Statistical Science
1996, Vol. 11, No. 3, 189–228

Bootstrap Confidence Intervals

Thomas J. DiCiccio and Bradley Efron

Abstract. This article surveys bootstrap methods for producing good approximate confidence intervals. The goal is to improve by an order of magnitude upon the accuracy of the standard intervals $\hat{\theta} \pm z^{(\alpha)}\hat{\sigma}$, in a way that allows routine application even to very complicated problems. Both theory and examples are used to show how this is done. The first seven sections provide a heuristic overview of four bootstrap confidence interval procedures: BC_a, bootstrap-t, ABC and calibration. Sections 8 and 9 describe the theory behind these methods, and their close connection with the likelihood-based confidence interval theory developed by Barndorff-Nielsen, Cox and Reid and others.

Key words and phrases: Bootstrap-t, BC_a and ABC methods, calibration, second-order accuracy

1. INTRODUCTION

Confidence intervals have become familiar friends in the applied statistician's collection of data-analytic tools. They combine point estimation and hypothesis testing into a single inferential statement of great intuitive appeal. Recent advances in statistical methodology allow the construction of highly accurate approximate confidence intervals, even for very complicated probability models and elaborate data structures. This article discusses bootstrap methods for constructing such intervals in a routine, automatic way.

Two distinct approaches have guided confidence interval construction since the 1930's. A small catalogue of exact intervals has been built up for special situations, like the ratio of normal means or a single binomial parameter. However, most confidence intervals are approximate, with by far the favorite approximation being the *standard interval*

$$\hat{\theta} \pm z^{(\alpha)}\hat{\sigma}. \tag{1.1}$$

Here $\hat{\theta}$ is a point estimate of the parameter of interest θ, $\hat{\sigma}$ is an estimate of $\hat{\theta}$'s standard deviation, and $z^{(\alpha)}$ is the 100αth percentile of a normal deviate, $z^{(0.95)} = 1.645$ and so on. Often, and always in this paper, $\hat{\theta}$ and $\hat{\sigma}$ are obtained by maximum likelihood theory.

The standard intervals, as implemented by maximum likelihood theory, are a remarkably useful tool. The method is completely automatic: the statistician inputs the data, the class of possible probability models and the parameter of interest; a computer algorithm outputs the intervals (1.1), with no further intervention required. This is in notable contrast to the construction of an exact interval, which requires clever thought on a problem-by-problem basis when it is possible at all.

The trouble with standard intervals is that they are based on an asymptotic approximation that can be quite inaccurate in practice. The example below illustrates what every applied statistician knows, that (1.1) can considerably differ from exact intervals in those cases where exact intervals exist. Over the years statisticians have developed tricks for improving (1.1), involving bias-corrections and parameter transformations. The bootstrap confidence intervals that we will discuss here can be thought of as automatic algorithms for carrying out these improvements without human intervention. Of course they apply as well to situations so complicated that they lie beyond the power of traditional analysis.

We begin with a simple example, where we can compute the bootstrap methods with an exact interval. Figure 1 shows the *cd4 data*: 20 HIV-positive subjects received an experimental antiviral drug; cd4 counts in hundreds were recorded for each subject at baseline and after one year of treatment, giv-

Thomas J. DiCiccio is Assocciate Professor, Department of Social Statistics, 358 Ives Hall, Cornell University, Ithaca, New York 14853-3901 (email: tjd9@cornell.edu). Bradley Efron is Professor, Department of Statistics and Department of Health Research and Policy, Stanford University, Stanford, California 94305-4065 (e-mail: brad@playfair.stanford.edu).

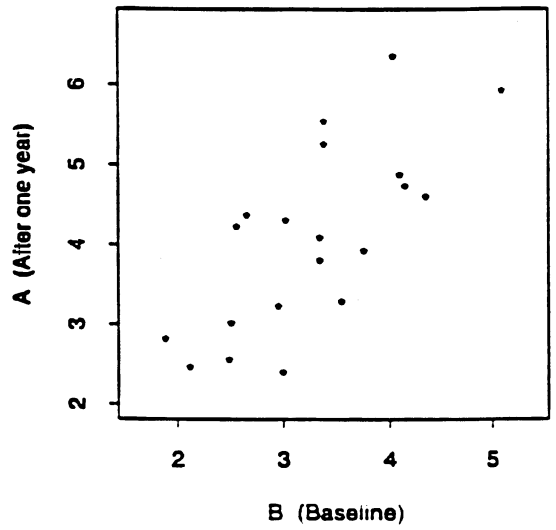

FIG. 1. *The cd4 data; cd4 counts in hundreds for* 20 *subjects, at baseline and after one year of treatment with an experimental anti-viral drug; numerical values appear in Table* 1.

ing data, say, $x_i = (B_i, A_i)$ for $i = 1, 2, \ldots, 20$. The data is listed in Table 1. The two measurements are highly correlated, having sample correlation coefficient $\hat{\theta} = 0.723$.

What if we wish to construct a confidence interval for the true correlation θ? We can find an exact interval for θ if we are willing to assume bivariate normality for the (B_i, A_i) pairs,

$$\binom{B_i}{A_i} \sim_{\text{i.i.d.}} N_2(\lambda, \Gamma) \quad \text{for} \quad i = 1, 2, \ldots, 20, \tag{1.2}$$

where λ and Γ are the unknown expectation vector and covariance matrix. The exact central 90% interval is

$$(\hat{\theta}_{\text{EXACT}}[0.05], \quad \hat{\theta}_{\text{EXACT}}[0.95]) = (0.47, 0.86). \tag{1.3}$$

This notation emphasizes that a two-sided interval is intended to give correct coverage at both endpoints, two 0.05 noncoverage probabilities in this case, not just an overall 0.10 noncoverage probability.

The left panel of Table 2 shows the exact and standard intervals for the correlation coefficient of the cd4 data, assuming the normal model (1.2). Also shown are approximate confidence intervals based on three different (but closely related) bootstrap methods: ABC, BC_a and bootstrap-t. The ABC and BC_a methods match the exact interval to two decimal places, and all of the bootstrap intervals are more accurate than the standard. The examples and theory that follow are intended to show that this is no accident. The bootstrap methods make computer-based adjustments to the standard interval endpoints that are guaranteed to improve the coverage accuracy by an order of magnitude, at least asymptotically.

TABLE 1
The cd4 data, as plotted in Figure 1

Subject	Baseline	One year	Subject	Baseline	One year
1	2.12	2.47	11	4.15	4.74
2	4.35	4.61	12	3.56	3.29
3	3.39	5.26	13	3.39	5.55
4	2.51	3.02	14	1.88	2.82
5	4.04	6.36	15	2.56	4.23
6	5.10	5.93	16	2.96	3.23
7	3.77	3.93	17	2.49	2.56
8	3.35	4.09	18	3.03	4.31
9	4.10	4.88	19	2.66	4.37
10	3.35	3.81	20	3.00	2.40

The exact interval endpoints [0.47, 0.86] are defined by the fact that they "cover" the observed value $\hat{\theta} = 0.723$ with the appropriate probabilities,

$$\text{Prob}_{\theta=0.47}\{\hat{\theta} > 0.723\} = 0.05 \tag{1.4}$$

and

$$\text{Prob}_{\theta=0.86}\{\hat{\theta} > 0.723\} = 0.95. \tag{1.5}$$

Table 2 shows that the corresponding probabilities for the standard endpoints [0.55, 0.90] are 0.12 and 0.99. The standard interval is far too liberal at its lower endpoint and far too cautious at its upper endpoint. This kind of error is particularly pernicious if the confidence interval is used to test a parameter value of interest like $\theta = 0$.

Table 2 describes the various confidence intervals in terms of their length and right–left asymmetry around the point estimate $\hat{\theta}$,

$$\begin{aligned} \text{length} &= \hat{\theta}[0.95] - \hat{\theta}[0.05], \\ \text{shape} &= \frac{\hat{\theta}[0.95] - \hat{\theta}}{\hat{\theta} - \hat{\theta}[0.05]}. \end{aligned} \tag{1.6}$$

The standard intervals always have shape equal to 1.00. It is in this way that they err most seriously. For example, the exact normal-theory interval for Corr has shape equal to 0.52, extending twice as far to the left of $\hat{\theta} = 0.723$ as to the right. The standard interval is much too optimistic about ruling out values of θ below $\hat{\theta}$, and much too pessimistic about ruling out values above $\hat{\theta}$. This kind of error is automatically identified and corrected by all the bootstrap confidence interval methods.

There is no compelling reason to assume bivariate normality for the data in Figure 1. A *nonparametric* version of (1.2) assumes that the pairs (B_i, A_i)

TABLE 2

Exact and approximate confidence intervals for the correlation coefficient, cd4 data; $\hat{\theta} = 0.723$: the bootstrap methods ABC, BC_a, bootstrap-t and calibrated ABC are explained in Sections 2–7; the ABC and BC_a intervals are close to exact in the normal theory situation (left panel); the standard interval errs badly at both endpoints, as can be seen from the coverage probabilities in the bottom rows

	Normal theory					Nonparametric				
	Exact	ABC	BC_a	Bootstrap-t	Standard	ABC	BC_a	Bootstrap-t	Calibrated	Standard
0.05	0.47	0.47	0.47	0.45	0.55	0.56	0.55	0.51	0.56	0.59
0.95	0.86	0.86	0.86	0.87	0.90	0.83	0.85	0.86	0.83	0.85
Length	0.39	0.39	0.39	0.42	0.35	0.27	0.30	0.35	0.27	0.26
Shape	0.52	0.52	0.54	0.52	1.00	0.67	0.70	0.63	0.67	1.00
Cov 05	0.05	0.05	0.05	0.04	0.12					
Cov 95	0.95	0.95	0.95	0.97	0.99					

are a random sample ("i.i.d.") from some unknown bivariate distribution F,

$$\binom{B_i}{A_i} \sim_{\text{i.i.d.}} F, \quad i = 1, 2, \ldots, n, \tag{1.7}$$

$n = 20$, without assuming that F belongs to any particular parametric family. Bootstrap-based confidence intervals such as ABC are available for nonparametric situations, as discussed in Section 6. In theory they enjoy the same second-order accuracy as in parametric problems. However, in some nonparametric confidence interval problems that have been examined carefully, the small-sample advantages of the bootstrap methods have been less striking than in parametric situations. Methods that give third-order accuracy, like the bootstrap calibration of an ABC interval, seem to be more worthwhile in the nonparametric framework (see Section 6).

In most problems and for most parameters there will not exist exact confidence intervals. This great gray area has been the province of the standard intervals for at least 70 years. Bootstrap confidence intervals provide a better approximation to exactness in most situations. Table 3 refers to the parameter θ defined as the maximum eigenvalue of the covariance matrix of (B, A) in the cd4 experiment,

$$\theta = \text{maximum eigenvalue } \{\text{cov}(B, A)\}. \tag{1.8}$$

The maximum likelihood estimate (MLE) of θ, assuming either model (1.2) or (1.7), is $\hat{\theta} = 1.68$. The bootstrap intervals extend further to the right than to the left of $\hat{\theta}$ in this case, more than 2.5 times as far under the normal model. Even though we have no exact endpoint to serve as a "gold standard" here, the theory that follows strongly suggests the superiority of the bootstrap intervals. Bootstrapping involves much more computation than the standard intervals, on the order of 1,000 times more, but the algorithms are completely automatic, requiring no more thought for the maximum eigenvalue than the correlation coefficient, or for any other parameter.

One of the achievements of the theory discussed in Section 8 is to provide a reasonable theoretical gold standard for approximate confidence intervals. Comparison with this gold standard shows that the bootstrap intervals are not only asymptotically more *accurate* than the standard intervals, they are also more *correct*. "Accuracy" refers to the coverage errors: a one-sided bootstrap interval of intended coverage α actually covers θ with probability $\alpha + O(1/n)$, where n is the sample size. This is second-order accuracy, compared to the slower first-order accuracy of the standard intervals, with coverage probabilites $\alpha + O(1/\sqrt{n})$. However confidence intervals are supposed to be inferentially correct as well as accurate. Correctness is a harder property to pin down, but it is easy to give examples of incorrectness: if $x_1, x_2, \ldots, x_n$ is a random sample from a normal distribution $N(\theta, 1)$, then $(\min(x_i), \max(x_i))$ is an exactly accurate two-sided confidence interval for θ of coverage probability $1 - 1/2^{n-1}$, but it is incorrect. The theory of Section 8 shows that all of our better confidence intervals are second-order correct as well as second-order accurate. We can see this improvement over the standard intervals on the left side of Table 2. The theory says that this improvement exists also in those cases like Table 3 where we cannot see it directly.

2. THE BC_a INTERVALS

The next six sections give a heuristic overview of bootstrap confidence intervals. More examples are presented, showing how bootstrap intervals can be routinely constructed even in very complicated and messy situations. Section 8 derives the second-order properties of the bootstrap intervals in terms of asymptotic expansions. Comparisons with likelihood-based methods are made in Section 9. The bootstrap can be thought of as a convenient way of executing the likelihood calculations in para-

TABLE 3
Approximate 90% central confidence intervals for the maximum eigenvalue parameter (1.7), cd4 data; the bootstrap intervals extend much further to the right of the MLE $\hat{\theta} = 1.68$ than to the left

	Normal theory			Nonparametric			
	ABC	BC_a	Standard	ABC	BC_a	Calibated	Standard
0.05	1.11	1.10	0.80	1.15	1.14	1.16	1.01
0.95	3.25	3.18	2.55	2.56	2.55	3.08	2.35
Length	2.13	2.08	1.74	1.42	1.41	1.92	1.34
Shape	2.80	2.62	1.00	1.70	1.64	2.73	1.00

metric exponential family situations and even in nonparametric problems.

The bootstrap was introduced as a nonparametric device for estimating standard errors and biases. Confidence intervals are inherently more delicate inference tools. A considerable amount of effort has gone into upgrading bootstrap methods to the level of precision required for confidence intervals.

The BC_a method is an automatic algorithm for producing highly accurate confidence limits from a bootstrap distribution. Its effectiveness was demonstrated in Table 2. References include Efron (1987), Hall (1988), DiCiccio (l984), DiCiccio and Romano (1995) and Efron and Tibshirani (1993). A program written in the language S is available [see the note in the second paragraph following (4.14)].

The goal of bootstrap confidence interval theory is to calculate dependable confidence limits for a parameter of interest θ from the bootstrap distribution of $\hat{\theta}$. Figure 2 shows two such bootstrap distributions relating to the maximum eigenvalue parameter θ for the cd4 data, (1.8). The nonparametric bootstrap distribution (on the right) will be discussed in Section 6.

The left panel is the histogram of 2,000 normal-theory bootstrap replications of $\hat{\theta}$. Each replication was obtained by drawing a bootstrap data set analogous to (1.2),

$$\begin{pmatrix} B_i^* \\ A_i^* \end{pmatrix} \sim_{\text{i.i.d.}} N_2(\hat{\lambda}, \hat{\Gamma}), \quad i = 1, 2, \ldots, 20, \tag{2.1}$$

and then computing $\hat{\theta}^*$, the maximum likelihood estimate (MLE) of θ based on the boostrap data. In other words $\hat{\theta}^*$ was the maximum eigenvalue of the empirical covariance matrix of the 20 pairs (B_i^*, A_i^*). The mean vector $\hat{\lambda}$ and covariance matrix $\hat{\Gamma}$ in (2.1) were the usual maximum likelihood estimates for λ and Γ, based on the original data in Figure 1. Relation (2.1) is a *parametric* bootstrap sample, obtained by sampling from a parametric MLE for the unknown distribution F. Section 6 discusses nonparametric bootstrap samples and confidence intervals.

The 2,000 bootstrap replications $\hat{\theta}^*$ had standard deviation 0.52. This is the bootstrap estimate of standard error for $\hat{\theta}$, generally a more dependable standard error estimate than the usual parametric delta-method value (see Efron, 1981). The mean of the 2,000 values was 1.61, compared to $\hat{\theta} = 1.68$, indicating a small downward bias in the Maxeig statistic. In this case it is easy to see that the downward bias comes from dividing by n instead of $n-1$ in obtaining the MLE $\hat{\Gamma}$ of the covariance matrix.

Two thousand bootstrap replications is 10 times too many for estimating a standard error, but not too many for the more delicate task of setting confidence intervals. These bootstrap sample size calculations appear in Efron (1987, Section 9).

The BC_a procedure is a method of setting approximate confidence intervals for θ from the percentiles of the bootstrap histogram. Suppose θ is a parameter of interest; $\hat{\theta}(\mathbf{x})$ is an estimate of θ based on the observed data $\mathbf{x}$; and $\hat{\theta}^* = \hat{\theta}(\mathbf{x}^*)$ is a bootstrap replication of $\hat{\theta}$ obtained by resampling $\mathbf{x}^*$ from an estimate of the distribution governing $\mathbf{x}$. Let $\hat{G}(c)$ be the cumulative distribution function (c.d.f.) of B bootstrap replications $\hat{\theta}^*(b)$,

$$\hat{G}(c) = \#\{\hat{\theta}^*(b) < c\}/B. \tag{2.2}$$

In our case $B = 2{,}000$. The upper endpoint $\hat{\theta}_{BC_a}[\alpha]$ of a one-sided level-α BC_a interval, $\theta \in (-\infty, \hat{\theta}_{BC_a}[\alpha])$ is defined in terms of $\hat{G}$ and two numerical parameters discussed below: the *bias-correction* z_0 and the *acceleration* a (BC_a stands for "bias-corrected and accelerated"). By definition the BC_a endpoint is

$$\hat{\theta}_{BC_a}[\alpha] = \hat{G}^{-1}\Phi\left(z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}\right). \tag{2.3}$$

Here Φ is the standard normal c.d.f, with $z^{(\alpha)} = \Phi^{-1}(\alpha)$ as before. The central 0.90 BC_a interval is given by $(\hat{\theta}_{BC_a}[0.05], \hat{\theta}_{BC_a}[0.95])$. Formula (2.3) looks strange, but it is well motivated by the transformation and asymptotic arguments that follow.

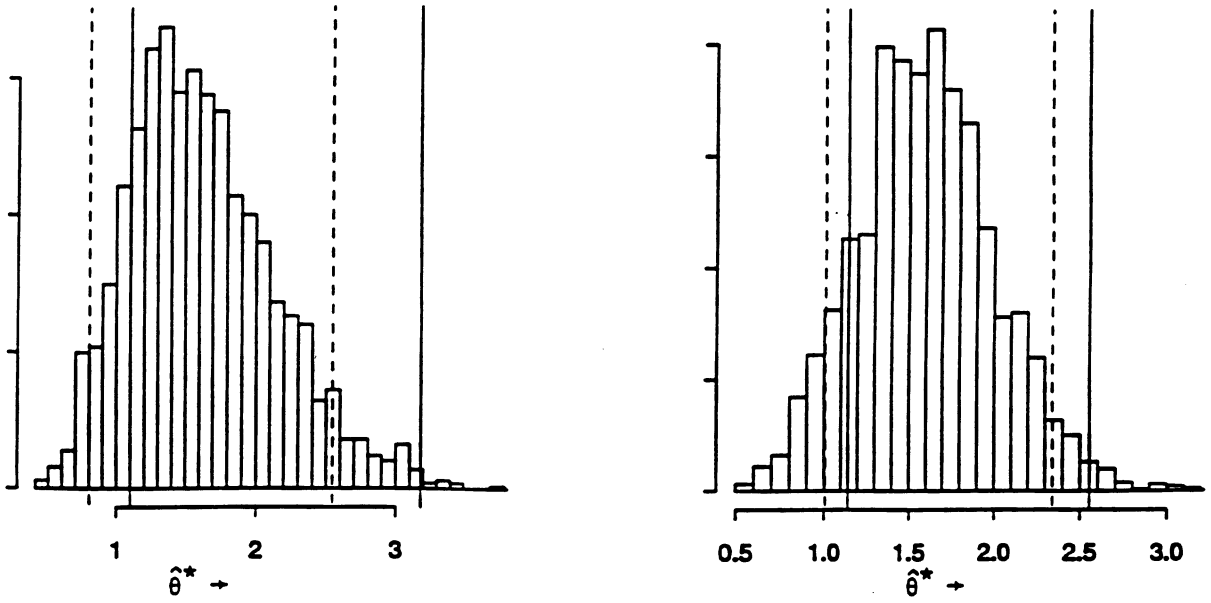

FIG. 2. *Bootstrap distributions for the maximum eigenvalue of the covariance matrix, cd4 data*: (*left*) 2,000 *parametric bootstrap replications assuming a bivariate normal distribution*; (*right*) 2,000 *nonparametric bootstrap replications, discussed in Section* 6. *The solid lines indicate the limits of the* BC_a 0.90 *central confidence intervals, compared to the standard intervals* (*dashed lines*).

If a and z_0 are zero, then $\hat{\theta}_{BC_a}[\alpha] = \hat{G}^{-1}(\alpha)$, the 100αth percentile of the bootstrap replications. In this case the 0.90 BC_a interval is the interval between the 5th and 95th percentiles of the bootstrap replications. If in addition $\hat{G}$ is perfectly normal, then $\hat{\theta}_{BC_a}[\alpha] = \hat{\theta} + z^{(\alpha)}\hat{\sigma}$, the standard interval endpoint. In general, (2.3) makes three distinct corrections to the standard intervals, improving their coverage accuracy from first to second order.

The c.d.f. $\hat{G}$ is markedly long-tailed to the right, on the normal-theory side of Figure 2. Also a and z_0 are both estimated to be positive, $(\hat{a}, \hat{z}_0) = (0.105, 0.226)$, further shifting $\hat{\theta}_{BC_a}[\alpha]$ to the right of $\hat{\theta}_{\text{STAN}}[\alpha] = \hat{\theta} + z^{(\alpha)}\hat{\sigma}$. The 0.90 BC_a interval for θ is

$$(\hat{G}^{-1}(0.157), \hat{G}^{-1}(0.995)) = (1.10, 3.18), \tag{2.4}$$

compared to the standard interval (0.80, 2.55).

The following argument motivates the BC_a definition (2.3), as well as the parameters a and z_0. Suppose that there exists a monotone increasing transformation $\phi = m(\theta)$ such that $\hat{\phi} = m(\hat{\theta})$ is normally distributed for every choice of θ, but possibly with a bias and a nonconstant variance,

$$\hat{\phi} \sim N(\phi - z_0\sigma_\phi, \sigma_\phi^2), \quad \sigma_\phi = 1 + a\phi. \tag{2.5}$$

Then (2.3) gives exactly accurate and correct confidence limits for θ having observed $\hat{\theta}$.

The argument in Section 3 of Efron (1987) shows that in situation (2.5) there is another monotone transformation, say $\xi = M(\theta)$ and $\hat{\xi} = M(\hat{\theta})$, such that $\hat{\xi} = \xi + W$ for all values of ξ, with W always having the same distribution. This is a translation problem so we know how to set confidence limits $\hat{\xi}[\alpha]$ for ξ,

$$\hat{\xi}[\alpha] = \xi - W^{(1-\alpha)}, \tag{2.6}$$

where $W^{(1-\alpha)}$ is the $100(1-\alpha)$th percentile of W. The BC_a interval (2.3) is exactly equivalent to the translation interval (2.6), and in this sense *it is correct as well as accurate*.

The bias-correction constant z_0 is easy to interpret in (2.5) since

$$\text{Prob}\{\hat{\phi} < \phi\} = \Phi(z_0). \tag{2.7}$$

Then $\text{Prob}\{\hat{\theta} < \theta\} = \Phi(z_0)$ because of monotonicity. The BC_a algorithm, in its simplest form, estimates z_0 by

$$\hat{z}_0 = \Phi^{-1}\left\{\frac{\#\{\hat{\theta}^*(b) < \hat{\theta}\}}{B}\right\}, \tag{2.8}$$

Φ^{-1} of the proportion of the bootstrap replications less than $\hat{\theta}$. Of the 2,000 normal-theory bootstrap replications $\hat{\theta}^*$ shown in the left panel of Figure 2, 1179 were less than $\hat{\theta} = 1.68$. This gave $\hat{z}_0 = \Phi^{-1}(0.593) = 0.226$, a positive bias correction since $\hat{\theta}^*$ is biased downward relative to $\hat{\theta}$. An often more accurate method of estimating z_0 is described in Section 4.

The acceleration a in (2.5) measures how quickly the standard error is changing on the normalized scale. The value $\hat{a} = 0.105$ in (2.4), obtained from

formula (4.9) of Section 4, is moderately large. Suppose we think we have moved 1.645 standard errors to the right of $\hat{\phi}$, to

$$\tilde{\phi} = \hat{\phi} + 1.645\sigma_{\hat{\phi}}.$$

Actually though, with $a = 0.105$,

$$\sigma_{\tilde{\phi}} = (1 + 1.645a)\sigma_{\hat{\phi}} = 1.173\sigma_{\hat{\phi}},$$

according to (2.5). For calculating a confidence level, $\tilde{\phi}$ is really only $1.645/1.173 = 1.40$ standard errors to the right of $\hat{\phi}$, considerably less than 1.645. Formula (2.3) automatically corrects for an accelerating standard error. The next section gives a geometrical interpretation of a, and also of the BC_a formula (2.3).

The peculiar-looking formula (2.3) for the BC_a endpoints is designed to give exactly the right answer in situation (2.5), and to give it automatically in terms of the bootstrap distribution of $\hat{\theta}^*$. Notice, for instance, that the normalizing transformation $\hat{\phi} = m(\hat{\theta})$ is not required in (2.3). By comparison, the standard interval works perfectly only under the more restrictive assumption that

$$\hat{\theta} \sim N(\theta, \sigma^2), \tag{2.9}$$

with σ^2 constant. In practice we do not expect either (2.9) or (2.5) to hold exactly, but the broader assumptions (2.5) are likely to be a better approximation to the truth. They produce intervals that are an order of magnitude more accurate, as shown in Section 8.

