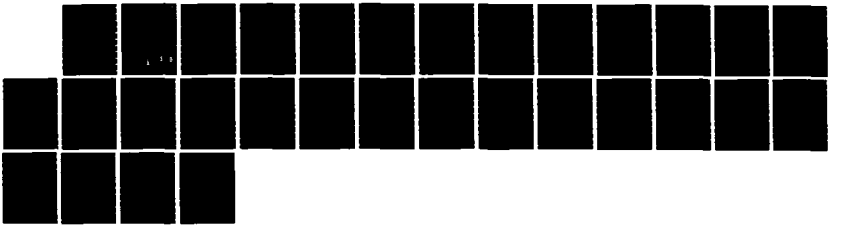
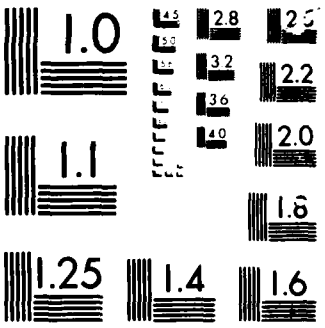
AD-A168 480 BOOTSTRAP CONFIDENCE INTERVALS AND BOOTSTRAP APPROXIMATIONS(U) STANFORD UNIV CA DEPT OF STATISTICS T DICICCO ET AL. 04 JUN 86 TR-375 N00014-86-K-0156 1/1

UNCLASSIFIED F/G 12/1 NL





MICROCOPY CHART

(12)

BOOTSTRAP CONFIDENCE INTERVALS AND BOOTSTRAP APPROXIMATIONS

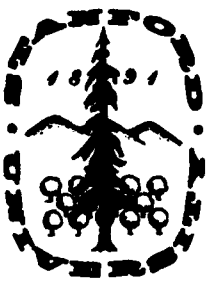
BY

THOMAS DiCICCIO and ROBERT TIBSHIRANI

TECHNICAL REPORT NO. 375
JUNE 4, 1986

PREPARED UNDER CONTRACT
N00014-86-K-0156 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA



12

BOOTSTRAP CONFIDENCE INTERVALS
AND
BOOTSTRAP APPROXIMATIONS

BY

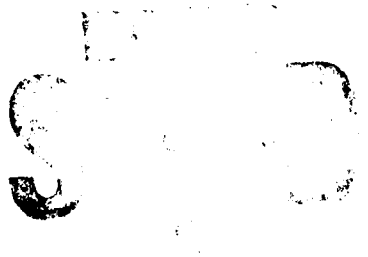
THOMAS DiCICCIO and ROBERT TIBSHIRANI

TECHNICAL REPORT NO. 375
JUNE 4, 1986

Prepared Under Contract
N00014-86-K-0156 (NR-042-267)
For the Office of Naval Research

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA



Bootstrap Confidence Intervals and Bootstrap Approximations

by

Thomas DiCiccio

and

Robert Tibshirani

1. Introduction.

In a recent series of papers ((1981), (1984a), and (1984b)) Bradley Efron has suggested a number of methods for constructing confidence intervals for a real valued parameter θ using the bootstrap. In increasing order of generality, these are the Percentile interval, the Bias Corrected Percentile (BC) interval and the Bias Corrected Percentile Acceleration (BC_a) interval. Each of these intervals is constructed from the bootstrap distribution of a statistic $\hat{\theta}$.

The usual (non-parametric) bootstrap works by sampling from the empirical distribution function $\hat{F}^n$; accordingly, confidence intervals derived from the bootstrap are designed for non-parametric problems. It is difficult, however, to define a "correct" confidence interval in the non-parametric setting and this quantity is needed in order to measure the performance of a confidence interval procedure. Thus to assess the quality of the bootstrap intervals, Efron moves to a different arena, that of one-parameter families. In this setting, one can construct an interval with the desired coverage by inverting the most powerful test at each parameter value. Efron takes this exact interval as the gold standard and considers the parametric versions of the bootstrap intervals, that is, those obtained from the "parametric" bootstrap (sampling from the parametric m.l.e instead of $\hat{F}^n$). Efron shows that the most general of these intervals, the BC_a interval, is second order correct; that is, its endpoints differ from the exact interval by $O_p(1/n)$.

This provides a strong justification for the BC_a interval. Standard confidence intervals of the form

$$(\hat{\theta} + z^{(\alpha)}\,\hat{\sigma}\,,\hat{\theta} + z^{(1-\alpha)}\,\hat{\sigma}\,) \tag{1.1}$$

differ from the exact interval by $O_p(1/n^{1/2})$. (In the above, $\hat{\sigma}$ is an estimate of the standard deviation of $\hat{\theta}$). The $O_p(1/n^{1/2})$ term can cause the exact interval to be asymmetric, an effect picked up by the BC_a interval but not by the standard intervals or by studentized intervals, both of which are symmetric by definition. While Efron does not show that the non-parametric BC_a interval is second order correct, he hypothesizes that given a reasonable definition of this notion, it will be.

Underlying the BC_a interval is a transformation of the problem to a Normal Scaled Translation Family (Efron (1982)) of the form $\theta+(1+a\theta)Z$ where Z is a N(0,1) random variable. Although computation of the BC_a interval doesn't require specification of this transformation, Efron shows that a) if such a transformation exists, the BC_a interval equals the exact interval, and b) the BC_a interval is second order correct in any one parameter problem, so that loosely speaking, to second order, such a transformation always exists.

In this paper we show how to construct this transformation in general. It turns out to be a variance stabilizing transformation followed by a skewness reducing transformation. This construction produces the following benefits: 1) it sheds light on how the BC_a interval works and 2) produces a new interval, (we call it the "$BC_a{}^0$" interval) equal to the BC_a interval (to 2nd order) which can be computed without bootstrap sampling. We also derive from (2) a second order approximation to the bootstrap distribution of the statistic that doesn't require bootstrap sampling. Both the new interval and the approximation require only n+2 evaluations of the statistic. The transformation generalizes the one constructed by Efron (1984b, section 10) for translation families.

The layout of this paper is as follows. In section 2 we concentrate on one parameter problems. We review the BC_a interval and its relation to the exact interval. The $BC_a{}^0$ interval is defined and shown to equal (to second order) the BC_a interval. Some numerical examples are given. In section 3 we discuss confidence intervals for multiparameter problems, and section 4 focusses on the non-parametric problem.

We show how the BC_a^0 interval can be computed without bootstrap sampling and give a number of examples. Section 5 shows how the bootstrap disribution of a statistic can be approximated using the tools developed earlier. Finally, in section 6 we provide proofs of the results quoted throughout.

