A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation

BRADLEY EFRON and GAIL GONG*

This is an invited expository article for *The American Statistician*. It reviews the nonparametric estimation of statistical error, mainly the bias and standard error of an estimator, or the error rate of a prediction rule. The presentation is written at a relaxed mathematical level, omitting most proofs, regularity conditions, and technical details.

KEY WORDS: Bias estimation; Variance estimation; Nonparametric standard errors; Nonparametric confidence intervals; Error rate prediction.

1. INTRODUCTION

This article is intended to cover lots of ground, but at a relaxed mathematical level that omits most proofs, regularity conditions, and technical details. The ground in question is the nonparametric estimation of statistical error. "Error" here refers mainly to the bias and standard error of an estimator, or to the error rate of a data-based prediction rule.

All of the methods we discuss share some attractive properties for the statistical practitioner: they require very little in the way of modeling, assumptions, or analysis, and can be applied in an automatic way to any situation, no matter how complicated. (We will give an example of a very complicated prediction rule indeed). An important theme of what follows is the substitution of raw computing power for theoretical analysis.

The references upon which this article is based (Efron 1979a,b, 1981a,b,c, 1982; Efron and Gong 1982) explore the connections between the various nonparametric methods, and also the relationship to familiar parametric techniques. Needless to say, there is no danger of parametric statistics going out of business. A good parametric analysis, when appropriate, can be far more efficient than its nonparametric counterpart. Often, though, parametric assumptions are difficult to justify, in which case it is reassuring to have available the comparatively crude but trustworthy nonparametric answers.

What are the bootstrap, the jackknife, and cross-validation? For a quick answer, before we begin the main exposition, we consider a problem where none of the three methods are necessary, estimating the standard error of a sample average.

The data set consists of a random sample of size n from an unknown probability distribution F on the real line,

$$X_1, X_2, \ldots, X_n \sim F. \tag{1}$$

Having observed $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, we compute the sample average $\bar{x} = \sum_{i=1}^n x_i/n$ for use as an estimate of the expectation of F.

An interesting fact, and a crucial one for statistical applications, is that the data set provides more than the estimate $\bar{x}$. It also gives an estimate for the accuracy of $\bar{x}$, namely

$$\hat{\sigma} = \left[\frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2\right]^{1/2}; \tag{2}$$

$\hat{\sigma}$ is the estimated standard error of $\bar{X} = \bar{x}$, the root mean squared error of estimation.

The trouble with formula (2) is that it does not, in any obvious way, extend to estimators other than $\bar{X}$, for example the sample median. The jackknife and the bootstrap are two ways of making this extension. Let

$$\bar{x}_{(i)} = \frac{n\bar{x} - x_i}{n-1} = \frac{1}{n-1} \sum_{j \neq i} x_i, \tag{3}$$

the sample average of the data set deleting the ith point. Also, let $\bar{x}_{(\cdot)} = \sum_{i=1}^n x_{(i)}/n$, the average of the deleted averages. (Actually $\bar{x}_{(\cdot)} = \bar{x}$, but we need the dot notation below.) The jackknife estimate of standard error is

$$\hat{\sigma}_J = \left[\frac{n-1}{n} \sum_{i=1}^n (\bar{x}_{(i)} - \bar{x}_{(\cdot)})^2\right]^{1/2}. \tag{4}$$

The reader can verify that this is the same as (2). The advantage of (4) is an easy generalizability to any estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$. The only change is to substitute $\hat{\theta}_{(i)} = \hat{\theta}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ for $\bar{x}_{(i)}$ and $\hat{\theta}_{(\cdot)} = \sum_{i=1}^n \hat{\theta}_{(i)}/n$ for $\bar{x}_{(\cdot)}$.

The bootstrap generalizes (2) in an apparently different way. Let $\hat{F}$ be the empirical probability distribution of the data, putting probability mass $1/n$ on each x_i, and let $X_1^*, X_2^*, \ldots, X_n^*$ be a random sample from $\hat{F}$,

$$X_1^*, X_2^*, \ldots, X_n^* \sim \hat{F}. \tag{5}$$

In other words each X_i^* is drawn independently with replacement and with equal probability from the set $\{x_1, x_2, \ldots, x_n\}$. Then $\bar{X}^* = \sum_{i=1}^n X_i^*/n$ has variance

*Bradley Efron is Professor of Statistics and Biostatistics at Stanford University. Gail Gong is Assistant Professor of Statistics at Carnegie-Mellon University. The authors are grateful to Rob Tibshirani who suggested the final example in Section 7; to Samprit Chatterjee and Werner Stuetzle who suggested looking at estimators like "BootAve" in Section 9; and to Dr. Peter Gregory of the Stanford Medical School who provided the original analysis as well as the data in Section 10. This work was partially supported by the National Science Foundation and the National Institutes of Health.

$$\text{var.}\,\overline{X}^* = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{6}$$

var. indicating variance under sampling scheme (5). The bootstrap estimate of standard error for an estimator $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is

$$\hat{\sigma}_B = [\text{var.}\,\hat{\theta}(X_1^*, X_2^*, \ldots, X_n^*)]^{1/2}. \tag{7}$$

Comparing (7) with (2) we see that $[n/(n-1)]^{1/2}\hat{\sigma}_B = \hat{\sigma}$ for $\hat{\theta} = \overline{X}$. We could make $\hat{\sigma}_B$ exactly equal $\hat{\sigma}$, for $\hat{\theta} = \overline{X}$, by adjusting definition (7) with the factor $[n/(n-1)]^{1/2}$, but there is no general advantage in doing so. A simple algorithm described in Section 2 allows the statistician to compute $\hat{\sigma}_B$ no matter how complicated $\hat{\theta}$ may be. Section 3 shows the close connection between $\hat{\sigma}_B$ and $\hat{\sigma}_J$.

Cross-validation relates to another, more difficult, problem in estimating statistical error. Going back to (1), suppose we try to predict a new observation from F, call it X_0, using the estimator $\overline{X}$ as a predictor. The expected squared error of prediction $E[X_0 - \overline{X}]^2$ equals $((n+1)/n)\mu_2$ where μ_2 is the variance of the distribution F. An unbiased estimate of $((n+1)/n)\mu_2$ is

$$(n+1)\hat{\sigma}^2. \tag{8}$$

Cross-validation is a way of obtaining nearly unbiased estimators of prediction error in much more complicated situations. The method consists of (a) deleting the points x_i from the data set one at a time; (b) recalculating the prediction rule on the basis of the remaining $n-1$ points; (c) seeing how well the recalculated rule predicts the deleted point; and (d) averaging these predictions over all n deletions of an x_i. In the simple case above, the cross-validated estimate of prediction error is

$$\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}_{(i)}]^2. \tag{9}$$

A little algebra shows that (9) equals (8) times $n^2/(n^2-1)$, this last factor being nearly equal to one.

The advantage of the cross-validation algorithm is that it can be applied to arbitrarily complicated prediction rules. The connection with the bootstrap and jackknife is shown in Section 9.

2. THE BOOTSTRAP

This section describes the simple idea of the bootstrap (Efron 1979a). We begin with an example. The 15 points in Figure 1 represent various entering classes at American law schools in 1973. The two coordinates for law school i are $x_i = (y_i, z_i)$,

y_i = average LSAT score of entering students at school i,

z_i = average undergraduate GPA score of entering students at school i.

(The LSAT is a national test similar to the Graduate Record Exam, while GPA refers to undergraduate grade point average.)

The observed Pearson correlation coefficient for these $n = 15$ pairs is $\hat{\rho}(x_1, x_2, \ldots, x_{15}) = .776$. We want to attach a nonparametric estimate of standard error to $\hat{\rho}$. The bootstrap idea is the following:

1. Suppose that the data points $x_1, x_2, \ldots, x_{15}$ are independent observations from some bivariate distribution F on the plane. Then the true standard error of $\hat{\rho}$ is a function of F, indicated $\sigma(F)$,

$$\sigma(F) = [\text{var}_F\,\hat{\rho}(X_1, X_2, \ldots, X_n)]^{1/2}.$$

(It is also a function of sample size n, and the functional form of the statistic $\hat{\rho}$, but both of these are known to the statistician.)

2. We don't know F, but we can estimate it by the empirical probability distribution $\hat{F}$,

$$\hat{F}\text{: mass } \frac{1}{n} \text{ on each observed data point } x_i,$$

$$i = 1, 2, \ldots, n.$$

3. The bootstrap estimate of $\sigma(F)$ is

$$\hat{\sigma}_B = \sigma(\hat{F}). \tag{10}$$

For the correlation coefficient and for most statistics, even very simple ones, the function $\sigma(F)$ is impossible to express in closed form. That is why the bootstrap is not in common use. However in these days of fast and cheap computation $\hat{\sigma}_B$ can easily be approximated by Monte Carlo methods:

(i) Construct $\hat{F}$, the empirical distribution function, as just described.

(ii) Draw a *bootstrap sample* $X_1^*, X_2^*, \ldots, X_n^*$ by independent random sampling from $\hat{F}$. In other words, make n random draws *with replacement* from $\{x_1, x_2, \ldots, x_n\}$. In the law school example a typical bootstrap sample might consist of 2 copies of point 1, 0 copies of point 2, 1 copy of point 3, and so on, the total number of copies adding up to $n = 15$. Compute the *bootstrap replication*, $\hat{\rho}^* = \hat{\rho}(X_1^*, X_2^*, \ldots, X_n^*)$, that is, the value of the statistic, in this case the correlation coefficient, evaluated for the bootstrap sample.

(iii) Do step (ii) some large number "B" of times,

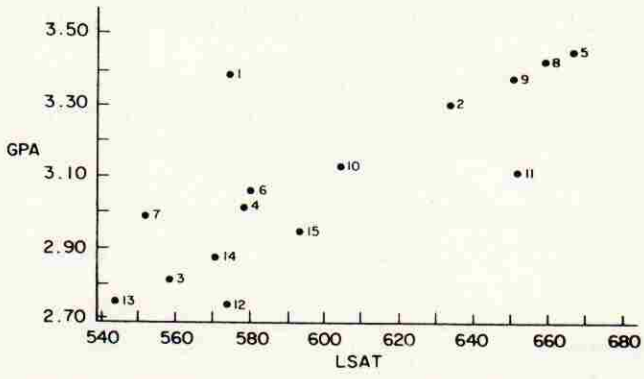

Figure 1. The law school data (Efron 1979B). The data points, beginning with School #1, are (576, 3.39), (635, 3.30), (558, 2.81), (578, 3.03), (666, 3.44), (580, 3.07), (555, 3.00), (661, 3.43), (651, 3.36), (605, 3.13), (653, 3.12), (575, 2.74), (545, 2.76), (572, 2.88), (594, 2.96).