Formula (2.5) generalizes (2.9) in three ways, by allowing bias, nonconstant standard error and a normalizing transformation. These three extensions are necessary and sufficient to give second-order accuracy,

$$\text{Prob}\{\theta < \hat{\theta}_{BC_a}[\alpha]\} = \alpha + O(1/n), \tag{2.10}$$

compared with $\text{Prob}\{\theta < \hat{\theta}_{\text{STAN}}[\alpha]\} = \alpha + O(1/\sqrt{n})$, where n is the sample size in an i.i.d. sampling situation. This result is stated more carefully in Section 8, which also shows the second-order correctness of the BC_a intervals. Hall (1988) was the first to establish (2.10).

The BC_a intervals are transformation invariant. If we change the parameter of interest from θ to some monotone function of θ, $\phi = m(\theta)$, likewise changing $\hat{\theta}$ to $\hat{\phi} = m(\hat{\theta})$ and $\hat{\theta}^*$ to $\hat{\phi}^* = m(\hat{\theta}^*)$, then the α-level BC_a endpoints change in the same way,

$$\hat{\phi}_{BC_a}[\alpha] = m(\hat{\theta}_{BC_a}[\alpha]). \tag{2.11}$$

The standard intervals are not transformation invariant, and this accounts for some of their practical difficulties. It is well known, for instance, that normal-theory standard intervals for the correlation coefficient are much more accurate if constructed on the scale $\phi = \tanh^{-1}(\theta)$ and then transformed back to give an interval for θ itself. Transformation invariance means that the BC_a intervals cannot be fooled by a bad choice of scale. To put it another way, the statistician does not have to search for a transformation like $\tanh^{-1}$ in applying the BC_a method.

In summary, BC_a produces confidence intervals for θ from the bootstrap distribution of $\hat{\theta}^*$, requiring on the order of 2,000 bootstrap replications $\hat{\theta}^*$. These intervals are transformation invariant and exactly correct under the normal transformation model (2.5); in general they are second-order accurate and correct.

3. THE ACCELERATION *a*

The acceleration parameter a appearing in the BC_a formula (3.2) looks mysterious. Its definition in (2.5) involves an idealized transformation to normality which will not be known in practice. Fortunately a enjoys a simple relationship with Fisher's score function which makes it easy to estimate. This section describes the relationship in the context of one-parameter families. In doing so it also allows us better motivation for the peculiar-looking BC_a formula (2.3).

Suppose then that we have a one-parameter family of c.d.f.'s $G_\theta(\hat{\theta})$ on the real line, with $\hat{\theta}$ being an estimate of θ. In the relationships below we assume that $\hat{\theta}$ behaves asymptotically like a maximum likelihood estimator, with respect to a notional sample size n, as made explicit in (5.3) of Efron (1987). As a particular example, we will consider the case

$$\hat{\theta} \sim \theta\frac{\text{Gamma}_n}{n}, \quad n = 10, \tag{3.1}$$

where Gamma indicates a standard gamma variate with density $t^{n-1}\exp\{-t\}/\Gamma(n)$ for $t > 0$.

Having observed $\hat{\theta}$, we wonder with what confidence we can reject a trial value θ_0 of the parameter $\hat{\theta}$. In the gamma example (3.1) we might have

$$\hat{\theta} = 1 \quad \text{and} \quad \theta_0 = 1.5. \tag{3.2}$$

The easy answer from the bootstrap point of view is given in terms of the bootstrap c.d.f. $\hat{G}(c) = G_{\hat{\theta}}(c)$. We can define the bootstrap *confidence value* to be

$$\tilde{\alpha} = \hat{G}(\theta_0) = G_{\hat{\theta}}(\theta_0). \tag{3.3}$$

However, this will usually not agree with the more familiar hypothesis-testing *confidence level* for a one-parameter problem, say

$$\hat{\alpha} = 1 - G_{\theta_0}(\hat{\theta}), \tag{3.4}$$

the probability under θ_0 of getting a less extreme observation than $\hat{\theta}$. (For convenience these definitions assume $\hat{\theta} < \theta_0$.) In the case of (3.1)–(3.2) we have $\tilde{\alpha} = 0.930$ while $\hat{\alpha} = 0.863$.

The BC_a formula (2.3) amounts to a rule for converting bootstrap confidence values $\tilde{\alpha}$ into hypothesis-testing confidence levels $\hat{\alpha}$. This becomes crucial as soon as we try to use the bootstrap on problems more complicated than one-parameter families. Define

$$\tilde{z} = \Phi^{-1}(\tilde{\alpha}) \quad \text{and} \quad \hat{z} = \Phi^{-1}(\hat{\alpha}). \tag{3.5}$$

For a given value of θ_0 and $\hat{\alpha}$ above, let $\alpha = \hat{\alpha}$ and $\hat{\theta}_{BC_a}[\alpha] = \theta_0$ in (2.3). If (2.3) works perfectly, then we have

$$\Phi^{-1}\hat{G}(\theta_0) = \tilde{z} = z_0 + \frac{z_0 + \hat{z}}{1 - a(z_0 + \hat{z})}, \tag{3.6}$$

or

$$\hat{z} = \frac{\tilde{z} - z_0}{1 + a(\hat{z} - z_0)} - z_0. \tag{3.7}$$

The fact that the BC_a intervals are second-order accurate implies that the conversion formula (3.7) itself must be quite accurate.

To use (3.7), or (2.3), we first must estimate the two parameters z_0 and a. The bias-correction z_0 is estimated by

$$\hat{z}_0 = \Phi^{-1}\hat{G}(\hat{\theta}) = \Phi^{-1}G_{\hat{\theta}}(\hat{\theta}) \tag{3.8}$$

as in (2.8). The acceleration a is estimated in terms of the skewness of the score function

$$\dot{\ell}_\theta(\hat{\theta}) = \frac{\partial}{\partial\theta}\log\{g_\theta(\hat{\theta})\}, \tag{3.9}$$

where $g_\theta(\hat{\theta})$ is the density $\partial G_\theta(\hat{\theta})/\partial\hat{\theta}$. Section 10 of Efron (1987) shows that one-sixth the skewness of $\dot{\ell}_\theta(\hat{\theta})$ evaluated at $\theta = \hat{\theta}$,

$$\hat{a} = \mathrm{SKEW}_{\theta=\hat{\theta}}\{\dot{\ell}_\theta(\hat{\theta})\}/6, \tag{3.10}$$

is an excellent estimate of a.

Both z_0 and a are of order $O(1/\sqrt{n})$, with the estimates $\hat{z}_0$ and $\hat{a}$ erring by $O(1/n)$. For the gamma problem (3.1) it is easy to calculate that

$$\hat{z}_0 = 0.106 \quad \text{and} \quad \hat{a} = 0.105. \tag{3.11}$$

If $\hat{\theta}$ is the MLE in a one-parameter family (but not in general), then $\hat{z}_0$ and $\hat{a}$ are nearly the same, as is the case here.

The usable form of (3.7) is

$$\hat{z} = \frac{\tilde{z} - \hat{z}_0}{1 + \hat{a}(\tilde{z} - z_0)} - \hat{z}_0. \tag{3.12}$$

We can list three important properties of the $(\tilde{z}, \hat{z})$ curve (3.12) near $\tilde{z} = \hat{z}_0$:

$$(\tilde{z}, \hat{z}) = (\hat{z}_0 - \hat{z}_0) \quad \text{at} \quad \tilde{z} = \hat{z}_0; \tag{3.13}$$

$$\frac{d\hat{z}}{d\tilde{z}} = 1 \quad \text{at} \quad \tilde{z} = \hat{z}_0, \tag{3.14}$$

and

$$\frac{d^2\hat{z}}{d\tilde{z}^2} = -2\hat{a} \quad \text{at} \quad \tilde{z} = \hat{z}_0. \tag{3.15}$$

The last of these relationships is of special interest here. It says that *the curvature of the* $(\tilde{z}, \hat{z})$ *curve at* $\hat{z}_0$ *is directly proportional to the acceleration* $\hat{a}$.

In any given one-parameter problem we can find the actual $(\tilde{z}, \hat{z})$ curve, at least in theory. This is obtained by keeping $\hat{\theta}$ fixed and varying the trial point θ_0 in (3.3)–(3.5). Figure 3 shows the $(\tilde{z}, \hat{z})$ curve for the gamma problem, with $\hat{\theta}$ any fixed value, say $\hat{\theta} = 1$. In this case the BC_a approximation formula (3.12) matches the actual $(\tilde{z}, \hat{z})$ curve to three decimal places over most of the range of the graph. At $\hat{\theta} = 1$, $\theta_0 = 1.5$ for example, $\hat{z}$ equals 1.092 both actually and from (3.15).

The fact that the BC_a formula (2.3) is second-order accurate implies that the conversion formula (3.12) errs only by $O(1/n)$. This means that relationships (3.13)–(3.15) must have the same order of accuracy, even in quite general problems. In particular, the curvature of the actual $(\tilde{z}, \hat{z})$ plot, if it were possible to compute it, would nearly equal $-2\hat{a}$, with $\hat{a}$ given by the skewness definition (3.10).

None of this is special to one-parameter families except for the skewness definition (3.10), which does not allow for nuisance parameters. The next section

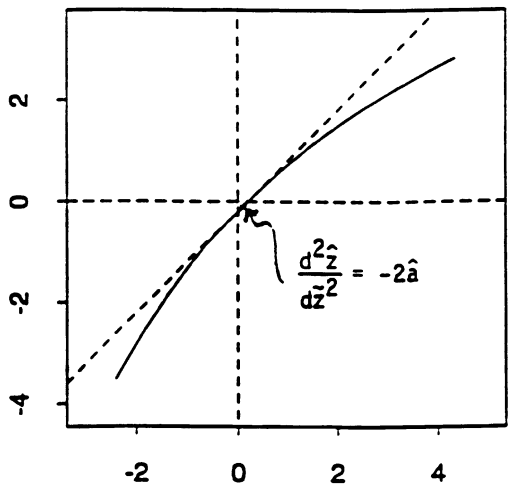

FIG. 3. *Plot of* $\hat{z}$ *versus* $\tilde{z}$ *in the gamma problem* (3.1)*; the* BC_a *approximation* (3.12) *or* (2.3)*, matches the actual curve to three decimal places. The central curvature of the* $(\tilde{z}, \hat{z})$ *plot is proportional to the acceleration* $\hat{a}$.

shows how to extend the skewness definition of $\hat{a}$ to multiparameter situations. This gives an estimate that is easy to evaluate, especially in exponential families, and that behaves well in practice. In fact a is usually easier to estimate than z_0, despite the latter's simpler definition.

4. THE ABC METHOD

We now leave one-parameter families and return to the more complicated situations that bootstrap methods are intended to deal with. In many such situations it is possible to approximate the BC_a interval endpoints analytically, entirely dispensing with Monte Carlo simulations. This reduces the computational burden by an enormous factor, and also makes it easier to understand how BC_a improves upon the standard intervals. The ABC method ("ABC" standing for approximate bootstrap confidence intervals) is an analytic version of BC_a applying to smoothly defined parameters in exponential families. It also applies to smoothly defined nonparametric problems, as shown in Section 6. DiCiccio and Efron (1992) introduced the ABC method, which is also discussed in Efron and Tibshirani (1993).

The BC_a endpoints (2.3) depend on the bootstrap c.d.f. $\hat{G}$ and estimates of the two parameters a and z_0. The ABC method requires one further estimate, of the *nonlinearity parameter* c_q, but it does not involve $\hat{G}$.

The standard interval (1.1) depends only on the two quantities $(\hat{\theta}, \hat{\sigma})$. The ABC intervals depend on the five quantities $(\hat{\theta}, \hat{\sigma}, \hat{a}, \hat{z}_0, \hat{c}_q)$. Each of the three extra numbers $(\hat{a}, \hat{z}_0, \hat{c}_q)$ corrects a deficiency of the standard method, making the ABC intervals second-order accurate as well as second-order correct.

The ABC system applies within multiparameter exponential families, which are briefly reviewed below. This framework includes most familiar parametric situations: normal, binomial, Poisson, gamma, multinomial, ANOVA, logistic regression, contingency tables, log-linear models, multivariate normal problems, Markov chains and also nonparametric situations as discussed in Section 6.

The density function for a p-parameter exponential family can be written as

$$g_\mu(\mathbf{x}) = \exp[\eta' y - \psi(\eta)] \tag{4.1}$$

where $\mathbf{x}$ is the observed data and $y = Y(\mathbf{x})$ is a p-dimensional vector of sufficient statistics; η is the p-dimensional natural parameter vector; μ is the expectation parameter $\mu = E_\mu\{y\}$; and $\psi(\eta)$, the cumulant generating function, is a normalizing factor that makes $g_\mu(\mathbf{x})$ integrate to 1.

The vectors μ and η are in one-to-one correspondence so that either can be used to index functions of interest. In (4.1), for example, we used μ to index the densities g, but η to index ψ. The ABC algorithm involves the mapping from η to μ, say

$$\mu = \mathrm{mu}(\eta), \tag{4.2}$$

which, fortunately, has a simple form in all of the common exponential families. Section 3 of DiCiccio and Efron (1992) gives function (4.2) for several families, as well as specifying the other inputs necessary for using the ABC algorithm.

The MLE of μ in (3.1) is $\hat{\mu} = y$, so that the MLE of a real-valued parameter of interest $\theta = t(\mu)$ is

$$\hat{\theta} = t(\hat{\mu}) = t(y). \tag{4.3}$$

As an example consider the bivariate normal model (1.2). Here $\mathbf{x} = ((B_1, A_1), (B_2, A_2), \ldots, (B_{20}, A_{20}))$ and $y = \sum_{i=1}^{20}(B_i, A_i, B_i^2, B_iA_i, A_i^2)'/20$. The bivariate normal is a five-parameter exponential family with

$$\mu = (\lambda_1, \lambda_2, \lambda_1^2 + \Gamma_{11}, \lambda_1\lambda_2 + \Gamma_{12}, \lambda_2^2 + \Gamma_{22})'. \tag{4.4}$$

Thus the correlation coefficient is the function $t(\mu)$ given by

$$\theta = \frac{\mu_4 - \mu_1\mu_2}{[(\mu_3 - \mu_1^2)(\mu_5 - \mu_2^2)]^{1/2}}; \tag{4.5}$$

$\hat{\theta} = t(\hat{\mu})$ is seen to be the usual sample correlation coefficient.

We denote the $p \times p$ covariance matrix of y by $\Sigma(\mu) = \mathrm{cov}_\mu\{y\}$, and let $\hat{\Sigma} = \Sigma(\hat{\mu})$, the MLE of Σ. The delta-method estimate of standard error for $\hat{\theta} = t(\hat{\mu})$ depends on $\hat{\Sigma}$. Let $\dot{t}$ denote the gradient vector of $\theta = t(\mu)$ at $\mu = \hat{\mu}$,

$$\dot{t} = \left(\ldots, \frac{\partial t}{\partial \mu_i}, \ldots\right)'_{\mu=\hat{\mu}}. \tag{4.6}$$

Then

$$\hat{\sigma} = (\dot{t}'\hat{\Sigma}\dot{t})^{1/2} \tag{4.7}$$

is the parametric delta-method estimate of standard error, and it is also the usual Fisher information standard error estimate.

The $\hat{\sigma}$ values for the standard intervals in Tables 2 and 3 were found by numerical differentiation, using

$$\left.\frac{\partial t}{\partial \mu_i}\right|_{\hat{\mu}} \doteq \frac{t(\hat{\mu} + \varepsilon e_i) - t(\hat{\mu} - \varepsilon e_i)}{2\varepsilon} \tag{4.8}$$

for a small value of ε, with e_i the ith coordinate vector. The covariance matrix $\hat{\Sigma}$ is simple to calculate in most of the familiar examples, as shown in

DiCiccio and Efron (1992, Section 3) giving $\hat{\sigma}$ from (4.7). This assumes that $t(\mu)$ is differentiable. In fact we need $t(\mu)$ to be twice differentiable in order to carry out the ABC computations.

The ABC algorithm begins by computing $\hat{\sigma}$ from (4.7)–(4.8). Then the parameters (a, z_0, c_q) are estimated by computing $p+2$ numerical second derivatives. The first of these is

$$\hat{a} = \frac{\partial^2}{\partial\varepsilon^2}[\dot{t}'\mathrm{mu}(\hat{\eta}+\varepsilon\dot{t})]_{\varepsilon=0}\Big/ 6\hat{\sigma}^3, \tag{4.9}$$

when $\hat{\eta}$ is the MLE of the natural parameter vector η. This turns out to be the same as the skewness definition of $\hat{a}$, (3.10), in the one-parameter family obtained from Stein's *least favorable family* construction [see Efron, 1987, (6.7)]. Formula (4.9) uses exponential family relationships to compute the skewness from a second derivative.

The second ABC numerical derivative is

$$\hat{c}_q = \frac{\partial^2}{\partial\varepsilon^2} t\left(\hat{\mu} + \frac{\varepsilon\hat{\Sigma}\dot{t}}{\hat{\sigma}}\right)\Big|_{\varepsilon=0}\Big/ 2\hat{\sigma}; \tag{4.10}$$

$\hat{c}_q$ measures how nonlinear the parameter of interest θ is, as a function of μ.

The final p second derivatives are required for the bias-correction parameter z_0. The parametric delta-method estimate of bias for $\hat{\theta} = t(\hat{\mu})$ can be expressed as

$$\hat{b} = \frac{1}{2}\sum_{i=1}^{p} \frac{\partial^2}{\partial\varepsilon^2} t(\hat{\mu} + \varepsilon d_i^{1/2}\gamma_i)\Big|_{\varepsilon=0}, \tag{4.11}$$

where d_i is the ith eigenvalue and γ_i is the ith eigenvector of $\hat{\Sigma}$. Then

$$\hat{z}_0 = \Phi^{-1}\big(2\cdot\Phi(\hat{a})\cdot\Phi(\hat{c}_q - \hat{b}/\hat{\sigma})\big) \doteq \hat{a} + \hat{c}_q - \hat{b}/\hat{\sigma}. \tag{4.12}$$

This involves terms other than $\hat{b}$ becuase z_0 relates to *median* bias. For the kind of smooth exponential family problems considered here, (4.12) is usually more accurate than the direct estimate (2.8).

The simplest form of the ABC intervals, called ABCquadratic or ABCq, gives the α-level endpoint directly as a function of the five numbers $(\hat{\theta}, \hat{\sigma}, \hat{a}, \hat{z}_0, \hat{c}_q)$:

$$\begin{aligned} \alpha &\rightarrow w \equiv \hat{z}_0 + z^{(\alpha)} \\ &\rightarrow \lambda \equiv \frac{w}{(1-\hat{a}w)^2} \rightarrow \xi \equiv \lambda + \hat{c}_q\lambda^2 \\ &\rightarrow \hat{\theta}_{\mathrm{ABCq}}[\alpha] = \hat{\theta} + \hat{\sigma}\xi. \end{aligned} \tag{4.13}$$

The *original ABC* endpoint, denoted $\hat{\theta}_{\mathrm{ABC}}[\alpha]$, requires one more recomputation of the function $t(\cdot)$:

$$\begin{aligned} \alpha &\rightarrow w = \hat{z}_0 + z^{(\alpha)} \rightarrow \lambda = \frac{w}{(1-\hat{a}w)^2} \\ &\rightarrow \hat{\theta}_{\mathrm{ABC}}[\alpha] = t\left(\hat{\mu} + \frac{\lambda\hat{\Sigma}\dot{t}}{\hat{\sigma}}\right). \end{aligned} \tag{4.14}$$

Notice that $\hat{c}_q$ is still required here, to estimate $\hat{z}_0$ in (4.12).

Formula (4.14) is the one used in Tables 2 and 3. It has the advantage of being transformation invariant, (2.11), and is sometimes more accurate than (4.13). However, (4.13) is *local*, all of the recomputations of $t(\mu)$ involved in (4.8)–(4.13) taking place infinitesimally near $\hat{\mu} = y$. In this sense ABCq is like the standard method. Nonlocality occasionally causes computational difficulties with boundary violations. In fact (4.13) is a simple quadratic approximation to (4.14), so ABC and ABCq usually agree reasonably well.

The main point of this article is that highly accurate approximate confidence intervals can now be calculated on a routine basis. The ABC intervals are implemented by a short computer algorithm. [The ABC intervals in Tables 2 and 3 were produced by the parametric and nonparametric ABC algorithms "abcpar" and "abcnon." These and the BC_a program are available in the language S: send electronic mail to *statlib@lib.stat.cmu.edu* with the one-line message: *send bootstrap.funs from S*.] There are five inputs to the algorithm: $\hat{\mu}$, $\hat{\Sigma}$, $\hat{\eta}$ and the functions $t(\cdot)$ and $\mathrm{mu}(\cdot)$. The outputs include $\hat{\theta}_{\mathrm{STAN}}[\alpha]$, $\hat{\theta}_{\mathrm{ABC}}[\alpha]$ and $\hat{\theta}_{\mathrm{ABCq}}[\alpha]$. Computational effort for the ABC intervals is two or three times that required for the standard intervals.

The ABC intervals can be useful even in very simple situations. Suppose that the data consists of a single observation x from a Poisson distribution with unknown expectation θ. In this case $\hat{\theta} = t(x) = x$ and $\hat{\sigma} = \sqrt{\hat{\theta}}$. Carrying through definitions (4.9)–(4.14) gives $\hat{a} = \hat{z}_0 = 1/(6\hat{\theta}^{1/2})$, $\hat{c}_q = 0$, and so

$$\hat{\theta}_{\mathrm{ABC}}[\alpha] = \hat{\theta} + \frac{w}{(1-\hat{a}w)^2}\sqrt{\hat{\theta}}, \quad w = \hat{z}_0 + z^{(\alpha)}.$$

For $x = 7$, the interval $(\hat{\theta}_{\mathrm{ABC}}[0.05], \hat{\theta}_{\mathrm{ABC}}[0.95])$ equals (3.54, 12.67). This compares with the exact interval (3.57, 12.58) for θ, splitting the atom of probability at $x = 7$, and with the standard interval (2.65, 11.35).

Here is a more realistic example of the ABC algorithm, used in a logistic regression context. Table 4 shows the data from an experiment concerning mammalian cell growth. The goal of this experiment was to quantify the effects of two factors on the success of a culture. Factor "r" measures the ratio of two key constituents of the culture plate, while factor "d" measures how many days were allowed for culture maturation. A total of 1,843 independent cultures were prepared, investigating 25 different (r_i, d_j) combinations. The table lists s_{ij} and n_{ij} for each combination, the num-

TABLE 4

Cell data: 1,843 *cell cultures were prepared, varying two factors, r (the ratio of two key constituents) and d (the number of days of culturing). Data shown are* s_{ij} *and* n_{ij}, *the number of successful cultures and the number of cultures attempted, at the ith level of r and the jth level of d*

	d_1	d_2	d_3	d_4	d_5	Total
r_1	5/31	3/28	20/45	24/47	29/35	81/186
r_2	15/77	36/78	43/71	56/71	66/74	216/371
r_3	48/126	68/116	145/171	98/119	114/129	473/661
r_4	29/92	35/52	57/85	38/50	72/77	231/356
r_5	11/53	20/52	20/48	40/55	52/61	143/269
Total	108/379	162/326	285/420	256/342	333/376	1144/1843

ber of successful cultures, compared to the number attempted.

We suppose that the number of successful cultures is a binomial variate,

$$s_{ij} \sim_{\text{i.i.d.}} \text{binomial}(n_{ij}, \pi_{ij}), \qquad i, j = 1, 2, 3, 4, 5, \tag{4.15}$$

with an additive logistic regression model for the unknown probabilities π_{ij},

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mu + \alpha_i + \beta_j, \qquad \sum_1^5 \alpha_i = \sum_1^5 \beta_j = 0. \tag{4.16}$$

For the example here we take the parameter of interest to be

$$\theta = \frac{\pi_{15}}{\pi_{51}}, \tag{4.17}$$

the success probability for the lowest r and highest d divided by the success probability for the highest r and lowest d. This typifies the kind of problem traditionally handled by the standard method.

A logistic regression program calculated maximum likelihood estimates $\hat\mu, \hat\alpha_i, \hat\beta_j$, from which we obtained

$$\hat\theta = \frac{1+\exp[-(\hat\mu+\hat\alpha_5+\hat\beta_1]}{1+\exp[-(\hat\mu+\hat\alpha_1+\hat\beta_5)]} = 4.16. \tag{4.18}$$

The output of the logistic regression program provided $\hat\mu$, $\hat\Sigma$ and $\hat\eta$ for the ABC algorithm. Section 3 of DiCiccio and Efron (1992) gives the exact specification for an ABC analysis of a logistic regression problem. Applied here, the algorithm gave standard and ABC 0.90 central intervals for θ,

$$(\hat\theta_{\text{STAN}}[0.05], \hat\theta_{\text{STAN}}[0.95]) = (3.06, 5.26), \qquad (\hat\theta_{\text{ABC}}[0.05], \hat\theta_{\text{ABC}}[0.95]) = (3.20, 5.43). \tag{4.19}$$

The ABC limits are shifted moderately upwards relative to the standard limits, enough to make the shape (1.6) equal 1.32. The standard intervals are not too bad in this case, although better performance might have been expected with $n = 1,843$ data points. In fact it is very difficult to guess a priori what constitutes a large enough sample size for adequate standard-interval performance.

The ABC formulas (4.13)–(4.14) were derived as second-order approximations to the BC_a endpoints by DiCiccio and Efron (1992). They showed that these formulas give second-order accuracy as in (2.10), and also second-order correctness. Section 8 reviews some of these results. There are many other expressions for ABC-like interval endpoints that enjoy equivalent second-order properties in theory, although they may be less dependable in practice. A particularly simple formula is

$$\hat\theta_{\text{ABC}}[\alpha] \doteq \hat\theta_{\text{STAN}}[\alpha] + \hat\sigma\{\hat z_0 + (2\hat a + \hat c_q)z^{(\alpha)^2}\}. \tag{4.20}$$

This shows that the ABC endpoints are not just a translation of $\hat\theta_{\text{STAN}}[\alpha]$.

