*J. R. Statist. Soc.* B (1988)
**50**, *No.* 3, *pp.* 321–337

# Bootstrap Methods

By DAVID V. HINKLEY†

*The University of Texas at Austin, USA*

[*Read before the* Royal Statistical Society *on Wednesday, March 16th, 1988, at a meeting organized by the* Birmingham Group, *Professor* J. B. Copas *in the Chair*]

SUMMARY

A survey of some developments in bootstrap methodology is given. Topics include confidence limits, significance tests, empirical likelihoods, conditioning, double bootstrapping, and numerical techniques. Special attention is given to regression problems. There are brief remarks about more complex problems, including variance component problems, time series and nonparametric regression.

*Keywords*: BALANCED SAMPLES; SADDLEPOINT METHODS; PIVOTS; CONFIDENCE LIMITS; SIGNIFICANCE TESTS; CONDITIONAL INFERENCE; MONTE CARLO METHODS; JACKKNIFE; LIKELIHOOD; PERMUTATION TEST; REGRESSION; VARIANCE COMPONENTS; TIME SERIES; SAMPLE SURVEYS; NONPARAMETRIC METHODS

## 1. INTRODUCTION

The essence of bootstrap methods is the simulation of relevant properties of a statistical procedure with minimal model assumptions. The word 'simulation' here is used in the widest possible sense, from simple substitution of an estimated distribution in a formula to complex Monte Carlo simulation of representative samples and their analysis. In any given context bootstrap methods may be similar variously to simulation methods, permutation methods, jackknife methods or other familiar 'resampling' methods. One major focus of research has been the search for reliable, automatic, empirical methods for calculating confidence limits. Because most bootstrap methods involve numerical approximation, potentially powerful techniques of theoretical and Monte Carlo approximation have been and continue to be studied. As to potential applications, considerable effort has been devoted to classical problems involving means, correlations and regression. But increasingly attention is directed to more complex problems such as those associated with variance components, time series, sample surveys and nonparametric curve fitting.

The aim of the present paper is to review and illustrate many of the developments in bootstrap methodology, so as to highlight key ideas and potential usefulness. The choice of material inevitably reflects personal interests, however, so that the paper is in no way a comprehensive review. The first sections deal with the relatively simple context of homogeneous samples; Sections 2–5 respectively discuss the basic bootstrap method, numerical techniques, confidence limit methods and significance test methods. Regression problems are considered in Section 6, and the idea of a conditional bootstrap introduced there is further discussed in Section 7. Section 8 looks at some recent suggestions for empirical likelihoods. Some more complex applications are outlined in Section 9. Finally, Section 10 contains some general discussion.

† *Address for correspondence*: Department of Mathematics, The University of Texas at Austin, Austin, TX 78712, USA.

 0035–9246/88/50321 $2.00

## 2. BASIC BOOTSTRAP METHOD

To begin with a very simple example, consider the sample of $n = 10$ measurements $x_1, \ldots, x_{10}$ in the first row of Table 1, whose average and standard deviation are $\bar{x} = 17.87$ and $s = 7.19$. Suppose that we wish to make statistical statements about the accuracy of the sample average $\bar{x}$ as an estimate of $\mu$, the mean of $X$ in the population from which the sample was drawn. For the sake of definiteness, suppose that we wish to know (a) the variance of $\bar{X}$, (b) $\Pr\{c \leqslant \bar{X} - \mu \leqslant d\}$ for specified $c$ and $d$, and (c) 95% confidence limits for $\mu$ on either side of $\bar{x}$.

One classical approach would be to describe random variation in sampled $X$ values by a distribution function $F(x \mid \theta) = \Pr\{X \leqslant x\}$, with $\theta$ an unknown parameter (vector or scalar) which includes $\mu$. Possible answers to problems (a)–(c) are found by theoretical calculation based on $F$ with an estimate $\hat{\theta}$ in place of $\theta$. For example, if $F$ is the cdf of the $N(\mu, \sigma^2)$ distribution, so that $\theta = (\mu, \sigma^2)$, then the variance of $\bar{X}$ is $\sigma^2/n$, which we usually calculate with $s^2 = (n-1)^{-1}\Sigma(x_i - \bar{x})^2$ in place of $\sigma^2$. In bootstrap terminology, this is a *parametric bootstrap* calculation.

The *nonparametric bootstrap*, more usually called simply *bootstrap*, approach is to not assume anything about the form of $F$, only that it exists. Then in place of $F(x \mid \hat{\theta})$ one might use the empirical cdf

$$\tilde{F}(x) = n^{-1} \sum hv(x - x_i),$$

TABLE 1

*A random sample and small bootstrap analyses of its mean*†

| *Bootstrap sample* | *Frequencies of datum values for the following data* | | | | | | | | | | $\bar{x}^*$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *9.6* | *10.4* | *13.0* | *15.0* | *16.6* | *17.2* | *17.3* | *21.8* | *24.0* | *33.8* | |
| *Simple bootstrap* | | | | | | | | | | | |
| 1 | 1 | 0 | 0 | 1 | 3 | 1 | 1 | 0 | 2 | 1 | 19.07 |
| 2 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 3 | 2 | 0 | 18.48 |
| 3 | 0 | 0 | 2 | 1 | 2 | 0 | 2 | 0 | 3 | 0 | 18.08 |
| 4 | 1 | 1 | 1 | 2 | 0 | 1 | 1 | 1 | 0 | 2 | 18.69 |
| 5 | 1 | 0 | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 0 | 16.77 |
| 6 | 1 | 1 | 2 | 0 | 0 | 1 | 1 | 2 | 1 | 1 | 18.19 |
| 7 | 0 | 1 | 3 | 1 | 0 | 1 | 3 | 0 | 1 | 0 | 15.75 |
| 8 | 2 | 1 | 0 | 0 | 2 | 1 | 0 | 0 | 2 | 2 | 19.56 |
| 9 | 1 | 1 | 1 | 2 | 0 | 0 | 1 | 1 | 1 | 2 | 19.37 |
| 10 | 0 | 1 | 2 | 0 | 2 | 1 | 0 | 3 | 1 | 0 | 17.62 |
| Sample average of $\bar{x}^*$s = 18.16, sample variance of $\bar{x}^*$s = 1.41 | | | | | | | | | | | |
| *Randomized block bootstrap* | | | | | | | | | | | |
| 1 | 0 | 0 | 1 | 1 | 3 | 1 | 0 | 0 | 2 | 2 | 21.06 |
| 2 | 1 | 3 | 1 | 0 | 1 | 0 | 0 | 1 | 2 | 1 | 17.40 |
| 3 | 2 | 0 | 0 | 0 | 1 | 1 | 0 | 2 | 3 | 1 | 20.24 |
| 4 | 1 | 0 | 1 | 0 | 0 | 3 | 3 | 1 | 0 | 1 | 18.17 |
| 5 | 1 | 2 | 1 | 0 | 2 | 0 | 2 | 2 | 0 | 0 | 15.48 |
| 6 | 0 | 2 | 2 | 0 | 1 | 0 | 1 | 2 | 1 | 1 | 18.21 |
| 7 | 1 | 0 | 0 | 3 | 1 | 3 | 0 | 0 | 1 | 1 | 18.06 |
| 8 | 2 | 1 | 1 | 2 | 0 | 0 | 2 | 0 | 1 | 1 | 16.50 |
| 9 | 2 | 1 | 1 | 1 | 0 | 0 | 2 | 1 | 0 | 2 | 18.16 |
| 10 | 0 | 1 | 2 | 3 | 1 | 2 | 0 | 1 | 0 | 0 | 15.42 |
| Sample average of $\bar{x}^*$s = 17.87, sample variance of $\bar{x}^*$s = 3.33 | | | | | | | | | | | |

†Average $\bar{x} = 17.87$.

where $hv(u) = 0$ $(u < 0)$, $1(u \geqslant 0)$; possibly one would consider a smoothed version of $\tilde{F}$ (Efron, 1982, ch. 5; Silverman and Young, 1987). For problem (a), $\sigma^2$ in the formula $\text{var}(\bar{X}) = \sigma^2/n$ would now be calculated with $\tilde{F}$ in place of $F$ as $\tilde{\sigma}^2 = \int x^2 d\tilde{F}(x) - (\int x d\tilde{F}(x))^2 = n^{-1}\Sigma(x_i - \bar{x})^2$, perhaps modified to its unbiased form $s^2$. There is nothing novel about this, of course, but there is about using $\tilde{F}$ to do the probability calculations for problems (b) and (c).

Consider problem (b) in detail, and rewrite the required probability in the more suggestive form

$$P = \Pr\{c \leqslant \text{mean}(data) - \text{mean}(F) \leqslant d\}. \tag{1}$$

If this is calculated with $\tilde{F}$ substituted for $F$ *everywhere*, the result is the estimate

$$\tilde{P} = \Pr\{c \leqslant \text{mean}(data^*) - \text{mean}(\tilde{F}) \leqslant d\}, \tag{2}$$

where $data^*$ is a random sample of size $n$ drawn from $\tilde{F}$. Because theoretical evaluation of $\tilde{P}$ appears impossible, one might well adopt the strategy of numerical simulation: draw repeated samples $data^*(1), \ldots, data^*(B)$ from $\tilde{F}$, and calculate

$$\tilde{P}_{\text{sim}} = \frac{\text{number of times } c \leqslant \text{mean}(data^*(i)) - \text{mean}(\tilde{F}) \leqslant d}{B}. \tag{3}$$

Table 1 illustrates this for $B = 10$. Each bootstrap sample $data^*(i)$ is recorded in the form of frequencies of original data values. For $c = -1$ and $d = +1$ we get $\tilde{P}_{\text{sim}} = 0.50$, a not very accurate approximation to $\tilde{P} = 0.37$ (see Section 3) resulting from the ridiculously small value of $B$: it would be customary to have $B$ well in excess of 100.

Note that in the simulation, drawing a random sample from $\tilde{F}$ means simply sampling $n$ values from *data* randomly with replacement. But is this a good numerical strategy? It would not be if we required only $\text{var}(\bar{X})$, because the simpler technique known as the *jackknife* (Miller, 1974; Efron, 1982) uses $n$ systematic samples from *data* and gives the correct answer—here meaning $\tilde{\sigma}^2/n$. Can $\tilde{P}$ itself be calculated without numerical simulation? Such questions are addressed in the next section.

A very different question concerns the accuracy of $\tilde{P}$ as an approximation to, or estimate of, $P$. If $\tilde{P}$ is very inaccurate, then choosing data-dependent values $c = \tilde{c}$ and $d = \tilde{d}$ to make $\tilde{P} = 0.95$, for example, would make the natural 0.95 bootstrap confidence limit formula

$$\text{mean}(data) - \tilde{d} \leqslant \text{mean}(F) \leqslant \text{mean}(data) - \tilde{c} \tag{4}$$

unreliable. We know from experience that this is likely to happen for small samples of, say, normal or gamma data: the reliable confidence limit methods are based on probabilities for $(\bar{x} - \mu)/s$ and $\bar{x}/\mu$ respectively, not $\bar{x} - \mu$. Is there some way of finding out that $\tilde{P}$ is inaccurate? Is there a general, reliable way to calculate confidence limits for $\mu$? To these questions we return in Section 4.

The example of the average illustrates a general type of problem to which considerable theoretical effort has been directed. Given a statistical estimate $T = t(\tilde{F})$ of population characteristic $\theta = t(F)$, we wish to calculate $Q = E\{R_t(F, \tilde{F}) \mid F\}$; here $E(\cdot \mid F)$ denotes the expectation with respect to $F$. The quantity $R_t(F, \tilde{F})$ might be simple, e.g. $(\bar{X} - \mu)^2$, or complicated, e.g. the indicator of whether or not $(\bar{X} - \mu)/S \leqslant a$. The nonparametric bootstrap approximation of $Q$ is $\tilde{Q} = E\{R_t(\tilde{F}, \tilde{F}^*) \mid \tilde{F}\}$, where $\tilde{F}^*$ is the empirical cdf of the bootstrap sample $X_1^*, \ldots, X_n^*$ which is drawn randomly

from $\tilde{F}$. While the consistency of $\tilde{Q}$ for $Q$ flows from the consistency of $\tilde{F}$ for $F$, a more detailed assessment of $\tilde{Q} - Q$ is often useful, especially if one is trying to compare confidence limit procedures or if alternative approximations to $Q$ are being considered. The majority of theoretical results (see Beran (1982, 1984) and Hall (1987a), and references therein) deal either with estimates $T$ which are functions of vector averages, so that standard expansion techniques apply, or with estimates representable by Volterra series,

$$T = \theta + n^{-1} \sum a_1(X_j; F) + n^{-2} \sum\sum a_2(X_i, X_j; F) + \ldots, \tag{5}$$

in which $a_1$ is the influence function of $T$. Some of the relevant results for confidence limit methods are reviewed by DiCiccio and Romano (1988).

In what follows, the discussion focuses first on some of the questions raised in this section, and then reviews a variety of bootstrap methods, in a rather non-technical way. Throughout we shall denote bootstrap samples of data by $X_1^*, \ldots, X_n^*$ and corresponding statistics by $T^*$.

## 3. NUMERICAL TECHNIQUES

The exact calculation of property $Q = E\{R_t(\tilde{F}, \tilde{F}^*) \mid \tilde{F}\}$ is ordinarily not possible. There are essentially two ways to proceed: theoretical approximation and purely numerical approximation.

The simplest type of theoretical approximation would be to replace $T = t(\tilde{F})$ by its linear approximation

$$T_L = t(F) + n^{-1} \sum a_1(X_j; F), \tag{6}$$

i.e. the first two terms on the right of (5). From $T_L$ is derived the $N(0, \tilde{V})$ approximation for the distribution of $n^{1/2}(T - \theta)$, with $\tilde{V} = n^{-1} \Sigma\{a_1(X_j; \tilde{F})\}^2$. This is the (infinitesimal) jackknife method, which may often be adequate, but which negates a potential advantage of bootstrap methods, namely high order or small sample accuracy.

The simplest example of numerical approximation, illustrated by (3), is the generation of $B$ samples $x_b^*$, $b = 1, \ldots, B$, from $\tilde{F}$ followed by calculation of

$$\tilde{P}_{\text{sim}} = B^{-1} \sum_{b=1}^{B} R_t(\tilde{F}, \tilde{F}_b^*).$$

The required magnitude of $B$ will depend on the form of $R_t$, but will often be at least 100.

A general discussion of improvement in numerical techniques by Thernau (1983) suggests several approaches, including the importance sampling and control methods familiar in Monte Carlo methodology. The different approach of balanced sampling has been studied in more detail (Obgonmwan and Wynn, 1986; Davison *et al.*, 1987; Graham *et al.*, 1987). The central idea here can be expressed in two ways, the more profitable of which is as follows. Write a simulated sample from $\tilde{F}$ as $(x_{\xi(1)}, \ldots, x_{\xi(n)})$, and $\xi = (\xi(1), \ldots, \xi(n))$. Then the $B$ vectors $\xi_1, \ldots, \xi_B$ which define the bootstrap simulation should cover the $n$-dimensional lattice cube $\{1, 2, \ldots, n\}^n$ in as uniform a manner as possible. Exact uniformity on one- and two-dimensional margins is achievable by use of classical experimental designs. For example, one-dimensional balance is achieved if the $B \times n$ matrix with $(b, i)$th element $\xi_b(i)$ defines a randomized block design with columns as blocks, entries as treatment labels. The second half of

Table 1 illustrates this with $B = 10$, corresponding to a single randomized block. Note that the average of the 10 $\bar{x}^*$s is necessarily equal to $\bar{x}$, thereby yielding a correct estimate of zero bias for $\bar{X}$:

$$\text{estimated bias} = B^{-1} \sum (\bar{x}_b^* - \bar{x}) = 0.$$

Also the variance of the $\bar{x}^*$s is closer to the correct value $n^{-1}\tilde{\sigma}^2$ for the variance of $\bar{X}$.

Two-dimensional balance can be achieved using orthogonal Latin squares, and a somewhat weaker form of balance, suitable for homogeneous data, is achievable using balanced incomplete block designs (Graham *et al.*, 1987). What two-dimensional balance gives is error-free approximation of bias and variance for the linear part of a statistic, which for large samples is adequate.

What do such designs achieve in practical terms? Probably a fourfold or fivefold reduction in $B$ for any given level of simulation error, if we are approximating moments of $T$. But for estimating the $100p$th percentile, say, of $T - \theta$ by the $(B+1)p$th ordered value of $T^* - T$, balanced designs are not so effective, especially for $p < 0.05$ or $p > 0.95$. It seems quite likely that a more effective strategy is to select among one-dimensional balanced designs using a rejection technique along the lines suggested by Ogbonmwan and Wynn (1986). Further research is needed in this area.

Switching now to theoretical approximation, particularly for the probability distribution of $T^*$, one elementary approach is to modify normal approximations with Edgeworth corrections. More interesting, and usually more effective, is the use of saddlepoint approximations (Davison and Hinkley, 1988). For example, consider again $T = \bar{X}$, and write the empirical cumulant generating function of $X$ as

$$\tilde{K}(\lambda) = \log \int e^{\lambda x} d\tilde{F}(x) = \log\left( n^{-1} \sum e^{\lambda x_i} \right).$$

Then a direct application of equation (4.9) of Daniels (1987) gives

$$\tilde{P} = \Pr(T^* - t \leqslant y \mid \tilde{F}) \doteq \Phi(w_y) + \phi(w_y)(w_y^{-1} - z_y^{-1}),$$

where

$$w_y = [2n\{\lambda_{t+y}(t+y) - \tilde{K}(\lambda_{t+y})\}]^{1/2} \operatorname{sgn}(\lambda_{t+y}),$$

$$z_y = \lambda_{t+y}\{n\tilde{K}''(\lambda_{t+y})\}^{1/2}$$

with $\lambda_{t+y}$ the unique solution of $\tilde{K}'(\lambda) = t + y$. Table 2 gives a brief summary of numerical results so obtained for the data of Table 1, in the form of percentile approximations. Comparison is made to exact results (simple numerical simulation with $B = 50\,000$) and normal approximation results. The saddlepoint approximation is excellent.

