



# Better Bootstrap Confidence Intervals

BRADLEY EFRON\*

We consider the problem of setting approximate confidence intervals for a single parameter  $\theta$  in a multiparameter family. The standard approximate intervals based on maximum likelihood theory,  $\hat{\theta} \pm \hat{\sigma}z^{(\alpha)}$ , can be quite misleading. In practice, tricks based on transformations, bias corrections, and so forth, are often used to improve their accuracy. The bootstrap confidence intervals discussed in this article automatically incorporate such tricks without requiring the statistician to think them through for each new application, at the price of a considerable increase in computational effort. The new intervals incorporate an improvement over previously suggested methods, which results in second-order correctness in a wide variety of problems. In addition to parametric families, bootstrap intervals are also developed for nonparametric situations.

KEY WORDS: Resampling methods; Approximate confidence intervals; Transformations; Nonparametric intervals; Second-order theory; Skewness corrections.

## 1. INTRODUCTION

This article concerns setting approximate confidence intervals for a real-valued parameter  $\theta$  in a multiparameter family. The nonparametric case, where the number of nuisance parameters is infinite, is also considered. The word "approximate" is important, because in only a few special situations can exact confidence intervals be constructed. Table 1 shows one such situation: the data  $(y_1, y_2)$  are bivariate normal with unknown mean vector  $(\eta_1, \eta_2)$ , covariance matrix =  $\mathbf{I}$  the identity; the parameters of interest are  $\theta = \eta_2/\eta_1$  and, in addition,  $\xi = 1/\theta$ . Fieller's construction (1954) gives central 90% interval (5% error in each tail) of  $[\cdot29, \cdot76]$  for  $\theta$ , having observed  $\mathbf{y} = (8, 4)$ . The corresponding interval for  $\xi = 1/\theta$  is the obvious mapping  $\xi \in [1/\cdot76, 1/\cdot29]$ .

Table 1 also shows the standard approximate intervals

$$\theta \in [\hat{\theta} + \hat{\sigma}z^{(\alpha)}, \hat{\theta} + \hat{\sigma}z^{(1-\alpha)}], \quad (1.1)$$

where  $\hat{\theta}$  is the maximum likelihood estimate (MLE) of  $\theta$ ,  $\hat{\sigma}$  is an estimate of its standard deviation, often based on derivatives of the log-likelihood function, and  $z^{(\alpha)}$  is the  $100 \cdot \alpha$  percentile point of a standard normal variate. In Table 1,  $\alpha = .05$  and  $z^{(\alpha)} = -z^{(1-\alpha)} = -1.645$ .

The standard intervals (1.1) are extremely useful in statistical practice because they can be applied in an automatic way to almost any parametric situation. However, they can be far from perfect, as the results for  $\xi$  show. Not only is the standard interval for  $\xi$  quite different from the exact interval, it is not even the obvious transformation  $[1/\cdot73, 1/\cdot27]$  of the standard interval for  $\theta$ .

Approximate confidence intervals based on bootstrap computations were introduced by Efron (1981, 1982a). Like the standard intervals, these can be applied automatically to almost any situation, though at greater computational expense than (1.1). Unlike (1.1), the bootstrap intervals transform correctly, so the interval for  $\xi = 1/\theta$

in the Fieller example is obtained by inverting the endpoints of the interval for  $\theta$ . They also tend to be more accurate than the standard intervals. In the situation of Table 1 the bootstrap intervals agree with the exact intervals to three decimal places. Efron (1985) showed that this is no accident; there is a wide class of problems for which the bootstrap intervals are an order of magnitude more accurate than the standard intervals.

In those problems where exact confidence limits exist the endpoints are typically of the form

$$\hat{\theta} + \hat{\sigma}(z^{(\alpha)} + A_n^{(\alpha)}/\sqrt{n} + B_n^{(\alpha)}/n + \dots), \quad (1.2)$$

where  $n$  is the sample size (see Efron 1985). The standard intervals (1.1) are *first-order correct* in the sense that the term  $\hat{\theta} + \hat{\sigma}z^{(\alpha)}$  asymptotically dominates (1.2). However, the second-order term  $\hat{\sigma}A_n^{(\alpha)}/\sqrt{n}$  can have a major effect in small-sample situations. It is this term that causes the asymmetry of the exact intervals about the MLE as illustrated in Table 1. As a point of comparison the Student- $t$  effect is of third-order magnitude, comparable with  $\hat{\sigma}B_n^{(\alpha)}/n$  in (1.2). The bootstrap method described in Efron (1985) was shown to be *second-order correct* in a certain class of problems, automatically producing intervals of correct second-order asymptotic form  $\hat{\theta} + \hat{\sigma}(z^{(\alpha)} + A_n^{(\alpha)}/\sqrt{n} + \dots)$ .

This article describes an improved bootstrap method that is second-order correct in a wider class of problems. This wider class includes all of the familiar parametric examples where there are no nuisance parameters and where the data have been reduced to a one-dimensional summary statistic, with asymptotic properties of the usual MLE form (see Sec. 5).

Several authors have developed higher-order correct approximate confidence intervals based on Edgeworth expansions (Abramovitch and Singh 1985; Beran 1984a,b; Hall 1983; Withers 1983), sometimes using bootstrap methods to reduce the theoretical computations. There is a close theoretical relationship between this line of work and the current article (see, e.g., Remark G, Sec. 11). However, the details of the various methods are considerably different, and they can give quite different numerical results. An important point, which will probably have to be settled by extensive simulations, is which method, if any, handles best the practical problems of day-to-day applied statistics.

## 2. OVERVIEW

The standard interval (1.1) is based on taking literally the asymptotic normal approximation

$$(\hat{\theta} - \theta)/\hat{\sigma} \sim N(0, 1), \quad (2.1)$$

\* Bradley Efron is Professor of Statistics and Biostatistics, Department of Statistics, Stanford University, Stanford, CA 94305. The author is grateful to Robert Tibshirani, Timothy Hesterberg, and John Tukey for several useful discussions, suggestions, and references.

Table 1. Central 90% Confidence Intervals for  $\theta = \eta_2/\eta_1$  and  $\zeta = 1/\theta$ , Having Observed  $(y_1, y_2) = (8, 4)$  From a Bivariate Normal Distribution  $\mathbf{y} \sim N_2(\boldsymbol{\eta}, \mathbf{I})$

	For $\theta$	(R/L)	For $\zeta$	(R/L)
Exact interval (also bootstrap)	[.29, .76]	(1.21)	[1.32, 3.50]	(2.20)
Standard approximation (1.1)	[.27, .73]	(1.00)	[1.08, 2.92]	(1.00)
MLE	$\hat{\theta} = .5$		$\hat{\zeta} = 2$	

NOTE: The exact intervals are based on Fieller's construction. R/L is the ratio of the right side of the interval, measured from the MLE, to the left side. The exact intervals are markedly asymmetric. The approximate bootstrap intervals of Efron (1982a) agree with the exact intervals to three decimal places in this case.

with the estimated standard error  $\hat{\sigma}$  considered to be a fixed constant. In certain cases it is well known that both convergence to normality and constancy of  $\sigma$  can be dramatically improved by considering instead of  $\hat{\theta}$  and  $\theta$  a monotone transformation  $\hat{\phi} = g(\hat{\theta})$  and  $\phi = g(\theta)$ . The classic example is that of the correlation coefficient from a bivariate normal sample, for which Fisher's inverse hyperbolic tangent transformation works beautifully (see Efron 1982b).

The bias-corrected bootstrap intervals previously introduced by Efron (1981, 1982a), called the BC intervals, assume that normality and constant standard error can be achieved by some transformation  $\hat{\phi} = g(\hat{\theta})$ ,  $\phi = g(\theta)$ , say

$$(\hat{\phi} - \phi)/\tau \sim N(-z_0, 1), \tag{2.2}$$

$\tau$  being the constant standard error of  $\hat{\phi}$ . Allowing the bias constant  $z_0$  in (2.2) considerably improves the approximation in many cases, including that of the normal correlation coefficient. Taking (2.2) literally gives the obvious confidence interval  $(\hat{\phi} + \tau z_0) \pm \tau z^{(\alpha)}$  for  $\phi$ , which can then be converted back to a confidence interval for  $\theta$  by the inverse transformation  $\theta = g^{-1}(\phi)$ . The advantage of the BC method is that all of this is done automatically from bootstrap calculations, without requiring the statistician to know the correct transformation  $g$ .

The improved bootstrap method introduced in this article, called  $BC_a$ , makes one further generalization on (2.1): it is assumed that for some monotone transformation  $g$ , some bias constant  $z_0$ , and some "acceleration constant"  $a$ , the transformation  $\hat{\phi} = g(\hat{\theta})$ ,  $\phi = g(\theta)$  results in

$$(\hat{\phi} - \phi)/\tau \sim N(-z_0\sigma_\phi, \sigma_\phi^2), \quad \sigma_\phi = 1 + a\phi. \tag{2.3}$$

Notice that (2.2) is the special case of (2.3), with  $a = 0$ .

Given (2.3), it is not difficult to find the correct confidence interval for  $\phi$  and then convert it back to an interval for  $\theta$  by  $\theta = g^{-1}(\phi)$ . The  $BC_a$  method produces this interval for  $\theta$  automatically, without requiring any knowledge of the transformation to form (2.3). This is the gist of Lemma 1 in Section 3.

The difference between (2.2) and (2.3) is greater than it seems. The hypothesized ideal transformation  $g$  leading to (2.2) must be both *normalizing* and *variance stabilizing*, whereas in (2.3)  $g$  need be only normalizing. Efron (1982b) shows that normalization and stabilization are partially antagonistic goals in familiar families such as the Poisson and the binomial. Schenker's counterexample to the BC method (1985), which helped motivate this article, is based

on a family (discussed in Sec. 3) for which (2.2) fails. The main purpose of this article, to produce automatically intervals that are second-order correct, generally requires assumption (2.3) rather than (2.2).

It is not surprising that a theory based on (2.3) is usually more accurate than a theory based on (2.1). In fact, applied statisticians make frequent use of devices like those in (2.3), transformations, bias corrections, and even acceleration adjustments, to improve the performance of the standard intervals. The advantage of the  $BC_a$  method is that it automates these improvements, so the statistician does not have to think them through anew for each new application.

The bootstrap was originally introduced as a nonparametric Monte Carlo device for estimating standard errors. The basic idea, however, can be applied to any statistical problem, including parametric ones, and does not necessarily require Monte Carlo simulations. We will begin our discussion of the  $BC_a$  method by considering the simplest type of parametric problem: where the data consists only of a single real-valued statistic  $\hat{\theta}$  in a one-parameter family of densities  $f_\theta(\hat{\theta})$ , say  $\hat{\theta} \sim f_\theta$ , and where we want a confidence interval for  $\theta$  based on  $\hat{\theta}$ .

Sections 3, 4, and 5 describe the  $BC_a$  method in this simple setting, show how to calculate it from bootstrap computations, and demonstrate that it gives second-order correct intervals for  $\theta$  under reasonable conditions.

Of course there is no need for the bootstrap in the simple situation  $\hat{\theta} \sim f_\theta$ , since then it is usually quite easy to calculate exact confidence intervals for  $\theta$ . There are three reasons for beginning the discussion with the simple situation: (a) it makes clear the logic of the  $BC_a$  method; (b) it makes possible the comparison of  $BC_a$  intervals with exact intervals, exact intervals usually not existing in complicated problems; (c) it then turns out to be quite easy to extend the  $BC_a$  method to complicated situations, where it is more likely to be needed.

The simple situation  $\hat{\theta} \sim f_\theta$  can be made more complicated, and more realistic, in two ways: the data can consist of a vector  $\mathbf{y}$  instead of a single summary statistic  $\hat{\theta}$ , and the parameter can be a vector  $\boldsymbol{\eta}$  instead of a single unknown scalar  $\theta$ . Section 6 considers multiparameter families  $f_{\boldsymbol{\eta}}(\mathbf{y})$ , where we wish to set an approximate confidence interval for a real-valued function  $\theta = t(\boldsymbol{\eta})$ .

Our approach is to reduce the problem back to the simple situation. The data vector  $\mathbf{y}$  is replaced by an efficient estimator  $\hat{\theta}$  of  $\theta$ , perhaps the MLE, and the multiparameter family  $f_{\boldsymbol{\eta}}$  is replaced by a *least favorable* one-parameter family. All of the calculations are handled automatically by the  $BC_a$  algorithm. For a class of examples, including the Fieller problem of Table 1, the  $BC_a$  method automatically produces second-order correct intervals, but a proof of general second-order correctness does not yet exist for multiparameter situations.

Section 7 returns to the original nonparametric setting of the bootstrap: the data  $\mathbf{y}$  is assumed to be a random sample  $x_1, x_2, \dots, x_n$  from a completely unknown probability distribution  $F$ . We wish to set an approximate confidence interval for  $\theta = t(F)$ , some real-valued function of  $F$ . The  $BC_a$  method extends in a natural way to the

nonparametric setting. In the case where  $\theta$  is the expectation, theoretical analysis shows the  $BC_a$  intervals performing reasonably. Except for the case of the expectation, not much is proved about nonparametric  $BC_a$  intervals, though the empirical results look promising. Section 8 develops a heuristic justification for the nonparametric  $BC_a$  method in terms of the geometry of multinomial sampling.

In the simple situation  $\hat{\theta} \sim f_{\hat{\theta}}$  the parametric bootstrap distribution  $\hat{\theta}^* \sim f_{\hat{\theta}}$  can often be written down explicitly, or at least approximated by standard parametric devices such as Edgeworth expansions. The number of bootstrap replications of  $\hat{\theta}^*$ , to use the terminology of previous papers, is then, effectively, infinity. For more complicated situations like the nonparametric confidence interval problem, Monte Carlo sampling is usually needed to calculate the  $BC_a$  intervals. How many bootstrap replications are necessary? The answer, on the order of 1,000, is derived in Section 9. This compares with only about 100 bootstrap replications necessary to adequately calculate a bootstrap standard error. Bootstrap confidence intervals require a lot more computation than bootstrap standard errors, if second-order accuracy is desired.

To get the main ideas across, some important technical points are deferred until Sections 10–12.

### 3. BOOTSTRAP CONFIDENCE INTERVALS FOR SIMPLE PARAMETRIC SITUATIONS

We first consider the simple situation  $\hat{\theta} \sim f_{\theta}$ , where we have a one-parameter family of densities  $f_{\theta}(\hat{\theta})$  for the real-valued statistic  $\hat{\theta}$ . We wish to set a confidence interval for  $\theta$  having observed only  $\hat{\theta}$ . The statistic  $\hat{\theta}$  estimates  $\theta$ . Later we will make specific assumptions about the properties of  $\hat{\theta}$  as an estimator of  $\theta$ —essentially that  $\hat{\theta}$  behaves like the MLE asymptotically, though  $\hat{\theta}$  may be some first-order efficient estimator other than the MLE.

By definition, the parametric bootstrap distribution in this simple situation is

$$\hat{\theta}^* \sim f_{\hat{\theta}}. \quad (3.1)$$

In other words it is the distribution of the statistic of interest when the unknown parameter  $\theta$  is set equal to the observed point estimate  $\hat{\theta}$ . We also need to define the cdf of the bootstrap distribution

$$\hat{G}(s) = \int_{-\infty}^s f_{\hat{\theta}}(\hat{\theta}^*) d\hat{\theta}^* = \Pr_{\hat{\theta}}\{\hat{\theta}^* < s\}. \quad (3.2)$$

The integral is replaced by a summation in discrete families. The goal of bootstrap theory is to make inferential statements on the basis of the bootstrap distribution. In this article the inferences are approximate confidence intervals for  $\theta$ .

