SIX QUESTIONS RAISED BY THE BOOTSTRAP

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 139

August 1990

PREPARED UNDER THE AUSPICES

OF

PUBLIC HEALTH SERVICE GRANT 5 R01 GM21215–16

DIVISION OF BIOSTATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

Six Questions Raised by the Bootstrap

By

Bradley Efron

Technical Report No. 139

August 1990

Prepared Under the Auspices

Of

Public Health Service Grant 5 R01 GM21215-16

Also supported by National Science Foundation Grant DMS89–05874 and issued as Technical Report No. 350, Department of Statistics, Stanford University.

Division of Biostatistics

Stanford University

Stanford, California

SIX QUESTIONS RAISED BY THE BOOTSTRAP

B. Efron

Abstract

Investigations of bootstrap methods often raise more general questions in statistical inference. This talk discusses six such questions: (1) Why do distributions estimated by maximum likeklihood tend to have too short tails? (2) Why does the delta method tend to underestimate standard errors? (3) Why are cross-validation estimates so variable? (4) What is a "correct" confidence interval? (5) What is a good nonparametric pivotal quantity? (6) Can we get bootstrap-like answers without Monte Carlo?

Introduction. Working on the bootstrap tends to raise broader questions of statistical theory. This paper considers six such questions. The first three worry about standard methods, and how they sometimes fail us. The second three questions concern matters not much examined by standard theory. Only the last question directly concerns the bootstrap, but bootstrap considerations appear in all six.

The title may give the impression that six answers will be provided. The actual number is closer to 1.5. My hope in presenting this paper is to attract more solutions, or at least more interest, in some questions that seem to me to be of considerable importance.

1. Why do maximum likelihood estimated distributions tend to be short-tailed? Estimates of probability distributions obtained by maximum likelihood tend to be more concentrated than the distributions themselves. Here is a familiar example: suppose that $y_1, y_2, \cdots, y_n$ is an independent and identically distributed (i.i.d.) sample from an unknown distribution F,

$$F \overset{\text{i.i.d.}}{\rightarrow} y_1, y_2, \cdots, y_n. \tag{1.1}$$

The nonparametric maximum likelihood estimate (MLE) of F is the empirical distribution $\hat{F}$,

$$\hat{F}: \quad \text{probability } 1/n \text{ on } y_i, \qquad i = 1, 2, \cdots, n. \tag{1.2}$$

The empirical distribution assigns probability

$$\text{Prob}_{\hat{F}}\{A\} = \frac{\#\{y_i \in A\}}{n} \tag{1.3}$$

to any set A in the sample space of the $y's$. This is an unbiased estimate of the true probability $\text{Prob}_F\{A\}$,

$$E\{\text{Prob}_{\hat{F}}\{A\}\} = \text{Prob}_F\{A\}. \tag{1.4}$$

However the same unbiasedness does not apply to the variance functional: $\mathrm{Var}_{\hat{F}}\{Y\} = \sum_{i=1}^{n}(y_i - \bar{y})^2/n$ has expectation

$$E\,\mathrm{Var}_{\hat{F}}\{Y\} = \frac{n-1}{n}\mathrm{Var}_F\{Y\}. \tag{1.5}$$

We see that the variance function is underestimated by maximum likelihood, albeit mildly so. Elementary statistics courses recommend estimating the variance by

$$\frac{n}{n-1}\mathrm{Var}_{\hat{F}}\{Y\} = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1), \tag{1.6}$$

rather than by $\mathrm{Var}_{\hat{F}}\{Y\}$ itself. This removes the bias, but doesn't explain it.

The underestimation effect seen in (1.5) is quite small from the point of view of asymptotic theory. Most often we are estimating $\sigma^2 = \mathrm{Var}_F\{Y\}$ in order to form a confidence interval for the expectation $\mu = E_F\{Y\}$, say of the form

$$\bar{y} \pm z^{(\alpha)}\widehat{\mathrm{se}}, \tag{1.7}$$

where $z^{(\alpha)}$ is the $100\alpha^{\text{th}}$ normal percentile, e.g. $z^{(.95)} = 1.645$, and $\widehat{\mathrm{se}} = \hat{\sigma}/\sqrt{n}$, the estimated standard error of $\bar{y}$.

The choice of the MLE versus the unbiased estimate of σ^2 makes a difference of magnitude only

$$O(\widehat{\mathrm{se}}/n) \tag{1.8}$$

in (1.7). This is a third order correction in the usual parlance. By comparison, ignoring the skewness of F in forming interval (1.7) makes a difference of $O(\widehat{\mathrm{se}}/\sqrt{n})$, a second order error; and using $\bar{y}$ instead of an efficient estimate $\hat{\mu}$, makes a difference of $O(\widehat{\mathrm{se}})$, a first order error. (See Efron (1987).) Third order notwithstanding, the underestimation effect in (1.5) can cause substantial problems when n is small.

We can exacerbate the phenomenon seen in (1.5) by considering a standard linear model,

$$y_i = x_i\beta + e_i, \tag{1.9}$$

where x_i is an observed p-dimensional covariate vector, β is an unknown p-dimensional parameter vector, and the e_i are an i.i.d. sample from some unknown distribution F with mean 0 and variance σ^2. Suppose we assume that F is normal, $F \sim N(0, \sigma^2)$. Then the MLE for σ^2 is

$$\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2/n, \tag{1.10}$$

having expectation

$$E\{\hat{\sigma}^2\} = \frac{n-p}{n}\sigma^2. \tag{1.11}$$

If p is a large function of n, then the MLE $\hat{\sigma}^2$ badly underestimates the dispersion parameter σ^2. Nevertheless, with p fixed and n going to infinity, (1.11) still represents a third-order error.

The underestimation seen in (1.5) gets worse as we consider measures of dispersion more extreme than the variance.

Table 1 shows the results of a Monte Carlo experiment involving 100 samples, each of which consisted of 10 independent observations from a standard exponential distribution. Four functionals of F were considered, the mean, the standard deviation, the skewness and the kurtosis. The sampling experiment shows that skew($\hat{F}$) badly underestimates skew(F), the situation being worse for the kurtosis. Fisher's theory of k-statistics, Kendall and Stuart, Chapter 12, (1958), reduces the bias of skew($\hat{F}$) and kurt($\hat{F}$), without explaining it.

	Mean	Sd	Skew	Kurt
True F:	1	1	2	6
Ave $\hat{F}$:	1.01	.863	.910	.052
%<True:	55%	71%	93%	100%

Table 1. Sampling Experiment. 100 i.i.d. samples of size $n = 10$, from a standard exponential distribution. For each sample, the mean, standard deviation, skewness, and kurtosis of $\hat{F}$ were evaluated. Shown are the average values of these four functionals, and the proportion of the 100 samples for which the sample functional was less than the true functional.

My own concern with the dispersion-reducing tendencies of maximum likelihood estimation stems from its effect on bootstrap confidence intervals. The simplest way to form a bootstrap confidence interval is to use the order statistics of the bootstrap replications. Suppose $\theta = t(F)$ is a real-valued parameter of interest, such as a mean, a correlation, an eigenvalue, etc., estimated by $\hat{\theta} = t(\hat{F})$. We draw some large number "B" of independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \cdots, \mathbf{x}^{*B}$, where in the one-sample situation (1.1), each $\mathbf{x}^*$ is a random sample of size n from $\hat{F}$,

$$\hat{F} \overset{\text{i.i.d.}}{\to} (y_1^*, y_2^*, \cdots, y_n^*) \equiv \mathbf{x}^*. \tag{1.12}$$

The y_i do not have to be real-valued; F can be a distribution on an arbitrary probability space.

Each bootstrap sample gives a bootstrap replication of $\hat{\theta} = t(\hat{F})$, namely $\hat{\theta}^{*b} = t(\hat{F}^{*b})$, where $\hat{F}^{*b}$ is the empirical distribution corresponding to the b^{th} bootstrap sample $\mathbf{x}^{*b}$. The percentile confidence interval, of approximate coverage 90%, is defined to be the central 90% range of the ordered bootstrap replication. In the usual order-statistic notation, the 90% percentile interval is $[\hat{\theta}^*_{(.05B)}, \hat{\theta}^*_{(.95B)}]$.

The length of the percentile interval,

$$\widehat{\text{length}} = \hat{\theta}^*_{(.95B)} - \hat{\theta}^*_{(.05B)} \tag{1.13}$$

is a dispersion functional of $\hat{F}$, similar to $\text{Var}_{\hat{F}}\{Y\}$ except that it is usually evaluated by a Monte Carlo process. Like $\text{Var}_{\hat{F}}$, $\widehat{\text{length}}$ often tends to be a little too small, leading to undercoverage for the percentile confidence interval.

