsilon = -1/(n-1)$ in the definition of U_i^0, while the infinitesimal jackknife lets $\varepsilon \to 0$, thereby earning the name.

The U_i^0 are values of what Mallows (1974) calls the empirical influence function. Their definition is a nonparametric estimate of the true influence function

$$IF(x) = \lim_{\varepsilon \to 0} \frac{\theta((1-\varepsilon)F + \varepsilon\delta_x) - \theta(F)}{\varepsilon},$$

δ_x being the degenerate distribution putting mass 1 on x. The right side of (10.12) is then the obvious estimate of the influence function approximation to the standard error of $\hat{\theta}$ (Hampel, 1974), $\sigma(F) \doteq [\int IF^2(x)\, dF(x)/n]^{1/2}$. The empirical influence function method and the infinitesimal jackknife give identical estimates of standard error.

How have statisticians gotten along for so many years without methods like the jackknife and the bootstrap? The answer is the delta method, which is still the most commonly used device for approximating standard errors. The method applies to statistics of the form $t(\bar{Q}_1, \bar{Q}_2, \cdots, \bar{Q}_A)$, where $t(\cdot, \cdot, \cdots, \cdot)$ is a known function and each $\bar{Q}_a$ is an observed average, $\bar{Q}_a = \sum_{i=1}^n Q_a(X_i)/n$. For example, the correlation $\hat{\theta}$ is a function of $A = 5$ such averages; the average of the first coordinate values, the second coordinates, the first coordinates squared, the second coordinates squared, and the cross-products.

In its nonparametric formulation, the delta method works by (a) expanding t in a linear Taylor series about the expectations of the $\bar{Q}_a$; (b) evaluating the standard error of the Taylor series using the usual expressions for variances and covariances of averages; and (c) substituting $\gamma(\hat{F})$ for any unknown quantity $\gamma(F)$ occurring in (b). For example, the nonparametric delta method estimates the standard error of the correlation $\hat{\theta}$ by

$$\left\{\frac{\hat{\theta}^2}{4n}\left[\frac{\hat{\mu}_{40}}{\hat{\mu}_{20}^2} + \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} + \frac{2\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} + \frac{4\hat{\mu}_{22}}{\hat{\mu}_{11}^2} - \frac{4\hat{\mu}_{31}}{\hat{\mu}_{11}\hat{\mu}_{02}} + \frac{4\hat{\mu}_{13}}{\hat{\mu}_{11}\hat{\mu}_{02}}\right]\right\}^{1/2}$$

where, in terms of $x_i = (y_i, z_i)$,

$$\hat{\mu}_{gh} \equiv \Sigma(y_i - \bar{y})^g (z_i - \bar{z})^h / n$$

(Cramér (1946), p. 359).

THEOREM. *For statistics of the form $\hat{\theta} = t(\bar{Q}_1, \cdots, \bar{Q}_A)$, the nonparametric delta method and the infinitesimal jackknife give the same estimate of standard error* (Efron, 1982c).

The infinitesimal jackknife, the delta method, and the empirical influence function approach are three names for the same method. Notice that the results reported in line 7 of Table 2 show a severe downward bias. Efron and Stein (1981) show that the ordinary jackknife is always biased upward, in a sense made precise in that paper. In the authors' opinion the ordinary jackknife is the method of choice if one does not want to do the bootstrap computations.

ACKNOWLEDGMENT

This paper is based on a previous review article appearing in *Behaviormetrika*. The authors and Editor are grateful to that journal for graciously allowing this revision to appear here.

REFERENCES

ANDERSON, O. D. (1975). *Time Series Analysis and Forecasting: The Box–Jenkins Approach.* Butterworth, London.

BAHADUR, R. and SAVAGE, L. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.* **27** 1115–1122.

BERAN, R. (1984). Bootstrap methods in statistics. *Jahrb. Math. Ver.* **86** 14–30.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. R. Statist. Soc. Ser. B* **26** 211–252.

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619.

COX, D. R. (1972). Regression models and life tables. *J. R. Statist. Soc. Ser. B* **34** 187–202.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press, Princeton, New Jersey.

EFRON, B. (1979a). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7** 1-26.

EFRON, B. (1979b). Computers and the theory of statistics: thinking the unthinkable. *Soc. Ind. Appl. Math.* **21** 460–480.

EFRON, B. (1981a). Censored data and the bootstrap. *J. Amer. Statist. Assoc.* **76** 312–319.

EFRON, B. (1981b). Nonparametric estimates of standard error: the jackknife, the bootstrap, and other resampling methods, *Biometrika* **68** 589–599.

EFRON, B. (1982a). The jackknife, the bootstrap, and other resampling plans. *Soc. Ind. Appl. Math. CBMS-Natl. Sci. Found. Monogr.* **38**.

EFRON, B. (1982b). Transformation theory: how normal is a one parameter family of distributions? *Ann. Statist.* **10** 323–339.

EFRON, B. (1982c). Maximum likelihood and decision theory. *Ann. Statist.* **10** 340–356.