There are two difficulties with the saddlepoint approximation method in this context. First is a technical difficulty associated with the discreteness of $\tilde{F}$; this makes formal proofs of asymptotic expansions complicated, but not impossible. More important is the limited range of problems to which known saddlepoint approximations apply, essentially those for which $T$ solves a linear estimating equation of the form $\Sigma\psi(X_j, T) = 0$ with $\psi(x, t)$ monotone in $t$. An *ad hoc* approximation can be obtained via series expansions of $T^* - T$, but the result does not have the degree of accuracy typical for saddlepoint methods. A key unsolved problem is to derive saddlepoint

TABLE 2

*Approximations to bootstrap percentage points for $\bar{X} - \mu$; data in Table 1*

| | $p$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *0.001* | *0.01* | *0.05* | *0.10* | *0.90* | *0.95* | *0.99* | *0.999* |
| Exact percentile† | −6.34 | −5.55 | −3.34 | −2.69 | 2.87 | 3.73 | 5.47 | 7.52 |
| Saddlepoint percentile | −6.31 | −5.52 | −3.33 | −2.69 | 2.85 | 3.75 | 5.48 | 7.46 |
| Normal percentile | −8.46 | −7.03 | −3.74 | −2.91 | 2.91 | 3.74 | 5.29 | 7.03 |
| Fisher–Cornish | −6.51 | −5.74 | −3.48 | −2.81 | 3.00 | 3.97 | 5.89 | 8.19 |

† From 50 000 random samples.

approximations for non-linear statistics such as $T = n^{-1}\Sigma a(X_j) + n^{-2}\Sigma\Sigma b(X_i, X_j)$: such approximations would give accurate results for statistics with expansion (5).

What of the other possible numerical techniques? The Monte Carlo control method can be applied to approximate moments of a statistic, for example using $T_L$ in (6) as control, since $T_L$ has known moments under sampling from $\tilde{F}$. Use of the Monte Carlo method of importance sampling is currently under investigation by Dr A. C. Davison. For approximation of probabilities, such as (2), one obvious approach is to apply smoothing techniques to the empirical distribution of simulated values of the relevant statistical quantities, such as $\bar{X}^* - \bar{X}$.

## 4. CONFIDENCE LIMIT METHODS

The most studied problem in (nonparametric) bootstrap methodology is the determination of reliable confidence limit procedures. This is the subject of the companion paper by DiCiccio and Romano (1988), so an exhaustive survey will not be attempted here.

The basic problem arises from the discrepancy between (1) and (2). In principle a confidence interval procedure for parameter $\theta$ based on estimate $T$ would be solved by finding $a_P$ such that $\Pr(T - \theta \leqslant a_P) = P$, for given $P$. Then, for example, equitailed $1 - \alpha$ limits for $\theta$ would be $T - a_{1-\alpha/2}$ and $T - a_{\alpha/2}$, cf. (4). Bootstrap estimates $\tilde{a}_P$ are usually not satisfactory, in essence because $T^* - T$ is not pivotal for $\tilde{F}$s within probable range of $F$. A useful analogy is the problem of setting confidence limits for a normal mean, where the $N(0, \tilde{\sigma}^2/n)$ approximation for $\bar{x} - \mu$ would not give a satisfactory confidence distribution for $\mu$ if $n$ were very small. Actually the solution to the latter problem suggests at least one of several possible approximate solutions for the nonparametric bootstrap problem.

One way to construct a reliable confidence limit procedure is to construct an invertible pivot, say $Q(T, \theta, \mathbf{S})$ with $\mathbf{S}$ containing relevant ancillary features. Familiar examples in classical statistics are Student's $t$ statistic for a normal mean, and $\bar{X}/\mu$ for an exponential mean. In the bootstrap context we would require that $Q^* = Q(T^*, T, \mathbf{S}^*)$ be very close to pivotal under sampling from $\tilde{F}$s within probable range of $F$. Analogy with the normal mean problem suggests trying $Q = (T^* - \theta)/S^*$ with $S^*$ a nonparametric estimate of standard error such as is provided by a jackknife method (Miller, 1974; Efron, 1982, ch. 6). In his detailed theoretical comparison of confidence limit procedures, Hall (1988) shows that this Studentized form leads to one-sided confidence limits whose coverage is correct to $O(n^{-1/2})$.

A different pivotal construction is offered by Beran (1987), who mimics the probability integral transform approach. Thus if $Q_0(T, \theta)$ has cdf $\tilde{G}_0$ under sampling from $\tilde{F}$, then $Q = \tilde{G}_0(Q_0(T, \theta))$ is very nearly pivotal. If $\tilde{G}$ is the distribution function of $Q$ under sampling from $\tilde{F}$, and if $Q$ is monotone in $\theta$, then solutions to

$$\tilde{G}(Q(T, \theta)) = \tfrac{1}{2}\alpha \text{ and } 1 - \tfrac{1}{2}\alpha$$

define approximate equitailed $1 - \alpha$ limits for $\theta$. The difficulty is that $\tilde{G}$ is based on second-level bootstrapping, i.e. sampling from samples from $\tilde{F}$: see below. On the surface this suggests the need for a rather extravagant numerical simulation, perhaps using $10^5$ or $10^6$ samples. The theoretical and numerical properties are comparable to those for the Studentized estimate approach outlined above. The method seems worthy of further study.

It may be appropriate here to say a little more about second-level bootstrapping, a process which has several potential uses. Suppose that one wants to check whether or not $T - \theta$ is pivotal, considering this in the first instance as the limited question as to whether or not $\text{var}(T \mid F) = \sigma^2(\theta)$ is in fact constant. (Note that this ignores the possibility of another parameter $\phi$ affecting the distribution of $T$.) An empirical strategy is to simulate several samples from each of several populations, each of which has a different value of $\theta$. For each population, then, one obtains an estimate of $\sigma^2(\theta)$: these estimates are compared to assess possible dependence on $\theta$. In the nonparametric bootstrap context, a population and its $\theta$ value are equated to a simulated sample $(x_1^*, \ldots, x_n^*)$ and its $\theta$ estimate $t^*$. Therefore $\sigma^2(t^*)$ is estimated by taking samples $(x_1^{**}, \ldots, x_n^{**})$ from $(x_1^*, \ldots, x_n^*)$ and computing the empirical variance $\tilde{\sigma}^2(t^*)$ of the $t^{**}$ values which are the $\theta$ estimates calculated from $(x_1^{**}, \ldots, x_n^{**})$. One might take 50 $t^{**}$s for each of 20 $t^*$s. This idea appears to be due to P. L. Chapman; see Chapman and Hinkley (1986).

By way of illustration, Fig. 1(a) shows estimated 5th and 95th percentiles of the error in sample correlation coefficient $r$ for 20 values of population correlation $\rho$, all obtained from two-level bootstrapping of one sample of $n = 20$ bivariate normal pairs. Fig. 1(b) gives corresponding results for Fisher's $z$ transform, $z = \tanh^{-1} r$. Note that in the first plot, the estimated percentiles of $r - \rho$ mimic the normal theory trend: the fitted curves are close to $\pm 1.645(1 - \rho^2)/\sqrt{n}$. The plot suggests strongly that $r - \rho$ is not pivotal. On the other hand, the near-horizontal trends of percentiles in the second plot suggest that error in $z$ is very nearly pivotal. This would imply that bootstrap results for $z$ are reliable approximations to theoretical properties of $z$.

Beran's pivotal construction is not the only confidence limit method based on second-level bootstrapping. More recently Tibshirani (1987) has considered explicit use of smoothed versions of $\tilde{\sigma}^2(t^*)$ to obtain a variance-stabilized estimate

$$U = h(T) = \int^T \{\tilde{\sigma}^2(t^*)\}^{-1/2} \mathrm{d}t^*,$$

to which is then applied a confidence limit procedure for the invertible function $h(\theta)$. Initial results show the method to be competitive with the best known methods in many problems.

The final method to be mentioned is the accelerated bias-corrected percentile method of Efron (1987), which attempts implicit rather than explicit use of variance stabilization, while at the same time recognizing that the variance-stabilized estimate

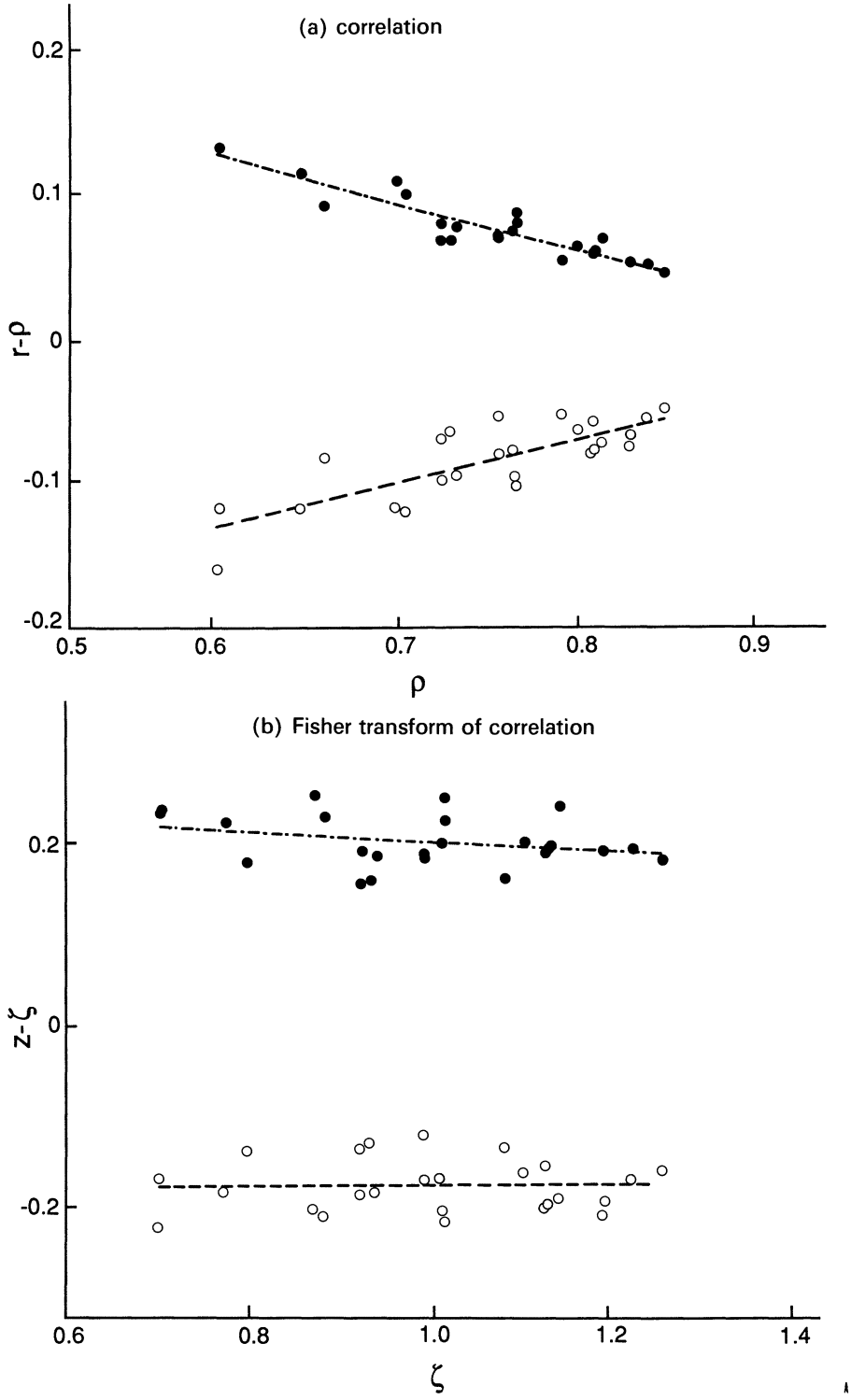


Fig. 1. Bootstrap estimates of 5% (○) and 95% (●) quantiles of (a) $r-\rho$ and (b) $z-\zeta=\tanh^{-1}r-\tanh^{-1}\rho$ obtained from analysis of one sample of $n=20$ pseudo-normal pairs: values of $\rho$ are $B=25$ $r^*$s; estimated quantiles of $r$ are quantiles of empirical cdf of $r^{**}$ from 100 second-level bootstrap samples

$h(T)$ may have bias of order $n^{-1}$ and standardized skewness of order $n^{-1/2}$. This leads to the working assumption that for appropriate $h(\cdot)$ and constants $\tau$, $\alpha$ and $\beta_0$,

$$Q = \frac{\tau\{h(T) - h(\theta)\}}{1 + \alpha\tau h(\theta)} + \beta$$

has a standard normal distribution. Efron's use of this assumption in the bootstrap context does not involve knowing $h(\cdot)$, $\tau$, $\alpha$ or $\beta$. The reliability of the resulting confidence limit method is rather uneven, albeit often very good. One obvious defect is that for large enough $\alpha$, $Q$ may not be monotone over an appropriately wide range for $\theta$. DiCiccio and Romano (1988) discuss the method in detail.

There are many empirical studies of the performances of bootstrap confidence limits, and the results shown in Table 3 seem quite representative. Here $T$ is the mean $\bar{X}$ of samples of size $n = 20$, artificially generated from the $\chi_1^2$ distribution. For each sample, bootstrap simulation with $B = 1000$ was used. Table 3, taken from Owen (1987) shows empirical error rates of nominal 90% equitailed intervals for mean $\theta$, based on 1000 data sets.

A different approach to bootstrap assessment of parameter uncertainty is via a nonparametric likelihood. This is discussed separately in Section 8.

## 5. SIGNIFICANCE TESTS

The connexion between confidence limits and significance tests (Cox and Hinkley, 1974) may be exploited to test certain kinds of hypotheses about parameters. But a direct approach is also possible using bootstrap techniques, particularly for 'pure significance tests' (Cox and Hinkley, 1974, ch. 3). There are, of course, connexions to other nonparametric methods of testing.

Suppose that $T$ is a statistic proposed for testing hypothesis $H$, large values of $T$ being evidence against $H$. We have indicated in Section 2 how the simple bootstrap approximates a probability such as $\Pr(T \leqslant d \mid F)$ by $\Pr(T^* \leqslant d \mid \tilde{F})$. Now a different sampling distribution is required, because the test $P$ value is calculated under the restriction imposed by $H$. If $\delta(\cdot,\cdot)$ is a distance measure between distributions, and if $\mathscr{F}_H$ is the set of all distributions satisfying $H$, then the bootstrap data distribution might be taken as

$$\tilde{F}_H \text{ minimizing } \delta(F, \tilde{F}) \text{ for } F \in \mathscr{F}_H.$$

TABLE 3

*Error rates of bootstrap 90% confidence intervals for mean $\theta$ of $\chi_1^2$, samples of size $n = 20$ (Owen, 1987)*

| *Method* | *Proportion of times $\theta <$ lower limit* | *Proportion of times $\theta >$ upper limit* | *Aggregate error rate* |
|---|---|---|---|
| Exact parametric | 0.051 | 0.056 | 0.107 |
| Bootstrap percentile | 0.023 | 0.150 | 0.173 |
| Efron's accelerated, bias-corrected bootstrap | 0.050 | 0.105 | 0.155 |
| Bootstrap Student $t$ | 0.038 | 0.072 | 0.112 |

The bootstrap test $P$ value corresponding to observed statistic $t_{\text{obs}}$ would be

$$\tilde{P}_H = \Pr\{T_H^* \geq t_{\text{obs}} \mid \tilde{F}_H\}, \tag{7}$$

where $T_H^*$ is the test statistic calculated under random sampling from $\tilde{F}_H$.

There are basically two ways to obtain $\tilde{F}_H$ from $\tilde{F}$, one being to change the probabilities at $x_1, \ldots, x_n$ from $n^{-1}$ to $w_1, \ldots, w_n$; the other being to redistribute the probabilities $n^{-1}$ to a wider support than $x_1, \ldots, x_n$. Efron (1982, ch. 10) discusses applications of the former, specifically embedding $\tilde{F}$ in an exponential family; see also Owen (1987).

Uses of modified support are described by Ducharme *et al.* (1985) and by Young (1986). For example, in one of Young's applications, the hypothesis $H$ asserts independence of the two components of $X = (Y, Z)$, and $\tilde{F}_H$ is naturally taken to be the product of the empirical marginal cdfs $\tilde{G}$ and $\tilde{H}$ of $Y$ and $Z$ respectively. The resulting test is therefore very similar to a randomization test, the difference being only that between sampling with and without replacement. The same phenomenon would occur in a two-sample comparison, where a common aggregate distribution would be defined by $\tilde{F}_H$.

A rather striking application of the bootstrap is Silverman's (1981) test for unimodality of a distribution, which uses smooth density estimates as the particular form of probability redistribution. This nicely illustrates the usefulness of bootstrap methods when classical theoretical approaches to calculation of the $P$ value are intractable. Another example is outlined in Section 6.

## 6. REGRESSION PROBLEMS

Application of bootstrap methods in regression is of potential importance because of the ever-increasing generality of regression methods, for which the classical methods of assessment used in textbook linear regression are inappropriate. Efron (1986) gives a useful introduction, and Wu (1986) with its accompanying discussion refers to much of what is known about properties of the bootstrap in regression analysis. There are two general types of problem, one the assessment of accuracy of regression coefficients or fitted values of mean response, the other being selection of variables or choice of model on the basis of some measure of model fit.

Suppose that we have a particular form of model $y_i = \mu(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i$ connecting continuous responses $y_i$ to explanatory variables $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})$, with $\varepsilon_i$s as random errors. Given some method of fitting the relationship, such as least squares or $M$ estimation, we obtain coefficient estimate $\hat{\boldsymbol{\beta}}$ and fitted values $\hat{\mu}_i = \mu(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$. Inspection of the residuals $\hat{e}_i = y_i - \hat{\mu}_i$, or prior evidence, may suggest that the errors $\varepsilon_i$ are homogeneous, with distribution $F$ estimable by the empirical distribution $\tilde{F}$ of residuals. If so, the bootstrap methods discussed earlier extend straightforwardly, simulated data sets *data** taking the form $\{(x_i, y_i^*), i = 1, \ldots, n\}$ with $y_i^* = \hat{\mu}_i + \varepsilon_i^*$, where $\varepsilon_i^*$ is randomly sampled from $(\hat{e}_1, \ldots, \hat{e}_n)$. Fitting the model to *data** gives simulated estimate $\hat{\boldsymbol{\beta}}^*$ and fitted values $\hat{\mu}_i^*$. Repeated simulation then leads to required assessments of uncertainty as in earlier sections.

As an example, consider the following significance testing problem. The mean relationship $\mu(x)$ is either linear (hypothesis $H$) or piecewise linear with two linear segments intersecting at $x = \gamma$. Statistic $T$ is the normal theory likelihood ratio test statistic, whose exact null distribution is intractable even if errors $\varepsilon$ are normal.