In repeated sampling situations the estimated constants $(\hat a, \hat z_0, \hat c_q)$ are of stochastic order $1/\sqrt{n}$ in the sample size, the same as $\hat\sigma$. They multiply $\hat\sigma$ in (4.20), resulting in corrections of order $\hat\sigma/\sqrt{n}$ to $\hat\theta_{\text{STAN}}[\alpha]$. If there were only 1/4 as much cell data, $n = 461$, but with the same proportion of successes in every cell of Table 4, then $(\hat a, \hat z_0, \hat c_q)$ would be twice as large. This would double the relative difference $(\hat\theta_{\text{ABC}}[\alpha] - \hat\theta_{\text{STAN}}[\alpha])/\hat\sigma$ according to (4.20), rendering $\hat\theta_{\text{STAN}}[\alpha]$ quite inaccurate.

Both $\hat a$ and $\hat z_0$ are transformation invariant, retaining the same numerical value under monotone parameter transformations $\phi = m(\theta)$. The nonlinearity constant $\hat c_q$ is not invariant, and it can be reduced by transformations that make ϕ more linear as a function of μ. Changing parameters from $\theta = \pi_{15}/\pi_{51}$ to $\phi = \log(\theta)$ changes $(\hat a, \hat z_0, \hat c_q)$ from $(-0.006, -0.025, 0.105)$ to $(-0.006, -0.025, 0.025)$ for the cell data. The standard intervals are nearly correct on the ϕ scale. The ABC and BC_a methods automate this kind of data-analytic trick.

We can visualize the relationship between the BC_a and ABC intervals in terms of Figure 3. The

BC_a method uses Monte Carlo bootstrapping to find $\tilde{z}$, as in (3.3) and (3.5), and then maps $\tilde{z}$ into an appropriate hypothesis-testing value $\hat{z}$ via formula (3.7). The ABC method also uses formula (3.7) [or, equivalently, (2.3)], but in order to avoid Monte Carlo computations it makes one further analytic approximation: $\tilde{z}$ itself, the point on the horizontal axis in Figure 3, is estimated from an Edgeworth expansion. The information needed for the Edgeworth expansion is obtained from the second derivatives (4.9)–(4.11).

5. BOOTSTRAP-*t* INTERVALS

The BC_a formula strikes some people as complicated, and also "unbootstraplike" since the estimate $\hat{a}$ is not obtained directly from bootstrap replications. The bootstrap-t method, another bootstrap algorithm for setting confidence intervals, is conceptually simpler than BC_a. The method was suggested in Efron (1979), but some poor numerical results reduced its appeal. Hall's (1988) paper showing the bootstrap-t's good second-order properties has revived interest in its use. Babu and Singh (1983) gave the first proof of second-order accuracy for the bootstrap-t.

Suppose that a data set $\mathbf{x}$ gives an estimate $\hat{\theta}(\mathbf{x})$ for a parameter of interest θ, and also an estimate $\hat{\sigma}(\mathbf{x})$ for the standard deviation of $\hat{\theta}$. By analogy with Student's t-statistic, we define

$$T = \frac{\hat{\theta} - \theta}{\hat{\sigma}} \tag{5.1}$$

and let $T^{(\alpha)}$ indicate the 100αth percentile of T. The upper endpoint of an α-level one-sided confidence inteval for θ is

$$\hat{\theta} - \hat{\sigma}T^{(1-\alpha)}. \tag{5.2}$$

This assumes we know the T-percentiles, as in the usual Student's-t case where $T^{(\alpha)}$ is the percentile of a t-distribution. However, the T-percentiles are unknown in most situations.

The idea of the bootstrap-t is to estimate the percentiles of T by bootstrapping. First, the distribution governing $\mathbf{x}$ is estimated and the bootstrap data sets $\mathbf{x}^*$ are drawn from the estimated distribution, as in (2.1). Each $\mathbf{x}^*$ gives both a $\hat{\theta}^*$ and a $\hat{\sigma}^*$, yielding

$$T^* = \frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*}, \tag{5.3}$$

a bootstrap replication of (5.1). A large number B of independent replications gives estimated percentiles

$$\hat{T}^{(\alpha)} = B \cdot \alpha\text{th ordered value of } \{T^*(b), b = 1, 2, \ldots, B\}. \tag{5.4}$$

[So if $B = 2{,}000$ and $\alpha = 0.95$, then $\hat{T}^{(\alpha)}$ is the 1,900th ordered $T^*(b)$.] The 100αth bootstrap-t confidence endpoint $\hat{\theta}_T[\alpha]$ is defined to be

$$\hat{\theta}_T[\alpha] = \hat{\theta} - \hat{\sigma}\hat{T}^{(1-\alpha)}, \tag{5.5}$$

following (5.2).

Figure 4 relates to the correlation coefficient for the cd4 data. The left panel shows 2,000 normal-theory bootstrap replications of

$$T = \frac{\hat{\theta} - \theta}{\hat{\sigma}}, \quad \hat{\sigma} = \frac{1 - \hat{\theta}^2}{\sqrt{20}}. \tag{5.6}$$

Each replication required drawing $((B_1^*, A_1^*), \ldots, (B_{20}^*, A_{20}^*))$ as in (2.1), computing $\hat{\theta}^*$ and $\hat{\sigma}^*$, and then calculating the bootstrap$-t$ replication $T^* = (\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$. The percentiles $(\hat{T}^{(0.05)}, \hat{T}^{(0.95)})$ equalled $(-1.38, 2.62)$, giving a 0.90 central bootstrap-t interval of (0.45, 0.87). This compares nicely with the exact interval (0.47, 0.86) in Table 2.

Hall (1988) showed that the bootstrap-t limits are second-order accurate, as in (2.10). DiCiccio and Efron (1992) showed that they are also second-order correct (see Section 8).

Definition (2.17) uses the fact that $(1 - \hat{\theta}^2)/\sqrt{n}$ is a reasonable normal-theory estimate of standard error for $\hat{\theta}$. In most situations $\hat{\sigma}^*$ must be numerically computed for each bootstrap data set $\mathbf{x}^*$, perhaps using the delta method. This multiplies the bootstrap computations by a factor of at least $p + 1$, where p is the number of parameters in the probability model for $\mathbf{x}$. The nonparametric bootstrap-t distribution on the right side of Figure 4 used $\hat{\sigma}^*$ equal to the nonparametric delta-method estimate. The main disadvantage of both BC_a and bootstrap-t is the large computational burden. This does not make much difference for the correlation coefficient, but it can become crucial for more complicated situations. The ABC method is particularly useful in complicated problems.

More serious, the bootstrap-t algorithm can be numerically unstable, resulting in very long confidence intervals. This is a particular danger in nonparametric situations. As a rough rule of thumb, the BC_a intervals are more conservative than bootstrap-t, tending to stay, if anything, too close to the standard intervals as opposed to deviating too much.

Bootstrap-t intervals are not transformation invariant. The method seems to work better if θ is a translation parameter, such as a median or an expectation. A successful application of the type appears in Efron (1981, Section 9). Tibshirani (1988) proposed an algorithm for transforming θ to a more translation-like parameter $\phi = m(\theta)$, before applying the bootstrap-t method. Then the resulting interval is transformed back to the θ scale via $\theta =$

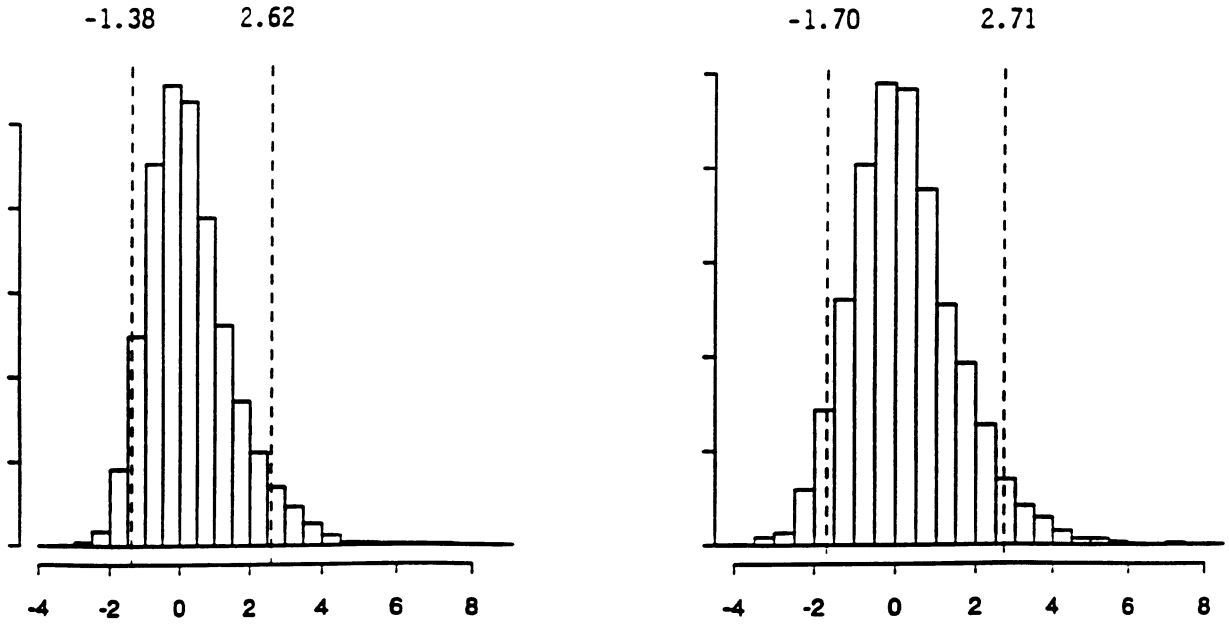

FIG. 4. *Bootstrap-t distributions relating to* θ *the cd*4 *data correlation*: (*left*) 2,000 *normal-theory bootstrap relications of* T *using* $\hat{\sigma}^* = (1 - \hat{\theta}^*)^2/\sqrt{20}$; (*right*) 2,000 *nonparametric bootstrap replications of* T *using* $\hat{\sigma}^*$ *given by the nonparametric delta method*; *dashed lines show* 5*th and* 95*th percentiles.*

$m^{-1}(\phi)$. See DiCiccio and Romano (1995, Section 2.b) or Efron and Tibshirani (1993, Section 12.6).

The bootstrap-t and BC_a methods look completely different. However, surprisingly, the ABC method connects them.

The ABC method was introduced as a non–Monte Carlo approximation to BC_a, but it can also be thought of as an approximation to the bootstrap-t method The relationships in (4.13) can be reversed to give the attained significance level (ASL) α for any observed data set. That is, we can find α such that $\hat{\theta}_{\mathrm{ABCq}}[\alpha]$ equals an hypothesized value θ for the parameter of interest:

$$\begin{aligned} \theta \;&\to\; \xi \;=\; \frac{\theta - \hat{\theta}}{\hat{\sigma}} \\ &\to\; \lambda = \frac{2\xi}{1 + (1 + 4\hat{c}_q\xi)^{1/2}} \\ &\to\; w = \frac{2\lambda}{(1 + 2\hat{a}\lambda) + (1 + 4\hat{a}\lambda)^{1/2}} \\ &\to\; \alpha = \Phi(w - \hat{z}_0). \end{aligned} \tag{5.7}$$

If the ABCq method works perfectly, then the ASL as defined by (5.7) will be uniformly distributed over [0, 1], so

$$Z = \Phi^{-1}(\alpha) \tag{5.8}$$

will be distributed as a $N(0, 1)$ variate.

Notice that T in (5.1) equals $-\xi$ in (5.7). The ABCq method amounts to assuming that

$$h_{\hat{a},\,\hat{z}_0,\,\hat{c}_q}(T) \sim N(0, 1) \tag{5.9}$$

for the transformation defined by (5.7)–(5.8). In other words, ABCq uses an estimated transformation of T to get a pivotal quantity. The bootstrap-t method assumes that T itself is pivotal, but then finds the pivotal distribution by bootstrapping. The calibration method discussed in Section 7 uses both an estimated transformation and bootstrapping, with the result being still more accurate intervals.

6. NONPARAMETRIC CONFIDENCE INTERVALS

The BC_a, bootstrap-t, and ABC methods can be applied to the construction of *nonparametric* confidence intervals. Here we will discuss the one-sample nonparametric situation where the observed data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ are a random sample from an arbitrary probability distribution F,

$$x_1, x_2, \ldots, x_n \sim_{\text{i.i.d.}} F. \tag{6.1}$$

The sample space $\mathscr{X}$ of the distribution can be anything at all; $\mathscr{X}$ is the two-dimensional Euclidean space $\mathbb{R}^2$ in (1.7) and on the right side of Table 1, and is an extended version of $\mathbb{R}^5$ in the missing-value example below. Multisample nonparametric problems are mentioned briefly at the end of this section.

The *empirical distribution* $\hat{F}$ puts probability $1/n$ on each sample point x_i in $\mathbf{x}$. A real-valued param-

eter of interest $\theta = t(F)$ has the nonparametric estimate

$$\hat{\theta} = t(\hat{F}), \tag{6.2}$$

also called the nonparametric maximum likelihood estimate. A nonparametric bootstrap sample $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ is a random sample of size n drawn from $\hat{F}$,

$$x_1^*, x_2^*, \ldots, x_n^* \sim \hat{F}. \tag{6.3}$$

In other words, $\mathbf{x}^*$ equals $(x_{j_1}, x_{j_2}, \ldots, x_{j_n})$ where $j_1, j_2, \ldots, j_n$ is a random sample drawn with replacement from $\{1, 2, \ldots, n\}$. Each bootstrap sample gives a nonparametric bootstrap replication of $\hat{\theta}$,

$$\hat{\theta}^* = t(\hat{F}^*), \tag{6.4}$$

where $\hat{F}^*$ is the empirical distribution of $\mathbf{x}^*$.

Nonparametric BC_a confidence intervals for θ are constructed the same way as the parametric intervals of Section 2. A large number of independent bootstrap replications $\hat{\theta}^*(1)$, $\hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)$ are drawn according to (4.3)–(4.4), $B \approx 2{,}000$, giving a bootstrap cumulative distribution function $\hat{G}(c) = \#\{\hat{\theta}^*(b) < c\}/B$. The BC_a endpoints $\hat{\theta}_{BC_a}[\alpha]$ are then calculated from formula (2.3), plugging in nonparametric estimates of z_0 and a.

Formula (2.8) gives $\hat{z}_0$, which can also be obtained from a nonparametric version of (4.12). The acceleration a is estimated using the *empirical influence function* of the statistic $\hat{\theta} = t(\hat{F})$,

$$U_i = \lim_{\varepsilon \to 0} \frac{t((1-\varepsilon)\hat{F} + \varepsilon\delta_i)}{\varepsilon}, \quad i = 1, 2, \ldots, n. \tag{6.5}$$

Here δ_i is a point mass on x_i, so $(1-\varepsilon)\hat{F} + \varepsilon\delta_i$ is a version of $\hat{F}$ putting extra weight on x_i and less weight on the other points. The usual nonparametric delta-method estimate of standard error is $[\Sigma U_i^2/n^2]^{1/2}$, this being the value used in our examples of the standard interval (1.1).

The estimate of a is

$$\hat{a} = \frac{1}{6} \frac{\sum_{i=1}^n U_i^3}{(\sum_{i=1}^n U_i^2)^{3/2}}. \tag{6.6}$$

This looks completely different than (4.9), but in fact it is the same formula, applied here in a multinomial framework appropriate to the nonparametric situation. The similarity of (6.6) to a skewness reflects the relationship of $\hat{a}$ to the skewness of the score function, (3.10). The connection of nonparametric confidence intervals with multinomial estimation problems appears in Efron (1987, Sections 7 and 8).

There is a simpler way to calculate the U_i and $\hat{a}$. Instead of (6.5) we can use the *jackknife influence function*

$$U_i = (n-1)(\hat{\theta}_{\cdot} - \hat{\theta}_{(i)}) \tag{6.7}$$

in (6.6), where $\hat{\theta}_{(i)}$ is the estimate of θ based on the reduced data set $\mathbf{x}_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. This makes it a little easier to calculate the BC_a limits since the statistic $\hat{\theta}(\mathbf{x})$ does not have to be reprogrammed in the functional form $\hat{\theta} = t(\hat{F})$.

The nonparametric BC_a method is unfazed by complicated sample spaces. Table 5 shows an artificial missing-data example discussed in Efron (1994). Twenty-two students have each taken five exams labelled A, B, C, D, E, but some of the A and E scores (marked "?") are missing. If there were no missing data, we would consider the rows of the matrix to be a random sample of size $n = 22$ from an unknown five-dimensional distribution F. Our goal is to estimate

$$\theta = \text{maximum eigenvalue of } \Sigma, \tag{6.8}$$

where Σ is the covariance matrix of F.

An easy way, though not necessarily the best way, to fill in Table 5 is to fit a standard two-way additive model $\nu + \alpha_i + \beta_j$ to the non-missing scores by least squares, and then to replace the missing values

TABLE 5

Twenty-two students have each taken five exams, labelled A, B, C, D, E. Some of the scores for A and E (indicated by "?") are missing. Original data set from Kent, Mardia and Bibby (1979)

Student	A	B	C	D	E
1	?	63	65	70	63
2	53	61	72	64	73
3	51	67	65	65	?
4	?	69	53	53	53
5	?	69	61	55	45
6	?	49	62	63	62
7	44	61	52	62	?
8	49	41	61	49	?
9	30	69	50	52	45
10	?	59	51	45	51
11	?	40	56	54	?
12	42	60	54	49	?
13	?	63	53	54	?
14	?	55	59	53	?
15	?	49	45	48	?
16	17	53	57	43	51
17	39	46	46	32	?
18	48	38	41	44	33
19	46	40	47	29	?
20	30	34	43	46	18
21	?	30	32	35	21
22	?	26	15	20	?

x_{ij} by

$$\hat{x}_{ij} = \hat{\nu} + \hat{\alpha}_i + \hat{\beta}_j. \tag{6.9}$$

The filled-in 22×5 data matrix has rows $\hat{x}_i$, $i = 1, 2, \ldots, 22$, from which we can calculate an empirical covariance matrix

$$\hat{\Sigma} = \tfrac{1}{22}\sum_{i=1}^{22}(\hat{x}_i - \hat{\mu}_i)(\hat{x}_i - \hat{\mu}_i)', \quad \hat{\mu} = \tfrac{1}{22}\sum_{1}^{22}\hat{x}_i, \tag{6.10}$$

giving the point estimate

$$\hat{\theta} = \text{maximum eigenvalue of } \hat{\Sigma} = 633.2. \tag{6.11}$$

How accurate is $\hat{\theta}$?

It is easy to carry out a nonparametric BC_a analysis. The "points" x_i in the data set $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $n = 22$, are the rows of Table 5, for instance $x_{22} = (?, 26, 15, 20, ?)$. A bootstrap data set $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$ is a 22×5 data matrix, each row of which has been randomly selected from the rows of Table 5. Having selected $\mathbf{x}^*$, the bootstrap replication $\hat{\theta}^*$ is computed by following the same steps (4.9)–(4.11) that gave $\hat{\theta}$. Figure 5 is a histogram of 2,200 bootstrap replications $\hat{\theta}^*$, the histogram being noticeably long-tailed toward the right. The 0.90 BC_a confidence interval for θ is

$$(\hat{\theta}_{BC_a}[0.05],\ \hat{\theta}_{BC_a}[.095]) = (379,1,164), \tag{6.12}$$

extending twice as far to the right of $\hat{\theta}$ as to the left.

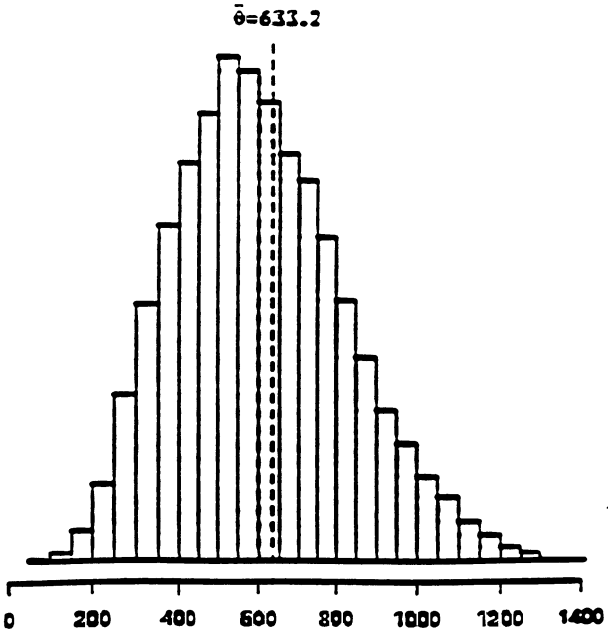

FIG. 5. *Histogram of* 2,200 *nonparametric bootstrap replications of the maximum eigenvalue statistic for the student score data; bootstrap standard error estimate* $\hat{\sigma} = 212.0$. *The histogram is long-tailed to the right, and so is the* BC_a *confidence interval* (6.12).

It is easy to extend the ABC method of Section 4 to nonparametric problems, greatly reducing the computational burden of the BC_a intervals. The formulas are basically the same as in (4.9)–(4.14), but they simplify somewhat in the nonparametric–multinomial framework. The statistic is expressed in the functional form $\hat{\theta} = t(\hat{F})$ and then reevaluated for values of F very near $\hat{F}$, as in (6.5). The ABC limits require only $2n + 4$ reevaluations of the statistic. By comparison, the BC_a method requires some 2,000 evaluations $\hat{\theta}^* = t(\hat{F}^*)$, where $\hat{F}^*$ is a bootstrap empirical distribution.

The nonparametric ABC algorithm "abcnon" was applied to the maximum eigenvalue statistic for the student score data. After 46 reevaluations of the statistic defined by (6.9)–(6.11), it gave 0.90 central confidence interval

$$(\hat{\theta}_{\mathrm{ABC}}[0.05],\ \hat{\theta}_{\mathrm{ABC}}[0.95]) = (379,1,172), \tag{6.13}$$

nearly the same as (6.12). The Statlib program abcnon used here appears in the appendix to Efron (1994); Efron (1994) also applied abcnon to the full normal theory MLE of θ, (6.8), rather than to the ad hoc estimator (6.9)–(6.11). The resulting ABC interval (353, 1307) was 20% longer than (6.13), perhaps undermining belief in the data's normality.

So far we have only discussed one-sample nonparametric problems. The K-sample nonparametric problem has data

$$x_{k1}, x_{k2}, \ldots, x_{kn_k} \sim_{\text{i.i.d.}} F_k \quad \text{for} \quad k = 1, 2, \ldots, K, \tag{6.14}$$

for arbitrary probability distributions F_k on possibly different sample spaces $\mathcal{X}_k$. The nonparametric MLE of a real-valued parameter of interest $\theta = t(F_1, F_2, \ldots, F_K)$ is

$$\hat{\theta} = t(\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_K), \tag{6.15}$$

where $\hat{F}_k$ is the empirical distribution corresponding to $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kn_k})$.

It turns out that K-sample nonparametric confidence intervals can easily be obtained from either abcnon or bcanon, its nonparametric BC_a counterpart. How to do so is explained in Remarks C and H of Efron (1994).

7. CALIBRATION

Calibration is a bootstrap technique for improving the coverage accuracy of any system of approximate confidence intervals. Here we will apply it to the nonparametric ABC intervals in Tables 2 and 3. The general theory is reviewed in Efron and Tibshirani (1993, Sections 18.3 and 25.6), following ideas of

Loh (1987), Beran (1987), Hall (1986) and Hall and Martin (1988).

Let $\hat{\theta}\ [\alpha]$ be the upper endpoint of a one-sided level-α approximate confidence interval for parameter θ. If the approximation is actually working perfectly then the true probability of coverage

$$\beta(\alpha) \equiv \text{Prob}\{\theta < \hat{\theta}\ [\alpha]\} \tag{7.1}$$

will equal α. If not, we could use the *calibration curve* $\beta(\alpha)$ to improve the approximate confidence intervals. For example, if $\beta[0.03] = 0.05$ and $\beta[0.98] = 0.95$, then we could use $(\hat{\theta}[0.03], \hat{\theta}[0.98])$ instead of $(\hat{\theta}[0.05], \hat{\theta}[0.95])$ as our approximate central 0.90 interval.

Of course we do not know the calibration curve $\beta(\alpha)$. The interesting fact is that we can apply the bootstrap to estimate $\beta(\alpha)$, and then use the estimate to improve our original approximate intervals. The estimated calibration curve is

$$\hat{\beta}(\alpha) = \text{Prob}_*\{\hat{\theta} < \hat{\theta}\ [\alpha]^*\}. \tag{7.2}$$

Prob_* indicates bootstrap sampling as in (2.1) or (6.3) (so $\hat{\theta}$ is fixed), where $\hat{\theta}\ [\alpha]^*$ is the upper α endpoint of an interval based on the bootstrap data.