EFRON, B. (1983). Estimating the error rate of a prediction rule: improvements in cross-validation. *J. Amer. Statist. Assoc.* **78** 316–331.

EFRON, B. (1984). Better bootstrap confidence intervals. Tech. Rep. Stanford Univ. Dept. Statist.

EFRON, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* **72** 45–58.

EFRON, B. and GONG, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *Amer. Statistician* **37** 36–48.

EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596.

FIELLER, E. C. (1954). Some problems in interval estimation. *J. R. Statist. Soc. Ser. B* **16** 175–183.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

FRIEDMAN, J. H. and TIBSHIRANI, R. J. (1984). The monotone smoothing of scatter-plots. *Technometrics* **26** 243-250.

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383-393.

HASTIE, T. J. and TIBSHIRANI, R. J. (1985). Discussion of Peter Huber's "Projection Pursuit." *Ann. Statist.* **13** 502-508.

HINKLEY, D. V. (1978). Improving the jackknife with special reference to correlation estimation. *Biometrika* **65**, 13-22.

HYDE, J. (1980). *Survival Analysis with Incomplete Observations. Biostatistics Casebook.* Wiley, New York.

JAECKEL, L. (1972). The infinitesimal jackknife. Memorandum MM 72-1215-11. Bell Laboratories, Murray Hill, New Jersey.

JOHNSON, N. and KOTZ, S. (1970). *Continuous Univariate Distributions.* Houghton Mifflin, Boston, Vol. 2.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete samples. *J. Amer. Statist. Assoc.* **53** 457-481.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887-906.

MALLOWS, C. (1974). On some topics in robustness. Memorandum, Bell Laboratories, Murray Hill, New Jersey.

MILLER, R. G. (1974). The jackknife—a review. *Biometrika* **61** 1-17.

MILLER, R. G. and HALPERN, J. (1982). Regression with censored data. *Biometrika* **69** 521-531.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications.* Wiley, New York.

SCHENKER, N. (1985). Qualms about bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **80** 360-361.

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187-1195.

THERNEAU, T. (1983). Variance reduction techniques for the bootstrap. Ph.D. thesis, Stanford University, Department of Statistics.

TIBSHIRANI, R. J. and HASTIE, T. J. (1984). Local likelihood estimation. Tech. Rep. Stanford Univ. Dept. Statist. **97**.

TUKEY, J. (1958). Bias and confidence in not quite large samples, abstract. *Ann. Math. Statist.* **29** 614.

Comment

J. A. Hartigan

Efron and Tibshirani are to be congratulated on a wide-ranging persuasive survey of the many uses of the boostrap technology. They are a bit cagey on what is or is not a bootstrap, but the description at the end of Section 4 seems to cover all the cases; some data y comes from an unknown probability distribution F; it is desired to estimate the distribution of some function $R(y, F)$ given F; and this is done by estimating the distribution of $R(y^*, \hat{F})$ given $\hat{F}$ where $\hat{F}$ is an estimate of F based on y, and y^* is sampled from the known $\hat{F}$.

There will be three problems in any application of the bootstrap: (1) how to choose the estimate $\hat{F}$? (2) how much sampling of y^* from $\hat{F}$? and (3) how close is the distribution of $R(y^*, \hat{F})$ given $\hat{F}$ to $R(y, F)$ given F?

Efron and Tibshirani suggest a variety of estimates $\hat{F}$ for simple random sampling, regression, and autoregression; their remarks about (3) are confined mainly to empirical demonstrations of the bootstrap in specific situations.

I have some general reservations about the bootstrap based on my experiences with subsampling techniques (Hartigan, 1969, 1975). Let $X_1, \ldots, X_n$ be a random sample from a distribution F, let F_n be the empirical distribution, and suppose that $t(F_n)$ is an estimate of some population parameter $t(F)$. The statistic $t(\hat{F}_n)$ is computed for several random subsamples (each observation appearing in the subsample with probability ½), and the set of $t(\hat{F}_n)$ values obtained is regarded as a sample from the posterior distribution of $t(F)$. For example, the standard deviation of the $t(\hat{F}_n)$ is an estimate of the standard error of $t(F_n)$ from $t(F)$; however, the procedure is not restricted to real valued t.

The procedure seems to work not too badly in getting at the first- and second-order behaviors of $t(F_n)$ when $t(F_n)$ is near normal, but it not effective in handling third-order behavior, bias, and skewness. Thus there is not much point in taking huge samples $t(\hat{F}_n)$ since the third-order behavior is not relevant; and if the procedure works only for $t(F_n)$ near normal, there are less fancy procedures for estimating standard error such as dividing the sample up into 10 subsamples of equal size and computing their standard deviation. (True, this introduces more bias than having random subsamples each containing about half the observations.) Indeed, even if $t(F_n)$ is not normal, we can obtain exact confidence intervals for the median of $t(F_{n/10})$ using the 10 subsamples. Even five subsamples will give a respectable idea of the standard error.

Transferring back to the bootstrap: (A) is the boot-

J. A. Hartigan is Eugene Higgins Professor of Statistics, Yale University, Box 2179 Yale Station, New Haven, CT 06520.