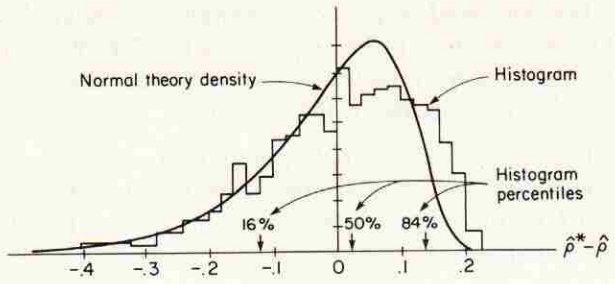

Figure 2. Histogram of $B = 1000$ bootstrap replications $\hat{\rho}^*$ for the law school data. The normal theory density curve has a similar shape, but falls off more quickly at the upper tail.

obtaining independent bootstrap replications $\hat{\rho}^{*1}$, $\hat{\rho}^{*2}, \ldots, \hat{\rho}^{*B}$, and approximate $\hat{\sigma}_B$ by

$$\hat{\sigma}_B = \left[\left(\sum_{b=1}^{B} (\hat{\rho}^{*b} - \hat{\rho}^{*\cdot})^2\right)\Big/(B-1)\right]^{1/2}, \quad \hat{\rho}^{*\cdot} \equiv \frac{\sum \hat{\rho}^{*b}}{B} \quad (11)$$

As $B \to \infty$, (11) approaches the original definition (10). The choice of B is further discussed below, but meanwhile we won't distinguish between (10) and (11), calling both estimates $\hat{\sigma}_B$.

Figure 2 shows $B = 1000$ bootstrap replications $\hat{\rho}^{*1}, \ldots, \hat{\rho}^{*1000}$ for the law school data. The abscissa is plotted in terms of $\hat{\rho}^* - \hat{\rho} = \hat{\rho}^* - .776$. Formula (11) gives $\hat{\sigma}_B = .127$. This can be compared with the normal theory estimate of standard error for $\hat{\rho}$, (Johnson and Kotz 1970, p. 229),

$$\hat{\sigma}_{\text{NORM}} \equiv \frac{1-\hat{\rho}^2}{\sqrt{n-3}} = .115.$$

One thing is obvious about the bootstrap procedure: it can be applied just as well to any statistic, simple or complicated, as to the correlation coefficient. In Table 1 the statistic is the 25 percent trimmed mean for a sample of size $n = 15$. The true distribution F (now defined on the line rather than on the plane) is the standard normal $\mathcal{N}(0, 1)$ for the left side of the table, or one-sided negative exponential for the right side. The true standard errors $\sigma(F)$ are .286 and .232, respectively. In both cases, $\hat{\sigma}_B$, calculated with $B = 200$ bootstrap replications, is nearly unbiased for $\sigma(F)$.

The jackknife estimate of standard error $\hat{\sigma}_J$, described in Section 3, is also nearly unbiased in both cases, but has higher variability than $\hat{\sigma}_B$, as shown by its higher coefficient of variation. The minimum possible coefficient of variation (C.V.), for a scale-invariant estimate of $\sigma(F)$, assuming full knowledge of the parametric model, is shown in brackets. In the normal case, for example, .19 is the C.V. of $[\Sigma(x_i - \bar{x})^2/14]^{1/2}$. The bootstrap estimate performs well by this standard considering its totally nonparametric character and the small sample size.

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	Coeff Var	Ave	Sd	Coeff Var
Bootstrap $\hat{\sigma}_B$: ($B = 200$)	.287	.071	.25	.242	.078	.32
Jackknife $\hat{\sigma}_J$:	.280	.084	.30	.224	.085	.38
True : [Minimum C.V.]	.286		[.19]	.232		[.27]

Table 2 returns to the case of $\hat{\rho}$, the correlation coefficient. Instead of real data we have a sampling experiment in which F is bivariate normal, true correlation $\rho = .5$, and the sample size is $n = 14$. The left side of Table 2 refers to $\hat{\rho}$, while the right side refers to the statistic $\hat{\phi} = \tanh^{-1} \hat{\rho} = .5 \log(1+\hat{\rho})/(1-\hat{\rho})$. For each estimator $\hat{\sigma}$, 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 $B = 512$, the latter value yielding only slightly better estimates $\hat{\sigma}_B$. Further increasing B would be pointless. It can be shown that $B = \infty$ would give $\sqrt{\text{MSE}} = .063$ in the $\hat{\rho}$ case, only .001 less than using $B = 512$. As a point of comparison, the normal theory estimate for the standard error of $\hat{\rho}$, $\hat{\sigma}_{\text{NORM}} = (1-\hat{\rho}^2)/(n-3)^{1/2}$, has $\sqrt{\text{MSE}} = .056$.

Why not generate the bootstrap observations from an estimate of $\hat{F}$ which is smoother than F? This is done in lines 3, 4, and 5 of Table 2. Let $\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'/n$ be the sample covariance matrix of the observed data. The *normal smoothed bootstrap* draws the bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{F} \oplus \mathcal{N}_2(0, .25\hat{\boldsymbol{\Sigma}})$, $\oplus$ indicating convolution. This amounts to estimating F by an equal mixture of the n distributions $\mathcal{N}_2(x_i, .25\hat{\boldsymbol{\Sigma}})$, that is by a normal window estimate. 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 $\hat{\phi} = \tanh^{-1} \hat{\rho}$ is specifically chosen to have nearly constant standard error in the bivariate-normal family. The *uniform smoothed bootstrap* samples $X_1^*, \ldots, X_n^*$ from $\hat{F} \oplus \mathcal{U}(0, .25\hat{\boldsymbol{\Sigma}})$, where $\mathcal{U}(0, .25\hat{\boldsymbol{\Sigma}})$ is the uniform distribution on a rhombus selected so $\mathcal{U}$ has mean vector 0 and covariance matrix $.25\hat{\boldsymbol{\Sigma}}$. It yields moderate reductions in $\sqrt{\text{MSE}}$ for both sides of the table.

The standard normal-theory estimates of line 8, Table 2, are themselves bootstrap estimates, carried out in a parametric framework. The bootstrap sample $X_1^*, \ldots, X_n^*$ is drawn from the parametric maximum likelihood distribution

$$\hat{F}_{\text{NORM}} \sim \mathcal{N}_2(\bar{x}, \hat{\boldsymbol{\Sigma}}),$$

rather than the nonparametric maximum likelihood distribution $\hat{F}$, and with only this change the bootstrap algorithm proceeds as previously described. In practice the bootstrap process is not actually carried out. If it were, and if $B \to \infty$, then a high-order Taylor series analysis shows that $\hat{\sigma}_B$ would equal approximately $(1-\hat{\rho}^2)/(n-3)^{1/2}$, the formula actually used to compute line 8 for the $\hat{\rho}$ side of Table 2. Notice that the normal

Table 2. Estimates of Standard Error for the Correlation Coefficient $\hat{\rho}$ and for $\hat{\phi} = \tanh^{-1} \hat{\rho}$; 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 $\hat{\rho}$				Standard Error Estimates for $\hat{\phi}$			
	Ave	Std Dev	CV	$\sqrt{MSE}$	Ave	Std Dev	CV	$\sqrt{MSE}$
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			

smoothed bootstrap can be thought of as a compromise between using $\hat{F}$ and $\hat{F}_{\text{NORM}}$ to begin the bootstrap process.

3. THE JACKKNIFE

The jackknife estimate of standard error was introduced by Tukey in 1958 (see Miller 1974). Let $\hat{\rho}_{(i)} = \hat{\rho}(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ be the value of the statistic when x_i is deleted from the data set, and let $\hat{\rho}_{(\cdot)} = (1/n) \sum_{i=1}^{n} \hat{\rho}_{(i)}$. The jackknife formula is

$$\hat{\sigma}_J = \left[((n-1)/n) \sum_{i=1}^{n} (\hat{\rho}_{(i)} - \hat{\rho}_{(\cdot)})^2 \right]^{1/2}.$$

Like the bootstrap, the jackknife can be applied to any statistic that is a function of n independent and identically distributed variables. It performs less well than the bootstrap in Tables 1 and 2, and in most cases investigated by the author (see Efron 1982), but requires less computation. In fact the two methods are closely related, which we shall now show.

Suppose the statistic of interest, which we will now call $\hat{\theta}(x_1, x_2, \ldots, x_n)$, 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. Both examples in Section 2 are of this form. Let $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ be a probability vector having nonnegative weights summing to one, and define the reweighted empirical distribution $\hat{F}(\mathbf{P})$: mass P_i on x_i, $i = 1, 2, \ldots, n$. Corresponding to $\mathbf{P}$ is a *resampled value* of the statistic of interest, say $\hat{\theta}(\mathbf{P}) = \theta(\hat{F}(\mathbf{P}))$. The shorthand notation $\hat{\theta}(\mathbf{P})$ assumes that the data points $x_1, x_2, \ldots, x_n$ are fixed at their observed values.

Another way to describe the bootstrap estimate $\hat{\sigma}_B$ is as follows. Let $\mathbf{P}^*$ indicate a vector drawn from the rescaled multinomial distribution

$$\mathbf{P}^* \sim \text{Mult}_n(n, \mathbf{P}^o)/n, \ (\mathbf{P}^o \equiv (1/n)\ (1, 1, \ldots, 1)'), \quad (12)$$

meaning the observed proportions from n random draws on n categories, with equal probability $1/n$ for each category. Then

$$\hat{\sigma}_B = [\text{var}_* \hat{\theta}(\mathbf{P}^*)]^{1/2}, \quad (13)$$

where var_* indicates variance under distribution (12). (This is true because we can take $P_i^* = \#\{X_j^* = x_i\}/n$ in step 2 of the bootstrap algorithm.)

