Bootstrap Methods: Another Look at the Jackknife

B. Efron

Annals of Statistics, Volume 7, Issue 1 (Jan., 1979), 1-26.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28197901%297%3A1%3C1%3ABMALAT%3E2.0.CO%3B2-6

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Annals of Statistics is published by Institute of Mathematical Statistics. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ims.html.

Annals of Statistics
©1979 Institute of Mathematical Statistics

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2002 JSTOR

http://www.jstor.org/
Sat Sep 28 17:52:33 2002

The Annals of Statistics
1979, Vol. 7, No. 1, 1–26

THE 1977 RIETZ LECTURE

BOOTSTRAP METHODS: ANOTHER LOOK AT THE JACKKNIFE

BY B. EFRON

Stanford University

We discuss the following problem: given a random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from an unknown probability distribution F, estimate the sampling distribution of some prespecified random variable $R(\mathbf{X}, F)$, on the basis of the observed data $\mathbf{x}$. (Standard jackknife theory gives an approximate mean and variance in the case $R(\mathbf{X}, F) = \theta(\hat{F}) - \theta(F)$, θ some parameter of interest.) A general method, called the "bootstrap," is introduced, and shown to work satisfactorily on a variety of estimation problems. The jackknife is shown to be a linear approximation method for the bootstrap. The exposition proceeds by a series of examples: variance of the sample median, error rates in a linear discriminant analysis, ratio estimation, estimating regression parameters, etc.

1. Introduction. The Quenouille–Tukey jackknife is an intriguing nonparametric method for estimating the bias and variance of a statistic of interest, and also for testing the null hypothesis that the distribution of a statistic is centered at some prespecified point. Miller [14] gives an excellent review of the subject.

This article attempts to explain the jackknife in terms of a more primitive method, named the "bootstrap" for reasons which will become obvious. In principle, bootstrap methods are more widely applicable than the jackknife, and also more dependable. In Section 3, for example, the bootstrap is shown to (asymptotically) correctly estimate the variance of the sample median, a case where the jackknife is known to fail. Section 4 shows the bootstrap doing well at estimating the error rates in a linear discrimination problem, outperforming "cross-validation," another nonparametric estimation method.

We will show that the jackknife can be thought of as a linear expansion method (i.e., a "delta method") for approximating the bootstrap. This helps clarify the theoretical basis of the jackknife, and suggests improvements and variations likely to be successful in various special situations. Section 3, for example, discusses jackknifing (or bootstrapping) when one is willing to assume symmetry or smoothness of the underlying probability distribution. This point reappears more emphatically in Section 7, which discusses bootstrap and jackknife methods for regression models.

The paper proceeds by a series of examples, with little offered in the way of general theory. Most of the examples concern estimation problems, except for Remark F of Section 8, which discusses Tukey's original idea for t-testing using the

Received June 1977; revised December 1977.

AMS 1970 *subject classifications.* Primary 62G05, 62G15; Secondary 62H30, 62J05.

Key words and phrases. Jackknife, bootstrap, resampling, subsample values, nonparametric variance estimation, error rate estimation, discriminant analysis, nonlinear regression.

jackknife. The bootstrap results on this point are mixed (and won't be reported here), offering only slight encouragement for the usual jackknife t tests.

John Hartigan, in an important series of papers [5, 6, 7], has explored ideas closely related to what is called bootstrap "Method 2" in the next section, see Remark I of Section 8. Maritz and Jarrett [13] have independently used bootstrap "Method 1" for estimating the variance of the sample median, deriving equation (3.4) of this paper and applying it to the variance calculation. Bootstrap "Method 3," the connection to the jackknife via linear expansions, relates closely to Jaeckel's work on the infinitesimal jackknife [10]. If we work in a parametric framework, this approach to the bootstrap gives Fisher's information bound for the asymptotic variance of the maximum likelihood estimator, see Remark K of Section 8.

2. Bootstrap methods. We discuss first the one-sample situation in which a random sample of size n is observed from a completely unspecified probability distribution F,

$$X_i = x_i, \qquad X_i \sim_{\text{ind}} F \qquad\qquad i = 1, 2, \ldots, n. \tag{2.1}$$

In all of our examples F will be a distribution on either the real line or the plane, but that plays no role in the theory. We let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ denote the random sample and its observed realization, respectively.

The problem we wish to solve is the following. Given a specified random variable $R(\mathbf{X}, F)$, possibly depending on both $\mathbf{X}$ and the unknown distribution F, estimate the sampling distribution of R on the basis of the observed data $\mathbf{x}$.

Traditional jackknife theory focuses on two particular choices of R. Let $\theta(F)$ be some parameter of interest such as the mean, correlation, or standard deviation of F, and $t(\mathbf{X})$ be an estimator of $\theta(F)$, such as the sample mean, sample correlation, or a multiple of the sample range. Then the sampling distribution of

$$R(\mathbf{X}, F) = t(\mathbf{X}) - \theta(F), \tag{2.2}$$

or more exactly its mean (the bias of t) and variance, is estimated using the standard jackknife theory, as described in Section 5. The bias and variance estimates say $\widehat{\text{Bias}}\,(t)$ and $\widehat{\text{Var}}\,(t)$, are cleverly constructed functions of $\mathbf{X}$ obtained by recomputing $t(\cdot)$ n times, each time removing one component of $\mathbf{X}$ from consideration. The second traditional choice of R is

$$R(\mathbf{X}, F) = \frac{t(\mathbf{X}) - \widehat{\text{Bias}}\,(t) - \theta(F)}{(\widehat{\text{Var}}\,(t))^{\frac{1}{2}}}. \tag{2.3}$$

Tukey's original suggestion was to treat (2.3) as having a standard Student's t distribution with $n - 1$ degrees of freedom. (See Remark F, Section 8.) Random variables (2.2), (2.3) play no special role in the bootstrap theory, and, as a matter of fact, some of our examples concern other choices of R.

The bootstrap method for the one-sample problem is extremely simple, at least in principle:

1. Construct the sample probability distribution $\hat{F}$, putting mass $1/n$ at each point $x_1, x_2, x_3, \ldots, x_n$.

2. With $\hat{F}$ fixed, draw a random sample of size n from $\hat{F}$, say

$$X_i^* = x_i^*, X_i^* \sim_{\text{ind}} \hat{F} \qquad i = 1, 2, \cdots, n. \tag{2.4}$$

Call this the *bootstrap sample*, $\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_n^*)$, $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$. Notice that we are not getting a permutation distribution since the values of $\mathbf{X}^*$ are selected *with* replacement from the set $\{x_1, x_2, \ldots, x_n\}$. As a point of comparison, the ordinary jackknife can be thought of as drawing samples of size $n - 1$ *without* replacement.

3. Approximate the sampling distribution of $R(\mathbf{X}, F)$ by the *bootstrap distribution* of

$$R^* = R(\mathbf{X}^*, \hat{F}), \tag{2.5}$$

i.e., the distribution of R^* induced by the random mechanism (2.4), with $\hat{F}$ held fixed at its observed value.

The point is that the distribution of R^*, which in theory can be calculated exactly once the data $\mathbf{x}$ is observed, equals the desired distribution of R if $F = \hat{F}$. Any nonparametric estimator of R's distribution, i.e., one that does a reasonably good estimation job without prior restrictions on the form of F, must give close to the right answer when $F = \hat{F}$, since $\hat{F}$ is a central point amongst the class of likely F's, having observed $\mathbf{X} = \mathbf{x}$. Making the answer exactly right for $F = \hat{F}$ is *Fisher consistency* applied to our particular estimation problem.

Just how well the distribution of R^* approximates that of R depends upon the form of R. For example, $R(\mathbf{X}, F) = t(\mathbf{X})$ might be expected to bootstrap less successfully than $R(\mathbf{X}, F) = [t(\mathbf{X}) - E_F t]/(\text{Var}_F t)^{\frac{1}{2}}$. This is an important question, related to the concept of pivotal quantities, Barnard [2], but is discussed only briefly here, in Section 8. Mostly we will be content to let the varying degrees of success of the examples speak for themselves.

As the simplest possible example of the bootstrap method, consider a probability distribution F putting all of its mass at zero or one, and let the parameter of interest be $\theta(F) = \text{Prob}_F\{X = 1\}$. The most obvious random variable of interest is

$$R(\mathbf{X}, F) = \bar{X} - \theta(F) \qquad \bar{X} = (\textstyle\sum_{i=1}^n X_i/n). \tag{2.6}$$

Having observed $\mathbf{X} = \mathbf{x}$, the bootstrap sample $\mathbf{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$ has each component independently equal to one with probability $\bar{x} = \theta(\hat{F})$, zero with probability $1 - \bar{x}$. Standard binomial results show that

$$R^* = R(\mathbf{X}^*, \hat{F}) = \bar{X}^* - \bar{x} \tag{2.7}$$

has mean and variance

$$E_*(\bar{X}^* - \bar{x}) = 0, \qquad \text{Var}_*(\bar{X}^* - \bar{x}) = \bar{x}(1 - \bar{x})/n. \tag{2.8}$$

(Notations such as "E_*," "Var_*," "Prob_*," etc. indicate probability calculations relating to the bootstrap distribution of $\mathbf{X}^*$, with $\mathbf{x}$ and $\hat{F}$ fixed.) The implication that $\bar{X}$ is unbiased for θ, with variance approximately equal to $\bar{x}(1-\bar{x})/n$, is universally familiar.

As a second example, consider estimating $\theta(F) = \mathrm{Var}_F X$, the variance of an arbitrary distribution on the real line, using the estimator $t(\mathbf{X}) = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$. Perhaps we wish to know the sampling distribution of

$$R(\mathbf{X}, F) = t(\mathbf{X}) - \theta(F). \tag{2.9}$$

Let $\mu_k(F)$ indicate the kth central moment of F, $\mu_k(F) = E_F(X - E_F X)^k$, and $\hat{\mu}_k = \mu_k(\hat{F})$, the kth central moment of $\hat{F}$. Standard sampling theory results, as in Cramér [3], Section 27.4, show that

$$R^* = R(\mathbf{X}^*, \hat{F}) = t(\mathbf{X}^*) - \theta(\hat{F})$$

has

$$E_* R^* = 0, \qquad \mathrm{Var}_* R^* = \frac{\hat{\mu}_4 - ((n-3)/(n-1))\hat{\mu}_2^2}{n}. \tag{2.10}$$

The approximation $\mathrm{Var}_F t(\mathbf{X}) \approx \mathrm{Var}_* R^*$ is (almost) the jackknife estimate for $\mathrm{Var}_F t$.

The difficult part of the bootstrap procedure is the actual calculation of the bootstrap distribution. Three methods of calculation are possible:

Method 1. Direct theoretical calculation, as in the two examples above and the example of the next section.

Method 2. Monte Carlo approximation to the bootstrap distribution. Repeated realizations of $\mathbf{X}^*$ are generated by taking random samples of size n from $\hat{F}$, say $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*N}$, and the histogram of the corresponding values $R(\mathbf{x}^{*1}, \hat{F}), R(\mathbf{x}^{*2}, \hat{F}), \ldots, R(\mathbf{x}^{*N}, \hat{F})$ is taken as an approximation to the actual bootstrap distribution. This approach is illustrated in Sections 3, 4 and 8.

Method 3. Taylor series expansion methods can be used to obtain the approximate mean and variance of the bootstrap distribution of R^*. This turns out to be the same as using some form of the jackknife, as shown in Section 5.