*Example (chi-squared scale family).* Suppose that

$$\hat{\theta} \sim \theta(\chi_{19}^2/19), \quad (3.3)$$

the example considered in Schenker (1985). Then

$$f_{\hat{\theta}}(\hat{\theta}) = c(\hat{\theta}/\theta)^{8.5} e^{-9.5(\hat{\theta}/\theta)} \quad \text{for } \hat{\theta} > 0 \\ [c = 9.5^{9.5}/\Gamma(9.5)]. \quad (3.4)$$

Having observed  $\hat{\theta}$ , the bootstrap distribution  $\hat{\theta}^* \sim \hat{\theta}(\chi_{19}^2/19)$  has density  $c(\hat{\theta}^*/\hat{\theta})^{8.5} e^{-9.5(\hat{\theta}^*/\hat{\theta})}$  for  $\hat{\theta}^* > 0$ . The bootstrap cdf is  $\hat{G}(s) = I_{9.5}(9.5s/\hat{\theta})$ , where  $I_{9.5}$  indicates the incomplete gamma function of degree 9.5.

Now suppose that for a family  $\hat{\theta} \sim f_{\theta}$  there exists a monotone-increasing transformation  $g$  and constants  $z_0$  and  $a$  such that

$$\hat{\phi} = g(\hat{\theta}), \quad \phi = g(\theta) \quad (3.5)$$

satisfy

$$\hat{\phi} = \phi + \sigma_{\phi}(Z - z_0), \quad Z \sim N(0, 1) \quad (3.6)$$

with

$$\sigma_{\phi} = 1 + a\phi. \quad (3.7)$$

This is of form (2.3), with  $\tau = 1$ . [Eq. (2.3) can always be reduced to the case  $\tau = 1$ ; see Remark A, Sec. 11.] We will assume that  $\phi > -1/a$  if  $a > 0$  in (3.7), so  $\sigma_{\phi} > 0$ , and likewise  $\phi < -1/a$  if  $a < 0$ . The constant  $a$  will typically be in the range  $|a| < .2$ , as will  $z_0$ .

Let  $\Phi$  denote the standard normal cdf, and let  $\hat{G}^{-1}(\alpha)$  denote the  $100 \cdot \alpha$  percentile of the bootstrap cdf (3.2).

*Lemma 1.* Under conditions (3.5)–(3.7), the correct central confidence interval of level  $1-2\alpha$  for  $\theta$  is

$$\theta \in [\hat{G}^{-1}(\Phi(z[\alpha])), \hat{G}^{-1}(\Phi(z[1 - \alpha]))], \quad (3.8)$$

where

$$z[\alpha] = z_0 + \frac{(z_0 + z^{(\alpha)})}{1 - a(z_0 + z^{(\alpha)})}, \quad (3.9)$$

and likewise for  $z[1 - \alpha]$ .

The proof of Lemma 1, at the end of this section, makes clear that interval (3.8) is correct in a strong sense: it is equivalent, under assumptions (3.5)–(3.7), to the usual obvious interval for a simple translation problem. Given the bootstrap cdf  $\hat{G}(s)$  and values of  $z_0$  and  $a$  derived from bootstrap calculations as in the following sections, we can form interval (3.8), (3.9) for  $\theta$  whether or not assumptions (3.5)–(3.7) apply. This by definition is the  $BC_a$  interval for  $\theta$ .

If  $z_0$  and  $a$  equal 0, then  $z[\alpha] = z^{(\alpha)}$  and (3.8) becomes  $\theta \in [\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1 - \alpha)]$ . In this case we just use the obvious percentiles of the bootstrap distribution to form an approximate confidence interval for  $\theta$ , an approach called the *percentile method* in Efron (1981, 1982a). In general  $z_0$  and  $a$  do not equal zero, and formulas (3.8), (3.9) make adjustments to the percentile method that are necessary to achieve second-order correctness.

Continuing example (3.3), the theory of Efron (1982b) shows that for the chi-squared scale family we can find a transformation  $g$  very nearly satisfying (3.5)–(3.7). Schenker (1985) proved the same result by a different method. The constants

$$z_0 = .1082, \quad a = .1077 \quad (3.10)$$

and the transformation  $g$  appropriate to family (3.3) are derived in Section 10 and Remark E of Section 11. Simple ways of approximating  $z_0$  and  $a$  for general families  $\hat{\theta} \sim f_{\theta}$  are given in Section 4.

Line 2 of Table 2 shows the central 90%  $BC_a$  interval,  $\alpha = .05$ , for family (3.3), with  $\hat{G}(s) = I_{9.5}(9.5s/\hat{\theta})$  and  $z_0$  and  $a$  as in (3.10). The exact confidence interval is  $\theta \in [19\hat{\theta}/\chi_{19}^{2(1-\alpha)}, 19\hat{\theta}/\chi_{19}^{2(\alpha)}]$ , where  $\chi_{19}^{2(\alpha)}$  is the 100 ·  $\alpha$  percentile point of a  $\chi_{19}^2$  distribution. Notice how closely the  $BC_a$  endpoints match those of the exact interval (see line 1). The standard interval (1.1) is quite inaccurate in this case.

Suppose that we set  $a = 0$  in (3.9), so  $z[\alpha] = 2z_0 + z^{(\alpha)}$ . Interval (3.8) with this definition of  $z[\alpha]$  and  $z[1 - \alpha]$  is called the *BC interval*, short for bias-corrected bootstrap interval, in Efron (1981, 1982a). In other words,  $BC = BC_a$ , with  $a = 0$ . The constant  $z_0$  is easier to obtain than the constant  $a$ , as discussed in the next section, which is why the BC interval might be used. Line 3 of Table 2 shows that for family (3.3) the BC interval is a definite improvement over the standard interval but goes only about half as far as it should toward achieving the asymmetry of the exact interval.

The Fieller situation of Table 1 is an example of a class of multiparameter problems for which  $a = 0$ , so the BC and  $BC_a$  intervals coincide. Efron (1985) showed that the BC intervals are second-order correct for this class, as discussed in Section 6. In general problems the full  $BC_a$  method is necessary to get second-order correctness, as shown in Section 5.

Bartlett (1953) and Schenker (1985) discussed problem (3.3). The  $BC_a$  method can be thought of as a computer-based way to carry out Bartlett's program of improved approximate confidence intervals without having to do his theoretical calculations.

*Proof of Lemma 1.* We begin by showing that the  $BC_a$  interval for  $\phi$  based on  $\hat{\phi}$  is correct in a certain obvious sense: notice that (3.6), (3.7) give

$$\{1 + a\hat{\phi}\} = \{1 + a\phi\}\{1 + a(Z - z_0)\}. \quad (3.11)$$

Taking logarithms puts the problem into standard translation form,

$$\hat{\zeta} = \zeta + W, \quad (3.12)$$

$\hat{\zeta} = \log\{1 + a\hat{\phi}\}$ ,  $\zeta = \log\{1 + a\phi\}$ , and  $W = \log\{1 + a(Z - z_0)\}$ . This example was discussed more carefully in Sections 4 and 8 of Efron (1982b), where the possibility of the bracketed terms in (3.11) being negative was dealt with. Here it will cause no trouble to assume them positive so that it is permissible to take logarithms. In fact the transformation to (3.12) is only for motivational purposes. A quicker but less informative proof of Lemma 1 is possible, working directly on the  $\phi$  scale.

Table 2. Central 90% Confidence Intervals for  $\theta$  Having Observed  $\hat{\theta} \sim \theta\chi_{19}^2/19$

		R/L
1. Exact	[.631 $\hat{\theta}$ , 1.88 $\hat{\theta}$ ]	2.38
2. $BC_a$ ( $a = .1077$ )	[.630 $\hat{\theta}$ , 1.88 $\hat{\theta}$ ]	2.37
3. BC ( $a = 0$ )	[.580 $\hat{\theta}$ , 1.69 $\hat{\theta}$ ]	1.64
4. Standard (1.1)	[.466 $\hat{\theta}$ , 1.53 $\hat{\theta}$ ]	1.00
5. Nonparametric $BC_a$	[.640 $\hat{\theta}$ , 1.68 $\hat{\theta}$ ]	1.88

NOTE: The  $BC_a$  interval, with  $a = .1077$ , the correct value, is nearly identical to the exact interval. The BC interval,  $a = 0$ , is only a partial improvement over the standard interval. The nonparametric  $BC_a$  interval is discussed in Section 7.

The translation problem (3.12) gives a natural central  $1 - 2\alpha$  interval for  $\zeta$  having observed  $\hat{\zeta}$ ,

$$\zeta \in [\hat{\zeta} - w^{(1-\alpha)}, \hat{\zeta} - w^{(\alpha)}], \quad (3.13)$$

where  $w^{(\alpha)}$  is the 100 ·  $\alpha$  percentile point for  $W$ ,  $\Pr\{W < w^{(\alpha)}\} = \alpha$ .

We will use the notation  $\theta[\alpha]$  for the  $\alpha$ -level endpoint of a confidence interval for a parameter  $\theta$ . For instance (3.13) says that  $\zeta[\alpha] = \hat{\zeta} - w^{(1-\alpha)}$ ,  $\zeta[1 - \alpha] = \hat{\zeta} - w^{(\alpha)}$ . The interval (3.13) can be transformed back to the  $\phi$  scale by the inverse mappings  $\hat{\phi} = (e^{\hat{\zeta}} - 1)/a$ ,  $\phi = (e^{\zeta} - 1)/a$ ,  $(Z - z_0) = (e^W - 1)/a$ . A little algebraic manipulation shows that the resulting interval for  $\phi$  has  $\alpha$ -level endpoint

$$\phi[\alpha] = \hat{\phi} + \sigma_{\hat{\phi}} \frac{(z_0 + z^{(\alpha)})}{1 - a(z_0 + z^{(\alpha)})}. \quad (3.14)$$

The cdf of  $\hat{\phi}$  according to (3.6) is  $\Phi((s - \phi)/\sigma_{\hat{\phi}} + z_0)$ , so the bootstrap cdf of  $\hat{\phi}^*$ , say  $\hat{H}$ , is  $\hat{H}(s) = \Phi((s - \hat{\phi})/\sigma_{\hat{\phi}} + z_0)$ . This has inverse  $\hat{H}^{-1}(\alpha) = \hat{\phi} + \sigma_{\hat{\phi}}\{\Phi^{-1}(\alpha) - z_0\}$ , which shows that  $\hat{H}^{-1}(\Phi(z[\alpha]))$  equals (3.14) [see definition (3.9)]. In other words, the  $BC_a$  interval for  $\phi$ , based on  $\hat{\phi}$ , coincides with the correct interval (3.14), "correct" meaning in agreement with the translation interval (3.13).

The  $BC_a$  intervals transform in the obvious way: if  $\hat{\phi} = g(\hat{\theta})$ ,  $\phi = g(\theta)$ , then the  $BC_a$  interval endpoints satisfy  $\phi[\alpha] = g(\theta[\alpha])$ . This follows directly from (3.8), (3.9) and the relationship  $\hat{H}(g(s)) = \hat{G}(s)$ , equivalently  $\hat{H}^{-1}(\alpha) = g(\hat{G}^{-1}(\alpha))$ , between the two bootstrap cdf's. Lemma 1 has now been verified: the transformations  $\hat{\theta} \rightarrow \hat{\phi} \rightarrow \hat{\zeta}$  and  $\theta \rightarrow \phi \rightarrow \zeta$  reduce the problem to translation form (3.12); the inverse transformations of the natural interval (3.13) for  $\zeta$  produce the  $BC_a$  interval (3.8), (3.9).

#### 4. THE TWO CONSTANTS

The  $BC_a$  intervals require the statistician to calculate the bootstrap distribution  $\hat{G}$  and also the two constants  $z_0$  and  $a$ . The bootstrap distribution is obtained directly from (3.2). This calculation does not require knowledge of the normalizing transformation  $g$  occurring in (3.5). The two constants  $z_0$  and  $a$  can also be obtained, or at least approximated, directly from the bootstrap distribution  $f_{\hat{\theta}}(\hat{\theta}^*)$ . These calculations, which are the subject of this section, assume that a transformation  $g$  to form (3.6), (3.7) exists, but do not require  $g$  to be known.

In fact the bias-correction constant  $z_0$  is

$$z_0 = \Phi^{-1}(\hat{G}(\hat{\theta})) \quad (4.1)$$

under assumptions (3.5)–(3.7), and so can be computed directly from  $\hat{G}$ . To verify (4.1) notice that

$$\Pr_{\theta}\{\hat{\phi} < \phi\} = \Pr\{Z < z_0\} = \Phi(z_0) \quad (4.2)$$

according to (3.6). However, (3.5) gives

$$\Pr_{\theta}\{\hat{\theta} < \theta\} = \Pr_{\phi}\{\hat{\phi} < \phi\} = \Phi(z_0) \quad (4.3)$$

for every value of  $\theta$ . Substituting  $\theta = \hat{\theta}$  gives  $\hat{G}(\hat{\theta}) = \Pr_{\theta}\{\hat{\theta}^* < \hat{\theta}\} = \Phi(z_0)$ , which is (4.1).

What about the acceleration constant  $a$ ? We will show that a good approximation for  $a$  is

$$a \doteq \frac{1}{6} \text{SKEW}_{\theta=\hat{\theta}}(\hat{\theta}), \quad (4.4)$$

where  $\text{SKEW}_{\theta=\hat{\theta}}(X)$  indicates the skewness of a random variable  $X$ ,  $\mu_3(X)/\mu_2(X)^{3/2}$ , evaluated at parameter point  $\theta$  equal to  $\hat{\theta}$ , and  $l_\theta$  is the score function of the family  $f_\theta(\hat{\theta})$ ,

$$l_\theta(\hat{\theta}) = \partial/\partial\theta \log f_\theta(\hat{\theta}). \tag{4.5}$$

Formula (4.4) allows us to calculate  $a$  from the form of the given density  $f_\theta$  near  $\theta = \hat{\theta}$ , without knowing  $g$ . Sections 6 and 7 discuss the computation of  $a$  in families with nuisance parameters. Section 10 gives a deeper discussion of  $a$  and its relationship to other quantities of interest. See also Remark F, Section 11.

*Example.* For  $\hat{\theta} \sim \theta(\chi^2_{19}/19)$ , as in Table 2, standard  $\chi^2$  calculations give  $\text{SKEW}(l_\theta)/6 = [8/(19 \cdot 36)]^{1/2} = .1081$ , which is quite close to the actual value  $a = .1077$  derived in Section 10.

Here is a simple heuristic argument that indicates the role of the constant  $a$  in setting approximate confidence intervals. Suppose that  $z_0 = 0$  and  $a > 0$  in (3.6), (3.7). Having observed  $\hat{\phi} = 0$ , and noticing  $\sigma_\phi = 1$ , the naive interval for  $\phi$  [which is almost the same as the standard interval (1.1)] is  $\phi \in [z^{(a)}, z^{(1-a)}]$ . If, however, the statistician checks the situation at the right endpoint  $z^{(1-a)}$ , he finds that the hypothesized standard deviation of  $\hat{\phi}$  has increased from 1 to  $1 + az^{(1-a)}$ . This suggests increasing the right endpoint to  $z^{(1-a)}(1 + az^{(1-a)})$ . Now the hypothesized standard deviation has further increased to  $1 + az^{(1-a)}(1 + az^{(1-a)})$ , suggesting a still larger right endpoint, and so forth. Continuing on in this way results in formula (3.14), leading to Lemma 1. [Improving the standard interval (1.1) by recomputing  $\hat{\sigma}$  at its endpoints is a useful idea. It was brought to my attention by John Tukey, who pointed out its use by Bartlett (1953); see, e.g., Bartlett's eq. (17). Tukey's (1949) unpublished talk anticipated many of the same points.]