Figure 3 illustrates the situation for the case $n = 3$. There are 10 possible bootstrap points. For example, the point $\mathbf{P}^* = (\frac{2}{3}, \frac{1}{3}, 0)'$ is the second dot from the left on the lower side of the triangle, and occurs with bootstrap probability $\frac{1}{9}$, under (12). It indicates a bootstrap sample X_1^*, X_2^*, X_3^* consisting of two x_1's and one x_2. The center point $\mathbf{P}^o = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})'$ has bootstrap probability $\frac{2}{9}$.

The jackknife resamples the statistic at the n points

$$\mathbf{P}_{(i)} = (1/(n-1))\ (1, 1, \ldots, 1, 0, 1, \ldots, 1)' \quad (0 \text{ in } i\text{th place}),$$

$i = 1, 2, \ldots, n$. These are indicated by the open circles in Figure 3. In general there are n jackknife points, compared with $\binom{2n-1}{n}$ bootstrap points.

The trouble with bootstrap formula (13) is that $\hat{\theta}(\mathbf{P})$ is usually a complicated function of $\mathbf{P}$ (think of the examples in Sec. 2), and so $\text{var}_* \hat{\theta}(\mathbf{P}^*)$ cannot be evalu-

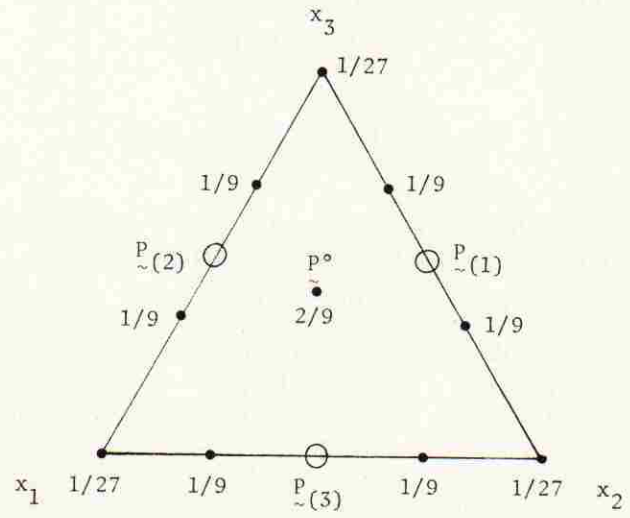

Figure 3. The bootstrap and jackknife sampling points in the case $n = 3$. The bootstrap points ($\cdot$) are shown with their probabilities.

ated except by Monte Carlo methods. The jackknife trick approximates $\hat{\theta}(\mathbf{P})$ by a linear function of $\mathbf{P}$, say $\hat{\theta}_L(\mathbf{P})$, and then uses the known covariance structure of (12) to evaluate var. $\hat{\theta}_L(\mathbf{P}^*)$. The approximator $\hat{\theta}_L(\mathbf{P})$ is chosen to match $\hat{\theta}(\mathbf{P})$ at the n points $\mathbf{P} = \mathbf{P}_{(i)}$. It is not hard to see that

$$\hat{\theta}_L(\mathbf{P}) = \hat{\theta}_{(\cdot)} + (\mathbf{P} - \mathbf{P}^o)'\mathbf{U} \tag{14}$$

where $\hat{\theta}_{(\cdot)} = (1/n)\ \Sigma\ \hat{\theta}_{(i)} = (1/n)\ \Sigma\ \hat{\theta}(\mathbf{P}_{(i)})$, and $\mathbf{U}$ is a column vector with coordinates $U_i = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})$.

Theorem. The jackknife estimate of standard error equals

$$\hat{\sigma}_J = \left[\frac{n}{n-1}\ \text{var.}\ \hat{\theta}_L(\mathbf{P}^*)\right]^{1/2},$$

which is $[n/(n-1)]^{1/2}$ times the bootstrap estimate of standard error for $\hat{\theta}_L$ (Efron 1982).

In other words the jackknife is, almost,[1] a bootstrap itself. The advantage of working with $\hat{\theta}_L$ rather than $\hat{\theta}$ is that there is no need for Monte Carlo: var. $\hat{\theta}_L(\mathbf{P}^*) = \text{var.}\ (\mathbf{P}^* - \mathbf{P}^o)'\mathbf{U} = \Sigma U_i^2/n^2$, using the covariance matrix for (12) and the fact that $\Sigma U_i = 0$. The disadvantage is (usually) increased error of estimation, as seen in Tables 1 and 2.

The fact that $\hat{\sigma}_J$ is almost $\hat{\sigma}_B$ for a linear approximation of $\hat{\theta}$ does not mean that $\hat{\sigma}_J$ is a reasonable approximation for the actual $\hat{\sigma}_B$. That depends on how well $\hat{\theta}_L$ approximates $\hat{\theta}$. In the case where $\hat{\theta}$ is the sample median, for instance, the approximation is very poor.

4. THE DELTA METHOD, INFLUENCE FUNCTIONS, AND THE INFINITESIMAL JACKKNIFE

There is a more obvious linear approximation to $\hat{\theta}(\mathbf{P})$ than $\hat{\theta}_L(\mathbf{P})$, (14). Why not use the first-order Taylor series expansion for $\hat{\theta}(\mathbf{P})$ about the point $\mathbf{P} = \mathbf{P}^o$? This is the idea of Jaeckel's *infinitesimal jackknife* (1972). The Taylor series approximation turns out to be

$$\hat{\theta}_T(\mathbf{P}) = \hat{\theta}(\mathbf{P}^o) + (\mathbf{P} - \mathbf{P}^o)'\mathbf{U}^o,$$

where

$$U_i^o = \lim_{\varepsilon \to 0} \frac{\hat{\theta}((1-\varepsilon)\mathbf{P}^o + \varepsilon\boldsymbol{\delta}_i) - \hat{\theta}(\mathbf{P}^o)}{\varepsilon},$$

$\boldsymbol{\delta}_i$ being the ith coordinate vector. This suggests the infinitesimal jackknife estimate of standard error

$$\hat{\sigma}_U = [\text{var.}\ \hat{\theta}_T(\mathbf{P}^*)]^{1/2} = [\Sigma U_i^{o\,2}/n^2]^{1/2}, \tag{15}$$

with var. still indicating variance under (12). The ordinary jackknife can be thought of as taking $\varepsilon = -1/(n-1)$ in the definition of U_i^o, while the infinitesimal jackknife lets $\varepsilon \to 0$, thereby earning the name.

The U_i^o 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 (15) 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 or 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 $t(\overline{Q}_1, \overline{Q}_2, \ldots, \overline{Q}_A)$, where $t(\cdot, \cdot, \ldots, \cdot)$ is a known function and each $\overline{Q}_a$ is an observed average, $\overline{Q}_a = \sum_{i=1}^n Q_a(X_i)/n$. For example, the correlation $\hat{\rho}$ 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 $\overline{Q}_a$; (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 $\hat{\rho}$ by

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

where, in terms of $x_i = (y_i,\ z_i)$, $\hat{\mu}_{gh} \equiv \Sigma(y_i - \overline{y})^g (z_i - \overline{z})^h/n$ (Cramér 1946, p. 359).

Theorem. For statistics of the form $\hat{\theta} = t(\overline{Q}_1, \ldots, \overline{Q}_A)$, the nonparametric delta method and the infinitesimal jackknife give the same estimate of standard error (Efron 1981b).

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 upwards, 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.

5. NONPARAMETRIC CONFIDENCE INTERVALS

In applied work, the usual purpose of estimating a standard error is to set confidence intervals for the un-

[1]The factor $[n/(n-1)]^{1/2}$ makes $\hat{\sigma}_J^2$ unbiased for σ^2 if $\hat{\theta}$ is a linear statistic, e.g., $\hat{\theta} = \overline{X}$. We could multiply $\hat{\sigma}_B$ by this same factor, and achieve the same unbiasedness, but there doesn't seem to be any general advantage to doing so.

known paramater. These are typically of the crude form $\hat{\theta} \pm z_\alpha \hat{\sigma}$, with z_α being the $100(1-\alpha)$ percentile point of a standard normal distribution. We can, and do, use the bootstrap and jackknife estimates $\hat{\sigma}_B$, $\hat{\sigma}_J$ in this way. However in small-sample parametric situations, where we can do exact calculations, confidence intervals are often highly asymmetric about the best point estimate $\hat{\theta}$. This asymmetry, which is $0(1/\sqrt{n})$ in magnitude, is substantially more important than the Student's t correction (replacing $\hat{\theta} \pm z_\alpha \hat{\sigma}$ by $\hat{\theta} \pm t_\alpha \hat{\sigma}$, with t_α the $100(1-\alpha)$ percentile point of the appropriate t distribution), which is only $0(1/n)$. This section discusses some nonparametric methods of assigning confidence intervals, which attempt to capture the correct asymmetry. It is abbreviated from a longer discussion in Efron (1981c), and also Chapter 10 of Efron (1982). All of this work is highly speculative, though encouraging.

We return to the law school example of Section 2. Suppose for the moment that we believe the data come from a bivariate normal distribution. The standard 68 percent central confidence interval (i.e., $\alpha = .16$, $1 - 2\alpha = .68$) for ρ in this case is $[.62, .87] = [\hat{\rho} - .16, \hat{\rho} + .09]$, obtained by inverting the approximation $\hat{\phi} \sim \mathcal{N}(\phi + \rho/(2(n-1)), 1/(n-3))$. Compared to the crude interval $\hat{\rho} \pm z_{.16}\, \hat{\sigma}_{\text{NORM}} = \hat{\rho} \pm \hat{\sigma}_{\text{NORM}} = [\hat{\rho} - .12, \hat{\rho} + .12]$, this demonstrates the magnitude of the asymmetry effect described previously.

The asymmetry of the confidence interval $[\hat{\rho} - .16, \hat{\rho} + .09]$ relates to the asymmetry of the normal-theory density curve for $\hat{\rho}$, as shown in Figure 2. The bootstrap histogram shows this same asymmetry. The striking similarity between the histogram and the density curve suggests that we can use the bootstrap results more ambitiously than simply to compute $\hat{\sigma}_B$.