In Section 4 we consider a two sample problem where the data consists of a random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ from F and an independent random sample $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ from G, F and G arbitrary probability distributions on a given space. In order to estimate the sampling distribution of a random variable $R((\mathbf{X}, \mathbf{Y}), (F, G))$, having observed $\mathbf{X} = \mathbf{x}$, $\mathbf{Y} = \mathbf{y}$, the one-sample bootstrap method can be extended in the obvious way: $\hat{F}$ and $\hat{G}$, the sample probability distributions corresponding to F and G, are constructed; bootstrap samples $X_i^* \sim \hat{F}$, $i = 1, 2, \ldots, m$, $Y_j^* \sim \hat{G}$, $j = 1, 2, \ldots, n$, are independently drawn; and finally the bootstrap distribution of $R^* = R((\mathbf{X}^*, Y^*), (\hat{F}, \hat{G}))$ is calculated, for use as an approximation to the actual distribution of R. The calculation of the bootstrap distribution proceeds by any of the three methods listed above. (The third method

makes clear the correct analogue of the jackknife procedure for nonsymmetric situations, such as the two sample problem; see the remarks of Section 6.)

So far we have only used nonparametric maximum likelihood estimators, $\hat{F}$ and $(\hat{F}, \hat{G})$, to begin the bootstrap procedure. This isn't crucial, and as the examples of Sections 3 and 7 show, it is sometimes more convenient to use other estimates of the underlying distributions.

3. Estimating the median. Suppose we are in the one-sample situation (2.1), with F a distribution on the real line, and we wish to estimate the median of F using the sample median. Let $\theta(F)$ indicate the median of F, and let $t(\mathbf{X})$ be the sample median,

$$t(\mathbf{X}) = X_{(m)}, \tag{3.1}$$

where $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ is the order statistic, and we have assumed an odd sample size $n = 2m - 1$ for convenience. Once again we take $R(\mathbf{X}, F) = t(\mathbf{X}) - \theta(F)$, and hope to say something about the sampling distribution of R on the basis of the observed random sample.

Having observed $\mathbf{X} = \mathbf{x}$, we construct the bootstrap sample $\mathbf{X}^* = \mathbf{x}^*$ as in (2.4). Let

$$N_i^* = \#\{X_i^* = x_i\}, \tag{3.2}$$

the number of times x_i is selected in the bootstrap sampling procedure. The vector $\mathbf{N}^* = (N_1^*, N_2^*, \cdots, N_n^*)$ has a multinomial distribution with expectation one in each of the n cells.

Denote the observed order statistic $x_{(1)} \leqslant x_{(2)} \leqslant x_{(3)} \leqslant \cdots \leqslant x_{(n)}$, and the corresponding N^* values $N_{(1)}^*, N_{(2)}^*, \cdots, N_{(n)}^*$. (Ties $x_i = x_{i'}$ can be broken by assigning the lower value of i, i' to the lower position in the order statistic.) The bootstrap value of R is

$$R^* = R(\mathbf{X}^*, \hat{F}) = X_{(m)}^* - x_{(m)}. \tag{3.3}$$

We notice that for any integer value l, $1 \leqslant l < n$,

$$\begin{aligned} \text{(3.4)} \quad \text{Prob}_*\{X_{(m)}^* > x_{(l)}\} &= \text{Prob}_*\{N_{(1)}^* + N_{(2)}^* + \cdots + N_{(l)}^* \leqslant m - 1\} \\ &= \text{Prob}\left\{\text{Binomial}\left(n, \frac{l}{n}\right) \leqslant m - 1\right\} \\ &= \Sigma_{j=0}^{m-1}\binom{n}{j}\left(\frac{l}{n}\right)^j\left(\frac{n-l}{n}\right)^{n-j}. \end{aligned}$$

Therefore

$$\begin{aligned} \text{(3.5)} \quad \text{Prob}_*\{R^* = x_{(l)} - x_{(m)}\} &= \text{Prob}\left\{\text{Binomial}\left(n, \frac{l-1}{n}\right) \leqslant m - 1\right\} \\ &\quad - \text{Prob}\left\{\text{Binomial}\left(n, \frac{l}{n}\right) \leqslant m - 1\right\}, \end{aligned}$$

a result derived independently by Maritz and Jarrett [13].

The case $n = 13$ ($m = 7$) gives the following bootstrap distribution for R^*:

(3.6)

$l =$	2 or 12	3 or 11	4 or 10	5 or 9	6 or 8	7
(3.5) =	.0015	.0142	.0550	.1242	.1936	.2230

For any given random sample of size 13 we can compute

$$E_*(R^*)^2 = \Sigma_{l=1}^{13}\big[x_{(l)} - x_{(7)}\big]^2 \operatorname{Prob}_*\{R^* = x_{(l)} - x_{(7)}\}, \tag{3.7}$$

and use this number as an estimate of $E_F R^2 = E_F[t(\mathbf{X}) - \theta(F)]^2$, the expected squared error of estimation for the sample median. Standard asymptotic theory, applied to the case where F has a bounded continuous density $f(x)$, shows that as the sample size n goes to infinity, the quantity $nE_*(R^*)^2$ approaches $1/4f^2(\theta)$, where $f(\theta)$ is the density evaluated at the median $\theta(F)$. This is the correct asymptotic value, see Kendall and Stuart [11], page 237. The standard jackknife applied to the sample median gives a variance estimate which is not even asymptotically consistent (Miller [14], page 8, is incorrect on this point): $n\,\widehat{\operatorname{Var}}(R) \to (1/4f^2(\theta))[\chi_2^2/2]^2$. The random variable $[\chi_2^2/2]^2$ has mean 2 and variance 20.

Suppose we happened to know that the probability distribution F was symmetric. In that case we could replace $\hat{F}$ by the symmetric probability distribution obtained from $\hat{F}$ by reflection about the median,

$$\hat{F}_{\mathrm{SYM}}\text{: probability mass } \frac{1}{2n-1} \text{ at } x_{(1)}, x_{(2)}, \ldots, x_{(n)} \text{ and } 2x_{(m)} - x_{(1)}, \ldots, 2x_{(m)} - x_{(n)}. \tag{3.8}$$

This is not the nonparametric maximum likelihood estimator for F, but has similar asymptotic properties, see Hinkley [8]. Let $z_{(1)} \leqslant z_{(2)} \leqslant \cdots \leqslant z_{(2n-1)}$ be the ordered values appearing in the distribution of $\hat{F}_{\mathrm{SYM}}$. The bootstrap procedure starting from $\hat{F}_{SYM}$ gives

$$\operatorname{Prob}_*\{R^* = z_{(l)} - x_{(m)}\} = \operatorname{Prob}\left\{\operatorname{Binomial}\left(n, \frac{l-1}{2n-1}\right) \leqslant m-1\right\} - \operatorname{Prob}\left\{\operatorname{Binomial}\left(n, \frac{l}{2n-1}\right) \leqslant m-1\right\}, \tag{3.9}$$

by the same argument leading to (3.5). For $n = 13$ the bootstrap probabilities (3.9) equal

(3.10)

$l =$	4 or 22	5 or 21	6 or 20	7 or 19	8 or 18
(3.9) =	.0016	.0051	.0125	.0245	.0414
$l =$	9 or 17	10 or 16	11 or 15	12 or 14	13
(3.9) =	.0614	.0820	.1002	.1125	.1170

The corresponding estimate of $E_F R^2$ would be $\Sigma_{l=1}^{25}[z_{(l)} - x_{(7)}]^2 \operatorname{Prob}_*\{R^* = z_{(l)} - x_{(7)}\}$.

Usually we would not be willing to assume F symmetric in a nonparametric estimation situation. However in dealing with continuous variables we might be

willing to attribute a moderate amount of smoothness to F. This can be incorporated into the bootstrap procedure at step (2.4). Instead of choosing each X_i^* randomly from the set $\{x_1, x_2, \ldots, x_n\}$, we can take

$$X_i^* = \bar{x} + c\left[x_{I_i} - \bar{x} + \hat{\sigma} Z_i\right] \tag{3.11}$$

where the I_i are chosen independently and randomly from the set $\{1, 2, \ldots, n\}$, and the Z_i are a random sample from some fixed distribution having mean 0 and variance σ_Z^2, for example the uniform distribution on $[-\frac{1}{2}, \frac{1}{2}]$, which has $\sigma_Z^2 = 1/12$. The most obvious choice is a normal distribution for the Z_i, but this would be self-serving in the Monte Carlo experiment which follows, where the X_i themselves are normally distributed. The quantities $\bar{x}$, $\hat{\sigma}$, and c appearing in (3.11) are the sample mean, sample standard deviation $(= (\hat{\mu}_2)^{\frac{1}{2}})$, and $[1 + \sigma_Z^2]^{-\frac{1}{2}}$, respectively, so that X_i^* has mean $\bar{x}$ and variance $\hat{\sigma}^2$ under the bootstrap sampling procedure. In using (3.11) in place of (2.4), we are replacing $\hat{F}$ with a smoothed "window" estimator, having the same mean and variance as $\hat{F}$.

A small Monte Carlo experiment was run to compare the various bootstrap methods suggested above. Instead of comparing the squared error of the sample median, the quantity bootstrapped was

$$R(\mathbf{X}, F) = \frac{|t(\mathbf{X}) - \theta(F)|}{\sigma(F)}, \tag{3.12}$$

the absolute error of the sample median relative to the population standard deviation. (This quantity is more stable numerically, because the absolute value is less sensitive than the square and also because $R^* = |t(\mathbf{X}^*) - \theta(\hat{F})|/\hat{\sigma}$ is scale invariant, which eliminates the component of variation due to $\hat{\sigma}$ differing from $\sigma(F)$. The stability of (3.12) greatly increased the effectiveness of the Monte Carlo trial.)

The Monte Carlo experiment was run with $n = 13$, $X_i \sim_{\text{ind}} \mathcal{N}(0, 1)$, $i = 1, 2, \ldots, n$. In this situation the true expectation of R is

$$E_F R = 0.95. \tag{3.13}$$

The first two columns of Table 1 show $E_F R^*$ for each trial, using the bootstrap probabilities (3.6), and then (3.10) for the symmetrized version. It is not possible to theoretically calculate $E_* R^*$ for the smoothed bootstrap (3.11), so these entries of Table 1 were obtained by a secondary Monte Carlo simulation, as described in "Method 2" of Section 2. A total of $N = 50$ replications $\mathbf{x}^{*j}$ were generated for each trial. This means that the values in the table are only unbiased estimates of the actual bootstrap expectations $E_* R^*$ (which could be obtained by letting $N \to \infty$); the standard error being about .15 for each entry. The effect of this approximation is seen in the column "$d = 0$," which would exactly equal column "(3.6)" if $N \to \infty$. (Within each trial, the same set of random numbers was used to generate the four different uniform distributions for Z_i, $d = 0, .25, .5, 1$.)

