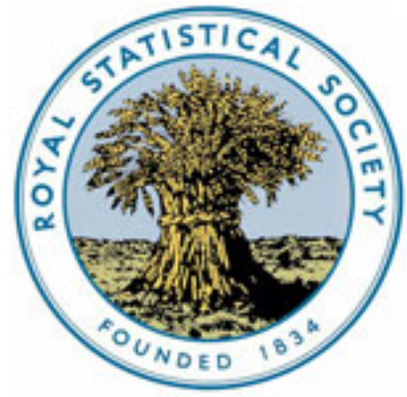

Jackknife-After-Bootstrap Standard Errors and Influence Functions
Author(s): Bradley Efron
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 54, No. 1 (1992), pp. 83-127
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2345949
Accessed: 14/10/2014 12:46

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Wiley and *Royal Statistical Society* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the Royal Statistical Society. Series B (Methodological)*.

http://www.jstor.org

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

J. R. Statist. Soc. B (1992)
54, *No.* 1, *pp.* 83–127

Jackknife-after-Bootstrap Standard Errors and Influence Functions

By BRADLEY EFRON†

Stanford University, USA

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, May 8th, 1991,* Dr F. Critchley *in the Chair*]

SUMMARY

This paper shows how to derive more information from a bootstrap analysis, information about the accuracy of the usual bootstrap estimates. Suppose that we observe data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, compute a statistic of interest $s(\mathbf{x})$ and further compute B bootstrap replications of s, say $s(\mathbf{x}^{*1})$ $s(\mathbf{x}^{*2}), \ldots, s(\mathbf{x}^{*B})$, where B is some large number like 1000. Various accuracy measures for $s(\mathbf{x})$ can be obtained from the bootstrap values, e.g. the bootstrap estimates of standard error and bias, or the length and shape of bootstrap confidence intervals. We might wonder how accurate these accuracy measures themselves are, or how sensitive they are to small changes in the individual data points x_i. It turns out that these questions can be answered from the information in the original bootstrap sample $s^{*1}, s^{*2}, \ldots, s^{*B}$, with no further resampling required. The answers, which make use of the jackknife and delta method influence functions, are easy to apply and can give informative results, as shown by several examples.

Keywords: BOOTSTRAP STATISTICS; CONFIDENCE INTERVAL INFLUENCE; IMPORTANCE SAMPLING; TESTING A PIVOTAL; TUNING AN ESTIMATOR

1. INTRODUCTION

The bootstrap is a computer-based technique for estimating standard errors, biases, confidence intervals and other measures of statistical accuracy. It automatically produces accuracy estimates in almost any situation, including very complicated ones, without requiring much thought from the statistician. This is a considerable virtue, but a virtue that can be abused. The danger lies in the possibility that the bootstrap estimates of accuracy, so easily produced, might be accepted uncritically.

This paper concerns thinking critically about quantities estimated by the bootstrap. To this end we shall use Tukey's jackknife to compute standard errors for bootstrap estimates. The jackknife has an interesting advantage here: the jackknife estimate of standard error for a bootstrap quantity can be computed from the original bootstrap replications, with no further resampling required. Moreover the jackknife calculations provide influence functions as well as standard errors. For example, we will be able to assess the influence of any one of the original data points on the length and shape of a bootstrap confidence interval.

Of what use are error estimates for bootstrap quantities? The examples of the following sections show jackknife-after-bootstrap standard errors playing a variety

† *Address for correspondence*: Department of Statistics, Stanford University, Sequoia Hall, Stanford, CA 94305, USA.

© 1992 Royal Statistical Society 0035–9246/92/54083 $2.00

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

of roles: a comparison of bootstrap variances selects 25% as the preferred trimming proportion in an estimation problem from particle physics; how well determined is this choice? A bootstrap-t analysis is used to form an approximate nonparametric confidence interval for a parameter of interest; how close to pivotal is the bootstrap-t distribution based on this particular small data set? A bootstrap bias estimate is positive; is it significantly positive, or might we easily obtain a negative bias estimate from a similarly constructed independent data set?

The paper proceeds as follows. Section 2 presents the basic jackknife and bootstrap definitions. These are developed in Section 3 to give standard errors and influence functions for bootstrap estimates. A pair of small data sets provides convenient illustrations of the ideas in Sections 2 and 3. The particle physics example of Section 4 shows the jackknife-after-bootstrap idea at work in a bigger, more complicated data analysis problem. Section 5 concerns the delta method, a closely related older cousin of the jackknife. Delta-after-bootstrap standard errors and influence functions are developed and shown to have a computational advantage over the jackknife in one particular situation. Section 6 discusses internal error, the inaccuracy in jackknife-after-bootstrap calculations arising from the limited number of bootstrap replications available in any particular situation.

What about bootstrap-after-bootstrap calculations? Applying a second level of bootstrapping is certainly the most direct and efficient way to assess the accuracy of the first-level results. It is also a more flexible approach than jackknife-after-bootstrap, answering questions that cannot be phrased in terms of standard errors. I am thinking here of Loh's (1987) approach to calibrating confidence limits, and related double-bootstrap hypothesis testing methods due to Chapman and Hinkley (1985), Beran (1988), Tibshirani (1988) and Hall and Martin (1988).

This paper concentrates on error estimates that do not require a second level of bootstrap replication. Jackknife-after-bootstrap and delta-after-bootstrap answers are obtained by simply rearranging the original bootstrap results. This reinforces the truism that bootstrap data, like real data, deserve a thorough examination. More to the point, jackknife-after-bootstrap requires perhaps 100–1000 times less computation than bootstrap-after-bootstrap. This advantage will certainly decrease as computers grow faster, or as more efficient bootstrap algorithms become available; see Hinkley and Shi (1989) and Efron (1990a). At present, bootstrap-after-bootstrap seems to be too computationally intensive for routine use.

2. ONE-SAMPLE NONPARAMETRIC SITUATION

In a one-sample problem the observed data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are obtained by independent and identically distributed (IID) sampling from an unknown distribution F,

$$F \xrightarrow{\text{IID}} (x_1, x_2, \ldots, x_n) = \mathbf{x}. \tag{2.1}$$

The individual data points x_i take their values in a sample space $\mathcal{X}$ which might be the real line, Euclidean vector space or a more general set of possible outcomes. This section defines the jackknife and bootstrap ideas that we shall need in terms of the one-sample nonparametric situation, nonparametric meaning that the unknown distribution F could be any probability distribution on $\mathcal{X}$. Section 4 extends the discussion to a multisample nonparametric situation. Section 5 of Efron (1990a) extends

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

the jackknife-after-bootstrap theory to parametric families, but parametric considerations appear here only in remarks 1 and 8.

Suppose that $s(\mathbf{x})$ is a real-valued statistic of interest, such as a mean, a correlation coefficient, or the maximum eigenvalue of a sample covariance matrix. Let $\mathbf{x}_{(i)}$ indicate the data set remaining after deletion of the ith point,

$$\mathbf{x}_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n), \tag{2.2}$$

and let $s_{(i)} = s(\mathbf{x}_{(i)})$, the corresponding deleted point value of the statistic of interest. The *jackknife influence function* for s is defined to be

$$u_i\{s\} = (n-1)(s_{(\,)} - s_{(i)}) \qquad \left(s_{(\,)} \equiv \sum_{i=1}^{n} s_{(i)}/n\right). \tag{2.3}$$

In the case where the x_i are real valued, and $s(\mathbf{x})$ equals the sample mean $\bar{x}$, formula (2.3) becomes $u_i\{\bar{x}\} = x_i - \bar{x}$. Intuitively, points with large positive or negative values of $x_i - \bar{x}$ have a large influence on the statistic $\bar{x}$. Formula (2.3) generalizes this notion to arbitrary statistics $s(\mathbf{x})$.

The *relative* jackknife influence function

$$u_i^{\dagger}\{s\} = u_i\{s\} \Big/ \left[\sum_j \frac{u_j\{s\}^2}{n-1}\right]^{1/2} \tag{2.4}$$

has a particularly easy interpretation: in the case of the mean, $u_i^{\dagger}\{\bar{x}\} = (x_i - \bar{x})/\hat{\sigma}$ (where $\hat{\sigma}^2 = \Sigma(x_i - \bar{x})^2/(n-1)$), the number of estimated standard deviations of x_i from $\bar{x}$. Moderate values of $u_i^{\dagger}\{s\}$, say $\sup_i(|u_i^{\dagger}\{s\}|) < 2$, assuage concerns about the robustness of $s(\mathbf{x})$. Hampel *et al.* (1986) is an excellent reference for influence functions and robustness.

Fig. 1 displays two small data sets used in most of our examples. Fig. 1(a) shows $n = 15$ data points $x_i = (y_i, z_i)$ pertaining to the entering classes of 1973 at 15 American law schools: y_i is the overall grade point average (GPA) for the class, while z_i is the class average on the national legal test, the law school achievement test (LSAT). Fig. 1 of Efron and Tibshirani (1986) lists the data. We shall be interested in various measures of accuracy concerning the Pearson correlation coefficient $s(\mathbf{x}) = 0.776$. The values plotted are the relative jackknife influence function $u_i^{\dagger}\{s\}$, equation (2.4). Point A is noticeably outlying and negatively influential (pulling down the value of s) $u_i^{\dagger}\{s\} = -2.97$, suggesting that the non-robustness of the Pearson correlation coefficient might have dangerous consequences for the law school data. Those consequences will become more evident in Section 3, when we assess the influence of point A on confidence intervals for the true correlation.

Fig. 1(b) shows $n = 8$ data points $x_i = (y_i, z_i)$ from a bioequivalence study. Each of eight patients was measured three times for the blood level of a certain hormone: once after taking placebo medication, once after taking a compound known to raise the hormone blood level and once after taking a new version of the same compound. Then $y_i =$ blood level after compound *minus* blood level after placebo, and $z_i =$ blood level after new compound *minus* blood level after compound. (The data values are $(y_i, z_i) = (8406, -1200), (2342, 2601), (8187, -2705), (8459, 1982), (4795, -1290), (3516, 351), (4796, -638), (10238, -2719)$.) The statistic of interest here is the ratio

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

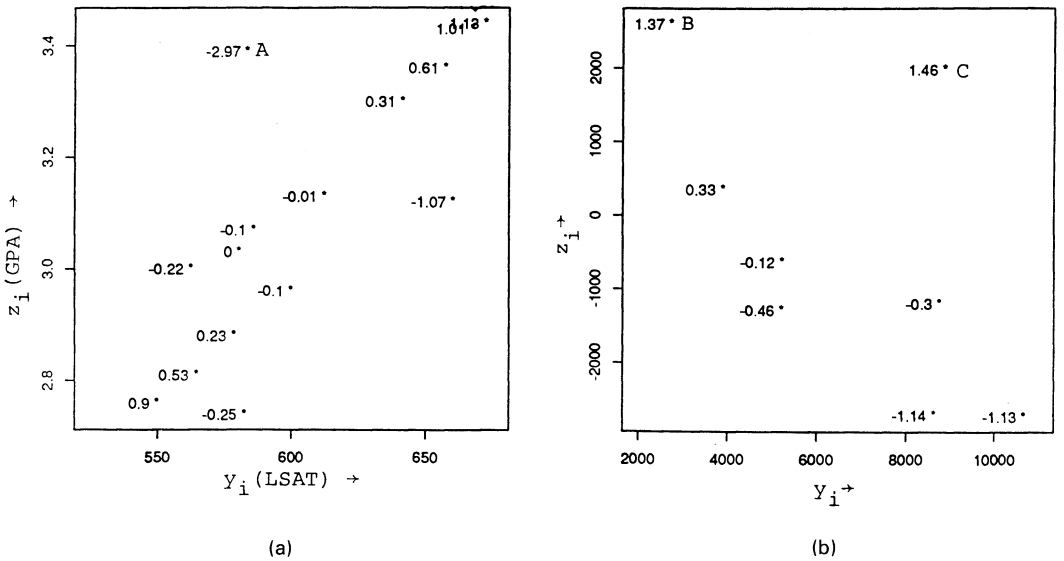

Fig. 1. Two small data sets: (a) the law school data, $n=15$ points $x_i=(y_i, z_i)$, where y_i and z_i are performance measures for the 1973 entering classes at 15 American law schools (the statistic of interest is the Pearson correlation coefficient $s(\mathbf{x})=0.776$; the values plotted are the relative jackknife influence function (2.4); the point labelled A is a noticeable outlier); (b) the bioequivalence data, $n=8$ points $x_i=(y_i, z_i)$ relating to a bioequivalence study described in the text (the statistic of interest is the ratio $s(\mathbf{x})=\bar{z}/\bar{y}=-0.071$; there are no outstanding outliers; the points labelled B and C are moderately influential in the positive direction)

$s(\mathbf{x})=\bar{z}/\bar{y}=-0.071$. We see that there are no flagrant outliers, only points B and C being even moderately influential.

Tukey's jackknife estimate for the standard error of $s(\mathbf{x})$, which followed Quenouille's original suggestion for the jackknife bias estimate, is

$$\mathrm{se}_{\mathrm{jack}}\{s\} \equiv \left[\sum_{i=1}^{n} u_i\{s\}^2/n(n-1)\right]^{1/2}, \tag{2.5}$$

reducing to the usual estimate $\{\Sigma_i(x_i-\bar{x})^2/n(n-1)\}^{1/2}$ for the standard error of $\bar{x}$ (see chapters 3 and 6 of Efron (1982)); $\mathrm{se}_{\mathrm{jack}}\{s\}$ equals 0.143 for the law school correlation coefficient and 0.105 for the bioequivalence ratio statistic. Note that points contribute to the estimated standard error proportional to $u_i\{s\}^2$. We see that $u_i^{\dagger}\{s\}^2/(n-1)=u_i\{s\}^2/\Sigma_j u_j\{s\}^2$; the proportion point i contributes to the estimated standard error. Point A contributes 63% of $\mathrm{se}_{\mathrm{jack}}\{s\}$ for the law school data.

The usual nonparametric estimate of F is $\hat{F}$, the *empirical probability distribution*, putting probability $1/n$ on each point x_i,

$$\hat{F}\text{: probability } 1/n \text{ on } x_i, \qquad i=1, 2, \ldots, n. \tag{2.6}$$

A *bootstrap sample* $\mathbf{x}^*=(x_1^*, x_2^*, \ldots, x_n^*)$ is a random sample of size n drawn from $\hat{F}$,

$$\hat{F} \xrightarrow{\mathrm{IID}} (x_1^*, x_2^*, \ldots, x_n^*)=\mathbf{x}^*. \tag{2.7}$$

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

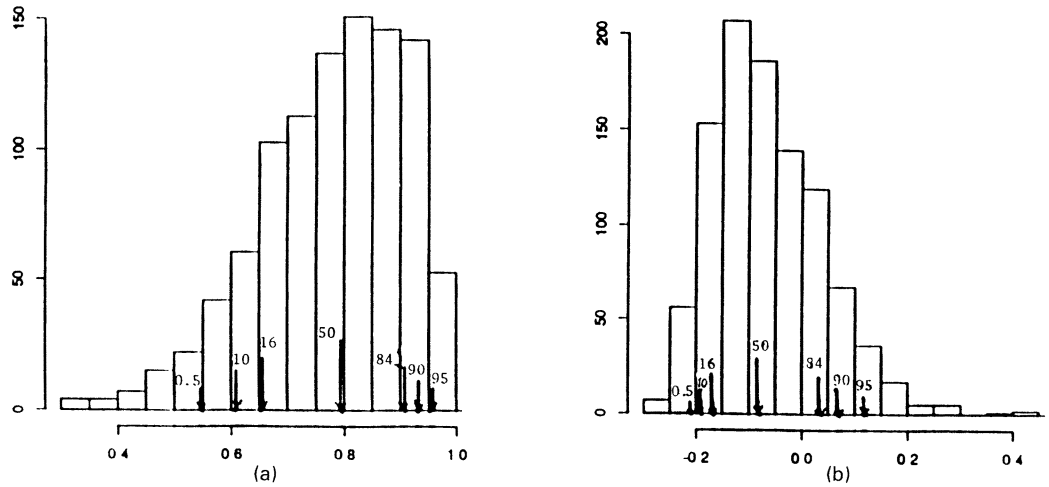

Fig. 2. Histograms of $B = 1000$ bootstrap replications (a) for the correlation coefficient, law school data and (b) for the ratio statistic, bioequivalence data; various percentile points of the bootstrap distributions are indicated

Then $s^* = s(\mathbf{x}^*)$, the statistic of interest evaluated for data set $\mathbf{x}^*$, is a *bootstrap replication* of s. A typical bootstrap analysis consists of independently drawing a large number B of independent bootstrap samples, evaluating the bootstrap replicates $s^{*b} = s(\mathbf{x}^{*b})$ for $b = 1, 2, \ldots, B$ and using summary statistics of the s^{*b} values to assess the accuracy of the original statistic $s(\mathbf{x})$. The two best known summaries are

$$\text{se}_{\text{boot}}\{s\} = \left\{ \sum_b (s^{*b} - s^{*\cdot})^2/(B-1) \right\}^{1/2}, \tag{2.8}$$

($s^{*\cdot} \equiv \Sigma_b s^{*b}/B$), the bootstrap estimate of standard error for s, and

$$\text{bias}_{\text{boot}}\{s\} = s^{*\cdot} - s(\mathbf{x}), \tag{2.9}$$

the bootstrap estimate of bias. See Efron and Tibshirani (1986).

$B = 1000$ bootstrap replications were computed for each of the two examples in Fig. 1. Fig. 2 displays the bootstrap histograms, both of which look noticeably asymmetric. Table 1 gives various summary statistics pertaining to the bootstrap analyses. Two of these statistics are properties of the *percentile confidence interval* $[s^{*(0.05)}, s^{*(0.95)}]$, its length and shape respectively; see Efron (1987). Here $s^{*(\alpha)}$ denotes the 100αth percentile of the empirical distribution of the $B = 1000$ bootstrap replications s^{*b}.

The jackknife-after-bootstrap method of Section 3 allows us to attach standard errors to the statistics in Table 1. These are standard errors in the usual sense: they indicate how much the statistic varies under random sampling (2.1) (not how much it varies because of the limitations of using only B bootstrap replications). We shall see for example that the estimated standard error for the bioequivalence shape estimate 0.440 is 0.313, so that the estimated shape is only 1.41 standard errors above 0; it is plausible that another sample of eight subjects might yield a negative shape for the histogram in Fig. 2(b).

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

TABLE 1

Bootstrap statistics for the law school and bioequivalence bootstrap analyses, $B = 1000$ bootstrap replications each†

Bootstrap statistic	*Law school correlation* $s = 0.776$	*Bioequivalence ratio* $s = -0.071$	*Definition*
se_{boot}	0.128	0.104	Formula (2.8)
$bias_{boot}$	0.002	0.0053	Formula (2.9)
Length	0.402	0.329	$s^{*(0.95)} - s^{*(0.05)}$
(normalized)	(0.95)	(0.96)	(divided by $2 \times 1.645\ se_{boot}$)
Shape	−0.470	0.440	$\log\{(s^{*(0.95)} - s^{*(0.5)})/(s^{*(0.5)} - s^{*(0.05)})\}$
$T^{*(0.95)}$	2.93		95th percentile for bootstrap-t statistic: see Section 3

†The length and shape statistics relate to the central 90% percentile interval $[s^{*(0.05)}, s^{*(0.95)}]$, where $s^{*(\alpha)}$ is the 100αth percentile of the bootstrap replications. Percentile intervals are the simplest form of bootstrap approximate confidence intervals; see Efron (1987).

The entries in Table 1 are *bootstrap statistics*, i.e. functions of $\mathbf{x}$ that are evaluated in terms of bootstrap sampling (2.7). Here is the general definition of a bootstrap statistic for one-sample nonparametric situations.

(a) Begin with a random variable $T(\mathbf{x}, F)$, a function of $\mathbf{x}$ and F, for example

$$T(\mathbf{x}, F) = s(\mathbf{x}) - \theta(F), \tag{2.10}$$

where $\theta(F)$ is a parameter of interest, e.g. the Pearson correlation coefficient (for F bivariate), and $s(\mathbf{x})$ is an estimator of $\theta(F)$, e.g. Spearman's rank correlation. We shall take $T(\mathbf{x}, F)$ real valued, but this is not a necessity of the theory.

(b) Let $[T(\mathbf{X}, F)]$ indicate the probability distribution of $T(\mathbf{X}, F)$, for $\mathbf{X} = (X_1, \ldots, X_n)$ an IID sample from F, and let $\phi[T(\mathbf{X}, F)]$ be some functional of this distribution, e.g. its expectation, its standard deviation or its 90th percentile point.

(c) Finally, set $\gamma(F) \equiv \phi[T(\mathbf{X}, F)]$, and define the *bootstrap statistic*

$$\hat{\gamma}(\mathbf{x}) \equiv \gamma(\hat{F}), \tag{2.11}$$

where $\hat{F}$ is the *empirical probability distribution* (2.6).

Definition (2.11) of a bootstrap statistic is equivalent to

$$\hat{\gamma}(\mathbf{x}) = \phi[T(\mathbf{x}^*, \hat{F})]. \tag{2.12}$$

In the chain of definitions leading to definition (2.12), $\mathbf{x}$ determines $\hat{F}$, (2.6), $\hat{F}$ gives $\mathbf{x}^*$ by random sampling (2.7), $\mathbf{x}^*$ and $\hat{F}$ determine a bootstrap replication of T, e.g. $T(\mathbf{x}^*, F) = s(\mathbf{x}^*) - \theta(\hat{F})$ in case (2.10), and finally the bootstrap distribution of T (i.e. the distribution $[T(\mathbf{x}^*, \hat{F})]$, when $\hat{F}$ is held fixed at its observed value and only $\mathbf{x}^*$ is considered random) determines $\hat{\gamma}$ according to the functional ϕ (2.12).

In practice the numerical value of $\hat{\gamma}(\mathbf{x})$ must be approximated by Monte Carlo methods. We generate B independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ according

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

to formula (2.7), calculate the corresponding bootstrap replications of T, say $T^{*b} = T(\mathbf{x}^{*b}, \hat{F})$, and approximate $\hat{\gamma}(\mathbf{x})$, definition (2.12), by

$$\tilde{\gamma}(\mathbf{x}) = \phi\,[\,T(\mathbf{x}^{*b}, \hat{F}),\quad b = 1, 2, \ldots, B\,]. \tag{2.13}$$

where $[\,T(\mathbf{x}^{*b}, \hat{F}),\ \ b = 1, 2, \ldots, B\,]$ indicates the empirical distribution putting probability $1/B$ on each value T^{*b}. The entries of Table 1 were calculated in this way.

3. JACKKNIFE-AFTER-BOOTSTRAP

This section discusses the application of the jackknife to bootstrap statistics $\hat{\gamma}(\mathbf{x})$, definition (2.12). First we need to calculate the deleted point values $\hat{\gamma}_{(i)} = \hat{\gamma}(\mathbf{x}_{(i)})$, equation (2.2), to compute the jackknife influence function $u_i\{\hat{\gamma}\}$ and standard error estimate $\mathrm{se}_{\mathrm{jack}}\{\hat{\gamma}\}$, equations (2.3) and (2.5). Let $\hat{F}_{(i)}$ indicate the deleted point empirical distribution,

$$\hat{F}_{(i)}\text{: probability } 1/(n-1) \text{ on } x_j, \qquad j = 1, 2, \ldots, i-1, i+1, \ldots, n. \tag{3.1}$$

The following obvious lemma leads directly to the computation of $\hat{\gamma}_{(i)}$.

Lemma 1. An IID sample of size n from $\hat{F}_{(i)}$,

$$\hat{F}_{(i)} \xrightarrow{\text{IID}} (x_1^*, x_2^*, \ldots, x_n^*) = \mathbf{x}^*, \tag{3.2}$$

has the same distribution as a bootstrap sample from $\hat{F}$, sample (2.7), in which none of the x_j^* values equals x_i.