Two ways of forming nonparametric confidence intervals from the bootstrap histogram are discussed in Efron (1981c). The first, called the *percentile method,* uses the 100α and $100(1-\alpha)$ percentiles of the bootstrap histogram, say

$$\theta \in [\hat{\theta}(\alpha), \hat{\theta}(1-\alpha)], \tag{16}$$

as a putative $1 - 2\alpha$ central confidence interval for the unknown parameter θ. Letting

$$\hat{C}(t) \equiv \frac{\#\{\hat{\theta}^{*b} < t\}}{B},$$

then $\hat{\theta}(\alpha) = \hat{C}^{-1}(\alpha)$, $\hat{\theta}(1-\alpha) = \hat{C}^{-1}(1-\alpha)$. In the law school example, with $B = 1000$ and $\alpha = .16$, the 68 percent interval is $\rho \in [.65, .91] = [\hat{\rho} - .12, \hat{\rho} + .13]$, almost exactly the same as the crude normal-theory interval $\hat{\rho} \pm \hat{\sigma}_{\text{NORM}}$.

Notice that the median of the bootstrap histogram is substantially higher than $\hat{\rho}$ in Figure 2. In fact, $\hat{C}(\hat{\rho}) = .433$, only 433 out of 1000 bootstrap replications having $\hat{\rho}^* < \hat{\rho}$. The *bias-corrected percentile method* makes an adjustment for this type of bias. Let $\Phi(z)$ indicate the CDF of the standard normal distribution, so $\Phi(z_\alpha) = 1 - \alpha$, and define

$$z_0 \equiv \Phi^{-1}\{\hat{C}(\hat{\theta})\}.$$

The bias-corrected putative $1 - 2\alpha$ central confidence interval is defined to be

$$\theta \in [\hat{C}^{-1}\{\Phi(2z_0 - z_\alpha)\},\ \hat{C}^{-1}\{\Phi(2z_0 + z_\alpha)\}]. \tag{17}$$

If $\hat{C}(\hat{\theta}) = .50$, the median unbiased case, then $z_0 = 0$ and (8) reduce to the uncorrected percentile interval (16). Otherwise the results can be quite different. In the law school example $z_0 = \Phi(.433) = -.17$, and for $\alpha = .16$, (8) gives $\rho \in [\hat{C}^{-1}\{\Phi(-1.34)\},\ \hat{C}^{-1}\{\Phi(.66)\}] = [\hat{\rho} - .17, \hat{\rho} + .10]$. This agrees nicely with the normal-theory interval $[\hat{\rho} - .16, \hat{\rho} + .09]$.

Table 3 shows the results of a small sampling experiment, only 10 trials, in which the true distribution F was bivariate normal, $\rho = .5$. The bias-corrected percentile method shows impressive agreement with the normal-theory intervals. Even better are the smoothed intervals, last column. Here the bootstrap replications were obtained by sampling from $\hat{F} \oplus \mathcal{N}(0, .25\hat{\Sigma})$, as in line 3 of Table 2, and then applying (17) to the resulting histogram.

There are some theoretical arguments supporting (16) and (17). If there exists a normalizing transformation, in the same sense as $\hat{\phi} = \tanh^{-1}\hat{\rho}$ is normalizing for the correlation coefficient under bivariate-normal sampling, then the bias-corrected percentile method automatically produces the appropriate confidence intervals. This is interesting since we do not have to know the form of the normalizing transformation to apply (17). Bayesian and frequentist justifications are given also in Efron (1981c). None of these arguments is overwhelming, and in fact (17) and (16) sometimes perform poorly. Some other methods are suggested in Efron (1981c), but the appropriate theory is still far from clear.

6. BIAS ESTIMATION

Quenouille (1949) originally introduced the jackknife as a nonparametric device for estimating bias. Let us denote the bias of a functional statistic $\hat{\theta} = \theta(\hat{F})$ by

Table 3. Central 68% Confidence Intervals for ρ, 10 Trials of $X_1, X_2, \ldots, X_{15}$ Bivariate Normal With True $\rho = .5$. Each Interval Has $\hat{\rho}$ Subtracted From Both Endpoints

Trial	$\hat{\rho}$	Normal Theory	Percentile Method	Bias-Corrected Percentile Method	Smoothed and Bias-Corrected Percentile Method
1	.16	(−.29, .26)	(−.29, .24)	(−.28, .25)	(−.28, .24)
2	.75	(−.17, .09)	(−.05, .08)	(−.13, .04)	(−.12, .08)
3	.55	(−.25, .16)	(−.24, .16)	(−.34, .12)	(−.27, .15)
4	.53	(−.26, .17)	(−.16, .16)	(−.19, .13)	(−.21, .16)
5	.73	(−.18, .10)	(−.12, .14)	(−.16, .10)	(−.20, .10)
6	.50	(−.26, .18)	(−.18, .18)	(−.22, .15)	(−.26, .14)
7	.70	(−.20, .11)	(−.17, .12)	(−.21, .10)	(−.18, .11)
8	.30	(−.29, .23)	(−.29, .25)	(−.33, .24)	(−.29, .25)
9	.33	(−.29, .22)	(−.36, .24)	(−.30, .27)	(−.30, .26)
10	.22	(−.29, .24)	(−.50, .34)	(−.48, .36)	(−.38, .34)
AVE	.48	(−.25, .18)	(−.21, .19)	(−.26, .18)	(−.25, .18)

β, $\beta = E\{\theta(\hat{F}) - \theta(F)\}$. In the notation of Section 3, Quenouille's estimate is

$$\hat{\beta}_J = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}). \tag{18}$$

Subtracting $\hat{\beta}_J$ from $\hat{\theta}$, to correct the bias leads to the *jackknife estimate* of θ, $\hat{\theta}_J = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}$, see Miller (1974), and also Schucany, Gray, and Owen (1971).

There are many ways to justify (18). Here we follow the same line of argument as in the justification of $\hat{\sigma}_J$. The bootstrap estimate of β, which has an obvious motivation, is introduced, and then (18) is related to the bootstrap estimate by a Taylor series argument.

The bias can be thought of as a function of the unknown probability distribution F, $\beta = \beta(F)$. The bootstrap estimate of bias is simply

$$\hat{\beta}_B = \beta(\hat{F}) = E_*\{\theta(\hat{F}^*) - \theta(\hat{F})\}. \tag{19}$$

Here E_* indicates expectation with respect to bootstrap sampling, and $\hat{F}^*$ is the empirical distribution of the bootstrap sample.

In practice $\hat{\beta}_B$ must be approximated by Monte Carlo methods. The only change in the algorithm described in Section 2 is at step (iii), when instead of (or in addition to) $\hat{\sigma}_B$ we calculate

$$\hat{\beta}_B = \hat{\theta}^{*\cdot} - \hat{\theta} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta}).$$

In the sampling experiment of Table 2 the true bias, of $\hat{\rho}$ for estimating ρ, is $\beta = -.014$. The bootstrap estimate $\hat{\beta}_B$, taking $B = 128$, has expectation $-.014$ and standard deviation .031 in this case, while $\hat{\beta}_J$ has expectation $-.017$, standard deviation .040. Bias is a negligible source of statistical error in this situation compared with variability. In applications this is usually made clear by comparison of $\hat{\beta}_B$ with $\hat{\sigma}_B$.

The estimates (18) and (19) are closely related to each other. The argument is the same as in Section 3, except that we approximate $\hat{\theta}(\mathbf{P})$ with a quadratic rather than a linear function of P, say $\hat{\theta}_Q(\mathbf{P}) = a + (\mathbf{P} - \mathbf{P}^o)'\mathbf{b} + \frac{1}{2}(\mathbf{P} - \mathbf{P}^o)'\mathbf{c}(\mathbf{P} - \mathbf{P}^o)$. Let $\hat{\theta}_Q(\mathbf{P})$ be any such quadratic satisfying

$$\hat{\theta}_Q(\mathbf{P}^o) = \hat{\theta}(\mathbf{P}^o) = \hat{\theta} \text{ and } \hat{\theta}_Q(\mathbf{P}_{(i)}) = \hat{\theta}(\mathbf{P}_{(i)}),\ i = 1, 2, \ldots, n.$$

Theorem. The jackknife estimate of bias equals

$$\hat{\beta}_J = \frac{n}{n-1}\,[E_*\{\hat{\theta}_Q(\mathbf{P}^*) - \hat{\theta}\}],$$

which is $n/(n-1)$ times the bootstrap estimate of bias for $\hat{\theta}_Q$ (Efron 1982).

Once again, the jackknife is, almost, a bootstrap estimate itself, except applied to a convenient approximation of $\hat{\theta}(\mathbf{P})$.

More general problems. There is nothing special about bias and standard error as far as the bootstrap is concerned. The bootstrap procedure can be applied to almost any estimation problem.

Suppose that $R(X_1, X_2, \ldots, X_n; F)$ is a random variable, and we are interested in estimating some aspect of R's distribution. (So far we have taken $R = \theta(\hat{F}) - \theta(F)$ and have been interested in the expectation β and the standard deviation σ of R.) The bootstrap algorithm proceeds as described in Section 2, with these two changes: at step (ii), we calculate the bootstrap replication $R^* = R(X_1^*, X_2^*, \ldots, X_n^*; \hat{F})$, and at step (iii) we calculate the distributional property of interest from the empirical distribution of the bootstrap replications R^{*1}, $R^{*2}, \ldots, R^{*B}$.

For example, we might be interested in the probability that the usual t statistic $\sqrt{n}(\bar{X} - \mu)/S$ exceeds 2, where $\mu = E\{X\}$ and $S^2 = \Sigma(X_i - \bar{X})^2/(n-1)$. Then $R^* = \sqrt{n}(\bar{X}^* - \bar{x})/S^*$, and the bootstrap estimate is $\#\{R^{*b} > 2\}/B$. This calculation is used in Section 9 of Efron (1981c) to get confidence intervals for the mean μ in a situation where normality is suspect.

The cross-validation problem of Sections 8 and 9 involves a different type of error random variable R. It will be useful there to use a jackknife-type approximation to the bootstrap expectation of R,

$$E_*\{R^*\} \doteq R^o + (n-1)(R_{(\cdot)} - R^o). \tag{20}$$

Here $R^o = R(x_1, x_2, \ldots, x_n; \hat{F})$ and $R_{(\cdot)} = (1/n)\Sigma R_{(i)}$, $R_{(i)} = R(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n; \hat{F})$. The justification of (20) is the same as for the theorem of this section, being based on a quadratic approximation formula.