2. Confidence Intervals for One Parameter Problems.

2.1 The Bootstrap Method

We begin with a statement of the bootstrap method. The notation in this paper will follow that of Efron (1984b) as closely as possible.

Let $y=(x_1,x_2...x_n)$ represent the available data with each x_i assumed to be an independent realization from an unknown probability distribution F_η. Here η is the parameter vector and the parameter of interest is some functional $\theta=t(F_\eta)$. We have a point estimate $\hat{\theta}=t(\hat{F}_\eta)$ where $\hat{F}_\eta$ is some estimate of F_η and would like a confidence interval for θ. The bootstrap method works by resampling from $\hat{F}_\eta$. There are three distinct resampling strategies depending on the choice of $\hat{F}_\eta$:

1) One parameter problems. Here we assume that θ is the only unknown parameter, so that each x_i has distribution F_θ. Resampling is done from $F_{\hat{\theta}}$ where $\hat{\theta}$ is typically the maximum likelihood estimate of θ. This is known as the "parametric bootstrap".

2) Multiparameter problems. We take $\hat{\eta}$ equal to the maximum likelihood estimate of η and resample from $F_{\hat{\eta}}$. This is a multiparameter parametric bootstrap.

3) Non-parametric problems. F_η can be any distribution, so we estimate it by the empirical distribution function $\hat{F}^n$, the non-parametric maximum likelihood estimator of F_η. Resampling from $\hat{F}^n$ is equivalent to sampling with replacement from the original data $x_1,x_2,...x_n$. This is the usual (non-parametric) bootstrap.

2.2 The BC_a Interval.

Efron's BC_a interval uses bootstrap sampling to construct an approximate $1-2\alpha$ confidence interval for θ. Depending on the choice of $\hat{F}_\eta$ in steps a) and b) of the following algorithm, the intervals will apply to situations 1), 2) or 3). The BC_a interval is computed as follows:

a) Bootstrap data sets $y_1^*, y_2^*, \ldots y_B^*$ are created by resampling from $\hat{F}_\eta$.

b) For each y_b^*, b=1,2,...B, the bootstrap estimate $\hat{\theta}_b^* = t(\hat{F}_\eta^*)$ is calculated, where $\hat{F}_\eta^*$ is the estimate of F_η based on y_b^*.

c) The bootstrap distribution of the $\hat{\theta}_b^*$ values is constructed,

$$\hat{G}(s) = \#\{\hat{\theta}_b^* < s\}/B \tag{2.1}$$

d) The bias correction

$$z_0 = \Phi^{-1}(\hat{G}(\hat{\theta})) \tag{2.2}$$

is computed, $\Phi(.)$ being the cdf of the standard normal.

e) The acceleration constant a is computed (details later).

f) The BC_a interval is then given by

$$[\hat{G}^{-1}(\Phi(z[\alpha])), \hat{G}^{-1}(\Phi(z[1-\alpha]))] \tag{2.3}$$

where $z[\alpha] = z_0 + (z_0 + z^{(\alpha)})/(1 - a(z_0 + z^{(\alpha)}))$ and $z^{(\alpha)} = \Phi^{-1}(\alpha)$.

We note that when a=0, (2.3) reduces to Efron's BC (Bias-corrected) percentile interval, and if also $z_0=0$, then (2.3) is simply $[\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)]$, the percentile interval.

For the remainder of this section, we will be discussing the parametric BC_a interval, that is, with $\hat{F}_\eta = F_{\hat{\theta}}$. Sections 3 and 4 will discuss the multiparameter parametric BC_a and the non-parametric BC_a respectively.

Where does the complicated looking formula (2.3) come from? Recall that standard confidence intervals (1.1) are based on the assumption

$$\frac{\hat{\theta}-\theta}{\hat{\sigma}} \sim N(0,1) \tag{2.4}$$

The BC_a interval is based on a more general assumption:

$$g(\hat{\theta})-g(\theta) \sim N(-z_0[1+ag(\theta)]^2) \tag{2.5}$$

where g(.) is a monotone transformation. In (2.4) it is assumed that on the given scale, the standardized statistic is normal with constant variance. In (2.5), we only assume that on some transformed scale, the standardized statistic is normal, possibly with some bias and possibly with a standard deviation changing linearly with the parameter. Efron proves two facts about the BC_a interval:

1) If (2.5) holds for some g(.), then the BC_a interval is correct.

2) For any one parameter problem, the BC_a interval is second order correct. This means roughly that any one parameter problem can be approximately put in form (2.5).

Here's in more detail what's meant by 1) and 2). One can show that if (2.5) holds then the problem can be further transformed into a translation problem. The transformation used is $h(t)=(1/a)\log(1+at)$. The transformed problem is

where

$$\hat{\zeta} = \zeta + W$$

$$\hat{\zeta} = (1/a)\log(1+ag(\hat{\theta}))$$

$$\zeta = (1/a)\log(1+ag(\theta))$$

$$W = (1/a)\log(1+a(Z-z_0)) \tag{2.6}$$

Z being a N(0,1) random variable. On the ζ scale an "exact" interval can be constructed by inverting the pivotal $\hat{\zeta} - \zeta$. Transforming back to the g(.) scale then gives the BC_a interval. This is the meaning of 1). Fact 2) refers to a comparison of the BC_a interval with the exact interval for any one parameter problem. If we are in a one-parameter problem, then the statistic $\hat{\theta}$ has a distribution depending only on θ, say f_θ. Now suppose that the 100(1-α)th percentile of $\hat{\theta}$ as a function of θ, say $\theta(\alpha)$, is a continuously increasing function of θ for any fixed α. Then the usual exact confidence interval (constructed by inverting the size α most powerful test at each θ) is $(\theta_{ex}[\alpha], \theta_{ex}[1-\alpha])$ where $\theta_{ex}[\alpha]$ is the value of θ satisfying $\theta(\alpha)=\theta$. Then Efron shows

$$\frac{\theta_{BCa}[\alpha] - \theta_{ex}[\alpha]}{\hat{\sigma}} = O_p(1/n) \qquad (2.7)$$

where $\theta_{BCa}[\alpha]$ is the endpoint of the BC_a interval. By comparison, the endpoints of the standard interval (1.1) differ from the exact ones by $O_p(n^{-1/2})$.