Once again, this effect is tiny from the asymptotic point of view. The standard confidence interval for a parameter θ,

$$\hat{\theta} \pm z^{(\alpha)}\widehat{\text{se}}, \tag{1.14}$$

where $\hat{\theta} = \theta(\hat{F})$ and $\widehat{\text{se}}$ is any reasonably efficient estimate of standard error for $\hat{\theta}$, is typically first-order accurate: its non-coverage probability, in each tail, approaches the nominal value $1 - \alpha$

at rate $1/\sqrt{n}$. The bootstrap confidence interval "BC_a", Efron (1987), which is an improved version of the percentile interval, is second-order accurate: its non-coverage probabilities approach $(1-\alpha)$ at the factor rate $1/n$. Its coverage errors are of the third order, $O(\frac{1}{n})$, corresponding to the third order shortening effect seen in (1.5). However, even third-order errors can be consequential in small samples. This is especially true if the confidence intervals are used for hypothesis testing purposes, where small errors in the end-points can change the verdict of the test.

There are other third-order effects that produce coverage errors of order $1/n$ in the BC_a intervals. Efron (1985) details the third-order errors for bootstrap confidence intervals in a certain class of parametric problems. Perhaps the moral here is that "plug-in" methods, like maximum likelihood and the bootstrap, give answers accurate to the second-order but not to the third; and that the tendency toward shortness of maximum likelihood estimated distributions accounts for part of the third-order error.

2. Why does the delta method tend to underestimate standard errors? The delta method is the oldest device for assessing the standard errors of complicated statistical estimates, predating the jackknife and bootstrap methods by at least 150 years. The delta method is also known as the method of statistical differentials, propagation of errors formula, and the Taylor series method. Despite its ancient pedigree, I have found the delta method to be less reliable than the jackknife or the bootstrap, with an occasional tendency to badly underestimate the true standard error.

Table 2 shows the results of all four sampling experiments in Efron (1982) that compare the delta method, bootstrap, and jackknife estimates of standard error with the true standard error. In the first experiment, each sample consisted of ten independent pairs (x_i, y_i), with x_i having a uniform distribution on $[0,1]$, independent of y_i, which had a G_1 (one-sided exponential) distribution. The ratio statistic $\hat{\theta} = \bar{y}/\bar{x}$ then had true standard error .37, as verified by extensive Monte Carlo sampling. Each sample in the experiment gave a delta method, bootstrap, and jackknife estimate of standard error. In this case all three methods were nearly unbiased, giving average standard error estimates of .35, .37, and .37 respectively. Details of the four sampling experiments appear in Efron (1982), Section 3.

In the last three sampling experiments, the delta method has a noticeable downward bias. This is particularly evident in the last case, where the statistic of interest was the $\tanh^{-1}$ transformation of a simple correlation coefficient from 15 independent bivariate normal points.

A puzzling aspect of the differences seen in Table 2 is that the delta method is intimately related to both the jackknife and the bootstrap. In fact the delta method is identical to Jaeckel's (1972) infinitesimal jackknife. Suppose $\hat{\theta}$ is a functional statistic $\hat{\theta} = S(\hat{F})$, such as the ordinary mean $S(\hat{F}) = \int x d\hat{F} = \bar{x}$. Let $\hat{F}_i^\epsilon$ be a distorted version of the empirical distribution $\hat{F}$ that puts extra probability on the ith data point,

$$\hat{F}_i^\epsilon = \begin{cases} (1-\epsilon)/n + \epsilon & \text{probability on } x_i \\ (1-\epsilon)/n & \text{probability on } x_j,\ j \neq i. \end{cases} \tag{2.1}$$

The empirical influence function is the derivative

statistic	se delta	estimate boot	jack	True se	n	Sample consists of n independent pairs (x_i, y_i), with distribution
$\bar{y}/\bar{x}$	.35	.37	.37	.37	10	$x_i \sim U(0,1)$ indep. of $y_i \sim G_1$
$\bar{y}/\bar{x}$	.53	.64	.70	.67	10	$x_i \sim U(0,1)$ indep. of $y_i \sim G_1^2/2$
sample correlation	.175	.206	.223	.218	14	$(x_i, y_i) \sim N_2(\mu, \Sigma)$ true corr = .5
$\tanh^{-1}$ (corr)	.244	.301	.314	.297	14	$(x_i, y_i) \sim N_2(\mu, \Sigma)$, true corre = .5

Table 2. Four sampling experiments comparing the delta method, bootstrap, and jackknife estimates of standard error, from Efron (1982). Number tabled is the average estimate in the sampling experiment. The delta method gives the smallest estimates. In the last three cases, the delta method estimates are considerably too small.

$$U_i \equiv \frac{\partial}{\partial \epsilon} S(\hat{F}_i^{\epsilon})|_{\epsilon=0}. \tag{2.2}$$

Efron (1981) showed that the usual nonparametric delta function estimate of standard error using a linear Taylor series expansion of $\hat{\theta}$ is identical to the infinitesimal jackknife formula suggested by Jaeckel,

$$\text{se}_{\text{delta}}\{\hat{\theta}\} \equiv [\sum_{i=1}^{n} U^2/n^2]^{1/2}. \tag{2.3}$$

Notice that setting $\epsilon = -1/(n-1)$ in (2.1) gives the empirical distribution for the reduced data set with the ith point removed,

$$\hat{F}_i^{-(1/(n-1))} \equiv \hat{F}_{(i)} = \begin{cases} 0 & \text{probability on } x_i \\ \frac{1}{n-1} & \text{probability on } x_j, \ j \neq i, \end{cases} \tag{2.4}$$

Let $\hat{\theta}_{(i)} = S(\hat{F}_{(i)})$ and $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$. Tukey's jackknife estimate of standard error is

$$\text{se}_{\text{jack}}\{\hat{\theta}\} = \{\frac{n-1}{n} \Sigma[\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}]^2\}^{1/2}. \tag{2.5}$$

This is almost the same as (2.3), except that ϵ in (2.2) has been set equal to $-1/(n-1)$ instead of going to zero. (Tukey's formula also incorporates an extra factor of $n/(n-1)$, in order that $\text{se}_{\text{jack}}\{\bar{x}\}$ exactly equal the usual formula for the standard error of the mean, $[\Sigma(x_i - \bar{x})^2/n(n-1)]^{1/2}$.)

Efron and Stein (1981) showed that the jackknife estimates of variance, $\text{se}_{\text{jack}}\{\hat{\theta}\}^2$, tends to be biased upward as an estimate of the variance. A moderate upward bias is discernable in the jackknife column of Table 2. The close relationship of definitions (2.5) and (2.3) suggests that the jackknife and delta method should behave similarly, but such is not always the case.

The bootstrap method gives the nonparametric maximum likelihood estimate (MLE) of standard error: let $\text{se}\{t; F\}$ indicate the true standard error of a statistic $\hat{\theta} = t(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is an i.i.d. sample from F; then the bootstrap estimate of standard error is

$$\text{se}_{\text{boot}}\{\hat{\theta}\} = \text{se}\{t; \hat{F}\}, \tag{2.6}$$

where $\hat{F}$ is the nonparametric MLE of F, i.e. the empirical distribution of the sample $x_1, x_2, \cdots, x_n$. In other words, $\text{se}_{\text{boot}}\{\hat{\theta}\}$ is the nonparametric MLE of the true standard error.

We expect maximum likelihood estimates to be nearly unbiased, as demonstrated by the bootstrap column of Table 2. The next argument shows the close theoretical connection between all three estimators, the bootstrap, the delta method, and the jackknife. This deepens the puzzle of the delta method's poor performance.

A bootstrap sample is a random sample of size n from $\hat{F}$,

$$\hat{F} \overset{\text{i.i.d.}}{\to} (x_1^*, x_2^*, \cdots, x_n^*) \equiv \mathbf{x}^*. \tag{2.7}$$

Corresponding to $\mathbf{x}^*$ is the bootstrap replication $\hat{\theta}^* = t(\mathbf{x}^*)$. Definition (2.6) is an ideal version of $\text{se}_{\text{boot}}\{\hat{\theta}\}$, which is usually approximated by Monte Carlo: independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \cdots, \mathbf{x}^{*B}$ give independent bootstrap replications $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \cdots, \hat{\theta}^{*B}$; then

$$\text{se}_{\text{boot}}\{\hat{\theta}\} \doteq [\sum_{b=1}^{B} (\hat{\theta}^{*b} - \hat{\theta}^{*\cdot})^2/(B-1)]^{1/2}, \tag{2.8}$$

where $\hat{\theta}^{*\cdot} \equiv \sum_b \hat{\theta}^{*b}/B$. Expression (2.8) approaches (2.6) as $B \to \infty$. Usually B in the range $50 - 200$ gives good results, see Efron (1987).