7. MORE COMPLICATED DATA SETS

So far we have considered the simplest kind of data sets, where all the observations come from the same distribution F. The bootstrap idea, and jackknife-type approximations (which are not discussed here), can be applied to much more complicated situations. We begin with a two-sample problem.

The data in our first example consist of two independent random samples,

$$X_1, X_2, \ldots, X_m \sim F \text{ and } Y_1, Y_2, \ldots, Y_n \sim G,$$

F and G being two possibly different distributions on the real line. The statistic of interest is the Hodges-Lehmann shift estimate

$$\hat{\theta} = \text{median}\,\{y_j - x_i;\ i = 1, \ldots, m,\ j = 1, \ldots, n\}.$$

We desire an estimate of the standard error $\sigma(F, G)$.

The bootstrap estimate is simply

$$\hat{\sigma}_B = \sigma(\hat{F}, \hat{G}),$$

$\hat{G}$ being the empirical distribution of the y_i. This is evaluated by Monte Carlo, as in Section 3, with obvious modifications: a bootstrap sample now consists of a random sample $X_1^*, X_2^*, \ldots, X_m^*$ drawn from $\hat{F}$ *and* an independent random sample $Y_1^*, \ldots, Y_n^*$ drawn from $\hat{G}$. (In other words, m draws with replacement from $\{x_1, x_2, \ldots, x_m\}$, and n draws with replacement from $\{y_1, y_2, \ldots, y_n\}$.) The bootstrap replication $\hat{\theta}^*$ is the median of the mn differences $Y_j^* - X_i^*$. Then $\hat{\sigma}_B$ is approximated from B independent such replications as on the right side of (11).

Table 4 shows the results of a sampling experiment in

Table 4. Bootstrap Estimates of Standard Error for the Hodges-Lehmann Two-Sample Shift Estimate; $m = 6$, $n = 9$; True Distributions Both F and G Uniform [0, 1]

		Expectation	St. Dev.	C.V.	$\sqrt{MSE}$
Separate	B = 100	.165	.030	.18	.030
	B = 200	.166	.031	.19	.031
Combined	B = 100	.145	.028	.19	.036
	B = 200	.149	.025	.17	.031
True Standard Error		.167			

which $m = 6$, $n = 9$, and both F and G were uniform distributions on the interval [0, 1]. The table is based on 100 trials of the situation. The true standard error is $\sigma(F, G) = .167$. "Separate" refers to $\hat{\sigma}_B$ calculated exactly as described in the previous paragraph. The improvement in going from $B = 100$ to $B = 200$ is too small to show up in the table.

"Combined" refers to the following idea: suppose we believe that G is really a translate of F. Then it wastes information to estimate F and G separately. Instead we can form the combined empirical distribution

$$\hat{H}: \text{mass } \frac{1}{m+n} \text{ on}$$

$$x_1, x_2, \ldots, x_m, y_1 - \hat{\theta}, y_2 - \hat{\theta}, \ldots, y_n - \hat{\theta}.$$

All $m + n$ bootstrap variates $X_1^*, \ldots, X_m^*, Y_1^*, \ldots, Y_n^*$ are then sampled independently from $\hat{H}$. (We could add $\hat{\theta}$ back to the Y_j^* values, but this has no effect on the bootstrap standard error estimate, since it just adds the constant $\hat{\theta}$ to each bootstrap replication $\hat{\theta}^*$.)

The combined method gives no improvement here, but it might be valuable in a many-sample problem where there are small numbers of observations in each sample, a situation that arises in stratified sampling. (See Efron 1982, Ch. 8.) The main point here is that "bootstrap" is not a well-defined verb, and that there may be more than one way to proceed in complicated situations. Next we consider regression problems, where again there is a choice of bootstrapping methods.

In a typical regression problem we observe n independent real-valued quantitives $Y_i = y_i$,

$$Y_i = g_i(\beta) + \varepsilon_i, \; i = 1, 2, \ldots, n. \tag{21}$$

The functions $g_i(\cdot)$ are of known form, usually $g_i(\beta) = g(\beta; t_i)$, where t_i is an observed **p**-dimensional vector of covariates; β is a vector of unknown parameters we wish to estimate. The ε_i are an independent and identically distributed random sample from some distribution F on the real line,

$$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \sim F,$$

where F is assumed to be centered at zero in some sense, perhaps $E\{\varepsilon\} = 0$ or $\text{Prob}\{\varepsilon < 0\} = 0.5$.

Having observed the data vector $\mathbf{Y} = \mathbf{y} = (y_1, \ldots, y_n)$, we estimate β by minimizing some measure of distance between $\mathbf{y}$ and the vector of predicted values $\boldsymbol{\eta}(\beta) = (g_1(\beta), \ldots, g_n(\beta))$,

$$\hat{\beta}: \min_{\beta} D(\mathbf{y}, \boldsymbol{\eta}(\beta)).$$

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

Having calculated $\hat{\beta}$, we can modify the one-sample bootstrap algorithm of Section 2, and obtain an estimate of $\hat{\beta}$'s variability:

(i) Construct $\hat{F}$ putting mass $1/n$ at each observed residual,

$$\hat{F}: \text{mass } 1/n \text{ on } \hat{\epsilon}_i = y_i - g_i(\hat{\beta}).$$

(ii) Construct a bootstrap data set

$$Y_i^* = g_i(\hat{\beta}) + \epsilon_i^*, \; i = 1, 2, \ldots, n,$$

where the ϵ_i^* are drawn independently from $\hat{F}$, and calculate

$$\hat{\beta}^*: \min_{\beta} D(\mathbf{Y}^*, \boldsymbol{\eta}(\beta)).$$

(iii) Do step (ii) some large number B of times, obtaining independent bootstrap replications $\hat{\beta}^{*1}$, $\hat{\beta}^{*2}$, $\ldots$, $\hat{\beta}^{*B}$, and estimate the covariance matrix of $\hat{\beta}$ by

$$\hat{\boldsymbol{\Sigma}}_B = \left[\left(\sum_{b=1}^{B} (\hat{\beta}^{*b} - \hat{\beta}^{*\cdot})(\hat{\beta}^{*b} - \hat{\beta}^{*\cdot})'\right) \Big/ (B-1)\right],$$

$$(\hat{\beta}^{*\cdot} = \frac{1}{B} \Sigma \hat{\beta}^{*b}).$$

In ordinary linear regression we have $g_i(\beta) = t_i'\beta$ and $D(\mathbf{y}, \boldsymbol{\eta}) = \Sigma(y_i - \eta_i)^2$. Section 7 of Efron (1979a) shows that in this case the algorithm above can be carried out theoretically, $B = \infty$, and yields

$$\hat{\boldsymbol{\Sigma}}_B = \hat{\sigma}^2 \left(\sum_{i=1}^{n} t_i t_i'\right)^{-1}, \; \hat{\sigma}^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2/n. \tag{22}$$

This is the usual answer, except for dividing by n instead of $n - p$ in $\hat{\sigma}^2$. Of course the advantage of the bootstrap approach is that $\hat{\boldsymbol{\Sigma}}_B$ can just as well be calculated if, say, $g_i(\beta) = \exp(t_i\beta)$ and $D(\mathbf{y}, \boldsymbol{\eta}) = \sum_{i=1}^n |y_i - \eta_i|$.

There is another simpler way to bootstrap the regression problem. We can consider each covariate-response pair $x_i = (t_i, y_i)$ to be a single data point obtained by random sampling from a distribution F on $p + 1$ dimension space. Then we apply the one-sample bootstrap of Section 2 to the data set $x_1, x_2, \ldots, x_n$.

The two bootstrap methods for the regression problem are asymptotically equivalent, but can perform quite differently in small-sample situations. The simple method, described last, takes less advantage of the special structure of the regression problem. It does *not* give answer (22) 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 (21) is not correct.* For this reason we use the simple method of bootstrapping on the error rate prediction problem of Sections 9 and 10.

As a final example of bootstrapping complicated data

we consider a two-sample problem with censored data. The data are the leukemia remission times listed in Table 1 of Cox (1972). The sample sizes are $m = n = 21$. Treatment-group remission times (weeks) are 6+, 6, 6, 6, 7, 9+, 10+, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+; control-group remission times (weeks) are 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23. Here 6+ indicates a *censored* remission time, known only to exceed 6 weeks, while 6 is an uncensored remission time of exactly 6 weeks. None of the control-group times were censored.

We assume Cox's proportional hazards model, the hazard rate in the control group equaling e^{β} times that in the Treatment group. The partial likelihood estimate of β is $\hat{\beta} = 1.51$, and we want to estimate the standard error of $\hat{\beta}$. (Cox gets 1.65, not 1.51. Here we are using Breslow's convention for ties (1972), which accounts for the discrepancy.)

Figure 4 shows the histogram for 1000 bootstrap replications of $\hat{\beta}^*$. Each replication was obtained by the two-sample method described for the Hodges-Lehmann estimate:

(i) Construct $\hat{F}$ putting mass $\frac{1}{21}$ at each point 6+, 6, 6, ..., 35+, and $\hat{G}$ putting mass $\frac{1}{21}$ at each point 1, 1, ..., 23. (Notice that the "points" in $\hat{F}$ include the censoring information.)

(ii) Draw $X_1^*, X_2^*, \ldots, X_{21}^*$ by random sampling from $\hat{F}$, and likewise $Y_1^*, Y_2^*, \ldots, Y_{21}^*$ by random sampling from $\hat{G}$. Calculate $\hat{\beta}^*$ by applying the partial-likelihood method to the bootstrap data.

The bootstrap estimate of standard error for $\hat{\beta}$, as given by (11), is $\hat{\sigma}_B = .42$. This agrees nicely with Cox's asymptotic estimate $\hat{\sigma} = .41$. However, the percentile method gives quite different confidence intervals from those obtained by the usual method. For $\alpha = .05$, $1 - 2\alpha = .90$, the latter interval is $1.51 \pm 1.65 \cdot .41 = [.83, 2.19]$. The percentile method gives the 90 percent central interval [.98, 2.35]. Notice that $(2.35 - 1.51)/(1.51 - .98) = 1.58$, so that the percentile interval is considerably larger to the right of $\hat{\beta}$ than to the left. (The bias-corrected percentile method gives almost the same answers as the uncorrected method in this case since $\hat{C}(\hat{\beta}) = .49$.)