What makes the BC_a interval attractive is that one doesn't need to know the transformation g(.) to construct the interval! Looking back at (2.3), we see that 3 things are needed: the bootstrap distribution of $\hat{\theta}^*$ ($\hat{G}$), the bias constant z_0 and the acceleration constant a. As mentioned earlier , the bias term z_0 is estimated by $\Phi^{-1}(P(\hat{\theta}^* < \hat{\theta}))$. Note that $P(g(\hat{\theta}^*) < g(\hat{\theta})) = P(\hat{\theta}^* < \hat{\theta})$ for any monotone g(.) so bias is transformation invariant. It turns out that z_0 is typically $O_p(n^{-1/2})$.

We have still to discuss the acceleration constant a. From (2.5) we see that a measures how fast the standard deviation of $g(\hat{\theta})$ is changing with respect to $g(\theta)$. Like z_0, a is typically $O_p(n^{-1/2})$. Efron shows that a can be estimated by

$$a = \frac{SKEW_{\theta=\hat{\theta}}(\dot{l}_\theta(\hat{\theta}))}{6} \qquad (2.8)$$

Here $l_\theta(\theta)=d/\,d\theta\ (\log f_\theta)$ evaluated at $\theta=\hat{\theta}$ and $SKEW_{\theta=\hat{\theta}}(Z)$ represents the skewness of the random variable Z under the distribution governed by $\theta=\hat{\theta}$. As is the case with the other two components, computation of (2.8) doesn't require knowledge of g(.). It can be computed analytically for some simple cases and requires parametric bootstrap calculations in general. Note also that because the likelihood is invariant under monotone reparametrizations so is the right hand side of (2.8).

2.3 Example 1.

Table 1 illustrates the exact, standard and bootstrap confidence intervals for a familiar problem. The data $x_1, x_2, \ldots x_n$ are i.i.d N(0,1). The parameter of interest is $\theta=Var(x_i)$. Level $1-2\alpha$ confidence intervals are to be based on the unbiassed estimate $\hat{\theta} = \Sigma(x_i-\bar{x})^2/(n-1)$. The sample size n was taken to be 20 and $\alpha=.05$. The exact interval is based on inverting the pivotal $\hat{\theta}/\theta$ around its chi-squared (n-1) distribution. The standard interval (line 2) is of the form (1.1) with $\hat{\sigma} = \hat{\theta}\,(2/n)^{1/2}$ the estimated asymptotic standard error of $\hat{\theta}$. The BC_a interval (line 5) is based on formula (2.5). The BC interval (line 4) is based on (2.5) with a equal to 0 and the percentile interval (line 3) has a and z_0 equal to 0. The bootstrapping was performed parametrically, that is, resampling was done from $N(0,\hat{\theta})$. The remaining lines are discussed in section 4 . The lower and upper values in Table 1 refer to averages over 300 monte carlo simulations of the intervals. The level column indicates the proportion of trials in which each interval didn't contain the true value $\theta=1$.

Table 1
Confidence intervals for the variance

		Average Lower	Average Upper	Level (%)
	(1) Exact	.630	1.878	10.0
	(2) Standard	.466	1.531	11.0
Parametric	(3) Percentile	.520	1.585	10.7
	(4) BC	.578	1.670	10.7
	(5) BC_a	.628	1.860	9.7
	(6) BC_a^0	.629	1.877	10.0
Non Parametric	(7) Percentile	.484	1.363	24.3
	(8) BC	.592	1.467	19.3
	(9) BC_a	.617	1.524	19.3
	(10) BC_a^0	.633	1.540	18.7

Of the intervals (1)– (5), only the BC_a interval captures the assymetry of the exact interval. The standard interval (2) undercovers on the right but overcovers on the left so the overall level is about right. This illustrates why coverage alone is not a good way to assess confidence intervals. Efron (1984b) also considers this example and shows that to a high order of approximation one can transform the problem into form (2.5) with z_0=.1082 and a=(1/6)(8/19)$^{1/2}$ = .1081. Hence it is not surprising that the percentile and BC intervals perform poorly because the bias and acceleration components are non-negligible.

Remarks.

a) Efron begins by assuming that only $\hat{\theta}$ has been observed , having density f_θ. Bootstrap values $\hat{\theta}^*$ are generated from $f_{\hat{\theta}}$. We have assumed that a data vector y has been observed but confidence intervals will be based on ly on the m.l.e. $\hat{\theta}$. The two notions are equivalent and it is easy to see that the distribution of $\hat{\theta}^*$ for $y^* \sim F_{\hat{\theta}}$ is $f_{\hat{\theta}}$. By starting with the data vector y , the one-parameter, multi-parameter and non-parametric problems can all be presented in a unified fashion.

b). Let $l_y(\theta)$ be the log likelihood for θ based on y. Then as Efron notes (Remark F), $l_y(\theta)$ could be used in place of $l_{\hat{\theta}}(\theta)$ in the formula for a, for their skewnesses differ by only $O_p(1/n)$. The formula based on $l_y(\theta)$ will sometimes be easier to compute in the one-parameter case and is used in the multi-parameter and non-parametric problems in Sections 3 and 4.

2.4 A different view of the BC_a Interval: the BC_a^0 Interval.

It seems that the computation of the bootstrap distribution G alleviates the need to know g(.), yet the second order correctness of the BC_a interval suggests that a g(.) always exists approximately satisfying (2.5). Indeed this is the case as we will show in this section.

Let $l_y(\theta)$ be the log likelihood for θ based on y. Let $\kappa_2(\theta)=E(d^2 l_y(\theta) / d\theta^2)$ be the expected Fisher information for θ and let $\hat{\sigma}= [\kappa_2(\hat{\theta})]^{-1/2}$. Then the variance stabilizing transformation for $\hat{\theta}$ is $g_1(\hat{\theta})$ where

$$g_1(t) = c\int^t [\kappa_2(u)]^{1/2} du \tag{2.9}$$

Let $g_A(s)=(e^{As}-1)/ A$, *a skewness reducing transformation for strategically chosen A.* And finally let $g(t)=g_A(g_1(t))$. Then the following theorem asserts that this g(.) puts any one parameter problem into approximately form (2.5).