We call  $a$  the acceleration constant because of its effect of constantly changing the natural units of measurement as we move along the  $\phi$  (or  $\theta$ ) axis. Notice that we can write (3.7) as

$$\sigma_\phi = \sigma_{\phi_0} [1 + a(\phi - \phi_0)/\sigma_{\phi_0}], \tag{4.6}$$

so

$$a = \frac{d(\sigma_\phi/\sigma_{\phi_0})}{d((\phi - \phi_0)/\sigma_{\phi_0})} \tag{4.7}$$

for any fixed value of  $\phi_0$ . This shows that  $a$  is the relative change in  $\sigma_\phi$  per unit standard deviation change in  $\phi$ , no matter what value  $\phi$  has.

The point  $\phi_0 = 0$  is favored in definition (3.7), since  $\sigma_0$  has been set equal to the convenient value 1. There is no harm in thinking of 0 as the true value of  $\phi$ , the value actually governing the distribution of  $\hat{\phi}$  in (3.8), because in theory we can always choose the transformation  $g$  so that this is the case and, in addition, so that  $\sigma_0 = 1$  (see Remark A, Sec. 11). The restriction  $1 + a\phi > 0$  in (3.7) causes no practical trouble for  $|a| \leq .2$ , since it is then at least 5 standard deviations to the boundary of the permissible  $\phi$  region.

The remainder of this section is devoted to verifying

(4.4). The discussion is fairly technical and can be deferred until Section 10 at the reader's preference.

If we make smooth one-to-one transformations  $\hat{\phi} = g(\hat{\theta})$ ,  $\phi = h(\theta)$ , then  $l_\phi(\hat{\phi}) = l_\theta(\hat{\theta})/h'(\theta)$  and  $\text{SKEW}(l_\phi) = \text{SKEW}(l_\theta)$ . In other words, the right side of (4.4) is invariant under all mappings of this type. Suppose that for some choice of  $g$  and  $h$ , we can represent the family of distributions of  $\hat{\phi}$  as

$$\hat{\phi} = \phi + \sigma_\phi q(Z), \quad Z \sim N(0, 1), \tag{4.8}$$

where  $\sigma_\phi$  and  $q(Z)$  are functions of  $\phi$  and  $z$ , having at least one and two derivatives, respectively,  $q'(Z) > 0$ . Situation (4.8), with the added conditions  $q(0) = 0$ ,  $q'(0) = 1$ , is called a general scaled transformation family (GSTF) in Efron (1982b). [Please note the corrigenda to Efron (1982b).]

*Lemma 2.* The family (4.8) has score function  $l_\phi(\hat{\phi})$  satisfying

$$\sigma_\phi l_\phi(\hat{\phi}) \sim \left[ Z + \frac{q''(Z)}{q'(Z)} \right] \left[ \frac{1 + \dot{\sigma}_\phi q(Z)}{q'(Z)} \right] - \dot{\sigma}_\phi, \tag{4.9}$$

$Z \sim N(0, 1)$ .

Here  $\dot{\sigma}_\phi = d\sigma_\phi/d\phi$  and  $q'$  and  $q''$  are the first two derivatives of  $q$ .

Before presenting the proof of Lemma 2, we note that it verifies (4.4): in situation (3.6), (3.7), where  $\dot{\sigma}_\phi = a$ ,  $q'(Z) = 1$ ,  $q''(Z) = 0$ , the distributional relationship (4.9) becomes

$$\sigma_\phi l_\phi(\hat{\phi}) \sim (1 - az_0) \left[ Z + \frac{a}{1 - az_0} (Z^2 - 1) \right]. \tag{4.10}$$

Let

$$\varepsilon_0 = \frac{a}{1 - az_0}, \tag{4.11}$$

a quantity discussed in Section 10. From the moments of  $Z \sim N(0, 1)$ , (4.10) gives

$$\frac{\text{SKEW}(l_\phi)}{6} = \varepsilon_0 \frac{1 + \frac{3}{2} \varepsilon_0^2}{(1 + 2\varepsilon_0^2)^{3/2}}. \tag{4.12}$$

We will see in Section 10 that for the usual repeated sampling situation both  $a$  and  $z_0$  are order of magnitude  $O(n^{-1/2})$  in the sample size  $n$ . This means that  $\varepsilon_0 = a \cdot [1 + O(n^{-1})]$ , (4.11), and that  $\text{SKEW}(l_\theta)/6 = \text{SKEW}(l_\phi)/6 = a[1 + O(n^{-1})]$ , (4.12), justifying approximation (4.4). The "constant"  $a$  actually depends on  $\theta$ , but substituting  $\theta = \hat{\theta}$  in (4.4) causes errors only at the third-order level, like  $\hat{\sigma} B_n^{(a)}/n$  in (1.2), and so does not affect the second-order properties of the  $BC_a$  intervals.

*Proof of Lemma 2.* Starting from (4.8), the cdf of  $\hat{\phi}$  is  $\Phi(q^{-1}((\hat{\phi} - \phi)/\sigma_\phi))$ , so  $\hat{\phi}$  has density  $f_\phi(\hat{\phi}) = \exp(-\frac{1}{2}Z_\phi^2)/(\sqrt{2\pi} \sigma_\phi q'(Z_\phi))$ , where  $Z_\phi \equiv q^{-1}((\hat{\phi} - \phi)/\sigma_\phi)$ . This gives log-likelihood function

$$l_\phi(\hat{\phi}) = -\frac{1}{2}Z_\phi^2 - \log(q'(Z_\phi)) - \log(\sigma_\phi). \tag{4.13}$$

Lemma 2 follows by differentiating (4.13) with respect to  $\phi$  and noting that  $Z_\phi \sim N(0, 1)$  when sampling from (4.8).

### 5. SECOND-ORDER CORRECTNESS OF THE $BC_a$ INTERVALS

The standard intervals are based on approximation (2.1). The  $BC_a$  intervals, which improved considerably on the standard intervals in Tables 1 and 2, are based on the more general approximation (2.3). Is it possible to go beyond (2.3), to find still further improvements over the standard intervals? The answer is no, at least not in terms of second-order asymptotics. The theorem of this section states that for simple one-parameter problems the  $BC_a$  intervals coincide through second order with the exact intervals. In terms of (1.2), the  $BC_a$  intervals have the correct second-order asymptotic form  $\hat{\theta} + \hat{\sigma}(z^{(\alpha)} + A_n^{(\alpha)}/\sqrt{n} + \dots)$ .

We continue to consider the simple one-parameter problem  $\hat{\theta} \sim f_\theta$ . Suppose that the  $100 \cdot \alpha$  percentile of  $\hat{\theta}$  as a function of  $\theta$ , say  $\hat{\theta}_\alpha^{(1)}$ , is a continuously increasing function of  $\theta$  for any fixed  $\alpha$ . In this case the usual confidence interval construction gives an exact  $1 - 2\alpha$  central interval for  $\theta$  having observed  $\hat{\theta}$ , say  $[\theta_{\text{Ex}}[\alpha], \theta_{\text{Ex}}[1 - \alpha]]$ , where  $\theta_{\text{Ex}}[\alpha]$  is the value of  $\theta$  satisfying  $\hat{\theta}_\alpha^{(1-\alpha)} = \hat{\theta}$ . The exact interval in Table 2 is an example of this construction.

It is not necessary that  $\hat{\theta}$  be the MLE of  $\theta$ . In (3.6), for instance,  $\hat{\phi}$  is not the MLE of  $\phi$ . The  $BC_a$  method is quite insensitive to small changes in the form of the estimator (see Remark B, Sec. 11). It will be assumed, however, that  $\hat{\theta}$  behaves asymptotically like the MLE in terms of the orders of magnitude of its bias, standard deviation, skewness, and kurtosis,

$$\hat{\theta} - \theta \sim (B_\theta/n, C_\theta/\sqrt{n}, D_\theta/\sqrt{n}, E_\theta/n). \tag{5.1}$$

Here  $n$  is the sample size upon which the summary statistic  $\hat{\theta}$  is based;  $B_\theta, C_\theta, D_\theta,$  and  $E_\theta$  are bounded functions of  $\theta$  (and of  $n$ , which is suppressed in the notation). Then (5.1) says that the bias of  $\hat{\theta}$ ,  $B_\theta/n$ , is  $O(n^{-1})$ , the standard deviation  $C_\theta/\sqrt{n}$  is  $O(n^{-1/2})$ , skewness  $O(n^{-1/2})$ , and kurtosis  $O(n^{-1})$ . Higher cumulants, which are typically of order smaller than  $O(n^{-1})$ , will be assumed negligible in proving the results that follow (see DiCiccio 1984; Hougaard 1982).

In the simple situation  $\hat{\theta} \sim f_\theta$ ,  $\hat{\theta}$  is a sufficient statistic for  $\theta$ . Later when we consider more complicated problems we will take  $\hat{\theta}$  to be the MLE of  $\theta$ . This guarantees that  $\hat{\theta}$  is first-order efficient and asymptotically sufficient (Efron 1975).

The asymptotics of this article are stated relative to the size of the estimated standard error  $\hat{\sigma}$  of  $\hat{\theta}$ , as in (1.2). It is often convenient in what follows to have  $\hat{\sigma}$  be  $O_p(1)$ . This is easy to accomplish by transforming to  $\hat{\phi} \equiv \sqrt{n}\hat{\theta}$ ,  $\phi \equiv \sqrt{n}\theta$ , so (5.1) becomes

$$\hat{\phi} - \phi \sim (\beta_\phi, \sigma_\phi, \gamma_\phi, \delta_\phi), \tag{5.2}$$

where  $\beta_\phi = B_{\hat{\theta}}/n^{1/2}$ ,  $\sigma_\phi = C_{\hat{\theta}}/n^{1/2}$ ,  $\gamma_\phi = D_{\hat{\theta}}/n^{1/2}$ , and  $\delta_\phi = E_{\hat{\theta}}/n^{1/2}$ . Notice that  $\beta_\phi = O(n^{-1/2})$ ,  $\beta_\phi \equiv d\beta_\phi/d\phi = O(n^{-1})$ , and so forth. We can just assume to begin with that  $\hat{\theta}$  and  $\theta$  are the rescaled quantities previously called

$\hat{\phi}$  and  $\phi$ . Then the following orders of magnitude apply:

$$O(1) \quad O(n^{-1/2}) \quad O(n^{-1}) \quad O(n^{-3/2}) \\ \sigma_\theta \quad \hat{\sigma}_\theta, \beta_\theta, \gamma_\theta \quad \ddot{\sigma}_\theta, \hat{\beta}_\theta, \hat{\gamma}_\theta, \delta_\theta \quad \hat{\beta}_\theta, \hat{\gamma}_\theta, \hat{\delta}_\theta \tag{5.3}$$

*Theorem 1.* If  $\hat{\theta}$  has bias  $\beta_\theta$ , standard error  $\sigma_\theta$ , skewness  $\gamma_\theta$ , and kurtosis  $\delta_\theta$  satisfying (5.3), then the  $BC_a$  intervals are second-order correct.

The theorem states that  $\theta_{BC_a}[\alpha]$ , the  $\alpha$  endpoint of the  $BC_a$  interval, is asymptotically close to the exact endpoint,

$$(\theta_{BC_a}[\alpha] - \theta_{\text{Ex}}[\alpha])/\hat{\sigma} = O_p(n^{-1}). \tag{5.4}$$

This is not true for the standard intervals (1.1) or the  $BC$  intervals,  $a = 0$ . The proof of Theorem 1, which appears in Section 12, makes it clear that all three of the elements in (2.3), the transformation  $g$ , the bias-correction constant  $z_0$ , and the acceleration constant  $a$ , make necessary corrections of  $O_p(n^{-1/2})$  to the standard intervals.

### 6. NUISANCE PARAMETERS

The discussion so far has centered on the simple case  $\hat{\theta} \sim f_\theta$ , where we have only a real-valued parameter  $\theta$  and a real-valued summary statistic  $\hat{\theta}$  from which we are trying to construct a confidence interval for  $\theta$ . We have been able to show favorable properties of the  $BC_a$  intervals for the simple case, but of course the simple case is where we least need a general method like the bootstrap.

This section discusses the more difficult situation where there are nuisance parameters besides the parameter of interest  $\theta$ . Section 7 discusses the nonparametric situation, where the number of nuisance parameters is effectively infinite. Because of the inherently simple nature of the bootstrap it will be easy to extend the  $BC_a$  method to cover these cases, though we will not be able to provide as strong a justification for the correctness of the resulting intervals.

Suppose then that the data  $\mathbf{y}$  comes from a parametric family  $\mathcal{F}$  of density functions  $f_\eta$ , say  $\mathbf{y} \sim f_\eta$ , where  $\eta$  is an unknown vector of parameters, and we want a confidence interval for the real-valued parameter  $\theta = t(\eta)$ . In Efron (1985), the multivariate normal case  $\mathbf{y} \sim N_k(\eta, \mathbf{I})$  is examined in detail.

From  $\mathbf{y}$  we obtain  $\hat{\eta}$ , the MLE of  $\eta$ , and  $\hat{\theta} = t(\hat{\eta})$ , the MLE of  $\theta$ . The parametric bootstrap distribution of  $\mathbf{y}$  is defined to be

$$\mathbf{y}^* \sim f_{\hat{\eta}}, \tag{6.1}$$

the distribution of the data when  $\eta$  equals  $\hat{\eta}$ . From  $\mathbf{y}^*$  we obtain  $\hat{\eta}^*$ , the bootstrap MLE of  $\eta$ , and then  $\hat{\theta}^* = t(\hat{\eta}^*)$ .

The distribution of  $\hat{\theta}^*$  under model (6.1) is the parametric bootstrap distribution of  $\hat{\theta}$ , generalizing (3.1). This gives the bootstrap cdf

$$\hat{G}(s) = \text{Pr}_{\hat{\eta}}\{\hat{\theta}^* < s\}, \tag{6.2}$$

as in (3.2). The bias-correction constant  $z_0$  equals  $\Phi^{-1}(\hat{G}(\hat{\theta}))$ , as in (4.1).

To compute the  $BC_a$  intervals (3.8), (3.9), we also need to know the appropriate value of the acceleration constant  $a$ . We will find  $a$  by following Stein's (1956) construction,

which replaces the multiparameter family  $\mathcal{F} = \{f_{\boldsymbol{\eta}}\}$  by a *least favorable* one-parameter family  $\hat{\mathcal{F}}$ .