Each member of a bootstrap sample $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$ is randomly selected from the original data set $(x_1, x_2, \cdots, x_n)$. Let $\mathbf{P}^*$ be the resampling vector $(P_1^*, P_2^*, \cdots, P_n^*)$, where

$$P_i^* = \#\{x_j^* = x_i\}/n, \tag{2.9}$$

the proportion of the bootstrap sample equalling x_i. Then $\mathbf{P}^*$ has a rescaled multinomial distribution, of n draws on n categories each with probability $1/n$,

$$\mathbf{P}^* \sim \text{Mult}(n, \mathbf{P}^o)/n \qquad [\mathbf{P}^o \equiv (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})]. \tag{2.10}$$

With the original sample $\mathbf{x}$ fixed, we can think of $\hat{\theta}^* = t(\mathbf{x}^*)$ as a function of $\mathbf{P}^*$, say $\hat{\theta}(\mathbf{P}^*)$. Another way to state (2.6) is

$$\text{se}_{\text{boot}}\{\hat{\theta}\} = [\text{Var}_*\{\hat{\theta}(\mathbf{P}^*)\}]^{1/2}, \tag{2.11}$$

where Var_* indicates variance under the multinomial distribution (2.11).

Figure 1 schematically represents (2.11). The prone triangle is the simplex $\mathcal{S}$ in which $\mathbf{P}^*$ takes its values. The function $\hat{\theta}(\mathbf{P}^*)$ is represented by a curved surface over $\mathcal{S}$. At the center of $\mathcal{S}$ is $\mathbf{P}^o$, with $\hat{\theta}(\mathbf{P}^o) = t(\mathbf{x})$, the actual simple value of the statistic. We usually need to approximate (2.11) by Monte Carlo because there is no closed formula for the variance of a non-linear function of a multinomial vector.

Both the jackknife and the delta method avoid Monte Carlo by approximating $\hat{\theta}(\mathbf{P}^*)$ with a linear function of $\mathbf{P}^*$, say $\hat{\theta}_{\text{LIN}}(\mathbf{P}^*)$; and then theoretically evaluating $[\text{Var}_*\{\hat{\theta}_{\text{LIN}}(\mathbf{P}^*)\}]^{1/2}$, instead of (2.11), from the usual formula for the variance of a linear function of a multinomial.

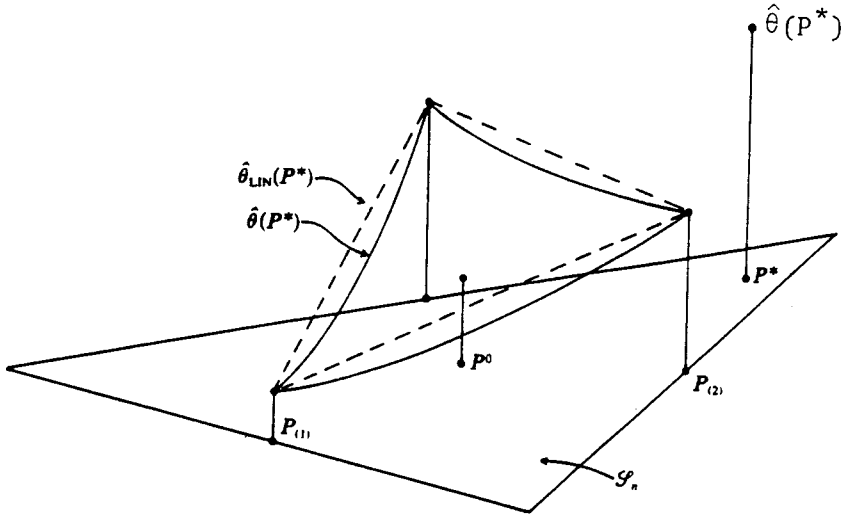

Figure 1. Schematic representation of bootstrap resampling; resampling vector $\mathbf{P}^*$ takes its values in $\mathcal{S}$; $\hat{\theta}(\mathbf{P}^*)$ is the curved surface over $\mathcal{S}$; bootstrap standard error is square root of variance of $\hat{\theta}(\mathbf{P}^*)$ where $\mathbf{P}^*$ has multinomial distribution (2.10). Dashed lines indicate the linear approximation to $\hat{\theta}(\mathbf{P}^*)$ used by the jackknife. The delta method uses the linear function tangent to the resampling surface at the central point $(\mathbf{P}^o, \hat{\theta}(\mathbf{P}^o))$.

The delta method standard error is based on the most obvious linear approximation to $\hat{\theta}(\mathbf{P}^*)$: the tangent hyperplane to the surface $\hat{\theta}(\mathbf{P}^*)$ through the central point $(\mathbf{P}^o, \hat{\theta})$. The jackknife uses the hyperplane which equals $\hat{\theta}(\mathbf{P}^*)$ at the n "jackknife points" $\mathbf{P}_{(1)}, \mathbf{P}_{(2)}, \cdots, \mathbf{P}_{(n)}$, where $\mathbf{P}_{(i)}$ is

$$(1, 1, \cdots, 1, 0, 1, \cdots, 1)/(n-1), \tag{2.12}$$

0 in the ith place. Once again, it is difficult to see why se_{jack} should be biased upwards and se_{delta} biased downwards.

We can set up an artificial estimation problem which gives some insight into the relationship between se_{delta} and se_{boot}. Consider the data $\mathbf{x}$ as fixed, and suppose that we observe a multinomial vector $\mathbf{P} \sim \text{Mult}(n, \underset{\sim}{\pi})/n$, where now $\underset{\sim}{\pi}$ can be any vector in the simplex $\mathcal{S}$, not just $\mathbf{P}^o$ as in (2.10). The artificial problem is to estimate $\hat{\theta}(\underset{\sim}{\pi})$ having observed $\mathbf{P}$.

The MLE of $\hat{\theta}(\underset{\sim}{\pi})$ in the artificial problem is $\hat{\theta}(\mathbf{P})$. It then turns out that $[\text{se}_{\text{boot}}\{\hat{\theta}\}]^2$ is the variance of the MLE when $\underset{\sim}{\pi} = \mathbf{P}^o$, while $[\text{se}_{\text{delta}}\{\hat{\theta}\}]^2$ is the Cramer-Rao lower bound for the unbiased estimation of $\hat{\theta}(\underset{\sim}{\pi})$, at $\underset{\sim}{\pi} = \mathbf{P}^o$. This argument makes it plausible that se_{delta} would usually be less than se_{boot}. "Plausible" isn't a proof though, and it isn't true that $\text{se}_{\text{delta}} < \text{se}_{\text{boot}}$ in every case.

The delta method is much too useful a tool to throw away. However, it's numerical results shouldn't be accepted uncritically, since they seem liable to underestimation. The jackknife and bootstrap standard errors are both more dependable.

3. Why are cross-validation estimators so variable? Cross-validation, like the delta method, is a time-honored technique for assessing statistical error. Modern computational equipment has greatly increased the use of cross-validation for choosing estimation rules and estimating their prediction errors. "Time-honored" is not the same as "tried and true". My own experience, as described in this section, is that cross-validation can be undependable in some situations, and that substantially better methods are available. The discussion here is taken from Efron (1983) and Chapter 7 of Efron (1982). Good references for cross-validation include Stone (1974), Geisser (1975), and Lachenbruch and Mickey (1968).

Figure 2 shows a sample obtained in a simulation experiment: we observe a training set of data (y_i, z_i), $i = 1, 2, \cdots, 14$, where

$$y_i = \begin{cases} 0 & \text{with probability } 1/2 \\ 1 & \text{with probability } 1/2, \end{cases} \tag{3.1}$$

and z_i is bivariate normal, with identity covariance, and mean depending on y_i,

$$z_i | y_i \sim N_2((y_i - \frac{1}{2}, 0), I) \tag{3.2}$$

The two means, $(\frac{1}{2}, 0)$ and $(-\frac{1}{2}, 0)$, are indicated by stars in Figure 2. In real practice, of course, we wouldn't know the probability mechanism generating the data.

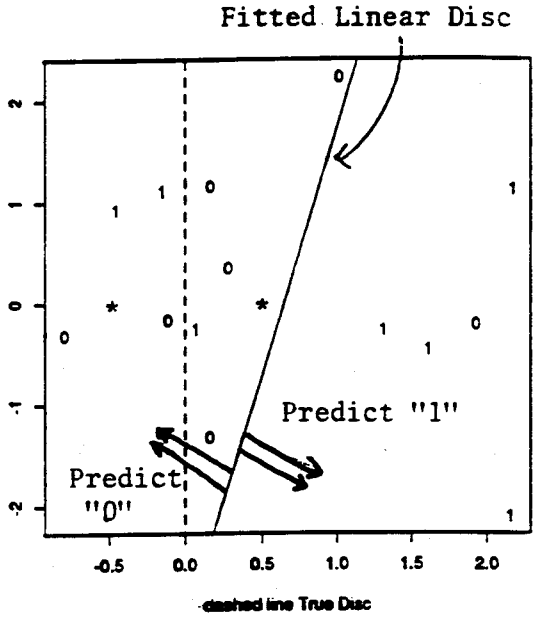

Figure 2. Fisher's linear discriminant (solid line) fit to 14 data points generated according to (3.1), (3.2); stars indicate mean values of the "0" and "1" distributions; the apparent error rate is 4/14 in this case.