Theorem 2.1

If $\hat{\theta}\sim f_\theta$, and g(t) is as defined above, then with regularity conditions on the derivatives of the log-likelihood,

$$E(g(\hat{\theta}) - g(\theta)) = -z_0 + O(n^{-1})$$

and

$$Var(g(\hat{\theta}) - g(\theta)) = (1 + A\,g(\theta)) + O(n^{-1})$$

Furthermore, if $A= SKEW_{\theta=\hat{\theta}}(l_\theta(\theta))/ 6$, then

$$SKEW(g(\hat{\theta})-g(\theta)) = O(n^{-1})$$

What use is theorem 2.1? For one, it enables us to construct a confidence interval on the original θ scale. For simplicity, choose c in (2.9) so that $g_1(\hat{\theta})=0$ and hence $g(\hat{\theta})=0$. If (2.5) holds, then Efron shows that the endpoints of the correct interval on the g-scale are

$$g(\hat{\theta}) + [1 + a g(\hat{\theta})] \frac{(z_0 + z^{(\alpha)})}{(1-a\,(z_0 + z^{(\alpha)}))} \tag{2.10}$$

which equals $(z_0+z^{(\alpha)})/(1-a(z_0+z^{(\alpha)})$ since $g(\hat{\theta})=0$. The corresponding endpoints on the θ scale are thus

$$g^{-1}\left[\frac{(z_0+z^{(\alpha)})}{1-a(z_0+z^{(\alpha)})}\right] \tag{2.11}$$

We will call this interval the BC_a^0 interval and denote its endpoints by $\theta_{BCa}^0[\alpha]$. Given theorem 2.1, it is not surprising that the endpoints of BC_a^0 and BC_a agree up to $O_p(n^{-1})$.

Theorem 2.2

$$\frac{\theta_{BCa}^0[\alpha]-\theta_{BCa}[\alpha]}{\hat{\sigma}} = O_p(n^{-1})$$

Together with Efron's result (5.4), it also establishes the second order correctness of the BC_a^0 interval.

Note that the BC_a^0 interval, like the BC_a interval, maps in the obvious way under reparametrization because the variance stabilizing transformation also maps correctly.

2.5 Example 1 continued.

Line 6 in Table 1 shows the results of the BC_a^0 interval applied to the variance problem. The overall results are very similar to the BC_a numbers and on an individual basis the BC_a^0 and the BC_a intervals were very close. We used the values $z_0=.1082$ and $a=(1/6)(8/19)^{1/2}\approx.1081$ computed analytically by Efron. The transformation $g_1(s)$ works out to $[(n-1)/2]^{1/2}\log(s)$ and hence $g(s)=g_a(g_1(t))=k_1t^c+k_2$ where $c=[(n-1)/2]^{1/2}a = 1/3$. Thus the procedure has reproduced the Wilson-Hifferty cube root transformation. Efron (1984b, Remark E) makes a similar calculation.

2.6. Example 2. The correlation coefficient.

As a second example we consider the correlation coefficient problem discussed in Efron and Hinkley (1977). The data (x_i,y_i) are i.i.d bivariate normal with means 0, variance 1 and correlation θ. We will base central 90% confidence intervals for θ on the m.l.e $\hat{\theta}$. Note that the sample correlation $\rho=\Sigma x_i y_i / (\Sigma x_i^2 \Sigma y_i^2)^{1/2}$ is not the m.l.e. Standard calculations show $a=-(1/3)(\theta(3+\theta^2))/[n^{1/2}(1+\theta^2)^{3/2}]$. We will consider the case $n=15$, $\theta=.9$ for which $a=-.12119$. Table 2 shows the results of 300 monte carlo runs for a number of intervals.

Table 2
Results for correlation coefficient example.

	Average Lower	Average Upper	Level (%)
Standard (based on ρ)	.816	.954	7.0
Standard (based on $\tanh^{-1}(\rho)$)	.757	.958	7.3
Percentile	.761	.930	18.0
BC	.742	.922	23.3
BC_a	.701	.914	29.3
BC_a^0	.763	.931	14.0

The first two intervals are based on the sample correlation coefficient (using the observed Fisher information for the variance). The second interval was obtained by transforming by $\tanh^{-1}$, computing the interval, then transforming back. The bootstrap intervals are all based on $\hat{\theta}$ and parametric bootstrap sampling. The variance stabilizing transformation turns out to be

$$g_1(\theta) = n^{1/2}\{\tanh^{-1}[2^{1/2}\theta/(1+\theta^2)^{1/2}] - \tanh^{-1}[\theta/(1+\theta^2)^{1/2}]\} \qquad (2.12)$$

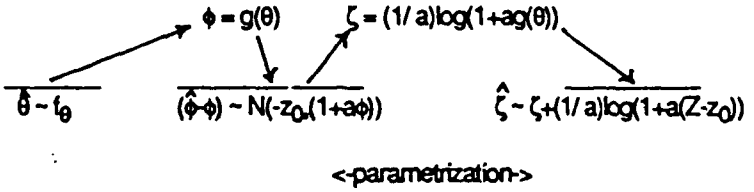
The results are surprising. The BC and BC_a intervals seem to pull percentile interval in the wrong direction and hence the coverage gets worse. The BC_a^0 interval performs quite well and seems to agree with the interval based on the $\tanh^{-1}$ transformation.

2.7 More on the transformations.

Recall the discussion of the BC_a interval in section 2.2. A monotone transformation g(.) that mapped the problem into the form $g(\hat{\theta})-g(\theta) \sim N(-z_0(1+ag(\theta))^2)$ was assumed to exist. Let $\hat{\phi}=g(\hat{\theta})$ and $\phi=g(\theta)$. Once the problem was mapped to the ϕ scale, the transformation (1/ a) log(1+at) was used to further map the problem into a translation family and thereby obtain an exact confidence interval. The two transformations were then inverted to produce the desired interval on the θ scale. This is summarized in Figure 1.

Figure 1.

Transformations Implicitly used by the BC_a interval



The BC_a procedure automatically achieves this working only on the θ scale with no knowledge of g(.). The BC_a^0 interval, on the other hand, gives an explicit construction for g(.), namely $g(t)=g_1(g_a(t))$ where $g_1(t)=c\int^t [\kappa_2(u)]^{1/2}du$ and $g_a(t)=(e^{at}-1)/a$. Notice that the transformation $(e^{at}-1)/a$ is just the inverse of the transformation (1/ a)log(1+at). Hence we have a simpler description of the intervals: the transformation $g_1(t)$ is used to map the problem into the translation form $\hat{\zeta}=\zeta+(1/a)\log(1+a(Z-z_0))$. The BC_a^0 procedure computes $g_1(t)$ explicitly while the the BC_a procedure avoids computation of $g_1(t)$ through use of the bootstrap distribution $\hat{G}$.

3. Confidence intervals in multiparameter problems.

In section 2 we concentrated on one-parameter problems although early on we discussed the multiparameter parametric bootstrap. Here we will briefly describe the extension of the BC_a and BC_a^0 intervals to multiparameter problems. The main purpose of the discussion will be to provide a framework for the non-parametric problem addressed in the next section.