Let  $\dot{l}_{\boldsymbol{\eta}}$  be the vector with  $i$ th coordinate  $\partial/\partial\eta_i \log f_{\boldsymbol{\eta}}(\mathbf{y})$ , so  $\dot{l}_{\hat{\boldsymbol{\eta}}}(\mathbf{y}) = 0$  by definition of the MLE  $\hat{\boldsymbol{\eta}}$ , and let  $\ddot{l}_{\hat{\boldsymbol{\eta}}}$  be the  $k \times k$  matrix with  $ij$ th entry  $\partial^2/(\partial\eta_i\partial\eta_j) \log f_{\boldsymbol{\eta}}(\mathbf{y})|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}}$ . In addition, let  $\hat{\nabla}$  be the gradient vector of  $\theta = t(\boldsymbol{\eta})$  evaluated at the MLE,  $\hat{\nabla}_i = (\partial/\partial\eta_i)t(\boldsymbol{\eta})|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}}$ . The *least favorable direction* at  $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$  is defined to be

$$\hat{\boldsymbol{\mu}} \equiv (-\ddot{l}_{\hat{\boldsymbol{\eta}}})^{-1}\hat{\nabla}. \tag{6.3}$$

Then the least favorable family  $\hat{\mathcal{F}}$  is the one-parameter subfamily of  $\mathcal{F}$  passing through  $\hat{\boldsymbol{\eta}}$  in the direction  $\hat{\boldsymbol{\mu}}$ ,

$$\hat{\mathcal{F}} = \{\hat{f}_{\lambda}(\mathbf{y}^*) \equiv f_{\hat{\boldsymbol{\eta}}+\lambda\hat{\boldsymbol{\mu}}}(\mathbf{y}^*)\}. \tag{6.4}$$

Using  $\mathbf{y}^*$  to denote a hypothetical data vector from  $\hat{f}_{\lambda}$  is intended to avoid confusion with the actual data vector  $\mathbf{y}$  that gave  $\hat{\boldsymbol{\eta}}$ ;  $\hat{\boldsymbol{\eta}}$  and  $\hat{\boldsymbol{\mu}}$  are fixed in (6.4), only  $\lambda$  being unknown.

Consider the problem of estimating  $\theta(\lambda) \equiv t(\hat{\boldsymbol{\eta}} + \lambda\hat{\boldsymbol{\mu}})$  having observed  $\mathbf{y}^* \sim \hat{f}_{\lambda}$ . The Fisher information bound for an unbiased estimate of  $\theta$  in this one-parameter family evaluated at  $\lambda = 0$  is  $\hat{\nabla}'(-\ddot{l}_{\hat{\boldsymbol{\eta}}})^{-1}\hat{\nabla}$ , which is the same as the corresponding bound for estimating  $\theta = t(\boldsymbol{\eta})$ , at  $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$ , in the multiparameter family  $\mathcal{F}$ . This is Stein's reason for calling  $\hat{\mathcal{F}}$  least favorable.

We will use  $\hat{\mathcal{F}}$  to calculate an approximate value for the acceleration constant  $a$ ,

$$a \doteq \{\text{SKEW}_{\lambda=0}[\partial \log f_{\hat{\boldsymbol{\eta}}+\lambda\hat{\boldsymbol{\mu}}}(\mathbf{y}^*)/\partial\lambda]/6\}. \tag{6.5}$$

This is formula (4.4) applied to  $\hat{\mathcal{F}}$ , assuming that  $\hat{\lambda} = 0$  (which is the MLE of  $\lambda$  in  $\hat{\mathcal{F}}$  when  $\mathbf{y}^* = \mathbf{y}$ , the actual data vector). See Remark F, Section 11.

Formula (6.5) is especially simple in the exponential family case where the densities  $f_{\boldsymbol{\eta}}(\mathbf{y})$  are of the form

$$f_{\boldsymbol{\eta}}(\mathbf{y}) = e^{n(\boldsymbol{\eta}'\mathbf{y} - \psi(\boldsymbol{\eta}))}f_0(\mathbf{y}). \tag{6.6}$$

The factor  $n$  in the exponent of (6.6) is not necessary, but it is included to agree with the situation where the data consists of iid observations  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ , each with density  $\exp(\boldsymbol{\eta}'\mathbf{x} - \psi(\boldsymbol{\eta}))$ , and  $\mathbf{y}$  is the sufficient vector  $\sum_{i=1}^n \mathbf{x}_i/n$ .

*Lemma 3.* For the exponential family (6.6), formula (6.5) gives

$$a = \frac{1}{6\sqrt{n}} \frac{\hat{\psi}^{(3)}(0)}{(\hat{\psi}^{(2)}(0))^{3/2}}, \tag{6.7}$$

where

$$\hat{\psi}^{(j)}(0) = \left. \frac{\partial^j \psi(\hat{\boldsymbol{\eta}} + \lambda\hat{\boldsymbol{\mu}})}{\partial \lambda^j} \right|_{\lambda=0}. \tag{6.8}$$

*Proof.* We have

$$\left. \frac{\partial \log f_{\hat{\boldsymbol{\eta}}+\lambda\hat{\boldsymbol{\mu}}}(\mathbf{y}^*)}{\partial \lambda} \right|_{\lambda=0} = n\hat{\boldsymbol{\mu}}'(\mathbf{y}^* - \hat{\psi}(\hat{\boldsymbol{\eta}})), \tag{6.9}$$

so  $\text{SKEW}_{\lambda=0}[(\partial \log f_{\hat{\boldsymbol{\eta}}+\lambda\hat{\boldsymbol{\mu}}}(\mathbf{y}^*)/\partial\lambda)]$  equals the skewness of  $\hat{\boldsymbol{\mu}}'\mathbf{y}^*$  for  $\mathbf{y}^* \sim \hat{f}_{\hat{\boldsymbol{\eta}}}$ . The fact that  $\text{SKEW}(\hat{\boldsymbol{\mu}}'\mathbf{y}^*)$  equals  $[\hat{\psi}^{(3)}(0)/(\hat{\psi}^{(2)}(0))^{3/2}]/\sqrt{n}$  is a standard exercise in exponential family theory. Note that Lemma 3 applies to  $\mathbf{y} \sim$

$N_k(\boldsymbol{\eta}, \mathbf{I})$ , the case considered in Efron (1985), and gives  $a = 0$ , which is why the unaccelerated BC intervals worked well there.

Table 3 relates to the following example:

$$\mathbf{y} \sim N_4(\boldsymbol{\eta}, \sigma_{\boldsymbol{\eta}}^2\mathbf{I}), \quad [\sigma_{\boldsymbol{\eta}} = 1 + a(\|\boldsymbol{\eta}\| - 8)], \tag{6.10}$$

where we observe  $\mathbf{y} = (8, 0, 0, 0)$  and wish to set confidence intervals for the parameter  $\theta = t(\boldsymbol{\eta}) = \|\boldsymbol{\eta}\|$ . The case  $a = 0$  amounts to finding a confidence interval for the noncentrality parameter of a noncentral  $\chi^2$  distribution and can be solved exactly. The theory of Efron (1985) applies to the  $a = 0$  case, and we see that the  $\text{BC}_0$  interval, that is, the BC interval, well-matches the exact interval.

Table 3 shows the result of varying the constant  $a$  from .10 to  $-.10$ . This example has a particularly simple geometry: the sphere  $C_{\hat{\theta}} = \{\boldsymbol{\eta} : \|\boldsymbol{\eta}\| = \hat{\theta}\}$  is the set of  $\boldsymbol{\eta}$  vectors having  $t(\boldsymbol{\eta})$  equal to the MLE value  $\hat{\theta} = t(\hat{\boldsymbol{\eta}})$ ; the least favorable direction  $\hat{\boldsymbol{\mu}}$  is orthogonal to  $C_{\hat{\theta}}$  at  $\hat{\boldsymbol{\eta}}$ ; the distribution of  $\hat{\theta}$  is nearly normal (see Efron 1985, Table 2), with standard deviation changing in the least favorable direction at a rate nearly equal to  $a$ , as in (4.7). The  $\text{BC}_a$  intervals alter predictably with  $a$ . For instance, comparing the upper endpoint at  $a = .10$  with  $a = 0$ , notice that  $(9.70 - 8.00)/(9.44 - 8.00) = 1.18$ , closely matching the expansion factor due to acceleration,  $1 + .10 \cdot 1.645 = 1.16$ .

We could disguise problem (6.10) by making nonlinear transformations

$$\tilde{\mathbf{y}} = g(\mathbf{y}), \quad \tilde{\boldsymbol{\eta}} = h(\boldsymbol{\eta}), \tag{6.11}$$

in which case the geometry of the  $\text{BC}_a$  intervals might not be obvious from the form of the parameter  $\theta = t(h^{-1}(\tilde{\boldsymbol{\eta}})) = \|h^{-1}(\tilde{\boldsymbol{\eta}})\|$  and the transformed densities  $\tilde{f}_{\tilde{\boldsymbol{\eta}}}(\tilde{\mathbf{y}})$ . However, the  $\text{BC}_a$  method is invariant under such transformations (see Remark C, Sec. 11), so the statistician would automatically get the same intervals as if he knew the normalizing transformations  $\mathbf{y} = g^{-1}(\tilde{\mathbf{y}})$ ,  $\boldsymbol{\eta} = h^{-1}(\tilde{\boldsymbol{\eta}})$ .

Currently we cannot justify the  $\text{BC}_a$  method as being second-order correct in the multiparameter context of this section, though it seems a likely conjecture that this is so. We know that it is so in the one-parameter case (see Sec. 5) and in the restricted multiparameter case of Efron (1985), where the  $\text{BC}_a$  and BC methods coincide, and that the  $\text{BC}_a$  method makes a rather obvious correction to the BC interval in the general multiparameter case.

*Table 3. Central 90% Confidence Intervals for  $\theta = \|\boldsymbol{\eta}\|$ , Having Observed  $\|\mathbf{y}\| = 8$ , From the Parametric Family  $\mathbf{y} \sim N_4(\boldsymbol{\eta}, \sigma_{\boldsymbol{\eta}}^2\mathbf{I})$ , With  $\sigma_{\boldsymbol{\eta}} = 1 + a(\|\boldsymbol{\eta}\| - 8)$*

	Exact	(R/L)	$\text{BC}_a$	(R/L)	(6.5)
$a = .10$	[6.46, 9.69]	(.96)	[6.47, 9.70]	(.97)	.0984
$a = .05$	[6.32, 9.57]	(.85)	[6.34, 9.56]	(.84)	.0498
$a = 0$	[6.14, 9.47]	(.74)	[6.19, 9.44]	(.75)	0
$a = -.05$	[5.92, 9.38]	(.65)	[6.03, 9.35]	(.66)	-.0498
$a = -.10$	[5.62, 9.30]	(.56)	[5.89, 9.27]	(.60)	-.0984

NOTE: The standard interval (1.1) is [6.36, 9.64] for all values of  $a$ . The last column shows that (6.5) nearly equals the constant  $a$  in this case. The exact intervals are based on the noncentral  $\chi^2$  distribution.

7. THE NONPARAMETRIC CASE

This section concerns the nonparametric case where the data  $\mathbf{y} = (x_1, x_2, \dots, x_n)$  consist of  $n$  iid observations  $x_i$  that may have come from any probability distribution  $F$  on their common sample space  $\mathcal{X}$ . There is a real-valued parameter  $\theta = t(F)$  for which we desire an approximate confidence interval. We will show how the  $BC_a$  method can be used to provide such an interval based on the obvious nonparametric estimate  $\hat{\theta} = t(\hat{F})$ . Here  $\hat{F}$  is the empirical probability distribution of the sample, putting mass  $1/n$  on each observed value  $x_i$ .

A bootstrap sample  $\mathbf{y}^* \sim \hat{F}$  consists in this case of an iid sample of size  $n$  from  $\hat{F}$ , say  $\mathbf{y}^* = (x_1^*, x_2^*, \dots, x_n^*)$ . In other words,  $\mathbf{y}^*$  is a random sample of size  $n$  drawn with replacement from  $\{x_1, x_2, \dots, x_n\}$ . The bootstrap sample  $\mathbf{y}^*$  gives a bootstrap replication of  $\hat{\theta}$ ,  $\hat{\theta}^* = t(\hat{F}^*)$ , where  $\hat{F}^*$  puts mass  $1/n$  on each  $x_i^*$ . The bootstrap cdf  $\hat{G}(s)$  is the probability that a bootstrap replication is less than  $s$ ,

$$\hat{G}(s) = \Pr_{\hat{F}}\{\hat{\theta}^* < s\}, \tag{7.1}$$

as in (6.2) and (3.2). The bias-correction constant  $z_0$  equals  $\Phi^{-1}(\hat{G}(\hat{\theta}))$ , as in (4.1).

For most nonparametric problems the bootstrap cdf  $\hat{G}$  has to be determined by Monte Carlo sampling. Section 9 discusses how many Monte Carlo replications of  $\hat{\theta}^*$  are necessary. Here we will continue to assume that  $\hat{G}$  has been computed exactly—in effect, that we have taken an infinite number of bootstrap replications  $\hat{\theta}^*$ .

At this point we could use  $\hat{G}$  to form the BC interval for  $\theta$ , but to obtain the  $BC_a$  interval (3.8), (3.9) we also need the value of the acceleration constant  $a$ . We will derive a simple approximation for  $a$ , based on Lemma 3. It depends on

$$U_i = \lim_{\Delta \rightarrow 0} \frac{t((1 - \Delta)\hat{F} + \Delta\delta_i) - t(\hat{F})}{\Delta}, \tag{7.2}$$

$i = 1, 2, \dots, n,$

the empirical influence function of  $\hat{\theta} = t(\hat{F})$ . Here  $\delta_i$  is a point mass at  $x_i$ , so  $U_i$  is the derivative of the estimate  $\hat{\theta}$  with respect to the mass on point  $x_i$ . [Jaeckel's infinitesimal jackknife estimate of standard error for  $\hat{\theta}$  is  $(\sum_1^n U_i^2)^{1/2}/n$ .] Definition (7.2) assumes that  $t(F)$  is smoothly defined for choices of  $F$  near  $\hat{F}$  [see Efron 1982a, (6.3), or Efron 1979, sec. 5]. Note that  $\sum_1^n U_i = 0$ .

The next section shows that Lemma 3, applied to a family appropriate to the nonparametric situation, gives the following approximation for the constant  $a$ ,

$$a \doteq \frac{1}{6} \left[ \frac{\left(\sum_{i=1}^n U_i^3\right)}{\left(\sum_{i=1}^n U_i^2\right)^{3/2}} \right]. \tag{7.3}$$

This is a convenient formula since the  $U_i$  can be evaluated easily by using finite differences in definition (7.2).

*Example 1: The Law School Data.* Table 4 shows two indexes of student excellence, LSAT and GPA, for each of 15 American law schools (see Efron 1982a, sec. 2.2). The Pearson correlation coefficient  $\hat{\rho}$  between LSAT and GPA equals .776; we want a confidence interval for the

Table 4. The Law School Data and Values of the Empirical Influence Function for the Correlation Coefficient  $\hat{\rho}$

$i$	(LSAT, GPA)	$U_i$	$i$	(LSAT, GPA)	$U_i$
1	(576, 3.39)	−1.507	9	(651, 3.36)	.310
2	(635, 3.30)	.168	10	(605, 3.13)	.004
3	(558, 2.81)	.273	11	(653, 3.12)	−.526
4	(578, 3.03)	.004	12	(575, 2.74)	−.091
5	(666, 3.44)	.525	13	(545, 2.76)	.434
6	(580, 3.07)	−.049	14	(572, 2.88)	.125
7	(555, 3.00)	−.100	15	(594, 2.96)	−.048
8	(661, 3.43)	.477			

true correlation  $\rho$ . Table 4 also shows the values of  $U_i$  for the statistic  $\hat{\rho}$ , from which formula (7.3) produces  $a \doteq - .0817$ .  $B = 100,000$  bootstrap replications (about 100 times more than actually needed; see Sec. 9) gave  $\hat{G}(\hat{\theta}) = .463$ , and so  $z_0 = - .0927$ . Using these values of  $a$  and  $z_0$  in (3.8), (3.9) resulted in the central 90% nonparametric  $BC_a$  interval [.43, .92] for  $\rho$ . The usual bivariate normal interval, based on Fisher's  $\tanh^{-1}$  transformation, is [.49, .90]. This is also the *parametric*  $BC_a$  interval based on the simple family  $\hat{\rho} \sim f_\rho$ , where  $f_\rho(\hat{\rho})$  is Fisher's density function for the correlation coefficient from bivariate normal data. The standard interval (1.1),  $\hat{\rho} \pm 1.645\hat{\sigma}$ , using the bootstrap estimate  $\hat{\sigma} = .133$ , is [.56, .99].