Proof. In either case, each of the n components of $\mathbf{x}^*$ independently equals x_j, $j \neq i$, with probability $1/(n-1)$. □

For a given bootstrap sample (2.7), let P_i denote the proportion of the bootstrap sample equalling x_i,

$$P_i = \#\{x_j^* = x_i\}/n, \tag{3.3}$$

and define the *resampling vector* $\mathbf{P} = (P_1, P_2, \ldots, P_n)'$. (We could, but will not, use the more consistent notation P_i^* and $\mathbf{P}^*$.) Then, according to lemma 1 and definition (2.12),

$$\hat{\gamma}_{(i)} = \phi\,[\,T(\mathbf{x}^*, \hat{F}_{(i)})\,|\,P_i = 0\,], \tag{3.4}$$

where $[\,T(\mathbf{x}^*, \hat{F}_{(i)})\,|\,P_i = 0\,]$ indicates the conditional bootstrap distribution of $T(\mathbf{x}^*, \hat{F}_{(i)})$ given that $P_i = 0$.

In practice, the value of $\hat{\gamma}_{(i)}$, like $\hat{\gamma}$ itself, must be approximated by Monte Carlo methods. We approximate $\hat{\gamma}_{(i)}$ by

$$\tilde{\gamma}_{(i)} = \phi\,[\,T_i^{*b},\ b \text{ such that } P_i^b = 0\,], \tag{3.5}$$

where the bracketed term indicates the empirical distribution of $T_i^{*b} = T(\mathbf{x}^{*b}, \hat{F}_{(i)})$ for those values of b such that the resampling vector $\mathbf{P}^b$ has $P_i^b = 0$. In most cases, including all those considered in this paper, $\tilde{\gamma}_{(i)}$ is easily computed by rearrangement of the original bootstrap calculations. No further resampling computations are required.

The approximations $\tilde{\gamma}_{(i)}$ usually converge to $\hat{\gamma}_{(i)}$ as the number of bootstrap replications $B \to \infty$. (Counter-examples exist for discontinuous functionals ϕ like the

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

TABLE 2

Estimated jackknife standard errors and influence functions for the bootstrap statistics length and shape of Table 1, formulae (3.5) and (3.6)†

		Law school correlation				*Bioequivalence ratio*		
		Correlation	*Length*	*Shape*		*Ratio*	*Length*	*Shape*
Statistic		0.776	0.402	−0.470		−0.071	0.329	0.440
Jackknife standard error		0.143	0.248	0.307		0.106	0.100	0.376
(corrected)			(0.242)	(0.000)			(0.095)	(0.313)
[smoothed]			[0.245]	[0.026]				
		$\tilde{u}_i^{\dagger}\{s\}$	$\tilde{u}_i\{\hat{\gamma}\}$ *(±0.20)*	$\tilde{u}_i\{\hat{\gamma}\}$ *(±1.17)*		$\tilde{u}_i^{\dagger}\{s\}$	$\tilde{u}_i\{\hat{\gamma}\}$ *(±0.086)*	$\tilde{u}_i\{\hat{\gamma}\}$ *(±0.55)*
Influence functions	A	−2.970	3.100	0.380		−1.140	−0.110	0.240
		−1.070	0.790	−1.940		−1.120	−0.200	1.490
		−0.250	−0.520	1.060		−0.460	−0.230	0.020
		−0.220	−0.120	−1.880		−0.310	−0.140	0.110
		−0.100	−0.270	1.540		−0.120	−0.140	−0.150
		−0.100	−0.170	1.020		0.330	0.010	0.040
		−0.010	0.430	−0.710	B	1.370	0.610	0.550
		0.000	−0.340	1.650	C	1.460	0.220	−2.300
		0.230	0.190	−1.040				
		0.310	−0.120	−0.900				
		0.530	−0.520	0.390				
		0.610	−0.500	0.690				
		0.900	−0.430	−0.110				
		1.010	−0.900	0.940				
		1.130	−0.620	−1.090				

†The ± values for $\tilde{u}_i$ and the corrected jackknife standard errors reflect the limitations of our Monte Carlo calculations, with $B = 1000$ rather than $B \to \infty$, as explained in Section 6. The smoothed estimates of the standard error for the law school correlation are explained in remark 11.

percentiles.) Using $\tilde{\gamma}_{(i)}$, we can approximate the jackknife influence function and standard error estimate for $\hat{\gamma}$ in the obvious way,

$$\tilde{u}_i\{\hat{\gamma}\} = (n-1)(\tilde{\gamma}_{(\,)} - \tilde{\gamma}_{(i)}) \qquad \left(\tilde{\gamma}_{(\,)} \equiv \sum_i \tilde{\gamma}_{(i)}/n\right)$$
$$\tilde{\mathrm{se}}_{\mathrm{jack}}\{\hat{\gamma}\} = \left[\sum_i \tilde{u}_i\{s\}^2/n(n-1)\right]^{1/2} \tag{3.6}$$

with, usually, $\tilde{u}_i\{\hat{\gamma}\} \to u_i\{\hat{\gamma}\}$ and $\tilde{\mathrm{se}}_{\mathrm{jack}}\{\hat{\gamma}\} \to \mathrm{se}_{\mathrm{jack}}\{\hat{\gamma}\}$ as $B \to \infty$. Section 6 considers how the number of bootstrap replications B affects the estimates $\tilde{u}_i$ and $\tilde{\mathrm{se}}_{\mathrm{jack}}$. Remark 9 discusses the difference between $\mathrm{se}\{\hat{\gamma}\}$ and $\mathrm{se}\{\tilde{\gamma}\}$, which we shall ignore for now.

Table 2 gives $\tilde{u}_i\{\hat{\gamma}\}$ and $\tilde{\mathrm{se}}_{\mathrm{jack}}\{\hat{\gamma}\}$ for the bootstrap statistic's length and shape, for both data sets. The relative influence functions for length and shape,

$$\tilde{u}_i^{\dagger}\{\hat{\gamma}\} = \tilde{u}_i\{\hat{\gamma}\} \Big/ \left(\sum_j \frac{\tilde{u}_j\{\hat{\gamma}\}^2}{n-1}\right)^{1/2} \tag{3.7}$$

are displayed in Fig. 3.

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

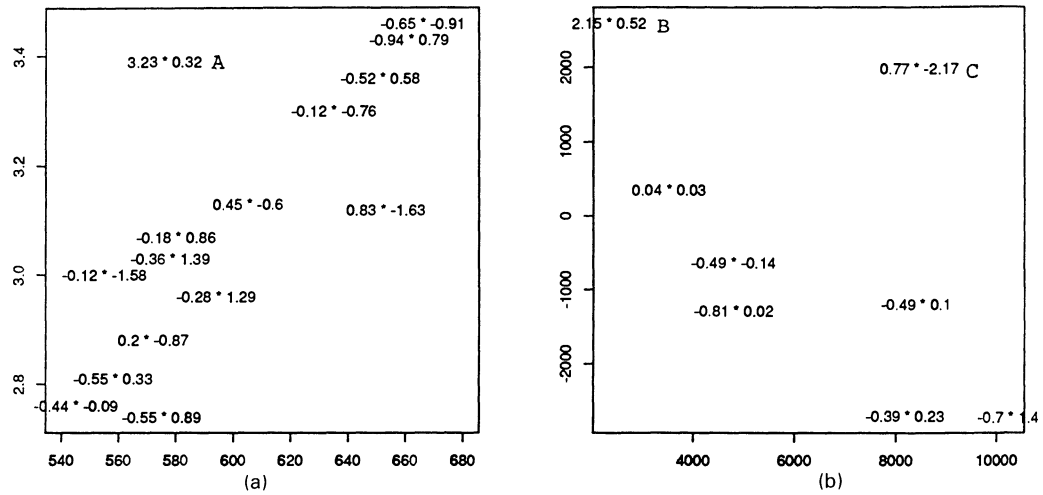

Fig. 3. Relative influence functions $\tilde{u}_i^{\dagger}\{\hat{\gamma}\}$, for two bootstrap statistics $\hat{\gamma}$ (the left-hand number is $\tilde{u}_i^{\dagger}$ for $\hat{\gamma}$ the length of the central 90% percentile interval, as in the third line of Table 1; the right-hand number is $\tilde{u}_i^{\dagger}$ for $\hat{\gamma}$ the shape statistic, fourth line of Table 1; point A is enormously influential for length, but not shape; point B is highly influential for length but not shape; point C is highly influential for shape but not length): (a) law school correlation; (b) bioequivalence ratio

Some non-obvious facts emerge: point A has little influence on shape, but is enormously influential, in a positive direction, for the length of the percentile interval (so using a robust correlation estimate that reduced the influence of point A would probably shorten the length of the relevant confidence interval); point B is highly influential on length but not shape, and vice versa for point C. The length of the percentile confidence interval is better estimated in the bioequivalence case than in the law school case, coefficient of variation $0.100/0.329 = 0.30$ compared with $0.248/0.402 = 0.62$, because of the overwhelming effect of point A. Conversely, shape is better estimated in the law school case.

Table 2 gives error values ($\pm$) for the estimates $\tilde{u}_i\{\hat{\gamma}\}$. These reflect the limitations of using the bootstrap data from only $B = 1000$ replications in equations (3.5) and (2.13), rather than $B \to \infty$; see Section 6. The sum of squares comprising $\widetilde{\text{se}}_{\text{jack}}\{\hat{\gamma}\}$, equations (3.6), is increased by these errors. Corrected values of the jackknife standard errors appear in the third line of Table 2. Even with this correction, the shape statistic for the bioequivalence data is not significantly different from 0, $0.440/0.313 = 1.41$. The $\pm$ errors for the law school shape statistic are so large that they easily account for all of $\widetilde{\text{se}}_{\text{jack}}\{\hat{\gamma}\} = 0.307$. (This can be seen in Fig. 3(a), where adjacent data points have completely different values of $\tilde{u}_{(i)}$. The smoothed estimate of standard error for $\hat{\gamma}$, explained in remark 11 of Section 6, is only 0.026.) Thus $\hat{\gamma}/\text{se}_{\text{jack}}\{\hat{\gamma}\}$ should be substantially bigger than $-0.470/0.307 = -1.53$. It is reasonable to conclude that the asymmetry seen for the law school correlation bootstrap histogram is genuine, but that the asymmetry for the bioequivalence ratio bootstrap histogram may well be an artefact of these particular eight data points.

The computations going into the left-hand side of Table 2 are graphically portrayed in Fig. 4. The horizontal broken lines are at heights $s^{*(\alpha)}$, the 100αth percentile of all 1000 bootstrap correlations for α values 0.05, 0.10, 0.16, 0.50, 0.84, 0.90, 0.95.

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

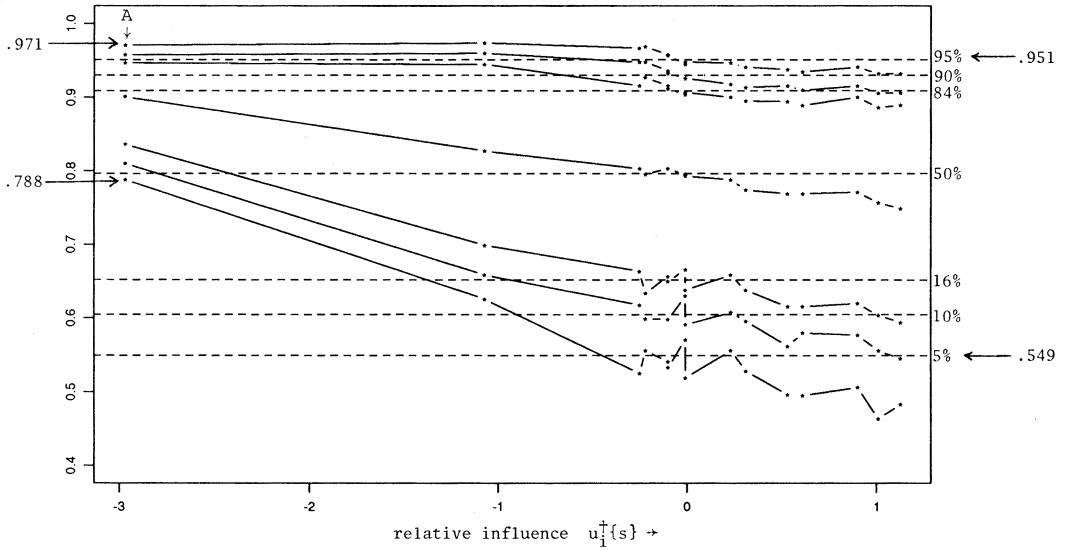

Fig. 4. Percentiles $B = 1000$ bootstrap replications of $s(\mathbf{x})$, the law school correlation coefficient: the broken horizontal lines indicate percentiles of all 1000 values s^{*b}, in ascending order 0.05, 0.10, 0.16, 0.50, 0.84, 0.90, 0.95; the broken curves are sample percentiles for samples $\mathbf{x}^{*b}$ having $P_i^b = 0$; index i is arranged in increasing order of $u_i^{\dagger}\{s\}$, as shown in Fig. 1; for example, 384 of the $\mathbf{x}^{*b}$ were missing point A, and the corresponding values of s^{*b} had fifth percentile 0.788, 95th percentile 0.971; point A is seen to have an enormous negative influence on the lower percentiles of the bootstrap distribution for s, since removing it greatly increases these percentiles

The broken curves trace these same percentiles as one point at a time is deleted from the original sample.

The points were deleted in order of their relative influence on $s(\mathbf{x})$, equation (2.4), and the deleted point percentiles were plotted against $u_i^{\dagger}\{s\}$. Point A of Fig. 1 is plotted at the far left-hand side, at $u_i^{\dagger}\{s\} = -2.97$. 384 of the 1000 bootstrap samples $\mathbf{x}^{*b}$ did not include point A. (The expected number is $1000(1-1/15)^{15} = 355$.) The 384 corresponding bootstrap replications $s(\mathbf{x}^{*b})$ have fifth percentile 0.788 and 95th percentile 0.971. We see that point A has an enormous positive influence on length, since deleting point A reduces length by half, from $0.402 = 0.951 - 0.549$ to $0.183 = 0.971 - 0.788$. This shows up in Fig. 3, where point A is seen to have relative influence 3.23 for length.

Next we shall apply jackknife-after-bootstrap calculations to a more ambitious bootstrap analysis. Let ρ be the true correlation between GPA and LSAT in the law school situation, the correlation of the unknown bivariate distribution F giving the data $\mathbf{x}$, sample (2.1). Suppose that we want a confidence interval for ρ, as was implicitly the reason for our interest in the percentile interval $[s^{*(0.05)}, s^{*(0.95)}]$. The bootstrap-$t$ approach (see Hall (1988)) begins with the definition of a Student t-like quantity pertaining to ρ,

$$T(\mathbf{x}, F) = \frac{s(\mathbf{x}) - \rho(F)}{d(\mathbf{x})}, \tag{3.8}$$

the denominator $d(\mathbf{x})$ being some estimate of standard error for $s(\mathbf{x})$.

In what follows, we take $s(\mathbf{x})$ to be the Pearson correlation coefficient and

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

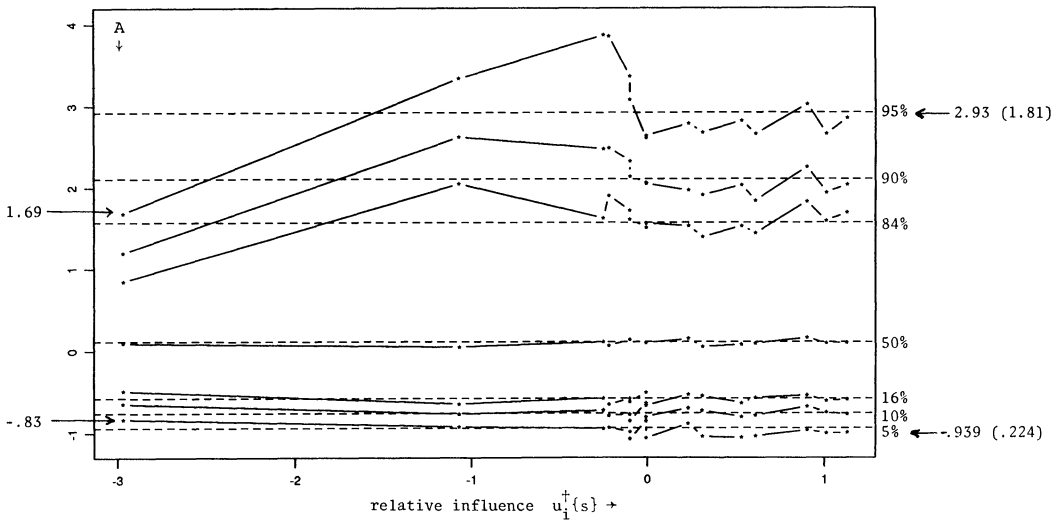

Fig. 5. Percentiles of $B = 1000$ bootstrap replications of the T-statistic (3.8), law school data: the broken horizontal lines indicate percentiles of all 1000 values T^{*b}; the broken curves are percentiles removing one point at a time from the data set, points removed in order of influence $u_i^\dagger\{s\}$, as in Fig. 4; the estimate $T^{*(0.95)} = 2.93$ has jackknife-after-bootstrap standard error 1.81, after correction for internal error

$$d(\mathbf{x}) = \{1 - s(\mathbf{x})^2\}/\sqrt{15} + 0.03. \tag{3.9}$$

If F is bivariate normal, then $\{1 - s(\mathbf{x})^2\}/\sqrt{15}$ is a reasonable approximation to the standard error of $s(\mathbf{x})$. The *ad hoc* constant 0.03 was added to stabilize the bootstrap distribution, which otherwise had a very heavy upper tail.

A central 90% confidence interval for ρ is

$$\rho \in [\, s(\mathbf{x}) - d(\mathbf{x})\, T^{(0.95)},\ s(\mathbf{x}) - d(\mathbf{x})\, T^{(0.05)} \,], \tag{3.10}$$

where $T^{(\alpha)}$ is the 100αth percentile of T. This is equivalent to the statement that T lies in $[\, T^{(0.05)},\ T^{(0.95)} \,]$ with 90% probability. However, formula (3.10) is unusable since the percentiles of T are unknown.

The bootstrap-t approach replaces $T^{(0.05)}$ and $T^{(0.95)}$ in formula (3.10) with their bootstrap estimates: each bootstrap sample $\mathbf{x}^{*b}$ yields a bootstrap replication of T,

$$T^{*b} = \frac{s(\mathbf{x}^{*b}) - s(\mathbf{x})}{d(\mathbf{x}^{*b})} \tag{3.11}$$

(where we have used $\rho(\hat{F}) = s(\mathbf{x})$, the sample correlation coefficient); then the percentiles $T^{*(0.05)}$ and $T^{*(0.95)}$ of the B bootstrap replications are substituted for $T^{(0.05)}$ and $T^{(0.95)}$ in expression (3.10). The $B = 1000$ bootstrap replications of the law school data give $T^{*(0.05)} = -0.939$ and $T^{*(0.95)} = 2.93$. Substituting these values, and $s(\mathbf{x}) = 0.776$ and $d(\mathbf{x}) = 0.133$, into expression (3.10) gives $\rho \in [0.388, 0.901]$ as the 90% bootstrap-t confidence interval for ρ.

Fig. 5 is the same as Fig. 4, except that now it is the percentiles of T^* (and of the deleted point values T_i^*) that are displayed. We see that the upper percentiles are quite unstable. The jackknife-after-bootstrap standard error for $T^{*(0.95)} = 2.93$ is 1.96, reduced to 1.81 after correction for internal error; see Section 6. The coefficient of variation for $T^{*(0.95)}$ is $0.62 = 1.81/2.93$, compared with $0.24 = 0.224/0.939$ for the lower percentile $T^{*(0.05)}$.

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

The bootstrap-t approach has excellent asymptotic properties as the sample size $n \to \infty$. As Hall (1988) shows, this is because $T = \{s(\mathbf{x}) - \rho(F)\}/d(\mathbf{x})$ approaches its limiting distribution at a high rate of convergence. Fig. 5 shows that the asymptotics have not yet taken over in this data set. Even small changes in F, from $\hat{F}$ to $\hat{F}_{(i)}$, expression (3.1), produce large changes in the distribution of T. Faced with this evidence, we might try another confidence interval approach, like the BC_a intervals of Efron (1987), or base our analysis on a more robust measure of correlation. Or, as suggested by Tibshirani (1988), we could try the bootstrap-t approach on a different scale, for example after making Fisher's $\tanh^{-1}$-transformation, and hope that the equivalent of Fig. 5 would show greater stability.

Remark 1. It is sometimes useful to apply the bootstrap in parametric problems, when traditional parametric methods of error assessment are too difficult to use or too approximate to trust. Jackknife-after-bootstrap methods can be applied within parametric families. Doing so shows the relation of lemma 1 to *importance sampling*; see Hammersley and Handscomb (1964).

Suppose that $\mathcal{F} = \{f_\eta(x), \eta \in N\}$ is a parametric family of density functions indexed by a parameter vector η and that we are in the one-sample situation, where the observed data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is an IID sample from some member of $\mathcal{F}$,

$$f_\eta \xrightarrow{\text{IID}} (x_1, x_2, \ldots, x_n) = \mathbf{x}. \tag{3.12}$$

Let $f_\eta(\mathbf{x}) \equiv \prod_{i=1}^n f_\eta(x_i)$ indicate the density of the whole sample. The parameter space N is a subset of k-dimensional Euclidean space. Given $\mathbf{x}$, we estimate η according to some rule $\hat{\eta} = \hat{\eta}(\mathbf{x})$. A parametric bootstrap sample is an IID sample of size n from $f_{\hat{\eta}}$,

$$f_{\hat{\eta}} \xrightarrow{\text{IID}} (x_1^*, x_2^*, \ldots, x_n^*) = \mathbf{x}^*. \tag{3.13}$$

The random variable $T(\mathbf{x}, F)$ can now be written as $T(\mathbf{x}, \eta)$. Definition (2.12) of a bootstrap statistic still applies,

$$\hat{\gamma}(\mathbf{x}) = \phi\,[\,T(\mathbf{x}^*, \hat{\eta})\,], \tag{3.14}$$

where $[\,T(\mathbf{x}^*, \hat{\eta})\,]$ indicates the distribution of $T(\mathbf{x}^*, \hat{\eta})$ with $\hat{\eta}$ fixed and $\mathbf{x}^*$ generated according to sample (3.13).

Suppose that we have generated B independent bootstrap samples according to rule (3.13), $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$. As in definition (2.13), we estimate the ideal value $\hat{\gamma}(\mathbf{x})$ by

$$\tilde{\gamma}(\mathbf{x}) = \phi\,[\,T(\mathbf{x}^{*b}, \hat{\eta}),\ \ b = 1, 2, \ldots, B\,], \tag{3.15}$$

where the bracketed term is the empirical distribution of the B bootstrap replications $T(\mathbf{x}^{*b}, \hat{\eta})$. However, formula (3.5) for estimating $\hat{\gamma}_{(i)}$ no longer makes sense. Its place is taken by an *importance sampling* formula.

Let $\hat{\eta}_{(i)} = \hat{\eta}(\mathbf{x}_{(i)})$ be the estimate of η based on the deleted point data set $\mathbf{x}_{(i)}$, equation (2.2). The bootstrap density ratio

$$R_i(\mathbf{x}^*) \equiv f_{\hat{\eta}_{(i)}}(\mathbf{x}^*)/f_{\hat{\eta}}(\mathbf{x}^*) \tag{3.16}$$

is assumed to be finite with probability 1 when sampling from $f_{\hat{\eta}_{(i)}}(\mathbf{x}^*)$. Then we have a standard importance sampling result.

Lemma 2. For any function $r(T)$, the expectation of $r\{T(\mathbf{x}^*, \hat{\eta}_{(i)})\}R_i(\mathbf{x}^*)$ under

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

bootstrap sampling (3.13) is the same as the expectation of $r\{T(\mathbf{x}^*, \hat{\boldsymbol{\eta}}_{(i)})\}$ under deleted point bootstrap sampling

$$f_{\hat{\boldsymbol{\eta}}_{(i)}} \xrightarrow{\text{IID}} (x_1^*, x_2^*, \ldots, x_n^*) = \mathbf{x}^*. \tag{3.17}$$

($\hat{\boldsymbol{\eta}}_{(i)}$ is considered fixed in both cases, only $\mathbf{x}^*$ varying.)