Suppose that our unknown probability mechanism is F_η where η is a k dimensional parameter. Denote the (real-valued) parameter of interest by $\theta=t(\eta)$. In order to apply the confidence interval procedures of section 2, we must first reduce the problem to a one-parameter problem. We will follow Efron and utilize Stein's least favourable family for this purpose.

Denote the density of F_η by f_η and let the m.l.e of η be $\hat{\eta}$. Let I_η be the k by k matrix with ijth entry $-(d^2 / d\eta_i d\eta_j) \log f_\eta$ evaluated at $\eta=\hat{\eta}$. Let $\hat{\nabla}$ be the gradient vector of $\theta=t(\eta)$ evaluated at $\hat{\eta}$, $\hat{\nabla}_i=(d/ d\eta_i)\, t(\eta)|_{\eta=\hat{\eta}}$. The least favourable direction through η is defined to be

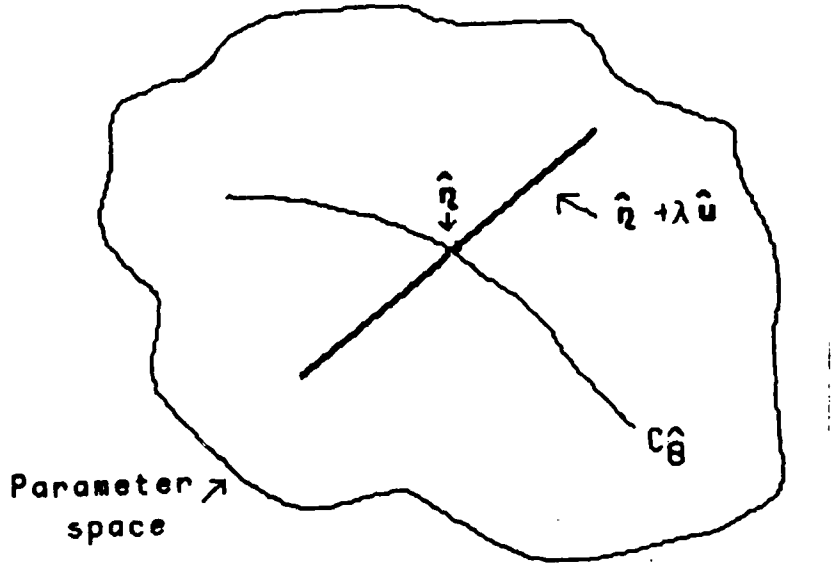
$$\hat{\mu} = (I_{\hat{\eta}})^{-1} \hat{\nabla} \tag{3.1}$$

The least favourable family $\hat{F}$ is the one-dimensional subfamily of F_η passing through $\hat{\eta}$ in the direction $\hat{\mu}$:

$$\hat{F}: f_{\hat{\eta}+\lambda\hat{\mu}} \tag{3.2}$$

Note that $\hat{\eta}$ and $\hat{\mu}$ are fixed, and λ is the parameter of the family. Why is this family called least favourable? Roughly speaking, this family points in the direction that θ is changing fastest in the information metric $(I_{\hat{\eta}})^{-1}$. More formally, consider estimation of $\theta(\lambda)=t(\hat{\eta}+\lambda\hat{\mu})$ in the family $f_{\hat{\eta}+\lambda\hat{\mu}}$. One can show that observed Fisher information for $\theta(\lambda)$ in this problem is the same as that for $\hat{\theta}=t(\hat{\eta})$ in the original k dimensional problem. Furthermore, any other subfamily has a greater Fisher information for θ. In this asymptotic sense the reduction of the full family to the least favourable family is the only reduction in which estimation of θ is not made artificially easier. Figure 2 illustrates the least favourable family.

Figure 2.
Stein's least favourable family
$\hat{\eta}$= m.l.e, $\hat{\theta}= t(\hat{\eta})$, $C_\theta = \{\eta \mid t(\eta)=\theta\}$,
the level surface of constant θ



Tibshirani and Wasserman (1985) and Diciccio and Tibshirani (1985) show that the least favourable family passes through $\hat{\eta}$ in the same direction as the profile likelihood and also that the two families differ by only $O_p(1/n)$.

Given this reduction we can now apply the BC_a method, acting as if our problem is the one parameter problem $f_{\hat{\eta}+\lambda\hat{u}}$. The algorithm of section 2.2 can be used with resampling performed parametrically from the m.l.e $F_{\hat{\eta}}$ (corresponding to the one dimensional m.l.e $\lambda=0$). The bias constant z_0 is estimated by $\Phi^{-1}(\hat{G}(\hat{\theta}))$ as before. The acceleration constant a will be different than before, however; it will involve the skewness of the log-likelihood in the least favourable family:

$$a = \frac{SKEW_{\lambda=0}(d/d\lambda)[\log f_{\hat{\eta}+\lambda\hat{u}}]}{6} \qquad (3.3)$$

Except for some simple cases, estimation of a will require bootstrap computations. Fortunately, an explicit formula for a will be available in the non-parametric case (next section).

The BC_a^0 method can also be used in this setting. Its definition is much the same as before. Here we use $g_1(t) = c\int^t [\kappa_2^{\lambda}(u)]^{1/2}du$, where $\kappa_2^{\lambda}(u)$ is the expected Fisher information for λ in the family $f_{\hat{\eta}+\lambda\hat{\mu}}$ and $g_a(t)=(e^{at}-1)/a$ as before. Using formula (3.3) for a and $z_0 = \Phi^{-1}(\hat{G}(\hat{\theta}))$ we obtain an interval (λ_L, λ_U) for λ. Finally this gives an interval for θ through the relationship $\theta(\lambda)=t(\hat{\eta}+\lambda\hat{\mu})$. Note that $g_1(t)$ will be difficult to calculate in general but like a, it is easily computed in the non-parametric case.

We have constructed the BC_a and BC_a^0 intervals for multiparameter problems by extending the one-parameter definition to the least favourable family. To justify their use we need to show that in some sense they are second order correct. It turns out that a "correct" interval is difficult to define; instead, we can resort to the weaker requirement that each of the intervals err in their coverage only by $O_p(1/n)$. Formally,

$$\text{Prob}_{\eta}(\theta_{BCa}[\alpha] < \theta < \theta_{BCa}[1-\alpha]) = 1-2\alpha + O_p(1/n) \tag{3.4}$$

and similarly for $\theta_{BCa}^0[\alpha]$. We conjecture this result and also

$$\frac{\theta_{BCa}^0[\alpha] - \theta_{BCa}[\alpha]}{\hat{\sigma}} = O_p(n^{-1}) \tag{3.5}$$

but so far we have been unable to proof these conjectures.