Formula (7.3) is invariant under monotone changes of the parameter of interest. This results in the  $BC_a$  intervals having correct transformation properties. Suppose, for example, that we change parameters from  $\rho$  to  $\phi = g(\rho) \equiv \tanh^{-1}(\rho)$ , with corresponding nonparametric estimate  $\hat{\phi} = g(\hat{\rho})$ . The central 90%  $BC_a$  interval for  $\phi$  based on  $\hat{\phi}$  is then the obvious transformation of the interval for  $\theta$  based on  $\hat{\theta}$ ,  $[g(.43), g(.92)] = [.46, 1.59]$ . This compares with Fisher's  $\tanh^{-1}$  interval  $[g(.49), g(.90)] = [.54, 1.47]$  and the standard interval  $\hat{\phi} \pm 1.645\hat{\sigma}_\phi = [.49, 1.59]$ . The standard interval is much more reasonable-looking on the  $\tanh^{-1}$  scale, as we might expect from Fisher's transformation theory. As commented before, a major advantage of the  $BC_a$  method is that the statistician need not know the correct scale on which to work. In effect the method effectively selects the best (most normal) scale and then transforms the interval back to the scale of interest.

*Example 2: The Mean.* Suppose that  $F$  is a distribution on the real line, and  $\theta = t(F)$  equals the expectation  $E_F X$ . The empirical influence function  $U_i = (x_i - \bar{x})$ , so (7.3) gives

$$a = \frac{1}{6} \frac{\sum (x_i - \bar{x})^3 / [\sum (x_i - \bar{x})^2]^{3/2}}{\sum (x_i - \bar{x})^2} = (1/6\sqrt{n})(\hat{\mu}_3/\hat{\mu}_2^{3/2}) = \hat{\gamma}/6\sqrt{n}. \tag{7.4}$$

Here  $\hat{\mu}_h = \sum (x_i - \bar{x})^h/n$ , the  $h$ th sample central moment, and  $\hat{\gamma} = \hat{\mu}_3/\hat{\mu}_2^{3/2}$  the sample skewness. It turns out also that  $z_0 \doteq \hat{\gamma}/6\sqrt{n}$  in this case, by standard Edgeworth arguments. Both  $a$  and  $z_0$  are typically of order  $n^{-1/2}$ .

Because the sample mean is such a simple statistic, we can use Edgeworth methods to get asymptotic expressions for the  $\alpha$ -level endpoint of the  $BC_a$  interval:

$$\theta_{BC_a}[\alpha] = \bar{x} + \hat{\sigma}\{z^{(\alpha)} + (\hat{\gamma}/6\sqrt{n})(2z^{(\alpha)^2} + 1) + O_p(n^{-1})\}, \tag{7.5}$$

$\hat{\sigma} \equiv (\hat{\mu}_2/n)^{1/2}$ . This compares with  $\theta_{BC}[\alpha] \doteq \bar{x} + \hat{\sigma}\{z^{(\omega)} + (\hat{\gamma}/6\sqrt{n})(z^{(\omega)^2} + 1) + O_p(n^{-1})\}$ , (7.6)

for the BC interval, so the  $BC_a$  intervals are shifted approximately  $(\hat{\gamma}/6\sqrt{n})z^{(\omega)^2}$  further right.

Johnson (1978) suggested modifying the usual  $t$  statistic  $T = (\bar{x} - \theta)/\hat{\sigma}$  to  $T_J = T + (\hat{\gamma}/6\sqrt{n})(2T^2 + 1)$  and then considering  $T_J$  to have a standard  $t_{n-1}$  distribution in order to obtain confidence intervals for  $\theta = E_F X$ . Efron (1981, sec. 10) showed that this is much like using the bootstrap distribution of  $T^* = (\bar{x}^* - \bar{x})/\hat{\sigma}^*$  as a pivotal quantity. Interestingly enough, *the Edgeworth expansion of  $\theta_J[\alpha]$ , the  $\alpha$  endpoint of Johnson's interval, coincides with (7.5)*. The  $BC_a$  method makes a “ $t$  correction” in the case of  $\theta = E_F X$ , but it is not the familiar Student- $t$  correction, which operates at third order in (1.2), but rather a second-order correction, coming from the correlation between  $\bar{x}$  and  $\hat{\sigma}$  in nonnormal populations (see Remark D, Sec. 11).

I conjecture that the nonparametric  $BC_a$  intervals will be second-order correct for any parameter  $\theta$ . There is no proof of this, a major difficulty being the definition of second-order correctness in the nonparametric situation. Whether or not it is true, small-sample nonparametric confidence intervals are far from well understood and, as emphasized in Schenker (1985), should be interpreted with some caution.

*Example 3: The Variance.* Suppose that  $\mathcal{X}$  is the real line and  $\theta = \text{var}_F X$ , the variance. Line 5 of Table 2 shows the result of applying the nonparametric  $BC_a$  method to data sets  $x_1, x_2, \dots, x_{20}$ , which were actually iid samples from an  $N(0, 1)$  distribution. The number .640, for example, is the average of  $\theta_{BC_a}[\.05]/\hat{\theta}$  over 40 such data sets,  $B = 4,000$  bootstrap replications per data set. The upper limit  $1.68 \cdot \hat{\theta}$  is noticeably small. The reason is simple: the nonparametric bootstrap distribution of  $\hat{\theta}^*$  has a short upper tail compared with the parametric bootstrap distribution, which is a scaled  $\chi^2_{19}$  random variable. The results of Beran (1984a), Bickel and Freedman (1981), and Singh (1981) show that the nonparametric bootstrap distribution is highly accurate asymptotically, but of course that is not a guarantee of good small-sample behavior. Bootstrapping from a smoothed version of  $\hat{F}$ , as in Efron (1982a, sec. 5.3), alleviates the problem in this particular example.

### 8. GEOMETRY OF THE NONPARAMETRIC CASE

Formula (7.3), which allows us to apply the  $BC_a$  method nonparametrically, is based on a simple heuristic argument: instead of the actual sample-space  $\mathcal{X}$  of the data points  $x_i$ , consider only distributions  $F$  supported on  $\hat{\mathcal{X}} = \{x_1, x_2, \dots, x_n\}$ , the observed data set. This is an  $n$ -category multinomial family, to which the results of Section 6 can be applied. Because the multinomial is an exponential family, Lemma 3 directly gives (7.3).

We will now examine this argument more carefully, with the help of a simple geometric representation. See Efron (1981, sec. 11) for further discussion of this approach to nonparametric confidence intervals.

A typical distribution supported on  $\hat{\mathcal{X}}$  is

$$F(\mathbf{w}) : \text{mass } w_i \text{ on } x_i, \quad (8.1)$$

where  $\mathbf{w} = (w_1, w_2, \dots, w_n)$  can be any vector in the simplex  $\mathfrak{S}_n = \{\mathbf{w} : w_i \geq 0 \forall i, \sum_1^n w_i = 1\}$ . The parameter  $\theta = t(F)$  is defined on  $\mathfrak{S}_n$  by  $\theta(\mathbf{w}) = t(F(\mathbf{w}))$ . The central point of the simplex,

$$\mathbf{w}^o \equiv \mathbf{1}/n = (1/n, 1/n, \dots, 1/n), \quad (8.2)$$

corresponds to  $F(\mathbf{w}^o) = \hat{F}$ , the usual empirical distribution;  $\theta(\mathbf{w}^o) = \hat{\theta} = t(\hat{F})$ , the nonparametric MLE of  $\theta$ . The curved surface

$$\mathcal{C}_\theta = \{\mathbf{w} : \theta(\mathbf{w}) = \theta(\mathbf{w}^o) = \hat{\theta}\} \quad (8.3)$$

comprises those distributions  $F(\mathbf{w})$  having  $\theta(\mathbf{w}) = \hat{\theta}$ . The vector  $\mathbf{U}_i$  is orthogonal to  $\mathcal{C}_\theta$  at  $\mathbf{w}^o$ , as shown in Figure 1, which follows from definition (7.2) of the empirical influence function.  $\mathbf{U}$  is essentially the gradient of  $\theta(\mathbf{w})$  at  $\mathbf{w}^o$  (see Efron 1982a, sec. 6.3).

With  $\mathbf{w}$  unknown, but  $\hat{\mathcal{X}} = \{x_1, \dots, x_n\}$  considered fixed, one can imagine setting a confidence interval for  $\theta(\mathbf{w})$  on the basis of a hypothetical sample  $x_1^*, x_2^*, \dots, x_n^* \stackrel{\text{iid}}{\sim} F(\mathbf{w})$ . A sufficient statistic is the vector of proportions  $P_i = \#\{x_j^* = x_i\}/n$ , say  $\mathbf{P} = (P_1, P_2, \dots, P_n)$ , with distribution

$$\mathbf{P} \sim \text{mult}_n(n, \mathbf{w})/n, \quad \mathbf{w} \in \mathfrak{S}_n. \quad (8.4)$$

The notation here indicates  $n$  draws from an  $n$ -category multinomial, having probability  $w_i$  for category  $i$ . We suppose that we have observed  $\mathbf{P} = \mathbf{w}^o$  in (8.4), that is, that the hypothetical sample  $x_1^*, \dots, x_n^*$  equals the actual sample  $x_1, \dots, x_n$ .

Distributions (8.4) form an  $n$ -parameter exponential family (6.6) with  $\mathbf{y} = \mathbf{P}$ ,  $\eta_i = \log(nw_i) + c$ , and  $\psi(\boldsymbol{\eta}) = \log(\sum_1^n \exp(\eta_i)/n)$ . Here  $c$  can be any constant, since all vectors  $\boldsymbol{\eta} + c\mathbf{1}$  correspond to the same probability vector  $\mathbf{w}$ , namely  $w_i = \exp(\eta_i)/\sum_1^n \exp(\eta_j)$ .

If one accepts the reduction of the original nonparametric problem to (8.4), with observed value  $\mathbf{P} = \mathbf{w}^o$ , then it is easy to carry through the least favorable family calculations (6.3)–(6.5): (i)  $\hat{\boldsymbol{\eta}} = \mathbf{0}$ ; (ii)  $\hat{\boldsymbol{\mu}} = \mathbf{U}$ ; (iii)  $\hat{f}_\lambda$  is the member of (7.4) corresponding to  $\hat{\boldsymbol{\eta}} + \lambda\hat{\boldsymbol{\mu}} = \lambda\mathbf{U}$ , namely

$$\mathbf{P}^* \sim \text{mult}(n, \mathbf{w}^\lambda)/n, \quad w_i^\lambda = \exp(\lambda U_i) / \sum_{j=1}^n \exp(\lambda U_j); \quad (8.5)$$

(iv) finally, formula (7.3) follows directly from Lemma 3, by differentiating  $\hat{\psi}(\lambda) = \log(\sum_1^n \exp(\lambda J_j)/n)$  (and remembering that  $\sum U_i = 0$ ).

Only step (ii) is not immediate, but it is a straightforward consequence of definition (6.3) and standard properties of the multinomial. It has already been noted that  $\mathbf{U}$  is orthogonal to  $\mathcal{C}_\theta$ , so  $\mathbf{U}$  is proportional to  $\hat{\mathbf{V}}$  in (6.3). However,  $-\hat{\mathbf{I}}_\theta = \mathbf{I} - \mathbf{1}\mathbf{1}'/n$ , which has pseudo-inverse  $\mathbf{I}$ . Thus  $\hat{\boldsymbol{\mu}}$  is proportional to  $\mathbf{U}$ . Since (6.7), (6.8) produce the same value of  $a$  if  $\hat{\boldsymbol{\mu}}$  is multiplied by any constant, this in effect gives  $\hat{\boldsymbol{\mu}} = \mathbf{U}$ .

An interesting case that provides some support for the

nonparametric  $BC_a$  method is that where the sample space is finite to begin with, say  $\mathcal{X} = \{1, 2, \dots, L\}$ . A typical distribution on  $\mathcal{X}$  is  $\mathbf{f} = (f_1, \dots, f_L)$ , where  $f_i = \Pr\{x_i = l\}$ . The observed sample proportions  $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_L)$ ,  $\hat{f}_i \equiv \#\{x_i = l\}/n$ , are sufficient, with distribution  $\hat{\mathbf{f}} \sim \text{mult}_L(n, \mathbf{f})/n$ . This is an  $L$ -parameter exponential family, so the theory of Section 6 applies. It turns out that Lemma 3 agrees with formula (7.3) in this case. *Nonparametric  $BC_a$  intervals are the same as parametric  $BC_a$  intervals when  $\mathcal{X}$  is finite.* See remarks G and H of Efron (1979) for the first-order bootstrap asymptotics of finite sample spaces.

Family (8.4) was used in Section 11 of Efron (1981) to motivate a method called *nonparametric tilting*, a nonparametric analog of the standard hypothesis-testing approach to confidence interval construction. The one-parameter tilting family, (11.12) of Efron (1981), is closely related to the least favorable family  $\hat{\mathcal{F}}$  in Figure 1. Efron (1981, table 5) considered samples of size  $n = 15$  for the one-sided exponential density  $f(x) = \exp[-(x + 1)]$  ( $x > -1$ ). Central 90% tilting intervals for  $\theta = E_F X$  were constructed for each of 10 such samples, averaging  $[-.34, .50]$ . The corresponding nonparametric  $BC_a$  intervals averaged  $[-.34, .52]$  and were quite similar to the tilting intervals on a sample-by-sample comparison. The nonparametric  $BC_a$  method is computationally simpler than nonparametric tilting and seems likely to give similar results in most problems.

We end this section with a useful approximation formula for the bias-correction constant  $z_0$ , developed jointly with Timothy Hesterberg. In addition to (7.2) we need the second-order empirical influence function

$$V_{ij} = \lim_{\Delta \rightarrow 0} \{t((1 - \Delta)\hat{F} + \Delta\delta_i + \Delta\delta_j) - t((1 - \Delta)\hat{F} + \Delta\delta_i) - t((1 - \Delta)\hat{F} + \Delta\delta_j) + t(\hat{F})\} / \Delta^2. \quad (8.6)$$

Define  $z_{01} \equiv (\frac{1}{6}) \sum_1^n U_i^3 / (\sum_1^n U_i^2)^{3/2}$  [approximation (7.3) for  $a$ ] and

$$z_{02} \equiv \left[ \frac{\mathbf{U}'\mathbf{V}\mathbf{U}}{\|\mathbf{U}\|^2} - \text{tr } \mathbf{V} \right] / (2n\|\mathbf{U}\|), \quad (8.7)$$

where  $\mathbf{V}$  is the  $n \times n$  matrix  $(V_{ij})$ .

*Lemma 4.* The bias-correction constant  $z_0$  approximately equals

$$\Phi^{-1}\{2\Phi(z_{01})\Phi(z_{02})\}. \quad (8.8)$$

For the law school data, Example 1 of Section 7,  $z_{01} = -.0817$  and  $z_{02} = -.0067$ , giving  $z_0 = -.0869$  from (8.8), compared with  $z_0 = -.0927 \pm .0039$  from  $B = 100,000$  bootstrap replications.