We want to estimate $\hat{\gamma}_{(i)} = \hat{\gamma}(\mathbf{x}_{(i)})$, the deleted point value of $\hat{\gamma}(\mathbf{x})$, from B bootstrap samples generated according to sample (3.13). Lemma 2 suggests the estimate

$$\tilde{\gamma}_{(i)} = \phi\,[\,T(\mathbf{x}^{*b}, \hat{\boldsymbol{\eta}}_{(i)}),\ \text{probabilities } R_i(\mathbf{x}^{*b})/B\,], \tag{3.18}$$

the bracketed term indicating the distribution putting probability $R_i(\mathbf{x}^{*b})/B$ on $T(\mathbf{x}^{*b}, \hat{\boldsymbol{\eta}}_{(i)})$. Remark L of Efron (1990a) discusses an improvement on equation (3.15) based on ratio estimation.

Parametric bootstrap calculations are often substantially more stable than their nonparametric counterparts. They trade off a reduction in variance for a possible increase in bias. Section 5 of Efron (1990a) re-does the bootstrap-t analysis (3.11), assuming a bivariate normal parametric family. The parametric analogue to Fig. 5 shows that the bootstrap-t statistic is now much closer to being pivotal.

4. CHOOSING AN ESTIMATOR FROM THE DATA

Suppose that the statistician is trying to choose from a family of possible estimators $s(\mathbf{x}, q)$. For example $s(\mathbf{x}, q)$ might be the q% trimmed mean for a real-valued data set $\mathbf{x}$, so $s(\mathbf{x}, 0)$ is the ordinary mean, $s(\mathbf{x}, 0.50)$ is the median and $s(\mathbf{x}, 0.25)$ is the 25% trimmed mean, the average of the middle 50% of the data. A reasonable selection method is to generate B bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$, compute the bootstrap variance estimators $v(q) = [\text{se}_{\text{boot}}\{s(\mathbf{x}, q)\}]^2$ for various choices of q and select $s(\mathbf{x}, q)$ corresponding to the smallest value of $v(q)$. How well determined is this choice? The jackknife-after-bootstrap method provides an answer, with no further bootstrap sampling required. This section illustrates the calculations in the context of an estimation problem from particle physics, described more fully in Efron (1988) and Hayes *et al.* (1989).

The tau particle is a heavy electron-like particle discovered in the 1970s by Martin Perl at the Stanford Linear Accelerator Center. Soon after its production the tau particle decays into various collections of more stable particles. About 86% of the time the decay involves just one charged particle. This rate, called decay$_1$, in Table 3, has been independently estimated 13 times, as shown at the top of the table. Each estimate represents a major research project involving several years of work. The mean of the 13 numbers is 85.962, the 10% trimmed mean is 85.947, etc., as shown just below the data. Estimated bootstrap variances are given for each estimator, all these being based on the same $B = 1000$ bootstrap samples $\mathbf{x}(1)^{*1}, \mathbf{x}(1)^{*2}, \ldots, \mathbf{x}(1)^{*1000}$, generated as in sample (2.7) from the decay$_1$ data set $\mathbf{x}(1) = (x_1(1), x_2(1), \ldots, x_{13}(1))'$.

The one-charged-particle event comprises four main decay modes, called ρ, π, e, and μ in Table 3, plus an uncertain catalogue of other events. There have been $n = 6$ independent measurements for the rate of occurrence of ρ, decay$_\rho$; $n = 7$ for decay$_\pi$; $n = 14$ for decay$_e$; $n = 19$ for decay$_\mu$. (A 20th observation for decay$_\mu$ has been excluded from consideration here. It is such an egregious outlier that it dominates the choice of an estimator if included.) Because of certain physical constraints,

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

TABLE 3
Tau data†

Decay_1 ($n=13$)	84.0	84.7	84.7	85.1	85.2	85.2	86.0	86.1	86.7	86.9
	87.2	87.8	87.9							
Means	85.962	85.947	85.892		85.877	85.846	85.785	86.000		
Variances	0.107	0.141	0.178		0.201	0.229	0.293	0.381		
Trims (%)	0	10	20		25	30	40	50		
Decay_ρ ($n=6$)	20.5	22.1	22.3	22.3	22.6	24.0				
Means	22.300	22.313	22.322		22.317	22.308	22.300	22.300		
Variances	0.178	0.178	0.182		0.168	0.157	0.159	0.159		
Decay_π ($n=7$)	8.0	9.0	9.9	10.0	10.7	11.7	11.8			
Means	10.160	10.220	10.243		10.221	10.193	10.086	10.000		
Variances	0.250	0.310	0.369		0.397	0.434	0.469	0.535		
Decay_e ($n=14$)	13.0	16.0	17.0	17.4	17.6	18.2	18.2	18.3	18.4	18.9
	19.0	19.1	20.4	22.4						
Means	18.136	18.209	18.240		18.257	18.268	18.264	18.250		
Variances	0.312	0.254	0.168		0.145	0.131	0.123	0.129		
Decay_μ ($n=19$)	12.9	15.0	17.1	17.4	17.5	17.6	17.7	17.7	17.8	18.0
	18.2	18.3	18.3	18.8	19.4	21.0	22.0	22.0	22.4	(35.0)
Means	18.374	18.454	18.156		18.066	18.016	18.000	18.000		
Variances	0.259	0.269	0.217		0.185	0.157	0.129	0.125		
$\Delta = \text{decay}_1 - (\text{decay}_\rho + \text{decay}_\pi + \text{decay}_e + \text{decay}_\mu)$										
Means	16.995	16.750	16.931		17.016	17.061	17.135	17.450		
Variances	1.106	1.152	1.114		1.096	1.108	1.173	1.329		
Trims (%)	0	10	20		25	30	40	50		

†13 independent measures of decay_1, the percentage of times that the tau particle decays into one charged particle, likewise 6, 7, 14 and 19 independent measurements of the decay modes decay_ρ, decay_π, decay_e and decay_μ; of particular interest is the difference $\Delta = \text{decay}_1 - (\text{decay}_\rho + \text{decay}_\pi + \text{decay}_e + \text{decay}_\mu)$; the *means* are trimmed means, with trim 0, 0.1, 0.2, 0.25, 0.3, 0.4 and 0.5 (the median); the *variances* are bootstrap variances obtained from $B = 1000$ bootstrap replications for each of five decay categories. Which estimator is preferred? The data are from Efron (1988), where the outlying value 35.0 for decay_μ was included in the analysis.

any one experiment provides only one estimate in the table: either an estimate for the composite rate decay_1, or for one of the four modes, decay_ρ, decay_π, decay_e, decay_μ.

The goal of Hayes *et al.* (1989) was to give a confidence interval for the difference parameter

$$\Delta \equiv \text{decay}_1 - (\text{decay}_\rho + \text{decay}_\pi + \text{decay}_e + \text{decay}_\mu). \tag{4.1}$$

The corresponding difference of the 25% trimmed mean was the estimator $\hat{\Delta}$ used in the bootstrap confidence interval construction; $\hat{\Delta} = 17.016$ for the data in Table 3.

The 25% trimmed mean was chosen on the basis of a preliminary bootstrap analysis: five independent bootstrap data sets $\{\mathbf{x}^{*b}(h),\ b = 1, 2, \ldots, 1000\}$ were generated as in expression (2.7), from the five original data vectors $\mathbf{x}(h)$ in Table 3, $h \equiv 1, \rho, \pi$, e, μ; these gave variance estimates $v(q, h)$ for the trimmed means $s\{\mathbf{x}(h), q\}$, q equal to 0%, 10%, 20%, 25%, 30%, 40%, 50%,

$$v(q, h) = \sum_{b=1}^{1000} \frac{\{s^{*b}(q, h) - s^{*\cdot}(q, h)\}^2}{B-1} \qquad (s^{*b}(q, h) \equiv s\{\mathbf{x}^{*b}(h), q\}). \tag{4.2a}$$

The sum

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

$$v(q, \Delta) \equiv \sum_h v(q, h) \tag{4.2b}$$

was used to estimate the variance of $\hat{\Delta}(q) = s\{\mathbf{x}(1), q\} - [s\{\mathbf{x}(\rho), q\} + s\{\mathbf{x}(\pi), q\} + s\{\mathbf{x}(e), q\} + s\{\mathbf{x}(\mu), q\}]$; the choice $q = 25\%$ minimized $v(q, \Delta)$ as seen at the bottom of Table 3, so the 25% trimmed mean was selected as the preferred estimator for subsequent calculations.

How well determined is the choice $q = 25\%$? Would we expect to derive nearly the same answer from another, independent, version of Table 3, or might we select a much different value of q? In fact, jackknife-after-bootstrap calculations show that the choice $q = 25\%$ is *not* well determined here. This is seen in Fig. 6, where $v(q, \Delta)$ is plotted as a function of q, along with the jackknife-after-bootstrap standard error interval $v(q, \Delta) \pm \check{se}(q, \Delta)$. These standard errors apply to the *differences* between $v(q, \Delta)$ for the various choices of q, as described later. We see that even the largest difference $v(0.50, \Delta) - v(0.25, \Delta)$ is less than one standard error greater than 0. The point here is not that the choice of $q = 25\%$ is foolish, but rather that it is not a choice strongly dictated by the data.

The remainder of this section describes the calculation of $\check{se}(q, \Delta)$. First consider the one-sample problem where we have just one data vector $\mathbf{x}$ (instead of five) and want to select from possible estimators $s(\mathbf{x}, q)$ on the basis of minimum bootstrap variance $v(q) = \Sigma_{b=1}^{B}\{s^{*b}(q) - s^{*\cdot}(q)\}^2/(B-1)$. Let $v_{(i)}(q)$ be the deleted point estimate (3.5), the empirical variance of the $s^{*b}(q)$ values corresponding to resampling vectors $\mathbf{P}^b$ with $P_i^b = 0$. Define $\mathbf{v}_{(i)} = (v_{(i)}(q_1), v_{(i)}(q_2), \ldots, v_{(i)}(q_K))$, where $q_1, q_2, \ldots, q_K$ are the allowed choices for q ($K = 7$ in Fig. 6). The jackknife estimate of covariance for $\mathbf{v} = (v(q_1), v(q_2), \ldots, v(q_K))$ is

$$\widehat{\mathbf{cov}}_{\text{jack}} = \frac{n-1}{n} \sum_{i=1}^{n} (\mathbf{v}_{(i)} - \mathbf{v}_{(\,)})'(\mathbf{v}_{(i)} - \mathbf{v}_{(\,)}), \tag{4.3}$$

this being the multivariate version of equation (3.6).

The $K \times K$ matrix $\widehat{\text{cov}}_{\text{jack}}$ tends to overestimate cov_{jack}, the ideal jackknife estimate that would be obtained from equation (4.3) if $B = \infty$. An approximate correction matrix, to be subtracted from $\widehat{\text{cov}}_{\text{jack}}$, is obtained as in formula (6.12) of Section 6,

$$\mathbf{cor} = \frac{(n-1)^2}{n} \frac{(e_n - 1)\tilde{\boldsymbol{\sigma}}^2}{B} \tag{4.4}$$

where $e_n = (1 - 1/n)^n$ and $\tilde{\boldsymbol{\sigma}}^2$ is the $K \times K$ matrix having elements

$$\tilde{\sigma}^2_{k_1, k_2} = \sum_b \{t^{*b}(q_{k_1}) - t^{*\cdot}(q_{k_1})\}\{t^{*b}(q_{k_2}) - t^{*\cdot}(q_{k_2})\}/(B-1), \tag{4.5}$$

$t^{*b}(q) \equiv \{s^{*b}(q) - 2s^{*\cdot}(q)\}s^{*b}(q)$. (The statistic $t^{*b}(q)$ arises from a Taylor expansion of the variance functional.) Correcting $\widehat{\text{cov}}_{\text{jack}}$ will turn out not to be very important for the tau data.

In comparing the bootstrap variances $v(q)$, we are more interested in their differences than their absolute values. The vector of differences of the $v(q)$ values from their mean is

$$\check{\mathbf{v}} = \mathbf{m}_K \mathbf{v}, \tag{4.6}$$

where $\mathbf{m}_K$ is the $K \times K$ projection matrix $\mathbf{I}_K - \mathbf{1}\mathbf{1}'/K$ ($\mathbf{I}$ is the identity matrix and

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

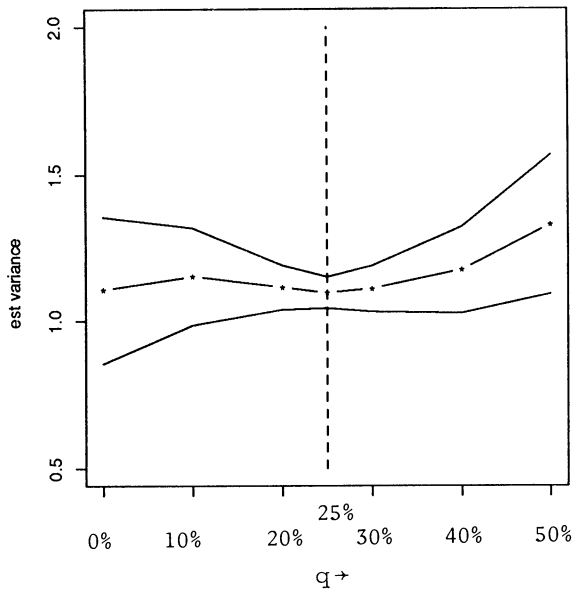

Fig. 6. Choice of an estimator for the tau data: —·—·, bootstrap variance $v(q, \Delta)$ for the q% trimmed mean estimator of the difference parameter Δ, equation (4.2b), q equal to 0%, 10%, 20%, 25%, 30%, 40% and 50%; ———, $v(q, \Delta) \pm \tilde{se}(q)$, where $\tilde{se}(q)$ is a jackknife-after-bootstrap estimate of the standard error of $v(q, \Delta)$, corrected for internal error (none of the variance estimates differ by as much as one standard error)

TABLE 4
Choice of estimators for the tau data†

q	$v(q, \Delta)$	$\tilde{se}(q, \Delta)$	$\tilde{se}(q, \Delta)$
0	1.106	0.261	0.251
0.1	1.152	0.175	0.167
0.2	1.114	0.082	0.076
0.25	1.096	0.059	0.053
0.3	1.109	0.084	0.078
0.4	1.173	0.157	0.148
0.5	1.329	0.251	0.238

†$v(q, \Delta)$ is the bootstrap variance estimate (4.2b); based on $B = 1000$ bootstrap replications for each of the five decay rate data sets; $\check{se}(q, \Delta)$ is the square root of the diagonal element of $\check{cov}(\Delta)$, equation (4.8); $\tilde{\check{se}}(q, \Delta)$ is the same quantity calculated without subtracting the correction for internal error. The $\check{se}(q, \Delta)$ are estimated standard errors for the differences between the $v(q, \Delta)$; see Fig. 6.

1 is the vector of K 1s). Correcting for internal errors, we estimate the covariance of $\check{\mathbf{v}}$ by

$$\check{\mathbf{cov}} \equiv \mathbf{m}_K(\widetilde{\mathbf{cov}}_{\text{jack}} - \mathbf{cor})\mathbf{m}_K. \tag{4.7}$$

Returning to the tau data, $\check{cov}(h)$ was calculated as in expression (4.7) for each set of data $\mathbf{x}(h)$, $h \equiv 1$, ρ, π, e, μ, giving the estimated covariance matrix for $\check{\mathbf{v}}(\Delta) = \Sigma_h \check{\mathbf{v}}(h)$

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

$$\widehat{\mathbf{cov}}(\Delta) = \sum_h \widehat{\mathbf{cov}}(h) \tag{4.8}$$

The entries $\check{se}(q, \Delta)$ in Table 4 are the square roots of the diagonal elements of $\widehat{\mathbf{cov}}(\Delta)$. The full curves in Fig. 6 are $v(q, \Delta) \pm \check{se}(q, \Delta)$. The uncorrected standard error estimates $\widehat{se}(q, \Delta)$ in Table 4 show that the correction for internal error is small in this case.

Remark 2. Including the 20th observation 35.0 in the decay$_\mu$ data set raises $v(0, \Delta)$, the variance estimate for the ordinary mean, to nearly 2.0 and raises $v(0.1, \Delta)$ to 1.4, without much changing $v(q, \Delta)$ for $q \geqslant 0.2$. With these values, the choice $q = 0.25$ looks considerably more convincing than in Fig. 6, but really is not, as an error analysis shows.

Remark 3. The methodology of this section offers, at least theoretically, a way to incorporate the data-based choice of an estimator into our estimate of its variability. Consider again the one-sample problem where we have available a range of estimators $s(\mathbf{x}, q)$, now letting q range over a continuum of possible values, for example $q \in [0, 0.5]$ in the trimmed mean case. The bootstrap estimate of variance $v(q)$ is minimized for some value $\hat{q}$, say $\hat{q} = q(\mathbf{x})$.

With $\hat{q}$ considered fixed, the jackknife influence function for $s(\mathbf{x}, \hat{q})$ is $u_i\{s(\mathbf{x}, \hat{q})\} = (n-1)\{s(\ , \hat{q}) - s(\mathbf{x}_{(i)}, \hat{q})\}$, $s(\ , \hat{q}) \equiv \Sigma_i s(\mathbf{x}_{(i)}, \hat{q})/n$. It is more realistic to take into account the variability of $q(\mathbf{x})$: $u_i[s\{\mathbf{x}, q(\mathbf{x})\}] = (n-1)[s(\ ) - s\{\mathbf{x}_{(i)}, q(\mathbf{x}_{(i)})\}]$, $s(\ ) \equiv \Sigma_i s\{\mathbf{x}_{(i)}, q(\mathbf{x}_{(i)})\}/n$. Here $q(\mathbf{x}_{(i)})$ is computed from the original bootstrap samples $\mathbf{x}^{*1}, \ldots, \mathbf{x}^{*B}$ by using formula (3.5). The variance estimate

$$\sum_i u_i[s\{\mathbf{x}, q(\mathbf{x})\}]^2/n(n-1)$$

takes into account the data-based choice of q.

5. DELTA METHOD FOR BOOTSTRAP STATISTICS

The delta method, or infinitesimal jackknife, is another approach to assessing influence functions and standard errors. This section discusses the application of the delta method to bootstrap statistics. In some special situations, this approach is more efficient than the jackknife in its use of a fixed number B of bootstrap replications.

The delta method, as it will be used here, applies to *functional statistics*, statistics $s(\mathbf{x})$ that are functions of the empirical distribution $\hat{F}$, say

$$s(\mathbf{x}) = S(\hat{F}). \tag{5.1}$$

The correlation and ratio statistics are functional, but the unbiased estimate of variance $s(\mathbf{x}) = \Sigma_i (x_i - \bar{x})^2/(n-1)$ is not: the data set which repeats each component of $\mathbf{x}$ twice yields the same $\hat{F}$ as $\mathbf{x}$, but a different value of $s(\mathbf{x})$. We assume below that the functional $S(F)$ is smoothly defined in a neighbourhood of $\hat{F}$; see section (6.3) of Efron (1982) and section 2.5 of Huber (1981).

Let $\hat{F}_{\epsilon, i}$ be a variant of the empirical distribution that puts extra probability on the ith point,

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

$$\hat{F}_{\epsilon,\,i}\text{: probability}\begin{cases}\dfrac{1-\epsilon}{n}+\epsilon & \text{on } x_i,\\[2ex] \dfrac{1-\epsilon}{n} & \text{on } x_j,\ j\neq i.\end{cases} \tag{5.2}$$

Keeping ϵ in the range $[-1/(n-1),\ 1]$ makes the probabilities (5.2) non-negative. The *delta method influence function* for S, or s, is defined as the derivative

$$U_i\{S\}=\frac{\partial S(\hat{F}_{\epsilon,\,i})}{\partial\epsilon}\bigg|_{\epsilon=0}=\lim_{\epsilon\to 0}\left\{\frac{S(\hat{F}_{\epsilon,\,i})-S(\hat{F})}{\epsilon}\right\}; \tag{5.3}$$

equation (5.3) is also called the empirical influence function (Mallows, 1974), or Jaeckel's (1972) infinitesimal jackknife influence function. The derivative in equation (5.3) can sometimes be evaluated theoretically, but it is usually easier just to substitute a small value of ϵ in the last expression.

The delta method estimate of standard error, for S or s, is

$$\text{se}_{\text{delta}}\{S\}\equiv\left(\sum_{i=1}^{n}U_i\{S\}^2/n^2\right)^{1/2}. \tag{5.4}$$

This definition agrees with the usual nonparametric delta method estimate of standard error when applied to functions of means like the ratio statistic $s(\mathbf{x})=\bar{z}/\bar{y}$; see section 6 of Efron (1981). The choice of denominator n^2 in expression (5.4), rather than $n(n-1)$ as in estimate (2.5), is a convention that makes this agreement perfect but needs to be kept in mind when comparing se_{delta} with se_{jack}. For the law school correlation, $\text{se}_{\text{delta}}=0.124$ compared with $\text{se}_{\text{jack}}=0.143$; for the bioequivalence ratio, $\text{se}_{\text{delta}}=0.098$ compared with $\text{se}_{\text{jack}}=0.105$.

We want to apply the delta method to bootstrap statistics $\hat{\gamma}(\mathbf{x})=\gamma(\hat{F})$, statistic (2.11). First we need a result relating ordinary bootstrap samples (2.7) to samples from $\hat{F}_{\epsilon,\,i}$,

$$\hat{F}_{\epsilon,\,i}\xrightarrow{\text{IID}}(x_1^*,\ x_2^*,\ \ldots,\ x_n^*)=\mathbf{x}^*. \tag{5.5}$$

There are n^n possible bootstrap samples $(x_1^*,\ x_2^*,\ \ldots,\ x_n^*)$, each of which has probability $\hat{f}(\mathbf{x}^*)=1/n^n$ under mechanism (2.7). Let $\hat{f}_{\epsilon,\,i}(\mathbf{x}^*)$ indicate the probability density of $\mathbf{x}^*$ under mechanism (5.5).

Lemma 3. The ratio of probability densities of a bootstrap sample $\mathbf{x}^*$, for mechanism (5.5) compared with mechanism (2.7), is

$$\frac{\hat{f}_{\epsilon,\,i}(\mathbf{x}^*)}{\hat{f}(\mathbf{x}^*)}=(1-\epsilon)^n\left(1+\frac{n\epsilon}{1-\epsilon}\right)^{nP_i}, \tag{5.6}$$

where $P_i=\#\{x_j^*=x_i\}/n$ as in equation (3.3).

The proof of lemma 3 is by direct computation from probabilities (5.2). By letting ϵ approach its lower limit $-1/(n-1)$, lemma 3 gives lemma 1.

We first consider bootstrap statistics $\hat{\gamma}(\mathbf{x})$, definition (2.12), of the *expectation form*

$$\hat{\gamma}(\mathbf{x})=\gamma(\hat{F})=E_{\hat{F}}[\,r\{T(\mathbf{x}^*,\ \hat{F})\}\,], \tag{5.7}$$

where $r(T)$ is a differentiable function of T, and $E_{\hat{F}}$ indicates the ordinary bootstrap expectation, with $\hat{F}$ fixed and $\mathbf{x}^*$ random as in mechanism (2.7). Form (5.7) looks rather

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

special, but it will lead to influence function expressions for all the bootstrap statistics in Table 1. The form $r\{T(\mathbf{x}^*, \hat{F})\}$ for the random variable in equation (5.7) is no more general than $T(\mathbf{x}^*, \hat{F})$ but is notationally convenient in the calculations that follow.