4. Non-parametric problems.

If we were to approach the non-parametric problem in its most general form we would have to consider all possible distributions F_η, that is, let η be *infinite dimensional*. This would obviously be infeasible. Following Efron, we simplify the problem substantially by assuming that F_η has support only on the observed data $x_1, x_2, \ldots x_n$. This makes the problem finite dimensional and the approach of section 3 can be used.

Consider the data $x_1, x_2, \ldots x_n$ to be fixed and let $\eta_i = \log(\text{Prob}(X=x_i))$, $i=1,2,\ldots n$. We can describe any realization from F_η by P^* where $P_i^* = \#\{X^*_k = x_i\}/n$. Then $F_{\hat{\eta}}$ is a rescaled multinomial distribution, that is $P^* \sim \text{Mult}(n, e^{\hat{\eta}})/n$. The observed sample gives rise to $\eta = \log(P^0)$ where $P^0 = (1/n, 1/n, \ldots 1/n)^t$ and hence $F_{\hat{\eta}} = \text{Mult}(n, P^0)/n$. The least favourable family through η turns out to be $P^* \sim \text{Mult}(n, w^\lambda)/n$, where $w_i^\lambda = e^{\lambda U_i} / \sum e^{\lambda U_j}$ and

$$U_i = \lim_{\varepsilon \to 0} \frac{t((1-\varepsilon)\hat{F}^n + \varepsilon\delta_i) - t(\hat{F}^n)}{\varepsilon} \tag{4.1}$$

(See Efron 1984b, section 7). Here δ_i is a point mass at x_i and the U_i are called the empirical influence components of $\hat{\theta} = t(\hat{F}^n)$.

We now have almost all we need to compute the BC_a interval for the non-parametric case. Resampling is done from $F_{\hat{\eta}} = \text{Mult}(n, P^0)/n$ and this is equivalent to sampling with replacement from $x_1, x_2, \ldots x_n$. The bias constant is estimated as $\Phi^{-1}(\hat{G}(\hat{\theta}))$ as before. We require only an estimate of the acceleration a. Applying formula (3.3) to the multinomial family gives

$$a = \frac{\sum U_i^3}{(\sum U_i^2)^{3/2}} \tag{4.2}$$

Table 1 line 9 shows the results of the non-parametric BC_a interval applied to the variance problem. It outperforms the (non-parametric) percentile and bias-corrected percentile intervals but doesn't fully capture the assymetry of the exact interval. This is due to the short tails of the bootstrap distribution of $\hat{\theta}$.

The BC_a^0 interval can also be used here. The transformation $g_1(t)=c\int^t [\kappa_2^{\lambda}(s)]^{1/2}ds$ requires an estimate of the expected Fisher information $\kappa_2^{\lambda}(s)$ for the multinomial subfamily (4.1). Straightforward calculations show that

$$\kappa_2^{\lambda}(s) = n[\sum U_i^2 e^{U_i s}/\sum e^{U_i s} - (\sum U_i e^{U_i s}/\sum e^{U_i s})^2] \qquad (4.3)$$

A simple numerical integration (like the trapezoid rule) can then be used to compute $g_1(t)$. Note that $k_2^{\lambda}(s)$ is a non-negative function by Jensen's inequality and is in fact positive unless all the U_i's are equal. Hence $g_1(t)$ will be monotone increasing and invertible.

Line 10 of Table 1 shows the results of the BC_a^0 procedure applied to the variance problem. As in the parametric case the results were very similar on an interval to interval basis to the BC_a results.

Actually, computation of the BC_a^0 intervals doesn't even require bootstrap sampling! The only component of the procedure that seems to require it is the estimation of z_0. But Efron (1984b section 7) provides an approximation for z_0 based on first and second order empirical influences. Let V be the n by n matrix of second order influences, define $z_{01}=(1/6)\sum U_i^3/[\sum U_i^2]^{3/2}$ (the approximation for a) and let $z_{02}=[U^tVU/||U||^2 - \text{trace}(V)]/2n||U||^2$. Then a good approximation for z_0 is

$$z_0 = \Phi^{-1}(2\Phi(z_{01})\Phi(z_{02})) \qquad (4.4)$$

Using the following method due to Tim Hesterberg of Stanford, z_{02} can be computed with only 2 additional evaluations of the statistic.Let U(i,ε) equal the expression in the right hand side of (4.1) for some small positive ε. Let D(i,ε) =U(i,ε) - U(ε) where U(ε) is the mean of the U(i,ε) 's. It is easy to show that $\text{trace}(V)=\varepsilon^2\Sigma\ U(i,\varepsilon)$. Using the notation $\theta(P^*)$ to denote $\theta=t(F)$ evaluated for the distribution F putting mass P_i^* on x_i (see e.g. Efron 1981), one can also show that $U^tVU = [\theta(P^0+\varepsilon U) - \theta(P^0-\varepsilon U) - 2\theta(P^0)]/\varepsilon^2$.

Thus a total of n+2 evaluations of the statistic are required to compute a and z_0. Note however that (4.4) is only an approximation; Hesterberg is presently studying its accuracy.

If the BC_a and BC_a^0 intervals can be shown to be second order correct, then they will also be second order correct in the non-parametric setting, if it is assumed that the number of categories in the support of the multinomial stays fixed as n goes to infinity. Combined with the assumption that the support of the distribution is confined to $x_1, x_2, \ldots x_n$, this is a less than ideal definition on "non-parametric second order correctness". We are currently looking at ways of making it more realistic.

Example 3. The Proportional Hazards model.

For illustration we applied these methods to the proportional hazards model of Cox (1972). The data we chose was mouse leukemia data analysed by Cox in that paper. It consists of the survival times (y_i) in weeks of mice in two groups (x_i), control (0) and treatment (1), as well as a censoring indicator (δ_i). The partial likelihood estimator $\hat{\beta}$ was 1.51. We applied the confidence interval procedures by considering (y_i, x_i, δ_i) as the sampling unit. Estimation of the BC_a^0 interval requires writing the statistic as a functional statistic— not necessary for the BC interval because it only evaluates the statistic on bootstrap samples. We define the partial likelihood estimator for sample weights w, $\beta(w)$, as the maximizer of

$$PL(w) = \prod_{i \in D} \exp(\Sigma \beta x_j w_j) / (\Sigma_{j \in R_i} w_j \exp(x_j \beta) \Sigma w_k) \qquad (4.5)$$

where D is the set indices of the failure times, R_i is the set of indices of the items at risk before the ith failure and each of the sums is over the items failing at the ith failure time. This definition is found in Tibshirani (1984). Finally, U and V were computed by substituting $\varepsilon = 1/(n+1)$ into their definitions. Table 3 shows the results of the various non-parametric confidence procedures.