Fisher's linear discriminant divides the prediction space, $\mathcal{R}^2$ in this case, into two regions, for the purpose of predicting the y value of a future pair (y, z), having observed only z. In practice, z

might be some observable predictors, like age, sex, weight, etc. and y some dichotomous outcome that we want to predict, like the success or failure of a medical procedure.

The linear discriminant predicts 1 or 0 depending on the size of a certain linear function,

$$\widehat{\text{pred}}(z) = \begin{cases} 1 & if \quad a + b'z \geq 0 \\ 0 & if \quad a + b'z < 0. \end{cases} \tag{3.3}$$

The constant a and the vector b are functions of the training set $\{(y_i, z_i), i = 1, 2, \cdots, 14\}$, see (2.13) of Efron (1983). One of the main uses of cross-validation is to assess the prediction error of a data- based prediction rule like (3.3).

The True Error Rate of a prediction rule is the probability that it misclassifies a future observation,

$$\text{True Error} = \text{Prob}\{y \neq \widehat{\text{pred}}(z)\}. \tag{3.4}$$

The probability in (3.4) is conditional on the training set; $\widehat{\text{pred}}()$ is fixed, only y and z being random. For the sample in Figure 2, True Error = .344, as calculated from (3.1), (3.2). True Error rates cannot be calculated in practice since we usually don't know the probability mechanism generating (y, z).

An obvious estimate of True Error is the Apparent Error rate

$$\text{AppError} = \#\{y_i \neq \widehat{\text{pred}}(z_i)\}/14, \tag{3.5}$$

the proportion of the training set points misclassified by $\widehat{\text{pred}}$. Four of the 14 points are misclassified in Figure 2, three 1's and one 0, so AppError = 4/14 = .286. In this case the difference between True and Apparent rates,

$$\text{Diff} = \text{True} - \text{App}, \tag{3.6}$$

is positive. Diff is usually positive, at least in expectation, because $\widehat{\text{pred}}$ is fit to the training set, and so does better in "predicting" the training set than in predicting a genuinely new point.

Cross-validation gives a way of using the training set to obtain a nearly unbiased estimate of Diff. Let $\widehat{\text{pred}}_{(i)}$ be the prediction rule obtained from the reduced training set that excludes point (y_i, z_i). In our example

$$\widehat{\text{pred}}_{(i)}(z) = \begin{cases} 1 & \text{if} \quad a_i + b'_{(i)}z \geq 0, \\ 0 & \text{if} \quad a_i + b'_{(i)}z < 0, \end{cases} \tag{3.7}$$

where $a_{(i)}, b_{(i)}$ are the coefficients of Fisher's linear discriminant based on $(y_1, z_1), (y_2, z_2), \cdots, (y_{i-1}, z_{i-1}), (y_{i+1}, z_{i+1}), \cdots, (y_{14}, z_{14})$. The cross-validation estimate of Diff is

$$\widehat{\text{Diff}}_{\text{CV}} = \#\{y_i \neq \widehat{\text{pred}}_{(i)}(z_i)\}/14 - \text{AppError}. \tag{3.8}$$

Intuitively, the first term of (3.8) should be a newly unbiased estimate of True Error, so $\widehat{\text{Diff}}$ should be nearly unbiased for Diff. We can illustrate this with a sampling experiment.

The sampling experiment comprised 100 independent replications of (3.1), (3.2). In other words, 100 independent realizations of Figure 2 were constructed. Table 3 shows the results. We see that the average of $\widehat{\text{Diff}}_{\text{CV}}$, the cross-validation estimate, averaged .091, compared to .096 for the true Diff. This corroborates the approximate unbiasedness of $\widehat{\text{Diff}}_{\text{CV}}$.

		True Err	App err	Diff	$\widehat{\text{Diff}}$ CV	$\widehat{\text{Diff}}$ .632
trial		.....		....	.....	.....
	[1,]	.458	.286	.172	.214	.083
	[2,]	.312	.357	-.045	.000	.068
	[3,]	.313	.357	-.044	.071	.095
	[4,]	.351	.429	-.078	.071	.051
	[5,]	.330	.357	-.027	.143	.094
	[6,]	.318	.143	.175	.214	.086
	[7,]	.310	.071	.239	.071	.083
	[8,]	.380	.286	.094	.071	.130
	[9,]	.360	.429	-.069	.071	.119
	[10,]	.335	.143	.192	.000	.042
		.....		....	.....	.....
all 100	mean:	.360	.264	.096	.091	.076
	sd:	.045	.123	.113	.073	.035
	RMSE:				.149	.117

Table 3. One hundred replications of (3.1), (3.2); shown are results of the first 10 replications, and summary statistics for all 100. The cross-validation estimte of Diff, "CV", is nearly unbiased for the true Diff, but has a large standard deviation.

Table 3 also shows that $\widehat{\text{Diff}}_{\text{CV}}$ is highly variable: its standard deviation over the 100 simulations was .073, about 80% as big as its mean .091. Unbiased or not, this makes $\widehat{\text{Diff}}_{\text{CV}}$ an undependable estimator of Diff.

Efron (1983) shows that cross-validation is closely related to the bootstrap, much as the delta method, jackknife, and bootstrap are related in Figure 1. This leads to several new estimators for Diff, based on variants of the underlying bootstrap argument. The most successful of these, "$\widehat{\text{Diff}}_{.632}$", also appears in Table 3. We see that it is moderately biased downward, but has much smaller standard deviation than $\widehat{\text{Diff}}_{\text{CV}}$. An objective way to compare the two procedures is in terms of their root mean square errors for estimating True Error = AppError + Diff,

$$\text{RMSE} = [E\{\text{True} - (\text{App} + \widehat{\text{Diff}})\}^2]^{1/2} \tag{3.9}$$

we see that RMSE was .149 for cross-validation, compared to .117 for the .632 rule.

Five sampling experiments are considered in Efron (1983), and eight estimators of Diff. $\widehat{\text{Diff}}_{.632}$ was clearly the winner in terms of RMSE, and cross-validation was even more clearly the loser. In two of the five experiments, using cross-validation was considerably worse than simply estimating True by App, i.e. taking $\widehat{\text{Diff}} = 0$.

We have another small mystery here: all of the Diff estimates, including cross-validation, are variants of the same bootstrap argument, and yet they perform quite differently in small-sample simulation experiments. My belief, or hope, is that future research will produce a dependable improvement over cross-validation. The problem of estimating the prediction error of a data-based prediction rule is important enough to discuss further study.

4. What is a correct confidence interval? Suppose we wish to construct a 90% central confidence interval for a real-valued parameter θ, having observed some data $\mathbf{x}$. A proposed interval $[\hat{\theta}_{\mathrm{LO}}(\mathbf{x}), \hat{\theta}_{\mathrm{UP}}(\mathbf{x})]$ is said to be accurate if it fails to cover θ exactly 5% of the time in each direction,

$$\mathrm{Prob}\{\theta > \hat{\theta}_{\mathrm{UP}}(\mathbf{x})\} = \mathrm{Prob}\{\theta < \hat{\theta}_{\mathrm{LO}}(\mathbf{x})\} = .05. \tag{4.1}$$

An accurate confidence interval is not necessarily a correct one, though. If $x_1, x_2, \cdots, x_{10}$ is a random sample from a $N(\theta, \gamma)$ distribution, with both the mean θ and the variance γ unknown, then the student's t interval for θ based on only $x_1, x_2, \cdots, x_5$ is obviously accurate in the sense of (4.1). Equally obvious, it is not the correct interval for θ. It is inferentially wrong, though probabilistically right.

The question of correctness arises forcibly in the construction of approximate confidence intervals. Various theories have been put forth to construct intervals accurate to a high degree of asymptotic approximation. We will discuss some of these theories in this section, and Sections 5 and 6 as well. Are the intervals they construct highly correct as well as highly accurate? Answering this question seems crucial if we are to avoid the pitfall of the student's t example above.

The notion of interval correctness is more difficult to pin down than interval accuracy. Correctness is clear-cut only in the simplest situations, where the data $\mathbf{x}$ can be reduced to a one-dimension sufficient statistic $\hat{\theta}$, and where the percentiles of $\hat{\theta}$ increase monotonically as a function of θ. Then the textbook method of confidence interval construction, taking $\hat{\theta}_{\mathrm{LO}}$ to be that θ for which the observed value $\hat{\theta}$ is at the 95th percentile of possible outcomes, and similarly for $\hat{\theta}_{\mathrm{HI}}$, gives what most statisticians would call the correct confidence interval for θ. Sometimes more complicated-looking situations can be reduced to the simple form. In a bivariate normal model for example, the maximum likelihood estimate $\hat{\rho}$ of the true correlation ρ has a distribution depending only on ρ.