Theorem 1. The delta method influence function for a bootstrap statistic of expectation form (5.7) is

$$U_i\{\hat{\gamma}\} = E_{\hat{F}}\left[n^2(P_i - 1/n)\, r\{T(\mathbf{x}^*, \hat{F})\} + r'\{T(\mathbf{x}^*, \hat{F})\} \frac{\partial T(\mathbf{x}^*, \hat{F}_{\epsilon, i})}{\partial \epsilon}\bigg|_{\epsilon=0} \right]$$

$$= n^2 \operatorname{cov}_{\hat{F}}\{P_i, r^*\} + E_{\hat{F}}[(r'^*)\, U_i\{T(\mathbf{x}^*, \hat{F})\}]. \tag{5.8}$$

Proof. The second expression in equation (5.8) is just an abbreviated version of the first, with $r^* \equiv r\{T(x^*, \hat{F})\}$, $r'^* \equiv r'\{T(x^*, \hat{F})\}$, $U_i\{T(\mathbf{x}^*, \hat{F})\} \equiv \partial T(\mathbf{x}^*, \hat{F}_{\epsilon, i})/\partial\epsilon\big|_{\epsilon=0}$ and $\operatorname{cov}_{\hat{F}}$ indicating covariance under bootstrap sampling (2.7). Suppose that $b = 1, 2, \ldots, n^n$ indexes all possible bootstrap samples $\mathbf{x}^*$. According to lemma 3

$$\gamma(\hat{F}_{\epsilon, i}) = \sum_{b=1}^{n^n} r\{T(\mathbf{x}^{*b}, \hat{F}_{\epsilon, i})\}(1-\epsilon)^n \left(1 + \frac{n\epsilon}{1-\epsilon}\right)^{nP_i^b} \hat{f}(\mathbf{x}^{*b}). \tag{5.9}$$

Expression (5.8) for $U_i\{\hat{\gamma}\} = \partial\gamma(\hat{F}_{\epsilon, i})/\partial\epsilon\big|_0$ is obtained by applying standard differentiation formulae to equation (5.9).

Let us apply theorem 1 to the case $T(\mathbf{x}, F) = s(\mathbf{x}) - \theta(F)$ and $r(T) = T$, for which $\gamma(F) = E_F\{s(\mathbf{X}) - \theta(F)\}$, the bias of $s(\mathbf{x})$ as an estimate of $\theta(F)$. Then

$$\hat{\gamma}(\mathbf{x}) = E_{\hat{F}}\{s(\mathbf{x}^*)\} - \theta(\hat{F}), \tag{5.10}$$

the bootstrap estimate of bias. (If $s(\mathbf{x})$ is the usual nonparametric estimator $s(\mathbf{x}) = \theta(\hat{F})$, as with the correlation and ratio statistics, then $\hat{\gamma}(\mathbf{x})$ is an idealized version of estimate (2.9), with $B \to \infty$.) Since $r'(T) = 1$, and $\partial T(\mathbf{x}^*, \hat{F}_{\epsilon, i})/\partial\epsilon\big|_0 = -\partial\theta(\hat{F}_{\epsilon, i})/\partial\epsilon\big|_0 = -U_i\{\hat{\theta}\}$, theorem 1 gives

$$U_i\{\hat{\gamma}\} = n^2 \operatorname{cov}_{\hat{F}}\{P_i, s^*\} - U_i\{\hat{\theta}\}, \tag{5.11}$$

for the bootstrap bias estimate $\hat{\gamma} = E_{\hat{F}}\{s^*\} - \hat{\theta}$.

Applying the familiar calculus of the delta method, theorem 1 can be extended to cover all the bootstrap statistics shown in Table 1.

(a) Suppose that $\gamma(\hat{F})$ is a differentiable function of a vector statistic $\boldsymbol{\lambda}(\hat{F}) \equiv (\lambda_1(\hat{F}), \lambda_2(\hat{F}), \ldots, \lambda_k(\hat{F}))$, say

$$\gamma(\hat{F}) = C\{\boldsymbol{\lambda}(\hat{F})\}. \tag{5.12}$$

Then

$$U_i\{\hat{\gamma}\} = \mathbf{U}_i\{\hat{\boldsymbol{\lambda}}\}\, \hat{\nabla} \tag{5.13}$$

where $\mathbf{U}_i\{\hat{\boldsymbol{\lambda}}\} = (U_i\{\hat{\lambda}_1\}, \ldots, U_i\{\hat{\lambda}_k\})$ and $\hat{\nabla} = (\ldots, \partial C(\boldsymbol{\lambda})/\partial\lambda_k, \ldots)'_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}$.

(b) Suppose that for a given real number c we define the statistic $\pi_c(\hat{F}) = \operatorname{prob}_{\hat{F}}\{T(\mathbf{x}^*, \hat{F}) < c\}$. Let $\gamma^{(\alpha)}(\hat{F}) \equiv T(\mathbf{x}^*, \hat{F})^{(\alpha)}$, the 100αth bootstrap percentile of $T(\mathbf{x}, F)$. Then if $\hat{G}(t)$ indicates the bootstrap cumulative distribution

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

function (CDF) of $T(\mathbf{x}^*, \hat{F})$, and $\hat{g}(t)$ is the density corresponding to $\hat{G}(t)$, we have

$$U_i\{\hat{\gamma}^{(\alpha)}\} = -U_i\{\hat{\pi}_c\}/\hat{g}(c), \tag{5.14}$$

for $c = \hat{\gamma}^{(\alpha)} = \hat{G}^{-1}(\alpha)$. (See remark 5 concerning the definition of $\hat{g}(c)$.)

As an example of equation (5.13) we consider evaluating $U_i\{\hat{\gamma}\}$ for $\hat{\gamma}$ the bootstrap estimate of standard error of a statistic $s(\mathbf{x})$. Let $\mathbf{t}(\mathbf{x}) \equiv (s(\mathbf{x}), s(\mathbf{x})^2)$, $\boldsymbol{\lambda}(\hat{F}) \equiv E_{\hat{F}}\{t(\mathbf{x}^*)\} = (E_{\hat{F}}\{s^*\}, E_{\hat{F}}\{s^{*2}\})$, and $\gamma(\hat{F}) = C\{\boldsymbol{\lambda}(\hat{F})\}$, where $C(\boldsymbol{\lambda}) = (\lambda_2 - \lambda_1^2)^{1/2}$. Then $\gamma(\hat{F}) = [E_{\hat{F}}(s^{*2}) - (E_{\hat{F}}\{s^*\})^2]^{1/2}$, which is $\mathrm{sd}_{\mathrm{boot}}\{s\}$, estimate (2.8), for the ideal case $B \to \infty$.

Theorem 1 applied to $r\{T(\mathbf{x}, \hat{F})\} = t_i(\mathbf{x})$ for $i = 1, 2$ gives

$$\mathbf{U}_i\{\hat{\boldsymbol{\lambda}}\} = \mathrm{cov}_{\hat{F}}\{P_i, \mathbf{t}^*\}; \tag{5.15}$$

the second term in equation (5.8) vanishes here because $T(\mathbf{x}, F)$ is not a function of F. We compute $\hat{\boldsymbol{\nabla}} = (-2\hat{\lambda}_1, 1)'/2\hat{\gamma}$. Then equation (5.13) gives

$$U_i\{\hat{\gamma}\} = n^2 \,\mathrm{cov}_{\hat{F}}\{P_i, t^*_{\nabla}\} \tag{5.16}$$

where

$$t^*_{\nabla} \equiv \frac{s^{*2} - 2\hat{\lambda}_1 s^*}{2\hat{\gamma}} = \frac{s^*(s^* - 2E_{\hat{F}}\{s^*\})}{2\,\mathrm{sd}_{\hat{F}}\{s^*\}}. \tag{5.17}$$

We can estimate $U_i\{\hat{\gamma}\}$, equation (5.16), from B bootstrap replications of $s(\mathbf{x})$ in the obvious way, replacing the ideal values $\hat{\lambda}_1$ and $\hat{\gamma}$ in equation (5.17) by $\tilde{\lambda}_1$ and $\tilde{\gamma}$, equation (2.13), and by using the empirical covariance

$$\frac{n^2}{B}\sum_{b=1}^{B} P_i^b(t^{*b}_{\nabla} - t^{*\cdot}_{\nabla}), \tag{5.18}$$

$t^{*\cdot}_{\nabla} \equiv \Sigma_b t^{*b}_{\nabla}/B$. As explained in remark 4, there is an alternative estimator of $U_i\{\hat{\gamma}\}$ which usually makes more efficient use of the B bootstrap replications, namely

$$\tilde{\mathbf{U}}\{\hat{\gamma}\} = \mathbf{m}(\mathbf{pp}')^{-1}\mathbf{pt}^*_{\nabla}, \tag{5.19}$$

$\tilde{\mathbf{U}}\{\hat{\gamma}\}$ denoting the entire vector $(\tilde{U}_1\{\hat{\gamma}\}, \tilde{U}_2\{\hat{\gamma}\}, \ldots, \tilde{U}_n\{\hat{\gamma}\})'$. Here $\mathbf{p}$ is the $n \times B$ matrix of $\mathbf{P}^b$ vectors, $\mathbf{t}^*_{\nabla}$ is the $B \times 1$ vector of t^{*b}_{∇} values and $\mathbf{m}$ is the $n \times n$ projection matrix

$$\left.\begin{aligned} \mathbf{p} &= (\mathbf{P}^1, \mathbf{P}^2, \ldots, \mathbf{P}^b, \ldots, \mathbf{P}^B), \\ \mathbf{t}^*_{\nabla} &= (t^{*1}_{\nabla}, t^{*2}_{\nabla}, \ldots, t^{*B}_{\nabla})', \\ \mathbf{m} &= (\mathbf{I}_n - \mathbf{11}'/n), \end{aligned}\right\} \tag{5.20}$$

$\mathbf{I}_n$ being the $n \times n$ identity matrix and $\mathbf{1}$ being the vector of n 1s. Likewise, an efficient estimate of the influence function (5.11) for the bootstrap bias is

$$\tilde{\mathbf{U}}\{\hat{\gamma}\} = \mathbf{m}(\mathbf{pp}')^{-1}\mathbf{ps}^* - \mathbf{U}\{\hat{\theta}\}, \tag{5.21}$$

$\mathbf{U}\{\hat{\theta}\} \equiv (U_1\{\hat{\theta}\}, \ldots, U_i\{\hat{\theta}\})'$.

Table 5, which concerns the bioequivalence ratio statistic, compares the delta method and jackknife influence functions for the bootstrap bias estimate, and also for the

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

TABLE 5

Estimated influence functions and standard errors of the bootstrap bias estimate and bootstrap standard error estimate, for the bioequivalence ratio statistic†

	$u_i^\dagger\{s\}$	*Results for bootstrap bias estimate 0.0053*				*Results for bootstrap standard error estimate 0.104*				
		Delta method		*Jackknife*		*Delta method*		*Jackknife*		
		$\tilde{U}_i\{\hat{\gamma}\}$	$(\pm)$	$\tilde{u}_i\{\hat{\gamma}\}$	$(\pm)$	$\tilde{U}_i\{\hat{\gamma}\}$	$(\pm)$	$\tilde{u}_i\{\hat{\gamma}\}$	$(\pm)$	
	−1.14	−0.0111	(0.0046)	−0.0479	(0.0297)	−0.054	(0.0220)	−0.048	(0.0263)	
	−1.12	−0.0043	(0.0053)	0.0281	(0.0297)	−0.102	(0.0225)	−0.098	(0.0263)	
	−0.46	−0.0160	(0.0043)	−0.0068	(0.0297)	−0.056	(0.0219)	−0.076	(0.0263)	
	−0.31	−0.0066	(0.0040)	−0.0017	(0.0297)	−0.077	(0.0185)	−0.101	(0.0263)	
	−0.12	−0.0124	(0.0040)	0.0056	(0.0297)	−0.041	(0.0176)	−0.056	(0.0263)	
	0.33	0.0146	(0.0063)	−0.0094	(0.0297)	0.039	(0.0239)	0.013	(0.0263)	
B	1.37	0.0566	(0.0081)	0.0507	(0.0297)	0.236	(0.0457)	0.259	(0.0263)	B
C	1.46	−0.0208	(0.0041)	−0.0186	(0.0297)	0.055	(0.0278)	0.107	(0.0263)	C
	$\widetilde{\mathrm{se}}\{\hat{\gamma}\}$ [corrected]	0.0083	[0.0081]	0.0105	[0]	0.036	[0.035]	0.044	[0.043]	

†The $\pm$ values for the influence functions reflect the limitations of using only $B = 1000$ bootstrap replications, as explained in Section 6. The delta method influence functions are more accurate than the jackknife influence functions (have smaller $\pm$ values) for the bias estimate, but not for the standard error estimates.

bootstrap standard error. Formulae (5.21) and (5.19) give $\tilde{U}_i\{\hat{\gamma}\}$. There is a component of error in both $\tilde{U}_i$ and $\tilde{u}_i$ that comes from using only $B = 1000$ bootstrap replications, rather than letting $B \to \infty$. These errors are indicated by the $\pm$ values in Table 5, as derived in Section 6. For the bootstrap bias estimate, but not for the bootstrap standard error estimate, the $\pm$ values are much smaller for the delta method than for the jackknife. Note that the bootstrap bias estimate for the ratio data, $\hat{\gamma} = 0.0053$, is less than one estimated standard error away from 0, even after correcting

$$\widetilde{\mathrm{se}}_{\mathrm{delta}}\{\hat{\gamma}\} = \Big(\sum_i \tilde{U}_i\{\hat{\gamma}\}^2/n^2\Big)^{1/2}$$

for the $\pm$ error component, from 0.0083 down to 0.0081.

Results (5.13) and (5.14) combine to give delta method influence functions for percentile statistics like length and shape in Table 1. Let $T(\mathbf{x}, F) = s(\mathbf{x})$ and $r_c(T)$ equal 1 or 0 as $T < c$ or $T \geqslant c$. Then $\hat{\pi}_c = \pi_c(\hat{F}) = E_{\hat{F}}\{r_c^*\}$, where r_c^* equals 1 or 0 as $s^* < c$ or $s^* \geqslant c$. Theorem 1 gives $U_i\{\hat{\pi}_c\} = n^2 \operatorname{cov}_{\hat{F}}\{P_i, r_c^*\}$. Then the length statistic $\hat{\gamma} = \hat{\gamma}^{(0.95)} - \hat{\gamma}^{(0.05)}$ has influence function

$$U_i\{\hat{\gamma}\} = n^2 \operatorname{cov}_{\hat{F}}\left[P_i, \frac{r^*_{c(0.05)}}{\hat{g}\{c(0.05)\}} - \frac{r^*_{c(0.95)}}{\hat{g}\{c(0.95)\}}\right] \tag{5.22}$$

according to results (5.13) and (5.14), where $c(\alpha) \equiv \hat{G}^{-1}(\alpha)$. Likewise the shape statistic $\hat{\gamma} = \log\{(\hat{\gamma}^{(0.95)} - \hat{\gamma}^{(0.5)})/(\hat{\gamma}^{(0.5)} - \hat{\gamma}^{(0.05)})\}$ has influence function

$$U_i\{\hat{\gamma}\} = n^2 \operatorname{cov}_{\hat{F}}\left[P_i - \frac{r^*_{c(0.05)}}{\hat{g}\{c(0.05)\}(\hat{\gamma}^{(0.5)} - \hat{\gamma}^{(0.05)})} + \frac{r^*_{c(0.5)}}{\hat{g}\{c(0.5)\}}\left(\frac{1}{\hat{\gamma}^{(0.5)} - \hat{\gamma}^{(0.05)}} + \frac{1}{\hat{\gamma}^{(0.95)} - \hat{\gamma}^{(0.5)}}\right) - \frac{r^*_{c(0.95)}}{\hat{g}\{c(0.95)\}(\hat{\gamma}^{(0.95)} - \hat{\gamma}^{(0.5)})}\right]. \tag{5.23}$$

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

Remark 4. Estimation formula (5.19) for $\tilde{U}_i\{\hat{\gamma}\}$ comes from the *bootstrap influence function*, or bootstrap Hajek projection, for a function $S(\mathbf{P})$ of the proportion vector $\mathbf{P}$, equation (3.3). The bootstrap replications of any random variable $T(\mathbf{x}^*, \hat{F})$ that is invariant under permutations of the co-ordinates of $\mathbf{x}^*$ can be expressed as $T(\mathbf{x}^*, \hat{F}) = S(\mathbf{P})$, e.g. $\bar{x}^* = \mathbf{x}'\mathbf{P} \equiv S(\mathbf{P})$. This includes all the examples of this paper, except for the multisample problem considered in Section 4. Functions $S(\mathbf{P})$ have an orthogonal decomposition in the bootstrap probability space,

$$S(\mathbf{P}) = m + \mathbf{P}'\mathbf{a} + \epsilon(\mathbf{P}), \tag{5.24}$$

where $m = E_{\hat{F}}\{S(\mathbf{P})\}$, $\mathbf{a} = n^2 \operatorname{cov}_{\hat{F}}\{\mathbf{P}, S(\mathbf{P})\}$ and the remainder term $\epsilon(\mathbf{P})$ is orthogonal to every linear function of $\mathbf{P}$: $E_{\hat{F}}\{(M + \mathbf{P}'\mathbf{A})\,\epsilon(\mathbf{P})\} = 0$. This is derived in Section 3 of Efron (1990b), where $\mathbf{a}$ is called the bootstrap influence function for $S(\mathbf{P})$. For $S(\mathbf{P}) = t_\nabla$ as in equation (5.16), the ith component of $\mathbf{a}$ equals $U_i\{\hat{\gamma}\}$, since both equal $n^2 \operatorname{cov}_{\hat{F}}\{P_i, S(\mathbf{P})\}$.

The sum of the first two terms of equation (5.24), $S_{\text{lin}}(\mathbf{P}) \equiv m + \mathbf{P}'\mathbf{a}$, minimizes the expected squared residual $E_{\hat{F}}\{S(\mathbf{P}) - (M + \mathbf{P}'\mathbf{A})\}^2$ among all choices of the constant M and vector $\mathbf{A}$. This suggests that we use ordinary least squares if we want to estimate m and $\mathbf{a}$ from B bootstrap replications $(\mathbf{P}^b, S(\mathbf{P}^b))$, $b = 1, 2, \ldots, B$. In fact, formula (5.19) is the ordinary least squares estimate of $\mathbf{a}$, as explained in Efron (1990b), sections 3 and 6, which also discusses the advantage of formula (5.19) over formula (5.18).

The amount of advantage depends on the linearity of $S(\mathbf{P})$, as measured by

$$R^2 \equiv \operatorname{var}_{\hat{F}}\{S_{\text{lin}}(\mathbf{P})\} / \operatorname{var}_{\hat{F}}\{S(\mathbf{P})\}. \tag{5.25}$$

The advantage is much bigger for the bootstrap bias estimate in Table 5, $S(\mathbf{P}) = s^*$, $R^2 = 0.970$, than for the bootstrap standard error, $S(\mathbf{P}) = t^*_\nabla$, equation (5.17), $R^2 = 0.154$.

Remark 5. Expression (5.14) looks awkward since $\hat{g}(c)$ is supposed to be a density function for the bootstrap variate $T(\mathbf{x}^*, \hat{F})$, which is always discrete. In practice all that we need is the *average* density of $T(\mathbf{x}^*, \hat{F})$ over small intervals of t. 'Small' means intervals $[t_1, t_2]$ such that $\hat{G}(t_2) - \hat{G}(t_1) = O(1/n)$. This amounts to taking ϵ of order $O(1/n)$ in the approximation $U_i\{\hat{\gamma}^{(\alpha)}\} \doteq \{\gamma^{(\alpha)}(\hat{F}_{\epsilon, i}) - \gamma^{(\alpha)}(\hat{F})\}/\epsilon$. We know that $O(1/n)$ is sufficiently small because of the generally good performance of the jackknife, which uses $\epsilon = -1/(n-1)$. The bootstrap CDF $\hat{G}$ has so many support points, typically $O(4^n/\sqrt{n})$, that it appears almost continuous over intervals of bootstrap probability $O(1/n)$. See the comment following equations (5.20) concerning the numerical computation of $\hat{g}(c)$ involved in Table 2.

Remark 6. It is interesting to compare $U_i\{\hat{\gamma}\}$ with $u_i\{\hat{\gamma}\}$ in the simple case $\hat{\gamma} \equiv E_{\hat{F}}\{s(\mathbf{x}^*)\}$, where the comparison can be made explicit. Let

$$\begin{aligned} b(j) &\equiv \operatorname{prob}\left\{P_i = \frac{j}{n}\right\} = \binom{n}{j}\left(\frac{1}{n}\right)^j\left(1 - \frac{1}{n}\right)^{n-j}, \\ e_i(j) &\equiv E_{\hat{F}}\{s(\mathbf{x}^*) \mid P_i = j/n\}. \end{aligned} \tag{5.26}$$

Then according to theorem 1

$$U_i\{\hat{\gamma}\} = n^2 \sum_{j=0}^{n} b(j)\left(\frac{j-1}{n}\right) e_i(j), \tag{5.27}$$

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

compared with

$$u_i\{\hat{\gamma}\} = \text{constant} - (n-1)\, e_i(0), \tag{5.28}$$

since $\hat{\gamma}_{(i)} = e_i(0)$. For $n=8$, $U_i\{\hat{\gamma}\} = \Sigma_{j=0}^{8} w(j)\, e_i(j)$, where the weights $w(j)$ are

j	0	1	2	3	4	5	6	7	8
$w(j)$	−2.75	0	1.87	0.898	0.240	0.037	0.003	0.000	0.000

(5.29)

compared with the weights $-7, 0, 0, 0, \ldots$ for $u_i\{\hat{\gamma}\}$, equation (5.28).

It looks as though $U_i\{\hat{\gamma}\}$ uses more of the bootstrap data since it involves all the conditional expectations $e_i(j)$ rather than just $e_i(0)$. The actual differences between $U_i\{\hat{\gamma}\}$ and $u_i\{\hat{\gamma}\}$ tend to be moderately small; see chapter 6 of Efron (1982). For *linear* functional statistics s, those for which $R^2 = 1$, it is easy to show that $U_i\{\hat{\gamma}\} = u_i\{\hat{\gamma}\}$, and that both equal a_i, the bootstrap influence function. Section 7 of Efron (1983) says more about the conditional bootstrap expectations $e_i(j)$.

Remark 7. The orthogonal expansion (5.24) and theorem 1 relate as follows: the bootstrap influence function a_i for $s(\mathbf{x}^*) = S(\mathbf{P})$ equals the delta method influence function $U_i\{\hat{\gamma}\}$ for $\hat{\gamma} = E_{\hat{F}}\{s^*\}$.

Remark 8. Theorem 2 of Efron (1990a) discusses delta-after-bootstrap computations in parametric families. The parametric analogue of theorem 1, equation (5.8), is connected with Stein's (1981) lemma for normal theory risk estimation.

6. INTERNAL ERRORS

The estimated influence functions based on B bootstrap replications $\tilde{u}_i\{\hat{\gamma}\}$ and $\tilde{U}_i\{\hat{\gamma}\}$ differ from the ideal values $u_i\{\hat{\gamma}\}$ and $U_i\{\hat{\gamma}\}$ that would be obtained if $B \to \infty$. The standard deviations of these differences, called *internal errors*, are the '$\pm$' quantities in Tables 2 and 5. Formulae for the internal errors are developed in this section. These calculations have a familiar appearance because the bootstrap data $(\mathbf{P}^b, s^{*b})$, $b = 1, 2, \ldots, B$, is an IID sequence of pairs drawn from the bootstrap distribution $(\mathbf{P}, s^*)$.