The term  $z_{01}$  relates to skewness in  $\hat{\mathcal{F}}$ , and  $z_{02}$  is a geometric term arising from the curvature of  $c_\theta$  at  $\mathbf{w}^0$ . It is analogous to formula (A15) of Efron (1985). Lemma 4 will not be proved here but is important in the sample size considerations of Section 9.

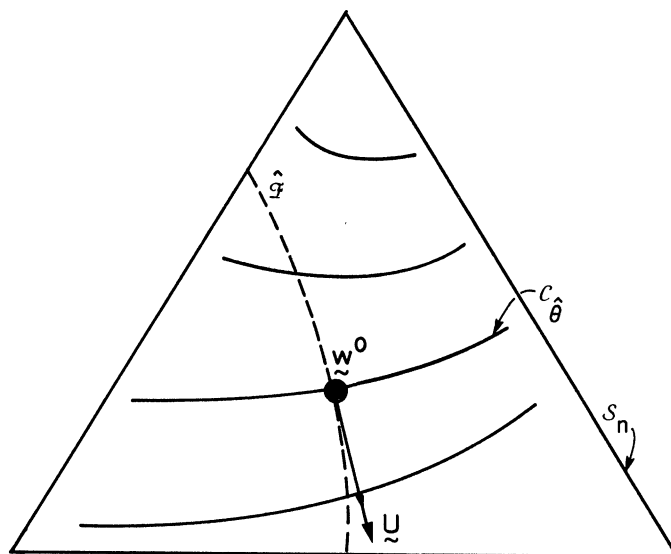


Figure 1. All probability distributions supported on  $\{x_1, x_2, \dots, x_n\}$  are represented as the simplex  $S_n$ . The central point  $\mathbf{w}^0$  corresponds to the empirical distribution  $\hat{F}$ . The curves indicate level surfaces of constant value of the parameter  $\theta$ . In particular  $c_{\hat{\theta}}$  comprises those probability distributions having  $\theta$  equal to  $\theta(\mathbf{w}^0) = \hat{\theta}$ , the MLE. The least favorable family  $\hat{\mathcal{F}}$  passes through  $\mathbf{w}^0$  in the direction  $\mathbf{U}$ , orthogonal to  $c_{\hat{\theta}}$ .

### 9. BOOTSTRAP SAMPLE SIZES

How many bootstrap replications of  $\hat{\theta}^*$  need we take? So far we have pretended that the number of replications  $B = \infty$ , but if Monte Carlo methods are necessary to obtain the bootstrap cdf  $\hat{G}$ , then  $B$  must be finite, usually the smaller the better. This section gives rough estimates of how small  $B$  may be taken in practice. The results are presented without proof, all being standard exercises in error estimation (see, e.g., Kendall and Stuart 1958, chap. 10). They apply to any situation, parametric or nonparametric, where  $\hat{G}$  is obtained by Monte Carlo sampling.

First consider the easy problem of estimating the standard error of  $\hat{\theta}$  via the bootstrap. The bootstrap estimate based on  $B$  replications,  $\hat{\sigma}_B = [\sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta}^*)^2 / (B - 1)]^{1/2}$ , has conditional coefficient of variation (standard deviation divided by expectation)

$$\text{CV}\{\hat{\sigma}_B \mid \mathbf{y}\} \doteq [(\hat{\delta} + 2)/4B]^{1/2}, \quad (9.1)$$

where  $\hat{\delta}$  is the kurtosis of the bootstrap distribution  $\hat{G}$ . The notation indicates that the observed data  $\mathbf{y}$  is fixed in this calculation. As  $B \rightarrow \infty$ , then (9.1)  $\rightarrow 0$  and  $\hat{\sigma}_B \rightarrow \hat{\sigma}$ , the ideal bootstrap estimate of standard error.

Of course  $\hat{\sigma}$  itself will usually not estimate the true standard error  $\sigma \equiv \text{SD}_{\theta}\{\hat{\theta}\}$  perfectly. Let  $\text{CV}(\hat{\sigma})$  be the coefficient of variation of  $\hat{\sigma}$ , unconditional now, averaging over the possible realizations of  $\mathbf{y}$  [e.g., if  $n = 20$ ,  $\hat{\theta} = \bar{x}$ ,  $x_i \stackrel{\text{iid}}{\sim} N(0, 1)$ , then  $\text{CV}(\hat{\sigma}) \doteq (1/40)^{1/2} = .16$ ]. The unconditional CV of  $\hat{\sigma}_B$  is then approximated by

$$\text{CV}(\hat{\sigma}_B) \doteq \left[ \text{CV}^2(\hat{\sigma}) + \frac{E\hat{\delta} + 2}{4B} \right]^{1/2}. \quad (9.2)$$

Table 5 displays  $\text{CV}(\hat{\sigma}_B)$  for various choices of  $B$  and  $\text{CV}(\hat{\sigma})$ , assuming that  $E\hat{\delta} = 0$ . For values of  $\text{CV}(\hat{\sigma}) \geq$

Table 5. Coefficient of Variation of  $\hat{\sigma}_B$ , the Bootstrap Estimate of Standard Error, as a Function of  $B$ , the Number of Bootstrap Replications, and  $CV(\hat{\sigma})$ , the Limiting CV as  $B \rightarrow \infty$

CV( $\hat{\sigma}$ )	B →				
	25	50	100	200	∞
.25	.29	.27	.26	.25	.25
.20	.24	.22	.21	.21	.20
.15	.21	.18	.17	.16	.15
.10	.17	.14	.12	.11	.10
.05	.15	.11	.09	.07	.05
0	.14	.10	.07	.05	0

NOTE: These data are based on (9.2), assuming that  $E\hat{\delta} = 0$ .

.10, typical in practice, there is little improvement past  $B = 100$ . In fact,  $B$  as small as 25 gives reasonable results.

Now we return to bootstrap confidence intervals. In the Monte Carlo situation the bootstrap cdf  $\hat{G}$  must be estimated from bootstrap replications  $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$ , say by

$$\hat{G}_B(s) = \#\{\hat{\theta}_b^* < s\} / B. \tag{9.3}$$

As  $B \rightarrow \infty$ , then  $\hat{G}_B \rightarrow \hat{G}$ , the ideal bootstrap cdf we have been using in the previous sections. Let  $\theta_B[\alpha]$  be the level  $\alpha$  endpoint of either the BC or  $BC_a$  interval obtained from  $\hat{G}_B(s)$  by substitution in (3.8), (3.9).

The following formula for the conditional CV of  $\theta_B[\alpha] - \hat{\theta}$  assumes that  $\hat{G}$  is roughly normal and that  $z_0$  and  $a$  are known, for example, from (8.8) and (6.5) or (7.3):

$$CV\{\theta_B[\alpha] - \hat{\theta} \mid \mathbf{y}\} \doteq \frac{1}{B^{1/2}|z^{(\alpha)}|} \left\{ \frac{\alpha(1-\alpha)}{\varphi(z^{(\alpha)})^2} \right\}^{1/2}, \tag{9.4}$$

$\varphi(z) \equiv \exp(-\frac{1}{2}z^2) / \sqrt{2\pi}$ . Notice that since we condition on  $\mathbf{y}$ , the only random quantity on the left side of (9.4) is  $\theta_B[\alpha]$ . Formula (9.4) measures the variability in  $\theta_B[\alpha] - \hat{\theta}$  due to taking only  $B$  bootstrap replications, rather than an infinite number.

Here is a brief tabulation of  $(9.4) \times B^{1/2}$ :

$\alpha$	:	.75	.90	.95	.975
$(9.4) \times B^{1/2}$	:	2.02	1.33	1.28	1.36

(9.5)

If  $B = 1,000$ , for instance, then  $CV\{\theta_B[.95] - \hat{\theta} \mid \mathbf{y}\} \doteq 1.28/1000^{1/2} = .040$ . Reducing  $B$  to 200 increases the conditional CV to .091. This last figure may be too big. The whole purpose of developing a theory better than (1.1) is to capture second-order effects. As the examples have indicated, these become interesting when the asymmetry ratio R/L is larger than say, 1.25, or smaller than .80. In such borderline situations, an extra 9% error in each tail due to inadequate bootstrap sampling may be unacceptable.

If the bias-correction constant  $z_0$  is estimated by Monte Carlo directly from  $z_0 = \Phi^{-1}(\hat{G}_B(\hat{\theta}))$ , rather than from (8.8), then

$$CV\{\theta_B[\alpha] - \hat{\theta} \mid \mathbf{y}\} \doteq \frac{1}{B^{1/2}z^{(\alpha)}} \left\{ \frac{1}{\varphi(0)^2} - \frac{2(1-\alpha)}{\varphi(0)\varphi(z^{(\alpha)})} + \frac{\alpha(1-\alpha)}{\varphi(z^{(\alpha)})^2} \right\}^{1/2} \tag{9.6}$$

for  $\alpha > .50$ . This gives larger CV's than (9.4):

$\alpha$	:	.75	.90	.95	.975
$(9.6) \times B^{1/2}$	:	3.04	1.97	1.75	1.71

(9.7)

Comparing (9.7) with (9.5) shows that we need  $B$  to be about twice as large to get the same CV if  $z_0$  is estimated rather than calculated. Formula (8.8) can be very helpful!

Both (9.4) and (9.6) assume that the bootstrap cdf is estimated by straightforward Monte Carlo sampling, as in (9.3). M. V. Johns (personal communication) has developed importance sampling methods that greatly accelerate the estimation of  $\hat{G}$  in some situations.

### 10. ONE-PARAMETER FAMILIES

We return to the simple situation  $\hat{\theta} \sim f_\theta$ , where there are no nuisance parameters and where we want a confidence interval for the real-valued parameter  $\theta$  based on a real-valued summary statistic  $\hat{\theta}$ . This section gives a more extensive discussion of the acceleration constant  $a$ , which has played a basic role in our considerations. Three familiar types of one-parameter families will be investigated: exponential families, translation families, and transformation families.

Efron (1982b) considered the following question: for a given family  $\hat{\theta} \sim f_\theta$ , do there exist mappings  $\hat{\phi} = g(\hat{\theta})$ ,  $\phi = h(\theta)$  such that  $\hat{\phi} = \phi + \sigma_\phi q(Z)$ ,  $Z \sim N(0, 1)$ , as in (4.8)? This last form, a General Scaled Transformation Family (GSTF), generalizes the concept of the ideal normalization, where  $\hat{\phi} = \phi + Z$ . [We now add the conditions  $q(0) = 0$ ,  $q'(0) = 1$ , as in Efron (1982b).]

The question is answered in terms of the diagnostic function  $D(z, \theta) \equiv [\varphi(0)/\varphi(z)][\hat{F}_\theta(\hat{\theta}_\theta^{(\alpha)})/\hat{F}_\theta(\mu_\theta)]$ . Here  $\varphi(z)$  is the standard normal density  $(2\pi)^{-1/2} \exp(-z^2/2)$ ;  $F_\theta$  is the cdf  $F_\theta(s) = \Pr\{\hat{\theta} \leq s\}$ ;  $\hat{F}_\theta(s) = (\partial/\partial\theta)F_\theta(s)$ ;  $\alpha = \Phi(z)$ ;  $\hat{\theta}_\theta^{(\alpha)}$  is the  $100 \cdot \alpha$  percentile of  $\hat{\theta}$  given  $\theta$ ,  $\hat{\theta}_\theta^{(\alpha)} = F_\theta^{-1}(\alpha)$ ; and  $\mu_\theta$  is the median of  $\hat{\theta}$  given  $\theta$ ,  $\mu_\theta = \hat{\theta}_\theta^{(.5)} = \hat{F}_\theta^{-1}(.5)$ . It is shown that the form of  $\sigma_\phi$  and  $q(z)$  in (4.8) can be inferred from  $D(z, \theta)$ , the main advantage being that  $D(z, \theta)$  is computed without knowledge of the normalizing transformations  $g, h$ .

The connection of transformation family theory with the acceleration constant  $a$  is the following: define

$$\varepsilon_\theta \equiv (\partial/\partial z)D(z, \theta)|_{z=0}. \tag{10.1}$$

If  $q(z)$  in (4.8) is symmetrically distributed about zero, a situation called a symmetric scaled transformation family (SSTF), then

$$\varepsilon_\theta = d\sigma_\phi/d\phi \tag{10.2}$$

(see Efron 1982b, eq. 4.11). A more complicated relationship holds for the GSTF case.

Notice that (10.2) is quite close to our original description of "a" as the rate of change of standard deviation on the normalized scale. As a matter of fact, we can transform (3.6), (3.7) into an SSTF by considering the statistic

$$\tilde{\phi} = \hat{\phi} + \frac{z_0}{1 - az_0} \sigma_\phi = \hat{\phi} + \frac{z_0}{1 - az_0} (1 + a\hat{\phi}), \tag{10.3}$$

instead of  $\hat{\phi}$  itself. Then it is easy to show that

$$\hat{\phi} = \phi + (1 + \varepsilon_0\phi)Z, \quad \varepsilon_0 = a/(1 - az_0), \quad (10.4)$$

an SSTF with  $\sigma_\phi = 1 + \varepsilon_0\phi$ ,  $\hat{\sigma}_\phi = \varepsilon_0$  for all  $\phi$ . [The quantity  $\varepsilon_0$  has the same definition in (10.4) as in (4.11).]

*Example.* For  $\hat{\theta} \sim \theta\chi^2_{19}/19$  as in Table 2,  $\varepsilon_\theta = .1090$  for all  $\theta$  [using Eq. (10.6)]. In addition,  $z_0 = \Phi^{-1} \Pr\{\chi^2_{19} < 19\} = .1082$ . The relationship  $a = \varepsilon_0/(1 + \varepsilon_0z_0)$  obtained by solving for  $a$  in (10.4) gives  $a = .1077$ , the value used in Table 2. This family is nearly in SSTF (see Remark E, Sec. 11).

We show below that under reasonable asymptotic conditions,

$$\text{SKEW}_\theta(\hat{l}_\theta)/6 \doteq \varepsilon_\theta, \quad (10.5)$$

where  $\varepsilon_\theta = (\partial/\partial z)D(z, \theta)|_{z=0}$ , as in (10.1). This last definition of  $\varepsilon_\theta$  can be evaluated for any family  $\hat{\theta} \sim f_\theta$ , assuming only that the necessary derivatives exist. The main point here is that  $\text{SKEW}_\theta(\hat{l}_\theta)/6$  always approximates  $\varepsilon_\theta$  (10.1), and in SST families  $\varepsilon_\theta$  has the acceleration interpretation (10.2).

Now to show (10.5). It is possible to reexpress (10.1) as

$$\varepsilon_\theta = -\frac{\varphi(0)}{\dot{\mu}_\theta f_\theta(\mu_\theta)} \dot{l}_\theta(\mu_\theta), \quad (10.6)$$

where  $\dot{\mu}_\theta = (d/d\theta)\mu_\theta$ , the rate of change of the median  $\mu_\theta$  with respect to  $\theta$ . For notational convenience suppose that  $\theta = 0$ . Instead of  $\hat{\theta}$ , consider the statistic  $X \equiv \hat{l}_0(\hat{\theta})/i_0$ , where  $i_0$  equals the Fisher information  $E_0 \dot{l}_0(\hat{\theta})^2$ . The parameter  $\varepsilon_\theta$  is invariant under one-to-one transformations of  $\hat{\theta}$ , so we can evaluate the right side of (10.6) in terms of  $X$ ,  $\varepsilon_\theta = -\varphi(0)\dot{l}_\theta^X(\mu_\theta^X)/\dot{\mu}_\theta^X f_\theta^X(\mu_\theta^X)$ .