TABLE 1*

	Unsmoothed Bootstrap		Smoothed Bootstrap (3.11)				
			Z_i uniform dist. on $[-d/2, d/2]$				Z_i triangular
Trial #	(3.6)	(3.10)	$d = 0$	$d = .25$	$d = .5$	$d = 1$	dist., $\sigma_Z^2 = 1/12$
1	1.07	1.18	1.09	1.10	1.12	1.11	1.16
2	.96	.74	1.10	1.10	1.08	1.09	1.15
3	1.22	.74	1.36	1.35	1.33	1.43	1.52
4	1.38	1.51	1.44	1.41	1.38	1.28	1.30
5	1.00	.83	1.03	1.05	1.09	1.14	1.17
6	1.13	1.21	1.27	1.26	1.23	1.20	1.26
7	1.07	.98	1.01	.94	.83	.79	.92
8	1.51	1.40	1.40	1.45	1.47	1.51	1.50
9	.56	.64	.69	.71	.74	.80	.81
10	1.05	.86	1.14	1.17	1.20	1.13	1.22
Ave.	1.09	1.01	1.15	1.15	1.15	1.15	1.20
S.D.	.26	.30	.23	.23	.23	.23	.22

*Ten Monte Carlo trials of $X_i \sim_{\text{ind}} \mathcal{U}(0, 1)$, $i = 1, 2, \ldots, 13$ were used to compare different bootstrap methods for estimating the expected value of random variable (3.12). The true expectation is 0.95. The quantities tabled are $E_* R^*$, the bootstrap expectation for that trial. The values in the first two columns are for the bootstrap as described originally, and for the symmetrized version (3.8)–(3.10). The smoothed bootstrap expectations were approximated using a secondary Monte Carlo simulation for each trial, $N = 50$, as described in "Method 2," Section 2. Each of these entries estimates the actual value of $E_* R^*$ unbiasedly with a standard error of about .15. The column "$d = 0$" would exactly equal column "(3.6)" if $N \to \infty$.

The most notable feature of Table 1 is that the simplest form of the bootstrap, "(3.6)," seems to do just as well as the symmetrical or smoothed versions. A larger Monte Carlo investigation of the same situation as in Table 1, 200 trials, 100 bootstrap replications per trial, was just a little more favorable to the smoothed bootstrap methods:

	(3.6)	(3.10)	$d = 0$	$d = .25$	$d = .5$	$d = 1$	$d = 2$
AVE.:	1.01	1.00	1.00	1.01	1.00	.99	.93
S.D.:	.31	.33	.32 [.31]	.32 [.30]	.32 [.30]	.30 [.29]	.26 [.25].

(The figures in square brackets are estimated standard deviations if N were increased from 100 to ∞, obtained by a components of variance calculation.) Remembering that we are trying to estimate the true value $E_F R = .95$, these seem like good performances for a nonparametric method based on a sample size of just 13.

The symmetrized version of the bootstrap might be expected to do relatively better than the unsymmetrized version if R itself was of a less symmetric form than (3.12), e.g., $R(\mathbf{X}, F) = \exp\{X_{(m)} - \theta(F)\}$. Likewise, the smoothed versions of the bootstrap might be expected to do relatively better if R itself were less smooth, e.g., $R(\mathbf{X}, F) = \text{Prob}\{X_{(m)} > \theta(F) + \sigma(F)\}$. However no evidence to support these guesses is available at present.

4. Error rate estimation in discriminant analysis. This section discusses the estimation of error rates in a standard linear discriminant analysis problem. There is a tremendous literature on this problem, nicely summarized in Toussaint [17]. In the two examples considered below, bootstrap methods outperform the commonly used "leave-one-out," or *cross-validation*, approach (Lachenbruch and Mickey [12]).

The data in the discriminant problem consists of independent random samples from two unknown continuous probability distributions F and G on some k-dimensional space R^k,

$$\begin{aligned} X_i &= x_i, \quad X_i \sim_{\text{ind}} F \qquad i = 1, 2, \ldots, m \\ Y_j &= y_j, \quad Y_j \sim_{\text{ind}} G \qquad j = 1, 2, \ldots, n. \end{aligned} \tag{4.1}$$

On the basis of the observed data $\mathbf{X} = \mathbf{x}$, $\mathbf{Y} = \mathbf{y}$ we use some method (linear discriminant analysis in the examples below) to partition R^k into two complementary regions A and B, the intent being to ascribe a future observation z to the F distribution if $z \in A$, or to the G distribution if $z \in B$.

The obvious estimate of the error rate, for the F distribution, associated with the partition (A, B) is

$$\widehat{\text{error}}_F = \frac{\#\{x_i \in B\}}{m}, \tag{4.2}$$

which will tend to underestimate the true error rate

$$\text{error}_F = \text{Prob}_F\{X \in B\}. \tag{4.3}$$

(In probability calculation (4.3), B is considered fixed, at its observed value, even though it is originally determined by a random mechanism.) We will be interested in the distribution of the difference

$$R((\mathbf{X}, \mathbf{Y}), (F, G)) = \text{error}_F - \widehat{\text{error}}_F, \tag{4.4}$$

and the corresponding quantity for the distribution G. We could directly consider the distribution of $\widehat{\text{error}}_F$, but concentrating on the difference (4.4) is much more efficient for comparing different estimation methods. This point is discussed briefly at the end of the section.

Given $\mathbf{x}$ and $\mathbf{y}$, we define the region B by

$$B = \left\{ z : (\bar{y} - \bar{x})' S^{-1} \left(z - \frac{\bar{x} + \bar{y}}{2} \right) > \log \frac{m}{n} \right\}, \tag{4.5}$$

where $\bar{x} = \Sigma x_i/m$, $\bar{y} = \Sigma y_j/n$, and $S = [\Sigma(x_i - \bar{x})(x_i - \bar{x})' + \Sigma(y_j - \bar{y})(y_j - \bar{y})']/(m + n)$. This is the maximum likelihood estimate of the optimum division under multivariate normal theory, and differs just slightly (in the definition of S) from the estimated version of the Fisher linear discriminant function discussed in Chapter 6 of Anderson [1].

"Method 2," the brute force application of the bootstrap via simulation, is implemented as follows: given the data **x**, **y**, bootstrap random samples

$$X_i^* = x_i^*, \quad X_i^* \sim_{\text{ind}} \hat{F} \qquad i = 1, 2, \ldots, m \tag{4.6}$$
$$Y_j^* = y_j^*, \quad Y_j^* \sim_{\text{ind}} \hat{G} \qquad j = 1, 2, \ldots, n$$

are generated, $\hat{F}$ and $\hat{G}$ being the sample probability distributions corresponding to F and G. This yields a region B^* defined by (4.5) with $\bar{x}^*, \bar{y}^*, S^*$ replacing $\bar{x}, \bar{y}, S$. The bootstrap random variable in this case is

$$R^* = R((\mathbf{X}^*, \mathbf{Y}^*), (\hat{F}, \hat{G})) = \frac{\#\{x_i \in B^*\}}{m} - \frac{\#\{x_i^* \in B^*\}}{m}. \tag{4.7}$$

In other words, (4.7) is the difference between the actual error rate, actual now being defined with respect to the "true" distribution $\hat{F}$, and the apparent error rate obtained by counting errors in the bootstrap sample.

Repeated independent generation of $(\mathbf{X}^*, \mathbf{Y}^*)$ yields a sequence of independent realizations of R^*, say $R^{*1}, R^{*2}, \ldots, R^{*N}$, which are then used to approximate the actual bootstrap distribution of R^*, this hopefully being a reasonable estimate of the unknown distribution of R. For example, the bootstrap expectation $E_* R^* = \sum_{j=1}^{N} R^{*j}/N$ can be used as an estimate of the true expectation $E_{F,G} R$.

To test out this theory, bivariate normal choices of F and G were investigated,

$$F\colon X \sim \mathcal{N}_2\left(\begin{pmatrix} -\frac{1}{2} \\ 0 \end{pmatrix}, \mathbf{I}\right) \quad G\colon Y \sim \mathcal{N}_2\left(\begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}, \mathbf{I}\right). \tag{4.8}$$

Two sets of sample sizes, $m = n = 10$ and $m = n = 20$, were looked at, with the results shown in Table 2. (The entries of Table 2 were themselves estimated by averaging over repeated Monte Carlo trials, which should not be confused with the

TABLE 2*

	$m = n = 10$			$m = n = 20$			
Random Variable	Mean	(S.E.)	S.D.	Mean	(S.E.)	S.D.	Remarks
Error Rate Diff. (4.4) R	**.062**	(.003)	**.143**	**.028**	(.002)	**.103**	Based on 1000 trials
Bootstrap Expectation $E_* R^*$	**.057**	(.002)	**.026**	**.029**	(.001)	**.015**	Based on 100 trials; $N = 100$ Bootstrap
			[.023]			[.011]	Replications per trial. (Figure in
Bootstrap Standard Deviation $SD_*(R^*)$	**.131**	(.0013)	**.016**	**.097**	(.002)	**.010**	brackets is S.D. if $N = \infty$.)
Cross–Validation Diff. $\tilde{R}$	**.054**	(.009)	**.078**	**.032**	(.002)	**.043**	Based on 40 trials

* The error rate difference (4.4) for linear discriminant analysis, investigated for bivariate normal samples (4.8). Sample sizes are $m = n = 10$ and $m = n = 20$. The values for the bootstrap method were obtained by Method 2, $N = 100$ bootstrap replications per trial. The bootstrap method gives useful estimates of both the mean and standard deviation of R. The cross-validation method was nearly unbiased for the expectation of R, but had about three times as large a standard deviation. All of the quantities in this table were estimated by repeated Monte Carlo trials; standard errors are given for the means.

Monte Carlo replications used in the bootstrap process. "Replications" will always refer to the bootstrap process, "trials" to repetitions of the basic situation.) Because situation (4.8) is symmetric, only random variable (4.4), and not the corresponding error rate for G, need be considered.

Table 2 shows that with $m = n = 10$, the random variable (4.4) has mean and standard deviation approximately (.062, .143). The apparent error rate underestimates the true error rate by about 6%, on the average, but the standard deviation of the difference is 14% from trial to trial, so bias is less troublesome than variability in this situation. The bootstrap method gave an average of .057 for E_*R^*, which, allowing for sampling error, shows that the statistic E_*R^* is nearly an unbiased estimator for $E_{F,G}R$. Unbiasedness is not enough, of course; we want E_*R^* to have a small standard deviation, ideally zero, so that we can rely on it as an estimate. The actual value of its standard deviation, .026, is not wonderful, but does indicate that most of the trials yielded E_*R^* in the range [.02, .09], which means that the statistician would have obtained a reasonably informative estimate of the true bias $E_{F,G}R = .062$.

As a point of comparison, consider the cross-validation estimate of R, say $\tilde{R}$, obtained by: deleting one x value at a time from the vector $\mathbf{x}$; recomputing B using (4.5), to get a new region $\tilde{B}$ (it is important *not* to change m to $m - 1$ in recomputing B—doing so results in badly biased estimation of R); seeing if the deleted x value is correctly classified by $\tilde{B}$; counting the proportion of x values misclassified in this way to get a cross-validated error rate $\widetilde{\text{error}}_F$; and finally, defining $\tilde{R} = \widetilde{\text{error}}_F - \widehat{\text{error}}_F$. The last row of Table 2 shows that $\tilde{R}$ has mean and standard deviation approximately (.054, .078). That is, *$\tilde{R}$ is three times as variable as E_*R^* as an estimator of $E_{F,G}R$.*

The bootstrap standard deviation of R^*, $SD_*(R^*) = \{\sum_{j=1}^{N}[R^{*j} - E_*R^*]^2/(N - 1)\}^{\frac{1}{2}}$, can be used as an estimate of $SD_{F,G}(R)$, the actual standard deviation of R. Table 2 shows that $SD_*(R^*)$ had mean and standard deviation (.131, .016) across the 100 trials. Remembering that $SD_{F,G}(R) = .143$, the bootstrap estimate $SD_*(R^*)$ is seen to be a quite useful estimator of the actual standard deviation of R.

How much better would the bootstrap estimator E_*R^* perform if the number of bootstrap replications N were increased from 100 to, say, 10,000? A components of variance analysis of all the data going into Table 2 showed that only moderate further improvement is possible. As $N \to \infty$, the trial-to-trial standard deviation of E_*R^* would decrease from .026 to about .023 (from .015 to .011 in the case $m = n = 20$).