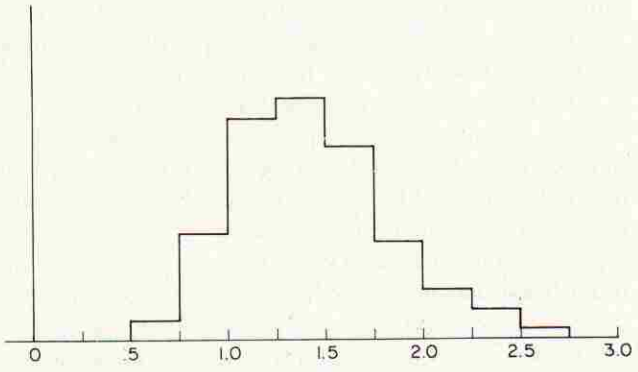

Figure 4. Histogram of 1000 bootstrap replications of $\hat{\beta}^*$ for the leukemia data, proportional hazards model. Courtesy of Rob Tibshirani, Stanford.

There are other reasonable ways to bootstrap censored data. One of these is described in Efron (1981a), which also contains a theoretical justification for the method used to construct Figure 4.

8. CROSS-VALIDATION

Cross-validation is an old but useful idea, whose time seems to have come again with the advent of modern computers. We discuss it in the context of estimating the error rate of a prediction rule. (There are other important uses; see Stone 1974; Geisser 1975.)

The prediction problem is as follows: each data point $x_i = (t_i, y_i)$ consists of a p-dimensional vector of explanatory variables t_i, and a response variable y_i. Here we assume y_i can take on only two possible values, say 0 or 1, indicating two possible responses, live or dead, male or female, success or failure, and so on. We observe $x_1, x_2, \ldots, x_n$, called collectively *the training set*, and indicated $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. We have in mind a formula $\eta(t; \mathbf{x})$ for constructing a *prediction rule* from the training set, also taking on values either 0 or 1. Given a new explanatory vector t_0, the value $\eta(t_0; \mathbf{x})$ is supposed to predict the corresponding response y_0.

We assume that each x_i is an independent realization of $X = (T, Y)$, a random vector having some distribution F on $p+1$-dimensional space, and likewise for the "new case" $X_0 = (T_0, Y_0)$. The *true error rate* err of the prediction rule $\eta(\cdot; \mathbf{x})$ is the expected probability of error over $X_0 \sim F$ with $\mathbf{x}$ fixed,

$$\text{err} \equiv E\{Q[Y_0, \eta(T_0, \mathbf{x})]\},$$

where $Q[y, \eta]$ is the error indicator

$$Q[y, \eta] = \begin{cases} 0 & \text{if } y = \eta \\ 1 & \text{if } y \neq \eta. \end{cases}$$

An obvious estimate of err is the *apparent error rate*

$$\overline{\text{err}} \equiv \hat{E}\{Q[Y_0, \eta(T_0; \mathbf{x})]\} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta(t_i; \mathbf{x})].$$

The symbol $\hat{E}$ indicates expectation with respect to the empirical distribution $\hat{F}$, putting mass $1/n$ on each x_i. The apparent error rate is likely to underestimate the true error rate, since we are evaluating $\eta(\cdot, \mathbf{x})$'s performance on the same set of data used in its construction. A random variable of interest is the *overoptimism*, true minus apparent error rate,

$$\begin{aligned} R(\mathbf{x}, F) &= \text{err} - \overline{\text{err}} \\ &= E\{Q[Y_0, \eta(T_0; \mathbf{x})]\} - \hat{E}\{Q[Y_0, \eta(T_0; \mathbf{x})]\}. \end{aligned} \quad (23)$$

The expectation of $R(\mathbf{X}, F)$ over the random choice of $X_1, X_2, \ldots, X_n$ from F,

$$\omega(F) \equiv ER(\mathbf{X}, F) \quad (24)$$

is the *expected overoptimism*.

The cross-validated estimate of err is

$$\text{err}^{\dagger} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \eta(t_i; \mathbf{x}_{(i)})],$$

$\eta(t_i; \mathbf{x}_i)$ being the prediction rule based on $\mathbf{x}_{(i)} =$

$(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. In other words $\text{err}^{\dagger}$ is the error rate over the observed data set, *not allowing* $x_i = (t_i, y_i)$ *to enter into the construction of the rule for its own prediction*. It is intuitively obvious that $\text{err}^{\dagger}$ is a less biased estimator of err than is $\widehat{\text{err}}$. In what follows we consider how well $\text{err}^{\dagger}$ estimates err, or equivalently how well

$$\omega^{\dagger} \equiv \text{err}^{\dagger} - \overline{\text{err}}$$

estimates $R(\mathbf{x}, F) = \text{err} - \overline{\text{err}}$. (These are equivalent problems since $\text{err}^{\dagger} - \text{err} = \omega^{\dagger} - R(\mathbf{x}, F)$.) We have used the notation $\omega^{\dagger}$, rather than $R^{\dagger}$, because it turns out later that it is actually ω being estimated.

We consider a sampling experiment involving Fisher's linear discriminant function. The dimension is $p = 2$ and the sample size of the training set is $n = 14$. The distribution F is as follows: $Y = 0$ or 1 with probability $\frac{1}{2}$, and given $Y = y$ the predictor vector T is bivariate normal with identity covariance matrix and mean vector $(y - \frac{1}{2}, 0)$. If F were known to the statistician, the ideal prediction rule would be to guess $y_0 = 0$ if the first component of t_0 was ≤ 0, and to guess $y_0 = 1$ otherwise. Since F is assumed unknown, we must estimate a prediction rule from the training set.

We use the prediction rule based on Fisher's estimated linear discriminant function (Efron 1975),

$$\eta(t; \mathbf{x}) = \begin{cases} 0 \\ 1 \end{cases} \text{ if } \hat{\alpha} + t'\hat{\beta} \text{ is } \begin{cases} \leq 0 \\ > 0 \end{cases}.$$

The quantities $\hat{\alpha}$ and $\hat{\beta}$ are defined in terms of n_0 and n_1, the number of y_i equal to zero and one, respectively; $\bar{t}_0$ and $\bar{t}_1$, the averages of the t_i corresponding to those y_i equaling zero and one, respectively; and $S = [\sum_{i=1}^{n} t_i t_i' - n_0 \bar{t}_0 \bar{t}_0' - n_1 \bar{t}_1 \bar{t}_1']/n$:

$$\hat{\alpha} = [\bar{t}_1' S^{-1} \bar{t}_1 - \bar{t}_2' S^{-1} \bar{t}_2]/2,$$

$$\hat{\beta} = (\bar{t}_2 - \bar{t}_1) S^{-1}.$$

Table 5 shows the results of 10 simulations ("trials") of this situation. The expected overoptimism, obtained from 100 trials, is $\omega = .098$, so that $R = \text{err} - \overline{\text{err}}$ is typically quite large. However, R is also quite variable from trial to trial, often being negative. The cross-validation estimate $\omega^{\dagger}$ is positive in all 10 cases, and does not correlate with R. This relates to the comment that $\omega^{\dagger}$ is trying to estimate ω rather than R. We will see later that $\omega^{\dagger}$ has expectation .091, and so is nearly unbiased for ω. However, $\omega^{\dagger}$ is too variable itself to be very useful for estimating R, which is to say that $\text{err}^{\dagger}$ is not a particularly good estimate of err. These points are discussed further in Section 9, where the two other estimates of ω appearing in Table 5, $\hat{\omega}_J$ and $\hat{\omega}_B$, are introduced.

Table 5. The First 10 Trials of a Sampling Experiment Involving Fisher's Linear Discriminant Function. The Training Set Has Size n = 14. The Expected Overoptimism is ω = .096, see Table 6

		Error Rates			Estimates of Overoptimism		
Trial	n_0, n_1	True $\overline{\text{err}}$	Apparent err	Overoptimism R	Cross-validation $\omega^{\dagger}$	Jackknife $\hat{\omega}_J$	Bootstrap ($B = 200$) $\hat{\omega}_B$
1	9,5	.458	.286	.172	.214	.214	.083
2	6,8	.312	.357	−.045	.000	.066	.098
3	7,7	.313	.357	−.044	.071	.066	.110
4	8,6	.351	.429	−.078	.071	.066	.107
5	8,6	.330	.357	−.027	.143	.148	.102
6	8,6	.318	.143	.175	.214	.194	.073
7	8,6	.310	.071	.239	.071	.066	.087
8	6,8	.382	.286	.094	.071	.056	.097
9	7,7	.360	.429	−.069	.071	.087	.127
10	8,6	.335	.143	−.192	.000	.010	.048

9. BOOTSTRAP AND JACKKNIFE ESTIMATES FOR THE PREDICTION PROBLEMS

At the end of Section 6 we described a method for applying the boostrap to any random variable $R(\mathbf{X}, F)$. Now we use that method on the overoptimism random variable (23), and obtain a bootstrap estimate of the expected overoptimism $\omega(F)$.

The bootstrap estimate of $\omega = \omega(F)$, (24), is simply

$$\hat{\omega}_B = \omega(\hat{F}).$$

As usual $\hat{\omega}_B$ must be approximated by Monte Carlo. We generate independent bootstrap replications $R^{*1}, R^{*2}, \ldots, R^{*B}$, and take

$$\hat{\omega}_B \doteq \frac{1}{B} \sum_{b=1}^{B} R^{*b}.$$

As B goes to infinity this last expression approaches $E_*\{R^*\}$, the expectation of R^* under bootstrap resampling, which is by definition the same quantity as $\omega(\hat{F}) = \hat{\omega}_B$. The bootstrap estimates $\hat{\omega}_B$ seen in the last column of Table 5 are considerably less variable than the cross-validation estimates $\omega^{\dagger}$.