We begin with the simple situation where $\hat{\gamma} = E_{\hat{F}}\{s(\mathbf{x}^*)\}$. In this case a deleted point value of $\hat{\gamma}$, $\hat{\gamma}_{(i)} = E_{\hat{F}}\{s^* \mid P_i = 0\}$, is estimated by the average of s^{*b} over those bootstrap samples missing x_i,

$$\tilde{\gamma}_{(i)} = s^{*\cdot}_{(i)} = \sum_{b=1}^{B} I_i^b s^{*b} \Big/ \sum_{b=1}^{B} I_i^b \qquad \left(I_i^b \equiv \begin{cases} 1 & \text{if } P_i^b = 0, \\ 0 & \text{if } P_i^b > 0 \end{cases} \right). \tag{6.1}$$

Letting $\mathbf{s}^{*\cdot}_{(\cdot)} \equiv (s^{*\cdot}_{(1)}, s^{*\cdot}_{(2)}, \ldots, s^{*\cdot}_{(n)})'$, the estimated jackknife influence function $\tilde{\mathbf{u}}\{\hat{\gamma}\} \equiv (\tilde{u}_1\{\hat{\gamma}\}, \tilde{u}_2\{\hat{\gamma}\}, \ldots, \tilde{u}_n\{\hat{\gamma}\})'$ is given by

$$\tilde{\mathbf{u}}\{\hat{\gamma}\} = -(n-1)\mathbf{m}\mathbf{s}^{*\cdot}_{(\cdot)}, \tag{6.2}$$

equations (3.6), with $\mathbf{m}$ the projection matrix (5.20). The delta method influence function is estimated by

$$\tilde{\mathbf{U}}\{\hat{\gamma}\} = \mathbf{m}(\mathbf{p}\mathbf{p}')^{-1}\mathbf{p}\mathbf{s}^*; \tag{6.3}$$

compare equation (5.19).

We want to estimate the internal standard errors of $\tilde{u}_i\{\hat{\gamma}\}$ and $\tilde{U}_i\{\hat{\gamma}\}$, the errors due

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

to using only B bootstrap replications, $B = 1000$ in this paper. In other words we want the variability of $\tilde{u}_i\{\hat{\gamma}\}$ and $\tilde{U}_i\{\hat{\gamma}\}$ arising from the Monte Carlo choice of the bootstrap vectors $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$, with the original data $\mathbf{x}$ held fixed. To this end we apply the jackknife in the framework of bootstrap sampling, deleting one pair $(\mathbf{P}^b, s^{*b})$ at a time from the entire bootstrap sample of B pairs.

Lemma 4. Let $\tilde{\mathbf{u}}^{(b)}$ and $\tilde{\mathbf{U}}^{(b)}$ denote $\tilde{\mathbf{u}}\{\hat{\gamma}\}$ and $\tilde{\mathbf{U}}\{\hat{\gamma}\}$ calculated from the $B-1$ bootstrap pairs $(\mathbf{P}^c, s^{*c})$, $c \neq b$. Then

$$\begin{pmatrix} \tilde{\mathbf{u}}^{(b)} \\ \tilde{\mathbf{U}}^{(b)} \end{pmatrix} - \begin{pmatrix} \tilde{\mathbf{u}}\{\hat{\gamma}\} \\ \tilde{\mathbf{U}}\{\hat{\gamma}\} \end{pmatrix} \doteq \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} \begin{pmatrix} M_3^b \\ M_4^b \end{pmatrix}. \tag{6.4}$$

Here M_1 and M_2 are $n \times n$ matrices, M_1 diagonal,

$$\begin{aligned} &M_1 = \mathbf{m}\,\mathrm{diag}((n-1)/B_i) \qquad \left(B_i \equiv \sum_{b=1}^{B} I_i^b, \text{ for } i = 1, 2, \ldots, n\right), \\ &M_2 = -\mathbf{m}(\mathbf{p}\mathbf{p}')^{-1}; \end{aligned} \tag{6.5}$$

M_3^b and M_4^b are the $n \times 1$ column vectors

$$M_3^b = (\ldots, I_i^b \cdot (s^{*b} - s_{(i)}^{*\cdot}), \ldots)', \quad M_4^b = \mathbf{P}^b \hat{\epsilon}^b, \tag{6.6}$$

where $\hat{\epsilon}^b$ is the bth component of the residual vector $\hat{\boldsymbol{\epsilon}}$,

$$\hat{\boldsymbol{\epsilon}} = (\mathbf{I}_b - \mathbf{p}'(\mathbf{p}\mathbf{p}')^{-1}\mathbf{p})\mathbf{s}^*. \tag{6.7}$$

Proof. A familiar matrix identity (see equation (3.12) of Efron (1982)) gives

$$\tilde{\mathbf{U}}^{(b)} - \tilde{\mathbf{U}}\{\hat{\gamma}\} = -\mathbf{m}\{(\mathbf{p}\mathbf{p}')^{-1}\mathbf{P}^b\hat{\epsilon}^b\}/\{1 - \mathbf{P}^{b\prime}(\mathbf{p}\mathbf{p}')^{-1}\mathbf{P}^b\} = M_2 M_4^b\{1 + O_p(1/B)\}. \tag{6.8}$$

The difference between $s_{(i)}^{*\cdot}$, equation (6.1), and the same quantity computed excluding the pair $(\mathbf{P}^b, s^{*b})$ is

$$s_{(i)}^{*\cdot(b)} - s_{(i)}^{*\cdot} = s_{(i)}^{*\cdot}\left\{\frac{1 - I_i^b s^b / s_{(i)}^{*\cdot} B_i}{1 - I_i^b / B_i} - 1\right\} = -\frac{I_i^b}{B_i}(s^{*b} - s_{(i)}^{*\cdot})\left\{1 + O_p\left(\frac{1}{B}\right)\right\}, \tag{6.9}$$

so

$$\tilde{\mathbf{u}}^{(b)} - \tilde{\mathbf{u}}\{\hat{\gamma}\} = -(n-1)\mathbf{m}(\mathbf{s}_{(\cdot)}^{*\cdot(b)} - \mathbf{s}_{(i)}^{*\cdot}) = M_1 M_3^b\{1 + O_p(1/B)\}. \quad \square \tag{6.10}$$

The $2n \times 1$ vector $(\tilde{\mathbf{u}}\{\hat{\gamma}\}', \tilde{\mathbf{U}}\{\hat{\gamma}\})'$ can be thought of as a random variable determined by the bootstrap data $\{(P^b, s^{*b}), b = 1, 2, \ldots, B\}$, with the original data $\mathbf{x}$ held fixed. The internal covariance of this vector, its covariance under the random choice of the bootstrap data, can be estimated from lemma 4 and Tukey's jackknife covariance formula,

$$\mathrm{cov}_{\mathrm{intern}}\begin{pmatrix} \tilde{\mathbf{u}}\{\hat{\gamma}\} \\ \tilde{\mathbf{U}}\{\hat{\gamma}\} \end{pmatrix} = \begin{pmatrix} M_1 M_3 M_3' M_1' & M_1 M_3 M_4' M_2 \\ M_2 M_4 M_3' M_1' & M_2 M_4 M_4' M_2 \end{pmatrix}, \tag{6.11}$$

M_3 and M_4 being the $n \times B$ matrices with bth columns M_3^b, M_4^b; see equation (3.13) of Efron (1982). (The right-hand side of equation (6.11) deviates from Tukey's definition by a factor $1 + O_p(1/B)$, which is negligible compared with the error $1 + O_p(1/\sqrt{B})$ of the jackknife formula for estimating the covariance matrix.) The internal covariance formula for $\hat{\gamma} = E_{\hat{F}}\{s^*\}$ also applies to the bootstrap bias estimate

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

TABLE 6

Internal standard error '±' for the delta method and jackknife influence functions, and internal correlation between $\tilde{U}_i\{\hat{\gamma}\}$ and $\tilde{u}_i\{\hat{\gamma}\}$, for the bioequivalence ratio statistic, from formula (6.11)†

	$u_i^\dagger\{s\}$	*Results for bootstrap bias estimate* $\tilde{U}_i\pm$	$\tilde{u}_i\pm$	*(correlation)*	*Results for bootstrap standard error estimate* $\tilde{U}_i\pm$	$\tilde{u}_i\pm$	*(correlation)*	
	−1.14	0.0046	0.0292	(0.12)	0.0220	0.0269	(0.75)	
	−1.12	0.0053	0.0316	(0.20)	0.0225	0.0290	(0.80)	
	−0.46	0.0043	0.0310	(0.16)	0.0219	0.0275	(0.78)	
	−0.31	0.0040	0.0311	(0.17)	0.0185	0.0268	(0.78)	
	−0.12	0.0040	0.0311	(0.09)	0.0176	0.0270	(0.69)	
	0.33	0.0063	0.0326	(0.16)	0.0239	0.0319	(0.63)	
B	1.37	0.0081	0.0275	(0.40)	0.0457	0.0289	(0.92)	B
C	1.46	0.0041	0.0293	(−0.09)	0.0278	0.0293	(0.76)	C
Table 5			0.0297			0.0263		

†$B=1000$. The ± values for $\tilde{U}_i$, but not $\tilde{u}_i$, are as in Table 5; $\tilde{U}_i\{\hat{\gamma}\}$ and $\tilde{u}_i\{\hat{\gamma}\}$ are highly correlated for the bootstrap standard error estimate $\hat{\gamma}$, but not for the bootstrap bias estimate $\hat{\gamma}$.

$\hat{\gamma}=E_{\hat{F}}\{s^*\}-\hat{\theta}$ since $\tilde{\mathbf{u}}\{\hat{\gamma}\}=\tilde{\mathbf{u}}\{E_{\hat{F}}s^*\}-\mathbf{u}\{\hat{\theta}\}$ and $\mathbf{u}\{\hat{\theta}\}$ is a fixed vector with covariance 0 under bootstrap sampling; likewise $\tilde{\mathbf{U}}\{\hat{\gamma}\}=\tilde{\mathbf{U}}\{E_{\hat{F}}s^*\}-\mathbf{U}\{\hat{\theta}\}$. Note that formula (6.11) extends to bootstrap statistics of form (5.12), in particular to the bootstrap estimate of standard error, by replacing $\mathbf{s}^*$ in definitions (6.6) and (6.7) with $\mathbf{t}^*_\nabla$, as defined in equation (5.17).

Table 6 reports the application of equation (6.11) to the bioequivalence ratio statistic, for the bootstrap bias estimate $\hat{\gamma}$ and also the bootstrap standard error $\hat{\gamma}$. The square root of the diagonal elements of $\text{cov}_{\text{intern}}$ are the ± values. Those for $\mathbf{U}_i$ agree with the ± values in Table 5, but the values for $\tilde{u}_i$ are different.

The ± values for $\tilde{u}_i$ in Table 5, actually just a single value for each of the two bootstrap statistics, were obtained from a simplified version of equation (6.11),

$$\text{cov}_{\text{intern}}\{\tilde{\mathbf{u}}(\hat{\gamma})\}\doteq \mathbf{m}(n-1)^2(e_n-1)\tilde{\sigma}^2/B, \tag{6.12}$$

$e_n=(1-1/n)^{-n}$, $\tilde{\sigma}^2\equiv\sum_{b=1}^{B}(s^{*b}-s^{*\cdot})^2/(B-1)$; approximation (6.12) is derived from $\text{cov}_{\text{intern}}\{\tilde{\mathbf{u}}\{\hat{\gamma}\}\}=M_1M_3M_3'M_1'$, equation (6.11), by using the following approximations:

$$\left.\begin{aligned} B_i &\doteq e_n^{-1}B,\\ \sum_b I_i^b(s^{*b}-s^{*\cdot}_{(i)})^2/B_i &\doteq \tilde{\sigma}^2,\\ \sum_b I_i^bI_j^b(s^{*b}-s^{*\cdot}_{(i)})(s^{*b}-s^{*\cdot}_{(j)})/\sum_b I_i^bI_j^b &\doteq \tilde{\sigma}^2 \qquad j\neq i. \end{aligned}\right\} \tag{6.13}$$

Table 6 shows that approximation (6.12) performs reasonably well.

The estimated influence functions $\tilde{\mathbf{u}}\{\hat{\gamma}\}$ and $\tilde{\mathbf{U}}\{\hat{\gamma}\}$, equations (6.2) and (6.3), are nearly unbiased for their ideal values $\mathbf{u}\{\hat{\gamma}\}$ and $\mathbf{U}\{\hat{\gamma}\}$,

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

$$E_{\hat F}\{\tilde{\mathbf{u}}\{\hat\gamma\}\} = \mathbf{u}\{\hat\gamma\}\{1 + O(1/B)\},$$
$$E_{\hat F}\{\tilde{\mathbf{U}}\{\hat\gamma\}\} = \mathbf{U}\{\hat\gamma\}\{1 + O(1/B)\}. \tag{6.14}$$

(As before, the expectations $E_{\hat F}$ are over the choice of the bootstrap data $(\mathbf{P}^b, s^{*b})$, $b = 1, \ldots, B$, with the original data $\mathbf{x}$ fixed. We could use the notation E_{intern} instead of $E_{\hat F}$.) A standard computation gives

$$E_{\hat F}\{\tilde{\text{se}}_{\text{jack}}\{\hat\gamma\}^2\} = \text{se}_{\text{jack}}\{\hat\gamma\}^2 + \frac{\text{trace}(\text{cov}_{\text{intern}}\{\tilde{\mathbf{u}}\})}{n(n-1)} + O\left(\frac{1}{B^2}\right)$$
$$\doteq \text{se}_{\text{jack}}\{\hat\gamma\}^2 + \frac{(n-1)^2}{n}\,\frac{(e_n - 1)\tilde\sigma^2}{B} + O\left(\frac{1}{B^2}\right), \tag{6.15}$$

the second formula coming from approximation (6.12). The $O(1/B)^2$ term is negligible compared with $O(1/B)$ of the other terms. Likewise

$$E_{\hat F}\{\tilde{\text{se}}_{\text{delta}}\{\hat\gamma\}^2\} = \text{se}_{\text{delta}}\{\hat\gamma\}^2 + \text{trace}(\text{cov}_{\text{intern}}\{\tilde{\mathbf{U}}\})/n^2 + O(1/B^2). \tag{6.16}$$

The corrected values for se_{delta} in Table 5 were obtained from

$$[\,\tilde{\text{se}}^2_{\text{delta}} - \text{trace}(\text{cov}_{\text{intern}}\{\tilde{\mathbf{U}}\}/n^2)\,]^{1/2},$$

and similarly for se_{jack} (using the second line of approximation (6.15)).

Internal error analyses can be done for almost any bootstrap statistic. They are no more than a standard error analysis, performed on the IID bootstrap sequence $(\mathbf{P}^b, s^{*b})$, $b = 1, 2, \ldots, B$. Suppose, for instance, that we are interested in the internal covariance analysis for $\hat\gamma = C(\hat{\boldsymbol\lambda})$, where $C(\boldsymbol\lambda)$ is a smooth function of a vector of percentiles of some random variable $T(\mathbf{x}, F)$, say $\boldsymbol\lambda = (T^{(\alpha_1)}, T^{(\alpha_2)}, \ldots, T^{(\alpha_k)})'$. This was the case in Table 1, where $T(\mathbf{x}, F) = s(\mathbf{x})$, $\boldsymbol\lambda = (T^{(0.05)}, T^{(0.5)}, T^{(0.95)})'$, and $C(\boldsymbol\lambda) = \lambda_3 - \lambda_1$ for the length statistic, $C(\boldsymbol\lambda) = \log\{(\lambda_3 - \lambda_2)/(\lambda_2 - \lambda_1)\}$ for the shape statistic.

A standard error analysis of the bootstrap percentiles gives results much like approximation (6.12):

$$\text{cov}_{\text{intern}}\{\tilde{\mathbf{u}}(\hat\gamma)\} \doteq \mathbf{m}(n-1)^2(e_n - 1)\tilde v' M \tilde v / B, \tag{6.17}$$

where M is the $k \times k$ matrix with ijth element $\alpha_{\min(i,j)}(1 - \alpha_{\max(i,j)})$ and

$$\tilde v' \equiv \left(\ldots, \frac{\partial C(\boldsymbol\lambda)}{\partial \lambda_i}\bigg|_{\boldsymbol\lambda = \tilde{\boldsymbol\lambda}} / \tilde g_{\alpha_i}, \ldots\right). \tag{6.18}$$

As in equation (5.14), $\hat g(t)$ is the density corresponding to $\hat G(t) = \text{prob}_{\hat F}\{T^* < t\}$, and $\tilde g_{\alpha_i}$ is an estimate, based on B bootstrap replications, of $\hat g(t)$ at $\tilde G^{-1}(\alpha_i)$. Formula (6.17), like formula (6.12), is a simplified version of a more careful but tedious result.

For the length statistic,

$$\tilde v' M \tilde v = \frac{0.05 \times 0.95}{\tilde g^2_{0.05}} - \frac{2 \times 0.05^2}{\tilde g_{0.05}\tilde g_{0.95}} + \frac{0.05 \times 0.95}{\tilde g^2_{0.95}}, \tag{6.19}$$

and for the shape statistic

$$\tilde v' = \left(\frac{1}{\tilde g_{0.05}}\frac{1}{\tilde\gamma^{(0.5)} - \tilde\gamma^{(0.05)}}, -\frac{1}{\tilde g_{0.5}}\left(\frac{1}{\tilde\gamma^{(0.95)} - \tilde\gamma^{(0.5)}} + \frac{1}{\tilde\gamma^{(0.5)} - \tilde\gamma^{(0.05)}}\right), \frac{1}{\tilde g_{0.95}}\frac{1}{\tilde\gamma^{(0.95)} - \tilde\gamma^{(0.5)}}\right) \tag{6.20}$$

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

M being the 3×3 symmetric matrix with elements $M_{11}=M_{33}=0.05\times 0.95$, $M_{22}=0.5^2$, $M_{12}=M_{23}=0.05\times 0.5$ and $M_{13}=0.05^2$.

Formulae (6.19) and (6.20) gave the $\pm$ values in Table 2. The estimated values of $\hat{g}(c)$ were obtained from standard density estimation techniques applied to the bootstrap replicates T^{*b}, $b=1, \ldots, 1000$. They are average empirical densities, using roughly the 10% of the bootstrap data nearest the value c of interest. The underlying discreteness of the bootstrap CDF caused no noticeable effects in these calculations.

Remark 9. Our formulae for estimating $\text{se}\{\hat{\gamma}\}$, corrected or not, ignore one fact: that the available estimate of the bootstrap statistic of interest is $\tilde{\gamma}$, definition (2.13), not the ideal value $\hat{\gamma}$, definition (2.12), and so we should be interested in $\text{se}\{\tilde{\gamma}\}$ rather than $\text{se}\{\hat{\gamma}\}$. However, the difference between $\text{se}\{\tilde{\gamma}\}$ and $\text{se}\{\hat{\gamma}\}$ tends to be small. Consider the case $\hat{\gamma}=E_{\hat{F}}\{s(\mathbf{x}^*)\}$, $\tilde{\gamma}=\Sigma_{b=1}^{B} s^{*b}/B$. The true standard errors, sampling over $\mathbf{x}$ as well as $\mathbf{x}^*$, have the relation

$$\text{se}\{\tilde{\gamma}\}^2=\text{se}\{\hat{\gamma}\}^2+E(\hat{\sigma}^2/B), \tag{6.21}$$

where $\hat{\sigma}^2=\text{var}_{\hat{F}}\{s^*\}$. Combining equation (6.21) with the last line of approximation (6.15) suggests that we estimate $\text{se}\{\tilde{\gamma}\}$ by

$$\left[\widetilde{\text{se}}_{\text{jack}}\{\hat{\gamma}\}^2-\left\{\frac{(n-1)^2}{n}\,(e_n-1)-1\right\}\frac{\tilde{\sigma}^2}{B}\right]^{1/2}. \tag{6.22}$$

For $n=8$, the bracketed factor is 10.70, compared with 11.70 ignoring the term coming from equation (6.21). This difference will be unimportant for most purposes.

Remark 10. The obvious estimate $\tilde{\gamma}$, definition (2.13), does not necessarily make the best use of B bootstrap replications. Efron (1990b) describes other estimators $\breve{\gamma}$ which converge faster than $\tilde{\gamma}$ to the ideal limiting value $\hat{\gamma}$, as $B\to\infty$. In particular, if $s(\mathbf{x})=\theta(\hat{F})$ and $\gamma(F)$ is the bias $E_F\{s(\mathbf{x})\}-\theta(F)$, define

$$\breve{\gamma}=s^{*\cdot}-\theta(\breve{F}),$$

where $\breve{F}$ is the distribution putting probability $\Sigma_{b=1}^{B} P_i^b/B$ on x_i. Then $\breve{\gamma}$ often converges much faster than $\tilde{\gamma}=s^{*\cdot}-\theta(\hat{F})$ to the ideal bootstrap bias estimate $\hat{\gamma}=E_{\hat{F}}\{s^*\}-\theta(\hat{F})$.

For the bioequivalence ratio statistic, $\breve{\gamma}=0.0077$ compared with the value $\tilde{\gamma}=0.0053$ given in Table 5. The estimated standard error $\widetilde{\text{se}}_{\text{delta}}\{\hat{\gamma}\}=0.0083$ applies at least as well to $\breve{\gamma}$ as to $\tilde{\gamma}$. (Formally, it applies to $\hat{\gamma}$, as in remark 9.)

Remark 11. The problem of estimating the influence function $u_i\{\hat{\gamma}\}$ seems to become more difficult as n increases. Consider the case $\lambda(F)=T(\mathbf{X}, F)^{(\alpha)}$, so λ is the 100αth percentile of $T(\mathbf{X}, F)$. The internal coefficient of variation CV for the estimated influence function is

$$\begin{aligned}\text{CV}_{\text{intern}}\{\tilde{u}_i\{\hat{\lambda}\}\}&\doteq\left(\frac{\text{var}_{\text{intern}}\{\tilde{u}_i\{\hat{\lambda}\}\}}{u_i\{\hat{\lambda}\}^2}\right)^{1/2}\\&\doteq\left[\frac{(n-1)^3(e_n-1)\alpha(1-\alpha)}{nB\tilde{g}_\alpha^2 u_i\{\hat{\lambda}\}^2}\right]^{1/2}=O\left(\frac{n}{\sqrt{B}}\right),\end{aligned} \tag{6.23}$$

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

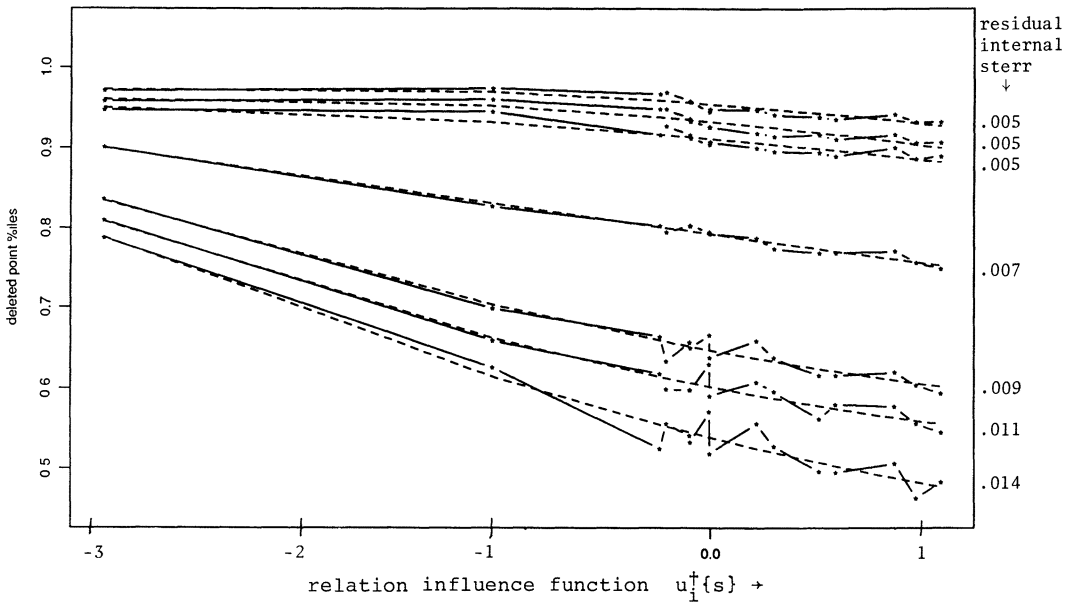

Fig. 7. Deleted point percentiles from the law school correlation, as in Fig. 4: -----, quadratic functions of $u_i^\dagger\{s\}$, fitted to the deleted point percentiles by ordinary least squares (numbers at the right-hand side indicate internal standard errors for the residuals from the quadratic fit (average standard errors excluding the two points at each end of the $u_i^\dagger\{s\}$ scale); the deviations of deleted point percentiles from fitted curves are commensurate with the internal standard errors)

as in expressions (6.17) and (6.18). It looks as though we need to take $B = O(n^2)$ bootstrap replications to maintain a reasonably small CV for $\tilde{u}_i\{\hat{\lambda}\}$.