It looks like we have to do separate bootstrap calculations in (7.2) for every value of α, but that is unnecessary if $\hat{\theta}\ [\alpha]$ is an increasing function of α, as it usually is. For a given bootstrap sample, let $\hat{\alpha}^*$ be the value of α that makes the upper endpoint equal $\hat{\theta}$,

$$\hat{\alpha}^*\colon \hat{\theta}[\hat{\alpha}^*] = \hat{\theta}. \tag{7.3}$$

Then the event $\{\hat{\alpha}^* < \alpha\}$ is equivalent to the event $\{\hat{\theta} < \hat{\theta}\ [\alpha]^*\}$, so

$$\hat{\beta}(\alpha) = \text{Prob}_*\{\hat{\alpha}^* < \alpha\}. \tag{7.4}$$

In order to calibrate a system of approximate confidence intervals we generate B bootstrap samples, and for each one we calculate $\hat{\alpha}^*$. The estimated calibration curve is

$$\hat{\beta}(\alpha) = \#\{\hat{\alpha}^*(b) < \alpha\}/B. \tag{7.5}$$

In other words, we estimate the c.d.f. of $\hat{\alpha}^*$. If the c.d.f. is nearly uniform, $\hat{\beta}(\alpha) \doteq \alpha$, then this indicates accurate coverage for our system of intervals. If not, we can use $\hat{\beta}(\alpha)$ to improve the original endpoints by calibration.

This idea was applied to the nonparametric ABC intervals of Tables 2 and 3, the correlation coefficient and maximum eigenvalue statistic for the cd4 data. Figure 6 shows the result of $B = 2{,}000$ bootstrap replications for each situation. The calibration shows good results for the correlation coefficient, with $\hat{\beta}(\alpha) \doteq \alpha$ over the full range of α. The story is less pleasant for the maximum eigenvalue. At the upper end of the scale we have $\hat{\beta}(\alpha) < \alpha$, indicating that we need to take $\alpha > 0.95$ to get actual 95% coverage. According to Table 6, which shows the percentiles of the $\hat{\alpha}^*$ distributions, we should take $\alpha = 0.994$. This kind of extreme correction is worrisome, but it produces an interesting result in Table 3: it moves the upper endpoint of the nonparametric interval much closer to the normal-theory value 3.25.

Calibrating the ABC intervals improves their accuracy from second to third order, with coverage errors, as in (2.10), reduced to $O(1/n^{3/2})$. We are talking about a lot of computation here, on the order of 1,000 times as much as for the ABC intervals themselves. The computational efficiency of ABC compared to BC_a becomes crucial in the calibration context. Calibrating the BC_a intervals would require on the order of 1,000,000 recomputations of the original statistic $\hat{\theta}$.

8. SECOND-ORDER ACCURACY AND CORRECTNESS

This section derives the second-order properties of the various bootstrap intervals. In order to validate the second-order accuracy and correctness of bootstrap confidence intervals we need asymptotic expansions for the cumulative distribution functions of $\hat{\theta}$ and $T = (\hat{\theta} - \theta)/\hat{\sigma}$. Later these expressions will be used to connect bootstrap theory to several other second-order confidence interval methods. In many situations, including those considered in the preceding sections, the asymptotic distribution of $U = (\hat{\theta} - \theta)/\sigma$ is standard normal, and the first three cumulants of U are given by

$$E(U) = \frac{k_1}{\sqrt{n}}, \quad \text{var}(U) = 1, \quad \text{skew}(U) = \frac{k_3}{\sqrt{n}},$$

where k_1 and k_3 are of order $O(1)$; the fourth- and higher-order cumulants are of order $O(n^{-1})$ or smaller. It follows that the first three cumulants of

$$T = \frac{(\hat{\theta} - \theta)}{\hat{\sigma}} = U\left\{1 - \frac{1}{2}\frac{(\hat{\sigma}^2 - \sigma^2)}{\sigma^2}\right\} + O_p(n^{-1})$$

are given by

$$E(T) = \frac{k_1 - \frac{1}{2}k_2}{\sqrt{n}} + O(n^{-1}),$$

$$\text{var}(T) = 1 + O(n^{-1}),$$

$$\text{skew}(T) = \frac{-(3k_2 - k_3)}{\sqrt{n}} + O(n^{-1}),$$

where

$$\frac{k_2}{\sqrt{n}} = E\left\{\frac{(\hat{\sigma}^2 - \sigma^2)(\hat{\theta} - \theta)}{\sigma^3}\right\}.$$

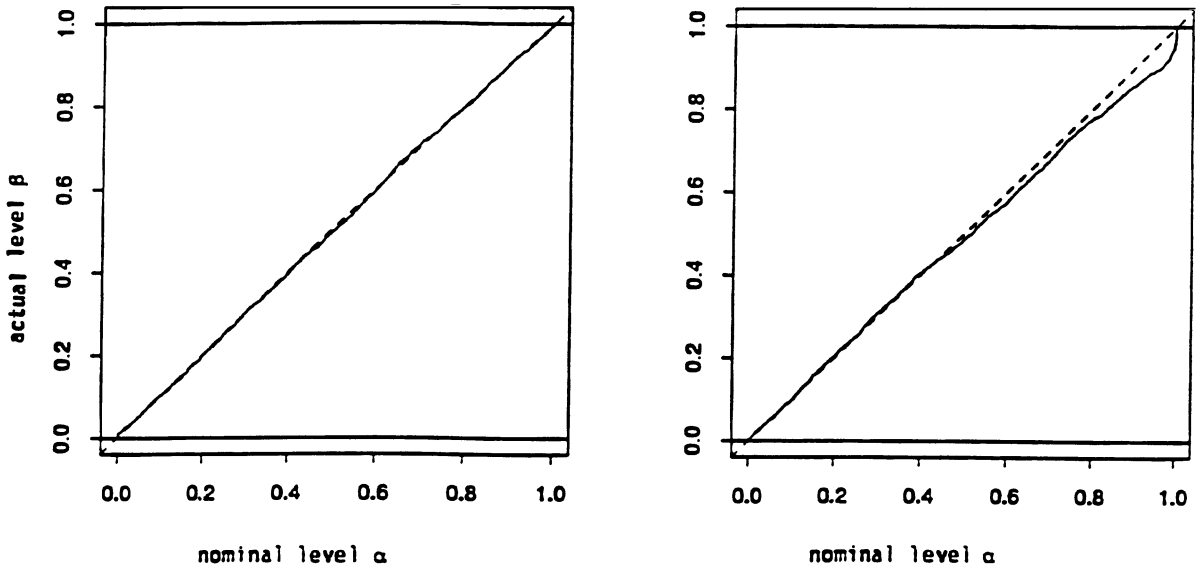

FIG. 6. *Estimated calibration curves for the nonparametric ABC method, cd4 data: (left panel) correlation coefficient as in Table* 2; *(right panel) maximum eigenvalue as in Table* 3; *each based on* 2,000 *bootstrap replications.*

TABLE 6
Percentiles of the distributions of $\hat{\alpha}^*$ *shown in Figure* 6; *the* 0.05 *and* 0.95 *values were used for the calibrated ABC endpoints in Tables* 2 *and* 3

Actual alpha	0.025	0.05	0.1	0.16	0.84	0.9	0.95	0.975
Nominal, corr	0.0196	0.0482	0.0984	0.164	0.843	0.898	0.953	0.980
Nominal, maxeig	0.0243	0.0515	0.1051	0.156	0.879	0.964	0.994	0.999

Observe that k_2 is of order $O(1)$, since σ^2 is of order $O(n^{-1})$ and $\hat{\sigma}^2$ generally differs from σ^2 by order $O_p(n^{-3/2})$. The fourth- and higher-order cumulants of T are of order $O(n^{-1})$ or smaller. Thus, when $\hat{\theta}$ is continuous, the cumulative distribution functions $H(u)$ and $K(t)$ of U and T typically have Cornish–Fisher expansions

$$\begin{aligned} H(u) &= \mathrm{pr}\{(\hat{\theta}-\theta)/\sigma \le u\} \\ &= \Phi\big[u - n^{-1/2}\{(k_1 - \tfrac{1}{6}k_3) + \tfrac{1}{6}k_3u^2\}\big] \\ &\quad + O(n^{-1}), \end{aligned} \tag{8.1}$$

$$\begin{aligned} K(t) &= \mathrm{pr}\{(\hat{\theta}-\theta)/\hat{\sigma} \le t\} \\ &= \Phi\big[t - n^{-1/2}\{(k_1 - \tfrac{1}{6}k_3) - (\tfrac{1}{2}k_2 - \tfrac{1}{6}k_3)t^2\}\big] \\ &\quad + O(n^{-1}). \end{aligned} \tag{8.2}$$

Furthermore, the inverse cumulative distribution functions $H^{-1}(\alpha)$ and $K^{-1}(\alpha)$ have expansions

$$\begin{aligned} H^{-1}(\alpha) &= z^{(\alpha)} + n^{-1/2}\big[(k_1 - \tfrac{1}{6}k_3) + \tfrac{1}{6}k_3\{z^{(\alpha)}\}^2\big] \\ &\quad + O(n^{-1}), \end{aligned} \tag{8.3}$$

$$\begin{aligned} K^{-1}(\alpha) &= z^{(\alpha)} + n^{-1/2} \\ &\quad \cdot \big[(k_1 - \tfrac{1}{6}k_3) - (\tfrac{1}{2}k_2 - \tfrac{1}{6}k_3)\{z^{(\alpha)}\}^2\big] \\ &\quad + O(n^{-1}). \end{aligned} \tag{8.4}$$

To compare approximate confidence limits, Hall (1988) defined an "exact" upper α confidence limit for θ as $\hat{\theta}_{\mathrm{exact}}[\alpha] = \hat{\theta} - \hat{\sigma}K^{-1}(1-\alpha)$. This limit is exact in the sense of coverage; note that $\mathrm{pr}\{K^{-1}(1-\alpha) \le (\hat{\theta}-\theta)/\hat{\sigma}\} = \alpha$ implies $\mathrm{pr}\{\theta \le \hat{\theta}_{\mathrm{exact}}[\alpha]\} = 1-\alpha$. It requires the cumulative distribution function K, which is rarely known in practice; however, although usually unavailable, $\hat{\theta}_{\mathrm{exact}}[\alpha]$ does provide a useful benchmark for making comparisons. By using (8.4), the exact limit is seen to satisfy

$$\begin{aligned} \hat{\theta}_{\mathrm{exact}}[\alpha] &= \hat{\theta} + \hat{\sigma}z^{(\alpha)} - n^{-1/2}\hat{\sigma} \\ &\quad \cdot \big[(k_1 - \tfrac{1}{6}k_3) - (\tfrac{1}{2}k_2 - \tfrac{1}{6}k_3)\{z^{(\alpha)}\}^2\big] \\ &\quad + O_p(n^{-3/2}). \end{aligned} \tag{8.5}$$

An approximate α confidence limit $\hat{\theta}[\alpha]$ is said to be second-order correct if it differs from $\hat{\theta}_{\mathrm{exact}}[\alpha]$ by order $O_p(n^{-3/2})$. It is easily seen from (8.2) that a *second-order correct* limit $\hat{\theta}[\alpha]$ is also *second-order accurate*, that is, $\mathrm{pr}\{\theta \le \hat{\theta}[\alpha]\} = \alpha + O(n^{-1})$.

Let $\hat{K}(t)$ be the bootstrap cumulative distribution function of T, so that $\hat{K}(t)$ is the cumulative distribution function of $T^* = (\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$. The first three cumulants of T^* typically differ from those of T by order $O_p(n^{-1})$, and $\hat{K}(t)$ has the expansion

$$\hat{K}(t) = \Phi\big[t - n^{-1/2}\big\{(\hat{k}_1 - \tfrac{1}{6}\hat{k}_3) - (\tfrac{1}{2}\hat{k}_2 - \tfrac{1}{6}\hat{k}_3)t^2\big\}\big] + O_p(n^{-1}),$$

where $\hat{k}_j = k_j + O_p(n^{-1/2})$. Hence, $\hat{K}(t) = K(t) + O_p(n^{-1})$ and $\hat{K}^{-1}(\alpha) = K^{-1}(\alpha) + O_p(n^{-1})$, and since $\hat{\sigma}$ is of order $O_p(n^{-1/2})$, the bootstrap-t confidence limit $\hat{\theta}_T[\alpha]$ satisfies

$$\begin{aligned}\hat{\theta}_T[\alpha] &= \hat{\theta} - \hat{\sigma}\hat{K}^{-1}(1-\alpha) \\ &= \hat{\theta} - \hat{\sigma}K^{-1}(1-\alpha) + O_p(n^{-3/2}) \\ &= \hat{\theta}_{\text{exact}}[\alpha] + O_p(n^{-3/2}).\end{aligned} \tag{8.6}$$

Expression (8.6) shows that the bootstrap-t method is second-order correct.

To demonstrate the second-order correctness of the BC_a method, let $\hat{H}(u)$ be the cumulative bootstrap distribution function of U, so that $\hat{H}(u)$ is the cumulative distribution function of $U^* = (\hat{\theta}^* - \hat{\theta})/\hat{\sigma}$. It is assumed that the estimator $\hat{\sigma}^2$ is such that the bootstrap distribution of $\hat{\theta}$ has variance that differs from $\hat{\sigma}^2$ by order $O_p(n^{-2})$, that is, $\operatorname{var}(\hat{\theta}^*) = \hat{\sigma}^2 + O_p(n^{-2})$. The first three cumulants of U^* typically differ from those of U by order $O_p(n^{-1})$, so $\hat{H}(u) = H(u) + O_p(n^{-1})$ and $\hat{H}^{-1}(\alpha) = H^{-1}(\alpha) + O_p(n^{-1})$. The bootstrap cumulative distribution function $\hat{G}(c)$ of $\hat{\theta}$ satisfies $\hat{G}(c) = \hat{H}\big\{(c - \hat{\theta})/\hat{\sigma}\big\}$, and $\hat{G}^{-1}(\alpha) = \hat{\theta} + \hat{\sigma}\hat{H}^{-1}(\alpha)$. Thus, (8.3) gives

$$\hat{G}^{-1}(\alpha) = \hat{\theta} + \hat{\sigma}z^{(\alpha)} + n^{-1/2}\hat{\sigma} \cdot\big[(k_1 - \tfrac{1}{6}k_3) + \tfrac{1}{6}k_3\big\{z^{(\alpha)}\big\}^2\big] + O_p(n^{-3/2}),$$

and, by definition (2.3),

$$\begin{aligned}\hat{\theta}_{BC_a}[\alpha] &= \hat{G}^{-1}\left[\Phi\left\{z_0 + \frac{z_0 + z^{(\alpha)}}{1 - a(z_0 + z^{(\alpha)})}\right\}\right] \\ &= \hat{G}^{-1}\big\{\Phi\big(z^{(\alpha)} + 2z_0 + a\big\{z^{(\alpha)}\big\}^2\big)\big\} + O_p(n^{-3/2}) \\ &= \hat{\theta} + \hat{\sigma}z^{(\alpha)} + n^{-1/2}\hat{\sigma} \cdot\left[2\sqrt{n}z_0 + \left(k_1 - \frac{1}{6}k_3\right) + \left(\sqrt{n}a + \frac{1}{6}k_3\right)\big\{z^{(\alpha)}\big\}^2\right] + O_p(n^{-3/2}).\end{aligned} \tag{8.7}$$

Comparison of (8.5) and (8.7) shows that $\hat{\theta}_{BC_a}[\alpha]$ is second-order correct when a and z_0 are defined by

$$a = (\tfrac{1}{2}k_2 - \tfrac{1}{3}k_3)/\sqrt{n}, \tag{8.8}$$

$$z_0 = -(k_1 - \tfrac{1}{6}k_3)/\sqrt{n}. \tag{8.9}$$

The quantities a and z_0 are of order $O(n^{-1/2})$. The quantity a satisfies

$$a = -\tfrac{1}{6}\big\{\operatorname{skew}(U) + \operatorname{skew}(T)\big\} + O(n^{-1}),$$

and interpretation (2.7) for z_0 is easily seen from (8.1), for

$$\begin{aligned}\Phi(z_0) &= \Phi\big\{-(k_1 - \tfrac{1}{6}k_3)/\sqrt{n}\big\} \\ &= H(0) + O(n^{-1}) \\ &= \operatorname{pr}\big\{\hat{\theta} \le \theta\big\} + O(n^{-1}).\end{aligned}$$

In practice, $\hat{\theta}_{BC_a}[\alpha]$ is calculated using estimates $\hat{a}$ and $\hat{z}_0$ that differ from a and z_0 by order $O_p(n^{-1})$; expression (8.7) shows that this change does not affect the second-order correctness of $\hat{\theta}_{BC_a}[\alpha]$. The estimate $\hat{z}_0$ given in expression (2.8) has this property, since

$$\begin{aligned}\hat{z}_0 &= \Phi^{-1}\big\{\hat{G}(\hat{\theta})\big\} = \Phi^{-1}\big\{\hat{H}(0)\big\} \\ &= \Phi^{-1}\big\{H(0)\big\} + O_p(n^{-1}) \\ &= \Phi^{-1}\big[\Phi\big\{-(k_1 - \tfrac{1}{6}k_3)/\sqrt{n}\big\}\big] + O_p(n^{-1}) \\ &= z_0 + O_p(n^{-1}).\end{aligned}$$

The second-order correctness of the bootstrap-t and the BC_a methods has been discussed by Efron (1987), Bickel (1987, 1988) Hall (1988) and DiCiccio and Romano (1995).

Definitions (8.8) and (8.9) for a and z_0 can be used to cast expansion (8.5) for $\hat{\theta}_{\text{exact}}[\alpha]$ into the form of (4.20). In particular,

$$\hat{\theta}_{\text{exact}}[\alpha] = \hat{\theta} + \hat{\sigma}z^{(\alpha)} + \hat{\sigma}\big[z_0 + (2a + c_q)\big\{z^{(\alpha)}\big\}^2\big] + O_p(n^{-3/2}), \tag{8.10}$$

where

$$c_q = -(\tfrac{1}{2}k_2 - \tfrac{1}{2}k_3)/\sqrt{n}. \tag{8.11}$$

The bias of $\hat{\theta}$ is

$$b = \sigma k_1/\sqrt{n}, \tag{8.12}$$

and z_0 can be expressed in terms of a, c_q and b by

$$\begin{aligned}z_0 &= a + c_q - b/\sigma \\ &= \Phi^{-1}\big(2\Phi(a)\Phi(c_q - b/\sigma)\big) + O(n^{-1}).\end{aligned} \tag{8.13}$$

If $\hat{c}_q$ and $\hat{b}$ are estimates that differ from c_q and b by order $O_p(n^{-1})$, then estimate (4.12),

$$\hat{z}_0 = \Phi^{-1}\big(2\Phi(\hat{a})\Phi(\hat{c}_q - \hat{b}/\hat{\sigma})\big) \tag{8.14}$$

differs from z_0 by the same order.

Once estimates $(\hat{\theta}, \hat{\sigma}, \hat{a}, \hat{z}_0, \hat{c}_q)$ are obtained, the quadratic version of the ABC confidence limit,

$\hat{\theta}_{\mathrm{ABC}_q}[\alpha] = \hat{\theta} + \hat{\sigma}\xi$, can be constructed according to definition (4.13). This limit is second-order correct. Since

$$\begin{aligned} w &= \hat{z}_0 + z^{(\alpha)} = z_0 + z^{(\alpha)} + O_p(n^{-1}), \\ \lambda &= w\big(1 - \hat{a}w\big)^{-2} \\ &= z^{(\alpha)} + z_0 + 2a\big\{z^{(\alpha)}\big\}^2 + O_p(n^{-1}), \\ \xi &= \lambda + \hat{c}_q\lambda^2 \\ &= z^{(\alpha)} + z_0 + (2a + c_q)\big\{z^{(\alpha)}\big\}^2 + O_p(n^{-1}), \end{aligned} \tag{8.15}$$

$\hat{\theta}_{ABC_q}[\alpha]$ agrees with (8.10) to error of order $O_p(n^{-3/2})$.

In many contexts, there exists a vector of parameters $\zeta = (\zeta_1, \ldots, \zeta_p)'$ and an estimator $\hat{\zeta} = (\hat{\zeta}_1, \ldots, \hat{\zeta}_p)'$ such that the parameter of interest is $\theta = t(\zeta)$, and the variance of the estimator $\hat{\theta} = t(\hat{\zeta})$ is of the form $\sigma^2 = v(\zeta) + O(n^{-2})$, so the variance is estimated by $\hat{\sigma}^2 = v(\hat{\zeta})$. This situation arises in parametric models and in the *smooth function of means model*. For the smooth function model, inference is based on independent and identically distributed vectors $x_1, \ldots, x_n$, each having mean μ; the parameter of interest is $\theta = t(\mu)$, which is estimated by $\hat{\theta} = t(\bar{x})$. In fact the smooth function model is closely related to exponential families, as shown in Section 4 of DiCiccio and Efron (1992).

Assume that $\sqrt{n}(\hat{\zeta} - \zeta)$ is normally distributed asymptotically. Typically, the first three joint cumulants of $\hat{\zeta}_1, \ldots, \hat{\zeta}_p$ are

$$\begin{gathered} E(\hat{\zeta}_i) = \zeta_i + \kappa_i, \quad \operatorname{cov}(\hat{\zeta}_i, \hat{\zeta}_j) = \kappa_{i,\,j}, \\ \operatorname{cum}(\hat{\zeta}_i, \hat{\zeta}_j, \hat{\zeta}_k) = \kappa_{i,\,j,\,k}, \quad i, j, k = 1, \ldots, p, \end{gathered}$$

where κ_i and $\kappa_{i,\,j}$ are of order $O(n^{-1})$ and $\kappa_{i,\,j,\,k}$ is of order $O(n^{-2})$, and the fourth- and higher-order joint cumulants are of order $O(n^{-3})$ or smaller. Straightforward calculations show that $\sigma^2 = \kappa_{i,\,j}t_it_j + O(n^{-2})$, where $t_i = \partial t(\zeta)/\partial\zeta_i$, $i = 1, \ldots, p$. In this expression and subsequently, the usual convention is used whereby summation over repeated indices is understood, with the range of summation being $1, \ldots, p$. Now, suppose ζ is sufficiently rich so that $\kappa_{i,\,j}$ depends on the underlying distribution only through ζ for indices i and j such that t_i and t_j are nonvanishing. Then it is possible to write

$$v(\zeta) = \kappa_{i,\,j}(\zeta)t_i(\zeta)t_j(\zeta) + O(n^{-2})$$

and

$$\hat{\sigma}^2 = v(\hat{\zeta}) = \kappa_{i,\,j}(\hat{\zeta})t_i(\hat{\zeta})t_j(\hat{\zeta}) + O_p(n^{-2}).$$

In this case, the quantities k_1, k_2, k_3 are given by

$$\begin{aligned} k_1 &= \sqrt{n}\big(\kappa_i t_i + \tfrac{1}{2}\kappa_{i,\,j}t_{ij}\big)/\big(\kappa_{i,\,j}t_it_j\big)^{1/2}, \\ k_2 &= \sqrt{n}\kappa_{i,\,j}v_it_j/\big(\kappa_{i,\,j}t_it_j\big)^{3/2} \\ &= \sqrt{n}\big(\kappa_{i,\,j/l}\kappa_{k,\,l}t_it_jt_k + 2\kappa_{i,\,j}\kappa_{k,\,l}t_it_kt_{jl}\big)/ \\ &\qquad \big(\kappa_{i,\,j}t_it_j\big)^{3/2}, \\ k_3 &= \sqrt{n}\big(\kappa_{i,\,j,\,k}t_it_jt_k + 3\kappa_{i,\,j}\kappa_{k,\,l}t_it_kt_{jl}\big)/ \\ &\qquad \big(\kappa_{i,\,j}t_it_j\big)^{3/2}, \end{aligned} \tag{8.16}$$

to error of order $O(n^{-1/2})$, where $t_{ij} = \partial^2 t(\zeta)/\partial\zeta_i\partial\zeta_j$, $v_i = \partial v(\zeta)/\partial\zeta_i$, $\kappa_{i,\,j/k} = \partial\kappa_{i,\,j}(\zeta)/\partial\zeta_k$, $i, j, k = 1, \ldots, p$. It follows from (8.8), (8.11) and (8.12) that

$$\begin{aligned} a &= \big(\tfrac{1}{2}\kappa_{i,\,j/l}\kappa_{k,\,l} - \tfrac{1}{3}\kappa_{i,\,j,\,k}\big)t_it_jt_k/ \\ &\qquad \big(\kappa_{i,\,j}t_it_j\big)^{3/2}, \\ b &= \kappa_i t_i + \tfrac{1}{2}\kappa_{i,\,j}t_{ij}, \\ c_q &= -\big(\tfrac{1}{2}\kappa_{i,\,j/l}\kappa_{k,\,l} - \tfrac{1}{2}\kappa_{i,\,j,\,k}\big)t_it_jt_k/ \\ &\qquad \big(\kappa_{i,\,j}t_it_j\big)^{3/2} \\ &\qquad + \tfrac{1}{2}\kappa_{i,\,j}\kappa_{k,\,l}t_it_kt_{jl}/\big(\kappa_{i,\,j}t_it_j\big)^{3/2} \end{aligned} \tag{8.17}$$

to error of order $O(n^{-1})$. An expression for z_0 having error of order $O(n^{-1})$ can be deduced from (8.17) by using (8.13).

The ABC method applies to both exponential families and the smooth function of means model. For these cases, $\hat{\zeta}$ is an unbiased estimate of ζ, and the cumulant generating function of $\hat{\zeta}$, $\Psi(\xi) = \log E\big\{\exp\big(\xi_i\hat{\zeta}_i\big)\big\}$, has an approximation $\hat{\Psi}(\xi)$ such that

$$\begin{aligned} \left.\frac{\partial\hat{\Psi}(\xi)}{\partial\xi_i}\right|_{\xi=0} &= \hat{\zeta}_i, \\ \left.\frac{\partial^2\hat{\Psi}(\xi)}{\partial\xi_i\partial\xi_j}\right|_{\xi=0} &= \kappa_{i,\,j}(\hat{\zeta}) + O_p(n^{-2}), \\ \left.\frac{\partial^3\hat{\Psi}(\xi)}{\partial\xi_i\partial\xi_j\partial\xi_k}\right|_{\xi=0} &= \kappa_{i,\,j,\,k} + O_p(n^{-5/2}). \end{aligned}$$

In particular, it is reasonable to take $\hat{\sigma}^2 = \hat{\Psi}_{ij}\hat{t}_i\hat{t}_j$, where $\hat{t}_i = t_i(\hat{\zeta})$, $i = 1, \ldots, p$. The ABC algorithm uses numerical differentiation of $t(\zeta)$ and $\hat{\Psi}_i(\xi)$ to facilitate calculation of estimates $\hat{\sigma}$, $\hat{a}$, $\hat{z}_0$, $\hat{c}_q$.