In the terminology of Section 5, the empirical distribution of residuals $\hat{e}_i$ from the linear regression is $\tilde{F}_H$, and significance probability $\tilde{P}_H$ in (7) is calculated using samples $y_i^* = \hat{\mu}_i + \varepsilon_i^*$ as described above with $\hat{\mu}_i$ the fitted linear regression values. This method was applied to a small set of data from a noise signal experiment in which the $n = 9$ values of $x$ were natural logarithms of 10, 20, 30, 50, 100, 150, 200, 300 and 500 with corresponding values of $y$ being 87.83, 86.50, 84.83, 83.50, 80.17, 79.50, 79.17, 78.67 and 78.67. The estimated point of intersection in the two-segment model is $\hat{\gamma} = 5.1$ and the test statistic is $t = 14.7$. From $B = 1000$ bootstrap samples, $\tilde{P}_H$ was calculated to be approximately 0.02. The null distribution of $T$, as estimated by the empirical distribution of $T_H^*$, is not at all close to the $\chi_2^2$ distribution which an (invalid) appeal to classical theory might suggest; see Feder (1975).

One might argue that raw residuals $\hat{e}_i$ should be modified prior to use as simulated errors, e.g. by standardizing to remove the effects of leverage and by adjusting to zero mean. Unpublished numerical evidence supports such modifications. Whether or not one need use complicated modifications such as those described by Cook and Tsai (1985) for non-linear models is unclear.

A more interesting context is that in which errors are not homogeneous, so that a single empirical error distribution is inappropriate. One simple approach is then to consider $(\mathbf{x}_i, y_i)$ as sampled from a joint distribution $F$, the implication being to sample vectors $(\mathbf{x}_i^*, y_i^*)$ from the data vectors in the bootstrap simulation. There are two drawbacks with this approach. First, it would often be the case that $\text{var}(\varepsilon_i)$ changes smoothly with $\mathbf{x}_i$ or $\mu_i$, and use might be made of this. Secondly, on the general grounds of requiring inference to be conditional on the design $D = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, one should not risk having simulated data sets whose designs $D^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*)$ are very different from $D$.

The last point could be dealt with separately either by pre-stratification or post-stratification of the sampling of data vectors, in either case forcing $D^*$ and $D$ to be close in a meaningful sense.

The design difficulty may be moot, of course, if some form of modelling for the errors is used. An example of this in nonparametric regression is given by Efron (1986). A local smoothing algorithm is used first to fit $\hat{\mu}(x)$, and is then applied to squared residuals $\hat{e}_i^2$ to fit a smooth relationship between $\sigma^2 = \text{var}(\varepsilon)$ and $x$, say $\hat{\sigma}^2(x)$. This permits calculation of homogeneous standardized residuals $r_i = \hat{e}_i / \hat{\sigma}(x_i)$, and thence defines a bootstrap model

$$y_i^* = \hat{\mu}(x_i) + \hat{\sigma}(x_i) r_i^*$$

with the $r_i^*$ randomly sampled from $(r_1, \ldots, r_n)$. Bootstrap samples are then used to obtain confidence bands for $\mu(x)$.

So far we have assumed that responses $y$ are continuous and that errors are additive. How might one apply bootstrap methods to responses which are counts, i.e. non-negative integers, say? One approach is to use the local linearization which GLIM uses for its iterative weighted least squares fitting of generalized linear models. But such an approach offers little more than jackknife methods. If count data are thought of as extended Poisson, that is with variance function $\phi(\mathbf{x})\mu$, then a locally smooth estimate of $\phi(\mathbf{x})$ could be produced and the data could be analysed appropriately in GLIM. More needs to be learned about the possible role of bootstrap methods in such situations.

Special mention should be made of cases where replication exists at every design point. In such cases it would be possible to estimate response distributions $F_i$ at each $\mathbf{x}_i$, and thence bootstrap by sampling from $\tilde{F}_i$ at each $\mathbf{x}_i$. This approach of course applies to multisample problems, unless separate variance components are involved (Section 9). An open question is how well the bootstrap will perform when each of very many $\tilde{F}_i$s is based on few responses. The results of Bickel and Freedman (1982) are probably relevant. There is a very useful series of papers by Freedman and Peters (1984 and references therein) on the performance of bootstrap methods in econometric regression models.

The rather different types of problems typified by model selection, variable selection and prediction assessment are problems to which cross-validation techniques (Stone, 1974) are often applied. A detailed analysis by Efron (1983) shows that cross-validation techniques may be inferior to bootstrap assessments in many cases; see also Bunke and Droge (1984). This important problem will not be discussed here.

## 7. CONDITIONAL BOOTSTRAP METHODS

In the preceding section the idea of conditioning was mentioned briefly. Conditioning on ancillary statistics is an important general component of statistical inference. As to whether or not relevant conditioning is generally possible in bootstrap methods, the situation is unclear.

A crucial issue may be the nature of the conditioning variable, or ancillary statistic. For example, suppose that $E(T-\theta \mid a,\ F) = b(a-\alpha,\ F)$ with $\alpha = E(A \mid F)$ or with $a = a(\tilde{F})$ and $\alpha = a(F)$. A bootstrap simulation can estimate $b(\cdot,\ F)$ by $b(\cdot,\ \tilde{F})$, but this cannot be used without knowing $\alpha$, at least with error negligible compared to $a-\alpha$. This difficulty seems to preclude conditional bootstrap analysis of the sample mean, for example. The regression application suggested in Section 6 is different in the sense that the effect of an ancillary measure $a$ of the design $D$ does not involve the mean of $A$.

There is also the difficulty of choosing $a$ in the absence of a model, accompanied by the difficulty of estimating properties conditional on $a$. For example, in a regression problem with non-homogeneous errors, the precise form of effect of the design $D$ on the variances of coefficients will usually be unknown. However, if the regression fit is approximately linear with weight $w_i$ attached to $(\mathbf{x}_i,\ y_i)$, and if $\mathrm{var}(y_i \mid \mathbf{x}_i)$ is estimated by $\hat{\sigma}_i^2$, then it would seem appropriate to define $a$ in terms of the elements of $\Sigma w_i^2 \hat{\sigma}_i^2 \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}$, by analogy with weighted least squares linear regression. Once $a$ is chosen, the required conditional property would be estimated using discrete partitions of the bootstrap simulation. For example, $\mathrm{var}(\hat{\beta} \mid a)$ could be approximated by a smoothed version of $\mathrm{var}(\hat{\beta}^* \mid a^*)$ evaluated at $a^* = a$.

In some, possibly rare, cases conditional distributions will be amenable to special numerical techniques, such as stratified simulation or conditional saddlepoint approximations. One example of the latter is given by Davison and Hinkley (1988).

It may be worth remarking that in classical statistics the likelihood function itself provides exact or approximate conditional inference (Barndorff-Nielsen, 1983; Cox and Reid, 1987). Quite possibly one might use the bootstrap likelihoods of Section 8 in the same way.

## 8. BOOTSTRAP PARTIAL LIKELIHOODS

Alchemy failed. But bootstrappers have produced likelihoods, or confidence

distributions. For want of something better, the term partial likelihood may be appropriate.

One direct approach by Hall (1987) is to derive a smooth density estimate from the bootstrap simulation values of the Studentized pivot $Q = (T - \theta)/S$ mentioned in Section 4. Such a partial likelihood has good properties when used to calculate confidence sets, and may show interesting features which standard normal approximations do not. A second approach is via the second-level bootstrap of Section 4, with likelihood evaluations at $\theta = t^*$ being calculated as approximate densities of $T^{**}$ at $t^*$.

A more classical analogy is pursued by Ogbonmwan and Wynn (1988) for problems involving contrast parameters. Suppose that data $\mathbf{y} = y_1, \ldots, y_n$ are such that, for the correct value of $\theta$, the transformed vector $g(\mathbf{y}, \theta) = g_1(\theta), \ldots, g_n(\theta)$ may be assumed to be a random sample from a fixed distribution function $F_0$. If $T = t(\mathbf{y})$ is the estimating function for $\theta$, define $T_\theta = t(g(\mathbf{y}, \theta))$ with observed value $t_\theta$. Then a partial likelihood for $\theta$ is the density of $T_\theta$ at $t_\theta$. The bootstrap version of this definition involves replacing $F_0$ by the empirical distribution function $\tilde{F}_\theta$ defined by data values $g(\mathbf{y}, \theta)$, and approximating the density of statistic $T_\theta^*$ obtained from samples generated by $\tilde{F}_\theta$. In some cases numerical simulation can be avoided, as in the following example, taken from Davison and Hinkley (1988).

Suppose that $\theta$ is the difference between means for two populations from which the following two samples were drawn

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *sample* 1: | 37.5 | 34.8 | 38.9 | 38.6 | 37.0 | 37.4 | 36.5 | 38.4 | 38.0 | 30.7 |
| *sample* 2: | 37.7 | 36.3 | 38.0 | 37.0 | 37.6 | 33.2 | 36.7 | 27.4 | 37.1 | 37.4 |

Denote general samples by $(x_1, \ldots, x_m)$ and $(x_{m+1}, \ldots, x_{m+n})$, and suppose that we choose to estimate $\theta$ by $t = m^{-1}\Sigma_{m+1}^{m+n} x_i - m^{-1}\Sigma_1^m x_i$. Since the two sample variances are nearly equal, it seems reasonable to take $g(\mathbf{x}, \theta) = (x_1, \ldots, x_m, x_{m+1} - \theta, \ldots, x_{m+n} - \theta)$. If $g_1^*, \ldots, g_{m+n}^*$ denotes a random sample from the elements of $g(\mathbf{x}, \theta)$, then $T_\theta^* = n^{-1}\Sigma_{m+1}^{m+n} g_i^* - m^{-1}\Sigma_1^m g_i^*$. A saddlepoint density approximation can be obtained for $T_\theta^*$, and its evaluation at $t_\theta = t - \theta$ defines the bootstrap partial likelihood. The result is graphed in Fig. 2, together with the normal theory modified profile likelihood.

Note that this type of bootstrap partial likelihood could just as easily be based on any estimate $T$, although the saddlepoint simplification requires that $T$ be defined by linear estimating equations. The method is very similar to the use of randomization distributions.

A more direct approach is taken by Owen (1987), who considers $\tilde{F}$ to be embedded in a class of distributions $\mathcal{F}_\mathbf{x}$ whose support is $x_1, \ldots, x_n$ in the simple case of homogeneous data. Then if $\theta = t(F)$, the bootstrap likelihood of $\theta$ is the profile likelihood under the 'model' $\mathcal{F}_\mathbf{x}$. More concretely, consider $F_\mathbf{w}$ to attach probabilities $w_1, \ldots, w_n$ at points $x_1, \ldots, x_n$; $\tilde{F}$ is the maximum likelihood estimate with $w_i \equiv n^{-1}$, $i = 1, \ldots, n$. Then define the bootstrap likelihood to be

$$BL(\theta) = \sup_{\mathbf{w}: t(F_\mathbf{w}) = \theta} \prod_{i=1}^{n} w_i.$$

Owen (1987) outlines and applies an algorithm for calculating $BL(\theta)$. He also demonstrates that, at least in simple cases, conventional chi-squared asymptotics apply to the log-likelihood ratio.

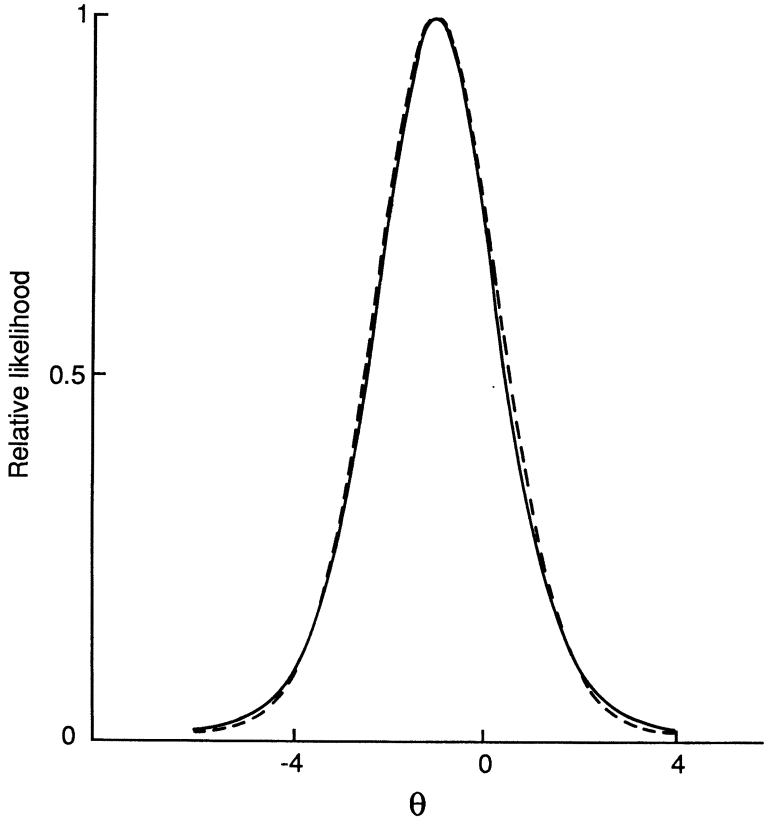


Fig. 2. Relative likelihoods for two-sample contrast parameter $\theta$: the full curve is the saddlepoint approximation to bootstrap likelihood; the broken curve is the normal theory modified profile likelihood

One unusually simple model where an empirical likelihood and resulting conditional analysis are possible is the change-point model. The basic theory and one application are described by Hinkley and Schechtman (1986). Another application is to the mean shift analysis of the series of UK coal-mining disasters (Andrews and Herzberg, 1985, p. 51). The model for count $x_i$ in the $i$th period of length one year is that $\Pr(X_i = r) = f_0(r)$, $i \leqslant \theta$, and $\Pr(X_i = r) = f_1(r)$, $i > \theta$, successve counts being independent. A nonparametric profile likelihood for $\theta$ is therefore

$$PL(t) = \prod_{i=1}^{t} \hat{f}_0(x_i \mid t) \prod_{i=t+1}^{n} \hat{f}_1(x_i \mid t),$$

where

$$\hat{f}_0(r \mid t) = t^{-1} \sum_{i=1}^{t} \delta(x_i - r), \quad \hat{f}_1(r \mid t) = (n-t)^{-1} \sum_{i=t+1}^{n} \delta(x_i - r).$$

Table 4 shows the crucial part of the data series and corresponding values of $PL(t)$ after normalizing to unit sum: the result is an approximate conditional distribution, in that if $\hat{\theta}$ maximizes $PL$ then

$$\Pr(\hat{\theta} - \theta = d \mid a) \propto PL(\hat{\theta} - d).$$

TABLE 4

*Part of the annual UK coal-mining disaster frequencies $x_t$ and corresponding normalized bootstrap likelihood PL (t), t = calendar year — 1850*

| Year $t$ | 34 (1884) | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency $x_t$ | 2 | 3 | 4 | 2 | 1 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 3 | 0 |
| Normalized $PL(t)$ | 0.002 | .003 | .189 | .199 | .048 | .100 | .130 | .220 | .061 | .021 | .008 | .004 | .008 | .001 |

The ancillary $a$ here is the set of likelihood ratio increments $PL(\hat{\theta}+k)/PL(\hat{\theta}+k-1)$, most influential being those for small $|k|$. These same increments could be used to partition a bootstrap simulation if a non-likelihood analysis were performed (Hinkley and Schechtman, 1987). Note that bootstrap simulation extends easily to more complicated models, such as first-order Markov processes.

## 9. OTHER APPLICATIONS

The types of applications mentioned thus far are mostly elementary, save for regression. There is a growing literature on other, more complex applications, some of which are mentioned in this section; see also the general remarks in Section 10.

One traditional area of application for subsampling techniques is the analysis of complex sample surveys. In the usual case where data sampling is without replacement from finite populations, ordinary bootstrapping (done with replacement) may produce inadmissible simulated samples. Partly for this reason, a series of special bootstrap techniques has been proposed in the sample survey literature. Some of the techniques are appraised by McCarthy and Snowden (1985), who give preliminary endorsement to the simple modification of increasing bootstrap sample size from $n$ to $n/(1-f)$, where $f$ is the data sampling fraction.

Problems involving time series, or more generally a stochastic process, raise the difficulty of the single realization. What plays the role of $\tilde{F}$? There are two possible elementary strategies: (i) split the realization into several pieces, and sample from these, or (ii) fit a model with independent innovations, and simulate realizations by adding sampled residuals to fitted values. More sophisticated versions of these strategies will be required for fairly general application.

Perhaps more conventional are problems involving variance components, such as occur in empirical Bayes models. The essential point here is that bootstrap simulation should, implicitly or explicitly, simulate each component of variability. Precisely how will depend on the application. Suppose, for example, that a notional model for the data matrix $x_{ij}$ of $p$ samples is $x_{ij}=\mu_i+\varepsilon_{ij}$, where the $\mu$s and $\varepsilon$s respectively have distributions $G$ and $F$. If we are interested in a statistic symmetric in the samples, such as $\bar{x}..$ or $\max_i \bar{x}_{i.}$, then we can simulate data $x_{ij}^*$ by $x_{ij}^*=\mu_i^*+\varepsilon_{ij}^*$, where $\mu_i^*$ is randomly sampled from estimates $(\hat{\mu}_1, \ldots, \hat{\mu}_p)$ and $\varepsilon_{ij}^*$ are randomly sampled from residuals $\{x_{ij}-\bar{x}_{i.}\}$. The estimates $\hat{\mu}_i$ would be of empirical Bayes type but corrected to have appropriate mean and variance, e.g. $\bar{x}..$ and the unbiased estimate of var($\mu$). Such a simulation would not be appropriate if, say, we were interested in mean $\mu_1$ *a priori*, for then one simulated sample should use $\mu^*=\hat{\mu}_1$. Further discussion of these kinds of applications will be found in Hill (1986) and Laird and Louis (1987).

## 10. GENERAL REMARKS

One might observe that bootstrap methods essentially embrace, or enlarge upon, familiar methods of simulation, subsampling and permutation. What is new is the generality of approach, the range of potential applications and the massive use of computer power.

It would be presumptive to dismiss the many simple applications because of existing classical methods: such applications are mere scale exercises, which help to tune the instruments and their players in the bootstrap orchestra so that they will perform better in the complex pieces of modern data analysis. Thus, for example, bootstrap methods may prove to be uniquely reliable tools for analysing nonparametric curve fits, complex pure significance test problems and nonstationary time series models. At the very least bootstrap methods provide a simple approach to assessment of the sensitivity of traditional methods to model assumptions. This thought also suggests the possible use of simulated samples to generate diagnostics, akin to the more usual case deletion diagnostics.