Fieller's construction for the ratio of normal means gives an example of a correct confidence interval in a genuinely multiparametric situation. Suppose y is bivariate normal with identity covariance matrix, $y \sim N_2(\mu, I)$, and we want a confidence interval for the ratio $\theta(\mu) = \mu_2/\mu_1$. The level surface $\{\theta(\mu) = \theta_0\}$ is a straight line passing through the origin at angle $\tan^{-1}(\theta_0)$, as shown in the left panel of Figure 3.

There is an obvious way to test the hypothesis $H_0 : \theta(\mu) = \theta_0$. Let D_0 be the signed distance from y to the straight line $\{\theta(\mu) = \theta_0\}$. [Almost any sign convention will do, for instance $\mathrm{sign}(D_0) = \mathrm{sign}(y_2 - y_{02})$, where y_0 is the nearest point to y on $\{\theta(\mu) = \theta_0\}$.] Then D_0 has a standard $N(0,1)$ distribution if $\mu \in \{\theta(\mu) = \theta_0\}$. The obvious .90 two-sided test rejects H_0 if y lies outside the band $|D_0| < 1.645$. Fieller's confidence interval comprises those values θ_0 such that the test accepts the hypothesis $\theta(\mu) = \theta_0$. In other words, it is these values θ_0 such that the distance from y to $\{\theta(\mu) = \theta_0\}$ is less than 1.645, as shown in the right panel of Figure 3. See Section 5.14 of Miller (1986). Most, though not all, statisticians find the pivotality of D_0 an irresistable argument for the correctness of the Fieller intervals.

Fieller's construction depends on the level surface $\{\theta(\mu) = \theta_0\}$ being straight lines. Suppose we consider a parameter for which the level surface $C_{\theta_0} = \{\theta(\mu) = \theta_0\}$ are curved, for example $\theta(\mu) = \mu_1\mu_2$, where the C_{θ_0} are hyperbolae. Efron (1985) discusses an approximate version of Fieller's construction applying to the curved case.

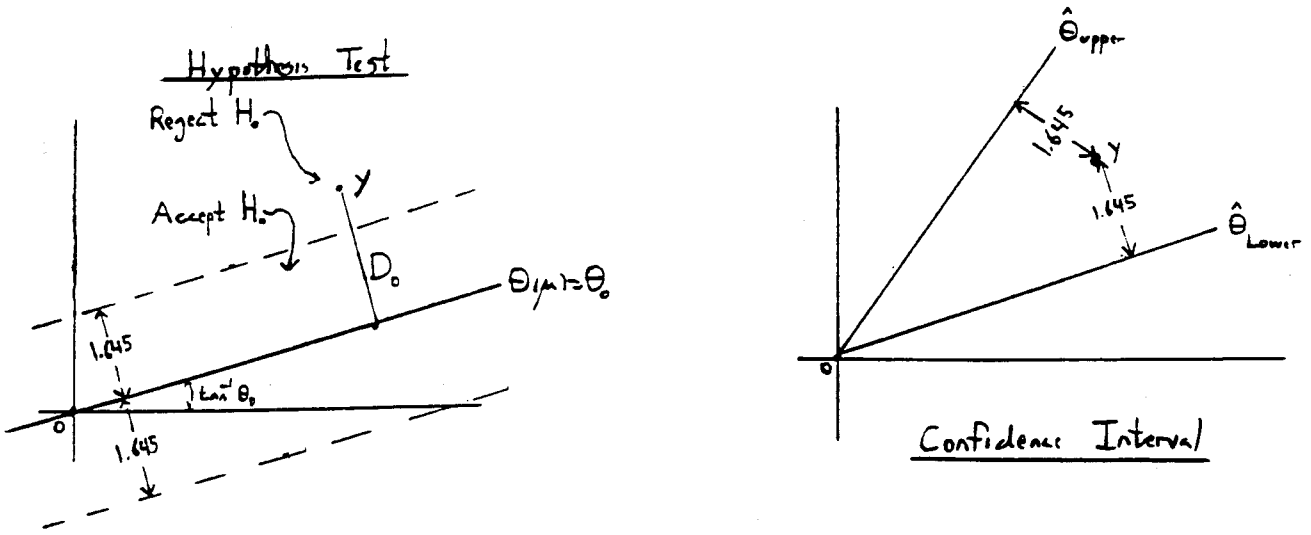

Figure 3. Fieller's construction of a 90% confidence interval for the rates of normal means; we observe $y \sim N_2(\mu, I)$ and want a confidence interval for $\theta(\mu) = \mu_2/\mu_1$; the level surface $\{\theta(\mu) = \theta_0\}$ is the straight line through the origin at angle $\tan^{-1}(\theta_0)$; hypothesis test for $H_0 : \theta = \theta_0$ accepts H_0 for $|D_0| < 1.645$, where D_0 is the signed distance from y to $\{\theta(\mu) = \theta_0\}$, left panel; the Fieller interval for θ is these values of θ_0 for which H_0 is accepted, right panel.

Let y_0 be the nearest point to the observed point $y \sim N_2(\mu, I)$ on C_{θ_0}, and let D_0 be the signed length of $y - y_0$ (using any reasonable sign convention). Also define curv_0 to be the curvature of C_{θ_0} at the point y_0. Then the adjusted signed distance,

$$D_0' \equiv \frac{D_0 - \text{curv}_0}{1 - \text{curv}_0^2} \tag{4.2}$$

is approximately normal. This approximation is very good, the cdf of D_0' differing from the standard normal cdf by only $O(n^{-3/2})$ if y is actually a sufficient statistic obtained from n observations $y_1, y_2, \cdots, y_n \overset{\text{i.i.d.}}{\sim} N_2(\mu, I)$. This is third- order accuracy, in the language of Section 1.

Inverting the approximate pivotal D_0' gives a third-order accurate approximate confidence interval for θ. Table 4 shows the results for the case $y \sim N_2(\mu, I)$, $\theta(\mu) = \|\mu\|$, when the observed vector y has length $\|y\| = 5$. (A version of (4.2) holds in higher dimensions.) In this case there is an exact confidence interval for θ based on inverting the non-central chi-square distribution of $\|y\|^2 \sim \chi_5^2(\theta^2)$.

We see that the approximate interval based on the adjusted signed distance gives excellent results in this case. The second order accurate bootstrap interval "BC_a" is not quite as good, while the first order accurate standard interval, (1.14), is far off.

	.05	.95	
Exact:	2.68	6.19	
D_0':	2.71	6.19	3rd order
BC_a:	2.94	6.06	2nd order
Standard:	3.36	6.64	1st order

Table 4. We observe $\|y\| = 5$, where $y \sim N_6(\mu, I)$, and want a confidence interval for $\theta(\mu) = \|\mu\|$; exact two-sided 90% interval (2.68, 6.19) is based on inverting the noncentral χ^2 distribution of $\|y\|^2$; interval based on inverting the approximate pivotal D_0' (4.2), is third order accurate; bootstrap method BC_a is second order accurate; standard method is first order accurate. From Efron (1985).

In this example, most statisticians, or at least most frequentists, would consider the non-central chi-square intervals to be correct, as well as exactly accurate. The D_0' intervals are third order correct, as well as third order accurate, in the sense that their endpoints differ from the exact ones by only $O(n^{-3/2})$ if y is obtained from an i.i.d. sampling situation.

Barndorff-Nielsen (1986) extends (4.2) to general parametric families. The signed distance D_0 is replaced by the signed square root of the likelihood ratio statistic. "Bartlett corrections" are made to the mean and standard deviation of D_0, to get an adjusted statistic D_0' which has a standard normal distribution to the third order of asymptotic accuracy. Then D_0' is inverted to give third-order accurate approximate confidence intervals for θ.

Are these intervals third order correct as well as third order accurate? Efron (1985) argues for the correctness of the D_0' intervals based on the fact that they come from an approximate pivotal that is geometrically reasonable. But (4.2) lacks the immediacy of Fieller's construction, even in the normal case. Moreover, there are other third order intervals available, Hall (1988), Cox and Reid (1987), Welch and Peers (1963), which may or may not agree with Barndorff-Nielsen's intervals.

DiCiccio and Efron (1990) show that all of these methods give confidence interval agreeing at the second order, and in a certain sense they all are second order correct, as well as accurate. The best situation would be if all the methods continued to agree at the third order. This happy result might very well be false. If so, the question of correctness will be a pressing one. Highly accurate confidence statements are not worth pursuing if they lead to inferential errors.

5. What is a good nonparametric pivotal quantity? Pivotal quantities play a crucial role in the theory of confidence intervals, as we saw in the previous section. Much of the recent bootstrap literature concerns the construction of approximate confidence intervals in nonparametric settings. This raises the question of this section: what is a good approximate pivotal quantity in nonparametric estimation problems?

We consider the one-sample situation, where the observed data $\mathbf{x}$ is an i.i.d. sample from some unknown probability distribution F

$$F \overset{\text{i.i.d.}}{\to} (x_1, x_2, \cdots, x_n) = \mathbf{x}, \tag{5.1}$$

and where we want an approximate confidence interval for a real-valued parameter $\theta(F)$. This becomes a nonparametric problem if we assume that F can be any distribution at all on the sample space of the x_i.