What does a typical bootstrap replication consist of in this situation? As in Section 3 let $\mathbf{P}^* = (P_1^*, P_2^*, \ldots, P_n^*)$ indicate the bootstrap resampling proportions $P_i^* = \#\{X_j^* = x_i\}/n$. (Notice that we are considering each vector $x_i = (t_i, y_i)$ as a single sample point for the purpose of carrying out the bootstrap algorithm.) Following through definition (13), it is not hard to see that

$$R^* = R(\mathbf{X}^*, \hat{F}) = \sum_{i=1}^{n} (P_i^o - \mathbf{P}_i^*)\, Q[y_i, \eta(t_i; \mathbf{X}^*)], \quad (25)$$

where $\mathbf{P}^o = (1, 1, \ldots, 1)'/n$ as before, and $\eta(\cdot, \mathbf{X}^*)$ is the prediction rule based on the bootstrap sample.

Table 6 shows the results of two simulation experiments (100 trials each) involving Fisher's linear discriminant fraction. The left side relates to the bivariate normal situation described in Section 8: sample size $n = 14$, dimension $d = 2$, mean vectors for the two randomly selected normal distributions $= (\pm\frac{1}{2}, 0)$. The right side still has $n = 14$, but the dimension has been raised to 5, with mean vectors $(\pm 1, 0, 0, 0, 0)$. Fuller descriptions appear in Chapter 7 of Efron (1982).

Seven estimates of overoptimism were considered. In the $d = 2$ situation, the cross-validation estimate $\omega^{\dagger}$, for example, had expectation .091, standard deviation .073, and correlation −.07 with R. This gave root mean

Table 6. Two Sampling Experiments Involving Fisher's Linear Discriminant Function. The Left Side of the Table Relates to the Situation of Table 5: $n = 14$, $d = 2$, True Mean Vectors $= (\pm\frac{1}{2}, 0)$. The Right Side Relates to $n = 14$, $d = 5$, True Mean Vectors $= (\pm 1, 0, 0, 0, 0)$

Overoptimism $R(\mathbf{X}, F)$	Dimension 2 Exp. $\omega = .096$	Sd. .113	Corr.	$\sqrt{MSE}$	Dimension 5 Exp. $\omega = .184$	Sd. .099	Corr.	$\sqrt{MSE}$
1. Ideal Constant	.096	0	0	.113	.184	0	0	.099
2. Cross-Validation	.091	.073	−.07	.139	.170	.094	−.15	.147
3. Jackknife	.093	.068	−.23	.145	.167	.089	−.26	.150
4. Bootstrap ($B = 200$)	.080	.028	−.64	.135	.103	.031	−.58	.145
5. BootRand ($B = 200$)	.087	.026	−.55	.130	.147	.020	−.31	.114
6. BootAve ($B = 200$)	.100	.036	−.18	.125	.172	.041	−.25	.118
7. Zero	0	0	0	.149	0	0	0	.209

squared error, of $\omega^\dagger$ for estimating R or equivalently of $\text{err}^\dagger$ for estimating err,

$$[E[\omega^\dagger - R]^2]^{\frac{1}{2}} = [E(\text{err}^\dagger - \text{err})^2]^{\frac{1}{2}} = .139.$$

The bootstrap, line 4, did only slightly better, $\sqrt{\text{MSE}} = .135$.

The zero estimate $\hat{\omega} \equiv 0$, line 7, had $\sqrt{\text{MSE}} = .149$, which is also $[E(\text{err} - \overline{\text{err}})^2]^{\frac{1}{2}}$, the $\sqrt{\text{MSE}}$ of estimating err by the apparent error $\overline{\text{err}}$, with zero correction for overoptimism. The "ideal constant" is ω itself. If we knew ω, which we don't in genuine applications, we would use the bias-corrected estimate $\overline{\text{err}} + \omega$. Line 1, left side, says that this ideal correction gives $\sqrt{\text{MSE}} = .113$.

We see that neither cross-validation nor the bootstrap are much of an improvement over making no correction at all, though the situation is more favorable on the right side of Table 6. Estimators 5 and 6, which will be described later, perform noticeably better.

The "jackknife," line 3, refers to the following idea: since $\hat{\omega}_B = E_*\{R^*\}$ is a bootstrap expectation, we can approximate that expectation by (19). In this case (25) gives $R^0 = 0$, so the jackknife approximation is simply $\hat{\omega}_J = (n-1)\, R_{(\cdot)}$. Evaluating this last expression, as in Chapter 7 of Efron (1982), gives

$$\hat{\omega}_J = \frac{1}{n}\sum_{i=1}^{n}\left\{ Q[y_i, \eta(t_i, \mathbf{x}_{(i)})] - \left(\sum_{j=1}^{n} Q[y_i, \eta(t_i, \mathbf{x}_{(j)})]\right)\Big/ n\right\}.$$

This looks very much like the cross-validation estimate, which can be written

$$\omega^\dagger = \frac{1}{n}\sum_{i=1}^{n}\{Q[y_i, \eta(t_i, \mathbf{x}_{(i)})] - Q[y_i, \eta(t_i, \mathbf{x})]\}.$$

As a matter of fact, $\hat{\omega}_J$ and $\omega^\dagger$ have asymptotic correlation one (Gong 1982). Their nearly perfect correlation can be seen in Table 5. In the sampling experiments of Table 6, $\text{corr}(\hat{\omega}_J, \omega^\dagger) = .93$ on the left side, and .98 on the right side. The point here is that *the cross-validation estimate $\omega^\dagger$ is, essentially, a Taylor series approximation to the bootstrap estimate $\hat{\omega}_B$.*

Even though $\hat{\omega}_B$ and $\omega^\dagger$ are closely related in theory and are asymptotically equivalent, they behave very differently in Table 6: $\omega^\dagger$ is nearly unbiased and uncorrelated with R, but has enormous variability; $\hat{\omega}_B$ has small variability, but is biased downwards, particularly in the right-hand case, and highly negatively correlated with R. The poor performances of the two estimators are due to different causes, and there are some grounds of hope for a favorable hybrid.

"BootRand," line 5, modified the bootstrap estimate in just one way: instead of drawing the bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ from $\hat{F}$, it was drawn from

$$\hat{F}_{\text{RAND}}\colon \text{mass} \begin{cases} \hat{\pi}_i/n & \text{on } (t_i, 1) \\ (1-\hat{\pi}_i)/n & \text{on } (t_i, 0) \end{cases} \qquad i = 1, 2, \ldots, n.$$

This is a distribution supported on $2n$ points, the observed points $x_i = (t_i, y_i)$ and also the complementary points $(t_i, 1 - y_i)$. The probabilities $\hat{\pi}_i$ were those naturally associated with the linear discriminant function,

$$\hat{\pi}_i = 1/[1 + \exp - (\hat{\alpha} + t_i'\hat{\beta})]$$

(see Efron 1975), except that $\hat{\pi}_i$ was always forced to lie in the interval $[.1, .9]$.

Drawing the bootstrap sample $X_1^*, \ldots, X_n^*$ from $\hat{F}_{\text{RAND}}$ instead of $\hat{F}$ is a form of smoothing, not unlike the smoothed bootstraps of Section 2. In both cases we support the estimate of F on points beyond those actually observed in the sample. Here the smoothing is entirely in the response variable y. In complicated problems, such as the one described in Section 10, t_i can have complex structure (censoring, missing values, cardinal and ordinal scales, discrete and continuous variates, etc.) making it difficult to smooth in the t space. Notice that in Table 6 BootRand is an improvement over the ordinary bootstrap in every way: it has smaller bias, smaller standard deviation, and smaller negative correlation with R. The decrease in $\sqrt{\text{MSE}}$ is especially impressive on the right side of the table.

"BootAve," line 6, involves a quantity we shall call $\tilde{\omega}_0$. Generating B bootstrap replications involves making nB predictions $\eta(t_i, \mathbf{X}^{*b})$, $i = 1, 2, \ldots, n$, $b = 1, 2, \ldots, B$. Let

$$I_{ib}^* = \begin{cases} 1 & \text{if } P_i^{*b} = 0 \\ 0 & \text{if } P_i^{*b} > 0 \end{cases}$$

Then

$$\tilde{\omega}_0 \equiv \Sigma_{i,b}\, I_{i,b}^*\, Q[y_i, \eta(t_i, \mathbf{X}^{*b})]/\Sigma_{ib}\, I_{ib}^* - \overline{\text{err}}.$$

In other words, $\tilde{\omega}_0 + \overline{\text{err}}$ is the observed bootstrap error rate for prediction of those y_i where x_i is not involved in the construction of $\eta(\cdot, \mathbf{X}^*)$. Theoretical arguments can be mustered to show that $\tilde{\omega}_0$ will usually have expectation greater than ω, while $\tilde{\omega}_B$ usually has expectation less than ω. "BootAve" is the compromise estimator $\hat{\omega}_{\text{AVE}} = (\hat{\omega}_B + \tilde{\omega}_0)/2$. It also performs well in Table 6, though there is not yet enough theoretical or numerical evidence to warrant unqualified enthusiasm.

The bootstrap is a general all-purpose device that can be applied to almost any problem. This is very handy,

Table 7. The Last 11 Liver Patients. Negative Numbers Indicate Missing Values

y	Constant 1	Age 2	Sex 3	Steroid 4	Antiviral 5	Fatigue 6	Malaise 7	Anorexia 8	Liver Big 9	Liver Firm 10	Spleen Palp 11	Spiders 12	Ascites 13	Varices 14	Bilirubin 15	Alk Phos 16	SGOT 17	Albumin 18	Protein 19	Histology 20	#
1	1	45	1	2	2	1	1	1	2	2	2	1	1	2	1.90	−1	114	2.4	−1	−3	145
0	1	31	1	1	2	1	2	2	2	2	2	2	2	2	1.20	75	193	4.2	54	2	146
1	1	41	1	2	2	1	2	2	2	1	1	1	2	1	4.20	65	120	3.4	−1	−3	147
1	1	70	1	1	2	1	1	1	−3	−3	−3	−3	−3	−3	1.70	109	528	2.8	35	2	148
0	1	20	1	1	2	2	2	2	2	−3	2	2	2	2	.90	89	152	4.0	−1	2	149
0	1	36	1	2	2	2	2	2	2	2	2	2	2	2	.60	120	30	4.0	−1	2	150
1	1	46	1	2	2	1	1	1	2	2	2	1	1	1	7.60	−1	242	3.3	50	−3	151
0	1	44	1	2	2	1	2	2	2	1	2	2	2	2	.90	126	142	4.3	−1	2	152
0	1	61	1	1	2	1	1	2	1	1	2	1	2	2	.80	95	20	4.1	−1	2	153
0	1	53	2	1	2	1	2	2	2	2	1	1	2	1	1.50	84	19	4.1	48	−3	154
1	1	43	1	2	2	1	2	2	2	2	1	1	1	2	1.20	100	19	3.1	42	2	155

but it implies that in situations with special structure the bootstrap may be outperformed by more specialized methods. Here we have done so in two different ways. BootRand uses an estimate of F that is better than the totally nonparametric estimate F. BootAve makes use of the particular form of R for the overoptimism problem.