Because bootstrap methods also apply in the arena of model assessment, they are pertinent to the larger, often neglected area of decision analysis under model uncertainty.

In the previous sections we have not commented on the non-negligible tendency for misapplication of bootstrap methods, in particular the misuse of simple random sampling from data sets. There is a very clear need to bring classical statistical theory to bear in the development of reliable methodology, as evidenced by the importance of pivots in confidence limit methods. In this context one should also consider Bayesian approaches to bootstrapping, which involve Dirichlet models; see Rubin (1981) and Banks (1987).

## REFERENCES

Andrews, D. F. and Herzberg, A. M. (1985) *Data: a Collection of Many Problems from Many Fields for the Student and Research Worker*. New York: Springer.

Banks, D. L. (1987) Improving the Bayesian bootstrap. Unpublished.

Barndorff-Nielsen, O. (1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 345–365.

Beran, R. J. (1982) Estimated sampling distributions: the bootstrap and its competitors. *Ann. Statist.*, **10**, 212–225.

———(1984) Bootstrap methods in statistics. *Jber. Dtsch. Math-Ver.*, **86**, 14–30.

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

Bickel, P. J. and Freedman, D. A. (1982) Bootstrapping regression models with many parameters. Unpublished, University of California at Berkeley.

Bunke, O. and Droge, B. (1984) Bootstrap and cross-validation estimates of the prediction error for linear regression models. *Ann. Statist.*, **12**, 1400–1424.

Chapman, P. L. and Hinkley, D. V. (1986) The double bootstrap, pivots and confidence limits. *Report 26*. Center for Statistical Sciences, University of Texas at Austin.

Cook, R. D. and Tsai, C. L. (1985) Residuals in nonlinear regression. *Biometrika*, **72**, 23–29.

Cox, D. R. and Hinkley, D. V. (1974) *Theoretical Statistics*. London: Chapman and Hall.

Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional likelihood. *J. R. Statist. Soc.* B, **49**, 1–39.

Daniels, H. E. (1987) Tail probability approximations. *Int. Statist. Rev.*, **55**, 37–48.

Davison, A. C. and Hinkley, D. V. (1988) Saddlepoint approximations in resampling methods. *Biometrika*, **75** (to appear).

Davison, A. C., Hinkley, D. V. and Schechtman, E. (1987) Efficient bootstrap simulation. *Biometrika*, **74**, 555–566.

DiCiccio, T. J. and Romano, J. P. (1988) A review of bootstrap confidence intervals. *J. R. Statist. Soc.* B, **50**, 338–354.

Ducharme, G. R., Jhun, M., Romano, J. P. and Truong, K. N. (1985) Bootstrap confidence cones for directional data. *Biometrika*, **72**, 637–645.

Efron, B. (1982) The jackknife, the bootstrap and other resampling plans. In *Regional Conference Series in Applied Mathematics*, No. 38. Philadephia: SIAM.

———(1983) Estimating the error rate of a prediction rule: improvements in cross-validation. *J. Amer. Statist. Ass.*, **78**, 316–331.
———(1986) Computer-intensive methods in statistical regression. Unpublished, Department of Statistics, Stanford University.
———(1987) Better bootstrap confidence intervals. *J. Amer. Statist. Ass.*, **82**, 171–200.
Feder, P. I. (1975) On asymptotic distribution theory in segmented regression problems: identified case. *Ann. Statist.*, **3**, 49–83.
Freedman, D. A. and Peters, S. C. (1983) Bootstrapping a regression equation: some empirical results.
Graham, R. L., Hinkley, D. V., John, P. W. M. and Shi, S. (1987) Balanced design of bootstrap simulations. *Report 48*. Center for Statistical Sciences, University of Texas at Austin.
Hall, P. (1987) On the bootstrap and likelihood-based confidence regions. *Biometrika*, **74**, 481–493.
———(1988) Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.*, to be published.
Hill, J. R. (1986) Empirical Bayes statistics: a comprehensive theory for data analysis. *PhD Thesis*. Department of Mathematics, University of Texas at Austin.
Hinkley, D. V. and Schechtman, E. (1987) Conditional bootstrap methods in the mean-shift model. *Biometrika*, **74**, 85–93.
Laird, N. M. and Louis, T. A. (1987) Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Ass.*, **82**, 739–750.
McCarthy, P. J. and Snowden, C. B. (1985) The bootstrap and finite population sampling. In *Vital and Health Statistics*, Series 2, No. 95. Washington DC: Public Health Service.
Miller, R. G., Jr (1974) The jackknife: a review. *Biometrika*, **61**, 1–17.
Ogbonmwan, S. M. and Wynn, H. P. (1986) Accelerated resampling codes with low discrepancy. Unpublished, Department of Statistics, Imperial College.
———(1988) Resampling generated likelihoods. In *Statistical Decision Theory and Related Topics IV* (eds S. S. Gupta and J. O. Berger), vol. 1, pp. 133–147. New York: Springer.
Owen, A. B. (1987) Empirical likelihood ratio confidence intervals for a single functional. Unpublished, Department of Statistics, Stanford University.
Rubin, D. B. (1981) The Bayesian bootstrap. *Ann. Statist.*, **9**, 130–134.
Silverman, B. W. (1981) Using kernel density estimates to investigate multimodality: *J. R. Statist. Soc.* B, **43**, 97–99.
Silverman, B. W. and Young, A. (1987) The bootstrap: to smooth or not to smooth? *Biometrika*, **74**, 469–479.
Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc.* B, **36**, 111–147.
Thernau, T. (1983) Variance reduction techniques for the bootstrap. *PhD Thesis*. Department of Statistics, Stanford University.
Tibshirani, R. (1987) Variance stabilization and the bootstrap. Unpublished, Department of Statistics, University of Toronto.
Wu, C. F. J. (1986) Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, **14**, 1261–1350.
Young, A. (1986) Conditioned data-based simulations: some examples from geometrical statistics. *Int. Statist. Rev.*, **54**, 1–13.

*J. R. Statist. Soc.* **B** (1988)
**50**, *No.* 3, *pp.* 338–354

# A Review of Bootstrap Confidence Intervals

By THOMAS J. DICICCIO and JOSEPH P. ROMANO†

*Stanford University, USA*

[*Read before the* Royal Statistical Society *on Wednesday, March 16th, 1988 at a meeting organized by the* Birmingham Group, *Professor* J. B. Copas *in the Chair*]

SUMMARY

A survey of bootstrap procedures for constructing confidence regions is given. In particular, several distinct bootstrap methods are considered, with emphasis on the mathematical correctness of these procedures. The percentile, bias-corrected percentile and accelerated bias-corrected percentile methods, developed by Efron, are reviewed in both parametric and nonparametric situations. A procedure related to the accelerated bias-corrected method, which avoids explicit calculation of the analytical corrections required in Efron's method, is also introduced. In the context of a functional approach for the construction of confidence regions, the bootstrap is motivated as a method to estimate the distributions of approximate pivots. Finally, iterative bootstrap methods are discussed as means to improve coverage accuracy.

*Keywords*: ASYMPTOTIC THEORY; BOOTSTRAP; CONFIDENCE REGIONS; PERCENTILE METHOD; BIAS-CORRECTED PERCENTILE METHOD; ACCELERATED BIAS-CORRECTED PERCENTILE METHOD; PERCENTILE-*t*; PIVOT; LEAST FAVOURABLE FAMILY; ORTHOGONAL PARAMETERS; SECOND-ORDER ACCURACY; PREPIVOTING

## 1. INTRODUCTION

This paper is a survey of bootstrap procedures for constructing confidence regions for parameters of interest. Such procedures rely on estimating the sampling distribution of a statistic or an approximate pivot. In general, bootstrap methods consist of estimating a characteristic of the unknown population by simulating the characteristic when the true population is replaced by an estimated one. The appeal of this approach is its wide applicability to complex data structures in both parametric and nonparametric problems. Several distinct bootstrap methods will be reviewed, with the emphasis on the mathematical correctness of these procedures.

In Section 2, Efron's percentile, bias-corrected (BC) percentile and accelerated bias-corrected ($BC_a$) percentile methods are developed in both parametric and nonparametric situations. These procedures arise from transformation theory considerations, and they have the property of invariance under reparameterization. The most promising of these techniques is the $BC_a$, which depends on the calculation of the acceleration constant. A general formula is given for this analytical adjustment for situations other than maximum likelihood estimation. A procedure related to the accelerated bias-corrected method is also introduced which avoids explicit calculation of analytical corrections required in Efron's method. These methods are compared and the percentile-*t* method is considered in some numerical examples.

In Section 3, some bootstrap methods, including the percentile-*t*, are motivated as a functional approach to the construction of confidence regions by using the bootstrap to estimate the distribution of approximate pivots. The mathematical analysis of these

† *Address for correspondence*: Department of Statistics, Stanford University, Sequoia Hall, Stanford, CA 94305–4065, USA.

 0035–9246/88/50338 $2.00

methods is outlined, including conditions required to justify them and a discussion of examples where they lead to inconsistencies.

Section 4 focuses on refined bootstrap methods with the goal being improved coverage accuracy. In particular, iterative methods proposed by Beran, Loh and Hall are discussed. Beran's method of prepivoting and Loh's calibrated confidence sets are seen to be equivalent. These methods offer the potential for increased accuracy of coverage for confidence sets, but their computational feasibility needs to be established for them to be considered viable general approaches.

In any given situation, the choice of bootstrap procedure depends on available theoretical results, computational considerations, the level of accuracy desired, simulation results and experience with similar problems. For example, both the $BC_a$ and the percentile-$t$ are second order correct; however, the $BC_a$ requires knowledge of an analytical constant while the percentile-$t$ requires a stable estimate of variance. Given the diversity of criteria in choosing a procedure, it is unlikely that a single procedure will emerge as a preferred method in all problems.

A common feature of all the procedures considered is the use of simulation to approximate a sampling distribution by treating an estimate of the population as the true one. The computational aspects of this problem are not treated here, but the reader is referred to section 3 of Hinkley (1988) and Johns (1988).

## 2. PERCENTILE METHOD AND RELATED PROCEDURES

### 2.1. *Introduction*

In a series of articles, Efron (1981, 1982, 1985, 1987) has introduced and refined the percentile method of using bootstrap calculations to set approximate confidence limits for scalar parameters. These refinements of the percentile method are the bias-corrected (BC) percentile method and the accelerated bias-corrected ($BC_a$) percentile method. Efron's approach is to first develop these procedures in the simple context of a parametric model indexed by a scalar parameter, for which there are no nuisance parameters present, and then to adapt them for application in multiparameter families and nonparametric situations. This development relies on transformation theory, and the resulting procedures have the desirable property of invariance under reparameterization.

For a review of the percentile method in the simplest case, suppose that $x_n = (X_1, \ldots, X_n)$ is a sample from a distribution having probability density function $f_\theta$ which depends upon the scalar parameter of interest $\theta$. Let $\hat{\theta}$ be an estimator of $\theta$ based on $x_n$ with distribution function $G_\theta(s) = P_\theta(\hat{\theta} \leqslant s)$. The exact upper $1-\alpha$ confidence limit for $\theta$ is taken to be that value $\theta[1-\alpha]$ satisfying $G_{\theta[1-\alpha]}(\hat{\theta}) = \alpha$, and the bootstrap distribution for $\hat{\theta}$ is $G_\theta$. Now suppose there exists a monotonically increasing transformation $g$ and a constant $\tau$ such that for all values of $\theta$

$$\tau\{g(\hat{\theta}) - g(\theta)\} \sim Z, \tag{2.1}$$

where $Z$ is symmetrically distributed about zero with distribution function $H$. Then

$$G_\theta(s) = H[\tau\{g(s) - g(\theta)\}];$$

hence, in terms of the $\alpha$ quantile $z^{(\alpha)} = H^{-1}(\alpha)$,

$$\theta[1-\alpha] = g^{-1}(g(\hat{\theta}) + z^{(1-\alpha)}/\tau)$$

and

$$G_\theta^{-1}(\alpha) = g^{-1}(g(\hat{\theta}) + z^{(\alpha)}/\tau).$$

Since $\theta[1-\alpha] = G_\theta^{-1}(1-\alpha)$, the confidence limits for $\theta$ can be obtained directly from the bootstrap distribution for $\hat{\theta}$ without explicit knowledge of the transformation $g$.

The quantity $G_{\hat{\theta}}^{-1}(1-\alpha)$ is called the percentile bootstrap confidence limit. Typically, no transformation $g$ exists for which (2.1) obtains exactly, and the difference between $G_{\hat{\theta}}^{-1}(1-\alpha)$ and $\theta[1-\alpha]$ is $O_p(n^{-1})$. Thus the percentile method can give poor approximations in small sample situations. For example, consider $n = 8$ observations from a bivariate normal distribution with known means and variances and unknown correlation coefficient $\theta$. For each of the values 0, 0.3 and 0.8 of the usual estimator $r$, the exact upper and lower 97.5% confidence limits for $\theta$ are compared with the corresponding percentile bootstrap limits in Table 1.

One approach for improvement of the percentile method is to account for bias in (2.1). Assume then that there exists a monotonically increasing transformation $g$ and constants $\tau$ and $z_0$ such that for all $\theta$

$$\tau\{g(\hat{\theta}) - g(\theta)\} + z_0 \sim Z,$$

where $Z$ is as described above. In this case, $G_\theta(s) = H(\tau\{g(\hat{\theta}) - g(\theta)\} + z_0)$, from which it follows that

$$\theta[1-\alpha] = g^{-1}(g(\hat{\theta}) + (z^{(1-\alpha)} + z_0)/\tau)$$

and

$$G_\theta^{-1}(\alpha) = g^{-1}(g(\hat{\theta}) + (z^{(\alpha)} - z_0)/\tau).$$

Since $\theta[1-\alpha] = G_\theta^{-1}\{H(z^{(1-\alpha)} + 2z_0)\}$, the confidence limits for $\theta$ can be obtained from the bootstrap distribution for $\hat{\theta}$. The bias correction $z_0$ can be determined similarly by $z_0 = H^{-1}\{G_\theta(\hat{\theta})\}$. Efron (1982) uses the standard normal distribution function $\Phi$ for $H$, and he calls $G_\theta^{-1}\{\Phi(z^{(1-\alpha)} + 2z_0)\}$ the bias-corrected percentile bootstrap confidence limit. If $z_0 = 0$, then the BC method reduces to the percentile method. The upper and lower 97.5% BC limits are shown for the correlation coefficient example in Table 1.

The BC limits, like the percentile limits, typically differ from the exact ones by terms of order $O_p(n^{-1})$. In the correlation coefficient example the BC method produces fairly accurate approximations; however, for an example in which the BC method performs poorly, consider a sample of size $n = 5$ from the exponential distribution with mean $\theta$, and take $\hat{\theta}$ to be the sample mean. Table 2 compares the exact upper and lower 97.5% confidence limits for $\theta$ with the corresponding approximations obtained by the percentile and BC methods. Although use of the bias correction produces an improvement over the percentile method, the BC limits are inadequate. This example is similar to one considered by Schenker (1985), which in part motivated the development of the $BC_a$ method.

To improve the BC method, Efron (1987) supposes there exists a monotonically increasing transformation $g$ and constants $\tau$, $z_0$ and $a$ such that for all values of $\theta$

$$\tau\left\{\frac{g(\hat{\theta}) - g(\theta)}{1 + a\tau g(\theta)}\right\} + z_0 \sim Z, \tag{2.2}$$

TABLE 1
*Upper and lower 97.5% confidence limits for the correlation coefficient ($n = 8$)*

| $r$ | $z_0$ | *Exact* | | *Percentile* | | *BC* | | *(2.4)* | | *Percentile-t* | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Lower* | *Upper* | *Lower* | *Upper* | *Lower* | *Upper* | *Lower* | *Upper* | *Lower* | *Upper* |
| 0 | 0 | −0.666 | 0.666 | −0.707 | 0.707 | −0.707 | 0.707 | −0.659 | 0.659 | −1.412 | 1.412 |
| | | (2.50) | (2.50) | (1.66) | (1.66) | (1.66) | (1.66) | (2.68) | (2.68) | | |
| 0.3 | −0.0611 | −0.479 | 0.797 | −0.495 | 0.838 | −0.539 | 0.819 | −0.475 | 0.792 | −1.342 | 1.257 |
| | | (2.50) | (2.50) | (2.24) | (1.30) | (1.61) | (1.80) | (2.56) | (2.71) | | |
| 0.8 | −0.1656 | 0.199 | 0.952 | 0.307 | 0.966 | 0.153 | 0.954 | 0.183 | 0.951 | −0.107 | 0.996 |
| | | (2.50) | (2.50) | (4.39) | (0.91) | (1.96) | (2.19) | (2.30) | (2.59) | (0.47) | (0.00) |

Below each lower limit $\theta_L$ and each upper limit $\theta_U$ is shown $1 - G_{\theta_L}(r)$ and $G_{\theta_U}(r)$, respectively, expressed as a percentage.

TABLE 2
*Upper and lower 97.5% confidence limits for the exponential mean ($n = 5$)*

| *Exact* | | *Percentile* | | *BC* | | $BC_a$ | |
|---|---|---|---|---|---|---|---|
| *Lower* | *Upper* | *Lower* | *Upper* | *Lower* | *Upper* | *Lower* | *Upper* |
| $0.488\hat{\theta}$ | $3.080\hat{\theta}$ | $0.325\hat{\theta}$ | $2.048\hat{\theta}$ | $0.390\hat{\theta}$ | $2.270\hat{\theta}$ | $0.488\hat{\theta}$ | $3.083\hat{\theta}$ |
| (2.50) | (2.50) | (0.06) | (10.11) | (0.43) | (7.28) | (2.50) | (2.49) |

Below each confidence limit is shown its true error rate in coverage expressed as a percentage. $z_0 = 0.1497$, $a = 0.1491$.

where $Z$ is as previously described. Under this assumption

$$G_\theta(s) = H\left[\tau\left\{g(s) - g(\theta)\right\} \Big/ \left\{1 + a\tau g(\theta)\right\} + z_0\right],$$

so that

$$\theta[1-\alpha] = g^{-1}\left[g(\hat{\theta}) + \frac{(z^{(1-\alpha)} + z_0)\{1 + a\tau g(\hat{\theta})\}}{\tau\{1 - a(z^{(1-\alpha)} - z_0)\}}\right]$$

and

$$G_\theta^{-1}(\alpha) = g^{-1}\left[g(\hat{\theta}) + \frac{(z^{(\alpha)} - z_0)\{1 + a\tau g(\hat{\theta})\}}{\tau}\right].$$

The confidence limits for $\theta$ can be obtained from the bootstrap distribution for $\hat{\theta}$, since

$$\theta[1-\alpha] = G_\theta^{-1}\left[H\left\{z_0 + \frac{(z^{(1-\alpha)} + z_0)}{1 - a(z^{(1-\alpha)} + z_0)}\right\}\right], \tag{2.3}$$

and $z_0 = H^{-1}\{G_\theta(\hat{\theta})\}$. Using the standard normal distribution function for $H$, Efron calls (2.3) the accelerated bias-corrected percentile bootstrap confidence limit. The $BC_a$ method reduces to the BC method if $a = 0$.