We begin with an example of a bad guess at a nonparametric pivotal quantity. The left panel of Figure 4 shows the lawschool data, 15 pairs of points $x_i = (a_i, b_i)$, where a_i and b_i are measures of excellence for the entering 1975 class at lawschool i, $i = 1, 2, \cdots, 5$. See Section 2.5 of Efron (1982). Suppose we want a confidence interval for $\theta(F)$ the Pearson correlation coefficient. The nonparametric MLE of θ is $\hat{\theta} = \theta(\hat{F}) = .776$, the sample correlation coefficient based on the 15 points $\mathbf{x} = (x_1, x_2, \cdots, x_{15})$.

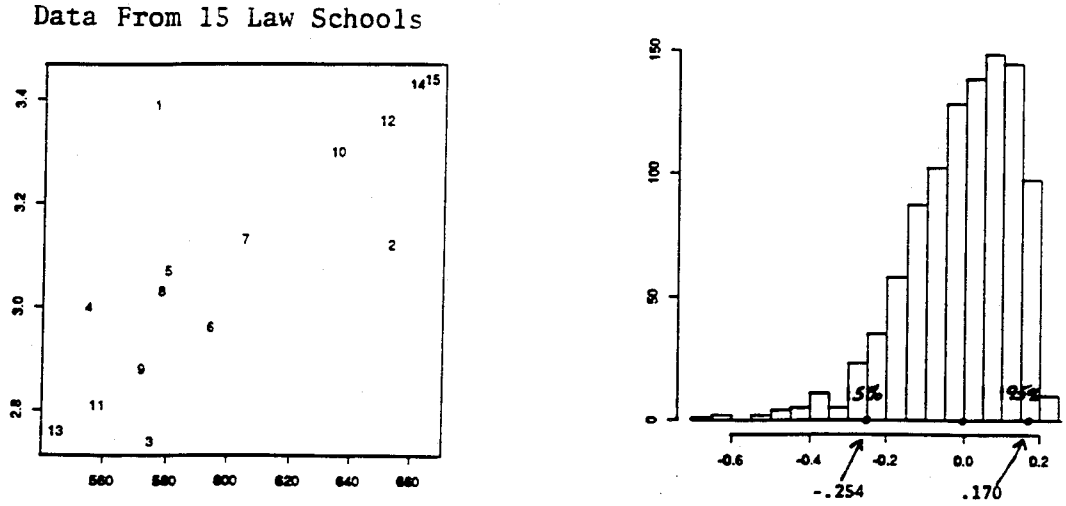

Figure 4. Lawschool data, $n = 15$ pairs of points, left panel. $B = 1000$ bootstrap replications of $\hat{\theta}^* - \hat{\theta}$, where $\hat{\theta}$ is the Pearson correlation, right panel. The quantity $\hat{\theta}^* - \hat{\theta}$ is a bad guess at a nonparametric pivotal quantity in this case.

The right side of Figure 4 shows the histogram of $B = 1000$ bootstrap replications of $\hat{\theta}^* - \hat{\theta}$; $B = 1000$ independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \cdots, \mathbf{x}^{*B}$ were generated as in (2.7), and for each one the difference of correlation coefficients $\hat{\theta}^{*b} - \hat{\theta}$ (= sample correlation coefficient of bth bootstrap sample $\mathbf{x}^{*b}$, minus .776) were calculated. We see that the 5th and 95th percentiles of the 1000 $\hat{\theta}^{*b} - \hat{\theta}$ values were $-.254$ and .170 respectively.

Suppose we believe that $\hat{\theta} - \theta$ is an approximate nonparametric pivotal quantity. Then $\hat{\theta}^* - \hat{\theta}$ should have approximately the same percentiles as $\hat{\theta} - \theta$, so that

$$\text{Prob}\{-.254 < \hat{\theta} - \theta < .170\} \doteq .90. \tag{5.2}$$

Inverting (5.2) gives a bad guess at a confidence interval for θ,

$$\theta \in [\hat{\theta} - .170, \hat{\theta} + .254] = [.606, 1.03]. \tag{5.3}$$

The quantity $\hat{\theta} - \theta$ is a poor choice for an approximate pivotal statistic in most situations, either parametric or nonparametric. If $\hat{\theta} - \theta$ is long-tailed to the left, as in the correlation example, then usually the confidence interval for θ will extend further left of $\hat{\theta}$ than right. (The reader can

check this by considering binomial or Poisson confidence intervals.) This is the opposite of what would happen if $\hat{\theta} - \theta$ were actually pivotal.

The second most obvious guess for an approximate nonparametric pivotal quantity is a t-like statistic, say

$$T = \frac{\hat{\theta} - \theta}{\hat{\sigma}}, \tag{5.4}$$

where $\hat{\sigma}(\mathbf{x})$ is some estimate of standard error for $\hat{\theta}(\mathbf{x})$, perhaps the jackknife or delta method estimates. If we believe in the pivotality of T, then we can use the bootstrap to construct a "bootstrap t" approximate confidence interval for θ; we generate some large number B of bootstrap replications of T,

$$T^* = \frac{\hat{\theta}(\mathbf{x}^*) - \hat{\theta}}{\hat{\sigma}(\mathbf{x}^*)}; \tag{5.5}$$

compute the 5th and 95th percentiles of the values T^{*b}, $b = 1, 2, \cdots, B$, say $T^{*(.05)}$ and $T^{*(.95)}$; and assign θ the approximate confidence interval

$$\theta \in [\hat{\theta} - T^{*(.95)}\hat{\sigma},\ \hat{\theta} - T^{*(.05)}\hat{\sigma}]. \tag{5.6}$$

A surprising and encouraging fact has emerged in the bootstrap literature. The bootstrap-t method gives second order accurate and correct intervals in a wide variety of situations. (See P. Hall (1988), Abramovitch and Singh (1985), and DiCiccio and Efron (1990).) There are good reasons to believe that the T statistic (5.4) is a nonparametric pivotal to the second order of asymptotic accuracy.

The favorable asymptotics of the bootstrap-t method are no guarantee of good small-sample behavior. In fact, there are practical difficulties connected with the choice of the denominator $\hat{\sigma}$ in (5.4). This can be seen in the lawschool correlation example. $B = 1000$ bootstrap replications of T were generated from the lawschool data, with

$$\hat{\sigma}(\mathbf{x}^*) = [(1 - \hat{\theta}(\mathbf{x}^*)^2)]/\sqrt{15} + .03. \tag{5.7}$$

Here $[1 - \hat{\theta}^2]/\sqrt{15}$ is an approximation to the standard error of $\hat{\theta}$, suggested by normal theory. The added quantity .03 was necessary to prevent occasional very large values of T^*. Section 5 of Efron (1990) gives a full explanation.

The 5th and 95th percentiles of the T^* values were $-.939$ and 2.93 respectively. Using (5.6), this gives the bootstrap-t confidence interval $[.40, .91]$ for θ. The 95th percentile 2.93 is suspiciously large, leading to a suspiciously low lower endpoint .40, see Table 5.

The same bootstrap replications that gave the bootstrap-t percentiles can be used to check the pivotality of T^*'s distribution. This relieves the statistician's reliance on asymptotic theory. Figure 5 shows the 5th, 10th, 16th, 50th, 84th, 90th, and 95th percentiles of the 1000 T^* bootstrap replications as dashed lines. So, for example, the 5% dashed line is at height $-.939$, and the 95% line at height 2.93.

		Nonparametric			Parametric		
	$\hat{\theta}^* - \hat{\theta}$	Standard	Boot-t	BC_a	Boot-t	BC_a	Exact
.05:	.61	.61	.40	.48	.53	.50	.49
.95:	1.03	.95	.91	.94	.93	.90	.90

Table 5. Approximate confidence intervals for the correlation coefficient, lawschool data; the bootstrap-t and BC_a methods are second order accurate and correct. Parametric intervals assume a bivariate normal model for the data. The lower endpoint of the nonparametric bootstrap-t interval is suspiciously low, compared with the BC_a answer and also with the parametric results.