In exponential families, the distribution of an observed random vector $y = (y_1, \ldots, y_p)'$ is indexed by an unknown parameter $\overline{\eta} = (\overline{\eta}_1, \ldots, \overline{\eta}_p)'$, and the log-likelihood function for $\overline{\eta}$ based on y has the form $l(\overline{\eta}; y) = n\big\{\overline{\eta}_i y_i - \overline{\psi}(\overline{\eta})\big\}$, where $y = E(y) + O_p(n^{-1/2})$ and both $\overline{\eta}$ and $\overline{\psi}(\overline{\eta})$ are of order $O(1)$. In this case, y plays the role of $\hat{\zeta}$, and ζ corresponds to the expectation parameter $\mu = E(y) = \partial\overline{\psi}(\overline{\eta})/\partial\overline{\eta}$.

Upon defining η and $\psi(\eta)$ by $\eta = n\overline{\eta}$ and $\psi(\eta) = n\overline{\psi}(\overline{\eta}) = n\overline{\psi}(\eta/n)$, the log-likelihood function for η based on y is $l(\eta; y) = \eta' y - \psi(\eta)$, which agrees with (3.1). The cumulant generating function for y is $\Psi(\xi) = \psi(\eta + \xi) - \psi(\eta)$, and the approximate cumulant generating function is

$$\hat{\Psi}(\xi) = \psi(\hat{\eta} + \xi) - \psi(\hat{\eta}),$$

where $\hat{\eta}$ is the maximum likelihood estimator obtained from the equations $\psi_i(\hat{\eta}) = y_i$, $i = 1, \ldots, p$. The usual information estimate of variance is $\hat{\sigma}^2 = \psi_{ij}(\hat{\eta})\hat{t}_i\hat{t}_j = \hat{\Psi}_{ij}\hat{t}_i\hat{t}_j$.

In the smooth function model, the cumulant generating function is approximated by

$$\hat{\Psi}(\xi) = n \log\left\{\frac{1}{n}\sum_{j=1}^{n} \exp\left(\frac{\xi_i x_{ij}}{n}\right)\right\},$$

which is the true cumulant generating function for the model that puts probability mass $1/n$ on each of the observed random vectors $x_j = (x_{1j}, \ldots, x_{pj})'$, $j = 1, \ldots, n$. The usual estimate of variance obtained from the delta-method is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n^2}\left\{\sum_{k=1}^{n}(x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)\right\}\hat{t}_i\hat{t}_j \\ &= \hat{\Psi}_{ij}\hat{t}_i\hat{t}_j,\end{aligned}$$

where $\overline{x}_i = \sum x_{ij}/n$.

Key features of exponential families and the smooth function model are that $\kappa_i = 0$ and $\kappa_{i,\,j/l}\kappa_{k,\,l} = \kappa_{i,\,j,\,k}$, $i, j, k = 1, \ldots, p$, so the expressions for a, b and c given in (5.17) undergo considerable simplification; in particular,

$$\begin{aligned}a &= \tfrac{1}{6}\kappa_{i,\,j,\,k}t_i t_j t_k/\left(\kappa_{i,\,j}t_i t_j\right)^{3/2}, \\ b &= \tfrac{1}{2}\kappa_{i,\,j}t_{ij}, \\ c_q &= \tfrac{1}{2}\kappa_{i,\,j}\kappa_{k,\,l}t_i t_k t_{jl}/\left(\kappa_{i,\,j}t_i t_j\right)^{3/2},\end{aligned}$$

to error of order $O(n^{-1})$.

The ABC method requires only that $t(\zeta)$ and $\hat{\Psi}_i(\xi)$ be specified; the estimates $\hat{\sigma}$, $\hat{a}$, $\hat{z}_0$, and $\hat{c}_q$ are obtained by numerical differentiation. The details are as follows. By definition,

$$\begin{aligned}\hat{t}_i &= \frac{d}{d\varepsilon}t(\hat{\zeta} + \varepsilon e_i)\bigg|_{\varepsilon=0}, \quad i = 1, \ldots, p, \\ \hat{\Psi}_{ij} &= \frac{d}{d\varepsilon}\hat{\Psi}_i(\varepsilon e_j)\bigg|_{\varepsilon=0}, \quad i, j = 1, \ldots, p,\end{aligned}$$

where e_i is the p-dimensional unit vector whose ith entry is 1. Let $\dot{t} = (\hat{t}_1, \ldots, \hat{t}_p)'$, $\hat{\Sigma} = (\hat{\Psi}_{ij})$, $\hat{\sigma}^2 = \hat{\Psi}_{ij}\hat{t}_i\hat{t}_j = \dot{t}'\hat{\Sigma}\dot{t}$. Then

$$\begin{aligned}\hat{a} &= \frac{\hat{\Psi}_{ijk}\hat{t}_i\hat{t}_j\hat{t}_k}{6\hat{\sigma}^3} = \frac{1}{6\hat{\sigma}^3}\frac{d^2}{d\varepsilon^2}\hat{t}_i\hat{\Psi}_i(\varepsilon\dot{t})\bigg|_{\varepsilon=0}, \\ \hat{c}_q &= \frac{\hat{\Psi}_{ij}\hat{\Psi}_{kl}\hat{t}_i\hat{t}_j\hat{t}_{kl}}{2\hat{\sigma}^3} = \frac{1}{2\hat{\sigma}}\frac{d^2}{d\varepsilon^2}t\left(\hat{\zeta} + \varepsilon\frac{\hat{\Sigma}\dot{t}}{\hat{\sigma}}\right)\bigg|_{\varepsilon=0}.\end{aligned}$$

Now $\hat{\Sigma} = \Gamma D \Gamma'$, where D is a diagonal matrix of eigenvalues of $\hat{\Sigma}$ and Γ is an orthogonal matrix whose columns are corresponding eigenvectors. Denote the ith diagonal element of D by d_i and the ith column of Γ by $\gamma_i = (\gamma_{1i}, \ldots, \gamma_{pi})'$, so that $\hat{\Psi}_{ij} = \sum_k d_k\gamma_{ik}\gamma_{jk}$. The quantity b can be estimated by

$$\hat{b} = \frac{\hat{\Psi}_{ij}\hat{t}_{ij}}{2} = \frac{1}{2}\sum_{i=1}^{p}\frac{d^2}{d\varepsilon^2}t\left(\hat{\zeta} + \varepsilon d_i^{1/2}\gamma_i\right)\bigg|_{\varepsilon=0}.$$

If calculating the eigenvalues and eigenvectors is too cumbersome, then $\hat{b}$ can be obtained from

$$\hat{b} = \frac{1}{2}\sum_{i=1}^{p}\frac{\partial^2}{\partial\varepsilon_1\partial\varepsilon_2}t\left(\hat{\zeta} + \varepsilon_1 e_i + \varepsilon_2\hat{\Sigma}e_i\right)\bigg|_{(\varepsilon_1,\,\varepsilon_2)=(0,0)}.$$

Once $\hat{\sigma}^2$, $\hat{a}$, $\hat{b}$, and $\hat{c}$ are calculated, then $\hat{z}_0$ can be obtained using (8.14).

The ABC confidence limit $\hat{\theta}_{ABC}[\alpha]$ is defined in (8.14) as

$$\hat{\theta}_{ABC}[\alpha] = t\left(\hat{\zeta} + \frac{\lambda\hat{\Sigma}\dot{t}}{\hat{\sigma}}\right).$$

This confidence limit is second-order correct; by (5.10) and (5.15),

$$\begin{aligned}\hat{\theta}_{ABC}[\alpha] &= \hat{\theta} + \lambda\frac{\hat{t}_i\hat{\Psi}_{ij}\hat{t}_j}{\hat{\sigma}} + \lambda^2\frac{\hat{t}_{ij}\hat{\Psi}_{ik}\hat{\Psi}_{jl}\hat{t}_k\hat{t}_l}{2\hat{\sigma}^2} \\ &\quad + O_p(n^{-3/2}) \\ &= \hat{\theta} + \hat{\sigma}\lambda + \hat{\sigma}\hat{c}_q\lambda^2 + O_p(n^{-3/2}) \\ &= \hat{\theta} + \hat{\sigma}\left[z^{(\alpha)} + z_0 + 2a\left\{z^{(\alpha)}\right\}^2\right] \\ &\quad + \hat{\sigma}c_q\left\{z^{(\alpha)}\right\}^2 + O_p(n^{-3/2}) \\ &= \hat{\theta}_{\text{exact}}[\alpha] + O_p(n^{-3/2}).\end{aligned}$$

The second-order correctness of the ABC method for exponential families was shown by DiCiccio and Efron (1992).

9. PARAMETRIC MODELS AND CONDITIONAL CONFIDENCE INTERVALS

An impressive likelihood-based theory of higher-order accurate confidence intervals has been developed during the past decade. This effort has involved many authors, including Barndorff-Nielsen (1986), Cox and Reid (1987), Pierce and Peters

(1992) and McCullagh and Tibshirani (1990). This section concerns the connection of bootstrap confidence intervals with the likelihood-based theory. We will see that in exponential families, including nonparametric situations, the bootstrap can be thought of as an easy, automatic way of constructing the likelihood intervals. However, in parametric families that are not exponential, the two theories diverge. There the likelihood intervals are second-order accurate in a conditional sense, while the bootstrap intervals' accuracy is only unconditional. To get good conditional properties, the bootstrap resampling would have to be done according to the appropriate conditional distribution, which would usually be difficult to implement.

Consider an observed random vector $y = (y_1, \ldots, y_n)'$ whose distribution depends on an unknown parameter $\zeta = (\zeta_1, \ldots, \zeta_p)'$, and let $l(\zeta) = l(\zeta; y)$ be the log-likelihood function for ζ based on y. Suppose the parameter $\theta = t(\zeta)$ is estimated by $\hat{\theta} = t(\hat{\zeta})$, where $\hat{\zeta} = (\hat{\zeta}_1, \ldots, \hat{\zeta}_p)'$ is the maximum likelihood estimator. Parametric bootstrap distributions are generally constructed using samples y^* drawn from the fitted distribution for y, that is, from the distribution having $\zeta = \hat{\zeta}$.

Asymptotic formulae for the first three cumulants of $\hat{\theta}$ are given by McCullagh (1987, Chapter 7), and using these formulae in conjunction with (8.16) shows that $\sigma^2 = \lambda^{i,j}t_it_j + O(n^{-2})$ and

$$
\begin{aligned}
k_1 &= -\sqrt{n}\big[\big(\tfrac{1}{2}\lambda_{i,j,k} + \tfrac{1}{2}\lambda_{ij,k}\big)\lambda^{i,j}\lambda^{k,l}t_l \\
&\qquad - \tfrac{1}{2}\lambda^{i,j}t_{ij}\big]/\big(\lambda^{i,j}t_it_j\big)^{1/2}, \\
k_2 &= -\sqrt{n}\big[\big(\lambda_{i,j,k} + 2\lambda_{ij,k}\big)\lambda^{i,l}\lambda^{j,m}\lambda^{k,n}t_lt_mt_n \\
&\qquad - 2\lambda^{i,j}\lambda^{k,l}t_it_kt_{jl}\big]/\big(\lambda^{i,j}t_it_j\big)^{3/2}, \\
k_3 &= -\sqrt{n}\big[\big(2\lambda_{i,j,k} + 3\lambda_{ij,k}\big)\lambda^{i,l}\lambda^{j,m}\lambda^{k,n}t_lt_mt_n \\
&\qquad - 3\lambda^{i,j}\lambda^{k,l}t_it_kt_{jl}\big]/\big(\lambda^{i,j}t_it_j\big)^{3/2},
\end{aligned}
\tag{9.1}
$$

to error of order $O(n^{-1/2})$, where $\lambda_{i,j} = E\big(l_il_j\big)$, $\lambda_{ij,k} = E\big(l_{ij}l_k\big)$, $\lambda_{i,j,k} = E\big(l_il_jl_k\big)$, with $l_i = \partial l(\zeta)/\partial\zeta_i$ and $l_{ij} = \partial^2 l(\zeta)/\partial\zeta_i\partial\zeta_j$, and $(\lambda^{i,j})$ is the $p \times p$ matrix inverse of $(\lambda_{i,j})$. The quantities $\lambda_{i,j}$, $\lambda_{ij,k}$ and $\lambda_{i,j,k}$ are assumed to be of order $O(n)$. The expected information estimate of variance is $\hat{\sigma}^2 = \hat{\lambda}^{i,j}\hat{t}_i\hat{t}_j$, where $\hat{\lambda}^{i,j} = \lambda^{i,j}(\hat{\zeta})$, and the variance of the bootstrap distribution of $\hat{\theta}$ satisfies $\operatorname{var}(\hat{\theta}^*) = \hat{\sigma}^2 + O_p(n^{-2})$. Thus, if the Studentized statistic is defined using the expected information estimate of variance, say $T_E = (\hat{\theta} - \theta)/\hat{\sigma}$, then the results of Section 5 show that the BC_a method is second-order correct with respect to T_E. Using (8.8) in conjunction with (9.1) to calculate a yields

$$
a = \tfrac{1}{6}\lambda_{i,j,k}\lambda^{i,l}\lambda^{j,m}\lambda^{k,n}t_lt_mt_n/\big(\lambda^{ij}t_it_j\big)^{3/2}, \tag{9.2}
$$

to error of order $O(n^{-1})$. This formula for a was given by Efron (1987).

If nuisance parameters are absent ($p = 1$) and $\theta = \zeta$, then (8.9), (9.1), and (9.2) show that

$$
\begin{aligned}
a = z_0 &= \tfrac{1}{6}\lambda_{1,1,1}(\lambda_{1,1})^{-3/2} \\
&= \tfrac{1}{6}\operatorname{skew}\big(\partial l(\theta)/\partial\theta\big),
\end{aligned}
\tag{9.3}
$$

to error of order $O(n^{-1})$. The equality of z_0 and a in this context was demonstrated by Efron (1987).

In addition to being invariant under monotonically increasing transformations of the parameter of interest as described in Section 3, the quantities a and z_0 are also invariant under reparameterizations $\eta = \eta(\zeta)$ of the model. Expression (9.2) for a is invariant under reparameterizations of the model, as is the formula for z_0 obtained by substituting (9.1) into (8.9). There is no restriction then in assuming the model is parameterized so that $\theta = \zeta^1$ and the nuisance parameters $\zeta^2, \ldots, \zeta^p$ are orthogonal to θ. Here, orthogonality means $\lambda_{1,a} = \lambda^{1,a} = 0$ $(a = 2, \ldots, p)$; see Cox and Reid (1987). In this case, (6.2) becomes

$$
a = \tfrac{1}{6}\lambda_{1,1,1}(\lambda_{1,1})^{-3/2} = \tfrac{1}{6}\operatorname{skew}\big(\partial l(\zeta)/\partial\zeta^1\big). \tag{9.4}
$$

Comparison of (9.4) with (9.3) indicates that, to error of order $O(n^{-1})$, a coincides with its version that would apply if the orthogonal nuisance parameters were known. In this sense, a can be regarded as unaffected by the presence of nuisance parameters. In contrast, for the orthogonal case,

$$
\begin{aligned}
z_0 &= \big(\tfrac{1}{2}\lambda_{a,b,1} + \tfrac{1}{2}\lambda_{ab,1}\big)\lambda^{a,b}(\lambda_{1,1})^{-1/2} \\
&\qquad + \tfrac{1}{6}\lambda_{1,1,1}(\lambda_{1,1})^{-3/2},
\end{aligned}
\tag{9.5}
$$

to error of order $O(n^{-1})$, where, for purpose of the summation convention, the indices a and b range over $2, \ldots, p$. Expression (9.5) shows that z_0 reflects the presence of unknown nuisance parameters.

Another possibility for Studentizing is to use the observed information estimate of variance, $\bar{\sigma}^2 = -\hat{l}^{ij}\hat{t}_i\hat{t}_j$, where $\big(\hat{l}^{ij}\big)$ is the $p \times p$ matrix inverse of $\big(\hat{l}_{ij}\big)$ and $\hat{l}_{ij} = l_{ij}(\hat{\zeta})$. Let $T_O = (\hat{\theta} - \theta)/\bar{\sigma}$. Using the bootstrap-$t$ method with T_E and T_O produces approximate confidence limits $\hat{\theta}_{T_E}[\alpha]$ and $\hat{\theta}_{T_O}[\alpha]$, which both have coverage error of order $O(n^{-1})$. However, $\bar{\sigma} = \hat{\sigma} + O_p(n^{-1})$, so $T_O = T_E + O_p(n^{-1/2})$, and $\hat{\theta}_{T_E}[\alpha]$ and $\hat{\theta}_{T_O}[\alpha]$ typically differ by order $O_p(n^{-1})$. The Studentized quantities T_E and T_O produce different definitions of second-order correctness. In particular, $\hat{\theta}_{BC_a}[\alpha]$ differs from $\hat{\theta}_{T_O}[\alpha]$ by order $O_p(n^{-1})$, and the BC_a method, which is second-order correct with respect to T_E, fails to be second-order correct with respect to T_O. For exponential families, $\hat{\sigma}^2 = \bar{\sigma}^2$ since

$\lambda^{i,j} = -l^{ij}$, and no distinction arises between T_E and T_O in the definition of second-order correctness.

Although T_E and T_O generally differ by order $O_p(n^{-1/2})$, their first three cumulants agree to error of order $O(n^{-1})$. It follows then from (5.5) that $\hat\theta_{T_E}[\alpha]$ and $\hat\theta_{T_O}[\alpha]$ have expansions

$$\begin{aligned}\hat\theta_{T_E}[\alpha] &= \hat\theta + \hat\sigma z^{(\alpha)} - n^{-1/2}\hat\sigma \\ &\quad \cdot \left[(k_1 - \tfrac{1}{6}k_3) - (\tfrac{1}{2}k_2 - \tfrac{1}{6}k_3)\{z^{(\alpha)}\}^2\right] \\ &\quad + O_p(n^{-3/2}), \\ \hat\theta_{T_O}[\alpha] &= \hat\theta + \overline{\sigma} z^{(\alpha)} - n^{-1/2}\overline{\sigma} \\ &\quad \cdot \left[(k_1 - \tfrac{1}{6}k_3) - (\tfrac{1}{2}k_2 - \tfrac{1}{6}k_3)\{z^{(\alpha)}\}^2\right] \\ &\quad + O_p(n^{-3/2}),\end{aligned} \tag{9.6}$$

where k_1, k_2, k_3 are given by (9.1). Expression (9.6) shows that if $\hat\theta_E[\alpha]$ is a second-order correct confidence limit with respect to T_E, such as $\hat\theta_{BC_a}[\alpha]$, then

$$\hat\theta_O[\alpha] = \hat\theta + \frac{\overline{\sigma}}{\hat\sigma}\left(\hat\theta_E[\alpha] - \hat\theta\right)$$

is second-order correct with respect to T_O.

Confidence limits that are second-order correct with respect to T_O agree closely with second-order accurate confidence limits obtained from likelihood ratio statistics. The profile *log-likelihood function* for θ is $l_p(\theta) = l(\hat\zeta_\theta)$, where $\hat\zeta_\theta$ is the constrained maximum likelihood estimator of ζ given θ; that is, $\hat\zeta_\theta$ maximizes $l(\zeta)$ subject to the constraint $t(\zeta) = \theta$. Since $\hat\zeta_{\hat\theta}$ is the global maximum likelihood estimator $\hat\zeta$, $l_p(\theta)$ is maximized at $\hat\theta$. The likelihood ratio statistic for θ is

$$W_p(\theta) = 2\{l(\hat\zeta) - l(\hat\zeta_\theta)\} = 2\{l_p(\hat\theta) - l_p(\theta)\},$$

and the signed root of the likelihood ratio statistic is

$$R_p(\theta) = \operatorname{sgn}(\hat\theta - \theta)\sqrt{W_p(\theta)}.$$

In wide generality, $W_p(\theta)$ and $R_p(\theta)$ are asymptotically distributed as χ^2_1 and $N(0,1)$, respectively.

Straightforward calculations show that the derivatives of $l_p(\theta)$ satisfy $l_p^{(1)}(\hat\theta) = 0$, $l_p^{(2)}(\hat\theta) = -\overline{\sigma}^2$ and

$$\begin{aligned}l_p^{(3)}(\hat\theta) &= \left(-\hat l_{ijk}\hat l^{il}\hat l^{jm}\hat l^{kn}\hat t_l\hat t_m\hat t_n + 3\hat l^{ij}\hat l^{kl}\hat t_i\hat t_k\hat t_{jl}\right)/\overline{\sigma}^6 \\ &= \left(\lambda_{ijk}\lambda^{i,l}\lambda^{j,m}\lambda^{k,n}t_l t_m t_n + 3\lambda^{i,j}\lambda^{k,l}t_i t_k t_{jl}\right)/ \\ &\quad \sigma^6 + O_p(n^{1/2}) \\ &= n^{-1/2}(3k_2 - k_3)/\sigma^3 + O_p(n^{1/2}) \\ &= (2a + c_q)/\sigma^3 + O_p(n^{1/2});\end{aligned}$$

these calculations make use of the Bartlett identities $\lambda_{ij} = E(l_{ij}) = -\lambda_{i,j}$ and

$$\lambda_{ijk} = E(l_{ijk}) = -\lambda_{i,j,k} - \lambda_{ij,k} - \lambda_{ik,j} - \lambda_{jk,i}.$$

Consequently, $W_p(\theta)$ and $R_p(\theta)$ have expansions

$$\begin{aligned}W_p(\theta) &= T_O^2 + n^{-1/2}(k_2 - \tfrac{1}{3}k_3)T_O^3 \\ &\quad + O_p(n^{-1}), \\ R_p(\theta) &= T_O + n^{-1/2}(\tfrac{1}{2}k_2 - \tfrac{1}{6}k_3)T_O^2 \\ &\quad + O_p(n^{-1}).\end{aligned} \tag{9.7}$$

Expansion (9.7) shows that

$$\begin{aligned}E(R_p) &= n^{-1/2}(k_1 - \tfrac{1}{6}k_3) + O(n^{-1}) \\ &= -z_0 + O(n^{-1}),\end{aligned} \tag{9.8}$$

$$\operatorname{var}(R_p) = 1 + O(n^{-1}), \quad \operatorname{skew}(R_p) = O(n^{-1}).$$

Thus, the distribution of $R_p(\theta) + \hat z_0$ is standard normal to error of order $O(n^{-1})$, and the approximate limit $\hat\theta_p[\alpha]$ that satisfies

$$R_p(\hat\theta_p[\alpha]) + \hat z_0 = -z^{(\alpha)} \tag{9.9}$$

is second-order accurate. Moreover, comparing (9.7) with the Cornish–Fisher expansion in (8.2) shows that this limit is second-order correct with respect to T_O. Approximate confidence limits obtained using (9.9) have been discussed by several authors, including Lawley (1956), Sprott (1980), McCullagh (1984) and Barndorff-Nielsen (1986). McCullagh (1984) and Barndorff-Nielsen (1986) have shown that these limits are second-order accurate conditionally; that is, they have conditional coverage error of order $O(n^{-1})$ given exact or approximate ancillary statistics. It follows that second-order conditional coverage accuracy is a property of all approximate confidence limits that are second-order correct with respect to T_O. In contrast, limits that are second-order correct with respect to T_E typically have conditional coverage error of order $O(n^{-1/2})$. Conditional validity provides a reason for preferring T_O over T_E to define "exact" confidence limits.

The profile log likelihood function $l_p(\theta)$ is not a genuine likelihood. In particular, the expectation of the profile score, $l_p^{(1)}(\theta)$, is not identically 0 and is generally of order $O(1)$. It can be shown that

$$E\{l_p^{(1)}(\theta)\} = (a - z_0)/\sigma + O(n^{-1}),$$

and hence, the estimating equation $l_p^{(1)}(\theta) = 0$, which yields the estimate $\hat\theta$, is not unbiased. To eliminate this bias, several authors, including Barndorff-Nielsen (1983, 1994), Cox and Reid (1987, 1993) and McCullagh and Tibshirani (1990), have recommended that the profile log-likelihood function $l_p(\theta)$ be replaced by an adjusted version

$$l_{ap}(\theta) = l_p(\theta) + d(\theta),$$

where the adjustment function $d(\theta)$ satisfies

$$d(\theta) = (\hat{a} - \hat{z}_0)T_O + O_p(n^{-1}), \tag{9.10}$$

so that

$$d^{(1)}(\theta) = -E\{l_p^{(1)}(\theta)\} + O_p(n^{-1}).$$

Hence, $E\{l_{ap}^{(1)}(\theta)\} = O(n^{-1})$, and $l_{ap}(\theta)$ behaves more like a genuine likelihood than does $l_p(\theta)$. For instance, McCullagh and Tibshirani (1990) suggested the adjustment

$$m(\theta) = -\int_{\hat{\theta}}^{\theta} \{a(\hat{\zeta}_u) - z_0(\hat{\zeta}_u)\}/\sigma(\hat{\zeta}_u)\, du. \tag{9.11}$$

The estimator $\hat{\theta}_{ap}$ that maximizes $l_{ap}(\theta)$ satisfies

$$\hat{\theta}_{ap} = \hat{\theta} + (z_0 - a)\sigma + O_p(n^{-3/2}).$$

The adjusted likelihood ratio statistic arising from $l_{ap}(\theta)$ is

$$W_{ap}(\theta) = 2\{l_{ap}(\hat{\theta}_{ap}) - l_{ap}(\theta)\},$$

and its signed root is $R_{ap}(\theta) = \operatorname{sgn}(\hat{\theta}_{ap} - \theta)\sqrt{W_{ap}(\theta)}$. It can be shown that

$$\begin{aligned} W_{ap}(\theta) &= W_p(\theta) + (z_0 - a)T_O + O_p(n^{-1}) \\ R_{ap}(\theta) &= R_p(\theta) + (z_0 - a) + O_p(n^{-1}), \end{aligned} \tag{9.12}$$

so it follows from (6.8) that

$$\begin{aligned} E(R_{ap}) &= -a + O(n^{-1}), \\ \operatorname{var}(R_{ap}) &= 1 + O(n^{-1}), \\ \operatorname{skew}(R_{ap}) &= O(n^{-1}). \end{aligned}$$

Consequently, the approximate confidence limit $\hat{\theta}_{ap}[\alpha]$ that satisfies

$$R_{ap}(\hat{\theta}_{ap}[\alpha]) + \hat{a} = -z^{(\alpha)} \tag{9.13}$$

is a second-order accurate confidence limit. Expansion (9.12) shows that $\hat{\theta}_{ap}[\alpha] = \hat{\theta}_p[\alpha] + O_p(n^{-3/2})$, so $\hat{\theta}_{ap}[\alpha]$ is also second-order correct with respect to T_O. Confidence limits obtained by (9.13) have been discussed by DiCiccio and Efron (1992), DiCiccio and Martin (1993), Efron (1993) and Barndorff-Nielsen and Chamberlin (1994).