Efron (1987) shows that for suitable choice of $a$ the $BC_a$ confidence limits are second order correct, which means that the $BC_a$ limits differ from the corresponding

exact limits by terms of order $O_p(n^{-3/2})$. Unlike $z_0$, the acceleration adjustment $a$ cannot be easily determined from the bootstrap distribution $G_{\hat{\theta}}$. If to error of order $O(n^{-1})$ the first three cumulants of $n^{1/2}(\hat{\theta}-\theta)$ are $n^{-1/2}\lambda_1(\theta)$, $\lambda_2(\theta)$ and $n^{-1/2}\lambda_3(\theta)$, and if

$$\gamma = n^{-1/2}\left\{3\frac{\lambda_2'}{\lambda_2^{1/2}} - 2\frac{\lambda_3}{\lambda_2^{3/2}}\right\}$$

and

$$\delta = n^{-1/2}\left\{\frac{\lambda_3}{\lambda_2^{3/2}} - 6\frac{\lambda_1}{\lambda_2^{1/2}}\right\},$$

where $\lambda_2' = \mathrm{d}\lambda_2(\theta)/\mathrm{d}\theta$, then $a$ should be chosen to satisfy $a = \gamma(\hat{\theta})/6 + O_p(n^{-1})$, while $z_0 = \delta(\hat{\theta})/6 + O_p(n^{-1})$. When $\hat{\theta}$ is the maximum likelihood estimator, $\gamma$ and $\delta$ are both equal to the skewness of the score function for $\theta$ based on $x_n$. DiCiccio and Tibshirani (1987) consider the construction of a transformation $g$ which approximately satisfies (2.2).

As shown in Table 2, the $\mathrm{BC_a}$ method is very accurate in the exponential mean example. In this case, $\gamma = \delta = 2/n^{1/2}$. For the correlation coefficient example, these formulae give $\gamma = 0$ and $\delta = -3\theta/n^{1/2}$. Thus in this example, $a = 0$ and the BC and $\mathrm{BC_a}$ limits are equal.

### 2.2. *Multiparameter Families*

Suppose that $x_n = (X_1, \ldots, X_n)$ is a sample from a distribution having probability density function $f_\eta$ which depends upon a vector parameter $\eta = (\eta^1, \ldots, \eta^p)$. Let $\hat{\eta}$ be an estimator of $\eta$ based on $x_n$, and suppose that the scalar parameter $\theta = t(\eta)$ is of interest. Let the distribution function of the estimator $\hat{\theta} = t(\hat{\eta})$ be $G_\eta(s) = P_\eta(\hat{\theta} \leqslant s)$; then the bootstrap distribution of $\hat{\theta}$ is $G_{\hat{\eta}}$.

Having calculated the bootstrap distribution $G_{\hat{\eta}}$, the upper $1-\alpha$ percentile bootstrap confidence limit for $\theta$ is $G_{\hat{\eta}}^{-1}(1-\alpha)$, and the BC limit is $G_{\hat{\eta}}^{-1}\{\Phi(z^{(1-\alpha)} + 2z_0)\}$, where $z_0 = \Phi^{-1}\{G_{\hat{\eta}}(\hat{\theta})\}$. The implementation of the $\mathrm{BC_a}$ method is less straightforward in multiparameter situations because of complications arising in the calculation of the acceleration adjustment $a$. In this case, Efron (1987) restricts attention to maximum likelihood estimators.

To introduce the notation required for Efron's formulation of $a$, let $l(\eta; x_n)$ be the log-likelihood function for $\eta$ based on $x_n$, and let $\kappa_{ij} = E\{l_i l_j\}$ and $\kappa_{ijk} = E\{l_i l_j l_k\}$, where $l_i = \partial l(\eta; x_n)/\partial\eta^i$. For brevity of notation in the expressions that follow, the usual convention is used whereby summation is understood over indices that appear as both subscripts and superscripts. Take $(\kappa^{ij}) = (\kappa_{ij})^{-1}$, $t_i = \partial t(\eta)/\partial\eta^i$, and set $\mu^i = \kappa^{ij}t_j$. Efron reduces the multiparameter family to a scalar parameter one by restricting attention to the line $\eta(\tau) = \hat{\eta} + \tau\hat{\mu}$, called the least favourable family, where $\hat{\mu} = (\hat{\mu}^1, \ldots, \hat{\mu}^p)$ and $\hat{\mu}^i = \mu^i(\hat{\eta})$. In analogy with the scalar parameter case, Efron takes $\gamma(\tau)$ to be the skewness of $\partial l(\eta(\tau); x_n)/\partial\tau$ evaluated at $\eta(\tau)$, so that $\gamma(0) = \hat{\kappa}_{ijk}\hat{\mu}^i\hat{\mu}^j\hat{\mu}^k/(\hat{\kappa}_{ij}\hat{\mu}^i\hat{\mu}^j)^{3/2}$, and then he recommends using $a = \gamma(0)/6$. Alternatively, observed information can be used in place of expected information for the calculation of $a$.

The difference between the true and nominal coverage levels of the percentile and BC confidence limits is typically $O(n^{-1/2})$, and for the $\mathrm{BC_a}$ method this difference is

$O(n^{-1})$. For multiparameter families, comparison of the approximate limits with exact ones is difficult, since a definition for exactness is not clear cut. See Bickel (1987) and Hall (1988) for further discussion concerning accuracy of the $BC_a$ limits. Efron shows that, in certain circumstances, the $BC_a$ limits agree to order $O_p(n^{-1})$ with limits derived by Cox (1980) and McCullagh (1984).

It is possible to obtain an expression for $a$ which is appropriate for more general estimators. Let $\delta^i = n^{1/2}(\hat{\eta}^i - \eta^i)$, and suppose that to error of order $O(n^{-1})$, $E(\delta^i) = n^{-1/2}\lambda^i$, $\operatorname{cov}(\delta^i, \delta^j) = \lambda^{ij}$, and $\operatorname{cum}(\delta^i, \delta^j, \delta^k) = n^{-1/2}\lambda^{ijk}$, where $\operatorname{cum}(\delta^i, \delta^j, \delta^k)$ is the third order cumulant of $\delta^i$, $\delta^j$ and $\delta^k$. Let

$$\gamma(\eta) = n^{-1/2}\{(3\lambda_l^{ij}\lambda^{lk} - 2\lambda^{ijk})t_i t_j t_k\}/(\lambda^{ij}t_i t_j)^{3/2},$$

where $\lambda_l^{ij} = \partial\lambda^{ij}(\eta)/\partial\eta^l$; then $a$ should be chosen to satisfy $a = \gamma(\hat{\eta})/6 + O_p(n^{-1})$.

For an example that illustrates the accuracy of the various approximate procedures, consider a sample from the bivariate distribution having probability density function $f(x, y) = \eta^1(\eta^2)^c \exp\{-(\eta^1 x + \eta^2 y)\}/\{y^{1-c}\Gamma(c)\}$ for $x > 0$, $y > 0$, and let $\theta = \eta^2/\eta^1$. In this case, $\gamma = 2(c-1)/\{n(c^2+c)\}^{1/2}$. Table 3 compares the upper and lower 97.5% percentile, BC and $BC_a$ limits with the exact limits in the case $n = 5$ and $c = 0.4$. The $BC_a$ method is very accurate in this situation. In the case $c = 1$, the percentile limits are exact.

### 2.3. *Related Procedures*

The $BC_a$ method has been criticized because it is not fully automatic; that is, it requires the calculation of the analytic adjustment $a$. DiCiccio and Romano (1987) have considered related procedures which closely approximate the $BC_a$ method, but do not require the explicit calculation of adjustments like $z_0$ and $a$.

Consider the simple case of a family indexed by a scalar parameter $\theta$, and suppose there exists a transformation satisfying (2.2). The exact upper $1-\alpha$ confidence limit $\theta[1-\alpha]$ can be found by the formula

$$G_{\hat{\theta}}^{-1}\{G_{\theta_0'}(\theta_0)\}, \tag{2.4}$$

where $\theta_0$ is any value of the parameter $\theta$ and $\theta_0' = G_{\theta_0}^{-1}(\alpha)$. Moreover, formula (2.4) is often exact when (2.2) is not exactly satisfied. As an example, this formula produces exact confidence limits for $\theta$ when $\hat{\theta}/\theta$ is pivotal; the $BC_a$ method is not exact in this case.

In practice, a reasonable choice for the initial value $\theta_0$ is the percentile limit $G_{\hat{\theta}}^{-1}(1-\alpha)$. Table 1 shows the approximate limits obtained by using this choice and

TABLE 3

*Upper and lower 97.5% confidence limits for $\theta = \eta^2/\eta^1$ ($n = 5$)*

| *Exact* | | *Percentile* | | *BC* | | *$BC_a$* | |
|---|---|---|---|---|---|---|---|
| *Lower* | *Upper* | *Lower* | *Upper* | *Lower* | *Upper* | *Lower* | *Upper* |
| $0.113\hat{\theta}$ (2.50) | $4.468\hat{\theta}$ (2.50) | $0.224\hat{\theta}$ (8.10) | $8.844\hat{\theta}$ (0.25) | $0.184\hat{\theta}$ (5.87) | $6.506\hat{\theta}$ (0.76) | $0.105\hat{\theta}$ (2.18) | $4.404\hat{\theta}$ (2.61) |

Below each confidence limit is shown its true error rate in coverage expressed as a percentage. $z_0 = -0.1217$, $a = -0.1195$.

(2.4) for the correlation coefficient example. In the exponential mean example, (2.4) is exact. An important feature of this procedure is that it can be iterated according to $\theta_{i+1} = G_{\theta}^{-1}\{G_{\theta_i'}(\theta_i)\}(i = 0, 1, \ldots)$, where $\theta_i' = G_{\theta_i}^{-1}(\alpha)$. Under certain conditions the difference between $\theta_i$ and $\theta[1-\alpha]$ is $O_p(n^{-1-i/2})$. Moreover, the numbers of calculations involved in this iterative procedure are linear in $i$, unlike other iterative methods discussed in Section 4.

For the extension of this method to multiparameter families, suppose that the model has been parameterized so as to be indexed by $(\theta, \psi)$, where $\psi = (\psi^1, \ldots, \psi^{p-1})$ consists of nuisance parameters orthogonal to $\theta$, i.e. $\operatorname{cov}\{n^{1/2}(\hat{\theta} - \theta), n^{1/2}(\hat{\psi}^i - \psi^i)\} = O(n^{-1})$. Cox and Reid (1987) give a detailed account of orthogonal parameters with reference to maximum likelihood estimation. To implement the method, commence with some initial value $\theta_0$, perhaps the percentile limit $\theta_0 = G_{(\hat{\theta},\hat{\psi})}^{-1}(1-\alpha)$, let $\theta_0' = G_{(\theta_0,\hat{\psi})}^{-1}(\alpha)$, and then take $\theta_1 = G_{(\hat{\theta},\hat{\psi})}^{-1}\{G_{(\theta_0',\hat{\psi})}(\theta_0)\}$. The difference between the $\mathrm{BC_a}$ limit and $\theta_1$ is $O_p(n^{-3/2})$, and the error in coverage of the approximate limit $\theta_1$ is $O(n^{-1})$. However, further iteration in the multiparameter case does not improve coverage accuracy.

In practice, it can be inconvenient to determine an orthogonal parameterization directly, and the preceding procedure can be sufficiently well approximated by using the least favourable family. In terms of the notation introduced previously for an estimator $\hat{\eta}$, consider the line $\eta(\tau) = \hat{\eta} + \tau\hat{\mu}$, where $\mu^i = \lambda^{ij}t_j$, $\hat{\mu}^i = \mu^i(\hat{\eta})$, and $\hat{\mu} = (\hat{\mu}^1, \ldots, \hat{\mu}^p)$. Let $G_\tau(s) = P_{\hat{\eta}+\tau\hat{\mu}}(\hat{\theta} \leqslant s)$, so that $G_0$ is the bootstrap distribution for $\hat{\theta} = t(\hat{\eta})$. Starting with an initial value $\theta_0$, perhaps taken to be the percentile limit $G_0^{-1}(1-\alpha)$, the approximate limit is $G_0^{-1}\{G_{\tau_0}(\theta_0)\}$, where $\tau_0$ is the value of $\tau$ such that $\theta_0 = t\{\eta(\tau)\}$, $\theta_0' = G_{\tau_0}^{-1}(\alpha)$, and $\tau_0'$ is the value of $\tau$ satisfying $\theta_0' = t\{\eta(\tau)\}$. This approximate limit differs from the one obtained using the orthogonal parameterization by terms of order $O_p(n^{-3/2})$.

For the multiparameter problem considered in Table 3, this procedure is exact. However, in the case of a sample drawn from the normal distribution with unknown mean $\mu$ and variance $\sigma^2$ where $\mu$ is the parameter of interest, the percentile method and related procedures perform poorly for small sample sizes. The approximate limits for $\mu$ given by these methods are the exact ones that would be obtained if the variance was known and equal to $\hat{\sigma}^2$.

A procedure which does give the correct limits in the normal mean example is the percentile-$t$ method. This method, further discussed in Sections 3 and 4, makes use of the bootstrap distribution of an approximately pivotal quantity instead of using the bootstrap distribution for $\hat{\theta}$. In the multiparameter context, consider the approximate pivot $(\hat{\theta} - \theta)/(\hat{\lambda}^{ij}\hat{t}_i\hat{t}_j)^{1/2}$, and let $K_\eta$ be its distribution function. Thus $K_\eta(s) = P_\eta\{(\hat{\theta} - \theta)/(\hat{\lambda}^{ij}\hat{t}_i\hat{t}_j)^{1/2} \leqslant s\}$, and the bootstrap distribution is $K_{\hat{\eta}}$. The percentile-$t$ approximation to an upper $1-\alpha$ confidence limit for $\theta$ is $\hat{\theta} - (\hat{\lambda}^{ij}\hat{t}_i\hat{t}_j)^{1/2}K_{\hat{\eta}}^{-1}(\alpha)$. The difference between the true and nominal coverage levels of the percentile-$t$ limits is $O(n^{-1})$. Although these limits are not invariant under reparameterization, the difference between the percentile-$t$ and $\mathrm{BC_a}$ limits is $O_p(n^{-3/2})$. For the examples considered in Tables 2 and 3, the percentile-$t$ method produces exact confidence limits. However, for the correlation coefficient example, with $\operatorname{var}(\hat{\theta}) = (1-\theta^2)^2/(n-1) + O(n^{-2})$, this method produces poor approximations as shown in Table 1.

The percentile-$t$ method can be extended to the construction of approximate confidence regions for vector-valued parameters of interest, which has been discussed by Hall (1988); such an extension for the $\mathrm{BC_a}$ method has yet to be as fully developed.

For the case of maximum likelihood estimation, it may be appealing to Studentize by using observed rather than expected information in deriving the percentile-$t$ limits. Beran (1987) considers the use of bootstrap distributions for log-likelihood ratio test statistics.

### 2.4. *Nonparametric Inference*

Suppose that $x_n = (X_1, \ldots, X_n)$ is a sample from an unknown distribution $F$, and suppose that $\theta = T(F)$ is the scalar parameter of interest. Let $\hat{F}_n$ be the empirical distribution function of $x_n$, and consider the estimator $\hat{\theta} = T(\hat{F}_n)$ of $\theta$. Let $G_F(s) = P_F(\hat{\theta} \leqslant s)$ be the distribution function of $\hat{\theta}$ under $F$; then the bootstrap distribution for $\hat{\theta}$ is $G_{\hat{F}_n}$.

The percentile and BC limits are easily found in this situation. The approximate upper $1-\alpha$ confidence limits for $\theta$ obtained by the percentile and BC methods are $G_{\hat{F}_n}^{-1}(1-\alpha)$ and $G_{\hat{F}_n}^{-1}\{\Phi(z^{(1-\alpha)} + 2z_0)\}$, respectively, where $z_0 = \Phi^{-1}\{G_{\hat{F}_n}(\hat{\theta})\}$. The implementation of the $\mathrm{BC}_a$ method is less straightforward since it requires the calculation of the adjustment $a$.

To calculate $a$, Efron reduces the nonparametric situation to a multiparameter one by restricting attention to distribution functions with support on $X_1, \ldots, X_n$. Corresponding to each such distribution $F$ is a vector $w = (w^1, \ldots, w^n)$, where $w^i$ is the probability mass assigned to $X_i$. Efron considers the problem of setting confidence limits for $\theta = \theta(w)$ having made $n$ draws from such an $n$-category multinomial distribution and observed each category to appear once; that is $\hat{w} = (n^{-1}, \ldots, n^{-1})$. The least favourable family in this situation is $w(\tau) = (w^1(\tau), \ldots, w^n(\tau))$, where

$$w^i(\tau) = \mathrm{e}^{\tau U_i} \Big/ \left\{ \sum_{j=1}^{n} \mathrm{e}^{\tau U_j} \right\}$$

and

$$U_i = \lim_{\Delta \to 0} \frac{T\{(1-\Delta)\hat{F}_n + \Delta\delta_i\} - T(\hat{F}_n)}{\Delta}$$

is the $i$th component of the empirical influence function with $\delta_i$ denoting a point mass at $X_i$. The formula for the acceleration adjustment is

$$a = \left( \sum_{i=1}^{n} U_i^3 \right) \Big/ \left\{ 6 \left( \sum_{i=1}^{n} U_i^2 \right)^{3/2} \right\}.$$

Efron suggests the $U_i$s be calculated numerically, and he has used $\Delta = 0.001$ in practice.