The percentiles of the T distribution when sampling from F, as in (5.1), are functions of F. Bootstrap sampling gives these function values for $F = \hat{F}$, the empirical distribution (1.2). If T

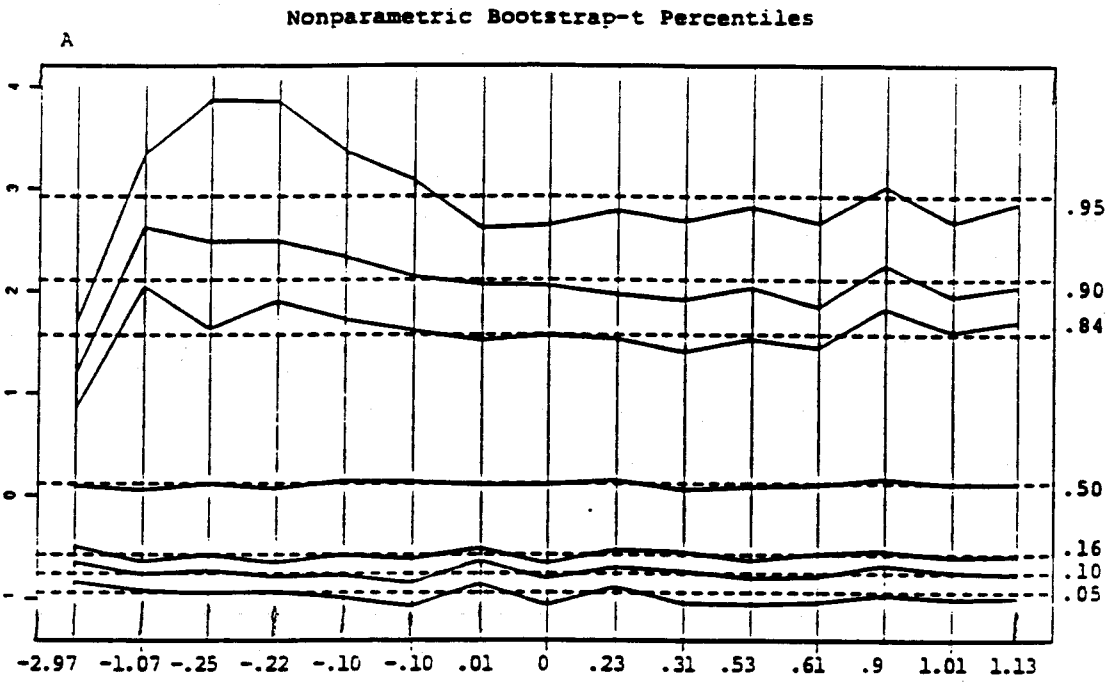

Figure 5. The percentiles of the 1000 bootstrap-t repliactions T^*, dashed lines; also the percentiles of T^*, successively removing one point at a time from the original data set, jagged lines. The upper percentiles are highly sensitive to removing single data points, indicating a lack of pivotality. From Section 5 of Efron (1990a). Bottom numbers show point removed, numbered as in Figure 4.

is approximately pivotal, then we expect the percentiles not to change much if we change F from $\hat{F}$ to some other nearby distribution. The jagged lines in Figure 5 trace the percentiles of T^* as we change $\hat{F}$ to the deleted-point distributions $\hat{F}_{(i)}$, (2.4). That is, we successively remove one point at a time from the original data set $\mathbf{x}$, and compute the percentiles of the bootstrap-t distribution when sampling from $F = \hat{F}_{(i)}$, the empirical distribution of the remaining 14 points. Efron (1990a) shows how this computation can be done without requiring any bootstrap samples beyond the original 1000.

The upper bootstrap-t percentiles are seen to be quite variable under small changes in $\hat{F}$. This argues against the pivotality of T, in this situation. The jackknife-after-bootstrap theory in Efron (1990a) uses the jagged line variability to assign a standard error of ± 1.8 to the estimated value 2.93 for the 95th percentile. We simply do not have enough data to estimate the nonparametric bootstrap-t percentiles very well in this problem.

The T statistic (5.4) is a reasonable answer to the question "what is a good nonparametric pivotal quantity?", at least in theory. Further development will be needed to make the bootstrap-t a dependable method in practice. It is not yet clear when bootstrap-t methods, even with the bugs worked out, will be preferable to other nonparametric confidence interval methods like the BC_a.

6. What are computationally efficient ways to bootstrap? Typical problems require $50 - 200$ bootstrap replications to estimate a standard error, and $1000 - 2000$ replications to compute a bootstrap confidence interval, see Section 9 of Efron (1987). These numbers assume that the bootstrap estimation is done in the most obvious way. Various computational and probabilistic methods have been suggested to reduce the number of replications required. The promise of such methods is not only a reduction of the computational burden, but also a deeper understanding of the bootstrap. This section discusses just two of the methods, with references to many others.

Suppose again that we are in the one-sample nonparametric situation (5.1), and that we wish to assess the bias of the statistic $\hat{\theta}(\mathbf{x})$ as an estimate of the parameter $\theta(F)$. The straightforward bootstrap estimate of bias is calculated as follows: B bootstrap samples $\mathbf{x}^{*b}$ give replications $\hat{\theta}^{*b} = \hat{\theta}(\mathbf{x}^{*b})$, $b = 1, 2, \cdots, B$; then the bias of $\hat{\theta}$ is estimated by

$$\widehat{\text{bias}}_B = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{*b} - \hat{\theta}(\mathbf{x}). \tag{6.1}$$

In the resampling vector notation of (2.9)-(2.11),

$$\widehat{\text{bias}}_B = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}(\mathbf{P}^{*b}) - \hat{\theta}(\mathbf{P}^{o}), \tag{6.2}$$

where $\mathbf{P}^{*b}$ is the resampling vector corresponding to $\mathbf{x}^{*b}$, and $\mathbf{P}^o$ is $(1, 1, \cdots, 1)'/n$, the center point of the resampling simplex $\mathcal{S}$ in Figure 1. As B goes to infinity, $\widehat{\text{bias}}_B$ approaches the ideal bootstrap bias estimate

$$\text{bias}_\infty = E_*\{\hat{\theta}(\mathbf{P}^*)\} - \hat{\theta}(\mathbf{P}^o), \tag{6.3}$$

E_* indicating expectation with respect to the multinomial distribution (2.10).

Section 2 of Efron (1990b) discusses an improvement on $\bar{\text{bias}}_B$,

$$\widehat{\text{bias}}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}(\mathbf{P}^{*b}) - \hat{\theta}(\bar{\mathbf{P}}), \tag{6.4}$$

where $\bar{\mathbf{P}} = \sum_b \mathbf{P}^{*b}/B$, the average of the B resampling vectors. (This assumes that $\hat{\theta}(\mathbf{P}^*)$ is continuously defined as a function of $\mathbf{P}^*$, so that $\hat{\theta}(\bar{\mathbf{P}})$ is well specified.) As B increases, the improved estimate $\widehat{\text{bias}}_B$ approaches the ideal value bias_∞ more quickly than does $\bar{\text{bias}}_B$.

Table 6 refers to the following simulation experiment: ten samples were selected, each consisting of $n = 10$ independent bivariate normal points $x_i = (y_i, z_i)$, drawn from the distribution $F = N_2((8,4)'/\sqrt{10}, I)$. We are interested in the bias of $\hat{\theta}(\mathbf{x}) = \bar{z}/\bar{y}$ as an estimate of $\theta(F) = E\{y\}/E\{z\}$.

Table 6 compares $\widehat{\text{bias}}_B$ with $\bar{\text{bias}}_B$ for $B = 20$. The bias estimates are multiplied by 1000 for each reading. Also shown is $\bar{\text{bias}}_{8000}$, which we use in place of bias_{20} as the ideal bootstrap bias estimate. We see that $\widehat{\text{bias}}_{20}$ is much better than $\bar{\text{bias}}_{20}$ in matching this ideal, the ratio of squared errors for the 10 samples being

$$\Sigma[\bar{\text{bias}}_{20} - \bar{\text{bias}}_{8000}]^2/\Sigma[\widehat{\text{bias}}_{20} - \bar{\text{bias}}_{8000}] = 23.9.$$

In a much larger version of Table 6, this ratio was estimated to be 50.3. In effect, this means that $\widehat{\text{bias}}_{20}$ is about as effective as $\bar{\text{bias}}_{20\times 50}$!

Sample	$\overline{bias}_{8,000}$	$\hat{bias}_{20}$	$\widehat{bias}_{20}$
1	7.35	4.08	−16.53
2	23.45	26.78	31.90
3	10.90	11.49	27.19
4	17.30	15.00	−31.12
5	.95	−.60	2.56
6	3.50	3.35	−10.00
7	14.55	13.99	.56
8	3.90	3.18	9.82
9	4.45	6.68	.70
10	12.85	23.90	3.55

NOTE: Entries are 1,000 × bias estimate. In this case $\hat{bias}_{20}$ is about 50 times as good an estimator as $\widehat{bias}_{20}$.

Table 6. Estimates of bias for $\hat{\theta} = \bar{z}/\bar{y}$; ten samples, each of size $n = 10$ drawn from $(y_i, z_i) \sim N_2/(8,4)'/\sqrt{10}, I)$; $\widehat{\text{bias}}_{20}$ compared with $\bar{\text{bias}}_{20}$ as an estimate of $\bar{\text{bias}}_{8000} \doteq \text{bias}_\infty$; $\widehat{\text{bias}}_{20}$ tends to be closer to $\bar{\text{bias}}_{8000}$. Entries are 1000 × bias estimate. From Section 2 of Efron (1990b).