This pessimistic result assumes that the deleted point values $\hat{\lambda}_{(i)}$ are computed in the obvious nonparametric way of approximation (3.5). However, there are often better ways to approximate $\hat{\lambda}_{(i)}$, especially if n is large. Fig. 7 shows the deleted point percentiles for the law school correlation bootstrap analysis, as in Fig. 4. The broken lines in Fig. 7 are ordinary quadratic regressions, used to smooth the jagged deleted point percentile lines. The deviations between the jagged and smooth curves are roughly commensurate with the internal residual error, obtained by using approximation (6.17).

Now we can read off smoothed estimates $\check{\lambda}_{(i)}$, for $\hat{\lambda}_{(i)}$, from the broken lines, leading to smoothed estimates $\check{u}_i\{\hat{\gamma}\}$ for the influence functions of the length and shape statistics. Doing this had little effect on the estimated influence function for the length statistic. However, the $\check{u}_i\{\hat{\gamma}\}$ values for the shape statistic were much more stable than the $\tilde{u}_i\{\hat{\gamma}\}$, having less than 1/100th the variance:

$$\check{\text{se}}_{\text{jack}}\{\hat{\gamma}\} = [\,\Sigma_i \check{u}_i\{\hat{\gamma}\}^2/n(n-1)\,]^{1/2} = 0.026$$

compared with

$$\tilde{\text{se}}_{\text{jack}}\{\hat{\gamma}\} = \{\Sigma\, \tilde{u}_i\{\hat{\gamma}\}^2/n(n-1)\}^{1/2} = 0.307,$$

as reported in Table 2.

Remark 12. Section 5 of Efron (1990b) introduces an improved method for estimating bootstrap percentiles, based on the bootstrap Hajek projection

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

formula (5.24). This method works well when R^2, expression (5.25), is near 1, as is likely to be so when the number n of original observations grows large. (Typically, $1-R^2$ is $O(1/n)$.) The same method can be applied to estimating the deleted point percentiles, yielding improved estimates $\check{u}_i\{\hat{\gamma}\}$ as in remark 11.

Remark 13. We might try to reduce internal errors by averaging $\tilde{U}_i\{\hat{\gamma}\}$ and $\tilde{u}_i\{\hat{\gamma}\}$. Table 6 is mildly discouraging: for the bootstrap standard error $\hat{\gamma}$, we can see that averaging will not reduce the internal error much since the internal correlations are so high. The correlations are low for the bootstrap bias estimate $\hat{\gamma}$, but now $\text{var}_{\text{intern}}\{\tilde{U}_i\{\hat{\gamma}\}\}$ is so much smaller than $\text{var}_{\text{intern}}\{\tilde{u}_i\{\hat{\gamma}\}\}$ that averaging is still ineffective.

REFERENCES

Beran, R. (1988) Prepivoting test statistics: a bootstrap view of the Behrens–Fisher problem, the Bartlett adjustment, and nonparametric analogs. *J. Am. Statist. Ass.*, to be published.

Chapman, D. and Hinkley, D. V. (1985) *Technical Report*. University of Texas, Austin.

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

——— (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

——— (1983) Estimating the error rate of a prediction rule: improvements in cross-validation. *J. Am. Statist. Ass.*, **78**, 316–331.

——— (1987) Better bootstrap confidence intervals and bootstrap approximations. *J. Am. Statist. Ass.*, **82**, 171–185.

——— (1988) Three examples of computer-intensive statistical inference. *Sankhya* A, **50**, 338–362.

——— (1990a) Jackknife after bootstrap standard errors and influence functions. *Technical Report 134*. Stanford University.

——— (1990b) More efficient bootstrap computations. *J. Am. Statist. Ass.*, **85**, 79–89.

Efron, B. and Tibshirani, R. (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.*, **1**, 54–77.

Hall, P. (1988) Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.*, **16**, 927–953.

Hall, P. and Martin, M. A. (1988) On bootstrap resampling and iteration. *Biometrika*, **75**, 661–671.

Hammersley, I. and Handscomb, D. (1964) *Monte Carlo Methods*. London: Chapman and Hall.

Hampel, F., Ronchetti, E., Rousseeuw, P. and Stahel, W. (1986) *Robust Statistics, the Approach Based on Influence Functions*. New York: Wiley.

Hayes, K. G., Perl, M. L. and Efron, B. (1989) Application of the bootstrap statistical method to the tau-decay-mode problem. *Phys. Rev.* D, **39**, 274–279.

Hinkley, D. V. and Shi, S. (1989) Importance sampling and the nested bootstrap. *Biometrika*, **76**, 435–446.

Huber, P. (1981) *Robust Statistics*. New York: Wiley.

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

Loh, W.-Y. (1987) Calibrating confidence coefficients. *J. Am. Statist. Ass.*, **82**, 155–162.

Mallows, C. (1974) On some topics in robustness. *Memorandum*. Bell Laboratories, Murray Hill.

Stein, C. (1981) Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**, 586–596.

Tibshirani, R. (1988) Variance stabilization and the bootstrap. *Biometrika*, **75**, 433–444.

DISCUSSION OF THE PAPER BY EFRON

J. B. Copas (University of Birmingham): Although ideas of the bootstrap and jackknife go back to work of Quenouille and early papers by Barnard and Tukey and others, it is only recently that the full implications of these ideas have begun to be realized. It is tonight's author who has been very much associated with recent developments and it is therefore a pleasure to welcome Professor Efron and his latest contribution to this intriguing topic. As we have come to expect from his previous papers, his ideas are innovative and presented with his usual elegance and style.

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

The bootstrap is a very clever idea. As if by magic we can produce replications of our data, almost like lifting ourselves up by our bootstraps. In large samples this is convincing and clearly useful. But levitation is not for the doubter and with small samples we may well be hesitant. The good news of this paper is that we can put our doubts to rest with a little statistical surgery. The thought of severing our means of support may be alarming, but with Professor Efron as the surgeon with the jackknife we find that the surgery is only minor, even infinitesimal, and the outcome thoroughly therapeutic. The main idea of the paper, like all good ideas, is very simple: we just select those bootstrap replications that happen to omit the ith observation and hence reconstruct all the usual jackknife statistics. The roles of the bootstrap and jackknife here are intriguing. In the first part of the paper the jackknife is applied to the sample but estimated by the bootstrap, but then there is a subtle reversal of roles in Section 6.

A thread running through the paper is the correlation example of Fig. 1. A Q–Q plot of the normalized distances between the points and the mean for a bivariate normal distribution highlights point A, but the others seem sensible. Even point A's value of 7.3, however, is not all that unusual as the largest of a sample of 15 from χ^2 on 2 degrees of freedom. If the correlation is large then there would seem to be more scope for deletion of points to increase r than to decrease it, and this raises the question of the skewness of the distribution of the relative jackknife influences; perhaps -2.97 is not as unusual as it may seem.

So the only doubt for a bivariate normal distribution is point A. With it in, the 90% confidence interval is (0.51, 0.91). With it out it is (0.74, 0.96), both ends going up as we would expect. So I was a little surprised to see that the bootstrap-t approach makes both limits go down, to (0.39, 0.90). It is also odd that the largest value of the upper limit in Fig. 5 occurs for a point with nearly no relative influence. Perhaps this is something to do with the fudge factor which has crept into equation (3.9).

This bootstrap interval is based on the premise that the statistic T in equation (3.8) is approximately pivotal. The other bootstrap interval, mentioned in Section 2, is the 'percentile confidence interval'. This is the distribution of the sample estimate in a bootstrap setting, and it is unclear how this can be interpreted directly as a confidence interval except in the special case of the sample estimate, possibly after transformation, being unbiased with a symmetrical error distribution. But even then the argument is only clear for the parametric version of the bootstrap. A better approach is Professor Efron's own BC_a method but this is not used in defining the length and shape statistics which feature prominently in the paper. Some guidance on what these statistics mean would be welcome. Maybe we are tempted to interpret shape as an estimate of the shape of a likelihood function, a temptation that we should resist without further evidence.

Let us take what is perhaps the simplest non-trivial problem for which we can all agree on the correct inference, the single-sample location parameter model:

$$x = \theta + \epsilon, \qquad \epsilon \sim f(\epsilon).$$

The likelihood is $\Pi f(x_i - \theta)$. The sample spacings are ancillary and so we look to the conditional distribution of any location invariant statistic given those spacings. This is just a relocated and reversed version of the likelihood and so any confidence limit corresponds to taking the appropriate relative area under the likelihood function—essentially a Bayes analysis with a uniform prior. We note that the sampling distribution which generates these confidence limits is a random rigid translation of the sample vector which is quite unlike a bootstrap sample which picks and chooses data values from within a fixed set.

For the parametric version of the bootstrap explained at the end of Section 3, clearly,

$$P(\hat{\theta}^* - \hat{\theta}) \equiv P(\hat{\theta} - \theta)$$

and hence if $\theta^{*(\alpha)}$ is the α-quantile of the bootstrap distribution

$$P\{\theta > \hat{\theta} - (\theta^{*(\alpha)} - \hat{\theta})\} = \alpha,$$

defining an exact, but marginal, upper confidence limit for θ. What can we say about the approximation to this first equation if the parametric bootstrap is replaced by the nonparametric bootstrap? How close then are we to the second equation, which is after all the defining equation for a confidence limit? Can the conditional-on-P_i^b device of the paper be generalized to conditioning on bootstrap samples which are, in a relevant sense, reasonably comparable with the observed data?

A tough test related to my first question is the guarantee time exponential model with

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

$$f(\epsilon) = \begin{cases} 0 & \epsilon < 0, \\ \lambda \exp(-\lambda\epsilon) & \epsilon \geqslant 0. \end{cases}$$

Here θ is the least possible value of x, $\hat{\theta}$ is $x_{(1)}$ and $\hat{\theta}^*$ is $x^*_{(1)}$. As, approximately,

$$P(\hat{\theta}^* = x_{(i)}) = (e-1)\exp(-i),$$

the lower α confidence limit for θ is roughly

$$x_{(1)} - (x_{(k)} - x_{(1)})$$

where $k = 1 + [-\log \alpha]$. But

$$E(x_{(k)} - x_{(1)}) \simeq \frac{k-1}{n\lambda} = \frac{[-\log \alpha]}{n\lambda},$$

which can be compared with the correct confidence limit

$$x_{(1)} + \frac{\log \alpha}{n\lambda}.$$

Thus the bootstrap limit has the right expectation but is estimated rather imprecisely. As might be expected, the jackknife fails to pick this up, the jackknife standard error being out by a factor of $\sqrt{n}$.

Section 4 of the paper has considerable practical potential, as some of our most tricky statistical problems are of this type. In the tau data of Table 3, the variances increase with more trimming for decays 1 and π, but decrease for ρ, e and μ. This reflects the differing lengths of tails in these data sets—we would appear to do even better by allowing for different trimming percentages. It is all a matter of kurtosis, and can we expect small samples to tell us much about tail length? Surely not, and so it is reassuring that the jackknife confirms that little can be said about the best trimming rate. Disappointing it may be, but sensible.

This is just one of the many good things in the paper and it gives me great pleasure to propose the vote of thanks.

Alastair Young (University of Cambridge): Professor Efron is to be congratulated on producing a paper which forces us to think in critical terms of the bootstrap estimators that we construct. He has shown how questions about the accuracy of bootstrap estimators which may be expressed in terms of standard errors can be tackled without the need for further resampling, but by methods which require instead careful sorting of the original bootstrap simulation: but what of the accuracy of the standard errors constructed? I personally feel dubious about the utility of standard error estimates of 0, such as those in Tables 2 and 5 of the paper. Has Professor Efron attempted to validate his accuracy measures for circumstances other than the specific examples of this paper?

An issue that worries me concerns the number of bootstrap simulations B. Perceived wisdom would suggest that quite moderate values of B are adequate in the construction of the basic bootstrap quantities, such as bias estimates, though this paper appears to indicate that somewhat larger B is required to reduce the internal error in jackknife-after-bootstrap calculations to acceptable levels.

The primary reason for estimating the accuracy of a bootstrap quantity is usually to correct for error in that quantity. It seems to me that the techniques of Professor Efron's paper are probably not sufficiently flexible to be able, in general, to handle such error correction. Iterated levels of bootstrapping may be necessary if we are to provide anything more than rather crude error estimates for our bootstrap quantities. In this respect, I feel that Professor Efron is unduly pessimistic when he expresses the view that iterated bootstrap calculations are computationally too intensive for routine use. Much recent research effort has focused on methods of analytic approximation which reduce the computational demands of bootstrap methods, especially iterated bootstrap techniques. These methods deserve mention here.

A typical and important use of bootstrap-after-bootstrap calculations is for estimation of coverage error in bootstrap confidence intervals. I should like to indicate how this can be done efficiently and accurately, in a quite general setting, without the need for a second level of resampling.

Suppose that we are interested, on the basis of sample data $\mathscr{X} = \{X_1, \ldots, X_n\}$ from an unknown distribution, in inference for a parameter θ which is expressible as a smooth function of means, $\theta = g(\mu_1, \ldots, \mu_k)$. The parameter θ is estimated by $\hat{\theta} = g(\bar{Z}_1, \ldots, \bar{Z}_k)$, with $\bar{Z}_l = n^{-1}\Sigma_{j=1}^{n} f_l(X_j)$, for smooth

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

TABLE 7

Type of interval	*Coverage, $n=20$*	*Coverage, $n=35$*	*Coverage, $n=100$*
Percentile	0.80	0.83	0.88
Double bootstrap	0.87	0.88	0.90
Approximation 1	0.84	0.87	0.89
Approximation 2	0.87	0.89	0.90

functions $f_1, \ldots, f_k$. Examples of such parameters include variances, ratios of means, correlation coefficients etc.

Let $I_0(\alpha; \mathscr{X}, \mathscr{X}^*)$ denote a bootstrap confidence interval for θ of nominal coverage α. Here the notation indicates that I_0 is constructed from $\mathscr{X}$, using information provided by bootstrap samples $\mathscr{X}^*$ drawn from $\mathscr{X}$. The coverage $\pi(\alpha) = P\{\theta \in I_0(\alpha; \mathscr{X}, \mathscr{X}^*)\}$ will, in general, differ from α and may be estimated by

$$\hat{\pi}(\alpha) = P\{\hat{\theta} \in I_0(\alpha; \mathscr{X}^*, \mathscr{X}^{**}) \mid \mathscr{X}\}.$$

Here $\mathscr{X}^{**}$ denotes a generic resample from the first-level bootstrap sample $\mathscr{X}^*$. The idea of coverage correction is to use instead of $I_0(\alpha; \mathscr{X}, \mathscr{X}^*)$ the confidence interval $I_0(\delta_\alpha; \mathscr{X}, \mathscr{X}^*)$, where δ_α solves $\hat{\pi}(\delta_\alpha) = \alpha$. If I_0 is the percentile method interval, for instance, $\hat{\pi}(\alpha)$ can be well approximated by the proportion of times over B resamples $\mathscr{X}^*$ from $\mathscr{X}$ that

$$(1-\alpha)/2 \leqslant P(\hat{\theta}^{**} < \hat{\theta} \mid \mathscr{X}^*) \leqslant (1+\alpha)/2,$$

where $\hat{\theta}^{**} = g(\bar{Z}_1^{**}, \ldots, \bar{Z}_k^{**})$ denotes the version of $\hat{\theta}$ computed from a sample $\mathscr{X}^{**}$ drawn from $\mathscr{X}^*$.

In the smooth function model above, we can estimate $P(\hat{\theta}^{**} < \hat{\theta} \mid \mathscr{X}^*)$ analytically, thereby achieving the coverage correction on the basis of only the first-level bootstrap samples. This is done by using a saddlepoint approximation to the joint distribution of $\bar{Z}_1^{**}, \ldots, \bar{Z}_k^{**}$, given $\mathscr{X}^*$, in conjunction with a tail area approximation formula to the marginal distribution of $g(\bar{Z}_1^{**}, \ldots, \bar{Z}_k^{**})$: for details see DiCiccio *et al.* (1990a, b). This analysis replaces the need for a complete second level of resampling with the computationally less demanding exercise of solving a simple system of $2k+1$ non-linear equations in as many unknowns.

As a simple illustration, consider construction of nonparametric bootstrap confidence intervals for the variance $\theta = E(X^2) - E(X)^2 = \mu_2 - \mu_1^2$ of a normal distribution. The percentile method is known to perform badly in this example, and what is needed is not a means of recognizing this unreliability but some means of correcting the method to improve its performance. A simulation of 1600 confidence intervals for each of three sample sizes $n = 20$, 35 and 100 gave the estimates in Table 7 of coverage (standard error 0.01) for the percentile and double-bootstrap confidence intervals of nominal coverage 0.90. Each interval was based on 1000 resamples at the outer level, the double-bootstrap intervals using the analytic approximation method at the inner level. Such an approximation reduces the computational load of the bootstrap-after-bootstrap construction by a factor of about 7/8. DiCiccio *et al.* (1991) introduce two further approximations to these double-bootstrap intervals, based on approximate rather than exact solutions to the system of non-linear equations, and which result in substantial computational savings without much loss in accuracy. The first (approximation 1) is about 70 times faster than nested resampling, and the second (approximation 2) is about 25 times faster than nested resampling (Table 7).

In summary, I contend that more flexible bootstrap-after-bootstrap calculations can be applied easily and routinely without the need for the second level of bootstrapping. It would be of interest to know how the methods of Professor Efron's paper compare, in both scope and accuracy, with the kinds of procedure that I have sketched above.

It should be clear from my remarks that I found this a most stimulating paper, which raises many interesting issues. It gives me great pleasure to second the vote of thanks.

The vote of thanks was passed by acclamation.

A. C. Davison (University of Oxford): This paper emphasizes what are essentially single-case deletion diagnostics, which in other contexts are prone to masking. In principle the idea of subdividing bootstrap output to assess the effect of deleting cases from the original sample could be extended to any subset,

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

provided that the bootstrap is sufficiently large, but the 'combinatorial explosion' limits the usefulness of this. Recent work stemming from Cook (1986) attempts to overcome this problem by local perturbations of all the observations simultaneously. A difficulty with local perturbations in the present context is that resampling is a discrete procedure whereas smoothness seems necessary for local influence calculations. One possibility is to study empirical likelihood, a close relative of the bootstrap. The log(empirical likelihood) for the mean μ of a sample $x_1, \ldots, x_n$ is (Owen, 1988, 1990)

$$L_E(\mu) = -\sum_{j=1}^{n} \log\{1+\lambda_\mu(x_j-\mu)\}, \qquad \min(x_j) < \mu < \max(x_j),$$

where λ_μ is a Lagrange multiplier that solves the equation

$$\sum_{j=1}^{n} \frac{x_j-\mu}{1+\lambda(x_j-\mu)} = 0.$$

Consider local perturbation of the sample, $x_j \to x_j + \epsilon a_j$, where $(a_1, \ldots, a_n)$ is a fixed vector of unit length. Then following Cook's recipe slavishly, the choice of a_j that maximizes the local change in $L_E(\mu)$ is the vector $\mathbf{a}$ that maximizes

$$\left.\frac{\partial^2 L_E(\mu)}{\partial \epsilon^2}\right|_{\epsilon=0},$$

namely $a_j \propto \{1+\lambda_\mu(x_j-\mu)\}^{-2}$. This rather bizarre perturbation is useless at the maximum empirical likelihood estimate of μ, for which $a_j \propto 1$. There may be some use in ideas like this, but since in a sense empirical likelihood makes the data the model, it seems in principle difficult to check the model on the basis of the data.

A different way to combine perturbations and the bootstrap would be to add to a bootstrap resample $X_1^*, \ldots, X_n^*$ a perturbation $\epsilon_1^*, \ldots, \epsilon_n^*$, suitably normalized, and then somehow to assess the impact of the perturbation. The lack of an objective function against which to do so seems inevitably to lead to the double bootstrap, however. Is there a way to do this that avoids nested bootstrapping?

The most critical assumption in resampling is surely independence, and despite the optimistic tone of the second sentence of the paper bootstrapping with dependent data is insufficiently understood in theory and problematic in applications. Is there any way to check independence that is specific to the bootstrap?

There are important issues still to be tackled, but Professor Efron is to be congratulated for putting the question of sensitivity of bootstrap calculations firmly on the agenda. I add my thanks to those of the other discussants.

C. Chatfield (University of Bath): The author's clear writings have done much to promote the acceptance of bootstrapping as a useful addition to the statistician's toolkit. However, in practice, it is often unclear when bootstrapping should be used, and, if it is, what assumptions are implicitly made, and whether the results are to be regarded as descriptive or inferential. I hope that guidelines can be clarified. The technique appears most useful for analysing smallish expensive-to-collect data sets where prior information is sparse, distributional assumptions are unclear and where further data may be difficult to acquire. The data in Table 3 may well satisfy these criteria. The technique captures the 'spirit of the age' in that it is computationally intensive and is designed to squeeze as much as possible out of a single set of data. The possible extension to bootstrap-after-bootstrap methods, which is mentioned in the paper, seems to go even further.

Rather than to comment on the details of jackknife-after-bootstrapping, I would like to air my concern that bootstrapping may become overused. It carries to new extremes the statistician's tendency to overconcentrate on a single set of data. Thus I would like to remind the reader that, where possible, it is often better to devote one's effort towards getting more data. Classical statistical inference is essentially concerned with making inferences about the parameters of an *assumed* family of probability models from a *single* set of data. In practice model building is an iterative procedure involving model formulation, estimation (usually the easy part!) and model validation. Typically a model is formulated, fitted and checked on the *same* data set. It is 'well known' that this will affect *P*-values, and introduce selection biases (e.g. Miller (1990)), but people still do it and largely ignore the problems. In contrast scientific inference is typically concerned with collecting *many* data sets and establishing a relationship which *generalizes* to different conditions (i.e. we look for *significant sameness*, rather than for differences).

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

I realize that the data in Fig. 1(a) are only meant as an illustrative example, but I am worried that the proposed analysis will mislead the reader. In my view bootstrapping should not be used for these data, which do not satisfy the requisite criteria. I am *not* interested in evaluating SE(correlation) for one year's data. I *would* be interested in asking various other questions, such as which law school is the outlier and why. More importantly I would want to look at data for 1974, 1975 . . . (which presumably are readily available) and to see whether similar results arise. For example is the same law school an outlier each year? With the spread-out data in Fig. 1(b), I would also prefer to seek more replications rather than to try to squeeze more out of the very small data set than may actually be there.

In short, I suggest that bootstrapping should only be used in rather exceptional cases and that the statistician should always be clear what question is to be answered, be ready to ask searching questions if necessary, and give high priority to getting more data.

Robin Henderson (University of Newcastle upon Tyne): Professor Efron's interesting paper is certain to motivate several lines of future research. I would like to comment on the suggestion that the jackknife-after-bootstrap (JAB) approximation (3.5) should be used in nonparametric problems and estimate (3.18) in parametric problems, the latter trading off a reduction in variance for a possible increase in bias. In certain semiparametric situations there is a choice between the two within the same framework of assumptions, and it is not clear which should be preferred. To illustrate, consider the modelling of survival time t as a function of covariates x through a proportional hazards model for the survival distribution, namely

$$S(t|x) = S_0(t)^{\exp(\beta x)}$$

with $S_0(t)$ unspecified. Bootstrap methods can be very useful when interest is in estimating $S(t_0 \mid x_0)$ for fixed (t_0, x_0), and the additional information given by the JAB is therefore potentially valuable. However, we have to decide how to obtain the bootstrap sample and which JAB to use. Assume for simplicity no censoring so that the bootstrap sample can be obtained by either

(a) sampling from the observations (t_i, x_i) directly,
(b) considering the x_i as being fixed and, for each, sampling from the empirical conditional distribution of t given x_i, which is discrete with support at $t_1, t_2, \ldots, t_n$, or
(c) sampling from the empirical distribution $\hat{F}$ of the x_i and *then* from the conditional distribution of t given x_i, which is not the same as (a).