Hall (1988) has considered these procedures in the nonparametric context of a 'smooth function' model for which the estimator can be expressed as a function of multivariate vector means. In such cases, the difference between the nominal and true coverage levels of the approximate limits is $O(n^{-1/2})$ for the percentile and BC methods and is $O(n^{-1})$ for the $\mathrm{BC}_a$ method. Hall (1988) also considers the percentile-$t$ method for this model, and he shows that the percentile-$t$ limits differ from the $\mathrm{BC}_a$ limits by terms of order $O_P(n^{-3/2})$. Hall (1987) considers the percentile-$t$ method for vector-valued parameters in the nonparametric context.

The procedure given by expression (2.4) and extended to the multiparameter families in Section 2.3 can be applied in an obvious way to the nonparametric case by making use of the least favourable family $w(\tau)$.

## 3. FUNCTIONAL APPROACH

### 3.1. *Consistency*

In this section, the bootstrap is motivated as a natural functional approach to the construction of a confidence region. The development begins by focusing on the independent identically distributed case.

Let $x_n = (X_1, \ldots, X_n)$ be a sample of $n$ random variables taking values in a sample space $S$ and having unknown distribution $F$, where $F$ is assumed to belong to a certain collection $\mathbf{F}$ of distributions. The collection $\mathbf{F}$ may be finite or infinite dimensional. The interest lies in constructing a confidence interval for some parameter $T(F)$, whose range $\{T(F): F \in \mathbf{F}\}$ will be denoted $\mathbf{T}$. This leads to considering a root $R_n(x_n, T(F))$, which is just some functional depending on both $x_n$ and $T(F)$. For example, an estimator $T_n$ of a real-valued parameter $T(F)$ might be given so that a natural choice is $R_n(x_n, T(F)) = T_n - T(F)$, or alternatively, $R_n(x_n, T(F)) = [T_n - T(F)]/s_n$, where $s_n$ is some estimate of the standard deviation of $T_n$.

When $\mathbf{F}$ is suitably large, a natural construction for an estimator $T_n$ of $T(F)$ is $T_n = T(\hat{F}_n)$, where $\hat{F}_n$ is the empirical measure of $X_1, \ldots, X_n$. In regular parametric problems for which $\mathbf{F}$ is indexed by a parameter $\eta$ belonging to a subset of $\mathbb{R}^p$, $T(F)$ can be described as a parameter $t(\eta)$, and hence $T_n$ is often taken to be $T_n = t(\hat{\eta}_n)$, where $\hat{\eta}_n$ is some desirable estimate of $\eta$, such as a maximum likelihood estimate, a one-step maximum likelihood estimate, a minimum distance estimate, etc.

Let $J_n(F)$ be the law of $R_n(x_n, T(F))$ when $x_n = (X_1, \ldots, X_n)$ is a random sample from $F$, and let $J_n(x, F)$ be the corresponding cumulative distribution function. Also, let $J_n^{-1}(\alpha, F) = \inf\{x: J_n(x, F) \geqslant \alpha\}$ be an $\alpha$ quantile of the law $J_n(F)$. In order to construct a confidence region for $T(F)$, the sampling distribution or the appropriate quantiles of $J_n(F)$ must be known or estimated. The bootstrap procedure is to estimate $J_n(F)$ by $J_n(\hat{G}_n)$, where $\hat{G}_n$ is some estimate of $F$, and then estimate the appropriate quantiles of $J_n(F)$ by those of $J_n(\hat{G}_n)$. In nonparametric problems, $\hat{G}_n$ is typically (but not always) taken to be the empirical distribution $\hat{F}_n$; in parametric problems, $\hat{G}_n$ is usually $G_{\hat{\eta}_n}$. A resulting bootstrap confidence region for $T(F)$ takes the form

$$B_n(\alpha, x_n) = \{t \in \mathbf{T}: R_n(x_n, t) < J_n^{-1}(1-\alpha, \hat{G}_n)\}. \tag{3.1}$$

Note that when $J_n(F)$ is independent of $F$, the root $R_n(x_n, T(F))$ is said to be a pivot, in which case the bootstrap procedure is clearly valid. In general, it may not be possible to find an exact pivot. Some of the bootstrap literature refers to this procedure of estimating $J_n(F)$ by $J_n(\hat{G}_n)$ as the bootstrap pivotal method, but the term root is adapted here from Beran (1987) to distinguish a general root from a pivot defined in the classical sense.

How well does this bootstrap procedure work? Is it consistent, or perhaps optimal in any sense? Typically, one can show $J_n(F)$ converges weakly to a continuous limit law $J(F)$. In order for the bootstrap to be valid, $J_n(F)$ must be smooth in $F$. Smoothness in $F$ can often be described in terms of a suitable metric $d$, depending on the choice of root $R_n$. Specifically, let $d$ be a metric on $\mathbf{F}$ such that the estimate $\hat{G}_n$ of $F$ satisfies $d(\hat{G}_n, F) \to 0$ in probability as $n \to \infty$. Then, a sufficient condition for the bootstrap to be consistent is that the convergence of $J_n(F)$ to $J(F)$ be locally uniform in $F$, where uniformity is described in terms of $d$. Formally, the following uniform weak convergence result (denoted TAC from Beran (1987) for triangular array convergence) must hold.

*Triangular Array Convergence.* If $\{F_n, F_n \in \mathbf{F}\}$ is any sequence of distributions in $\mathbf{F}$ satisfying $d(F_n, F) \to 0$, then $J_n(F_n)$ converges weakly to a continuous limit law $J(F)$, which depends only on $F$.

When TAC holds and $d(\hat{G}_n, F) \to 0$ in probability, it follows that

$$\sup_x |J_n(x, \hat{G}_n) - J_n(x, F)| \to 0 \text{ in probability.} \tag{3.2}$$

Moreover, the continuity of the limit law $J(F)$ entails the convergence in probability of $J_n^{-1}(\alpha, \hat{G}_n)$ to the $\alpha$ quantile, $J^{-1}(\alpha, F)$ of $J(F)$. It follows (see Beran (1984a), theorem 1) that a bootstrap confidence region $B_n$ given by (3.1) satisfies, for any $F$ in $\mathbf{F}$,

$$P_F\{T(F) \in B_n(\alpha, x_n)\} \to 1 - \alpha \quad \text{as} \quad n \to \infty. \tag{3.3}$$

Several papers in the bootstrap literature show that the TAC condition holds. See, for example, Babu and Singh (1983), Beran (1984a), Beran and Millar (1985), Beran and Srivastava (1985), Bickel and Freedman (1981), Ducharme *et al.* (1985), Freedman (1981), Romano (1987) and Singh (1981). Two well-studied examples are the following.

*Example 3.1.* Let $\mathbf{F}$ be the collection of distributions on the line with finite variance. The problem is to construct a confidence interval for $T(F) = \int x \, dF(x)$. Here,

$$R_n(x_n, F) = n^{1/2} |T(\hat{F}_n) - T(F)|,$$

where $\hat{F}_n$ is the empirical distribution so that $T(\hat{F}_n)$ is the sample mean. Then, TAC holds when $d$ is, for example, Mallow's metric $d_2$ defined by: $d_2^2(F, G)$ is the infimum of $E(X - Y)^2$ over all joint distributions of $X$ and $Y$ whose fixed marginals are $F$ and $G$, respectively. Convergence of $d(F_n, F)$ to 0 is equivalent to $F_n$ converging weakly to $F$ and $\mathrm{Var}(F_n) \to \mathrm{Var}(F)$. See Bickel and Freedman (1981).

*Example 3.2.* Let $\mathbf{F}$ be the collection of all distributions $F$ on the line. The problem is to construct a confidence region for $F$ based on the root

$$R_n(x_n, F) = n^{1/2} \sup_t |\hat{F}_n(t) - F(t)|,$$

where $\hat{F}_n$ is the empirical measure based on $x_n = (X_1, \ldots, X_n)$. Then, the TAC holds with

$$d(F, G) = \sup_t |F(t) - G(t)|,$$

the usual Kolmogorov distance. This example was first considered in Bickel and Freedman (1981) and has been generalized to constructing a confidence set for a measure in an arbitrary sample space $S$ based on a supremum distance over a Vapnik–Cervonenkis class of sets; see Beran and Millar (1987).

Based on the validity of the bootstrap in example 2, it is natural to expect the asymptotic correctness of bootstrap confidence intervals for smooth functionals $T(F)$ of $F$, since the distribution of $n^{1/2}\{T(\hat{F}_n) - T(F)\}$ can often be approximated by some smooth functional of the empirical process. In particular, suppose $T$ is a real-valued functional defined on a large class $\mathbf{F}$ of distribution functions on the real line. Assume that $T$ is Fréchet differentiable; that is, for each fixed $F$, there exists a function $\psi_F$

such that

$$T(G) - T(F) = \int \psi_F \mathrm{d}(G - F) + o(|G - F|_\infty)$$

where $|G - F|_\infty$ is the supremum norm between distribution functions. Suppose further that

$$\int \psi_F^2 \, \mathrm{d}F < \infty \tag{3.4}$$

and

$$\int (\psi_G - \psi_F)^2 \mathrm{d}G = O(|G - F|_\infty). \tag{3.5}$$

*Proposition 3.1.* If $T$ is Fréchet differentiable with derivative $\psi_F$ satisfying (3.4) and (3.5), then TAC holds for the root

$$R_n(x_n, T(F)) = n^{1/2}\{T(\hat{F}_n) - T(F)\}$$

when $d(F, G) = |F - G|_\infty$. Hence, (3.2) and (3.3) hold when resampling from $\hat{G}_n = \hat{F}_n$, the empirical distribution.

An argument for this proposition is essentially given in Bickel and Freedman (1981). In fact, the result holds under the weaker assumption that $T$ is compactly differentiable with derivative $\psi_F$ satisfying (3.4) and (3.5). See Liu *et al.* (1986) for a discussion of the negligibility of the remainder term, although the TAC condition is not made explicit. Proposition 3.1 applies to many $M$ estimators, $L$ estimators and $U$ statistics, for example.

Extensions of the bootstrap outside of the independent and identically distributed framework are clearly possible. Given a root $R(x_n, T(F))$ based on data $x_n$ with distribution $F$, let $J_n(F)$ be the distribution of the root. As before, the bootstrap procedure is to estimate $J_n(F)$ by $J_n(\hat{G})$ for some suitable estimate $\hat{G}$. Freedman (1981) discusses regression models, Freedman (1984) and Haycock (1986) discuss time series models, and Sugahara (1987) discusses Markov chain models. The simplest example outside of the independent and identically distributed context is confidence intervals for two-sample problems.

*Example 3.3.* Let $x_n$ be a random sample of size $n$ from $F$ and let $y_m$ be an independent random sample of size $m$ from $G$. The problem is to construct a confidence interval for $T(F) - T(G)$ where $T$ is some functional (such as the mean or a quantile). Let

$$R_{nm}(x_n, y_m, T(F), T(G)) = [T(\hat{F}_n) - T(\hat{G}_m)] - [T(F) - T(G)],$$

where $\hat{F}_n$ is an estimate of $F$ based on $x_n$ and $\hat{G}_m$ is an estimate of $G$ based on $y_m$. Let $J_{nm}(F, G)$ be the law of $R_{nm}(x_n, y_m, T(F), T(G))$ when $x_n$ is a sample of size $n$ from $F$ and $y_m$ is a sample of size $m$ from $G$. The bootstrap procedure approximates the appropriate quantiles of $J_{nm}(F, G)$ by those of $J_{nm}(\hat{F}_n, \hat{G}_m)$. As in proposition 3.1, smoothness assumptions on $T$ yield consistency by taking $\hat{F}_n$ and $\hat{G}_m$ the empirical distributions corresponding to $x_n$ and $y_m$. With respect to example 3.2, one may also construct a confidence band for the difference $F - G$; see Beran (1984a).

### 3.2. *Optimality*

We have seen that smoothness of $J_n(F)$ in $F$ entails the consistency of the bootstrap. A natural question to ask is whether the bootstrap estimate $J_n(\hat{G}_n)$ of $J_n(F)$ is optimal in any sense. In nonparametric problems where $\mathbf{F}$ is the class of all distribution functions, it is well known that the empirical distribution function $\hat{F}_n$ is an optimal estimator of $F$ in a local asymptotic minimax (LAM) sense. One would expect similar optimality results for smooth functionals of $F$, such as $J_n(F)$; such results are well established in the literature, with a unifying approach developed in Millar (1983). In a decision-theoretic framework, Beran (1982) establishes the LAM property for the bootstrap distribution estimate $J_n(\hat{F}_n)$, assuming a locally uniform first-order Edgeworth expansion for $J_n(F)$ (which entails a certain differentiability property in $F$ for $J_n(F)$). Bootstrap estimates of bias, variance and skewness, which can be viewed as functionals of $J_n(F)$, are also typically LAM; see Beran (1984a) for details.

Having derived optimality properties for the bootstrap estimate $J_n(\hat{F}_n)$ of $J_n(F)$, a natural question is: do the resulting confidence sets possess any optimality property? In Beran and Millar (1985), a general asymptotic theory of optimal confidence sets is presented. In their framework, confidence sets should not only have the approximate stated level, but should also be properly centred and small in size. In particular, suppose the problem is to construct a confidence set for $T(F)$, where $T$ takes values in a Banach space $B$ with norm $|\cdot|_B$. If $\hat{C}_n$ is a confidence set for $T(F)$, the risk of $\hat{C}_n$ is given by

$$\mathbf{E}_F\{\sup\, g(|\, t - T(F)\,|_B) : t \in \hat{C}_n\}$$

for some increasing function $g$. The goal is to find a procedure $\hat{C}_n$, a ball in $B$ specified by its centre and radius, that achieves the minimum LAM risk subject to the constraint on level. Uner suitable conditions, the confidence set

$$\hat{C}_n = \{t : |\, t - T(\hat{F}_n)\,|_B \leqslant \hat{r}_n\}$$

is LAM, where $\hat{r}_n$ is determined by bootstrapping the root $R_n(x_n, T(F)) = |\, T(\hat{F}_n) - T(F)\,|_B$. Their framework applies to both nonparametric and parametric problems.

### 3.3. *Inconsistency*

The bootstrap need not be consistent, even in the weakest sense (3.3). The problem is due to lack of uniformity in $F$ of $J_n(F)$. Bickel and Freedman (1981) give two such examples. Other counterexamples may be found in Beran (1982) and Athreya (1987). Ghosh *et al.* (1984) discuss the bootstrap estimate of the variance of the sample median $T(\hat{F}_n)$, where $T(F) = \inf\{t : F(t) > 1/2\}$. They show it is possible for (3.3) to hold based on the root

$$R_n(x_n, T(F)) = n^{1/2}\{T(\hat{F}_n) - T(F)\};$$

however, the bootstrap estimate of variance of $R_n(x_n, F)$ is inconsistent. Although Studentization typically improves coverage accuracy, their results suggest that using a root based on Studentization can be inappropriate. The problem lies in the fact that the variance functional is not weakly continuous. The following example shows the problems that can occur when constructing bootstrap confidence intervals for functionals of a density.

*Example 3.4.* Given a sample $x_n = (X_1, \ldots, X_n)$ of $n$ observations from a distribution $F$ having density $f$, the problem is to construct a confidence interval for $T(F) = f(t)$ for some fixed $t$. Let $\hat{f}_{nh}(t)$ be a kernel density estimate of $f$ given by

$$\hat{f}_{nh}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h}\right)$$

for some appropriate smooth kernel $K$, and let $\hat{F}_{nh}$ be the corresponding distribution satisfying $\hat{F}_{nh}^{(1)} = \hat{f}_{nh}$. The density $f$ is assumed to be smooth (say, having two continuous derivatives) but is otherwise unknown. Let

$$R_{nh}(x_n, T(F)) = (nh)^{1/2}\{\hat{f}_{nh}(t) - f(t)\}.$$

Let $J_{nh}(F)$ be the distribution of $R_{nh}(x_n, T(F))$ under $F$. The optimal rate at which $h = h_n$ should be chosen is well known to be $nh_n^5 \to q$ for some $q > 0$. For such a choice of $h_n$, the following uniformity result holds (which can easily be translated into the TAC condition). If $G_n$ is a distribution having density $g_n$ such that the first two derivatives of $g$ converge uniformly to those of $f$ in a neighbourhood of $t$, then $J_{nh_n}(G_n)$ converges weakly to a continuous limit law $J(F)$. Hence, if $nb_n^5/\log(n) \to \infty$ and $b_n \to 0$, then $J_{nh_n}(\hat{F}_{nb_n})$ is a consistent bootstrap estimate so that (3.2) and (3.3) hold. On the other hand, when $b_n = h_n$, the second derivative of the kernel density estimate $\hat{f}_{nh_n}$ is not a consistent estimate of $f^{(2)}$ and failure of the bootstrap results. Unfortunately, the appropriate class of distributions which one may resample from often depends heavily on the parameter of interest when this parameter is a local property of the distribution. This emphasizes the need to be careful in a naive application of the bootstrap when bootstrapping functionals of a density; see Romano (1988).

Finally, we remark that bootstrap confidence intervals may be consistent in the sense that (3.3) holds for any fixed $F$, but the convergence is not uniform over $F$. That is, define the level of a confidence set $\hat{C}_n$ for $T(F)$ to be

$$\inf_F P_F\{T(F) \in \hat{C}_n\}.$$

For fixed $F$, one may have $P_F(T(F) \in \hat{C}_n)$ tends to $1 - \alpha$ (or at least $1 - \alpha$), but the level converges to a number less than $1 - \alpha$. The following example, adapted from Romano (1986) shows this can happen even in a parametric setting; see Romano (1986) for a nonparametric example.

*Example 3.5.* Let $x_n = (X_1, \ldots, X_n)$ be independent Bernoulli variables such that $P_\theta(X_i = 1) = \theta$. The problem is to construct a confidence interval for $\theta$. Let

$$R_n(x_n, \theta) = n^{1/2}(\hat{\theta}_n - \theta)$$

where $\hat{\theta}_n$ is the sample mean. Let $J_n(\theta)$ be the distribution of $R_n(x_n, \theta)$ under $\theta$, and let $B_n$ be the bootstrap confidence set

$$B_n(\alpha, x_n) = \{\theta: R_n(x_n, \theta) \leqslant J_n^{-1}(1 - \alpha, \hat{\theta}_n)\}.$$

Fix any $\alpha$ in (0, 1), and $\beta$ in $(\alpha, 1)$ and choose $\theta_n > 0$ so $(1 - \theta_n)^n \geqslant \beta$. If $x_n$ is a sample from $P_{\theta_n}$, the chance that all observations are zero is at least $\beta$. In such a case, the resulting bootstrap confidence set is just the set $\{0\}$ and does not contain the true

value $\theta_n$. It follows that

$$\limsup_{n\to\infty} P_{\theta_n}\{\theta \in B_n(\alpha, x_n)\} \leqslant 1-\beta.$$

For fixed $\alpha$, $\beta$ was arbitrary, so in fact the level tends to zero.