Numerical examples, especially in cases where the number of nuisance parameters is large, indicate that the standard normal approximation for $R_{ap}(\theta) + \hat{a}$ can be much more accurate than for $R_p(\theta) + \hat{z}_0$, and hence the limits obtained from (9.13) have better coverage accuracy than limits obtained from (9.12). Now, (9.8) suggests that the distribution of $R_p(\theta)$ is affected by the presence of nuisance parameters at the $O(n^{-1/2})$ level through the quantity z_0. However, the distribution of $R_{ap}(\theta)$ is insensitive to the presence of nuisance parameters at that level, because of the remarks made about a at (9.4).

Consider again the orthogonal case with $\theta = \zeta^1$. Let $R(\theta)$ be the signed root of the likelihood ratio statistic that would apply if the nuisance parameters $\zeta^2, \ldots, \zeta^p$ were known. It follows from the comparison of (9.3) and (9.4) that the distributions of $R(\theta)$ and $R_{ap}(\theta)$ agree to order $O(n^{-1})$, while the distributions of $R(\theta)$ and $R_p(\theta)$ agree only to order $O(n^{-1/2})$. Since $R(\theta)$ does not require estimation of nuisance parameters, its distribution is likely to be fairly close to standard normal. On the other hand, because of presence of nuisance parameters, the distribution of $R_p(\theta)$ can be far from standard normal, and asymptotic corrections can fail to remedy adequately the standard normal approximation.

These remarks can be illustrated by taking θ to be the variance in a normal linear regression model with q regression coefficients. In this case, θ is orthogonal to the regression coefficients, and

$$\begin{aligned} \sigma^2 &= \frac{2\theta^2}{n}, \quad a = \frac{2}{3\sqrt{2n}} + O(n^{-1}), \\ z_0 &= \frac{q}{\sqrt{2n}} + \frac{2}{3\sqrt{2n}} + O(n^{-1}), \end{aligned}$$

by (9.4) and (9.5). Note that a does not involve the nuisance parameters, while z_0 reflects the nuisance parameters through its dependence on q. In this case, $(a - z_0)/\sigma = -q/(2\theta)$, and (9.11) produces the adjustment function $d(\theta) = (q/2)\log\theta$. The effect making this adjustment to the profile log-likelihood is to account for the degrees of freedom; in particular, $\hat{\theta}_{ap} = n\hat{\theta}/(n-q)$. Table 7 shows, in the case $n = 8$ and $q = 3$, the true left-hand tail probabilities of approximate quantiles for R_p, R_{ap}, R and their mean-adjusted versions obtained using the standard normal approximation. Note the accuracy and the closeness of the approximation for R_{ap} and R; in constrast, the approximation for R_p is very poor.

Approximate confidence limits that are second-order correct with respect to T_O can be used to recover the profile and adjusted profile log-likelihoods, at least to error of order $O_p(n^{-1})$. Suppose that $\hat{\theta}_O[\alpha]$ is second-order correct; then, by (6.9),

$$R_p(\hat{\theta}_O[\alpha]) + \hat{z}_0 = -z^{(\alpha)} + O_p(n^{-1}).$$

It follows that

$$\begin{aligned} l_p(\hat{\theta}_O[\alpha]) = \text{constant} &- \tfrac{1}{2}(z^{(\alpha)} + z_0)^2 \\ &+ O_p(n^{-1}), \end{aligned} \tag{9.14}$$

and, by (6.10),

$$\begin{aligned} l_{ap}(\hat{\theta}_O[\alpha]) = \text{constant} &- \tfrac{1}{2}(z^{(\alpha)} + z_0)^2 \\ &- \{(\hat{a} - \hat{z}_0)/\overline{\sigma}\}\hat{\theta}_O[\alpha] \\ &+ O_p(n^{-1}). \end{aligned} \tag{9.15}$$

TABLE 7
True left-hand tail probabilities of approximate percentage points obtained from the standard normal approximation; table entries are percentages

Nominal	R_p	$R_p + \hat{z}_0$	R_{ap}	$R_{ap} + \hat{a}$	R	$R + \hat{a}$
1	13.67	2.85	1.81	1.19	1.60	1.04
2.5	22.10	5.69	4.18	2.90	3.75	2.58
5	31.34	9.62	7.83	5.68	7.12	5.13
10	43.68	16.32	14.54	11.09	13.44	10.18
50	84.38	56.81	58.41	51.90	56.65	50.08
90	98.60	91.09	93.06	90.58	92.51	89.88
95	99.44	95.38	96.71	95.30	96.42	94.90
97.5	99.77	97.59	98.43	97.65	98.28	97.43
99	99.93	98.98	99.40	99.06	99.34	98.96

Approximations (9.14) and (9.15) to $l_p(\theta)$ and $l_{ap}(\theta)$ are especially useful in complex situations. Efron (1993) discussed the use of second-order correct confidence limits, particularly the ABC limits, to construct implied likelihoods automatically in both parametric and nonparametric situations.

Second-order accurate confidence limits can also be constructed by using Bayesian methods with noninformative prior distributions. Assume $\theta = \zeta^1$, with the nuisance parameters $\zeta^2, \ldots, \zeta^p$ not necessarily orthogonal to θ, and consider Bayesian inference based on a prior density $\pi(\zeta)$. DiCiccio and Martin (1993) showed that the posterior distribution of

$$R_p + \frac{1}{R_p}\log\left(\frac{S}{R_p}\right), \tag{9.16}$$

is standard normal to error of order $O(n^{-3/2})$, where

$$S = l_1(\hat{\zeta}_\theta)\{-l^{11}(\hat{\zeta}_\theta)\}^{1/2}\frac{\left|-l_{ij}(\hat{\zeta}_\theta)\right|^{1/2}}{\left|-l_{ij}(\hat{\zeta})\right|^{1/2}}\frac{\pi(\hat{\zeta})}{\pi(\hat{\zeta}_\theta)},$$

and $\left|-l_{ij}(\hat{\zeta}_\theta)\right|$ denotes the determinant of the $p \times p$ matrix $\left(-l_{ij}(\hat{\zeta}_\theta)\right)$. Thus, the quantity $\hat{\theta}_\pi[\alpha]$ that satisfies

$$R_p(\hat{\theta}_\pi[\alpha]) + \frac{1}{R_p(\hat{\theta}_\pi[\alpha])} \cdot \log\left(\frac{S(\hat{\theta}_\pi[\alpha])}{R_p(\hat{\theta}_\pi[\alpha])}\right) = -z_0 \tag{9.17}$$

agrees with the posterior α quantile of θ to error of order $O(n^{-2})$.

From a frequentist perspective,

$$S = T_O + O_p(n^{-1/2}) = R_p + O_p(n^{-1/2}),$$

so the adjustment term $R_p^{-1}\log(S/R_p)$ in (6.16) is of order $O_p(n^{-1/2})$ under repeated sampling. Indeed, standard Taylor expansions show that

$$\frac{1}{R_p}\log\left(\frac{S}{R_p}\right) = z_0 + \sum_{i=1}^{p}\frac{\partial}{\partial\zeta^i}\{\lambda^{i,1}(\lambda^{1,1})^{-1/2}\} + \frac{\pi_i(\zeta)}{\pi(\zeta)}\lambda^{i,1}(\lambda^{1,1})^{-1/2} + O_p(n^{-1}), \tag{9.18}$$

where $\pi_i(\zeta) = \partial\pi(\zeta)/\partial\zeta^i$. It is apparent from (9.18) that if the prior density $\pi(\zeta)$ is chosen to satisfy

$$\frac{\pi_i(\zeta)}{\pi(\zeta)}\{\lambda^{i,1}(\lambda^{1,1})^{-1/2}\} = -\sum_{i=1}^{p}\frac{\partial}{\partial\zeta^i}\{\lambda^{i,1}(\lambda^{1,1})^{-1/2}\}, \tag{9.19}$$

then $R_p^{-1}\log(S/R_p) = z_0 + O_p(n^{-1})$. In this case, $\hat{\theta}_\pi[\alpha]$, the solution to (9.17), agrees to error of order $O_p(n^{-3/2})$ with $\hat{\theta}_p[\alpha]$, the solution to (9.9). Consequently, when the prior $\pi(\zeta)$ satisfies (9.19), $\hat{\theta}_\pi[\alpha]$ is second-order correct with respect to T_O, as is the posterior α quantile of θ. These approximate confidence limits also have conditional coverage error of order $O_p(n^{-1})$ given exact or approximate ancillary statistics. Prior distributions for which the posterior quantiles are second-order accurate approximate confidence limits under repeated sampling are usually called noninformative.

Equation (9.19) was given by Peers (1965). When the nuisance parameters $\zeta^2, \ldots, \zeta^p$ are orthogonal to $\theta = \zeta^1$, this equation reduces to

$$\frac{\pi_1(\zeta)}{\pi(\zeta)}(\lambda_{1,1})^{-1/2} = -\frac{\partial}{\partial\zeta^1}(\lambda_{1,1})^{-1/2}.$$

Tibshirani (1989) showed that this equation has solutions of the form

$$\pi(\zeta) \propto (\lambda_{1,1})^{1/2}g,$$

where g is arbitrary and depends only on the nuisance parameters.

REFERENCES

BABU, G. J. and SINGH, K. (1983). Inference on means using the bootstrap. *Ann. Statist.* **11** 999–1003.

BARNDORFF-NIELSEN, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70** 343–365.

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

BARNDORFF-NIELSEN, O. E. (1994). Adjusted versions of profile likelihood and likelihood roots, and extended likelihood. *J. Roy. Statist. Soc. Ser. B* **56** 125–140.

BARNDORFF-NIELSEN, O. E. and CHAMBERLIN, S. R. (1994). Stable and invariant adjusted directed likelihoods. *Biometrika* **81** 485–499.

BERAN, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74** 457–468.

BICKEL, P. J. (1987). Comment on "Better bootstrap confidence intervals" by B. Efron. *J. Amer. Statist. Assoc.* **82** 191.

BICKEL, P. J. (1988). Discussion of "Theoretical comparison of bootstrap confidence intervals" by P. Hall. *Ann. Statist.* **16** 959–961.

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49** 1–39.

COX, D. R. and REID, N. (1993). A note on the calculation of adjusted profile likelihood. *J. Roy. Statist. Soc. Ser. B* **55** 467–472.

DICICCIO, T. J. (1984). On parameter transformations and interval estimation. *Biometrika* **71** 477–485.

DICICCIO, T. J. and EFRON, B. (1992). More accurate confidence intervals in exponential families. *Biometrika* **79** 231–245.

DICICCIO, T. J. and MARTIN, M. (1993). Simple modifications for signed roots of likelihood ratio statistics. *J. Roy. Statist. Soc. Ser. B* **55** 305–316.

DICICCIO, T. J. and ROMANO, J. P. (1995). On bootstrap procedures for second-order accurate confidence limits in parametric models. *Statist. Sinica* **5** 141–160.

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

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

EFRON, B. (1987). Better bootstrap confidence intervals (with discussion). *J. Amer. Statist. Assoc.* **82** 171–200.

EFRON, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80** 3–26.

EFRON, B. (1994). Missing data, imputation, and the bootstrap (with comment and rejoinder). *J. Amer. Statist. Assoc.* **89** 463–478.

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

HALL, P. (1986). On the bootstrap and confidence intervals. *Ann. Statist.* **14** 1431–1452.

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

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

HOUGAARD, P. (1982). Parametrizations of non-linear models. *J. Roy. Statist. Soc. Ser. B* **44** 244–252.

LAWLEY, D. N. (1956). A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika* **43** 295–303.

LOH, W.-Y. (1987). Calibrating confidence coefficients. *J. Amer. Statist. Assoc.* **82** 155–162.

MCCULLAGH, P. (1984). Local sufficiency. *Biometrika* **71** 233–244.

MCCULLAGH, P. (1987). *Tensor Methods in Statistics*. Chapman and Hall, London.

MCCULLAGH, P. and TIBSHIRANI, R. (1990). A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. Ser. B* **52** 325–344.

PEERS, H. W. (1965). On confidence points and Bayesian probability points in the case of several parameters. *J. Roy. Statist. Soc. Ser. B* **27** 9–16.

PIERCE, D. and PETERS, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families (with discussion) *J. Roy. Stat. Soc. Ser. B* **54** 701–725.

SPROTT, D. A. (1980). Maximum likelihood in small samples: estimation in the presence of nuisance parameters. *Biometrika* **67** 515–523.

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

TIBSHIRANI, R. (1989). Noninformative priors for one parameter of many. *Biometrika* **76** 604–608.

Comment

Peter Hall and Michael A. Martin

Peter Hall is Professor and Michael A. Martin is Senior Lecturer, Centre for Mathematics and its Applications, Australian National University, Canberra, A.C.T. 0200, Australia (e-mail: halpstat@durras.anu.edu.au).

Professors DiCiccio and Efron have offered a compelling and insightful look at the current state of research into bootstrap confidence intervals. Their account is both timely and motivating, drawing together important connections between bootstrap confidence intervals and likelihood-based inference and pointing out that there are no uniformly superior methods. The paper also raises several issues that bear further comment, such as those below.

1. WHITHER CONFIDENCE INTERVALS?

As the authors point out in their Introduction, the bootstrap offers a highly accurate and attractive alternative to the "standard interval," which has dominated classical statistical inference for more than 70 years. However, we wonder if, like the standard interval, the whole notion of a confidence interval is not in need of reassessment. It provides a restrictive, one-dimensional level of information about the error associated with a point estimate. Indeed, the information conveyed by confidence intervals is so tied to the classical notion of the standard interval that practitioners have difficulty interpreting confidence intervals in other contexts. For example, it is natural, given two interval endpoints, to imagine that the true parameter value has greatest a posteriori likelihood of lying close to the middle of the interval.

One could incorporate numerical information about left–right asymmetry, for example, in terms of the skewness of the bootstrap estimate of the distribution of a statistic. Information about asymmetry is implicit in so-called short confidence intervals, such as those described by Hall (1988). But why not replace them altogether with more informative tools? The bootstrap affords a unique opportunity for obtaining a large amount of information very simply. The process of setting confidence intervals merely picks two points off a bootstrap histogram, ignoring much relevant information about shape and other important features.

"Confidence pictures" (e.g., Hall, 1992, Appendix III), essentially smoothed and transformed bootstrap histograms, are one alternative to confidence intervals. Graphics such as these provide a simple but powerful way to convey information lost in numerical summaries. The opportunities offered by dynamic graphics are also attractive, particularly when confidence information needs to be passed to a lay audience. (Consider, e.g., the need to provide information about the errors associated with predictions from opinion polls.) Bootstrap methods and new graphical ways of presenting information offer, together, exciting prospects for conveying information about uncertainty.

2. HOW AUTOMATIC SHOULD THE BOOTSTRAP BE?

While an "automatic" procedure, such as some forms of the bootstrap, has advantages, there are also potential problems, just as there may be with "automatic" statistical software in the hands of untrained users. Like any multipurpose tool, the bootstrap can be, and often is, bested by a special-purpose technique, and where such techniques are available, they should be promoted.

Nevertheless, the generality with which the bootstrap applies lends itself readily to solution of problems for which special techniques might not exist. The use of bootstrap methods has recently been greatly facilitated by the publication by Efron and Tibshirani, in their excellent monograph (Efron and Tibshirani, 1993) of a set of S-PLUS routines. However, if use of the bootstrap is to become truly widespread, it should make the leap into mainstream statistical packages.

3. NONPARAMETRIC LIKELIHOOD

The parallels that DiCiccio and Efron draw to likelihood-based parametric theory might be completed by mentioning extensive recent work in the area of nonparametric likelihood. Owen's (1988, 1990) empirical likelihood, Davison, Hinkley and Worton's (1992) bootstrap partial likelihood, and Efron's (1993) implied likelihood could be mentioned in this regard. Efron and Tibshirani (1993b, Chapter 24) provide an excellent review. Perhaps some of the theoretical development given in Section 9 of the paper could be brought to bear in the case of nonparametric likelihood.

We should mention in particular the ties that exist between parametric and empirical likelihood. The nonparametric bootstrap estimator is "maximum likelihood," in that it maximizes Owen's empirical likelihood. Empirical likelihood confidence regions are nonparametric analogues of profile likelihood regions, and the parallels extend to high-order features such as Bartlett correction.

4. PERCENTILE-*t* VERSUS BC_a

One of the more interesting aspects of the development of bootstrap methods has been the debate about relative merits of BC_a and percentile-t. Both methods stem from a simple philosophy that underlies much of statistical theory: inference should ideally be based on statistics whose distributions depend as little as possible on unknown parameters. Percentile-t is based on bootstrapping a quantity whose distribution depends very little on unknowns, and BC_a works by correcting for unknowns in a transformation to normality. The former approach is arguably simpler to use and understand, but not always a good performer.

Indeed, much has been said about the erratic behavior of percentile-t in problems where no obvious, good variance estimator exists. Moreover, there is empirical evidence that in such cases BC_a usually works well. Asymptotic theory is not obviously of

help in solving this mystery, although perhaps the inferior performance of Studentized statistics in approximating large deviation probabilities is at the root of the matter.

5. THE DOUBLE BOOTSTRAP

One might summarize the respective theoretical drawbacks of percentile-t and BC_a methods by noting that the former are not transformation invariant, and the latter are not monotone in coverage level. As a utilitarian procedure we favor a calibrated version of a simple method such as percentile. The percentile method is transformation respecting; its calibrated form "almost" respects transformations and is monotone in coverage level. Also, it is not hindered by problems associated with ratios of random variables, which are sometimes the downfall of percentile-t.

An oft-stated drawback of the double bootstrap is its computational expense. This is perhaps overstated, however, since the amount of computation needed to obtain a *single* double bootstrap interval is really not onerous in today's world of fast, inexpensive computers. Nonetheless, a significant amount of recent work has resulted in development of analytical approximations to double bootstrap confidence intervals that are accurate and that can be computed in a small fraction of the time needed for a double bootstrap calculation carried out by simulation alone. See, for example, Davison and Hinkley (1988), Daniels and Young (1991), DiCiccio, Martin and Young (1992, 1993) and Lee and Young (1995, 1996a). The latter papers by Lee and Young seem particularly promising, as they propose methods for producing approximate double bootstrap confidence intervals without the need for any resampling. A drawback of such analytical methods is that a measure of user intervention is required in setting up and calculating the necessary numerical adjustments, although that would greatly diminish if algorithmic support were provided by readily available software.

Finally, harking back to our original point about the appropriateness of the confidence interval paradigm itself, we note that the double bootstrap is flexible. When applied to the confidence interval problem, it targets a particular feature of interval performance, say coverage error, and uses the bootstrap to estimate and correct for error in that area. If we move from confidence intervals to another form of inference, provided we can quantify the notion of error in our procedure, there is every chance we can still use the double bootstrap to provide accurate inferences.

Comment

A. J. Canty, A. C. Davison and D. V. Hinkley

INTRODUCTION

Both authors have played important roles in developing and deepening our understanding of small-sample confidence interval methods, and we are grateful for the chance to comment on this paper. Time and space are limited, so we shall confine our remarks to the question "What makes a confidence interval reliable?" in the context of a nonparametric bootstrap analysis of the cd4 data.

A. J. Canty is Research Assistant and A. C. Davison is Professor of Statistics, both at the Swiss Federal Institute of Technology, Department of Mathematics, Lausanne, Switzerland. D. V. Hinkley is Professor of Statistics, University of California, Santa Barbara, California (e-mail: hinkley@pstat.ucsb.edu).

DATA ANALYSIS

If the data are of high enough quality to address the substantive question, and the statistic of interest, (here taken to be the largest eigenvalue t) bears on that issue, an applied worker who has constructed a confidence interval will want to know its sensitivity to underlying assumptions, to slight changes in the data and so forth. For a bootstrap interval, these questions can be addressed, to some extent, by examining the simulation output. To illustrate this, we performed 999 nonparametric bootstrap simulations from the cd4 data and, for each simulated dataset, obtained the largest eigenvalue t^* and an estimate v_L^* of its variance. The top left panel of Figure 1 contains the plot of the v_L^* against t^* and shows that the variance is roughly a linear function of the eigenvalue; we explain the plotting

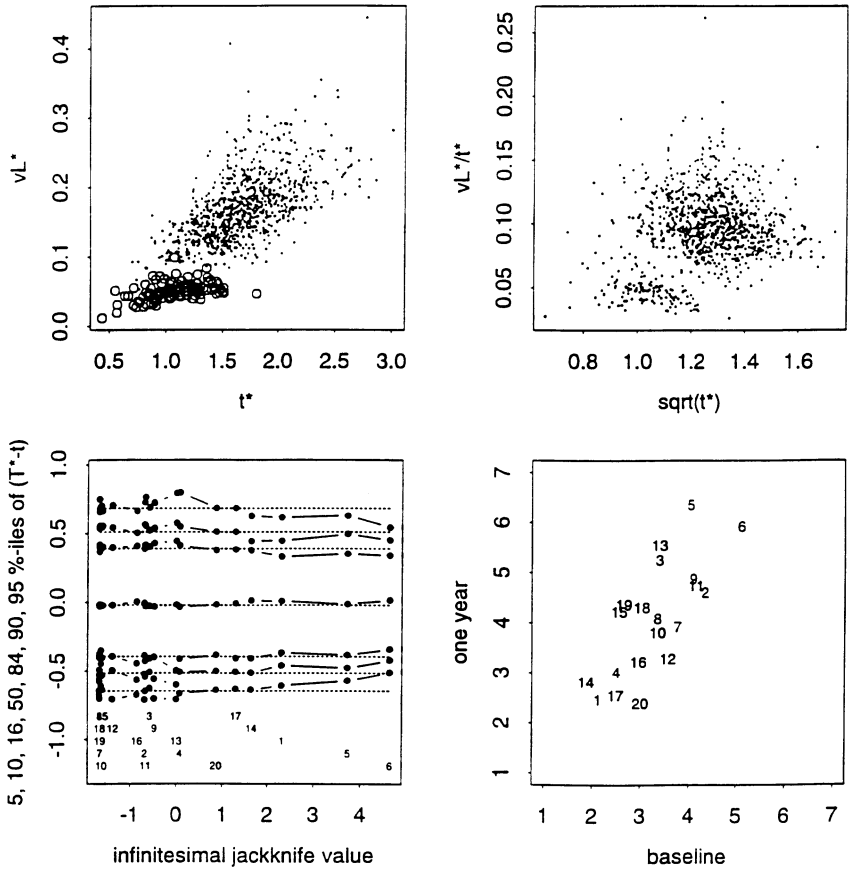

FIG. 1. *Nonparametric bootstrap results for largest eigenvalue of cd4 data; based on 999 nonparametric simulations: (top left) approximate variance v_L^* plotted against bootstrap statistics, t^*, with simulations in which neither case 5 nor case 6 appears marked by circles; (top right) corresponding plot for $t^{*1/2}$; (bottom left) jackknife-after-bootstrap plot for largest eigenvalue; (bottom right) data plot showing case numbers. See text for details.*

symbols below. This suggests that a square root transformation of the eigenvalue will be variance-stabilizing, and this impression is confirmed by the plot of v_L^*/t^* against $t^{*1/2}$ in the top right panel, which shows a weaker relation between the transformed statistic and its variance. This suggests that the square root scale should be used for calculation of any confidence intervals that are not scale-invariant.

However, there is a further difficulty: there is clear bunching in the lower part of the top panels, which suggests some problem with the simulation. The lower left panel of the figure shows a jackknife-after-bootstrap plot for t^* (Efron, 1992). The ingenious idea that underlies this is that we can get the effect of bootstrapping the reduced data set $y_1, \ldots, y_{j-1}, y_{j+1}, \ldots, y_n$ by considering only those bootstrap samples in which y_j did not appear. The horizontal dotted lines are quantiles of $t^* - t$ for all 999 bootstrap replicates, while the solid lines join the corresponding quantiles for the subsets of bootstrap replicates in which each of the 20 observations did not appear. The x-axis shows empirical influence values l_j, which measure the effect on t of putting more mass on each of the observations separately. If $\hat{F}$ represents the empirical distribution function of the data, which puts mass n^{-1} on each of the observations $y_1, \ldots, y_n$, and $t(\hat{F})$ is the corresponding statistic, we can write

$$l_j \doteq \frac{t\{(1-\varepsilon)\hat{F} + \varepsilon 1_j\} - t(\hat{F})}{\varepsilon},$$

where 1_j puts unit mass on y_j and ε is a suitably small value: thus 1_j is the instantaneous change in t when the sample is perturbed in the direction of y_j. Although numerical differentiation could have been used, to save computing time we used the formula

$$l_j = \left\{e^T(y_j - \bar{y})\right\}^2 - t,$$

where e is the eigenvalue corresponding to t, and $\bar{y}$ is the average of the y_j. The l_j play an important role in nonparametric statistics. In particular, they provide an approximate variance for t through the expression

$$v_L = n^{-2} \sum_{j=1}^{n} l_j^2 ,$$

the bootstrap version of which was used above.