10. A COMPLICATED PREDICTION PROBLEM

We end this article with the bootstrap analysis of a genuine prediction problem, involving many of the complexities and difficulties typical of genuine problems. The bootstrap is not necessarily the best method here, as discussed in Section 9, but it is impressive to see how much information this simple idea, combined with massive computation, can extract from a situation that is hopelessly beyond traditional theoretical solutions. A fuller discussion appears in Efron and Gong (1981).

Among $n = 155$ acute chronic hepatitis patients, 33 were observed to die from the disease, while 122 survived. Each patient had associated a vector of 20 covariates. On the basis of this training set it was desired to produce a rule for predicting, from the covariates, whether a given patient would live or die. If an effective prediction rule were available, it would be useful in choosing among alternative treatments. For example, patients with a very low predicted probability of death could be given less rigorous treatment.

Let $x_i = (t_i, y_i)$ represent the data for patient i, $i = 1, 2, \ldots, 155$. Here t_i is the 20-dimensional vector of covariates, and y_i equals 1 or 0 as the patient died or lived. Table 7 shows the data for the last 11 patients. Negative numbers represent missing values. Variable 1 is the constant 1, included for convenience. The meaning of the 19 other predictors, and their coding in Table 7, will not be explained here.

A prediction rule was constructed in 3 steps:

1. An $\alpha = .05$ test of the importance of predictor j, $H_0: \beta_j = 0$ versus $H_1: \beta_j \neq 0$, was run separately for $j = 2, 3, \ldots, 20$, based on the logistic model

$$\log \frac{\pi(t_i)}{1 - \pi(t_i)} = \beta_i + \beta_j t_{ij},$$

$$\pi(t_i) \equiv \text{Prob}\{\text{patient } i \text{ dies}\}.$$

Among these 19 tests, 13 predictors indicated predictive power by rejecting $H_0: j = 18, 13, 15, 12, 14, 7, 6, 19, 20, 11, 2, 5, 3$. These are listed in order of achieved significance level, $j = 18$ attaining the smallest alpha.

2. These 13 predictors were tested in a forward multiple-logistic-regression program, which added predictors one at a time (beginning with the constant) until no further single addition achieved significance level $\alpha = .10$. Five predictors besides the constant survived this step, $j = 13, 20, 15, 7, 2$.

3. A final forward, stepwise multiple-logistic-regression program on these five predictors, stopping this time at level $\alpha = .05$, retained four predictors besides the constant, $j = 13, 15, 7, 20$.

At each of the three steps, only those patients having no relevant data missing were included in the hypothesis tests. At step 2 for example, a patient was included only if all 13 variables were available.

The final prediction rule was based on the estimated logistic regression

$$\log \frac{\pi(t_i)}{1 - \pi(t_i)} = \sum_{J=1,\ 13,\ 15,\ 7,\ 20} \hat{\beta}_j\, t_{ij},$$

where $\hat{\beta}_j$ was the maximum likelihood estimate in this model. The prediction rule was

$$\eta(t;\, \mathbf{x}) = \begin{cases} 1 \\ 0 \end{cases} \text{ if } \Sigma_j\, \hat{\beta}_j\, t_{ij} \begin{cases} \geq c \\ < c \end{cases}, \tag{26}$$

$c = \log 33/122$.

Among the 155 patients, 133 had none of the predictors 13, 15, 7, 20 missing. When the rule $\eta(t;\, \mathbf{x})$ was applied to these 133 patients, it misclassified 21 of them, for an apparent error rate $\overline{\text{err}} = 21/133 = .158$. We would like to estimate how overoptimistic $\overline{\text{err}}$ is.

To answer this question, the simple bootstrap was applied as described in Section 9. A typical bootstrap sample consisted of $X_1^*, X_2^*, \ldots, X_{155}^*$, randomly drawn with replacement from the training set $x_1, x_2, \ldots, x_{155}$. The bootstrap sample was used to construct the bootstrap prediction rule $\eta(\cdot, \mathbf{X}^*)$, following the same three steps used in the construction of $\eta(\cdot, \mathbf{x})$, (26). This gives a bootstrap replication R^* for the overoptimism random variable $R = \text{err} - \overline{\text{err}}$, essentially as in (25), but with a modification to allow for difficulties caused by missing predictor values.

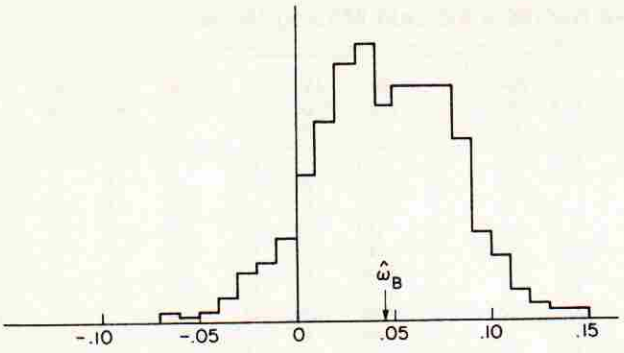

Figure 5. Histogram of 500 bootstrap replications of overoptimism for the hepatitis problem.

Figure 5 shows the histogram of $B = 500$ such replications. 95 percent of these fall in the range $0 \le R^* \le .12$. This indicates that the unobservable true overoptimism $\text{err} - \overline{\text{err}}$ is likely to be positive. The average value is

$$\hat{\omega}_B = \frac{1}{B}\sum_{b=1}^{B} R^{*b} = .045,$$

suggesting that the expected overoptimism is about $\frac{1}{3}$ as large as the apparent error rate .158. Taken literally, this gives the bias-corrected estimated error rate $.158 + .045 = .203$. There is obviously plenty of room for error in this last estimate, given the spread of values in Figure 5, but at least we now have some idea of the possible bias in err.

The bootstrap analysis provided more than just an estimate of $\omega(F)$. For example, the standard deviation of the histogram in Figure 5 is .036. This is a dependable estimate of the true standard deviation of R

13	7	20	15			
13	19	6				
20	16	19				
20	19					
14	18	7	16	2		
18	20	7	11			
20	19	15				
20						
13	12	15	8	18	7	19
15	13	19				
13	4					
12	15	3				
15	16	3				
15	20	4				
16	13	2	19			
18	20	3				
13	15	20				
15	13					
15	20	7				
13						
15						
13	14					
12	20	18				
2	20	15	7	19	12	
13	20	15	19			

Figure 6. Predictors selected in the last 25 bootstrap replications for the hepatitis program. The predictors selected by the actual data were 13, 15, 7, 20.

(see Efron 1982, Ch. VII), which by definition equals $[E(\text{err} - \overline{\text{err}} - \omega)^2]^{1/2}$, the $\sqrt{\text{MSE}}$ of $\overline{\text{err}} + \omega$ as an estimate of err. Comparing line 1 with line 4 in Table 6, we expect $\overline{\text{err}} + \hat{\omega}_B = .203$ to have $\sqrt{\text{MSE}}$ at least this big for estimating err.

Figure 6 illustrates another use of the bootstrap replications. The predictions chosen by the three-step selection procedure, applied to the bootstrap training set $\mathbf{X}^*$, are shown for the last 25 of the 500 replications. Among all 500 replications, predictor 13 was selected 37 percent of the time, predictor 15 selected 48 percent, predictor 7 selected 35 percent, and predictor 20 selected 59 percent. No other predictor was selected more than 50 percent of the time. No theory exists for interpreting Figure 6, but the results certainly discourage confidence in the casual nature of the predictors 13, 15, 7, 20.

[*Received January 1982. Revised May 1982.*]

REFERENCES

BRESLOW, N. (1972), Discussion of Cox (1974), *Journal of the Royal Statistical Society*, Ser. B, 34, 216–217.

COX, D.R. (1972), "Regression Models With Life Tables," *Journal of the Royal Statistical Society*, Ser. B, 34, 187–000.

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

EFRON, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association*, 70, 897–898.

——— (1979a), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1–26.

——— (1979b), "Computers and the Theory of Statistics: Thinking the Unthinkable," *SIAM Review*, 21, 460–480.

——— (1981a), "Censored Data and the Bootstrap," *Journal of the American Statistical Association*, 76, 312–319.

——— (1981b), "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and Other Resampling Methods," *Biometrika*, 00, 000–000.

——— (1981c), "Nonparametric Standard Errors and Confidence Intervals," *Canadian Journal of Statistics*, 9, 139–172.

——— (1982), "The Jackknife, the Bootstrap, and Other Resampling Plans," *SIAM*, monograph #38, CBMS–NSF.

EFRON, B., and GONG, G. (1981), "Statistical Theory and the Computer," unpublished manuscript.

GEISSER, S. (1975), "The Predictive Sample Reuse Method With Applications," *Journal of the American Statistical Association*, 70, 320–328.

GONG, G. (1982), "Cross-validation, the Jackknife, and the Bootstrap: Excess Error Estimation in Forward Logistic Regression", Ph.D. dissertation, Dept. of Statistics, Stanford University.

HAMPEL, F. (1974), "The Influence Curve and its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

JAECKEL, L. (1972), "The Infinitesimal Jackknife," Bell Laboratories Memorandum #MM 72-1215-11.

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

MALLOWS, C.L. (1974), "On Some Topics in Robustness", Memorandum, Bell Laboratories, Murray Hill, New Jersey.

QUENOUILLE, M. (1949), "Approximate Tests of Correlation in Time Series," *Journal of The Royal Statistical Society*, Ser. B, 11, 18–84.

SHUCANY, W.; BRAY, H.; and OWEN, O. (1971), "On Bias Reduction in Estimation," *Journal of the American Statistical Association*, 66, 524–533.

STONE, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.