## 4. ITERATIVE BOOTSTRAP REFINEMENTS

### 4.1. *Introduction and Studentized Roots*

Bootstrap methods discussed in Section 3 offer a widely applicable construction of confidence sets. In this Section, we review some refinements of the method, with the goal of improved coverage accuracy. Other methods, based in part on analytical corrections, are discussed in Abramovitch and Singh (1985), Hall (1983) and Withers (1983).

In order to construct a bootstrap confidence set (3.1), a choice of root must be specified. In particular, consider the root

$$R_n(x_n, T(F)) = n^{1/2}(T_n - T(F))$$

and the Studentized root

$$S_n(x_n, T(F)) = n^{1/2}(T_n - T(F))/s_n,$$

where $T_n$ is some estimate of $T(F)$ and $n^{1/2}s_n$ is some (consistent) estimate of the (asymptotic) standard deviation of $T_n$. In large samples, the distribution of $R_n$ under $F$, say $J_n(\cdot, F)$, is typically normal with mean 0 and variance $\sigma^2(F)$, which depends on $F$. On the other hand, the distribution of $S_n$ under $F$, say $K_n(\cdot, F)$, is asymptotically standard normal, and so is at least asymptotically pivotal. This suggests the (finite sample) distribution of $S_n$ is less dependent on $F$ than that of $R_n$. Indeed (see Beran (1982) or Hall (1988), for example), in regular cases, bootstrapping a Studentized root results in improved coverage accuracy. In particular,

$$J_n(x, F) - J_n(x, \hat{F}_n) = O_F(n^{-1/2})$$

and

$$K_n(x, F) - K_n(x, \hat{F}_n) = O_F(n^{-1}),$$

and the same orders of coverage error apply to the corresponding one-sided bootstrap confidence intervals. It should be noted, however, that coverage error of $O_F(n^{-1/2})$ based on the root $R_n$ is not really a fault of the bootstrap as Beran (1982) has shown that there does not exist an estimate of $J_n(\cdot, F)$ that can improve upon this rate of convergence. Rather, the problem lies in a perhaps poor choice of root.

Of course, the drawback of bootstrapping a Studentized root is that one must have a consistent estimate of $\sigma(F)$. One possibility is to use a bootstrap estimate of variance, but the resulting confidence procedure would involve nested bootstrap calculations, and so may be computationally undesirable. Moreover, as remarked in Section 3.3, bootstrap estimates of variance typically are, but need not be, consistent.

### 4.2. *Prepivoting*

As in Section 3, let $J_n(F)$ be the distribution function of a root $R_n(x_n, T(F))$. The resulting bootstrap confidence region $B_n(\alpha, x_n)$ is given by (3.1). Letting

$$R_{n1}(x_n, T(F)) = J_n\{R_n(x_n, T(F)), \hat{G}_n\},$$

we may write $B_n(\alpha, x_n)$ as

$$B_n(\alpha, x_n) = \{t \in \mathbf{T}: R_{n1}(x_n, t) < 1 - \alpha\}. \tag{4.1}$$

Let $J_{n1}(\cdot, F)$ be the distribution function of $R_{n1}(x_n, T(F))$ under $F$. In effect, confidence set (4.1) approximates the distribution $J_{n1}(\cdot, F)$ by the uniform distribution. Instead, analogous to (3.1), one may estimate the distribution $J_{n1}(\cdot, F)$ by $J_{n1}(\cdot, \hat{G}_n)$ to obtain a new confidence set:

$$B_{n1}(\alpha, x_n) = \{t \in \mathbf{T}: R_{n1}(x_n, t) < J_{n1}^{-1}(1 - \alpha, \hat{G}_n)\}. \tag{4.2}$$

This construction is Beran's (1987) method of prepivoting. Instead of bootstrapping the root $R_n(x_n, t)$, a new root, $R_{n1}(x_n, t)$, is formed (by using the bootstrapping distribution of $R_n(x_n, t)$!) and the method of Section 3 is applied to form a confidence set based on this new root. Beran argues that $R_{n1}$ is more nearly pivotal than $R_n$; that is, the distribution $J_{n1}(\cdot, F)$ is less dependent on $F$ than is $J_n(\cdot, F)$. This seems plausible because one typically begins by choosing an asymptotically normal root $R_n$ whose asymptotic variance depends on $F$. The asymptotic distribution of $R_{n1}$ is uniform and does not depend on $F$. Moreover, beginning with an asymptotically normal root

$$R_n(x_n, T(F)) = n^{1/2}\{T_n - T(F)\},$$

Beran argues that the prepivoting operation is asymptotically equivalent to Studentizing. The beauty of prepivoting is the generality of the approach.

The motivation for prepivoting may be derived from the point of view of Loh's (1987) calibrated confidence sets. In fact, we show these methods to be the same. In particular, consider the following refinement of the bootstrap confidence set (4.1). The exact coverage probability that $T(F)$ is contained in $B_n(\alpha, x_n)$ under $F$ is given by $J_{n1}(1 - \alpha, F)$, and is unknown because $F$ is unknown. We can estimate this coverage probability by $J_{n1}(1 - \alpha, \hat{G}_n)$. Now, choose $\alpha_1$ so that this estimated coverage is $1 - \alpha$; that is, solve

$$J_{n1}(1 - \alpha_1, \hat{G}_n) = 1 - \alpha. \tag{4.3}$$

Compute a new confidence set as before, except use $\alpha_1$ instead of $\alpha$ to obtain

$$\{t \in \mathbf{T}: R_{n1}(x_n, t) < 1 - \alpha_1\}. \tag{4.4}$$

By (4.3), $1 - \alpha_1 = J_{n1}^{-1}(1 - \alpha, \hat{G}_n)$, so that (4.2) and (4.4) agree.

The method of prepivoting may be iterated as follows. Given a root $R_{nj}(x_n, T(F))$, let $J_{nj}(\cdot, F)$ be its distribution function under $F$. Form a new root $R_{n,j+1}$ by

$$R_{n,j+1}(x_n, T(F)) = J_{nj}\{R_{nj}(x_n, T(F)), \hat{G}_n\}$$

and let

$$B_{nj}(\alpha, x_n) = \{t \in \mathbf{T}: J_{nj}(R_{n,j+1}(x_n, t), \hat{G}_n) < 1 - \alpha\}.$$

Beran argues that, when Edgeworth expansions for $J_n(\cdot, F)$ exist, the error level of $B_{nj}(\alpha, x_n)$ decreases as $j$ increases.

### 4.3. *Hall's Iterative Additive Correction*

Let $S_{n1}(x_n, t) = R_n(x_n, t) - J_n^{-1}(1-\alpha, \hat{G}_n)$. Rewrite confidence set (3.1) in the form

$$B_n(\alpha, x_n) = \{t \in \mathbf{T}: S_{n1}(x_n, t) < 0\}. \tag{4.5}$$

Let $L_{n1}(\cdot, F)$ be the distribution function of $S_{n1}(x_n, T(F))$ under $F$. The confidence set (4.5) pretends that $L_{n1}^{-1}(1-\alpha, F) = 0$. Instead one may estimate $L_{n1}^{-1}(1-\alpha, F)$ by $L_{n1}^{-1}(1-\alpha, \hat{G}_n)$ and form a new confidence set

$$A_{n1}(\alpha, x_n) = \{t \in \mathbf{T}: S_{n1}(x_n, t) < L_{n1}^{-1}(1-\alpha, \hat{G}_n)\}.$$

This is Hall's (1986a) additive correction. Like prepivoting, this procedure may be iterated as follows. Let

$$S_{n,j+1}(x_n, t) = S_{nj}(x_n, t) - L_{nj}^{-1}(1-\alpha, \hat{G}_n).$$

Let $L_{n,j+1}(\cdot, F)$ be the distribution function of $S_{n,j+1}(x_n, t)$ under $F$ and form

$$A_{n,j+1}(\alpha, x_n) = \{t \in \mathbf{T}: S_{n,j+1}(x_n, t) < L_{n,j+1}^{-1}(1-\alpha, \hat{G}_n)\}.$$

Like prepivoting, each round of iteration reduces the level error in regular cases.

### 4.4. *Conclusions*

Bootstrap iterative methods offer the potential of high accuracy, at least in regular problems. Of course, for fixed sample size, iteration may not improve accuracy and can make the situation worse. It is important to note that the currently available supporting theory is asymptotic in the sample size. Prepivoted confidence sets, unlike Hall's method, offer the advantage of being invariant under monotone transformations of the root and reparameterizations of $T(F)$. This can, in part, be explained by the fact that Beran's roots $R_{nj}$ obtained by prepivoting do not depend on the initial fixed choice of $\alpha$ whereas Hall's $S_{nj}$ do depend on $\alpha$. The current drawback with iterative methods is the inherent computational complexity. In some cases, however, prepivoting has analytical approximations; see Beran (1987). Otherwise, each round of iteration involves a nested bootstrap calculation, so that the number of computations is typically exponential in the number of iterations. In addition, even disregarding the computational problems, how many iterations are possible? Current asymptotics assume the number of iterations $j$ is fixed while $n \to \infty$. Clearly, $j$ cannot be comparable with $n$, but what is feasible? Two important open questions are the following. Can clever computational algorithms be designed to efficiently reduce the number of calculations as to make these methods viable? Second, do these iterative procedures converge and, if so, can one directly approximate the limiting algorithm?

## REFERENCES

Abramovitch, L. and Singh, K. (1985) Edgeworth corrected pivotal statistics and the bootstrap. *Ann. Statist.*, **13**, 116–132.

Athreya, K. (1987) Bootstrap of the mean in the infinite variance case. *Ann. Statist.*, **15**, 724–731.

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

Beran, R. (1982) Estimated sampling distributions: the bootstrap and competitors. *Ann. Statist.*, **10**, 212–225.

———(1984a) Bootstrap methods in statistics. *Jber. Dtsch. Math-Ver.*, **86**, 14–30.

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

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

Beran, R. and Millar, P. (1985) Asymptotic theory of confidence sets. In *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer*, II (eds L. Le Cam and R. Olshen), pp. 865–886. Belmont: Wadsworth.

———(1986) Confidence sets for a multivariate distribution. *Ann. Statist.*, **14**, 431–443.

Beran, R. and Srivastava, M. (1985) Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.*, **13**, 95–115.

Bickel, P. J. (1987) Discussion on Efron. *J. Amer. Statist. Ass.*, **82**, 191.

Bickel, P. J. and Freedman, D. A. (1981) Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**, 1196–1217.

Cox, D. R. (1980) Local ancillarity. *Biometrika*, **67**, 273–278.

Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc.* B, **49**, 1–39.

DiCiccio, T. and Romano, J. (1987) Accurate bootstrap confidence limits in parametric models. *Technical Report No. 281.* Department of Statistics, Stanford University.

DiCiccio, T. and Tibshirani, R. (1987) Bootstrap confidence intervals and bootstrap approximations. *J. Amer. Statist. Ass.*, **82**, 163–170.

Ducharme, G., Jhun, M., Romano, J. and Truong, K. (1985) Bootstrap confidence cones for directional data. *Biometrika*, **72**, 637–645.

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

———(1981) Non-parametric standard errors and confidence intervals. *Can. J. Statist.*, **9**, 139–172.

———(1982) The jackknife, the bootstrap, and other resampling plans. In *Regional Conference Series in Applied Mathematics*, No. 38. Philadelphia: SIAM.

———(1985) Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, **72**, 45–58.

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

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

Freedman, D. A. (1981) Bootstrapping regression models. *Ann. Statist.*, **9**, 1218–1228.

Freedman, D. (1984) On bootstrapping two-stage least-squares estimates in stationary linear models. *Ann. Statist.*, **12**, 827–842.

Ghosh, M., Parr, W., Singh, K. and Babu, G. (1984) A note on bootstrapping the sample median. *Ann. Statist.*, **12**, 1130–1135.

Hall, P. (1983) Inverting an Edgeworth expansion. *Ann. Statist.*, **11**, 569–576.

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

———(1986b) On the number of bootstrap simulations required to construct a confidence interval. *Ann. Statist.*, **4**, 1453–1462.

———(1987) On the bootstrap and likelihood-based confidence regions. *Biometrika*, **74**, 481–493.

———(1988) Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.*, to be published.

Haycock, K. (1986) Bootstrapping prediction error estimates in dynamic linear models. *PhD Dissertation.* Department of Statistics, University of California at Berkeley.

Hinkley, D. (1988) Bootstrap methods. *J. R. Statist. Soc.* B, **50**, 321–337.

Johns, V. (1988) Importance sampling for bootstrap confidence intervals. *J. Amer. Statist. Ass.*, to be published.

Liu, R., Singh, K. and Lo, S. (1986) On a representation related to the bootstrap. Preprint.

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

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

Millar, P. W. (1983) The minimax principle in asymptotic statistical theory. *Springer Lecture Notes*, **976**, 76–265.

Romano, J. (1986) A note on uniform convergence of the empirical measure with applications to simulation techniques. *Technical Report No. 263.* Department of Statistics, Stanford University.

———(1988) Bootstrapping the mode. *Ann. Inst. Statist. Math.*, to be published.

Schenker, N. (1985) Qualms about bootstrap confidence intervals. *J. Amer. Statist. Ass.*, **80**, 360–361.

———(1987) Discussion on Efron. *J. Amer. Statist. Ass.*, **82**, 192–194.

Singh, K. (1981) On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.*, **9**, 1187–1195.

Sugahara, C. N. (1987) Bootstrapping Markov chains. Forthcoming *PhD Dissertation.* Department of Statistics, University of California at Berkeley.

Withers, C. (1983) Expansion for the distribution and quantiles of a regular functional of the empirical distribution with applications to non-parametric confidence intervals. *Ann. Statist.*, **11**, 577–587.

*J. R. Statist. Soc.* **B** (1988)
**50**, *No.* 3, *pp.* 355–370

# DISCUSSION OF THE PAPERS BY HINKLEY AND DICICCIO AND ROMANO

**Dr J. T. Kent** (University of Leeds): The story of the bootstrap as a well-defined area of study starts with Efron's (1979) paper. Over the past 10 years the bootstrap has become one of the most popular recent developments in statistics. I can think of four reasons to help to explain why it has caught the imagination of the statistical public.

(a) Elegance: like many of the best ideas in mathematics, the principle behind the bootstrap, that of resampling from the empirical distribution function, is simple and elegant, yet very powerful.
(b) Packaging: the catchy name 'bootstrap' makes it easy for people to focus on the field, though it is perhaps confusing that the meaning of the term bootstrap has been expanded to include parametric resampling as well.
(c) Mathematical subtlety: there is sufficient complexity behind the bootstrap to lure the mathematical intellectuals on to the scene. The area has already attracted many deep thinkers, as these papers illustrate.
(d) Ease of use: in contrast, for the practitioner, there is the hope of a fairly automatic and straightforward methodology that can be used without the need for any thought.

Next I would like to comment on the language used for asymptotic bootstrap confidence intervals, which is a potential source of confusion. Let $\theta_{BS}[\alpha]$ be a bootstrap estimate of an exact $\alpha$ confidence limit $\theta_{EX}[\alpha]$ in a one-parameter model, based on $n$ observations. The following statements are roughly equivalent.

(a) $\theta_{BS}[\alpha]$ is accurate to second order.
(b) $\theta_{BS}[\alpha] - \theta_{EX}[\alpha] = O_p(n^{-3/2})$.
(c) $n^{1/2}\{\theta_{BS}[\alpha] - \theta_{EX}[\alpha]\} = O_p(n^{-1})$.
(d) $\theta_{BS}[\alpha]$ is accurate to order $n^{-1}$, i.e. the error is of higher order.
(e) The error in coverage probability using $\theta_{BS}[\alpha]$ is $O_p(n^{-1})$.

More specifically (a)–(c) are equivalent and imply (d) and (e). Amid this confusing choice of powers of $n$, DiCiccio and Romano emphasize the expressions (b) and (e), and I agree that these are the clearest way to express accuracy here.

Lastly, let me address the question of whether it is even meaningful to talk about second-order accuracy for confidence intervals in a nonparametric problem. To set the scene, consider observations from the model

$$N(\theta, (\sigma + a\theta)^2).$$

If the acceleration parameter $a$ is known then Efron's $BC_a$ percentile method is precisely what is needed here to produce confidence intervals accurate to second order. However, the case of unknown $a$ is perhaps more interesting because it mirrors the situation typically arising in a nonparametric problem. Here there is nothing that we can do to achieve second-order accuracy. Efron's suggestion in this case is effectively to take $a = 0$ (see Section 2.4 in the paper by DiCiccio and Romano), but it is not clear that this will always be a sensible suggestion.

Similar doubts arise over the use of the double bootstrap in Hinkley's paper to find a pivot in the normal correlation example (Fig. 1). If $r$ is the sample correlation from an elliptically symmetric bivariate distribution with correlation $\rho$ and kurtosis $\kappa$ then asymptotically

$$n^{1/2}(z - \xi) \sim N(0, 1 + \kappa)$$

where $z = \frac{1}{2}\log[(1 + r)/(1 - r)]$ and $\xi = \frac{1}{2}\log[(1 + \rho)/(1 - \rho)]$ (Muirhead, 1982). In Hinkley's example, based on the normal distribution we have $\kappa \equiv 0$. However, if $\kappa = \kappa(\rho)$ depends on $\rho$ and is not constant, then we see that $z - \xi$ is no longer a pivot, thus invalidating the interpretation of Hinkley's Fig. 1(b).

In other words the use of the bootstrap to produce confidence intervals accurate to second order involves assumptions about how the empirical distribution should be embedded into a neighbouring family of distributions. Schenker makes a similar point in the discussion of Efron's (1987) paper. Efron

 0035–9246/88/50355 $2.00

himself uses the concept of a 'least favorable family'. However, it is not clear that his embedding will always be meaningful in practice. More work is needed here to understand the implications of these embeddings.

I have found both papers a stimulating source for further discussion. I therefore have great pleasure in proposing the vote of thanks.