Table 3
Confidence intervals for
Proportional hazards example

Standard	(.84, 2.18)
Percentile	(.93, 2.34)
BC	(.95, 2.36)
BC_a	(.75, 2.15)
BC_a^0	(.87, 2.03)

Interestingly, the percentile and BC intervals shifted the standard interval to the right, but the negative acceleration (a=-.152) caused the BC_a and BC_a^0 intervals to shiftback to the left. The BC_a^0 is also somewhat shorter than the BC_a interval.

5. Approximating the bootstrap distribution of a statistic.

The results of sections 2 and 3 show (and conjecture) respectively, that

$$\hat{G}^{-1}(z[\alpha]) \qquad (z[\alpha]=z_0+ (z_0+z^{(\alpha)})/(1-a(z_0+z^{(\alpha)}))$$

and

$$g^{-1}[(z_0+z^{(\alpha)})/(1-a(z_0+z^{(\alpha)}))] \tag{5.1}$$

differ by only $O_p(n^{-1})$. We can use this to estimate $\hat{G}^{-1}(p)$ (for any p), without bootstrap sampling, as follows. First we find $z^{(\alpha)}$ such that $p=z[\alpha]$, i.e. $z^{(\alpha)}=p/(1+ap) -z_0$. Then we substitute this into (5.1) and thus get an approximation to $\hat{G}^{-1}(p)$.

If instead we want a density that closely approximates the bootstrap histogram, we recall that $\hat{g(\theta)}=g(\theta)+a(Z-z_0)$ where Z is a N(0,1) random variable. Hence a good approximating density is the density of $g^{-1}(g(\theta)+a(Z-z_0))$. After a little algebra this can be expressed as

$$j(s) = \psi[(e^{g_1(s)a}-1)/a + z_0]\, e^{g_1(s)a}\, (k_2(s))^{1/2} \tag{5.2}$$

where ψ is the density function of N(0,1). In the non-parametric case, (5.2) gives the density of λ and must be multiplied by $d\lambda/d\theta = n/k_2^{\lambda}(s)$ to obtain the density for $\hat{\theta}^*$.

For the Cox model example, Figure 3 shows a histogram of 1000 bootstrap values along with the approximating density j(s) (renormalized) and Table 4 shows the approximation based on (5.2). In both cases the agreement is quite good.

Figure 3
Bootstrap histogram and
approximation based on (5.2)

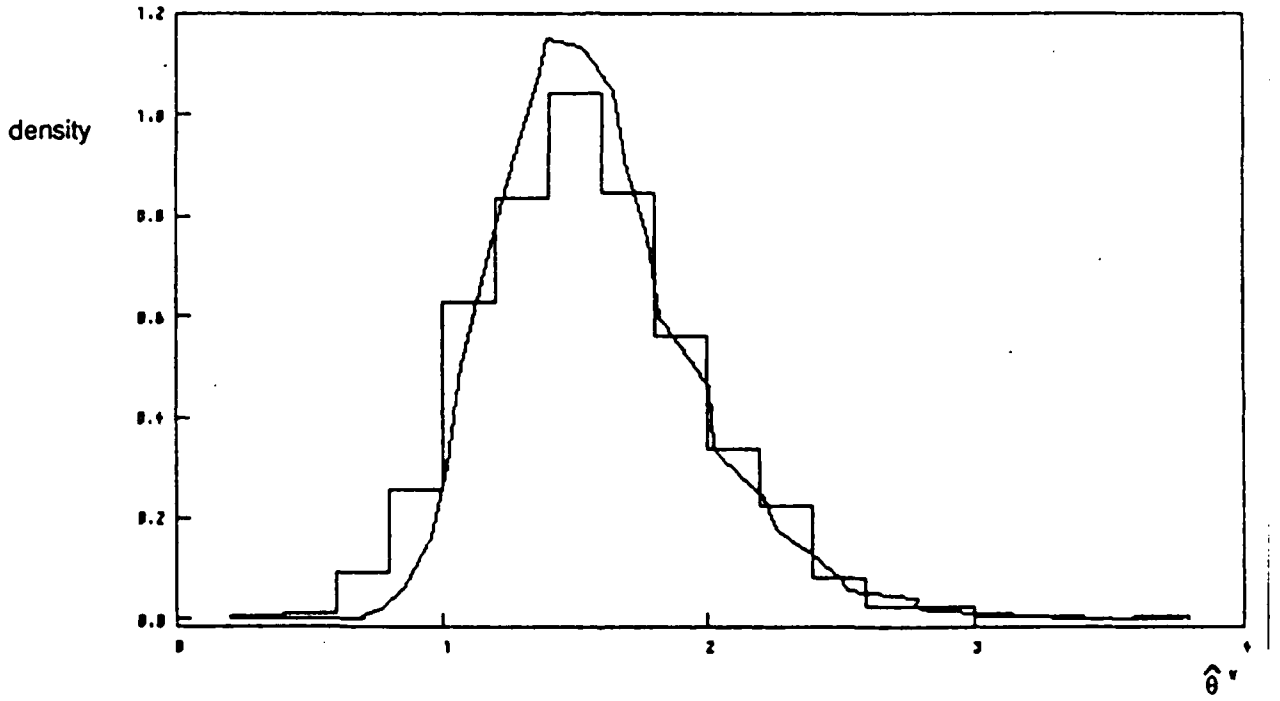


Table 4.
Approximations to $\hat{G}^{-1}(p)$

p	Bootstrap B=1000	Formula (5.1)
.025	0.80	0.86
.05	0.92	0.93
.10	1.04	1.04
.25	1.25	1.27
.50	1.53	1.52
.75	1.80	1.77
.90	2.34	2.24
.975	2.47	2.47

This approximating procedure can be thought of as a refinement of the usual central limit theorem approximation $N(\hat{\theta}, k_2(\hat{\theta})^{-1})$, correct to order $n^{-1/2}$. The new approximation

$$g^{-1}[N(g(\hat{\theta}) - z_0(1 + a g(\hat{\theta}))^2] \tag{5.4}$$

incorporates three order $n^{-1/2}$ components: g(.), z_0 and a. In a parametric setting, (5.2) could prove to be a useful alternative to an edgeworth expansion. It has two distinct advantages over edgeworth expansions: 1) it is always non-negative because g(.) is monotone increasing and 2) it is computable (albeit not often by hand) for general first order efficient statistics $\hat{\theta}$.