The reader may wonder which is the best estimator of the error rate error_F itself, rather than of the difference R. In terms of expected squared error, the order of preference is $\widehat{\text{error}}_F + E_*R^*$ (the bias-corrected value based on the bootstrap), $\widehat{\text{error}}_F$, and lastly $\widetilde{\text{error}}_F$, but the differences are quite small in the two situations of Table 2. The large variability of $\widehat{\text{error}}_F$, compared to its relatively small bias, makes

bias correction an almost fruitless chore in these two situations. (Of course, this might not be so in more difficult discriminant problems.) *The bootstrap estimates of* $E_{F,G}R$ *and* $SD_{F,G}(R)$ *considered together make it clear that this is the case,* which is a good recommendation for the bootstrap approach.

5. Relationship with the jackknife. This section concerns "Method 3" of approximating the bootstrap distribution, Taylor series expansion (or the *delta method*), which turns out to be the same as the usual jackknife theory. To be precise, it is the same as Jaeckel's *infinitesimal jackknife* [10, 14], a useful mathematical device which differs only in detail from the standard jackknife. Many of the calculations below, and in Remarks G—K of Section 8, can be found in Jaeckel's excellent paper, which offers considerable insight into the workings of jackknife methods.

Returning to the one-sample situation, define $P_i^* = N_i^*/n$, where $N_i^* = \#\{X_i^* = x_i\}$ as at (3.2), and the corresponding vector

$$\mathbf{P}^* = (P_1^*, P_2^*, \cdots, P_n^*). \tag{5.1}$$

By the properties of the multinomial distribution, $\mathbf{P}^*$ has mean vector and covariance matrix

$$E_*\mathbf{P}^* = \mathbf{e}/n, \qquad \mathrm{Cov}_*\mathbf{P}^* = \mathbf{I}/n^2 - \mathbf{e}'\mathbf{e}/n^3 \tag{5.2}$$

under the bootstrap sampling procedure, where $\mathbf{I}$ is the identity matrix and $\mathbf{e} = (1, 1, 1, \ldots, 1)$.

Given the observed data vector $\mathbf{X} = \mathbf{x}$, and therefore $\hat{F}$, we can use the abbreviated notation

$$R(\mathbf{P}^*) = R(\mathbf{X}^*, \hat{F}) \tag{5.3}$$

for the bootstrap realization of R corresponding to $\mathbf{P}^*$. In making this definition we assume that the random variable of interest, $R(\mathbf{X}, F)$, is symmetrically defined in the sense that its value is invariant under any permutation of the components of X, so that it is sufficient to know $\mathbf{N}^* = n\mathbf{P}^*$ in order to evaluate $R(\mathbf{X}^*, \hat{F})$. This is always the case in standard applications of the jackknife.

We can approximate the bootstrap distribution of $R(\mathbf{X}^*, \hat{F})$ by expanding $R(\mathbf{P}^*)$ in a Taylor series about the value $\mathbf{P}^* = \mathbf{e}/n$, say

$$R(\mathbf{P}^*) \doteq R(\mathbf{e}/n) + (\mathbf{P}^* - \mathbf{e}/n)\mathbf{U} + \tfrac{1}{2}(\mathbf{P}^* - \mathbf{e}/n)\mathbf{V}(\mathbf{P}^* - \mathbf{e}/n)'. \tag{5.4}$$

Here

$$\mathbf{U} = \left[\begin{array}{c} \vdots \\ \dfrac{\partial R(\mathbf{P}^*)}{\partial P_i^*} \\ \vdots \end{array}\right]_{\mathbf{P}^* = \mathbf{e}/n} \qquad \mathbf{V} = \left[\begin{array}{ccc} \vdots & & \\ \cdots & \dfrac{\partial^2 R(\mathbf{P}^*)}{\partial P_i^* \partial P_j^*} & \cdots \\ \vdots & & \end{array}\right]_{\mathbf{P}^* = \mathbf{e}/n}. \tag{5.5}$$

Expansion (5.4), and definitions (5.5), assume that the definition of $R(\mathbf{P}^*)$ can be smoothly interpolated between the lattice point values originally contemplated for $\mathbf{P}^*$. How to do so will be obvious in most specific cases, but a general recipe is difficult to provide. See Remarks G and H of Section 8.

The restriction $\sum P_i^* = 1$ has been ignored in (5.4), (5.5). This computational convenience is justified by extending the definition of $R(\mathbf{P}^*)$ to all vectors $\mathbf{P}^*$ having nonnegative components, at least one positive, by the homogeneous extension

$$R(\mathbf{P}^*) = R\left(\frac{\mathbf{P}^*}{\sum_{i=1}^n P_i^*}\right). \tag{5.6}$$

It is easily shown that the homogeneity of definition (5.6) implies

$$\mathbf{eU} = 0, \qquad \mathbf{eV} = -n\mathbf{U}', \qquad \mathbf{eVe}' = 0. \tag{5.7}$$

From (5.2) and (5.4) we get the approximation to the bootstrap expectation

$$E_* R(\mathbf{P}^*) \doteq R(\mathbf{e}/n) + \frac{1}{2}\operatorname{trace} \mathbf{V}\left[\mathbf{I}/n^2 - \mathbf{e}'\mathbf{e}/n^3\right] = R(\mathbf{e}/n) + \frac{1}{2n}\bar{V}, \tag{5.8}$$

where

$$\bar{V} = \sum_{i=1}^n V_{ii}/n. \tag{5.9}$$

Ignoring the last term in (5.4) gives a cruder approximation for the bootstrap variance,

$$\operatorname{Var}_* R(\mathbf{P}^*) \doteq \mathbf{U}'\left[\mathbf{I}/n^2 - \mathbf{e}'\mathbf{e}/n^3\right]\mathbf{U} = \sum_{i=1}^n U_i^2/n^2. \tag{5.10}$$

(Both (5.8) and (5.10) involve the use of (5.7).)

Results (5.8) and (5.10) are essentially the jackknife expressions for bias and variance. The usual jackknife theory considers $R(\mathbf{X}, F) = \theta(\hat{F}) - \theta(F)$, the difference between the obvious nonparametric estimator of some parameter $\theta(F)$ and $\theta(F)$ itself. In this case $R(\mathbf{X}^*, F) = \theta(\hat{F}^*) - \theta(\hat{F})$, $\hat{F}^*$ being the empirical distribution of the bootstrap sample, so that $R(\mathbf{e}/n) = \theta(\hat{F}) - \theta(\hat{F}) = 0$. Then (5.8) becomes $E_*[\theta(\hat{F}^*) - \theta(\hat{F})] \doteq (1/2n)\bar{V}$, suggesting $E_F[\theta(\hat{F}) - \theta(F)] \approx (1/2n)\bar{V}$; likewise (5.10) becomes $\operatorname{Var}_*[\theta(\hat{F}^*) - \theta(\hat{F})] \doteq \sum U_i^2/n^2$, suggesting $\operatorname{Var}_F \theta(\hat{F}) \approx \sum U_i^2/n^2$.

The approximations

$$\operatorname{Bias}_F \theta(\hat{F}) \approx \frac{1}{2n}\bar{V}, \qquad \operatorname{Var}_F \theta(\hat{F}) \approx \sum_{i=1}^n U_i^2/n^2 \tag{5.11}$$

exactly agree with those given by Jaeckel's infinitesimal jackknife [10], which themselves differ only slightly from the ordinary jackknife expressions. Without going into details, which are given in Jaeckel [10] and Miller [14], the ordinary jackknife replaces the derivatives $U_i = \partial R(\mathbf{P}^*)/\partial P_i$ with finite differences

$$\tilde{U}_i = (n-1)(R^* - R^*_{(i)}) \tag{5.12}$$

where $R^*_{(i)} = R(\mathbf{e}_{(i)}/(n-1))$, $\mathbf{e}_{(i)}$ being the vector with zero in the ith coordinate and ones elsewhere, and $R^*_{\cdot} = \sum_{i=1}^n R^*_{(i)}/n$. Expansion (5.4) combines with (5.7) to give

$$\tilde{U}_i \doteq \frac{n-2}{n-1} U_i - \frac{1}{2(n-1)}(V_{ii} - \bar{V}), \tag{5.13}$$

so that $\tilde{U}_i / U_i = 1 + O(1/n)$. The ordinary jackknife estimate of variance is $\sum_{i=1}^n \tilde{U}_i^2/n \cdot (n-1)$, differing from the variance expression in (5.11) by a factor $1 + O(1/n)$, the same statement being true for the bias. (In the familiar case $R = \theta(\hat{F}) - \theta(F)$, definition (5.12) becomes $\tilde{U}_i = (n-1)(\hat{\theta} - \hat{\theta}_{(i)})$, where $\hat{\theta}_{(i)}$ is the estimate of θ with x_i removed from the sample, and $\hat{\theta}_{\cdot} = \sum_i \hat{\theta}_{(i)}/n$; the jackknife estimate of θ is $\tilde{\theta} = \hat{\theta} + (n-1)(\hat{\theta} - \hat{\theta}_{\cdot})$, and $\tilde{\theta}_i = \tilde{\theta} + \tilde{U}_i$ is the ith *pseudo-value*, to use the standard terminology.)

As an example of Method 3, consider *ratio estimation*, where the X_i are bivariate observations, say $X_i = (Y_i, Z_i)$, and we wish to estimate $\theta(F) = E_F Y / E_F Z$. (Take $Y, Z > 0$ for convenience.) Let $t(\mathbf{X}) = \bar{Y}/\bar{Z}$, and $R(\mathbf{X}, F) = t(\mathbf{X})/\theta(F)$. It is easily verified that

$$U_i = \frac{y_i}{\bar{y}} - \frac{z_i}{\bar{z}}, \qquad V_{ij} = 2\frac{z_i}{\bar{z}}\frac{z_j}{\bar{z}} - \left(\frac{y_i}{\bar{y}}\frac{z_j}{\bar{z}} + \frac{y_j}{\bar{y}}\frac{z_i}{\bar{z}}\right), \tag{5.14}$$

and that (5.8), (5.10) give

$$E_* R^* \doteq 1 - \frac{1}{n^2}\left\{\Sigma_i\left(\frac{y_i}{\bar{y}} - 1\right)\left(\frac{z_i}{\bar{z}} - 1\right) - \Sigma_i\left(\frac{z_i}{\bar{z}} - 1\right)^2\right\}, \tag{5.15}$$

$$\mathrm{Var}_* R^* \doteq \frac{1}{n^2}\Sigma_i\left[\frac{y_i}{\bar{y}_{\cdot}} - \frac{z_i}{\bar{z}}\right]^2.$$

The biased corrected estimate for $\theta(F)$ is $t(\mathbf{X})/E_* R^*$, with approximate variance $(\hat{\theta}/n)^2 \Sigma[y_i/\bar{y} - z_i/\bar{z}]^2$. If the statistician feels uneasy about expressions (5.15) for any particular data set, perhaps because of outlying values, Method 2 can be invoked to check the bootstrap distribution of $t(\mathbf{X}^*)$ directly.