For  $\theta = 0$ ,  $X$  has expectation  $E_0 X = 0$  and standard deviation  $\sigma_0^X = i_0^{-1/2}$ ; in addition,  $\dot{l}_0^X(0) = 0$ , since  $X = 0$  implies that  $\theta = 0$  is a solution of the MLE equation. Assuming the usual asymptotic convergence properties, as in (5.1), (5.3), we have the following approximations:  $\dot{\mu}_0^X \doteq 1$ ;  $\mu_0^X \doteq -\gamma_w^X i_0^{-1/2}/6$ ;  $f_0^X(\mu_0^X) \doteq \varphi(0)i_0^{1/2}$ ;  $\dot{l}_0^X(\mu_0^X) \doteq -\sqrt{i_0} \gamma_w^X/6$ . These are derived from standard Edgeworth and Taylor series arguments, which will not be presented here. Taken together they give  $\varepsilon_0 \doteq \text{SKEW}_0(\dot{l}_0^X)/6 = \text{SKEW}_0(\hat{l}_0)/6$ , which is (10.5). The quantity  $\text{SKEW}_0(\hat{l}_0)/6$  is  $O(n^{-1/2})$ , and the error of approximation in (10.5) is quite small,

$$\varepsilon_0 = [\text{SKEW}_0(\hat{l}_0)/6][1 + O(n^{-1})]. \quad (10.7)$$

Approximation (10.5) is particularly easy to understand in one-parameter exponential families. Suppose that  $x_1, x_2, \dots, x_n$  are iid observations from such a family, with sufficient statistic  $y = \bar{x}$  having density  $f_\theta(y) = \exp\{n[\theta y - \psi(\theta)]\}f_0(y)$ . In this case formula (10.6) becomes

$$\varepsilon_\theta = \frac{\sigma_\theta^Y \varphi(0)}{\dot{\mu}_\theta^Y f_\theta^Y(\mu_\theta^Y)} \left[ \frac{\lambda_\theta^Y - \mu_\theta^Y}{\sigma_\theta^Y} \right], \quad (10.8)$$

where  $\lambda_\theta^Y = E_\theta\{y\}$ ,  $\mu_\theta^Y = \text{median}_\theta\{y\}$ ,  $\dot{\mu}_\theta^Y = \partial\mu_\theta^Y/\partial\theta$ , and so forth. The term  $[(\lambda_\theta^Y - \mu_\theta^Y)/\sigma_\theta^Y] = \gamma_\theta^Y/6[1 + O(n^{-1})]$ , and  $\sigma_\theta^Y \varphi(0)/\dot{\mu}_\theta^Y f_\theta^Y(\mu_\theta^Y) = 1 + O(n^{-1})$ , both of the calcu-

lations being quite straightforward. Thus  $\varepsilon_\theta = \gamma_\theta^Y/6[1 + O(n^{-1})]$ . Since  $\dot{l}_\theta(y) = n[y - \lambda_\theta]$ , we have  $\text{SKEW}_\theta(\hat{l}_\theta(y)) = \text{SKEW}_\theta(y) = \gamma_\theta^Y$ , verifying (10.5) for one-parameter exponential families.

*Example.* If  $Y \sim \text{Poisson}(\theta)$ ,  $\theta = 15$ , then  $\text{SKEW}_\theta(\hat{l}_\theta)/6 = 1/(6 \cdot \theta^{1/2}) = .0430$ . For the continued version of the Poisson family used in Efron (1982b; note Corrigenda, p. 1032),  $(\partial/\partial z)D(z, \theta)|_{z=0} = .0425$  for  $\theta = 15$ .

*Translation Families.* Suppose that we observe a translation family  $\hat{\zeta} = \zeta + W$ , as in (3.12). Express  $W$  as a function  $q(Z)$  of  $Z \sim N(0, 1)$ , for simplicity assuming that  $q(0) = 0$  and  $q'(0) = 1$ , as in Efron (1982b). Then  $z_0 = \Phi^{-1}\Pr\{\hat{\zeta} < \zeta\} = 0$ . In this case it looks like methods based on the percentiles of the bootstrap distribution must give wrong answers, since if  $W$  is long-tailed to the right then the correct interval (3.13) is long-tailed to the left, and vice versa. However, the  $BC_a$  method produces at least roughly correct intervals, as we saw in the proof of Lemma 1.

What happens is the following: for any constant  $A$  the transformation  $g_A(t) \equiv (\exp(At) - 1)/A$  gives  $\hat{\phi} = g_A(\hat{\zeta})$ ,  $\phi = g_A(\zeta)$ , and  $Z_A = g_A(W)$  satisfying

$$\hat{\phi} = \phi + \sigma_\phi^A \cdot Z_A, \quad \sigma_\phi^A = 1 + A\phi. \quad (10.9)$$

The Taylor series for  $W = q(Z)$  begins  $W = Z + (\gamma_w/6)Z^2 + \dots$ , where  $\gamma_w = \text{SKEW}(W)$ . Then  $Z_A = Z + (\gamma_w/6)^2 Z^2 + (A/2)Z^2 + \dots$ .

The choice  $A = a \equiv -\gamma_w/3$  results in  $Z_a = Z + cZ^3 + \dots$ , the quadratic term canceling out;  $Z_a$  is then approximately normal, so (10.9) is approximately situation (3.6), (3.7), with  $z_0 = 0$ ,  $a = -\gamma_w/3$ . But we know that the  $BC_a$  intervals are correct if we can transform to situation (3.6), (3.7). An application of Lemma 2, assuming that  $Z_a \sim N(0, 1)$ , shows that  $a = -\gamma_w/3 \doteq \text{SKEW}(\hat{l}_\zeta(\hat{\zeta}))/6$  for the translation family  $\hat{\zeta} = \zeta + W$ , reverifying (4.4). [If  $Z_a \sim N(0, 1)$  in (10.9), then  $a$  must equal  $\varepsilon$ , the constant value of  $\varepsilon_\zeta$ , (10.1), for the translation family  $\hat{\zeta} = \zeta + W$ ; one can show directly that  $\varepsilon \doteq -\gamma_w/3$  for such a family.]

In the example  $\hat{\theta} \sim \theta\chi^2_{19}/19$ , the two constants  $z_0$  and  $a$  are nearly equal. This is no fluke.

*Theorem 2.* If  $\hat{\theta}$  is the MLE of  $\theta$  in a one-parameter problem having standard asymptotic properties (5.1) or (5.3), then  $z_0 \doteq a$ ,

$$z_0 = \Phi^{-1}\Pr_\theta\{\hat{\theta} < \theta\} = \frac{\text{SKEW}_\theta(\hat{l}_\theta)}{6} [1 + O(n^{-1})]. \quad (10.10)$$

*Proof.* We follow the notation and results of DiCiccio (1984): thus  $k_1, k_2, k_3$  equal the first three cumulants of  $\hat{l}_\theta$  under  $\theta$ ;  $k_{01}, k_{02}, k_{03}$  the first three cumulants of  $\hat{l}_\theta$ ;  $k_{001}$ , the first cumulant of  $\hat{l}_\theta$ ; and  $k_{11} = \text{cov}_\theta(\hat{l}_\theta, \hat{l}_\theta)$ . (So  $k_2 = i_\theta$ , the Fisher information.) All cumulants are assumed to be  $O(n)$ . Then the relative bias of  $\hat{\theta}$  is

$$b \equiv \frac{E_\theta(\hat{\theta} - \theta)}{\text{var}_\theta(\hat{\theta})^{1/2}} = \frac{k_{001} - 2k_3}{6k_2^{3/2}} + O(n^{-3/2}), \quad (10.11)$$

and  $\hat{\theta}$  has skewness

$$\gamma_\theta = \frac{k_{001} - k_3}{k_2^{3/2}} + O(n^{-3/2}). \tag{10.12}$$

Both  $b$  and  $\gamma_\theta$  are  $O(n^{-1/2})$ .

Standard Edgeworth theory now gives

$$\begin{aligned} \Pr_{\theta}\{\hat{\theta} < \theta\} &= \Phi(-b) - \frac{\gamma}{6} \varphi(b)(b^2 - 1) + O(n^{-3/2}) \\ &= .5 + \varphi(0) \frac{(2k_3 - k_{001}) + (k_{001} - k_3)}{6k_2^{3/2}} \\ &\quad + O(n^{-3/2}) \\ &= .5 + \varphi(0) \frac{k_3}{6k_2^{3/2}} + O(n^{-3/2}). \end{aligned}$$

Since  $\text{SKEW}_\theta(\hat{\theta}) = k_3/k_2^{3/2}$ , this verifies (10.10).

In multiparameter problems it is no longer true that  $z_0 \doteq a$ . The geometry of the level surface  $c_\theta$  adds another term to  $z_0$ , as in (8.8).

### 11. REMARKS

*Remark A.* Suppose that instead of (3.6), (3.7) we have  $\sigma_\phi = \tau(1 + A\phi)$ , so  $\sigma_0 = \tau$  ( $\tau \neq 1$ ). The transformations  $\hat{\phi}' \equiv \hat{\phi}/\tau$ ,  $\phi' \equiv \phi/\tau$ , give  $\hat{\phi}' = \phi' + \sigma'_{\phi'}(Z - z_0)$ , where  $\sigma'_{\phi'} = 1 + a\phi'$  and  $a = A\tau$ , so we are back in form (3.6), (3.7). Notice that the derivative  $d(\sigma_\phi/\sigma_0)/d(\phi/\sigma_0) = a$ , as in (4.7). In a similar way we can transform (3.6), (3.7) so that  $\sigma_{\phi_0} = 1$  at any point  $\phi_0$ ; the resulting value of  $a$  satisfies (4.7).

*Remark B.* Instead of using  $\hat{\phi}$  to estimate  $\phi$  in (3.6), (3.7) we might change to the estimator  $\hat{\phi}^{(c)} \equiv \hat{\phi} - c\sigma_\phi$ , for some constant  $c$ . It turns out that we are still in situation (3.6), (3.7):  $\hat{\phi}^{(c)} = \phi + \sigma_\phi^{(c)}(Z - z_0^{(c)})$ , where

$$\sigma_\phi^{(c)} = 1 + a^{(c)}(\phi - \phi_0^{(c)}), \quad \phi_0^{(c)} = c/(1 - ac), \tag{11.1}$$

and  $a^{(c)} = a(1 - ac)$ ,  $z_0^{(c)} = z_0 + \phi_0^{(c)}$ . The choice  $c = -z_0/(1 - az_0)$  gives  $z_0^{(c)} = 0$ , as in (10.3), (10.4). The choice  $c = a$  gives approximately the MLE of  $\phi$ . Interestingly enough, the  $BC_a$  interval for  $\phi$  based on  $\hat{\phi}^{(c)}$  is the same for all choices of  $c$ . Minor changes in the choice of estimator seem to have little effect on the  $BC_a$  intervals in general, though for computational reasons it is best not to use very biased estimators having large values of  $z_0$ .

*Remark C.* Section 6 uses the MLE  $\hat{\theta} = t(\hat{\eta})$ . This has one major advantage: the  $BC_a$  interval for  $\theta$ , based on  $\hat{\theta}$ , stays the same under all multivariate transformations (6.11). Stein (1956) noted that the least favorable direction  $\hat{\mu}$  transforms in the obvious way under (6.11),  $\hat{\mu} = \mathbf{D}\hat{\mu}$ , where  $\mathbf{D}$  is the matrix with  $ij$ th element  $\partial\hat{\eta}_j/\partial\eta_i|_{\eta=\hat{\eta}}$ , from which it is easy to check that formula (6.5) is invariant: the constant  $a$  is assigned the same value no matter what transformations (6.11) are applied. The bootstrap distribution  $\hat{G}$  is similarly invariant, as shown in Efron (1985), and so is  $z_0$ . This implies that the  $BC_a$  intervals are invariant under transformations (6.11).

*Remark D.* The multiparametric theory of Section 5 gives an interesting result when applied to location-scale families;  $y = (x, s)$ ,  $\eta = (\theta, \sigma)$ , and family of densities  $f_\eta(y)$  of the form

$$f_{\theta,\sigma}(x, s) = (1/\sigma^2)f_{01}((x - \theta)/\sigma, s/\sigma), \tag{11.2}$$

$f_{01}(x, s)$  being a known bivariate density function.

Suppose that we wish to set a confidence interval for the location parameter  $\theta$  on the basis of its MLE  $\hat{\theta}$ . Parametric bootstrap intervals are based on the distribution of  $\hat{\theta}^*$  when sampling from  $f_{\theta,\sigma}(x^*, s^*)$ . The BC interval essentially amounts to pretending that  $\sigma$  is known (and equal to  $\hat{\sigma}$ ) in (11.2) and that we have only a location problem to deal with, rather than a location-scale problem. In contrast, the  $BC_a$  interval takes account of the fact that  $\sigma$  is unknown. In particular the least favorable direction  $\hat{\mu}$ , plotted in the  $(\theta, \sigma)$  plane, is *not* parallel to the  $\theta$  axis. It has a component in the  $\sigma$  direction, whose magnitude is determined by the correlation between  $x$  and  $s$ . This means that Stein's least favorable family (6.4) does not treat  $\sigma$  as a constant.

Table 6 relates to the following choice of  $f_{01}(x, s)$ :

$$x \sim \chi_{30}^2/30 - 1, \quad s | x \sim (1 + x)(\chi_{14}^2/14)^{1/2}, \tag{11.3}$$

the two  $\chi^2$  variates being independent. This is a computationally more tractable version of the problem discussed in Efron (1982, tables 4 and 5). Approximate central 90% intervals are given for  $\theta$ , having observed  $(x, s) = (0, 1)$ . For any other observed  $(x, s)$  the intervals transform in the obvious way,  $\theta_{xs}[\alpha] = x + s\theta_{01}[\alpha]$ . Line 3 of Table 6 shows the exact interval, based on inverting the distribution of the pivotal quantity  $T = (\hat{\theta} - \theta)/\hat{\sigma}$  for situations (11.2), (11.3).

In this case the  $BC_a$  method makes a large "second-order  $t$  correction," as in Example 2 of Section 6, shifting the BC interval a considerable way rightward and achieving the correct R/L ratio. The length of the  $BC_a$  interval is 90% the length of the  $T$  interval. This deficiency is a third-order effect, in the spirit of the familiar Student- $t$  correction. It arises from the variability of  $\hat{\sigma}$  as an estimate of  $\sigma$ , rather than the second-order effect due to the correlation of  $\hat{\sigma}$  with  $\hat{\theta}$ .

*Remark E.* Section 3 says that the family  $\hat{\theta} \sim \theta\chi_{19}^2/19$  can be mapped into form (3.6), (3.7). What are the appropriate mappings? It simplifies the problem to consider the equivalent family  $\hat{\theta} \sim \theta(\chi_{19}^2/c_0)$ , where  $c_0 = 18.3337 = \text{median}(\chi_{19}^2)$ . Then  $\hat{\zeta} \equiv g_1(\hat{\theta})$ ,  $\zeta \equiv g_1(\theta)$ , and  $W \equiv g_1(\chi_{19}^2/c_0)$  give a translation family (3.12), with  $\text{median}(W)$

Table 6. Central 90% Intervals for  $\theta$ , Having Observed  $(x, s) = (0, 1)$  From the Location-Scale Family (11.2), (11.3) so  $\hat{\theta} = 0$  and  $\hat{\sigma} = .966$

		RL	Length
1. BC interval	[-.336, .501]	1.49	.837
2. $BC_a$ interval	[-.303, .603]	1.99	.906
3. $T$ interval	[-.336, .670]	1.99	1.006

NOTE: Line 3 is based on the actual distribution of the pivotal quantity  $T = (\hat{\theta} - \theta)/\hat{\sigma}$ .