We see that when case 1, 5 or 6 is deleted, the distribution of t^* shifts to the left (l_1, l_5 and l_6 are positive) and becomes more peaked (the quantiles are closer together). The circles in the top left panel show the roughly $999 \times (1 - 2/20)^{20}$ simulations in which neither case 5 nor case 6 appears: the estimated variances are small and the values of t^* are shifted left and are less variable, as we had already surmised. The lower right panel explains this: values of t^* for samples where cases 5 and 6 do not appear will be less elliptical than those where they do.

What do we learn from this? A general lesson is that a bootstrap is a simulation study and should be treated as such. We need to think of informative displays of the output, to inspect them and to act accordingly. A particular lesson is that, for this dataset, arguments that rely heavily on smoothness assumptions, expansions, and so forth, are not trustworthy, as the statistic is overly dependent on a few observations. We would need to know more about the context of the example to say whether this can be fixed. Perhaps the authors could say more about the data in their Rejoinder.

METHOD ANALYSIS

To a mathematical statistician, there can be no such thing as a reliable confidence interval, only a reliable confidence interval method; that is, one giving intervals whose actual coverage probability is close to the nominal value. From this point of view, we want to construct a random interval $I_{1-2\alpha}$ with nominal coverage $1 - 2\alpha$ such that, when θ is the true parameter value,

$$\text{pr}(\theta \in I_{1-2\alpha}) = 1 - 2\alpha ,$$

with the probability calculation conditioned on a suitable ancillary statistic when one exists. Ancillaries usually arise from the particular parametric model being used. As pointed out in the paper, they can be difficult to identify in the nonparametric context, so we shall ignore them below. Here is a small selection from the smörgåsbord of bootstrap confidence interval methods:

- *normal intervals* $t \pm z_\alpha v^{*1/2}$, where v^* is the variance of the bootstrap replicates t^* and z_α is the α-quantile of the standard normal distribution;
- *transformed normal intervals*

  $$h^{-1}\{h(t) \pm z_\alpha v^{*1/2}\},$$

  where v^* is the variance of the bootstrap replicates $h(t^*)$, with $h(\cdot)$ the "variance-stabilizing" square root transformation;
- *basic bootstrap intervals*

  $$\left(2t - t^*_{[(R+1)(1-\alpha)]}, 2t - t^*_{[(R+1)\alpha]}\right),$$

  which are based on the assumed pivotality of $T - \theta$; here, T is the random variable of which t is the observed value;
- *transformed basic bootstrap intervals*, basic bootstrap intervals calculated on the transformed scale, then back-transformed;
- *Studentized bootstrap confidence intervals*

  $$\left(t - v_L^{*1/2} z^*_{[(R+1)(1-\alpha)]}, t - v_L^{*1/2} z^*_{[(R+1)\alpha]}\right),$$

  where $z^* = (t^* - t)/v_L^{*1/2}$ is the Studentized version of t^*, and $z^*_{(r)}$ is the rth order statistic of the simulations $z_1^*, \ldots, z_R^*$;
- *transformed Studentized bootstrap confidence intervals*, studentized bootstrap confidence intervals computed using the transformed scale, then back-transformed;
- *percentile confidence intervals*

  $$(t^*_{[(R+1)\alpha]}),\ t^*_{[(R+1)(1-\alpha)]}),$$

  based on assuming that there is a (unknown) transformation $g(\cdot)$ such that the distribution of $g(T) - g(\theta)$ is pivotal and also symmetric about zero;
- *BC_a confidence intervals*, as described in the paper;
- *ABC confidence intervals*, as described in the paper.

Our normal intervals are the standard intervals of the paper, except that we use a bootstrap estimate of variance, and our Studentized bootstrap intervals are the bootstrap-t intervals of the paper. More details of the methods above, and descriptions of other bootstrap confidence interval methods, can be found in Chapter 5 of Davison and Hinkley (1996) as well as in the paper.

Our Table 1 augments Table 3 of the paper by giving these intervals for the cd4 data, based on $R = 999$ nonparametric bootstrap simulations. We calculated the Studentized intervals using

$$v_L^* = n^{-2} \sum_{j=1}^{n} l_j^{*2} ,$$

TABLE 1
Lower and upper 90% bootstrap confidence limits (L, U) for the largest eigenvalue of the covariance matrix underlying the cd4 data, calculated on the original and the square root scale; all but the ABC method are based on 999 simulations

	Original scale		Transformed scale	
	L	U	L	U
Normal	1.00	2.35	1.06	2.44
Basic	1.07	2.41	1.16	2.62
Studentized	1.14	2.93	1.15	2.93
Percentile	0.94	2.28		
BC_a	1.18	2.64		
ABC	1.15	2.56		

where l_j^* is the jth empirical influence value for a value of t^* based on a bootstrap sample $y_1^*, \ldots, y_n^*$. When the studentized bootstrap method is numerically unstable, it is often because v_L^* is too small, but in this example v_L^* is typically slightly larger than the variance of t^* estimated using a small double bootstrap.

The BC_a interval uses the 148th and 990th ordered values of the 999 t^*, as opposed to the 50th and 950th used by the percentile interval. This is the large correction that we saw in Figure 2 of the paper, so large that a bigger simulation is needed to get a more accurate estimate of the upper limit. The BC_a interval in Table 3 of the paper has less Monte Carlo error and is very close to the endpoints 1.16 and 2.52 we obtained with 2,499 simulations. When a correction of this size is needed, the Studentized bootstrap requires a smaller simulation because it uses less extreme quantiles of the simulated z^*, in this case $z^*_{(50)}$ and $z^*_{(950)}$.

Our table shows that the more sophisticated methods—studentized, BC_a and ABC—give higher upper endpoints, and the effect of transformation is to shift intervals slightly rightward. This was also the effect of calibrating the ABC intervals, as we see from Table 3 of the paper.

There are nine intervals in our Table 1. Which is most reliable, in the sense that it results from a method whose coverage is closest to nominal? We performed a small simulation study to estimate coverages for the eigenvalue example. We generated 1,600 samples of size 20 from the bivariate normal distribution fitted to the cd4 data, and for each we used $R = 999$ bootstrap simulations to obtain the intervals described above. Table 2 shows the empirical coverages from this experiment. All the methods tend to undercover—some dramatically. No method performs very well overall, but the Studentized method works best, with two-sided coverages only slightly less than nominal. The normal, basic and percentile methods do very poorly in the upper tail: the top endpoint of these intervals is too low. Unfortunately the same is true of the BC_a and the ABC methods, which do only as well as the much simpler normal and basic bootstrap intervals on the transformed scale. The Studentized intervals do best in the upper tail, although transformation has little effect on their coverage accuracy. This is consistent with Table 1.

Figure 2 shows boxplots of the lengths of the confidence intervals. The most pronounced feature is the long intervals for the two Studentized methods, which helps to account for their better error rates. Far from being a drawback, in this problem the fact that the Studentized bootstrap method can give long confidence intervals is precisely what gives it the best coverage of the methods considered in our simulation study, and the "conservativeness" of the BC_a method is what leads it to undercover.

Other numerical studies have led us to similar conclusions: in small samples, nonparametric bootstrap methods typically undercover somewhat, and the Studentized bootstrap method can work better than the BC_a or ABC method, particularly if combined with a transformation. See Davison and Hinkley (1996, Chapter 5).

CONCLUDING COMMENTS

Any reader still with us will realize that we are uneasy about describing any confidence interval method as "automatic." Too much can go wrong: the free lunch arrives with an unidentified flying object in the soup. Data must be carefully scrutinized and simulation output checked for oddities, particularly when a nonparametric bootstrap analysis has been performed with a small sample. Simulation methods have made good confidence intervals easier to get, in the sense that fearsome mathematics need not be used, but they have not removed the need for thoughtful data analysis. A corollary of this point is that we are nervous about attempts (including our own) to replace nonparametric bootstrap simulation by analytical calculation, as in that case there are no simulation results to be inspected.

In the eigenvalue example, the nonparametric BC_a and ABC methods give intervals whose coverage is only slightly better than the much simpler normal and basic methods used with transformation. It is true that the transformation was guessed by a "trick," but the trick required just a few lines of code, in addition to the calculation of t, and in fact we could have used Monte Carlo ideas described in Chapter 9 of Davison and Hinkley (1996) to est-

TABLE 2

Empirical coverages (percent) for nonparametric bootstrap confidence limits in eigenvalue estimation; $R = 999$ for all simulation methods; 1,600 data sets generated from bivariate normal distribution; approximate standard errors for the results are also given

	Nominal coverage								
	Lower limit			Upper limit			Overall		
Method	2.5	5	10	90	95	97.5	80	90	95
Normal	0.3	1.6	5.2	76.2	81.4	85.1	71.0	79.8	84.8
transformed	0.9	2.3	6.1	78.5	84.5	88.6	72.4	82.2	87.7
Basic	0.8	2.6	7.4	77.9	82.1	84.4	70.5	79.5	83.6
transformed	2.5	5.2	10.4	82.9	87.6	90.8	72.5	82.4	88.3
Studentized	1.5	4.2	9.4	88.4	92.4	95.6	79.0	88.2	94.1
transformed	1.9	4.6	9.9	88.4	92.5	95.6	78.5	87.9	93.7
Bootstrap percentile	0.3	1.0	3.2	73.6	80.7	85.8	70.4	79.7	85.5
BC_a	2.3	5.5	10.0	83.7	88.8	91.6	73.7	83.3	89.3
ABC	2.5	5.6	10.8	83.8	88.7	91.2	73.8	83.1	88.7
Standard error	0.4	0.5	0.8	0.8	0.5	0.4	1.5	0.8	0.5

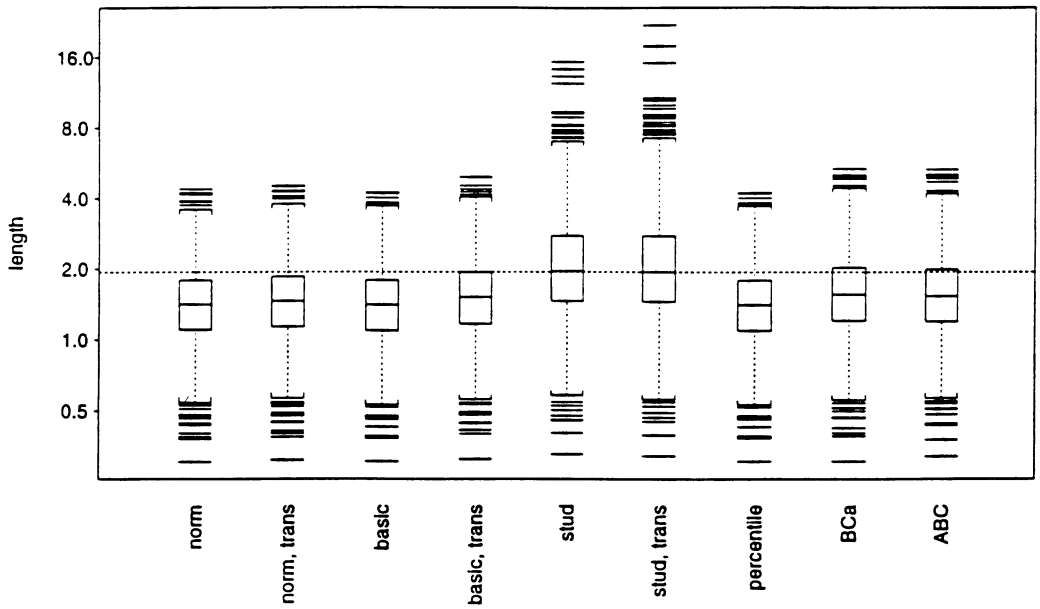

FIG. 2. *Boxplots of confidence interval lengths for the* 1,600 *simulated samples in the numerical experiment with bivariate normal data; the dotted horizontal line is at the median length of the transformed Studentized bootstrap interval; note the log scale of the vertical axis.*

mate it without calculating the estimated variances v_L^* or needing a double bootstrap.

The simple intervals share with the more accurate Studentized bootstrap the drawback that they are not scale-invariant, and of course this reduces their appeal. Choice among confidence interval methods is partly aesthetic, with some researchers insisting more strongly than others on the importance of parametrization-invariance. Our view is that in this example, the gain in coverage accuracy from using the Studentized bootstrap intervals outweighs the disadvantage that they are not invariant.

Our limited simulation underlines the unfortunate fact that the impressive theoretical analysis of confidence interval methods outlined in Sections 8 and 9 of the paper is not the whole story. In principle, the BC_a, ABC and Studentized methods are all second-order accurate, but for normal samples of size 20 the coverage of the Studentized method is better than the others by some margin. It turns out that, in practice, the ABC intervals can give a poor approximation to Studentized bootstrap intervals, and although this can be fixed by calibration, our Table 2 suggests that calibration cannot improve much on using the Studentized bootstrap,

which itself requires less effort than does calibrating an ABC interval.

While we admire the authors's efforts to find the Holy Grail of accurate, invariant, reliable confidence intervals for small-sample problems, and hope that they will continue their quest, our numerical work suggests that the end is not yet in sight.

Our comments above imply a need for methods of "post bootstrap" analysis. Some are described in Efron (1992), with a general discussion in Chapter 3 of Davison and Hinkley (1996). We have developed a library of bootstrap functions in S-PLUS which facilitates this type of analysis. The library may be obtained by anonymous ftp to markov.stats.ox.ac.uk and retrieving the file pub/canty/bootlib.sh.Z.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the U.K. Engineering and Physical Sciences Research Council through a research grant and through an Advanced Research Fellowship to the second author.

Comment

Leon Jay Gleser

The term "bootstrap confidence interval" concerns me because the use of the word "confidence" promises that a lower bound for the coverage probability of the interval is being maintained regardless of the true value of the parameter(s) or the choice of distribution within the family of distributions. In fact, the bootstrap method cannot always achieve that goal and in "routine, automatic" application may appear to apply when it actually does not.

A case in point is the problem of estimating the ratio of two means, the so-called Fieller problem. In their technical report (DiCiccio and Efron, 1995), which appears by its title to have been an earlier version of the present paper, the authors discussed using their methods for this problem in the context of the simple example of Figure 1, but apparently decided not to present this application here. Perhaps the reason for this decision is that they became aware that their bootstrap intervals cannot achieve the goal of maintaining a positive lower bound for coverage probability in this problem. A proof of this assertion is given in Gleser and Hwang (1987), where it is shown that for both ratios of means problems and a wide class of other estimation problems there does not exist an interval estimator which produces intervals with both almost surely finite length and coverage probability bounded below by a positive number. Put another way, any confidence interval procedure (such as the authors' various bootstrap procedures) that produces finite intervals with probability 1 must have minimum coverage probability *zero*!

Yet the bootstrap methods presented by DiCiccio and Efron are said to have coverage probability equal to the desired confidence level $1 - \alpha$ up to an approximation whose error goes to 0 as $n \to \infty$ irrespective of what the true value of the parameter (and the true distribution of the data) may be. This assertion is correct, but the problem is that the value N of n for which $n \geq N$ guarantees that the error of the approximation is less than a specified value ε may depend on the true value of the parameter (or the true distribution). That is, the order-in n terms displayed by the authors *are not necessarily uniform in the parameters* (or true distribution). This fact is not mentioned by the authors and is rarely discussed in the bootstrap and Edgeworth expansion literature (a notable exception being Hall and Jing, 1995), but is crucial to analytical evaluation of the applicability of the bootstrap methodology.

It is important to note that this nonuniformity problem can occur in the simplest and most innocuous of parameter estimation problems. Consider, for example, the problem where we observe i.i.d. observations from a normal distribution with unknown (but nonzero) mean μ and variance 1, and wish to estimate $1/\mu$. Using the methods in Gleser and Hwang (1987), it can be shown that any interval estimator for $1/\mu$ that takes on only finite intervals as values must have 0 confidence. It is not hard to see that the difficulty occurs because the parameter

Leon Gleser is Professor, Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania, 15260 (e-mail: gleser@vms.cis.pitt.edu).

space contains both positive and negative values of μ that are arbitrarily close to 0. Nothing in the bootstrap methodology gives warning of this problem, and thus naive users misled by the claims made for the bootstrap approach may apply this methodology to the problem of estimating $1/\mu$ in the belief that they can achieve a guaranteed coverage probability regardless of the value of μ.

Bootstrap procedures are not alone in having this problem. Almost all "automatic" large-sample interval estimation procedures advocated in the literature share similar difficulties (as was noted long ago by Jack Wolfowitz in the case of Wald's method). It is strange that students in my elementary statistics classes are quick to question "how large must n be?" in order that a certain approximate statistical method have its claimed performance (to, say, two-decimal accuracy), but this question is rarely answered (much less asked) in the literature. Granted, these are hard questions. But I would think that the advent of modern computers now would make it possible to provide useful answers to such questions.

ACCURACY AND CORRECTNESS

DiCiccio and Efron talk about the coverage accuracy and correctness of their confidence intervals. I have already discussed the coverage accuracy and how the advertised orders of magnitude of the error as functions of the sample size n overlook the fact that such errors also depend on the parameter (and perhaps also the true distribution as a whole). Nevertheless, the concept of accuracy in determining (or achieving) coverage probability is clear.

It is less clear what the authors mean by "correctness" (they themselves admit this). They appear to be talking about the closeness of the bootstrap interval endpoints to certain *ideal* confidence interval endpoints, for example, those corresponding to most accurate or smallest expected length confidence intervals for the given problem. There are, however, many such choices of ideal confidence intervals, depending upon what restrictions are placed upon the underlying distributions. The theoretical material in Section 8 of DiCiccio and Efron's paper tries to show how bootstrap methods approximate ideal exact confidence intervals based on a "mean-like" (in the sense of having cumulants of the same order in the sample size n as the sample mean) estimator in very general distributional contexts. Section 9 uses likelihood-based intervals for exponential families as the benchmark. In both cases, the asymptotic orders of accuracy are again not necessarily uniform in the parameters. Thus, although such theoretical comparisons are interesting, they do not answer the question of greatest interest to the practitioner, namely, "Is the sample size I have large enough for the bootstrap procedure to be, say, within 5% of being "correct"?"

THE FIRST LAW OF APPLIED STATISTICS

In his classic paper, Efron (1981) presented the bootstrap as a unified way of looking at many ad hoc techniques that were part of the bag of tricks of most applied statisticians. One of the insights provided by this overview was that such methods as cross-validation could be viewed as crude Monte Carlo approximations to functionals of the sample cumulative distribution function (c.d.f.). Modern computational power made it possible to replace such Monte Carlo approximations by better ones, even to the extent of being able to evaluate such functionals exactly. Later work by Efron and others, however, seems to have abandoned exact calculations in favor of Monte Carlo approximations, perhaps because the iterative nature of the bootstrap methods being studied (which placed a premium on quick computation) precluded exact calculation. The resulting emphasis on (re)sampling, and accompanying terminology, has tended to obscure the concept of the bootstrap as an evaluated functional of the sample c.d.f.

Thus, it should be noted that Monte Carlo and other resampling algorithms introduce variability that is not present in the data. Unless this extra variability is negligible, the consequence can be a violation of what, in my graduate statistics lectures, I call "the first law of applied statistics":

> Two individuals using the same statistical method on the same data should arrive at the same conclusion.

This requirement is clearly fundamental to scientific reasoning, for otherwise how can scientists check each others' work? From my reading of the literature, adherence to this law largely explains why applied statisticians almost unanimously reject randomized hypothesis testing procedures such as the Fisher–Irwin–Tocher test for 2×2 contingency tables.

Consider now the bootstrap procedures such as the ABC in which there is a series of Monte Carlo approximations to functionals of the sample c.d.f. Although each individual Monte Carlo approximation may be fairly accurate, the ensemble of such approximations can add a nontrivial amount of extraneous sampling error. Consequently, bootstrap confidence intervals formed using the same method from the same data by two different individuals can differ in a noticeable way. How much attention has been paid

to this possible problem? As bootstrap methods increase in sophistication and complexity, greater attention needs to be paid to increasing the accuracy of each Monte Carlo approximation; otherwise the greater accuracy achieved by the more sophisticated method may be undone by its greater unreliability (variation).

CONFIDENCE AND ACCURACY: A SUGGESTED APPROACH

DiCiccio and Efron seem to be attempting to create their confidence intervals by use of a pivotal approach. Such an approach appealed to R. A. Fisher because he thought it allowed him to transfer the variability of the estimate and assign it to the unknown parameter (fiducial inference). His theory floundered in part because pivotals were not always unique (something that may also be of concern for bootstrap pivotals). Neyman found pivotals useful because they were often optimal test statistics (thus leading to uniformly most accurate confidence regions) and because they simplified calculation of exact coverage probabilities. The resulting intervals answer the following question about a point estimator: "What accuracy can I obtain for a specified confidence?" In most practical contexts, the confidence interval derived from a pivotal has a random length; this length may have little relevance to the accuracy the practitioner wishes to obtain. Instead, I think most users of confidence interval methodology want to know, "Approximately how likely is it that I can achieve a specified accuracy d with my point estimator?"

Using the bootstrap methodology (specifically the bootstrap c.d.f. of an estimator), one can straightforwardly and directly estimate $P_L(d)$ and $P_U(d)$, the respective probabilities that an estimator is d units or more below and d units or more above the true value of the parameter being estimated. An estimator and its two estimated accuracy functions $P_L(d)$ and $P_U(d)$ are an extension of the (estimator, estimated loss) summary advocated by Lu and Berger (1989a, b) and others. This melding of bootstrap and decision theory should suggest to my mathematical statistical colleagues some new problems on which to try their techniques. More important, particularly because there is some hope of obtaining *uniform* (in the parameters) estimates of rates of convergence in n, it may give practitioners an applicable estimation methodology.

Comment

Stephen M. S. Lee and G. Alastair Young

This is a timely and provoking article. Recent years have seen enormous research efforts into the properties and scope of the bootstrap. While substantial attention has been paid to extending the seminal ideas of Efron (1979) to complicated and wide-ranging problems, it is the context of the paper by DiCiccio and Efron that has seen most progress in the development of practical and effective bootstrap inference.

Stephen M. S. Lee is at the Department of Statistics, University of Hong Kong, Pokfulam Road, Hong Kong. G. Alastair Young is Lecturer, Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, United Kingdom (e-mail: g.a.young@statslab.cam.ac.uk).

AN AGREED SOLUTION?

Efron and LePage (1992) remark that "the goal of automatically producing highly accurate confidence intervals seems to be moving towards a practical solution." DiCiccio and Efron offer us the solution. Their paper, while providing a beautiful general exposition of the principles behind the bootstrap solution to confidence interval construction, amounts to a strong piece of advocacy for one particular method, the ABC method. The reasons behind their view that this method constitutes the sought-for practical solution to the confidence interval problem are clear, and well-argued in their paper. The ABC method approximates the theoretically favored BC_a and bootstrap-t methods and therefore enjoys good accuracy and correctness properties, eliminates the need for Monte Carlo simulation and works well in practice, as the examples of the paper illustrate. With bootstrap calibration, even better performance can

be squeezed, and we can diagnose potential problems for the method.

But there is another solution to the problem, as sketched by Hall (1992, Section 3.11.1). Instead of using a refined bootstrap procedure such as BC_a or ABC, use bootstrap calibration directly on the crude percentile-based procedures these methods refine, and which seem currently favored in published applications of the bootstrap, as any literature search confirms. In doing so, we retain the desirable properties of these basic procedures (stability of length and endpoints, invariance under parametrization etc.) yet improve their coverage accuracy. The price is one of great computational expense, although, as is demonstrated by Lee and Young (1995), there are approximations which can bring such bootstrap iteration within the reach of even a modest computational budget. An advantage of this solution lies in its simplicity: there is no need to explain the mechanics of the method, in the way that is done for the BC_a and ABC methods in Sections 2–4 of DiCiccio and Efron's paper.

Which solution is best? To answer this requires a careful analysis of what we believe the bootstrap methodology to be. Our view is that willingness to use extensive computation to extract information from a data sample, by simulation or resampling, is quite fundamental. In other words, in comparing different methods, computational expense should not be a factor. All things being equal, we naturally look for computational efficiency, but things are hardly ever equal. How do the two solutions, that provided by DiCiccio and Efron and that involving the iterated percentile bootstrap, compare? There are two concerns here, theoretical performance and empirical performance, and the two might conflict. We demonstrate by considering the simple problem of constructing a two-sided nonparametric bootstrap confidence interval for a scalar population mean.

CALIBRATION AND COVERAGE PROPERTIES

All the common two-sided bootstrap intervals, including the percentile and ABC methods, have, for the "smooth function" model of Hall (1988), coverage error of order n^{-1}, where n is the sample size. The order of coverage error may be reduced by calibration, typically to order n^{-2}. In terms of the order of coverage error, we prefer the calibrated percentile method over the ABC method, although there is no immediate preference for the calibrated percentile interval over the calibrated ABC method.