Under (a) either equation (3.5) or equation (3.18) based on $\hat{S}_0$ can be used for the JAB, but under (b) and (c) only approximation (3.18) is available since the bootstrap samples contain combinations (t_i, x_j) that do not appear in the original data. Though not conclusive, preliminary simulation results indicate that the combination of (a) with approximation (3.5) leads to jackknife estimates of standard error which are considerably smaller than obtained by (a) with approximation (3.18) or either of (b) or (c). Otherwise, there seems to be little difference between the four methods. More work is required before a firm recommendation can be made but it is clear that the problem of choice of JAB is not trivial.

B. J. Worton (University of Oxford): I would like to comment on the tau data set analysis. As Professor Efron points out, it is possible to use a second level of bootstrapping to assess the variability of the bootstrap variance $v(q, \Delta)$ of the q% estimator based on the trimmed mean of the contrast parameter Δ shown in Fig. 6. This has been done and the results are shown in Fig. 8. Each curve was obtained by bootstrapping first-level bootstrap samples in the same way as the original data were bootstrapped. The plot shows that there is little to choose between the estimators, especially for $q \leqslant 0.3$.

I would now like to mention an alternative use for the first- and second-level bootstrap statistics, in the construction of a bootstrap-generated likelihood. The basic method of calculating a bootstrap likelihood is, in brief, as follows. Assume for the moment that we have a single independent random sample $\mathbf{X} = (X_1, \ldots, X_n)'$ from an unknown distribution F, and the population characteristic of interest is $\theta = t(F)$ which is estimated by $T = t(\hat{F})$, where $\hat{F}$ is the empirical distribution calculated from $\mathbf{X}$. If t_0 is the observed value of the statistic T calculated from the observed data, the idea is to use the second-level bootstrapping to estimate the density of T at t_0 for various $\theta = t^*$ first-level bootstrap statistic values. The algorithm proceeds by generating first-level bootstrap data sets, $\mathbf{x}_1^*, \ldots, \mathbf{x}_M^*$ and from these calculating corresponding statistics $t_1^*, \ldots, t_M^*$ in the usual way. For each of the first-level samples $\mathbf{x}_i^*$, generate second-level data sets, $\mathbf{x}_{i1}^{**}, \ldots, \mathbf{x}_{iN}^{**}$ and from these calculate statistics $t_{i1}^{**}, \ldots, t_{iN}^{**}$, smooth the $t_{i1}^{**}, \ldots, t_{iN}^{**}$ with a kernel density smoother and evaluate at t_0. This gives

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

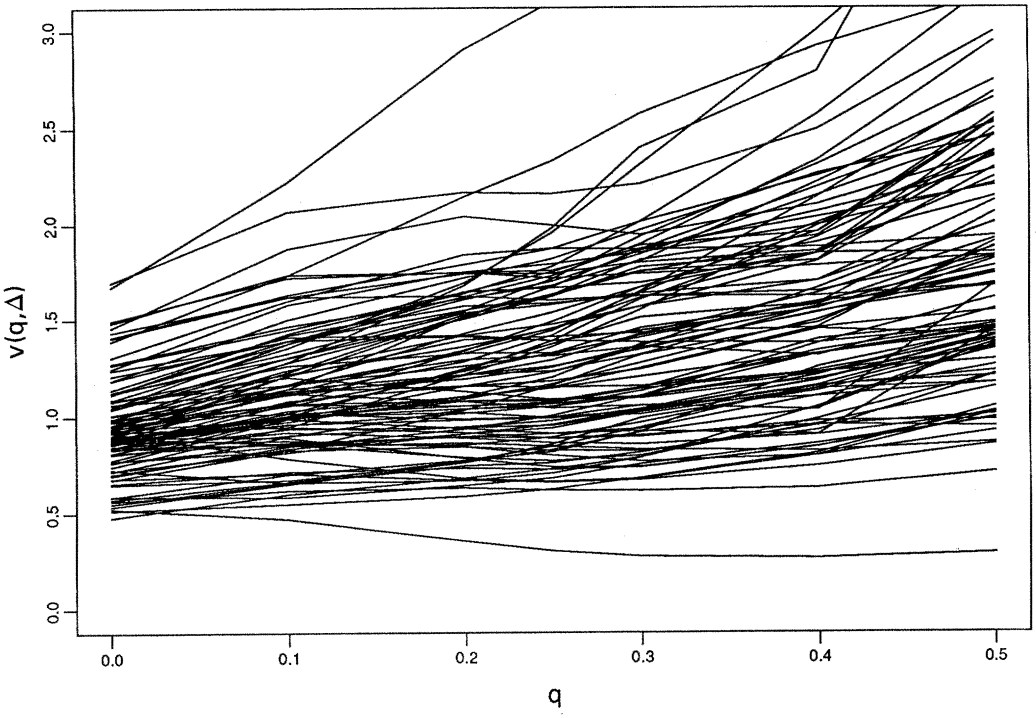

Fig. 8. Double bootstrap $v(q, \Delta)$ curves

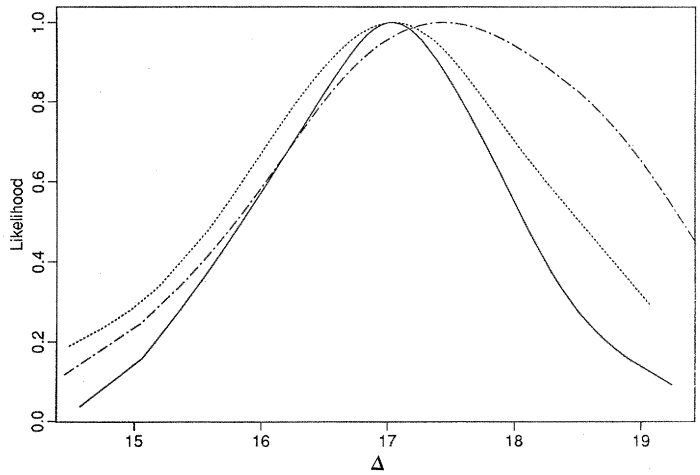

Fig. 9. Double-bootstrap likelihoods for the Δ estimator based on means (———), 25% trimmed means (· · · · ·) and medians (—·—·)

M point estimates of likelihood at $\theta = t_1^*, \ldots, t_M^*$. These points can then be scatterplot smoothed (on the log-scale) to produce a bootstrap likelihood curve estimate. The basic method extends easily to situations where the characteristic of interest is a function of several sample populations, as in the tau decay example.

Applying the bootstrap likelihood algorithm to the tau data example, taking the statistics of interest as being based on the means, the 25% trimmed means and the medians, we obtain the likelihoods shown

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

in Fig. 9. These likelihoods can be used to assess how informative the various estimators are about Δ. Evidently the median-based statistic seems to contain less information than the estimators based on the mean and 25% trimmed mean. We can also use the likelihoods to assess the plausibility of various Δ values.

The double bootstrap may appear to be a rather extravagant use of computer resources in this example. However, if we compare this cost against the effort that went into collecting the original data it seems quite negligible. It is possible in certain problems to reduce the computation time taken to construct a bootstrap likelihood dramatically by replacing the Monte Carlo simulation at the second level by accurate approximations.

Further details of bootstrap likelihood are given in Davison *et al.* (1991).

A. P. Dawid (University College London): The jackknife-after-bootstrap method involves removing from the collection of all bootstrap samples those for which any of the values take on the one that we want to omit. An alternative is simply to remove from each single bootstrap sample all those elements equal to the value that we want to omit. Admittedly, this would give a random bootstrap sample size, but I cannot see this as a major problem. This method would have the advantage of yielding a larger basis for performing calculations.

A. C. Atkinson (London School of Economics and Political Science): I have two brief comments on Professor Efron's interesting paper.

(a) Why could we not transform the correlation coefficient to approximate symmetry by using the z-transformation, as do Graham *et al.* (1990)? Smaller bootstrap samples should then be needed for the same degree of accuracy.
(b) Has Professor Efron any advice on the design of bootstrap studies? For whole sample studies references to the advantages of various forms of balanced design include Wynn and Ogbonmwan (1986), Davison *et al.* (1986) and Graham *et al.* (1990). The desirability of balance when individual observations are deleted appears to raise new problems.

My third comment elaborates on a remark by another discussant who mentions masking. The paper's statistics are examples of 'leave-one-out' diagnostics. It is well known that such methods may fail in the presence of groups of outliers or influential observations, as also may the local influence methods of Cook (1986). An example in regression is given by Atkinson and Weisberg (1991). To overcome these problems my colleague Dr Shephard and I have been using methods incorporating genetic algorithms (Goldberg, 1989; Davis, 1991). Initial results are encouraging, although the computer requirements are such as to satisfy the appetite of even the most ardent advocate of computer-intensive methods in statistics.

Frank Critchley (Warwick University, Coventry): In welcoming this paper, I would like to make four comments.

(a) Designer bootstraps: balance considerations suggest that, in the context of this paper (and also more generally, as others have remarked), there are gains to be made from designing bootstrap samples rather than producing them by Monte Carlo methods. For example, a design can be chosen so that the resampling vector $\mathbf{P}$ has all its elements equal. This contrasts with the Monte Carlo approach under which $\mathbf{P}$ is stochastic. In this latter case, by chance, some accuracy-of-accuracy measures are more accurate than others. Designs with higher order balance could helpfully be employed to extend the present one-at-a-time approach to the study of the joint effect of subsets of observations.
(b) Multiple re-use . . . of the *same* bootstrap samples induces (potentially very high) dependence between the resulting statistics. Where this is likely to be a problem, but further resampling is prohibitive for some reason, one approach might be to reduce this dependence—at some loss of efficiency—by calculating each statistic only on some (designed) subset of the relevant bootstrap samples.
(c) The simple random sample model (2.1), mimicked by the bootstrap samples, can be pressed too far. For example, A is clearly an outlier in the law school data. In such cases, it may seem inappropriate, and lead to at best inefficiency, and at worst error, to retain model (2.1). Put positively, there seems to me to be potential in developing a general data analysis strategy based on a constructive interplay between diagnostic and bootstrap methods.

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

(d) Delta *versus* jackknife: the author poses the interesting question about what determines which of these approaches makes the more efficient use of the fixed available number B of bootstrap replications. It seems that the answer may depend essentially on a trade-off between non-linearity and 'sample' size. The delta method calculations can use all B 'observations', unlike the jackknife method. However, since it is based on first-order derivatives, the efficiency of the delta method calculations depends *inter alia* on linearity considerations, which will vary from one functional to another (cf. remark 4).

C. Jennison (University of Bath): The basic idea of the bootstrap is summarized on p. 88 of the paper: a functional $\gamma(F)$ of the true data distribution is to be approximated by the bootstrap version, $\hat{\gamma}(\mathbf{x}) = \gamma(\hat{F})$. The adequacy of this approximation relies on asymptotic theory, but this theory can be of only limited value for small data sets, especially if the true F is not well behaved and convergence to the asymptotic result is slow.

I wonder what the true F looks like in the law school data example. There is one notable outlier in the data set, so let us suppose that F is a mixture of something well behaved, like a bivariate normal distribution, plus a more dispersed outlier distribution. If the data contain only one observation from the 'outlier' part, it is obvious that asymptotic theory which relies on consistently estimating both the normal and the outlier parts of F will not be of much help.

Diagnostics such as the influence function can still be helpful in the absence of an accurate estimate of F. But the paper also discusses estimates of quantities such as the standard error of the estimated correlation, $\text{se}(\hat{\rho})$, which depend very much on the whole of F. Given the problem in estimating the outlier component of F, do the proposed quantities have any supporting theory which is relevant to such a small and awkward example? Even if we invoked parametric assumptions, modelling the outlier part of F as a second bivariate normal distribution, there is only one observation from which to estimate its location and dispersion, so no real progress can be made: it would be very surprising if an all-purpose, automatic bootstrap technique could succeed in this difficult problem.

On a more positive note, I am sure that the bootstrap has much to offer in the analysis of moderately large and fairly well-behaved data sets. Such mundane problems fall between the extremes of a mathematical asymptotic theory and the simple analyses of small examples that we seem to see rather often. Has Professor Efron any advice or encouragement to applied statisticians wishing to make use of bootstrap methods in these intermediate situations?

Robert J. Tibshirani (University of Toronto): We often forget that bootstrap estimates, like most random variables, are subject to sampling variation. In this interesting paper, Professor Efron gives a simple way of estimating the standard deviation of a functional ϕ of the bootstrap distribution of a statistic T. I like his idea because it is simple and can be applied in a straightforward way, without having to worry about the form of T or ϕ.

Assessment of the variability of prediction error estimates is a potentially important area of application for the jackknife-after-bootstrap. I carried out a small experiment to investigate how it works for that problem. 50 data pairs were generated from the model $z = \beta_0 + t\beta_1 + \epsilon$ where $\beta_0 = 0$, $\beta_1 = 1$, $t \sim N(0, 1)$ and $\epsilon \sim N(0, 0.75^2)$. Denote the average squared residual (ASR) by

$$\sum_{1}^{n} (z_i - \hat{\beta}_0 - t_i \hat{\beta}_1)^2/50$$

where $(\hat{\beta}_0, \hat{\beta}_1)$ are the least squares estimates. The quantity of interest is the optimism OP which is the (downward) bias in ASR as an estimate of the true prediction error (Efron, 1983). For my sample, ASR = 0.438 and OP = 0.030. I used only $B = 10$ bootstrap samples; since model fitting procedures can be very costly in complicated models, this is often practically reasonable. Table 8 shows some standard errors for OP.

TABLE 8

True jackknife: 0.033
Jackknife-after-bootstrap: 0.283
Simple bootstrap formula: 0.086
True jackknife for cross-validation: 0.056
Simple cross-validation formula: 0.086

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

The first line was obtained by leaving one point out at a time and explicitly recomputing OP for a new set of bootstrap samples. The second line is Efron's approximation (3.5). (To avoid the possibility of differences due to bootstrap sample variability, the first and second results are actually averages over 25 realizations of the bootstrap process, as is the third result.) The jackknife-after-bootstrap method seems to overestimate the standard error of OP badly. I believe that this is because of the small value of B: when I increased B to 50 and 100, the standard error fell to 0.153 and 0.095 respectively. However, as the first result indicates, this discrepancy does not reflect the true variability of OP but instead reflects error in the jackknife-after-bootstrap estimate. It appears then that the jackknife-after-bootstrap method may not be very useful unless B is large, say at least 100. This would restrict its usefulness to situations in which the estimator T is relatively simple. I would be grateful if Professor Efron could comment on this.

The third result uses the simple formula $\{(1+1/B)\tilde{\sigma}^2\}^{1/2}$, where $\tilde{\sigma}$ is the bootstrap variance of the B individual OP estimates (see above Efron's equation (6.13)). This seems to perform reasonably well; such an approach can be used for any statistic of the expectation type. The fourth and fifth results refer to leave-one-out cross-validation. The fourth result gives the true jackknife standard error of the cross-validation estimate, whereas the fifth is the standard error of the $n=50$ individual error estimates. The simple formula is quite accurate; intuitively, it is reasonable because the individual cross-validation estimates are fairly independent, more so than the individual bootstrap estimates. This method is used in the CART algorithm (Breiman *et al.*, 1985). A 'jackknife after cross-validation' method may not be needed.

As a final general point, should not formula (3.4) read

$$\hat{\gamma}_{(i)} = \phi\,[\,T(\mathbf{x}^*,\ \hat{F}^{-i})|\ P_i = 0\,]$$

where $\hat{F}^{-i}$ puts mass $1/(n-1)$ on $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ and similarly in formula (4.5), i.e. should not x_i be left out of $\hat{F}$ as well? I tried this change in the above example, but it made little difference.

John W. Tukey (Princeton University): I am pleased to see Professor Efron using a jackknife, even in the limited way that he proposes. It is difficult to reconcile 'a second level of bootstrapping is certainly the most direct and *efficient* way' (emphasis added) with 'jackknife-after-bootstrap requires perhaps 100–1000 times less computation than bootstrap-after-bootstrap' (both near the close of Section 1).

If $y-\theta$, centred near 0, is more often large positive than large negative (is right skewed) then $\theta-y$ is left skewed, and

$$2y^{(0.50)} - y^{(0.95)} < \theta < 2y^{(0.50)} - y^{(0.05)},$$

also left skewed, is a more reasonable 90% confidence interval for θ than

$$y^{(0.05)} < \theta < y^{(0.95)},$$

which is right skewed (and of the same length). It is, I assert, appropriate to call the latter interval a *seductive* interval, since its simplicity can seduce statisticians into forgetting the distinction between $y-\theta$ and $\theta-y$.

(a) If we can re-express and reparameterize our problem to make the distribution of $y-\theta$ more nearly symmetric, or more slowly changing with θ, we should.
(b) We should study the effectiveness of the first inequalities, especially in comparison with the seductive interval.
(c) All this applies to bootstrap sampling, where I believe that we badly need to avoid the seductive bootstrap interval (defined as the percentile confidence interval near Fig. 2). (Why should we want to know about the shape of the seductive bootstrap interval, rather than about that of the natural confidence interval?)

I am concerned with an apparent willingness to forget anything that we know or understand about statistical behaviour—just pushing on with the question as originally formulated. Although (just before remark 1) we are invited to follow Tibshirani's suggestion and to work with $z=\tanh^{-1}r$, the author chooses not to do this. Fisher's z often has a more stable variance and a more symmetric distribution than r, in non-Gaussian as well as Gaussian situations (Haldane, 1949). Thus we can expect a bootstrap interval for z to function better than one for r. Indeed, near the author's Fig. 4, the interval lengths for z, $1.22=1.84-0.62$ and $1.04=2.11-1.07$, are much more nearly equal than those for r. The relative influence of point A for length

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

in terms of z will be much smaller than it was, presumably as an artefact, for length in terms of r.

Rudolf Beran (University of California, Berkeley): Let F_n denote the empirical distribution of an independent, identically distributed sample of size n, drawn from an unknown distribution F. Then, the theoretical nonparametric bootstrap estimate of a functional $H(F)$ is just the plug-in estimate $H(F_n)$. In the present paper, $H(F)$ is a real-valued function of a sampling distribution, such as a quantile or a tuning parameter. If H is sufficiently smooth in F, we know that the usual jackknife and delta method estimates for the standard error of $H(F_n)$ are consistent, and even asymptotically efficient among all nonparametric competitors. But what can be said when only a Monte Carlo approximation is available for the theoretical bootstrap estimate $H(F_n)$? Professor Efron's paper makes the important point that valid jackknife and delta method estimates for the standard error of $H(F_n)$ can be calculated from the same bootstrap samples that were used to approximate $H(F_n)$.

It is natural to consider an extension of Efron's results. When H is smoother, consistent jackknife and delta method estimates exist for the bias, skewness and kurtosis of $H(F_n)$. By fitting an Edgeworth expansion to these moments, we can estimate the sampling distribution of $H(F_n)$. This estimate can be asymptotically equivalent to the theoretical bootstrap distribution of $H(F_n)$, as shown in Beran (1984). In Efron's context, this means that a second round of bootstrapping might be *completely* replaced by a suitable Edgeworth estimate, with coefficients obtained from the jackknife or delta method. Of course, difficulties may arise in this more ambitious approach: estimates of skewness or kurtosis are not very good in small samples, and the effects of a finite number of first-round bootstrap samples on the fitted Edgeworth expansions need to be analysed.

Efron's introduction rightly notes that jackknife methods for assessing the variability of bootstrap estimates are narrower in scope than double-bootstrap methods. In particular, jackknife methods are designed for asymptotically normal estimates $H(F_n)$. Even asymptotic normality is not enough. It would be tricky, for instance, to make jackknife methods work for a function of the nonparametric bootstrap distribution of a sample quantile.

Hall's (1986) analysis indicates that using $B = 1000$ bootstrap samples biases coverage probability of a 90% bootstrap confidence interval by about 0.001. Using $B = 999$ removes this Monte Carlo bias easily. It would be of interest to hear Professor Efron's reasons for choosing $B = 1000$ in his confidence interval examples.

The following contributions were received in writing after the meeting.

Hung Chen (State University of New York, Stony Brook) **and Hung Kung Liu** (National Institute of Standards and Technology, Gaithersburg): We shall only comment on Professor Efron's use of the jackknife to assess the accuracy of bootstrap statistics. As is well known, the jackknife method may not reveal the effect of multiple outliers in the data. Also, sequential deletions of observations cannot be relied on to reveal multiple outliers. A possible cure is to delete several observations at once. But this extension is a computational nightmare, since there are many possibilities to be explored. However, this approach is computationally feasible if equation (2.12) can still be used. Can the author comment on the use of equation (2.12) in this case?

Next, we compare the bootstrap-after-bootstrap with the jackknife-after-bootstrap by considering the estimation of the population mean by the sample mean under a gross error model. In our simulation, 50 observations are drawn from the standard normal distribution and then we add 3 to the first two observations, so that the final data consist of two widely separate clusters with size 48 and 2 respectively. The number of bootstrap replications in the jackknife-after-bootstrap is 6000 and in the bootstrap-after-bootstrap it is 1000 for the first-level bootstrap and 2000 for the second-level bootstrap. Fig. 10 and Fig. 11 illustrate our calculations for the jackknife-after-bootstrap with the deletion of one and two observations respectively. In both figures, we follow the same description as for Fig. 3 with the broken lines indicating the 5th, 10th, 16th and 32nd bootstrap percentiles in ascending order. Fig. 12 shows a different summary of the results in Fig. 11. The curves from left to right are the estimated density functions of the four percentiles. Each is based on 1225 observations. Fig. 13 is the same density plot based on results from the bootstrap-after-bootstrap.

One observation from Figs 12 and 13 is that the overlapping patterns are quite different. Consider these methods as repeated resampling procedures. An intuitive explanation lies in the number of outliers that may appear in each of the first-level samples. A jackknife sample may contain at most two outliers, whereas a bootstrap sample may contain any number of outliers with different probabilities.

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

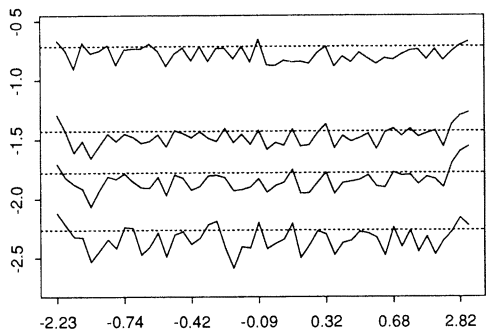

Fig. 10

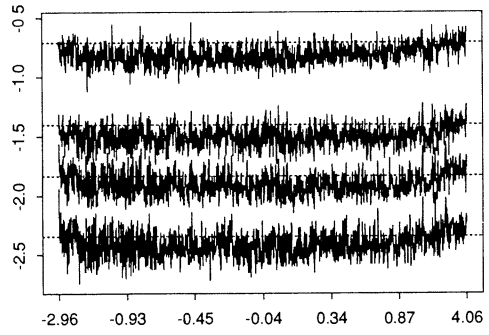

Fig. 11

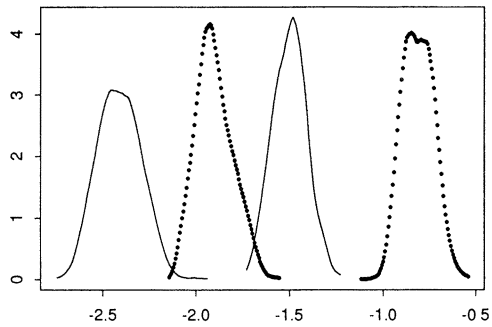

Fig. 12

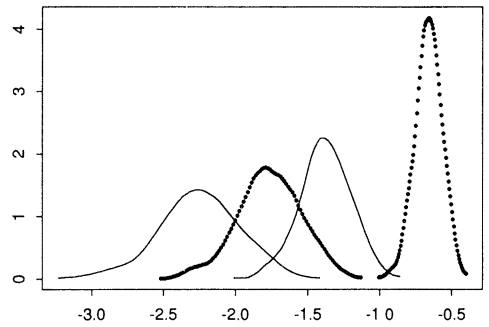

Fig. 13

The variability from bootstrapping a jackknife sample is expected to be less than the variability from bootstrapping a bootstrap sample. Therefore these results indicate that the bootstrap-after-bootstrap may provide information that cannot be revealed by the jackknife-after-bootstrap. However, the disadvantages of the bootstrap-after-bootstrap include the substantial extra computations and the difficulty in singling out atypical observations.