= 0, for any mapping  $g_1(t) = (\log t)/c_1$ . Choosing  $c_1 = .3292$  results in  $W = q(Z)$  having  $q(0) = 0, q'(0) = 1$ , as in the discussion of translation families in Section 10.

Section 10 suggests normalizing a translation family by  $g_A(t) = (\exp(At) - 1)/A$ , a good choice for  $A$  being the constant  $\varepsilon_\theta$ , (10.1), which equals .1090 for all  $\theta$  in the family  $\hat{\theta} \sim \theta(\chi^2_{19}/c_0)$ . The combined transformation  $g(t) = g_A(g_1(t))$  is  $g(t) = 9.1746[t^{.3311} - 1]$ . The transformed family  $\hat{\phi} = g(\hat{\theta}), \phi = g(\theta)$  is of form (3.6), (3.7),

$$\begin{aligned} \hat{\phi} &= \phi + (1 + .1090 \cdot \phi)Z, \\ Z &= 9.1746[(\chi^2_{19}/c_0)^{.3311} - 1]. \end{aligned} \tag{11.4}$$

Numerical calculations verify that  $Z$  as defined in (11.4) is very close to a standard normal variate. In fact we have automatically recovered, nearly, the Wilson–Hilferty cube root transformation (Johnson and Kotz 1970). Using (11.4), it is not difficult to show that  $g(t)$ , as defined previously, gives approximately (3.6), (3.7) when applied to the family  $\hat{\theta} \sim \theta(\chi^2_{19}/19)$  considered in Section 3, with constants  $z_0$  and  $a$  as stated. Schenker (1985) gave almost the same result.

*Remark F.* Suppose that  $\mathbf{y} = (x_1, x_2, \dots, x_n)$ , where the  $x_i$  are an iid sample from a regular one-parameter family  $f_\theta(x_i)$ , and that  $\hat{\theta}(\mathbf{y})$  is a first-order efficient estimator of  $\theta$ , like the MLE. The score function  $\dot{l}_\theta$  appearing in (4.4) is that based just on  $\hat{\theta}$ , rather than the score function based on the entire data set  $\mathbf{y}$ . However, it is easy to show from considerations like those in Efron (1975) that the two score functions are asymptotically identical. Their skewnesses differ only by amount  $O_p(n^{-1})$ . It is often more convenient to calculate  $a$  from the score function for  $\mathbf{y}$  rather than for  $\hat{\theta}$ , as was done, for example, in (6.5).

*Remark G.* McCullagh (1984) and Cox (1980) gave an interesting approximate confidence interval for  $\theta$ , which for the simple case  $\hat{\theta} \sim f_\theta$  has endpoint

$$\begin{aligned} \theta_{APP}[\alpha] &= \hat{\theta} + 1/\sqrt{\hat{k}_2} \\ &\times \left\{ z^{(\alpha)} + \frac{(3\hat{k}'_2 + 2\hat{k}_{001}) + \hat{k}_{001}z^{(\alpha)^2}}{6\hat{k}_2^{3/2}} \right\}. \end{aligned} \tag{11.5}$$

Here  $\hat{\theta}$  is the MLE of  $\theta$ ; if  $k_2(\theta) = E_\theta \dot{l}_\theta^2$ , the Fisher information, then  $\hat{k}_2 = k_2(\hat{\theta})$  and  $\hat{k}'_2 = dk_2(\theta)/d\theta|_{\theta=\hat{\theta}}$ ; and  $\hat{k}_{001} = (E_\theta \dot{l}_\theta^3)_{\theta=\hat{\theta}}$ . Formula (11.5) is based on higher-order asymptotic approximations to the distribution of the MLE (see also Barndorff-Nielsen 1984).

It can be shown, as indicated in Section 12, that  $\theta_{BC_a}[\alpha]$  also closely matches (11.5),  $(\theta_{BC_a}[\alpha] - \theta_{APP}[\alpha])/\hat{\sigma} = O_p(n^{-1})$ . We see again that the  $BC_a$  method offers a way to avoid theoretical effort, at the expense of increased numerical computations.

## 12. PROOF OF THEOREM 1

A monotonic mapping  $\hat{\phi} = g(\hat{\theta}), \phi = g(\theta)$  transforms the exact confidence interval in the obvious way,  $\phi_{EX}[\alpha] = g(\theta_{EX}[\alpha])$ ; likewise for the  $BC_a$  interval. By using such a mapping we can always make  $\hat{\phi} = 0$  and the distribution

of  $\hat{\phi}$  given  $\phi = 0$  perfectly normal. Because of (5.3), which says that the distributions of  $\hat{\theta}$  are approaching normality at the usual  $O(n^{-1/2})$  rate, the normalizing transformation  $g$  is asymptotically linear,  $g(\theta) = \theta + c_2\theta^2 + c_3\theta^3 + \dots$ ,  $c_2 = O(n^{-1/2}), c_3 = O(n^{-1})$ .

We will assume that the problem is already in the form  $\hat{\theta} = 0$ , with the cdf of  $\hat{\theta}$  for  $\theta = 0$  normal, say

$$G_0 \sim N(-z_0, 1). \tag{12.1}$$

Here  $z_0 = \Phi^{-1}P_0\{\hat{\theta} < 0\}$  must be included because it is not affected by any monotonic transformations;  $z_0 \doteq \gamma_\theta/6$  is  $O(n^{-1/2})$  by (5.3). A simple exercise, using the mean value theorem of calculus, shows that if (5.4) is true in the transformed problem (12.1), then it is true in the original problem.

Assuming (5.3),  $\hat{\theta} = 0$ , and (12.1), we will show that the exact interval has endpoint

$$\begin{aligned} \theta_{EX}[\alpha] &\doteq \frac{z_0 + z^{(\alpha)}}{1 - \dot{\sigma}_0 z^{(\alpha)} + \dot{\beta}_0 + (\dot{\gamma}_0/6)(z^{(\alpha)^2} - 1)} \\ &+ (\ddot{\sigma}_0/2)(z_0 + z^{(\alpha)})^3, \end{aligned} \tag{12.2}$$

compared with

$$\theta_{BC_a}[\alpha] \doteq \frac{z_0 + z^{(\alpha)}}{1 - \dot{\sigma}_0(z_0 + z^{(\alpha)})} \tag{12.3}$$

for the  $BC_a$  interval. In this section the symbol “ $\doteq$ ” indicates accuracy through  $O(n^{-1})$  or  $O_p(n^{-1})$ , with errors  $O(n^{-3/2})$  or  $O_p(n^{-3/2})$ . Then

$$\begin{aligned} &\frac{\theta_{BC_a}[\alpha] - \theta_{EX}[\alpha]}{\sigma_\theta} \\ &\doteq \theta_{BC_a}[\alpha]\{\dot{\sigma}_0 z_0 + \dot{\beta}_0 + (\dot{\gamma}_0/6)(z^{(\alpha)^2} - 1)\} \\ &- (\ddot{\sigma}_0/2)(z_0 + z^{(\alpha)})^3, \end{aligned} \tag{12.4}$$

which is  $O_p(n^{-1})$ , as claimed in Theorem 1.

The proof of (12.2) begins by noting that (12.1) implies that  $\beta_0 = -z_0, \sigma_0 = 1, \gamma_0 = 0, \delta_0 = 0$ . Then (5.3) gives

$$\begin{aligned} E_\theta \hat{\theta} &= \theta + \beta_\theta \doteq (1 + \dot{\beta}_0)\theta - z_0, \\ \sigma_\theta &\doteq 1 + \dot{\sigma}_0\theta + \ddot{\sigma}_0\theta^2/2, \\ \gamma_\theta &\doteq \dot{\gamma}_0\theta, \quad \delta_\theta \doteq 0, \end{aligned} \tag{12.5}$$

for  $\theta = O(1)$  [i.e., for  $\theta$  a bounded function of  $n$ , in the sequence of situations referred to in (5.3)]. The  $100 \cdot \alpha$  percentile of  $\hat{\theta}$  given  $\theta$  is

$$\begin{aligned} \hat{\theta}_\theta^{(\alpha)} &\doteq (\theta + \beta_\theta) + \sigma_\theta\{z^{(\alpha)} + (\gamma_0/6)(z^{(\alpha)^2} - 1)\} \\ &\doteq [(1 + \dot{\beta}_0)\theta - z_0] + [1 + \dot{\sigma}_0\theta + (\ddot{\sigma}_0/2)\theta^2] \\ &\times [z^{(\alpha)} + (\dot{\gamma}_0\theta/6)(z^{(\alpha)^2} - 1)], \end{aligned} \tag{12.6}$$

using a Cornish–Fisher expansion and (12.5). The  $\theta$ , however, that has  $\hat{\theta}_\theta^{(\alpha)} = 0$  is by definition  $\theta_{EX}[1 - \alpha]$ . Solving the lower expression in (12.6) for 0 and substituting  $1 - \alpha$  for  $\alpha$  gives (12.2).

The proof of (12.3) follows from (3.8), (3.9), and (12.1) [which says that  $\hat{G} \sim N(-z_0, 1)$ ], if we can establish that

$a \doteq \dot{\sigma}_0$ . In fact, we show below that

$$\varepsilon_\theta \doteq \dot{\sigma}_0 \quad \text{for } \theta = O(n^{-1/2}), \quad (12.7)$$

which combines with  $a = \varepsilon_0/(1 + \varepsilon_0 z_0) \doteq \varepsilon_0$  to give the required result.

Formula (12.7) follows from (12.5), which gives the simpler expressions

$$E_\theta \hat{\theta} \doteq \theta - z_0, \quad \sigma_\theta \doteq 1 + \dot{\sigma}_0 \theta, \quad \gamma_\theta \doteq 0, \quad \delta_\theta \doteq 0 \quad (12.8)$$

for  $\theta = O(n^{-1/2})$ . The cdf of  $\hat{\theta}$  given  $\theta$  is calculated to be

$$G_\theta(\hat{\theta}) \doteq \Phi(z_\theta) \dot{z}_\theta - (\dot{\gamma}_0/6)(z_\theta^2 - 1), \quad (12.9)$$

$z_\theta \equiv (\hat{\theta} - \theta - \beta_\theta)/\sigma_\theta$ ,  $\dot{z}_\theta = (\partial/\partial\theta)z_\theta$ . Straightforward expansions give

$$D(z^{(\alpha)}, \theta) \doteq \frac{1 + \dot{\sigma}_0 z^{(\alpha)} + \dot{\beta}_0 + (\dot{\gamma}_0/6)(z^{(\alpha)2} - 1)}{1 + \dot{\beta}_0 - \dot{\gamma}_0/6}, \quad (12.10)$$

from which  $\varepsilon_\theta = (\partial/\partial z)D(z, \theta)|_{z=0} \doteq \dot{\sigma}_0/(1 + \dot{\beta}_0 - \dot{\gamma}_0/6)$ , verifying (12.7), (12.3), and the main result (12.4).

The proof that  $\theta_{BC_a}[\alpha]$  also matches the Cox–McCullagh formula (11.5) is similar to the proof of Theorem 1 and will not be presented here. The main step is an expression for  $\theta_{BC_a}[\alpha]$  involving Lemma 5,

$$\begin{aligned} \theta_{BC_a}[\alpha] \doteq & z^{(\alpha)} + (\hat{k}_3/6\hat{k}_2^{3/2})\{z^{(\alpha)2} + 1\} \\ & + (\hat{k}_3/6\hat{k}_2^{3/2})^2\{2z^{(\alpha)} + z^{(\alpha)3}\}. \end{aligned} \quad (12.11)$$

[Received November 1984. Revised December 1985.]

## REFERENCES

- Abramovitch, L., and Singh, K. (1985), "Edgeworth Corrected Pivotal Statistics and the Bootstrap," *The Annals of Statistics*, 13, 116–132.
- Barndorff-Nielsen, O. E. (1984), "Confidence Limits From  $c|j|\bar{L}$ ," Report 104, University of Aarhus, Dept. of Theoretical Statistics.
- Bartlett, M. S. (1953), "Approximate Confidence Intervals," *Biometrika*, 40, 12–19.
- Beran, R. (1984a), "Bootstrap Methods in Statistics," *Jber. d. Dt. Math. Verein*, 86, 14–30.
- (1984b), "Jackknife Approximations to Bootstrap Estimates," *The Annals of Statistics*, 12, 101–118.
- Bickel, P. J., and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196–1217.
- Cox, D. R. (1980), "Local Ancillarity," *Biometrika*, 67, 279–286.
- DiCiccio, T. J. (1984), "On Parameter Transformations and Interval Estimation," technical report, McMaster University, Dept. of Mathematical Science.
- Efron, B. (1975), "Defining the Curvature of a Statistical Problem (With Applications to Second Order Efficiency)" (with discussion), *The Annals of Statistics*, 3, 1189–1242.
- (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.
- (1981), "Nonparametric Standard Errors and Confidence Intervals" (with discussion), *Canadian Journal of Statistics*, 9, 139–172.
- (1982a), "The Jackknife, the Bootstrap, and Other Resampling Plans," CBMS 38, SIAM-NSF.
- (1982b), "Transformation Theory: How Normal Is a Family of Distributions?," *The Annals of Statistics*, 10, 323–339. (NOTE Corrigenda, *The Annals of Statistics*, 10, 1032.)
- (1984), "Comparing Non-nested Linear Models," *Journal of the American Statistical Association*, 79, 791–803.
- (1985), "Bootstrap Confidence Intervals for a Class of Parametric Problems," *Biometrika*, 72, 45–58.
- Fieller, E. C. (1954), "Some Problems in Interval Estimation," *Journal of the Royal Statistical Society, Ser. B*, 16, 175–183.
- Hall, P. (1983), "Inverting an Edgeworth Expansion," *The Annals of Statistics*, 11, 569–576.
- Hougaard, P. (1982), "Parameterizations of Non-linear Models," *Journal of the Royal Statistical Society, Ser. B*, 44, 244–252.
- Johnson, N. J. (1978), "Modified  $t$  Tests and Confidence Intervals for Asymmetrical Populations," *Journal of the American Statistical Association*, 73, 536–544.
- Johnson, N. L., and Kotz, S. (1970), *Continuous Univariate Distributions—2*, Boston: Houghton-Mifflin.
- Kendall, M., and Stuart, A. (1958), *The Advanced Theory of Statistics*, London: Charles W. Griffin.
- McCullagh, P. (1984), "Local Sufficiency," *Biometrika*, 71, 233–244.
- Schenker, N. (1985), "Qualms About Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, 80, 360–361.
- Singh, K. (1981), "On the Asymptotic Accuracy of Efron's Bootstrap," *The Annals of Statistics*, 9, 1187–1195.
- Stein, C. (1956), "Efficient Nonparametric Testing and Estimation," in *Proceedings of the Third Berkeley Symposium*, Berkeley: University of California Press, pp. 187–196.
- Tukey, J. (1949), "Standard Confidence Points," Memorandum Report 26, unpublished address presented to the Institute of Mathematical Statistics.
- Withers, C. S. (1983), "Expansions for the Distribution and Quantiles of a Regular Functional of the Empirical Distribution With Applications to Nonparametric Confidence Intervals," *The Annals of Statistics*, 11, 577–587.