The infinitesimal jackknife and the ordinary jackknife can both be applied starting from $\hat{F}_{\mathrm{SYM}}$, (3.8), rather than from $\hat{F}$. It is easiest to see how for the infinitesimal jackknife. Expansion (5.4) is still valid except that $\mathbf{U}$ is now a $(2n-1) \times 1$ vector, $\mathbf{V}$ is a $(2n-1) \times (2n-1)$ matrix, and $\mathbf{P}^*$ has bootstrap mean $\mathbf{e}/(2n-1)$, covariance matrix $(1/n)[\mathbf{I}/(2n-1) - \mathbf{e}'\mathbf{e}(2n-1)^2]$. The variance approximation corresponding to (5.10) is

$$\mathrm{Var}_{*\mathrm{SYM}} R(\mathbf{P}^*) = \frac{\sum_{l=1}^{2n-1} U_l^2}{n(2n-1)}. \tag{5.16}$$

6. Wilcoxon's statistic. We again consider the two-sample situation (4.1), this time with F and G being continuous probability distributions on the real line. The

parameter of interest will be

$$\theta(F, G) = P_{F,G}(X < Y), \tag{6.1}$$

estimated by Wilcoxon's statistic

$$\hat{\theta} = \theta(\hat{F}, \hat{G}) = \frac{1}{mn}\Sigma_{i=1}^{m}\Sigma_{j=1}^{n} I(X_i, Y_j), \tag{6.2}$$

where

$$\begin{aligned} I(a, b) &= 1 \qquad a < b \\ &= 0 \qquad a \geqslant b. \end{aligned} \tag{6.3}$$

The bootstrap variance of $\hat{\theta}$ can be calculated directly by Method 1, and will turn out below to be the same as the standard variance approximation for Wilcoxon's statistic. The comparison with Method 3, the infinitesimal jackknife, illustrates how this theory works in a two-sample situation. More importantly, *it suggests the correct analogue of the ordinary jackknife for such situations.*

There has been considerable interest in extending the ordinary jackknife to "unbalanced" situations, i.e., those where it is not clear what the correct analogue of "leave one out" is, see Miller [15], Hinkley [9]. In the two-sample problem, for example, should we leave out one x_i at a time, then one y_j at a time, or should we leave out all mn pairs (x_i, y_j) one at a time? (The former turns out to be correct.) This problem gets more crucial in the next section, where we consider regression problems.

Let $R((\mathbf{X}, \mathbf{Y}), (F, G))$ be $\hat{\theta}$ itself, so that the bootstrap value of R corresponding to $(\mathbf{X}^*, \mathbf{Y}^*)$ is $R((\mathbf{X}^*, \mathbf{Y}^*), (\hat{F}, \hat{G})) = \hat{\theta}^*$,

$$\hat{\theta}^* = \frac{1}{mn}\Sigma_i \Sigma_j I(X_i^*, Y_j^*). \tag{6.4}$$

Letting $I_{ij}^* = I(X_i^*, Y_j^*)$, straightforward calculations familiar from standard nonparametric theory, give

$$E_* I_{ij}^* = \hat{\theta}, \qquad \mathrm{Var}_* I_{ij}^* = \hat{\theta}(1 - \hat{\theta}), \qquad E_* I_{ij}^* I_{i'j'}^* = \hat{\theta}^2 \quad i \neq i', j \neq j' \tag{6.5}$$

and

$$\begin{aligned} E_* I_{ij}^* I_{ij'}^* &= \int_{-\infty}^{\infty} [1 - \hat{G}(z)]^2 d\hat{F}(z) \equiv \hat{\alpha}, & j \neq j' \\ E_* I_{ij}^* I_{i'j}^* &= \int_{-\infty}^{\infty} \hat{F}^2(z) d\hat{G}(z) \equiv \hat{\beta}, & i \neq i'. \end{aligned} \tag{6.6}$$

Using these results in (6.4) gives

$$\mathrm{Var}_* \hat{\theta}^* = \frac{(n-1)(\hat{\alpha} - \hat{\theta}^2) + (m-1)(\hat{\beta} - \hat{\theta})^2 + \hat{\theta}(1 - \hat{\theta})}{mn}, \tag{6.7}$$

which is the usual estimate for the variance of the Wilcoxon statistic, see Noether [16], page 32.

Method 3, the Taylor series or infinitesimal jackknife, proceeds as in Section 5, with obvious modifications for the two-sample situation. Let $\mathbf{N}_F^* = (N_{F1}^*, N_{F2}^*, \cdots, N_{Fm}^*)$ be the numbers of times $x_1, x_2, \cdots, x_m$ occur in the bootstrap sample $\mathbf{X}^*$, likewise $\mathbf{N}_G^* = (N_{G1}^*, N_{G2}^*, \cdots, N_{Gn}^*)$ for $\mathbf{Y}^*$, and define $\mathbf{P}_F^* = \mathbf{N}_F^*/m$, $\mathbf{P}_G^* = \mathbf{N}_G^*/n$, these being independent random vectors with mean and covariance as in (5.2). The expansion corresponding to (5.4) is

$$
\begin{aligned}
(6.8)\qquad R(\mathbf{P}_F^*, \mathbf{P}_G^*) \doteq\ & R(\mathbf{e}/m, \mathbf{e}/n) + (\mathbf{P}_F^* - \mathbf{e}/m)\mathbf{U}_F + (\mathbf{P}_G^* - \mathbf{e}/n)\mathbf{U}_G \\
& + \frac{1}{2}\big[(\mathbf{P}_F^* - \mathbf{e}/m)\,V_F(\mathbf{P}_F^* - \mathbf{e}/m)' \\
& + 2(\mathbf{P}_F^* - \mathbf{e}/m)\,V_{FG}(\mathbf{P}_G^* - \mathbf{e}/n)' \\
& + (\mathbf{P}_G^* - \mathbf{e}/n)\,V_G(\mathbf{P}_G^* - \mathbf{e}/n)'\big],
\end{aligned}
$$

where

$$
(6.9)\qquad U_{Fi} = \partial R/\partial P_{Fi}^*, \qquad V_{Fii'} = \partial^2 R/\partial P_{Fi}^*\partial P_{Fi'}^*, \qquad V_{FGij} = \partial^2 R/\partial P_{Fi}^*\partial P_{Gj}^*,
$$

all the derivatives being evaluated at $(\mathbf{P}_F^*, \mathbf{P}_G^*) = (\mathbf{e}/m, \mathbf{e}/n)$, analogous definitions applying to $\mathbf{U}_G$ and $\mathbf{V}_G$.

The results corresponding to (5.8) and (5.10) are

$$
(6.10)\qquad E_* R^* \doteq R(\mathbf{e}/m, \mathbf{e}/n) + \frac{1}{2}\left[\frac{\bar{V}_F}{m} + \frac{\bar{V}_G}{n}\right]
$$

and

$$
(6.11)\qquad \mathrm{Var}_* R^* \doteq \sum_{i=1}^m U_{Fi}^2/m^2 + \sum_{j=1}^n U_{Gj}^2/n^2,
$$

$\bar{V}_F = \sum_i V_{Fii}/m$, $\bar{V}_G = \sum_j V_{Gjj}/n$. For $R = \theta(\hat{F}, \hat{G}) - \theta(F, G)$, the approximations corresponding to (5.11) are

(6.12)

$$
\mathrm{Bias}_{F,G}\theta(\hat{F}, \hat{G}) \approx \frac{1}{2}\left[\frac{\bar{V}_F}{m} + \frac{\bar{V}_G}{n}\right], \qquad \mathrm{Var}_{F,G}\theta(F, G) \approx \frac{\sum_{i=1}^m U_{Fi}^2}{m^2} + \frac{\sum_{j=1}^n U_{Gj}^2}{n^2}.
$$

For the case of the Wilcoxon statistic (6.11) (or (6.12)) gives

$$
(6.13)\qquad \mathrm{Var}_*\hat{\theta}^* \doteq \frac{n\big[\hat{\alpha} - \hat{\theta}^2\big] + m\big[\hat{\beta} - \hat{\theta}^2\big]}{mn},
$$

which should be compared with (6.7).

How can we use the ordinary jackknife to get results like (6.12)? A direct analogy of (5.12) can be carried through, but it is simpler to change definitions slightly, letting

$$
\begin{aligned}
(6.14)\qquad D_{(i,\ )} &= R(\mathbf{e}/m, \mathbf{e}/n) - R(\mathbf{e}_{(i)}/(m-1), \mathbf{e}/n) \\
D_{(\ ,j)} &= R(\mathbf{e}/m, \mathbf{e}/n) - R(\mathbf{e}/m, \mathbf{e}_{(j)}/(n-1)),
\end{aligned}
$$

the difference from $R((\mathbf{x}, \mathbf{y}), (\hat{F}, \hat{G}))$ obtained by deleting x_i from $\mathbf{x}$ or y_j from $\mathbf{y}$. Expansion (6.8) gives

$$D_{(i,\ )} \doteq \frac{m-2}{(m-1)^2} U_{Fi} - \frac{1}{2(m-1)^2} V_{Fii} \tag{6.15}$$

$$D_{(\ ,j)} \doteq \frac{(n-2)^2}{(n-1)^2} U_{Gj} - \frac{1}{2(n-1)^2} V_{Gjj}.$$

From (6.15) it is easy to obtain approximations for the bias and variance expressions in terms of the D's:

$$-\left[\sum_{i=1}^m D_{(i,\ )} + \sum_{j=1}^n D_{(\ ,j)}\right] \doteq \frac{1}{2}\left[\left(\frac{m}{m-1}\right)^2 \overline{V}_F + \left(\frac{n}{n-1}\right)^2 \overline{V}_G\right], \tag{6.16}$$

which, as m and n grow large, approaches the second term in (6.10). (For $R = \hat{\theta} - \theta$, this gives the bias-corrected estimate $\tilde{\theta} = (m + n - 1)\hat{\theta} - \sum_i \hat{\theta}_{(i,\ )} - \sum_j \hat{\theta}_{(\ ,j)}$.) Likewise, just using the first line of (6.8) gives

$$\sum_{i=1}^m D_{(i,\ )}^2 + \sum_{j=1}^n D_{(\ ,j)}^2 \doteq \frac{m^2(m-2)^2}{(m-1)^4} \frac{\sum_{i=1}^m U_{Fi}^2}{m^2} + \frac{n^2(n-2)^2}{(n-1)^2} \frac{\sum_{j=1}^n U_{Gj}^2}{n^2}, \tag{6.17}$$

which approaches (6.11) as $m, n \to \infty$.

The advantage of the D's over expressions like (5.12) is that no group averages, such as R^*, need be defined. Group averages are easy enough to define in the two-sample problem, but are less clear in more complicated situations such as regression. Expressions (6.16) and (6.17) are easy to extend to any situation (which doesn't necessarily mean they give good answers—see the remarks of the next section!).

7. Regression models. A reasonably general regression model is

$$X_i = g_i(\beta) + \epsilon_i \qquad i = 1, 2, \ldots, n, \tag{7.1}$$

the $g(\cdot)$ being known functions of the unknown parameter vector β, and

$$\epsilon_i \sim_{\text{ind}} F \qquad i = 1, 2, \ldots, n. \tag{7.2}$$

All that is assumed known about F is that it is centered at zero in some sense, perhaps $E_F\epsilon = 0$ or $\text{Median}_F\epsilon = 0$. Having observed $\mathbf{X} = \mathbf{x}$, we use some fitting technique to estimate β, perhaps least squares,

$$\hat{\beta} : \min_\beta \sum_{i=1}^n [x_i - g_i(\beta)]^2, \tag{7.3}$$

and wish to say something about the sampling distribution of $\hat{\beta}$.