The bias estimate $\widehat{\text{bias}}_B$ corrects $\bar{\text{bias}}_B$ by taking into account the discrepancy between $\bar{\mathbf{P}}$, the average of the observed resampling vectors, and $\mathbf{P}^o$, their theoretical expectation. Davison, Hinkley, and Schechtman (1987) make the correction another way: they draw the resampling vectors $\mathbf{P}^{*b}$, $b = 1, 2, \cdots, B$, in a manner which forces $\bar{\mathbf{P}}$ to equal $\mathbf{P}^o$. Then (6.2) performs very much like $\widehat{\text{bias}}_B$. Various methods of improved bootstrap sampling for reducing the number of bootstrap replications appear in Johns (1988), Graham et al. (1987), Ogbonmwan and Wynn (1986), Therneau (1983), and Hesterberg (1988).

In some situations it is possible to obtain bootstrap results without any Monte Carlo sampling at all. DiCiccio and Efron (1990) discuss a method for confidence interval construction within exponential families. Their method called "ABC", for "Approximate BC_a" or "Approximate Bootstrap Confidence" intervals, replaces the Monte Carlo bootstrap replications of the (parametric) BC_a method with simple analytic approximations. This is possible because exponential family problems tend to be very smooth. The ABC method requires only $4p + 6$ recomputations of the statistic of interest, compared to a few thousand for the full BC_a intervals, where p is the dimension of the exponential family.

Figure 6 shows two examples of the ABC method. The right panel concerns the lawschool data of Figure 4. Now we assume that the data is bivariate normal, $x_i \overset{\text{i.i.d.}}{\sim} N_2(\mu, \Sigma)$ for $i = 1, 2, \cdots, 15$, with the mean vector μ and covariance matrix Σ unknown. This is a $p = 5$ parameter exponential family. We want a confidence interval for the correlation coefficient. The first-order correct standard intervals, dashed lines, are compared with the second-order correct ABC intervals, solid lines. The ABC limits lie far below the standard limits. The solid points indicate the exact limits for the correlation coefficient, assuming bivariate normality. We see that the ABC limits are nearly, but not perfectly, correct.

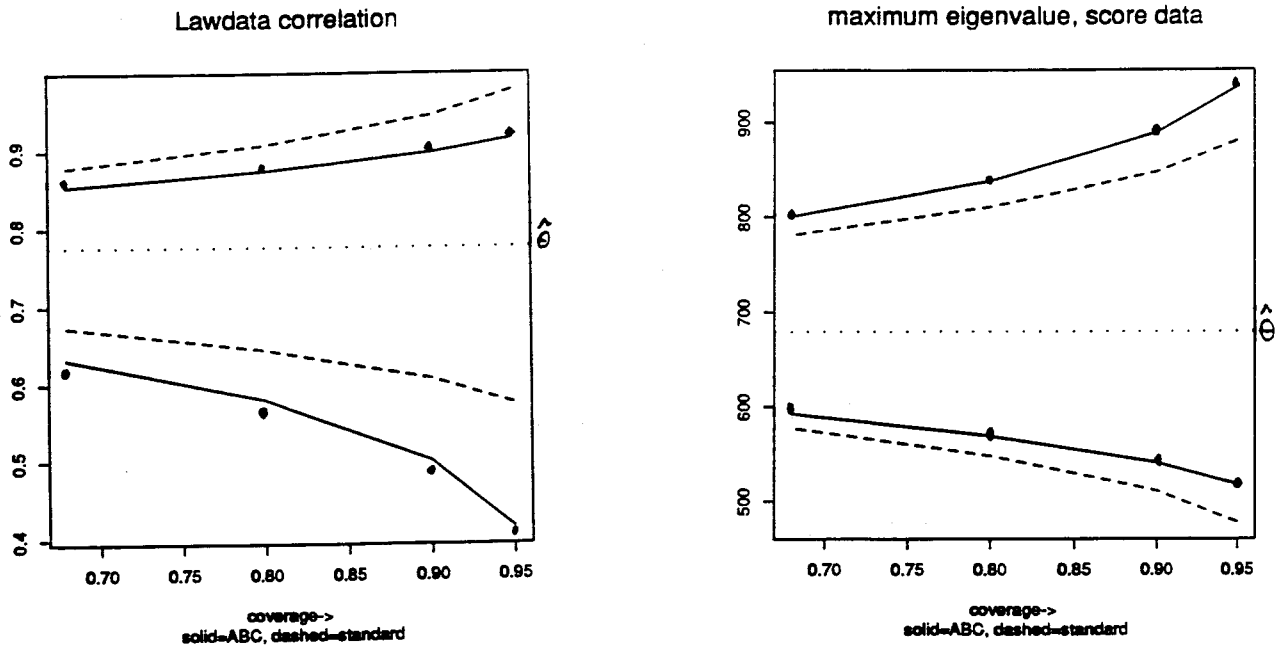

Figure 6. ABC second-order correct confidence intervals, solid lines, compared with the standard interval (1.14), dashed lines; horizontal axis is coverage probability of the two-sided interval; dotted line is MLE $\hat{\theta}$. Left panel lawschool data of Figure 4; assumes $x_i \overset{\text{i.i.d.}}{\sim} N_2(\mu, \Sigma)$, $i = 1, 2, \cdots, 15$; parameter of interest is correlation coefficient of Σ; solid points indicate exact confidence intervals for correlation. Right panel score data, Mardia, Kent, and Bibby (1979); assumes $x_i \overset{\text{i.i.d.}}{\sim} N_5(\mu, \Sigma)$, $i = 1, 2, \cdots, 88$; parameter of interest is maximum eigenvalue of Σ; solid points are BC_a limits, $B = 4000$ replications.

The right side of Figure 6 concerns the student score data from Mardia, Kent, and Bibby (1979). This comprises 5 test scores from each of $n = 88$ students. We assume that the scores have a multivariate normal distribution, $x_i \overset{\text{i.i.d.}}{\sim} N_5(\mu, \Sigma)$, $i = 1, 2, \cdots, 88$, a $p = 20$ parameter exponential family. In this case the ABC limits are shifted upwards from the standard limits. The solid points show the BC_a limits from a full bootstrap analysis involving $B = 4000$ parametric bootstrap samples. We see that the ABC limits are indeed good approximations to those from the BC_a, even though they require only a few percent as much computation.

REFERENCES

Abramovitch, L. and Singh, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap. *Ann. Stat.* **13**, 116-132.

Barnforff-Nielsen, O. E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika* **73**, 307-322.

Cox D. R., and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Royal Stat. Soc., Ser. B* **49**, 1-39.

Davison, A. C., Hinkley, D. V., and Schechtman, E. (1987). Efficient bootstrap simulation. *Biometrika* **73**, 555-561.

DiCiccio, T., and Efron, B. (1990). Better approximate confidence intervals in exponential families. Submitted to *JASA*.

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

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *CBMS* **38**, *SIAM-NSF*.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvements in cross-validation. *JASA* **78**, 316-331.

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

Efron, B. (1987). Better bootstrap confidence intervals. *JASA* **82**, 171-185.

Efron (1990a). Jackknife-after-bootstrap standard errors and influence functions. Submitted to *J. Royal Stat. Soc.*.

Efron, B. (1990b). More efficient bootstrap computations. *JASA* **85**, 79-89.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Ann. Stat.* **9**, 586-596.

Geisser, S. (1975). The predictive sample reuse method with applications. *JASA* **70**, 320-328.

Graham, R. L., Hinkley, D. V., John, P. W. M., and Shi, S. (1987). Balanced design of bootstrap simulations. Technical Report 48, University of Texas at Austin, Mathematics Dept.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Stat.* **16**, 927-985.

Hesterberg, T. (1988). Variance reduction techniques for bootstrap and other Monte Carlo simulations. Unpublished Ph.D. dissertation, Stanford University, Dept. of Statistics.

Jaeckel, L. (1972). The infinitesimal jackknife. *Memorandum MM72-1215-11*, Bell Labs, Murray Hill, NJ.

Johns, M. V. (1988). Importance sampling for bootstrap confidence intervals. *JASA* **83**, 709-714.

Kendall, M., and Stuart, A. (1958). *The Advanced Theory of Statistics.* London: Charles W. Griffin.

Lachenbruch, P., and Mickey, N, (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1-11.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* New York: Academic Press.

Miller, R. G. (1986). *Beyond ANOVA.* New York: John Wiley.

Ogbonmwan, S. M., and Wynn, H. P. (1986). Resampling generated likelihoods. *Proc. Fourth Purdue Symp. Decision Theory* (Vol. 1), eds. S. Gupta and J. Berger, New York: Springer-Verlag, pp. 137-147.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Royal Stat. Soc., Ser. B* **36**, 111-147.

Therneau, T. (1983). Variance reduction techniques for the bootstrap. Unpublished Ph.D. dissertation, Stanford University, Dept. of Statistics.

Welch, B. L., and Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Royal Stat. Soc., Ser. B* **25**, 318-329.