The reason that this procedure works in the non-parametric setting is that asymptotically, one has only to look at the bootstrap distribution of $\hat{\theta}^*$ projected onto U in order to compute $\hat{G}(.)$. It is easy to check that a (formula 4.2) equals the skewness of P^{*t} U and that z_0 takes into account both this skewness and the curvature of the level surfaces near P^0.

6. Proofs of theorems 2.1 and 2.2.

Suppose that the parameter θ has been rescaled to be of order $n^{1/2}$ as in Efron's (1984b) expression (4.5). Assume also the regularity conditions in Efron's (4.4). Consider now

$$\phi = g(\theta) = (e^{A\theta} - 1) / A \tag{6.1}$$

where A is understood to be a constant of order $n^{-1/2}$. Then

$$\hat{\phi} - \phi = (e^{A\theta} / A)(e^{A(\hat{\theta}-\theta)} - 1) \tag{6.2}$$

and from the moments of $\theta - \theta$ (see for example Welch 1965) it can be shown that

$$E(\hat{\phi} - \phi) = (1/2) n e^{A\theta} [(2\kappa_{11} + \kappa_{001}) / n^{1/2} + A / \kappa_2 + O(n^{-2})]$$

$$var(\hat{\phi} - \phi) = n e^{2A\theta} (1 / \kappa_2 + O(n^{-2}))$$

$$\gamma_1 (\hat{\phi} - \phi) = (3\kappa_{11} + 2\kappa_{001}) / \kappa_2^{3/2} + 3An^{1/2} / \kappa_2^{1/2} + O(n^{-1})$$

$$\gamma_2 (\hat{\phi} - \phi) = O(n^{-1}) \tag{6.3}$$

where γ_1 and γ_2 skewness and excess in kurtosis and the κ's are as defined in DiCiccio (1984). If the choice

$$A = -(1/3)(3\kappa_{11} + 2\kappa_{001}) / (n^{3/2}) \tag{6.4}$$

is made, then $\gamma_1(\hat{\phi} - \phi)$ is $O_p(n^{-1})$. By the relations attributed to Bartlett, $\kappa_3+3\kappa_{11}+\kappa_{001}=0$ and $\kappa_3=2\kappa_{001}+\kappa_2$, it follows that if θ is the variance stablized parameter with $\kappa_2=1$, then

$$A = (1/6)(\kappa_3 / \kappa_2^{3/2}) = (1/6)(\kappa_3 / n^{3/2}) \tag{6.5}$$

and

$$E(\hat{\phi} - \phi) = -z_0 + O(n^{-1})$$

$$var(\hat{\phi} - \phi) = e^{2A\theta} + O(n^{-1})$$

$$\gamma_1(\hat{\phi} - \phi) = O(n^{-1}) \tag{6.6}$$

Thus $\hat{\phi} - \phi$ is, to second order, normally distributed with mean $-z_0$ and standard deviation $e^{A\theta}=1+A\phi$. Although κ_3 at the true value θ_0 is unknown, $\kappa_3(\theta)$ may be used in its place for the calculation of A, without altering the orders of the preceding error terms. This establishes theorem (2.1). Theorem (2.2) then follows immediately from Efron's (11.3). In fact (11.3) holds exactly for $\theta_{BCa}0[\alpha]$.

Acknowledgements

We would like to thank Larry Wasserman for valuable discussions on profile likelihood and the non-parametric problem, Timothy Hesterberg for his z_0 formula and Bradley Efron whose research and encouragement stimulated this work.

REFERENCES

Cox, D.R. (1972). Regression models and life tables . J. R. Statist. Soc. B, 34, 187-202.

DiCiccio, T. J. (1984) On parameter.transformations and interval estimation. Biometrika 71, 3, 477- 485.

DiCiccio, T.J. and Tibshirani , R. (1985). Likelihood and least favourable families. In preparation.

Efron, B, and Hinkley, D. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with comments). Biometrika 65, 457-487.

Efron, B. (1981). Non-parametric standard errors and confidence intervals. (with discussion). Can. J. Stat. 9, 139- 172.

Efron , B. (1982). Transformation theory: how normal is a one-parameter family of distributions? Annals. Stat. 10, 323-339.

Efron, B. (1984a). Bootstrap confidence intervals for parametric problems. To appear, Biometrika.

Efron , B. (1984b). Better bootstrap confidence intervals. Tech. rep 14, Dept. of Statistics, Stanford University.

Stein, C. (1956). Efficient non-parametric estimation and testing. Proc. 3rd Berkeley Symp. 187-196.

Tibshirani, R. (1984). Local likelihood estimation. Tech. rep 97, Dept. of Statistics, Stanford University.

Tibshirani, R, and Wasserman, L. (1985) . A note on profile likelihood, least favourable families and Kullback-Leibler distance. Dept of Statistics Tech. rep 006, University of Toronto.

Welch, B.L. (1965). On comparisons between confidence point procedures in the case of a single parameter. J .R. Statist. Soc. B, 27, 1-8.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 375	2. GOVT ACCESSION NO. AD-A168480	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Bootstrap Confidence Intervals and Bootstrap Approximations		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL REPORT 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Thomas DiCiccio and Robert Tibshirani		8. CONTRACT OR GRANT NUMBER(s) N00014-86-K-0156
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 1111		12. REPORT DATE June 4, 1986 13. NUMBER OF PAGES 28
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bootstrap Approximations, Bootstrap Confidence Intervals, Scaled Transformation Families.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

PLEASE SEE FOLLOWING PAGE.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

TECHNICAL REPORT NO. 375

20. ABSTRACT

We study the "BC_a" bootstrap procedure (Efron 1984) for constructing parametric and non-parametric confidence intervals. The BC_a interval relies on the existence of a transformation that maps the problem into a "normal scaled transformation family". We show how to construct this tranformation in general. Exploiting this, we derive an interval that equals the BC_a interval to second order, computable without bootstrap sampling. As a further benefit, this construction provides a second order correct approximation to the bootstrap distribution of a statistic, computed without bootstrap sampling. Both the new interval and the approximation require only n+2 evaluations of the statistic, where n is the sample size.

END

DTIC

7-86