Method 2, the brute force application of the bootstrap, can be carried out by defining $\hat{F}$ as the sample probability distribution of the residuals $\hat{\epsilon}_i$,

$$\hat{F} : \text{mass}\frac{1}{n} \quad \text{at} \quad \hat{\epsilon}_i = x_i - g_i(\hat{\beta}), \qquad i = 1, 2, \ldots, n. \tag{7.4}$$

(If one of the components of β is a translation parameter for the functions $g(\cdot)$, then $\hat{F}$ has mean zero. If not, and if the assumption $E_F\epsilon = 0$ is very firm, one might still modify $\hat{F}$ by translation to achieve zero mean.) The bootstrap sample, given $(\hat{\beta}, \hat{F})$, is

$$X_i^* = g_i(\hat{\beta}) + \epsilon_i^*, \qquad \epsilon_i^* \sim_{\text{ind}} \hat{F} \qquad i = 1, 2, \ldots, n. \tag{7.5}$$

Each realization of (2.5) yields a realization of $\hat{\beta}^*$ by the same minimization process that gave $\hat{\beta}$,

$$\hat{\beta}^* : \min_\beta \sum_{i=1}^n [x_i^* - g_i(\beta)]^2. \tag{7.6}$$

Repeated independent bootstrap replications give a random sample $\hat{\beta}^{*1}, \hat{\beta}^{*2}, \hat{\beta}^{*3}, \ldots, \hat{\beta}^{*N}$ which can be used to estimate the bootstrap distribution of $\hat{\beta}^*$.

A handy test case is the familiar linear model, $g_i(\beta) = c_i\beta$, c_i a known $1 \times p$ vector, with first coordinate $c_{i1} = 1$ for convenience. Let $\mathbf{C}$ be the $n \times p$ matrix whose ith row is c_i, and $\mathbf{G}$ the $p \times p$ matrix $\mathbf{C}'\mathbf{C}$, assumed nonsingular. Then the least squares estimator $\hat{\beta} = \mathbf{G}^{-1}\mathbf{C}'\mathbf{X}$ has mean β and covariance matrix $\sigma_F^2\mathbf{G}^{-1}$ by the usual theory.

The bootstrap values ϵ_i^* used in (7.5) are independent with mean zero and variance $\hat{\sigma}^2 = \sum_{i=1}^n [x_i - g(\hat{\beta})]^2/n$. This implies that $\hat{\beta}^* = \mathbf{G}^{-1}\mathbf{C}'\mathbf{X}^*$ has bootstrap mean and variance

$$E_*\hat{\beta}^* = \hat{\beta}, \qquad \text{Cov}_*\hat{\beta}^* = \hat{\sigma}^2\mathbf{G}^{-1}. \tag{7.7}$$

The implication that $\hat{\beta}$ is unbiased for β, with covariance matrix approximately equal to $\hat{\sigma}^2\mathbf{G}^{-1}$, agrees with traditional theory, except perhaps in using the estimate $\hat{\sigma}^2$ for σ^2.

Miller [15] and Hinkley [9] have applied, respectively, the ordinary jackknife and infinitesimal jackknife to the linear regression problem. They formulate the situation as a one-sample problem, with (c_i, x_i) as the ith observed data point, essentially removing one row at a time from the model $\mathbf{X} = \mathbf{C}\beta + \epsilon$. The infinitesimal jackknife gives the approximation

$$\text{Cov}\,\hat{\beta} \approx \mathbf{G}^{-1}\Big[\sum_{i=1}^n c_i'c_i\hat{\epsilon}_i^2\Big]\mathbf{G}^{-1}, \tag{7.8}$$

(and the ordinary jackknife a quite similar expression) for the estimated covariance matrix. This doesn't look at all like (7.7)!

The trouble lies in the fact that the jackknife methods as used above ignore an important aspect of the regression model, namely that the errors ϵ_i are assumed to have the same distribution for every value of i. To make (7.8) agree with (7.7) it is only necessary to "symmetrize" the data set by adding hypothetical data points, corresponding to all the possible values of the residual $\hat{\epsilon}$, at each value of i, say

$$x_{ij} = c_i\hat{\beta} + \hat{\epsilon}_j \qquad j = 1, 2, \ldots, n \quad (i = 1, 2, \ldots, n). \tag{7.9}$$

Notice that the bootstrap implicitly does this at step (7.5). Applying the infinitesimal jackknife to data set (7.9), and remembering to take account of the artificially increased amount of data as at step (5.16), gives covariance estimate (7.7).

Returning to the nonlinear regression model (7.1), (7.2), where bootstrap-jackknife methods may really be necessary in order to get estimates of variability for $\hat{\beta}$, we now suspect that jackknife procedures like "leave out one row at a time" may be inefficient unless preceded by some form of data symmetrization such as (7.9). To put things the other way, as in Hinkley [9], such procedures tend to give consistent estimates of Cov $\hat{\beta}$ without assumption (7.2) that the residuals are identically distributed. The price of such complete generality is low efficiency. Usually assumption (7.2) can be roughly justified, perhaps after suitable transformations on X, in which case the bootstrap should give a better estimate of Cov $\hat{\beta}$.

8. Remarks.

REMARK A. Method 2, the straightforward calculation of the bootstrap distribution by repeated Monte Carlo sampling, is remarkably easy to implement on the computer. Given the original algorithm for computing R, only minor modifications are necessary to produce bootstrap replications $R^{*1}, R^{*2}, \ldots, R^{*N}$. The amount of computer time required is just about N times that for the original computations. For the discriminant analysis problem reported in Table 2, each trial of $N = 100$ replications, $m = n = 20$, took about 0.15 seconds and cost about 40 cents on Stanford's 370/168 computer. For a single real data set with $m = n = 20$, we might have taken $N = 1000$, at a cost of \$4.00.

REMARK B. Instead of estimating $\theta(F)$ with $t(\mathbf{X})$, we might make a transformation $\phi = g(\theta)$, $s = g(t)$, and estimate $\phi(F) = g(\theta(F))$ with $s(\mathbf{X}) = g(t(\mathbf{X}))$. That is, we might consider the random variable $S(\mathbf{X}, \mathbf{F}) = s(\mathbf{X}) - \phi(F)$ instead of $R(\mathbf{X}, \mathbf{F}) = t(\mathbf{X}) - \theta(F)$. The effect of such a transformation on the bootstrap is very simple: a bootstrap realization $R^* = R^*(\mathbf{X}^*, \hat{F}) = t(\mathbf{X}^*) - \theta(F)$ transforms into $S = S(\mathbf{X}^*, \hat{F}) = g(t(\mathbf{X}^*)) - g(\theta(\hat{F}))$, or more simply

$$S^* = g(R^* + \hat{\theta}) - g(\hat{\theta}), \tag{8.1}$$

so the bootstrap distribution of R^* transforms into that of S^* by (8.1).

Figure 1 illustrates a simple example. Miller [14], page 12, gives 9 pairs of numbers having sample Pearson correlation coefficient $\hat{\rho} = .945$. The top half of Figure 1 shows the histogram of $N = 1000$ bootstrap replications of $\hat{\rho}^* - \hat{\rho}$, the bottom half the corresponding histogram of $\tanh^{-1} \hat{\rho}^* - \tanh^{-1} \hat{\rho}$. The first distribution straggles off to the left, the second distribution to the right. The median is above zero, but only slightly so compared to the spread of the distributions, indicating that bias correction is not likely to be important in this example.

The purpose of making transformations is, presumably, to improve the inference process. In the example above we might be willing to believe, on the basis of normal theory, that $\tanh^{-1} \hat{\rho} - \tanh^{-1} \rho$ is more nearly *pivotal* than $\hat{\rho} - \rho$ (see

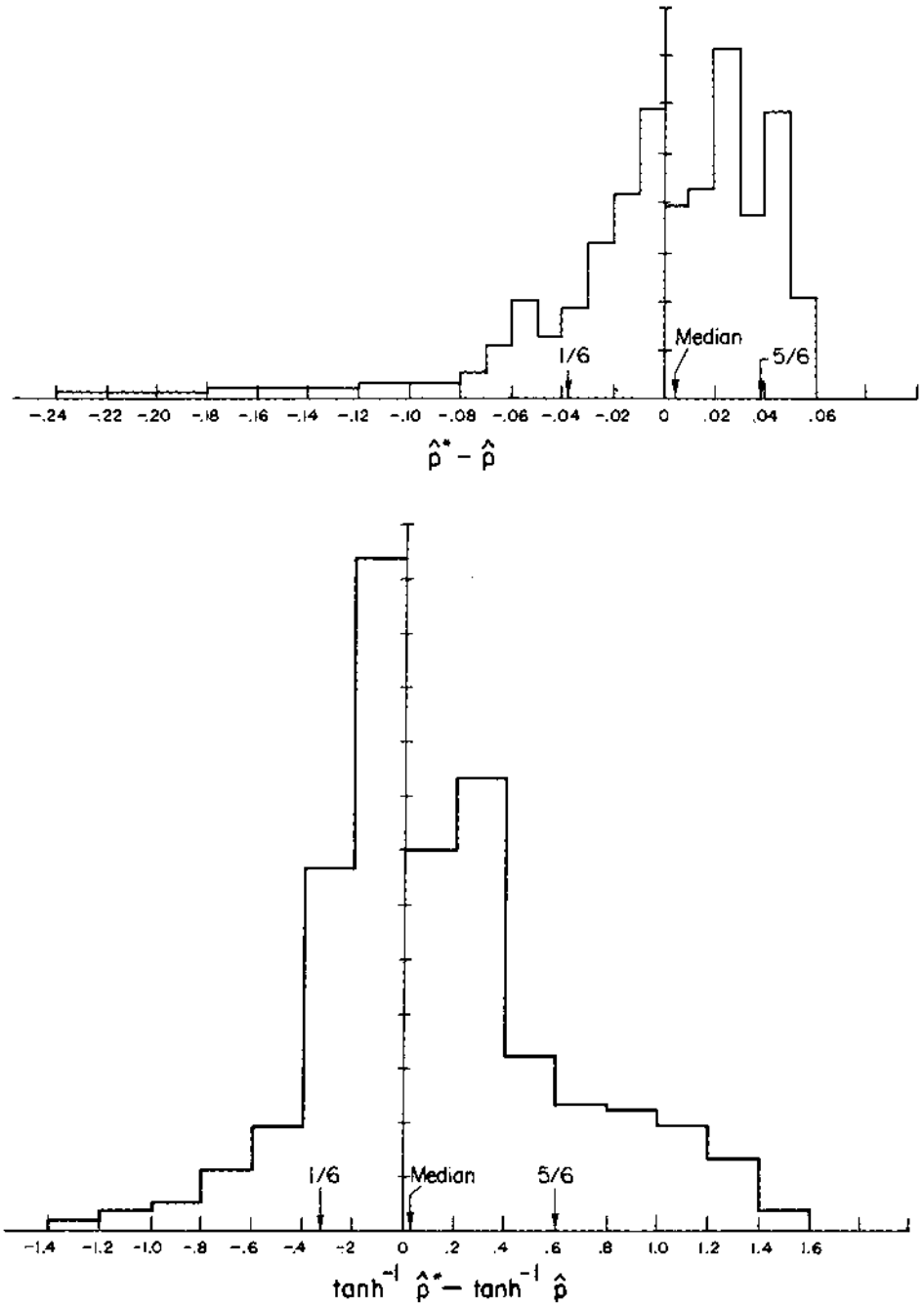

FIG. 1. The top histogram shows $N = 1000$ bootstrap replications of $\hat{\rho}^* - \hat{\rho}$ for the nine data pairs from Miller [10]: (1.15, 1.38), (1.70, 1.72), (1.42, 1.59), (1.38, 1.47), (2.80, 1.66), (4.70, 3.45), (4.80, 3.87), (1.41, 1.31), (3.90, 3.75). The bottom histogram shows the corresponding replications for $\tanh^{-1}\hat{\rho}^* - \tanh^{-1}\hat{\rho}$. The 1/6, 1/2, and 5/6 quantiles are shown for both distributions. All quantiles transform according to equation (8.1).