David Cox (Nuffield College, Oxford): Professor Efron's paper is characteristically impressive. The rather sweeping statement in the second sentence is presumably restricted to problems with a strong 'independent and identically distributed' component and the extension of his techniques to the assessment of precision in more complex systems raises interesting issues. The analysis of the nuclear physics data is dependent on an assumption such as symmetry which, although consistent with the data, seems rather strong. For otherwise the different trimmed means are estimating different parameters and there is no reason to think that because a parameter is estimated with maximum precision it is therefore the most meaningful.

Daniela De Angelis (Università di Roma 'La Sapienza'): Bootstrap methods are generally either implemented in a completely nonparametric fashion, by resampling from the observed data, or by parametric sampling from a fitted distribution. An intermediate approach, not discussed in this paper, but due to Professor Efron himself (Efron, 1979), involves resampling from a nonparametrically smoothed version of the empirical distribution. Use of such a smoothed bootstrap can lead to a considerable reduction in the mean-squared error of the bootstrap estimator in many problems, though in most common applications, where Hall's (1988) 'smooth function model' applies, the benefit due to smoothing is only of second order (De Angelis and Young, 1990a). First-order improvement in the performance of the bootstrap estimator can, however, be achieved in certain cases where the smooth function model does not apply, such as the problem of estimating the variance of a sample quantile: see Hall *et al.* (1989).

The ability to reproduce in practice the theoretical advantages of smoothing depends crucially on being able to choose an appropriate value for some smoothing parameter. A data-based procedure for such a choice is discussed by De Angelis and Young (1990a, b) and by Bowman and Hall (1991). The

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

key idea of the method is that of constructing, from the given sample data, a bootstrap estimate of the mean-squared error associated with the smoothed bootstrap estimation, and of choosing the smoothing parameter for the estimation itself to minimize this error estimate. The procedure has much in common with the method described in Section 4 of Professor Efron's paper, the main difference being that since the estimator is now a bootstrap estimator, which will generally require resampling in its construction, the estimation of the mean-squared error is already a bootstrap-after-bootstrap calculation. In many simple cases it is possible to construct explicitly, or at least to approximate, the error estimate without the need for a double level of bootstrap sampling, though the approach is perhaps most fruitful where computational short-cuts are most difficult to obtain (De Angelis and Young, 1990a). Although the method performs most effectively, Professor Efron's paper may provide the basis for a further refinement of the empirical smoothing, through the use of a jackknife-after-bootstrap-after-bootstrap calculation to investigate how well determined the choice of smoothing parameter is.

Thomas J. DiCiccio and Michael A. Martin (Stanford University): Professor Efron's development of the bootstrap has revolutionized the way that data are used. His article emphasizes that statisticians must think critically about the bootstrap quantities that they use, but that such an assessment need not incur much more computational cost than that expended in obtaining the original quantities.

How might jackknife-after-bootstrap calculations be applied generally? One possible way in which the technique might be adapted to a wider class of problems is the use of a delete-d jackknife-after-bootstrap, but such a technique might rival double bootstrapping in terms of computation.

The application of the jackknife to estimate the standard error of bootstrap-t percentiles is intriguing. There, the function ϕ is discontinuous, so there is some question about whether the jackknife-after-bootstrap estimate is consistent. It is well known that the delete-1 jackknife yields inconsistent variance estimates for samples quantiles. A useful heuristic approach is to view the jackknife-after-bootstrap technique as similar to a delete-d jackknife on the set of resamples T_b^*, $b=1, \ldots, B$, where $d \approx B/2$, rather than as a delete-1 jackknife on the original data. This value of d follows from noting that, typically, the number of distinct resamples is $\binom{2n-1}{n}$ and the number of resamples not containing the ith data point is $\binom{2n-2}{n}$. Consequently, for large n, each of the $\hat{\gamma}_{(i)}$ is based on approximately half of all possible resamples. Analogously, each of the $\tilde{\gamma}_{(i)}$ is based on roughly half of the B resamples drawn. Results of Shao and Wu (1989) indicate that, provided $B^{1/2}/d \to 0$, the delete-d jackknife quantile variance estimate based on all $\binom{B}{d}$ delete-d jackknife samples is consistent. Further, the variance estimate based on a *balanced* set of subsets of the $\binom{B}{d}$ jackknife samples is also consistent; see Shao and Wu (1989) for a discussion of balanced sets. Now, the set of subsets of resamples on which the $\hat{\gamma}_{(i)}$ are based is far from balanced—resamples containing n copies of any particular data point appear $n-1$ times in the set of subsets, whereas the most likely resample, the original data, does not appear. Similarly, the set of samples on which the $\tilde{\gamma}_{(i)}$ is based is unlikely to be balanced. In principle, the set of samples corresponding to $\tilde{\gamma}_{(i)}$, $i=1, \ldots, n$, could be embedded in a balanced set of subsamples, but at least B such subsamples are necessary to ensure balance. Can the requirement that the set of subsamples be balanced be relaxed? If not, it appears that at least B subsamples are needed to ensure consistency in this example. Procedures employing B subsamples are about as computationally demanding as double-bootstrap algorithms.

Nicholas I. Fisher (CSIRO Division of Mathematics and Statistics, Lindfield) **and Peter Hall** (Australian National University, Canberra): As always, Professor Efron has demonstrated his extraordinary ability to uncover and develop significant new research areas. His work on jackknife-after-bootstrap methods will undoubtedly prove seminal in stimulating research activity.

Along with a large number of challenging suggestions for future avenues of research, a number of gauntlets are thrown down. We wish to comment on just one of them here, stemming from the comparison of jackknife-after-bootstrap and bootstrap-after-bootstrap methods. Now, the efficacy of the jackknife is founded, analytically, on correcting a Taylor expansion for the 'linear' term. This presupposes a certain level of smoothness of the underlying statistical functional. In particular, jackknife methods do not perform well with indicator functions. For example, the leave-one-out jackknife estimate of the distribution function of a statistic is usually not consistent.

More specifically, if $\bar{X}$ denotes the mean of a random sample of size n from a distribution with mean μ and finite variance, and if $\bar{X}_i$ represents the mean of the $(n-1)$-sample obtained by deleting the ith data value, then the leave-one-out jackknife estimate of $p = \text{Prob}(\bar{X} \leqslant n\mu + n^{1/2}x)$ is

$$\hat{p} = n^{-1} \sum_{i=1}^{n} I(\bar{X}_i \leqslant n\mu + n^{1/2}x) = I(\bar{X} \leqslant n\mu + n_{1/2}x) + \Delta$$

where $\Delta \to 0$ in probability. Therefore $\hat{p}$ does not converge to p.

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

Similarly, the usefulness of jackknife methods for correcting errors in, say, the percentile method bootstrap estimate of a distribution function seems to be severely limited. It is in such cases, where estimation of the expected value of a discontinuous quantity is the main issue, that techniques based on the bootstrap-after-bootstrap can be deployed to good advantage.

Is the 'iterated bootstrap' as computationally impractical as some suggest? In samples of small to moderate size (where bootstrap methods come into their own) and for problems involving univariate statistics (where large values of B are relatively unimportant) the iterated bootstrap is already an eminently practical tool, using only the sort of late 20th-century technology which many statisticians currently have on their desks.

It will be helpful to give a new intuitive explanation of the way that the double bootstrap affects calibration, as the technique is not yet widely understood.

Let $\hat{\theta}$ be an estimate of a parameter θ, based on data $\mathscr{X}$, and suppose that we seek to approximate the sampling distribution of $\hat{\theta}$ by bootstrap methods. Let $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ be the corresponding bootstrap estimates calculated from resamples $\mathscr{X}_1^*, \ldots, \mathscr{X}_B^*$, and denote their order statistics by $\hat{\theta}_{(1)}^* \leqslant \ldots \leqslant \hat{\theta}_{(B)}^*$.

Then $\hat{\theta}_{(j)}^*$ purports to be a value such that

$$P(\theta \leqslant \hat{\theta}_{(j)}^*) = j/(B+1), \qquad j = 1, \ldots, B.$$

This may not be correct, to a satisfactory approximation, especially in small samples, so we might seek a closer approximation to $P(\theta \leqslant \hat{\theta}_{(j)}^*)$. For each resample $\mathscr{X}_i^*$, a complete bootstrap experiment based on b resamples from $\mathscr{X}_i^*$ can be carried out, leading to the B sets of ordered double-bootstrap estimates $\hat{\theta}_{i(1)}^{**} \leqslant \ldots \leqslant \hat{\theta}_{i(b)}^{**}$, $i = 1, \ldots, B$.

Sacrificing generality on the altar of simplicity, suppose that $b = B$. Then a better approximation to $P(\theta \leqslant \hat{\theta}_{(j)}^*)$ is given by

$$B^{-1} \sum_{i=1}^{B} I(\hat{\theta} \leqslant \hat{\theta}_{i(j)}^{**}), \qquad j = 1, \ldots, B.$$

The **author** replied later, in writing, as follows.

Let me restate the main idea of the paper. A bootstrap analysis has been run to assess the accuracy of some primary statistical results. This produces bootstrap statistics, like standard errors or confidence intervals, which are assessments of error for the primary results. By suitably rearranging the output of the bootstrap analysis we obtain jackknife estimates of accuracy for the bootstrap estimates themselves.

The advantage here is simplicity and ease of computation. The same analysis that gives the bootstrap statistics also gives their accuracies. The disadvantage is that we only derive approximate jackknife accuracy estimates.

Both adjectives, 'approximate' and 'jackknife', limit the jackknife-after-bootstrap's (JAB's) usefulness. For the approximations to be tolerable, the original bootstrap analysis must be fairly large, of the order of $B = 1000$ replications. Professor Tibshirani's example shows that small values of B can terribly inflate the JAB estimates of error.

Using the jackknife or the delta method for error analysis virtually restricts attention to standard errors and biases, which are the only error quantities easily obtained from influence functions. The bootstrap-after-bootstrap method is a more flexible approach, and likely to give more satisfactory answers, as shown for example by Dr Worton's reanalysis of the tau data. It was not included here because of my self-imposed restriction to error estimates that can be obtained from the original bootstrap computations. (This restriction rules out Professor Dawid's suggestion.) I apologize for any impression of negativity about iterated bootstrap methods. Several of the contributions show how promising this area has become: see Fisher and Hall, Young, De Angelis and Worton.

It was early 1990 when I wrote 'At present, bootstrap-after-bootstrap seems too computationally intensive for routine use'. This is probably still true, but an inexorable tide of inexpensive superfast computing continues to wash in. My Sun 3/50 personal computer is obsolete these days. I expect, and hope, that the quotation above will soon be rendered false.

This still leaves the question of how to use superintensive methods like the bootstrap-after-bootstrap. Some of the uses are clear. As Dr De Angelis points out, if an estimator is 'tuned' by using a bootstrap analysis, then evaluating its standard error probably involves a second level of bootstrapping. The bootstrap calibration idea, neatly explicated by Dr Young and Dr Fisher and Professor Hall, is another promising candidate. The Hinkley–Davison–Worton partial likelihood is a good example of new technology engendering interesting new theory.

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

A simpler question, in the spirit of Professor Atkinson's comments, is how to design an iterated bootstrap experiment. Given a constraint of say 100000 replications, is 1000×100 better or worse than 100×1000, etc.? Still another avenue is the elimination of Monte Carlo methods entirely from one or both of the bootstrap levels, as with Dr Young's saddlepoint methods, or Professor Beran's Edgeworth series.

All this is in apology for not going sufficiently far here in the pursuit of computer-based statistical methodology. Dr Chatfield, and to some extent Professor Tukey, seems to feel that I have gone too far. There can be no quarrel with the advice to get more data whenever possible, but eventually we are still left with a finite data set and inferences to make.

The more I work with the bootstrap and other resampling schemes the more they seem to be an application of classical statistical ideas (like plugging in $\hat{F}$ for F), carried out with the help of computers rather than mathematical approximations. This still may be torture, but it is torture with a classical provenance. In fact the paper offers a check on the bootstrap, to see whether it is generating more detail than the data support. Professor Copas states this point neatly in his last remarks.

Professor Tukey, who probably deserves most of the blame for the current interest in computational methods, worries that we (I) forget good statistical practice in pursuit of the computational muse. He, along with Professor Atkinson, wonders why the law school example was not analysed on the $\tanh^{-1}$-scale. An important point of research into the bootstrap, and other computer-based methods, is the automation of large parts of good statistical practice. For example the percentile intervals, being transformation invariant, work equally well or poorly on any scale. In complicated situations the statistician, even a very good applied statistician, will not know what transformation to make, so it is helpful to be using a transformation invariant method.

Fig. 4 is transformation invariant. Changing from the correlation s to $z = \tanh^{-1}(s)$ maps every point (u, s) in Fig. 4 to $(u, \tanh^{-1}(s))$. The lowest broken horizontal line goes from height 0.549 to height $0.617 = \tanh^{-1}(0.549)$. The lowest jagged curve has height $1.066 = \tanh^{-1}(0.788)$ at point A on the right, still just below the 50% horizontal line, etc. In this sense, point A has the same huge negative influence on the $\tanh^{-1}$-scale, or on any other scale. (The JAB numerical calculations will not be perfectly invariant, because the jackknife is not.)

Professor Copas and Professor Tukey raise an important point about translation families: if $X = \theta + Z$, where Z is centred at 0 but symmetric and long tailed to the *right*, should not the confidence interval for θ be long tailed to the *left* of the observed value $X = x$? This is commonsensical, and the opposite of what the percentile interval does. In fact, the translation model does not extend well to more general confidence interval situations. If X is Poisson(θ) for example, then its distribution is asymmetric to the right, but so is its confidence interval: likewise for the other familiar one-parameter families. The better version of the percentile method called BC_a does the right thing in these families, and also in translation families. See Section 10 of Efron (1987).

I could have plotted the BC_a confidence interval limits in Fig. 4 instead of the percentile limits, as Professor Copas hints. Length and shape in Table 1, and the subsequent JAB analysis, could just as easily (well, almost as easily; see the next paragraph) have referred to the BC_a intervals. For some reason there has not been much interest in influence calculations for confidence intervals. Even traditional non-bootstrap confidence intervals can be quite sensitive to outlying data points, often more so than a point estimate. The jackknife or delta method influence functions can be used to uncover these sensitivities.

Now is the time for confession. Professor Tibshirani caught a serious error in my original equation (3.4), since corrected. I originally wrote $\hat{\gamma}_{(i)} = \phi\,[\,T(\mathbf{x}^*, \hat{F})|\, P_i = 0\,]$, forgetting that '$\hat{F}$' needs be modified to '$\hat{F}_{(i)}$' in the deleted point computations. This affects the bootstrap-t calculations (3.11), where the deleted point statistics T_i^{*b} in approximation (3.5) are $\{s(\mathbf{x}^{*b}) - s(\mathbf{x}_i)\}/d(\mathbf{x}^{*b})$ and not $\{s(\mathbf{x}^{*b}) - s(\mathbf{x})\}/d(\mathbf{x}^{*b})$. (It is easy to calculate T_i^{*b} from the bootstrap statistics $(s(\mathbf{x}^{*b}), d(\mathbf{x}^{*b}))$ and $s(x_{(i)})$. As stated in the paper, no new bootstrap evaluations are required.) In answer to Professor Copas's reasonable concerns, Fig. 5 as shown is based on the correct version of equation (3.4).

The interesting thing about Fig. 5 is its clear demonstration that the T-statistic (3.8) is not even approximately pivotal here. This raises the prospect of a data-based search for an approximate pivotal, but I do not know how one would carry out such a search.

Fig. 14 shows the entire population of 82 law schools from which the 15 schools in Fig. 1(a) were selected. Point A still looks like an outlier, though less egregious. Dr Jennison's reasonable guess that it represents a separate mode of the bivariate distribution turns out to be unfounded. (Professor Copas's belief that point A is not so unusual is vindicated.) Our sample of 15 did not include the spectacular

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

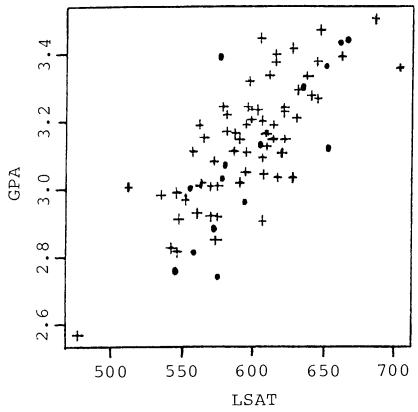

Fig. 14. (LSAT, GPA) scores for all 82 law schools: ●, school selected for Fig. 1(a)

outlier at the lower left-hand side, which is the only point of the 82 that looks incommensurate with bivariate normality.

Professor Cox raises a very reasonable question about the tau example: why should we be comparing different trimmed means according to their variances if they are estimating different quantities? One answer is that all trimmed means are translation equivariant, so that their power in detecting translation alternatives will be inversely proportional to their variances. Another possible answer concerns the parameter Δ, expression (4.1). The corresponding estimator $\hat{\Delta}$ is a contrast of two similar parts, $\widehat{\text{decay}}_1$ and $(\widehat{\text{decay}}_\rho + \ldots)$, and as such is likely to have nearly the same mean for any trimming proportions (so we need only worry about variance). For example, if X and Y are independent gamma variates of degrees 10 and 4 respectively, then the difference of their distribution trimmed means ranges from 6 for the true mean to 5.996 for the median.

The question of parametric *versus* nonparametric JAB analyses is raised by Dr Henderson, who makes an interesting point about intermediate semiparametric situations. It is worth noting that variability estimates do not necessarily increase as we go from parametric to nonparametric assessments. A parametric model that is wrong can easily force an overly large assessment of variance, as later revealed by a nonparametric, or semiparametric, bootstrap analysis.

I like Dr Critchley's suggestion for an interplay between diagnostic and bootstrap methods. Figs 4 and 5 were conceived in that spirit. Both Dr Critchley and Professor Atkinson wonder about designing bootstrap experiments that are more efficient than pure Monte Carlo methods. The ultimate designed bootstrap would be a Simpson's rule for integration over the resampling simplex, but alas this seems impossible. Less dramatic improvements over simple Monte Carlo methods can be found in Efron (1990), as well as in Professor Atkinson's references.

Bootstrap statistics (2.12) are more complicated functions of $\mathbf{x}$ than primary statistics like means, quantiles, etc. As a compensating virtue, they tend to be smoother functions of the underlying distributions F than their primary counterparts. (This makes them good candidates for jackknifing.) In particular, the quantiles of the bootstrap distribution of some statistic $\hat{\theta} = t(\hat{F})$ are usually more smoothly behaved than the quantiles of $\mathbf{x}$ itself. This is in answer to the concerns of Dr DiCiccio and Dr Martin, and also of Dr Fisher and Professor Hall, regarding some of the applications in the paper.

Dr Chen and Dr Liu, as well as Dr Davison and Professor Atkinson, worry about the masking of outliers by other outliers. Chen and Liu make a case for removing the points two at a time instead of singly. About 13% of the bootstrap samples will be missing any two given points, compared with 37% for any one point. We would need B to be of the order of 3000 to make the approach in approximation (3.5) feasible for two-at-a-time removal. The equivalent of Fig. 7 might help to reduce this number.

Dr Jennison overestimates my self-control if he thinks that I can resist answering his last question. I have started to use resampling methods regularly in my own applied work and find them very satisfactory for Jennison's 'mundane' problems. It is nice to be able to supplement a standard parametric error analysis with a nonparametric bootstrap investigation, if for no other reason than that often my clients better understand the bootstrap analysis. Resampling methods may look strange to us, because of

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions

our training in mathematical statistics, but they seem to strike most scientists as little more than common sense.

Bootstrap methods come into their own in complicated situations, which, as with the tau data, may not necessarily be large sample situations. There is a natural tendency to be less critical about a complicated analysis, e.g. a robust regression following a variable selection procedure, than a simple analysis, just because it is more difficult to do the complicated analysis. Resampling methods can help the applied statistician to avoid this obvious logical fallacy. Bootstrap methods themselves are complicated procedures, whose complications should not be an excuse for uncritical acceptance. That is the point of this paper, and many of the commentaries.

I am grateful to the Society and the commentators for arranging this discussion.

REFERENCES IN THE DISCUSSION

Atkinson, A. C. and Weisberg, S. (1991) Simulated annealing for the detection of multiple outliers using least squares and least median of squares fitting. In *Directions in Robust Statistics and Diagnostics I* (eds W. Stahel and S. Weisberg). New York: Springer.

Beran, R. (1984) Jackknife approximations to bootstrap estimates. *Ann. Statist.*, **12**, 101–118.

Bowman, A. W. and Hall, P. G. (1991) Empirical determination of smoothing for the bootstrap.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. J. (1985) *Classification and Regression Trees*. Belmont: Wadsworth.

Cook, R. D. (1986) Assessment of local influence (with discussion). *J. R. Statist. Soc.* B, **48**, 133–169.

Davis, L. (ed.) (1991) *Handbook of Genetic Algorithms*. New York: van Nostrand-Rheinhold.

Davison, A. C., Hinkley, D. V. and Schechtman, E. (1986) Efficient bootstrap simulation. *Biometrika*, **73**, 555–566.

Davison, A. C., Hinkley, D. V. and Worton, B. J. (1991) Bootstrap likelihoods, to be published.

De Angelis, D. and Young, G. A. (1990a) Smoothing the bootstrap. *Int. Statist. Rev.*, to be published.

——— (1990b) Bootstrapping the correlation coefficient: a comparison of smoothing strategies. *J. Statist. Computn Simuln*, to be published.

DiCiccio, T. J., Martin, M. A. and Young, G. A. (1990a) Analytic approximations to bootstrap distribution functions using saddlepoint methods. *Technical Report 356*. Department of Statistics, Stanford University.

——— (1990b) Analytic approximations for iterated bootstrap confidence intervals. *Technical Report 361*. Department of Statistics, Stanford University.

——— (1991) Fast and accurate approximate double bootstrap confidence intervals. *Biometrika*, to be published.

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

——— (1983) Estimating the error rate of a prediction rule: improvements in cross-validation. *J. Am. Statist. Ass.*, **78**, 316–331.

——— (1987) Better bootstrap confidence intervals and bootstrap approximations. *J. Am. Statist. Ass.*, **82**, 171–185.

——— (1990) More efficient bootstrap computations. *J. Am. Statist. Ass.*, **85**, 79–89.

Goldberg, D. E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading: Addison-Wesley.

Graham, R. L., Hinkley, D. V., John, P. W. M. and Shi, S. (1990) Balanced design of bootstrap simulations. *J. R. Statist. Soc.* B, **52**, 185–202.

Haldane, J. B. S. (1949) A note on non-normal correlation. *Biometrika*, **36**, 467–468.

Hall, P. (1986) On the number of bootstrap simulations required to construct a confidence interval. *Ann. Statist.*, **14**, 1453–1462.

——— (1988) Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.*, **16**, 927–985.

Hall, P. G., DiCiccio, T. J. and Romano, J. P. (1989) On smoothing and the bootstrap. *Ann. Statist.*, **17**, 692–704.

Miller, A. J. (1990) *Subset Selection in Regression*. London: Chapman and Hall.

Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

——— (1990) Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**, 90–120.

Shao, J. and Wu, C. F. J. (1989) A general theory for jackknife variance estimation. *Ann. Statist.*, **17**, 1176–1197.

Wynn, H. P. and Ogbonmwan, S. M. (1986) Discussion on Jackknife, bootstrap and other resampling methods in regression analysis (by C. F. J. Wu). *Ann. Statist.*, **14**, 1340–1343.

This content downloaded from 72.33.153.245 on Tue, 14 Oct 2014 12:46:52 PM
All use subject to JSTOR Terms and Conditions