For this context, the use of bootstrap iteration or calibration to reduce coverage error is due to Hall (1986) and Beran (1987). The calibration method of Loh (1987) corresponds to the method of Beran (1987) when applied to a bootstrap confidence interval. For the confidence interval problem the method of Hall (1986) amounts to making an additive adjustment, estimated by the bootstrap, to the endpoints of the confidence interval, while the method of Beran (1987) amounts to making an additive adjustment, again estimated by bootstrapping, to the nominal coverage level of the bootstrap interval. The method of calibration described by DiCiccio and Efron in Section 7 of their paper is a subtle variation on the latter procedure, and one which should be used with care. DiCiccio and Efron use a method in which the bootstrap is used to calibrate separately the nominal levels of the lower and upper limits of the interval, rather than the overall nominal level.

Theoretical and empirical evidence which we shall present elsewhere leads to the conclusion that, all things being taken into consideration, preference should be shown to methods which adjust nominal coverage, rather than the interval endpoints. We shall therefore focus on the question of how to calibrate the nominal coverage of a bootstrap confidence interval.

The major difference between the two approaches to adjusting nominal coverage is that the method as illustrated by DiCiccio and Efron is only effective in reducing coverage error of the two-sided interval to order n^{-2} when the one-sided coverage-corrected interval achieves a coverage error of order $n^{-3/2}$, as is the case with the ABC interval, but not the percentile interval. The effect of bootstrap calibration on the coverage error of one-sided intervals is discussed by Hall and Martin (1988) and by Martin (1990), who show that bootstrap coverage correction produces improvements in coverage accuracy of order $n^{-1/2}$, therefore reducing coverage error from order $n^{-1/2}$ to order n^{-1} for percentile intervals, but from order n^{-1} to order $n^{-3/2}$ for the ABC interval. If the one-sided corrected interval has coverage error of order $n^{-3/2}$, then separate correction of the upper and lower limits gives a two-sided interval with coverage error of order n^{-2}, due to the fact that the order $n^{-3/2}$ term involves an even polynomial. With the percentile interval, the coverage error, of order n^{-1}, of the coverage-corrected one-sided interval typically involves an odd polynomial, and terms of that order will not cancel when determining the coverage error of the two-sided interval, which remains of order n^{-1}. On the face of it, therefore, we should be wary of the calibration method described by DiCiccio and Efron, although the problems with it do not arise with the ABC interval.

A CLOSER EXAMINATION

The above discussion is phrased in terms of the magnitude of coverage error. Lee and Young (1996b) describe techniques by which we may obtain explicitly the leading term in an asymptotic expansion of the coverage error of a general confidence limit procedure: see also Martin (1990). Application of these methods to the intervals under consideration here allows closer examination of coverage error.

Table 1 gives information on the theoretical leading terms in expansions of the coverage error of the percentile interval (denoted I_P), iterated percentile interval (denoted I_{PITa} and I_{PITb}), ABC interval (denoted I_{ABC}) and iterated ABC interval (denoted by I_{ABCITa} and I_{ABCITb}). Figures refer to two-sided intervals of nominal coverage 90% and are shown for the two methods of nominal coverage calibration, for each of four underlying distributions. The intervals I_{PITa} and I_{ABCITa} calibrate the overall nominal coverage, while the other two iterated intervals use calibration in the way discussed by DiCiccio and Efron.

What is immediately obvious from the table is that the order of coverage error only tells part of the story. Compare the coefficients of n^{-1} for the interval I_{PITb} with the coefficients of n^{-2} for the other iterated intervals.

However, if we focus on those intervals that ensure a coverage error of order n^{-2}, it appears that the two types of iterated ABC interval are not significantly different, but that the iterated percentile interval has a leading error term consistently and significantly smaller than that of the ABC method. This same conclusion is true for any nominal coverage in the range 0.9–0.99.

THEORY AND PRACTICE

Theory and practice are two different things. Table 1 also reports a simulation by which we estimated the coverage probabilities of the various intervals, using 1,600 random samples of sizes $n =$ 15 and 30 drawn from each of the four distributions. The intervals I_P were each constructed from 1,000 (outer level) bootstrap samples. Each of the iterated intervals was calibrated by drawing 1,000 (inner level) bootstrap samples.

The simulation confirms clearly the advantages of calibration on coverage error. Without calibration the ABC method may have substantial coverage error and might be little better than the crude percentile method. Equally, however, the simulation demonstrates the theory to have only a strictly qualitative value in predicting the reduction in error obtained by calibration.

Considering the percentile intervals, we see little practical difference in coverage for the two calibration methods, although separate calibration of the upper and lower limits is strikingly more effective with a lognormal parent population. For the ABC limits, calibration of the overall nominal coverage seems distinctly preferable, contrary to the asymptotic conclusion. On the other hand, the empirical findings do match the theoretical conclusion that iterated percentile intervals are to be preferred over the calibrated ABC intervals.

CONCLUSIONS

A theoretical comparison of the coverage properties of bootstrap confidence intervals points strongly toward the use of calibration methods to reduce coverage error, in terms of a reduction in the order of coverage error. Closer inspection of the theory demonstrates that we should be careful in how we apply the notion of calibration and alerts us to the possibility that solution of the problem of producing bootstrap confidence intervals of low coverage error may require more than consideration of the theory. We should especially welcome therefore a paper such as that by DiCiccio and Efron, where the focus is not on the general properties of the methods, but rather on the behavior of the methods in particular well-chosen examples.

Rejoinder

Thomas J. DiCiccio and Bradley Efron

If the standard intervals were invented today, they might not be publishable. Simulation studies would show that they perform poorly in problems like those in our paper. In fact the standard intervals are immensely useful, and accurate enough to have been happily used by scientists on literally mil-

TABLE 1
Estimated coverage probabilities for mean, based on 1,600 *random samples of sizes* $n = 15$ *and* 30 *drawn from each of four different distributions and theoretical leading terms in expansion of coverage error*

	Interval					
	I_P	I_{PITa}	I_{PITb}	I_{ABC}	I_{ABCITa}	I_{ABCITb}
			Normal data $N(0,1)$			
$n = 15$	0.860	0.897	0.891	0.862	0.889	0.877
$n = 30$	0.892	0.902	0.899	0.893	0.902	0.892
Error	$-0.48395n^{-1}$	$-0.5410n^{-2}$	$0n^{-1}$	$-0.48395n^{-1}$	$-4.7845n^{-2}$	$-4.7845n^{-2}$
			Folded normal data $\lvert N(0,1)\rvert$			
$n = 15$	0.839	0.883	0.909	0.861	0.878	0.869
$n = 30$	0.869	0.888	0.910	0.875	0.888	0.882
Error	$-0.59605n^{-1}$	$-2.8452n^{-2}$	$0.084197n^{-1}$	$-0.38375n^{-}1$	$-7.7742n^{-2}$	$-5.4852n^{-2}$
			Negative exponential data $\exp(1)$			
$n = 15$	0.819	0.874	0.874	0.826	0.869	0.829
$n = 30$	0.876	0.901	0.900	0.875	0.898	0.889
Error	$-1.2079n^{-1}$	$-40.336n^{-2}$	$0n^{-1}$	$-1.0028n^{-1}$	$-99.900n^{-2}$	$-99.900n^{-2}$
			Log normal data $\exp(N(0,1))$			
$n = 15$	0.765	0.829	0.875	0.778	0.837	0.750
$n = 30$	0.815	0.853	0.889	0.819	0.858	0.824
Error	$-13.241n^{-1}$	$-132844n^{-2}$	$-14.251n^{-1}$	$-25.308n^{-1}$	$-665027n^{-2}$	$-805445n^{-2}$

lions of real problems. Statistical methods have to be judged by their competition, and until recently there has been no competition to the standard intervals for most situations that arise in practice.

Modern statistical theory combined with modern computation now allows us to improve upon the standard intervals, and to do so in a routine way that is suitable for day-to-day statistical applications. Our paper discusses several bootstrap-based methods for doing so. The BC_a and ABC methods are featured in the paper, mainly because their development shows clearly just what it is about the standard intervals that needs improvement. There is also the practical point that the BC_a and ABC methods *consistently* improve upon the standard intervals, although not always in dramatic fashion. Our particular focus here on the ABC intervals has a lot to do with their computational simplicity. The discussion of calibration in Section 7 involved a lot of computation, but it would have been immensly more if we had tried to calibrate the BC_a or bootstrap-t intevals.

So how well does the ABC method perform? Better than suggested by the commentaries, at least for smoothly continuous statistics like means, correlation and eigenvalues. Here is a closer look at Lee and Young's last example. We observe a random sample of size $n = 30$ from a normal distribution with unknown expectation and variance,

$$x_1, x_2, \ldots, x_{30} \sim_{\text{i.i.d.}} N(\lambda, \Gamma), \tag{1}$$

and wish to form confidence intervals for the parameter

$$\theta = \lambda + 5 \cdot \Gamma \tag{2}$$

or equivalently for

$$\gamma = e^{\theta}; \tag{3}$$

γ is the expectation of the lognormal variate $\exp\{X\}$, $X \sim N(\lambda, \Gamma)$. "Equivalently" in the previous sentence applies to the ABC method, which is transformation invariant, but not to the standard method, which will have different coverage probabilities for θ and γ.

The top half of Table 1 shows the results of 2,000 Monte Carlo simulations: situation (1) was replicated 2,000 times, with $\gamma = 0$, and $\Gamma = 1$, so $\theta = 1/2$. The parametric ABC and standard confidence interval endpoints $\hat{\theta}_{\text{ABC}}[\alpha]$ and $\hat{\theta}_{\text{STAN}}[\alpha]$ were computed for each simulation, as in Section 4, for various values of α. Also computed was $\hat{\gamma}_{\text{STAN}}[\alpha]$, the standard interval endpoint for γ. The table shows the actual coverage proportions in the 2,000 simulations, so, for example, 0.931 of the simulations had $\theta < \hat{\theta}_{\text{ABC}}[0.95]$. Also shown is the central 90% two-sided coverage, the proportion of simulations with $\hat{\theta}[0.05] < \theta < \hat{\theta}[0.95]$.

TABLE 1

Empirical coverage probabilities of the ABC and standard intervals $(-\infty, \hat{\theta}[\alpha])$ *for the lognormal expectation problem (lines* 1–3*); line* 3 *concerns* $\hat{\gamma}_{\mathrm{STAN}}[\alpha]$*, the standard interval applied to* γ *instead of* θ*;* 2,000 *Monte Carlo simulations for lines* 1–3*,* 1,000 *for lines* 4–5

	α								
Method	0.025	0.05	0.1	0.16	0.84	0.9	0.95	0.975	Central 0.90
1. parametric ABC	0.033	0.062	0.106	0.165	0.827	0.884	0.931	0.960	0.869
2. parametric standard	0.014	0.030	0.087	0.144	0.795	0.848	0.906	0.934	0.876
3. parametric standard γ	0.000	0.001	0.030	0.090	0.769	0.826	0.893	0.923	0.892
4. nonparametric ABC	0.028	0.070	0.118	0.172	0.809	0.863	0.916	0.943	0.846
5. nonparametric standard	0.019	0.043	0.102	0.150	0.780	0.844	0.901	0.921	0.858

Looking just at the two-sided 0.90 coverage probabilities, the clear winner is the parametric standard method applied to γ. It has empirical coverage 0.892 (for γ, or for θ taking the logs of the γ endpoints), nearly equal to the target value 0.90, compared to 0.869 for the parametric ABC intervals and 0.876 for the parametric standard method applied on the θ scale.

Wrong! In fact the standard method applied to γ performs dreadfully: $\hat{\gamma}_{\mathrm{STAN}}[0.05]$ was less than γ only 0.001 of the time, while γ exceeded $\hat{\gamma}_{\mathrm{STAN}}[0.95]$ in 0.107 of the cases. This kind of behavior, although not usually to this degree, is typical of the standard intervals, a too-liberal result at one end being balanced by a too-conservative result at the other. At the minimum, simulation studies must report a range of coverage probabilities, as in Table 1, and not just the central coverage. Hall and Martin make this point nicely in their "whither" comments.

However coverage probabilities by themselves are not enough. This is where the difficult notion of correctness comes in. Suppose that in situation (1) we desired a confidence interval for the expectation λ. The Student's-t endpoints based on the first 15 observations would be perfectly accurate, giving exactly the right coverage probabilities, but they would be inferentially incorrect. In this situation there is a correct answer, the Student's-t endpoints based on all 30 observations. We would expect a good approximate confidence interval method to track the correct endpoints closely, as well as having good coverage properties. The trouble is that in most situations, including the lognormal problem, we do not have a correct confidence method to use as a gold standard.

Section 8 follows Hall's way around this problem: an idealized Student's-t endpoint,

$$\hat{\theta}_{\mathrm{exact}}[\alpha] = \hat{\theta} - \hat{\sigma}K^{-1}(1-\alpha), \tag{4}$$

serves as the gold standard, where K is the c.d.f. of the t-like variable $(\hat{\theta} - \theta)/\hat{\sigma}$. The $\hat{\theta}_{\mathrm{exact}}[\alpha]$ endpoints cannot be used in practice, because we will not know K, but we can use them as reference points in a simulation study, where K can always be found by Monte Carlo. Section 8 shows that all of the second-order accurate methods agree to second order with $\hat{\theta}_{\mathrm{exact}}[\alpha]$, implying that $\hat{\theta}_{\mathrm{exact}}[\alpha]$ is a reasonable target for correct performance. The name "exact" is appropriate because (4) gives exactly the right coverage probability for every choice of α.

Table 2 applies this comparison to the parametric ABC and standard endpoints for $\theta = \lambda + 0.5 \cdot \Gamma$. The table shows the 0.05 and 0.95 endpoints for the first 7 of 100 simulations of (1), $(\lambda, \Gamma) = (0, 1)$. Notice that $\hat{\theta}_{\mathrm{ABC}}[\alpha]$ is always closer than $\hat{\theta}_{\mathrm{STAN}}[\alpha]$ to the gold standard value $\hat{\theta}_{\mathrm{exact}}[\alpha]$. This was true in all 100 simulations. Table 3 summarizes the endpoint differences $\hat{\theta}_{\mathrm{ABC}}[\alpha] - \hat{\theta}_{\mathrm{exact}}[\alpha]$ and $\hat{\theta}_{\mathrm{STAN}}[\alpha] - \hat{\theta}_{\mathrm{exact}}[\alpha]$ for all 100 simulations. We see that $\hat{\theta}_{\mathrm{ABC}}[\alpha]$ is almost an order magnitude better than $\hat{\theta}_{\mathrm{STAN}}[\alpha]$ at tracking the exact endpoints.

In other words, the ABC method gives a substantial and consistent improvement over the standard intervals. The same thing happens using the nonparametric ABC and standard intervals, although both methods are less accurate than they were parametrically.

In the authors' experience, the BC_a and ABC methods reliably improve upon the standard inter-

TABLE 2

Comparison of exact, ABC and standard parametric endpoints for θ*; first* 7 *of* 100 *simulations; the ABC endpoints are always closer to the exact endpoints*

$\alpha = 0.05$			$\alpha = 0.95$		
Exact	ABC	Standard	Exact	ABC	Standard
0.49	0.49	0.44	1.26	1.23	1.16
0.41	0.42	0.34	1.39	1.37	1.26
0.23	0.24	0.18	1.03	1.01	0.93
0.08	0.08	0.03	0.84	0.82	0.74
0.08	0.08	0.05	0.65	0.62	0.58
−0.02	−0.02	−0.06	0.54	0.51	0.47
0.20	0.20	0.16	0.87	0.84	0.78

TABLE 3
Summary statistics for $\hat{\theta}_{\mathrm{ABC}}[\alpha]-\hat{\theta}_{\mathrm{exact}}[\alpha]$ *and* $\hat{\theta}_{\mathrm{STAN}}[\alpha]-\hat{\theta}_{\mathrm{exact}}[\alpha]$; 100 *simulations of situation* (1)*, parametric methods*

	Difference			
	$\alpha = 0.05$		$\alpha = 0.95$	
	ABC	Standard	ABC	Standard
Mean	0.0059	−0.0509	−0.0252	−0.0989
Std. dev.	0.0062	0.0100	0.0070	0.0184

vals. That is why they were featured in our paper. They tend to be rather cautious improvements, sometimes not improving *enough* on the standard intervals. This is the case for the nonparametric upper limit in the maximum eigenvalue problem, Table 3 of the paper. (We disagree with Canty, Davison and Hinkley here: calibration is quite likely to improve the upper ABC endpoint substantially, as strongly suggested by the right panel of Figure 6.)

None of this is to say that the BC_a and ABC methods are the last word in approximate confidence intervals. This is a hot research area in both the bootstrap and the likelihood literatures. All four commentaries (and the paper) include interesting suggestions for doing better. Further improvements are likely to involve a deeper understanding of the confidence interval problem as well as better practical methods. From a theoretical point of view, estimates and hypothesis tests are much better understood than confidence intervals. There is no equivalent to the Cramér–Rao lower bound or the Neyman–Pearson lemma for confidence limits, but the methodological progress reported in our paper may foretell a theoretical breakthrough.

We note some specific points:

- The ABC intervals satisfy Gleser's "first law of applied statistics." In theory so do the BC_a intervals, and the other bootstrap methods, but in practice "ideal bootstrap" definitions like (2.3) have to be approximated by Monte Carlo calculations. The recommended value $B = 2{,}000$ for the number of bootstrap replications, based on simple binomial calculations, is sufficient to make the Monte Carlo error small relative to the underlying sampling error in most situations. Permutation tests, multiple imputation, the Gibbs sampler, and so forth also fail the first law of applied statistics, for the same reason as the bootstrap.
- Some practitioners are troubled by the failure of resampling methods to satisfy Gleser's first law (although comparing the Monte Carlo error in the bootstrap to randomized hypothesis tests does seem extreme). Consequently, higher-order methods that avoid simulation, such as the ABC for one-sided limits and the methods of Lee and Young for two-sided intervals, might be especially easy to market. Perhaps any shortcomings in their coverage accuracies would be offset in practice by their speed and widespread acceptability.
- Gleser's concerns about uniformity are certainly justified. A practical statement of this concern is "how accurate are my confidence interval coverages for my particular statistic and sample size?" The calibration methods of Section 7 provide at least a partial answer. Insisting on uniformity means you will never get an approximate confidence interval for some important problems, for example, the nonparametric estimation of an expectation.
- In the lognormal problem, the theory of similar tests applies, and Jensen (1986) has shown that the confidence limits obtained from the bias-adjusted signed root of the likelihood ratio statistic are second-order correct with respect to limits given by this theory. Consequently, for this problem, the ABC intervals are also second-order correct from the "similar test" point of view, partially answering Gleser's concerns.
- Two of the commentaries, by Hall and Martin and by Lee and Young, recommend a double bootstrap method that starts from the crude percentile method. This is the kind of suggestion that might turn out to be important in practice. The equivalent of Table 2 above, comparing the double bootstrap with the ABC, for example, would be most interesting. It would be particularly nice to see how well Lee and Young's intriguing "leading terms" (done one-sided) predict small-sample behavior for the various methods.
- In fact it is difficult to run a good simulation study of confidence intervals methods. Besides the pitfalls mentioned earlier, and the cruel computational burden, there is the question of interval length variability. One way to get better coverage accuracy is to make your intervals longer and more variable. As an extreme example, one could choose U uniform on $(0, 1)$ and define

$$\hat{\theta}[\alpha] = \begin{cases} \infty, & \text{if } U \le \alpha, \\ -\infty, & \text{if } U > \alpha. \end{cases}$$

Then the interval $(-\infty, \hat{\theta}[\alpha])$ would cover the true θ (or any other value) with probability α.

In Canty, Davison and Hinkley's simulation the Studentized intervals are longer and more variable than the others, raising some question about their better coverage values at the upper limit. This concern is really another question about correctness. The classical criterion of minimum coverage for untrue θ values weeds out silly examples like the one above, but seems hard to apply in general situations.

- Apart from the construction of confidence limits, one contribution of the ABC method is to identify the quantities $(\hat{a}, \hat{z}_0, \hat{c}_q)$, which are important in all second-order methods. For confidence interval procedures, Hall and Martin advocate the incorporation of information about the asymmetry of intervals based on skewness of bootstrap distributions. Indeed, according to expression (8.10), the asymmetry of the second-order correct intervals can be measured by the quantity in square brackets, $z_0+(2a+c_q)\{z^{(\alpha)}\}^2$. By the formula preceding (8.1), the skewness of the Studentized pivot is $-6(2a+c_q)+O(n^{-1})$, so the ABC method already offers such skewness information.
- Graphical analysis of a bootstrap simulation, even just printing out the bootstrap histogram, can be quite informative, as Canty, Davison and Hinkley show. Hall's "confidence pictures" are another nice device, being basically a fiducial description of the bootstrap-t inferences.
- Hall and Martin mention nonparametric likelihood. The theory of Section 9 extends immediately to the nonparametric framework. The basic property of likelihood needed in Section 9 is that the Bartlett identities are satisfied. Empirical likelihood and other versions of nonparametric likelihood do not satisfy the Bartlett identities exactly, but they do so to a sufficiently high order of accuracy for all the same arguments to go through. Such extensions of the theory were indicated by Efron (1993), and we are currently examining them more fully.
- Gleser notes that a potential problem for a theory of confidence intervals based on pivots is that pivotal quantities are not unique. The impact of this nonuniqueness is shown in the numerical results of Canty, Davison and Hinkley, who demonstrate that the performance of the bootstrap-t is substantially affected by the choice of parameterization for its implementation. Canty, Davison and Hinkley, in a long tradition, choose a variance-stabilizing reparameterization of the eigenvalue problem. However, in the parametric context, other authors (DiCiccio, 1984) have advocated the use of reparameterizations that reduce the skewness of the Studentized pivot. This approach would be consistent with the view of Hall and Martin, who suggest that the success of the bootstrap-t "is based on bootstrapping a quantity whose distribution depends very little on unknowns." Thus, the skewness expression $-6(2a + c_q)$ could be useful as a diagnostic for establishing appropriate parameterizations for the bootstrap-t. We are currently investigating this use of the ABC quantities.
- In line with Gleser's comments concerning nonuniqueness of confidence interval procedures, a goal of the paper was to show that many of the second-order accurate methods currently available, even likelihood-based and Bayesian ones, are somewhat similar. The numerical results of Canty, Davison and Hinkley and of Lee and Young show emphatically that there are appreciable higher-order differences between the methods. We are currently working on third-order procedures.
- Our paper features smooth statistics like correlations and eigenvalues, for which the ABC method tends to agree well with the BC_a, its parent method. The ABC method might not have looked so good if we had investigated rougher statistics like coefficients in a robust regression. As far as "automatic" usage is concerned, the BC_a intervals are easier for the statistician, if not for the computer. In nonparametric situations ABC requires an expression of the statistic $\hat{\theta}$ as a function of the bootstrap weights on the data points $x_1, x_2, \ldots, x_n$. This usually is not very hard to do, but it can be annoying. The BC_a method proceeds directly from the original definition of $\hat{\theta}$ as a function of the data $\mathbf{x}$ [using definition (6.7) to compute $\hat{a}$].

Are bootstrap confidence intervals ready for the prime time? If the question is one of always giving highly accurate coverage probabilities in small samples, the answer is no. But this would mean letting the perfect be the enemy of the possible. A more relevant question is whether we can reliably improve upon the standard intervals, and there the answer is yes.

ADDITIONAL REFERENCES

DANIELS, H. E. and YOUNG, G. A. (1991). Saddlepoint approximation for the Studentized mean, with an application to the bootstrap. *Biometrika* **78** 169–179.

DAVISON, A. C. and HINKLEY, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika* **75** 417–431.

DAVISON, A. C. and HINKLEY, D. V. (1996). *Bootstrap Methods and Their Application.* Cambridge Univ. Press.

DAVISON, A. C., HINKLEY, D. V. and WORTON, B. J. (1992). Bootstrap likelihoods. *Biometrika* **79** 113–130.

DICICCIO, T. J., MARTIN, M. A. and YOUNG, G. A. (1992). Fast and accurate approximate double bootstrap confidence intervals. *Biometrika* **79** 285–295.

DICICCIO, T. J., MARTIN, M. A. and YOUNG, G. A. (1993). Analytical approximations for iterated bootstrap confidence intervals. *Statistics and Computing* **2** 161–171.

EFRON, B. (1992). Jackknife-after-bootstrap standard errors and influence functions (with discussion). *J. Roy Statist. Soc. Ser. B* **54** 83–127.

EFRON, B. and LEPAGE, R. (1992). Introduction to bootstrap. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 3–10. Wiley, New York.

GLESER, L. J. and HWANG, J. T. (1987). The nonexistence of $100(1 - \alpha)$ percent confidence sets of finite expected diameter in errors-in-variables and related models. *Ann. Statist.* **15** 1351–1362.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

HALL, P. and JING, B-Y. (1995). Uniform coverage bounds for confidence intervals and Berry–Esseen theorems for Edgeworth expansion. *Ann. of Statist.* **23** 363–375.

JENSEN, J. L. (1986). Similar tests and the standardized log likelihood ratio statistic. *Biometrika* **73** 567–572.

LEE, S. M. S. and YOUNG, G. A. (1995). Asymptotic iterated bootstrap confidence intervals. *Ann. Statist.* **23** 1301–1330.

LEE, S. M. S. and YOUNG, G. A. (1996a). Sequential iterated bootstrap confidence intervals. *J. Roy. Statist. Soc. Ser. B* **58** 235–252.

LEE, S. M. S. and YOUNG, G. A. (1996b). Asymptotics and resampling methods. *Computing Science and Statistics*. To appear.

LU, K. L. and BERGER, J. O. (1989a). Estimation of normal means: frequentist estimation of loss. *Ann. Statist.* **17** 890–906.

LU, K. L. and BERGER, J. O. (1989b). Estimated confidence for multivariate normal mean confidence set. *J. Statist. Plann. Inference* **23** 1–20.

MARTIN, M. A. (1990). On bootstrap iteration for coverage correction in confidence intervals. *J. Amer. Statist. Assoc.* **85** 1105–1118.

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

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