Remark E) and so more worthwhile investigating by the bootstrap procedure. This does not mean that the bootstrap gives more accurate results, only that the results are more useful. Notice that if $g(\cdot)$ is monotone, then any quantile of the bootstrap distribution of R^* maps into the corresponding quantile of S^* via (8.1), and vice-versa. In particular, if we use the median (rather than the mean) to estimate the center of the bootstrap distribution, then we get the same answer working directly with $\hat{\theta}^* - \hat{\theta}$ ($\hat{\rho}^* - \hat{\rho}$ in the example), or first transforming to $\hat{\phi}^* - \hat{\phi}$ ($\tanh^{-1}\hat{\rho}^* - \tanh^{-1}\hat{\rho}$), taking the median, and finally transforming back to the original scale.

REMARK C. The bias and variance expressions (5.11) suggested by the infinitesimal jackknife transform exactly as in more familiar applications of the "delta method." That is, if $\phi = g(\theta)$, $\hat{\phi} = g(\hat{\theta})$ as above, and $\widehat{\text{Bias}}_F\,\hat{\theta}$, $\widehat{\text{Var}}_F\,\hat{\theta}$ are as given in formula (5.11), then it is easy to show that

$$\widehat{\text{Bias}}_F\,\hat{\phi} = g'(\hat{\theta})\widehat{\text{Bias}}_F\,\hat{\theta} + \frac{g''(\hat{\theta})}{2}\widehat{\text{Var}}_F\,\hat{\theta}, \tag{8.2}$$

$$\widehat{\text{Var}}_F\,\hat{\phi} = \left[\,g'(\hat{\theta})\,\right]^2\widehat{\text{Var}}_F\,\hat{\theta}.$$

In the context of this paper, the infinitesimal jackknife *is* the delta method; starting from a known distribution, that of $\mathbf{P}^*$, approximations to the moments of an arbitrary function $R(\mathbf{P}^*)$ are derived by Taylor series expansion. See Gray et al. [4] for a closely related result.

REMARK D. A standard nonparametric confidence statement for the median $\theta(F)$, $n = 13$, is

$$\text{Prob}_F\{x_{(4)} < \theta < x_{(10)}\} = \text{Prob}\left\{4 \leqslant \text{Bi}\left(13, \tfrac{1}{2}\right) \leqslant 9\right\} = .908. \tag{8.3}$$

If we make the continuity correction of halving the end point probabilities, (3.6) gives

$$\text{Prob}_*\{x_{(4)} < \hat{\theta}^* < x_{(10)}\} = .914, \tag{8.4}$$

where $\hat{\theta}^* = X^*_{(m)}$, the bootstrap value of the sample median. The agreement of (8.4) with (8.3) looks striking, until we try to use (8.4) for inference about θ; (8.4) can be rewritten as $\text{Prob}_*\{x_{(4)} - x_{(7)} < \hat{\theta}^* - \hat{\theta} < x_{(10)} - x_{(7)}\}$ (remembering that $\hat{\theta} = x_{(7)}$), which suggests

$$\text{Prob}_F\{x_{(4)} - x_{(7)} < \hat{\theta} - \theta < x_{(10)} - x_{(7)}\} \approx .914. \tag{8.5}$$

The resulting confidence interval statement for θ, again using $\hat{\theta} = x_{(7)}$, is

$$\text{Prob}_F\{2x_{(7)} - x_{(10)} < \theta < 2x_{(7)} - x_{(4)}\} \approx .914, \tag{8.6}$$

which is the reflection of interval (8.3) about the median!

The trouble here has nothing in particular to do with the bootstrap, and does not arise from the possibly large approximation error in statement (8.5), but rather in the inferential step from (8.5) to (8.6), which tries to use $\hat{\theta} - \theta$ as a *pivotal quantity*.

The same difficulty can be exhibited in parametric families: suppose we know that F is a translation of a standard exponential distribution (density e^{-x}, $x > 0$). Then there exist two positive numbers a and b, $a < b$, such that $\text{Prob}_F\{-a < \hat{\theta} - \theta < b\} = .91$. The corresponding interval statement $\text{Prob}_F\{x_{(7)} - b < \theta < x_{(7)} + a\} = .91$ will tend to look more like (8.6) than (8.3).

REMARK E. The difficulty above is a reminder that the bootstrap, and the jackknife, provide approximate *frequency* statements, not approximate *likelihood* statements. Fundamental inference problems remain, no matter how well the bootstrap works. For example, even if the bootstrap expectation $E_*(\hat{\theta}^* - \theta)^2$ very accurately estimates $E_F(\hat{\theta} - \theta)^2$, the resulting interval estimate for θ given $\hat{\theta}$ will be useless if small changes in F (or more exactly, in $\theta(F)$) result in large changes in $E_F(\hat{\theta} - \theta)^2$.

For the correlation coefficient, as discussed in Remark B, Fisher showed that $\tanh^{-1}\hat{\rho} - \tanh^{-1}\rho$ is nearly pivotal when sampling from bivariate normal populations. That is, its distribution is nearly the same for all bivariate normal populations, at least in the range $-.9 < \rho < .9$. This property tends to ameliorate inference difficulties, and is the principal reason for transforming variables, as in Remark B. The theory of pivotal quantities is well developed in parametric families, see Barnard [2], but not in the nonparametric context of this paper.

REMARK F. The classic pivotal quantity is Student's t-statistic. Tukey has suggested using the analogous quantity (2.3) for hypothesis testing purposes, relying on the standard t tables for significance points. This amounts to treating (2.3) as a pivotal quantity for all choices of F, $\theta(F)$, and $t(\mathbf{X})$. The only theoretical justifications for this rather optimistic assumption apply to large samples, where the Student t effect rapidly becomes negligible, see Miller [14]. Given the current state of the theory, one is as well justified in comparing (2.3) to a $\mathcal{N}(0, 1)$ distribution as to a Student's t distribution (except when $t(\mathbf{X}) = \bar{X}$).

An alternative approach is to bootstrap (2.3) by Method 2 to obtain a direct estimate of its distribution, instead of relying on the t distribution, and then compare the observed value of (2.3) to the bootstrap distribution.

REMARK G. The rationale for bootstrap methods becomes particularly clear when the sample space $\mathcal{X}$ of the X_i is a finite set, say

$$\mathcal{X} = \{1, 2, 3, \ldots, L\}. \tag{8.7}$$

The distribution F can now be represented by the vector of probabilities $\mathbf{f} = (f_1, f_2, \ldots, f_L)$, $f_l = \text{Prob}_F\{X_i = l\}$. For a given random sample $\mathbf{X}$ let $\hat{f}_l = \#\{X_i = l\}/n$ and $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_L)$, so that if $R(\mathbf{X}, F)$ is symmetrically defined in the components of $\mathbf{X}$ we can write it as a function of $\hat{\mathbf{f}}$ and $\mathbf{f}$, say

$$R(\mathbf{X}, F) = Q(\hat{\mathbf{f}}, \mathbf{f}). \tag{8.8}$$

Likewise $R(\mathbf{X}^*, \hat{F}) = Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})$, where $\hat{f}_l^* = \#\{X_i^* = l\}/n$ and $\hat{\mathbf{f}}^* = (\hat{f}_1^*, \hat{f}_2^*, \ldots, \hat{f}_L^*)$.

Bootstrap methods estimate the sampling distribution of $Q(\hat{\mathbf{f}}, \mathbf{f})$, given the true distribution $\mathbf{f}$, by the conditional distribution of $Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})$ given the observed value of $\hat{\mathbf{f}}$. This is plausible because

$$\hat{\mathbf{f}}|\mathbf{f} \sim \mathcal{M}_L(n, \mathbf{f}) \quad \text{and} \quad \hat{\mathbf{f}}^*|\hat{\mathbf{f}} \sim \mathcal{M}_L(n, \hat{\mathbf{f}}), \tag{8.9}$$

where $\mathcal{M}_L(n, \mathbf{f})$ is the L-category multinomial distribution with sample size n, probability vector $\mathbf{f}$. In large samples we expect $\hat{\mathbf{f}}$ to be close to $\mathbf{f}$, so that for reasonable functions $Q(\cdot, \cdot)$ (8.9) should imply the approximate validity of the bootstrap method.

The *asymptotic* validity of the bootstrap is easy to verify in this framework, assuming some regularity conditions on $Q(\cdot, \cdot)$. Suppose that $Q(\mathbf{f}, \mathbf{f}) = 0$ for all $\mathbf{f}$ (as it does in the usual jackknife situation where $R(\mathbf{X}, F) = \theta(\hat{F}) - \theta(F)$); that the vector $\mathbf{u}(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})$ with lth component equal to $\partial Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})/\partial \hat{f}_l^*$ exists continuously for $(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})$ in an open neighborhood of $(\mathbf{f}, \mathbf{f})$; and that $\mathbf{u} = \mathbf{u}(\mathbf{f}, \mathbf{f})$ does not equal zero. By Taylor's theorem, and the fact that $\hat{\mathbf{f}}^*$ and $\hat{\mathbf{f}}$ converge to $\mathbf{f}$ with probability one,

$$Q(\hat{\mathbf{f}}, \mathbf{f}) = (\hat{\mathbf{f}} - \mathbf{f})(\mathbf{u} + \boldsymbol{\epsilon}_n) \quad \text{and} \quad Q(\hat{\mathbf{f}}^*, \mathbf{f}) = (\hat{\mathbf{f}}^* - \hat{\mathbf{f}})(\mathbf{u} + \hat{\boldsymbol{\epsilon}}_n), \tag{8.10}$$

both $\boldsymbol{\epsilon}_n$ and $\hat{\boldsymbol{\epsilon}}_n$ converging to zero with probability one. From (8.9) and the fact that $\hat{\mathbf{f}}$ converges to $\mathbf{f}$ with probability one, we have

$$n^{\frac{1}{2}}(\hat{\mathbf{f}} - \mathbf{f})|\mathbf{f} \to \mathcal{M}_L(\mathbf{0}, \Sigma_f) \quad \text{and} \quad n^{\frac{1}{2}}(\hat{\mathbf{f}}^* - \hat{\mathbf{f}})|\hat{\mathbf{f}} \to \mathcal{M}_L(\mathbf{0}, \Sigma_f), \tag{8.11}$$

where Σ_f is the matrix with element (l, m) equal to $f_l(\delta_{lm} - f_m)$. Combining (8.10) and (8.11) shows that the bootstrap distribution of $n^{\frac{1}{2}}Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})$, given $\hat{\mathbf{f}}$, is asymptotically equivalent to the sampling distribution of $n^{\frac{1}{2}}Q(\hat{\mathbf{f}}, \mathbf{f})$, given the true probability distribution $\mathbf{f}$. Both have the limiting distribution $\mathcal{N}(0, \mathbf{u}'\Sigma_f\mathbf{u})$.

The argument above assumes that the form of $Q(\cdot, \cdot)$ does not depend upon n. More careful considerations are necessary in cases like (2.3) where $Q(\cdot, \cdot)$ does depend on n, but in a minor way. Some nondifferentiable functions such as the sample median (3.3) can also be handled by a smoothing argument, though direct calculation of the limiting distribution is easier in that particular case.

REMARK H. Taylor expansion (5.4) looks suspicious because the dimension of the vectors involved increases with the sample size n. However in situation (8.7), (8.8), it is easy to verify that (5.4) is the same as the second order Taylor expansion of $Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})$, for $\hat{\mathbf{f}}^*$ near $\hat{\mathbf{f}}$,

$$Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}}) \doteq Q(\hat{\mathbf{f}}, \hat{\mathbf{f}}) + (\hat{\mathbf{f}}^* - \hat{\mathbf{f}})\hat{\mathbf{u}} + \tfrac{1}{2}(\hat{\mathbf{f}}^* - \hat{\mathbf{f}})\hat{\mathbf{v}}(\hat{\mathbf{f}}^* - \hat{\mathbf{f}}). \tag{8.12}$$

Here $\hat{\mathbf{u}}$ has lth element $\partial Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})/\partial \hat{\mathbf{f}}_l^*|_{\hat{\mathbf{f}}^* = \hat{\mathbf{f}}}$ and $\hat{\mathbf{v}}$ has l,mth element $\partial^2 Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})/\partial \hat{f}_l^* \partial \hat{f}_m^*|_{\hat{\mathbf{f}}^* = \hat{\mathbf{f}}}$. The dimension of the vectors in (8.12) is L, and does not increase with sample size n. Expressions (5.8), (5.10) are the standard delta theory approximation for the mean and variance of $Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})$, given $\hat{\mathbf{f}}$, obtained from (8.12) and the distributional properties of $\hat{\mathbf{f}}^*|\hat{\mathbf{f}} \sim \mathcal{M}_L(n, \hat{\mathbf{f}})$.

REMARK I. Hartigan [5, 7] has suggested using *subsample* values to obtain confidence statements for an estimated parameter. His method consists of choosing a vector $\mathbf{x}^*$ whose components are a nonempty subset of the observed data vector $\mathbf{X} = \mathbf{x}$ (so each component x_i appears either zero or one time in $\mathbf{x}^*$). This process is repeated N times, where N is small compared to 2^n, giving vectors $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \cdots, \mathbf{x}^{*N}$ and corresponding subsample values $t(\mathbf{x}^{*1}), t(\mathbf{x}^{*2}), \cdots, t(\mathbf{x}^{*N})$ for some symmetric estimator $t(\cdot)$ defined for samples of an arbitrary size. By a clever choice of the vectors $\mathbf{x}^{*j}$, and for certain special estimation problems, the $t(\mathbf{x}^{*j})$ can be used to make precise confidence statements about an unknown parameter. More importantly in the context of this paper, Hartigan shows that by choosing the $\mathbf{x}^{*j}$ randomly, without replacement, from the $2^n - 1$ possible nonempty subsamples of x, asymptotically valid confidence statements can be made under fairly general conditions. This is very similar to bootstrap Method 2, except that the $\mathbf{x}^{*j}$ are selected by subsampling rather than bootstrapping.

In the finite case (8.7), let $\mathbf{x}^*$ be a randomly selected subsample vector, and let $\hat{f}_l^* = \#\{x_i^* = l\}/(\text{number of components of } \mathbf{x}^*)$, so $\hat{\mathbf{f}}^* = (\hat{f}_1^*, \hat{f}_2^*, \cdots, \hat{f}_L^*)$, as before, is the vector of proportions in the artificially created sample. It is easy to show that $n^{\frac{1}{2}}(\hat{\mathbf{f}}^* - \hat{\mathbf{f}})|\hat{\mathbf{f}} \to \mathcal{N}_L(\mathbf{0}, \Sigma_f)$, as at (8.11), which is all that is needed to get the same asymptotic properties obtained for the bootstrap. (Conversely, it can be shown that bootstrap samples have the same asymptotic "typicality" properties Hartigan discusses in [5, 7].) The bootstrap *may* give better small sample performance, because the similarity in (8.9), which is unique to the bootstrap, is a stronger property than the asymptotic equivalence (8.11), and also because the artificial samples used by the bootstrap are the same size as the original sample. However, no evidence one way or the other is available at the present time.

Hartigan's 1971 paper [6] introduces another method of resampling, useful for constructing prediction intervals, which only involves artificial samples of the same size as the real sample. Let $\{x_1^*, x_2^*, \cdots, x_n^*\}$ be a set of size n, each element of which is selected with replacement from $\{x_1, x_2, \cdots, x_n\}$. There are $\binom{2n-1}{n-1}$ distinct such sets, not counting differences in the order of selection. (For example $\{x_1, x_2\}$ yields the three sets $\{x_1, x_1\}$, $\{x_2, x_2\}$, $\{x_1, x_2\}$.) The random version of Hartigan's second method selects $\mathbf{x}^*$, or more exactly the set of components of $\mathbf{x}^*$, with equal probability from among these $\binom{2n-1}{n-1}$ possible choices. It can be shown that this results in $n^{\frac{1}{2}}(\hat{\mathbf{f}}^* - \hat{\mathbf{f}})|\hat{\mathbf{f}} \to \mathcal{N}_L(\mathbf{0}, 2\Sigma_f)$, so that the asymptotic covariance matrix is twice what it is in (8.11). Looking at (8.10), one sees that for this resampling scheme, $2^{-\frac{1}{2}}\, Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})$ has the same asymptotic distribution as $Q(\hat{\mathbf{f}}, \mathbf{f})$.

It is not difficult to construct other resampling schemes which give correct asymptotic properties. The important question, but one which has not been investigated, is which scheme is most efficient and reliable in small samples.

REMARK J. In situation (8.7), (8.8), the ordinary jackknife depends on evaluating $Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})$ for vectors $\hat{\mathbf{f}}^*$ of the form $\hat{\mathbf{f}}_{(l)}^*$,

$$(\hat{\mathbf{f}}_{(l)}^* - \hat{\mathbf{f}}) = \frac{1}{n-1}(\hat{\mathbf{f}} - \mathbf{e}_l), \tag{8.13}$$

$\mathbf{e}_l = (0, 0, \ldots, 1, 0, \ldots, 0)$, 1 in the lth place. (The values of l needed are those occurring in the observed sample $(x_1, x_2, \ldots, x_n)$; a maximum of $\min(n, L)$ different l values are possible.) Notice that

$$\|\hat{f}^*_{(l)} - \hat{f}\| \leqslant \frac{2^{\frac{1}{2}}}{n-1}. \tag{8.14}$$

The "resampling" vectors $\hat{f}^*_{(l)}$ are distance $O(1/n)$ away from $\hat{\mathbf{f}}$, as compared to $O_p(n^{-\frac{1}{2}})$ for the bootstrap vectors $\hat{\mathbf{f}}^*$, as seen in (8.11). In the case of the median, (3.3), the jackknife fails because of its overdependence on the behavior of $Q(\hat{\mathbf{f}}^*, \hat{\mathbf{f}})$ for $\hat{\mathbf{f}}^*$ very near $\hat{\mathbf{f}}$. In this case the derivative of the function $Q(\cdot, \cdot)$ is too irregular for the jackknife's quadratic extrapolation formulas to work. The grouped jackknife, in which the $\hat{\mathbf{f}}^*$ vectors are created by removing observations from $\mathbf{x}$ in groups of size g at a time, see page 1 of Miller [14], overcomes this objection if g is sufficiently large. (The calculations above suggest $g = O(n^{\frac{1}{2}})$.) As a matter of fact, the grouped jackknife gives the correct asymptotic variance for the median. If g is really large, say $g = n/2$, and the removal groups are chosen randomly, then this resampling method is almost the same as Hartigan's subsampling plan, discussed in Remark I.

REMARK K. We have applied the bootstrap in a nonparametric way, but there is no reason why it cannot be used in parametric problems. The only change necessary is that at (2.4), $\hat{F}$ is chosen to be the parametric m.l.e. for F, rather than the nonparametric m.l.e. As an example, suppose that F is known to be normal, with unknown mean and variance, and that we are interested in the expectation of $R(\mathbf{X}, F) = I_{[a,\,b]}(\bar{X})$, i.e., the probability that $\bar{X}$ occurs in a prespecified interval $[a, b]$. Then the *nonparametric* bootstrap estimate is $E_*R^* = \hat{G}^{(n)}(b) - \hat{G}^{(n)}(a)$, where $\hat{G}^{(n)}$ is the cdf of $\sum_{i=1}^n X_i^*/n$, obtained by convoluting the sample distribution $\hat{F}$ n times and then rescaling by division by n. The *parametric* bootstrap estimate is $E_*R^* = \Phi\big((b - \bar{x})/\big(\hat{\sigma} - n^{\frac{1}{2}}\big)\big) - \Phi\big((a - \bar{x})/\big(\hat{\sigma}/n^{\frac{1}{2}}\big)\big)$, where $\hat{\sigma} = \hat{\mu}_2^{\frac{1}{2}}$ and $\Phi(\cdot)$ is the standard normal cdf. If F is really normal and if n is moderately large, $n \geqslant 20$ according to standard Edgeworth series calculations, then the two estimates will usually be in close agreement.

It can be shown that the parametric version of Method 3 of the bootstrap, applied to estimating the variance of the m.l.e. in a one parameter family, gives the usual approximation: one over the Fisher information. The calculation is almost the same as that appearing in Section 3 of Jaeckel [10].

Acknowledgments. I am grateful to Professors Rupert Miller and David Hinkley for numerous discussions, suggestions and references, and to Joseph Verducci for help with the numerical computations. The referees contributed several helpful ideas, especially concerning the connection with Hartigan's work, and the large sample theory. I also wish to thank the many friends who suggested names more colorful than *Bootstrap*, including *Swiss Army Knife, Meat Axe, Swan-Dive, Jack-Rabbit*, and my personal favorite, the *Shotgun*, which, to paraphrase Tukey, "can blow the head off any problem if the statistician can stand the resulting mess."

REFERENCES

[1] ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

[2] BARNARD, B. (1974). Conditionality, pivotals, and robust estimation. *Proceedings of the Conference on Foundational Questions in Statistical Inference*. Memoirs No. 1, Dept. of Theoretical Statist., Univ. of Aarhus, Denmark.

[3] CRAMÉR, H. (1946). *Mathematical Methods in Statistics*. Princeton Univ. Press.

[4] GRAY, H., SCHUCANY, W. and WATKINS, T. (1975). On the generalized jackknife and its relation to statistical differentials. *Biometrika* **62** 637–642.

[5] HARTIGAN, J. A. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* **64** 1303–1317.

[6] HARTIGAN, J. A. (1971). Error analysis by replaced samples. *J. Roy. Statist. Soc. Ser. B* **33** 98–110.

[7] HARTIGAN, J. A. (1975). Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *Ann. Statist.* **3** 573–580.

[8] HINKLEY, D. (1976[a]). On estimating a symmetric distribution. *Biometrika* **63** 680.

[9] HINKLEY, D. (1976[b]). On jackknifing in unbalanced situations. Technical Report No. 22, Division of Biostatistics, Stanford Univ.

[10] JAECKEL, L. (1972). The infinitesimal jackknife. Bell Laboratories Memorandum ♯MM 72-1215-11.

[11] KENDALL, M. and STUART, A. (1950). *The Advanced Theory of Statistics*. Hafner, New York.

[12] LACHENBRUCH, P. and MICKEY, R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10** 1–11.

[13] MARITZ, J. S. and JARRETT, R. G. (1978). A note on estimating the variance of the sample median. *J. Amer. Statist. Assoc.* **73** 194–196.

[14] MILLER, R. G. (1974[a]). The jackknife—a review. *Biometrika* **61** 1–15.

[15] MILLER, R. G. (1974[b]). An unbalanced jackknife. *Ann. Statist.* **2** 880–891.

[16] NOETHER, G. (1967). *Elements of Nonparametric Statistics*. Wiley, New York.

[17] TOUSSAINT, G. (1974). Bibliography on estimation of misclassification. *IEEE Trans. Information Theory* **20** 472–479.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305