**A. C. Davison** (Imperial College, London): When the bootstrap was first advertised as a method which substituted computer power for that of the statistician's mind, the spectre of future unemployment seemed to loom for members of our profession. Suspicion was deepened by the discovery that much of standard statistical practice, such as the calculation of confidence intervals based on Fisher information, was henceforth to be termed the parametric bootstrap. It is thus a relief to welcome these excellent papers as evidence that, far from causing redundancies, the bootstrap has become, and looks set to continue to be, a potent statistical job creation scheme.

A first encounter with a bootstrap propagandist is usually something like this:

> Alice laughed. 'There's no use trying,' she said: 'one can't believe impossible things.'
>
> 'I daresay you haven't had much practice,' said the Queen. 'When I was your age, I always did it for half-an-hour a day. Why, sometimes I've believed as many as six impossible things before breakfast.'

However, the success of the idea forces us to think deeply about the interplay between data and model, and how we can use a model to make statements about the data. Its failure, for example when lack of appropriate conditioning leads to unsuitable sampling distribution and hence irrelevant inferences, forces us to think more clearly about general statistical principles such as that of conditionality: to clarify exactly which reference set the observed data are to be compared with. An example of the importance of this arises in Kendall and Kendall (1980), who use resampling distributions with modified support in a problem where naive sampling, either from the data or from a parametric model, would fail to reflect central features of the data. This issue arises both when conducting significance tests and when forming confidence regions, and I hope that Professor DiCiccio and Professor Romano will say what, if anything, can be said about conditional confidence regions in general settings. Is there a danger that the methods they describe could seriously mislead and, if so, how could we tell in practice? Professor Hinkley has pointed out that the issue of conditionality is implicit in much discussion of statistical methods, but it has had only passing consideration in the bootstrap literature.

Something else that is difficult to believe, even after conning the examples in Beran (1987), is that iterated sampling from bootstrap samples enables the calculation of ever more accurate confidence intervals. The theory is fascinating, but in many practical contexts the resulting gains in accuracy will be unnecessary. If the results are to be used in practice, it will be necessary to say at what level the iteration should stop. Do the authors have any theoretical or empirical guidance to offer?

I should like to rise to the bait offered by all three authors and to describe a method for efficient estimation of bootstrap tail probabilities based on importance sampling. Suppose that we wish to estimate the distribution of the asymptotically normal pivot $Z = (T-\theta)/\sqrt{V}$, where $V$ is the variance of $T-\theta$. The bootstrap estimate of this is the distribution of $Z^* = (T^* - t_{\text{obs}})/\sqrt{V^*}$, formed by resampling from the observed data $X_1, \ldots, X_n$. For suitably smooth functionals $T = t(F)$ we can approximate the behaviour of $Z^*$ by the linear terms

$$Z_L^* = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} L_j^*,$$

of its von Mises expansion, where $L_j^*$ is the infinitesimal jackknife pseudovalue corresponding to $X_j^*$, called $U_j$ by DiCiccio and Romano. Judicious choice of the scale of $T$ often makes $Z_L^*$ highly correlated with $Z^*$. We wish to estimate $P = \text{pr}(Z \leqslant z \mid F)$ by the Monte Carlo approximation

$$\frac{1}{S} \sum_{s=1}^{S} I(Z_s^* \leqslant z)$$

to the probability

$$P^* = \frac{1}{n^n} \sum\nolimits^* I(Z^* \leqslant z).$$

Here $I(A)$ is the indicator of event $A$, and $\Sigma^*$ indicates summation taken over all $n^n$ bootstrap samples. We can estimate $P^*$ directly, by resampling from the uniform distribution $F(\cdot)$ on the data $X_j$, or we can sample $X_j^*$ from a suitably chosen distribution $G(\cdot)$ on the $X_j$ and use the importance sampling estimate

$$\frac{1}{S}\sum_{s=1}^{S} I(Z_s^* \leqslant z) \prod_{j=1}^{n} \frac{\mathrm{d}F(X_j^*)}{\mathrm{d}G(X_j^*)}.$$

A little thought suggests that $G$ be chosen so that $Z_L^*$ has mean $z$, for example by choosing weights $w^j(\tau)$ for the $X_j^*$ so that $EZ_L^* = n^{-1/2}\Sigma w^j(\tau)L_j = z$. For $|z|$ in the range 1.5–2.5 likely to be useful in practice, the importance sampling estimate of probability has variance in the range 5–10 times smaller than the naive sampling estimate. Estimation at more extreme quantiles is more accurate. Pilot calculations suggest that use of this idea can reduce the size of a double bootstrap by a factor 10 or so. I have recently learned of the independent but similar work by Johns (1987).

One participant at a previous meeting remarked that he had thought the bootstrap was a simple method used to analyse complex problems, but that it seemed that more and more complex bootstraps were being invented to analyse ever simpler problems. Both are partly true: the bootstrap is found empirically to be the tool *par excellence* in problems that are too ill specified to tackle using classical tools, but to understand the underlying ideas requires theory more general, and hence potentially harder, than classical methods. For those with little faith in their distributional assumptions, the methods described by DiCiccio and Romano provide useful tools for confidence intervals, and, for those with more faith in their parametric models, the bootstrap aids assessment of the sensitivity of the conclusions drawn to the assumptions made, as Professor Hinkley points out. This attitude is summed up by the White Knight:

> 'I was wondering what the mousetrap was for,' said Alice. 'It isn't very likely there would be any mice on the horse's back.'
>
> 'Not very likely, perhaps,' said the Knight; 'but, if they do come, I don't choose to have them running all about.'

The authors have written papers which provoke thought, repay careful study and deserve to form the basis of a fruitful discussion. I congratulate them, and have much pleasure in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

**Professor B. W. Silverman** (University of Bath): One question that I have been interested in is the use of smoothing methods in a bootstrap context. In the problem considered in Silverman (1981, 1986) of testing whether a density was unimodal, some smoothing was inevitable, essentially because unimodality can only be defined by reference to the *density* underlying a population.

A more delicate situation arises if there is a possibility of smoothing or not smoothing in a bootstrap problem. Efron (1982) mentioned and explored the idea of using a smoothed version $\hat{F}$ of the empirical distribution $\tilde{F}$ as a basis for resampling. Efron's comparisons of this idea with the ordinary bootstrap were important, but they did not take account of the possibility of varying the amount of smoothing when constructing $\hat{F}$, and therefore they may have presented the smoothed bootstrap in a slightly unfair light. Dr Young and I (Silverman and Young, 1987) have looked into the question of whether smoothing is likely to help at all: in technical terms, is there any non-zero value whatever of the smoothing parameter such that using $\hat{F}$ instead of $\tilde{F}$ will give a better answer?

If, in Hinkley's notation, we regard the property $Q$ as a functional $Q(F)$ of the underlying distribution $F$, our question becomes a more general one about functionals of distributions: is $Q(\hat{F})$ ever better than $Q(\tilde{F})$ as an estimator of $Q(F)$? Superficially, we might imagine that the answer would depend on the smoothness of $F$ or perhaps of $Q$, but the criteria we developed to answer the question show that the whole matter is much more subtle and less well understood. For example, if $F$ is a normal distribution and $\hat{F}$ is constructed by standard kernel smoothing, then smoothing will have a deleterious effect on the data-based estimation of $\int x^2\,\mathrm{d}F(x)$ and of $\int x^4\,\mathrm{d}F(x)$, but not on that of $\int (x^4 - 7x^2)\,\mathrm{d}F(x)$.

One of Efron's standard bootstrap examples was the estimation of the standard error of the variance-stabilized sample correlation coefficient of a bivariate distribution. In this context our criterion showed that a suitable smoothed bootstrap procedure will give an improved estimate for a wide range of possible distributions $F$. Of course, the question of how much to smooth remains, and this is being

investigated by Dr Young, but it is clear from our work that the smoothed bootstrap may well have great unrealized potential, especially in the relatively small sample case, and I hope that it will be the subject of more research in the future.

**Dr G. Alastair Young** (University of Cambridge): I wish to add some further comments on smoothed bootstraps and to make a recommendation.

In the joint work referred to by Professor Silverman, simple theoretical conditions are obtained under which some smoothing will be advantageous in bootstrap estimation for linear functionals. Approximation ideas are then used, together with computer algebra, to extend such criteria to more complicated problems, such as the correlation coefficient example. The results obtained do not, however, provide any simple universal prescription for the smoothed bootstrap. The optimal degree of smoothing depends both on the particular functional being estimated and the underlying distribution. Further, it is not necessarily the case that the smoothed bootstrap should be based on a good estimate of the density itself. Investigations on simple problems show, however, that the application of standard density estimation methods, such as cross-validatory smoothing, is unlikely to cause much damage when used in the bootstrap context. Such procedures may nevertheless add greatly to the computational cost of the bootstrap estimation and it seems better to choose the smoothing with reference to the specific estimation problem in question. In the correlation coefficient example it is possible to use computer algebra to estimate the mean-squared error of the bootstrap estimator for all values of the smoothing parameter. An empirical smoothing procedure then involves data substitution to estimate the error function and a straightforward numerical minimization to choose the smoothing for the bootstrap estimation itself. For details see Young (1988).

A procedure related to the smoothed bootstrap is that due to David Kendall, described in Kendall and Kendall (1980). The key notion of the Kendall bootstrap is that of perturbing each data point by an independent amount drawn from a smoothing distribution, rather than resampling from a density estimate. The Kendall idea is most appropriate when the observed data are viewed as a realization of an underlying process, rather than as a set of independent identically distributed observations, and has so far only been used in bootstrap tests of geometrical and spatial pattern. In such applications, in contrast with the difficulties of smoothing Efron's bootstrap, the choice of smoothing for the Kendall procedure is simple: the smallest degree of smoothing is applied which ensures simulated data sets satisfying the null hypothesis constraint. The method seems much more general than applications would suggest and I believe this alternative smoothed bootstrap warrants further research as a means of implementing an approximately conditional bootstrap inference.

**Professor H. E. Daniels** (Cambridge Statistical Laboratory): Bootstrapping is not my natural habitat, but my curiosity was aroused by the fact that the bootstrapping procedure involves resampling from a data set which has finite support, whereas the underlying distribution could be, for example, normal. When estimating small tail probabilities in such cases, might this not lead to a noticeable bias? Dr Alastair Young and I decided to look into the question, starting with data sets from $N(\mu, 1)$ for which $\bar{X} - \mu$ is pivotal. Suppose the data set is $x_1, x_2, \ldots, x_N$ with mean $\bar{x}_N$ from which are taken bootstrap samples $x_1^*, x_2^*, \ldots, x_n^*$ with mean $\bar{x}_n^*$. Then $P(\bar{X} - \mu > a)$ is estimated by $\tilde{P} = P(\bar{X}_n^* - \bar{X}_N > a)$. When averaged over repeated data sets of $N$ to reduce the standard error to a reasonable size when $P$ is small, an enormous number of simulations is needed.

To obtain the results in Table 1, Dr Young took 50 000 bootstrap samples from each of 2000 data sets, i.e. $10^8$ samples in all! It is therefore useful to derive a saddlepoint approximation for $E\tilde{P}(\bar{x}_n^* - \bar{x}_N > a)$ in the following way.

For a particular data set the moment-generating function (MGF) of the empirical distribution is

$$M(T, x_1, \ldots, x_N) = [\exp(Tx_1) + \ldots + \exp(Tx_N)]/N,$$

and the MGF for $n\bar{z} = n(\bar{x}_n^* - \bar{x}_N)$ is

$$\exp\left[-\frac{n}{N}(x_1 + \ldots + x_N)T\right]M^n:$$

Davison and Hinkley use this to obtain a saddlepoint approximation to the bootstrap estimate of $\tilde{P}$. As we are interested in the bias of the estimate we have to average the MGF over all possible data sets before approximating. This can be done by making use of the fact that $M^n(T, x_1, \ldots, x_N)$ is the coefficient

of $\lambda^n/n!$ in

$$\prod_{j=1}^{N}\left[1+\frac{\lambda}{N}\exp(Tx_j)+\frac{\lambda^2}{N^2 2!}\exp(2Tx_j)+\ldots+\frac{\lambda^n}{N^n n!}\exp(nTx_j)\right].$$

When multiplied by $\exp[-(n/N)(x_1+\ldots+x_N)T]$ this splits into $N$ factors which can be averaged independently: then $E\exp(n\bar{z})$ is the coefficient of $\lambda^n/n!$ in $R^N(\lambda, T)$ where

$$R(\lambda, T)=\sum_{r=0}^{n}\frac{\lambda^r}{N^r r!}\exp\left[K\left(\left(r-\frac{n}{N}\right)T\right)\right].$$

Here $K(T)$ is the cumulant-generating function of the underlying distribution, which in this case can be taken to be $N(0, 1)$ with $K(T)=\frac{1}{2}T^2$. Then

$$E\tilde{P}=\frac{n!}{2\pi i}\frac{1}{2\pi i}\iint\frac{R^N(\lambda, T)}{\lambda^{n+1}}\exp(-naT)\,\mathrm{d}\lambda\,\frac{\mathrm{d}T}{T}.$$

With $\lambda=\exp\theta$, $R(\lambda, T)=\exp[\Omega(\theta, T)]$ it becomes

$$E(\tilde{P})=\frac{n!}{2\pi i}\frac{1}{2\pi i}\iint\exp[N\Omega(\theta, T)-n\theta-n\bar{z}T]\,\mathrm{d}\theta\,\frac{\mathrm{d}T}{T}$$

to which Skovgaard's extension of the Lugannani–Rice approximation can be applied.

As an example some results for $n=5$ and $N=10$ are given in Table 1. The bias seems to be least for $P$ around 0.01.

TABLE 1

| $\bar{z}$ | $E(\tilde{P})$ (SP) | $E(\tilde{P})$ (sim) | SE (sim) | $1-\Phi(\bar{z})$ |
|---|---|---|---|---|
| 0.1004 | 0.4005 | 0.4004 | 0.0007 | 0.4116 |
| 0.3143 | 0.2175 | 0.2186 | 0.0013 | 0.2411 |
| 0.5796 | 0.0804 | 0.0825 | 0.0011 | 0.0975 |
| 1.0029 | 0.0119 | 0.0123 | 0.0004 | 0.0125 |
| 1.3589 | 0.0022 | 0.0021 | 0.0001 | 0.0012 |

The next distribution to examine is the exponential distribution, using the ratio pivot, which exhibits quite different tail behaviours at either end.

**H. Tong** (University of Kent at Canterbury): The potential of bootstrap methods for non-linear time series analysis is enormous. I shall give a few examples which are based on naive experience at Canterbury and I would welcome our experts' guidance.

*Example 1.* As a test for multimodality, our experience suggests that Silverman's method may be applied to time series data. As an illustration, for the classic Canadian lynx data (logarithmically transformed) Table 2 suggests that the data follow a bimodal distribution (cf. the histogram in Tong (1983), p. 175.) It is interesting to note that the bimodality in our present context is connected with different singularities over the phase space of the underlying dynamical system.

TABLE 2

| No. of modes | $h_{\mathrm{crit}}$ | p value |
|---|---|---|
| 1 | 0.297 | 0.03 |
| 2 | 0.145 | 0.45 |
| 3 | 0.087 | 0.72 |
| 4 | 0.075 | 0.67 |

*Example 2.* Consider a threshold autoregressive model (cf. switching regression)

$$X_t + \theta X_{t-1} + \phi X_{t-1} I(X_{t-1} \leqslant r) = e_t, \quad e_t \sim N(0, \sigma^2),$$

where $I$ is the usual indicator function. Suppose that we wish to test $H_0$: $\phi = 0$. The nuisance parameter $r$ is absent under $H_0$, which invalidates application of classical theory. Although K. S. Chan of Chicago and I have shown that the *asymptotic* null distribution of the obvious likelihood ratio test statistic (Chan and Tong, 1988) is given by the first passage distribution of $B_s^2/(s - s^2)$, where $B_s$ denotes the Brownian bridge on (0, 1), for *finite* samples, we prefer to supplement our asymptotic results with Monte Carlo and bootstrap results. Chan and Tong (1984) represent *one* possible implementation. Another could be based on resampling with replacement of the fitted residuals under $H_0$.

*Example 3.* Bootstrap study of sampling properties of parameter estimates has led us to believe that there is frequently a need for non-linear time series models involving exponential terms, e.g.

$$X_t = [\alpha + (\beta + \gamma X_{t-1}) \exp(-\delta X_{t-1}^2)] X_{t-1} + \varepsilon_t,$$

to be reparameterized, such as by 'centring' $\exp(-\delta X_{t-1}^2)$ to $\exp[-\delta(X_{t-1}^2 - \Delta)]$, $\Delta = EX_t^2$, to avoid ill conditioning.

*Example 4.* The bootstrap method is also very useful in giving probability limits of $l$-step predictions based on threshold models.

**Dr P. H. Garthwaite** (University of Aberdeen) **and Dr S. T. Buckland** (Scottish Agricultural Statistics Service): Consider the parametric bootstrap, where the form of the population distribution is assumed known apart from the value of the parameter $\theta$. We might then use a method such as the percentile $t$ or the accelerated bias-corrected percentile method to determine a confidence interval, but in general the interval will have bias. If $\theta$ is a scalar, an alternative is to guess one end point of the confidence interval for $\theta$, resample once from the distribution this determines and use the result of the resampling to update the guess of the end point. An efficient method of searching for the true value of the end point in this way is based on the Robbins–Monro process (Robbins and Monro, 1951).

Suppose that $\hat{\theta}$ is the estimate of $\theta$ based on $n$-sample data and the upper end point of the central $100(1 - 2\alpha)\%$ confidence interval is sought. Let $U_m$ denote the $m$th estimate (guess) of the end point and generate a bootstrap sample of size $n$ under the assumption $\theta = U_m$. From this sample, estimate $\theta$ by $\hat{\theta}_m$. Set

$$U_{m+1} = \begin{cases} U_m - c\alpha/m, & \hat{\theta}_m > \hat{\theta} \\ U_m + c(1 - \alpha)/m, & \hat{\theta}_m \leqslant \hat{\theta}, \end{cases}$$

where $c$ is a 'step length constant'. If $U_m$ is currently equal to the upper $100\alpha^*\%$ point, the expected distance we step is $[\alpha^* c(1 - \alpha)/m] - [(1 - \alpha^*)c\alpha/m]$. This expression is zero for $\alpha^* = \alpha$, positive when $\alpha^* > \alpha$ and negative for $\alpha^* < \alpha$, so every step reduces the expected distance from the solution. An independent search is carried out for the lower limit. Implementation of the method is considered by Garthwaite and Buckland (1988).

The Robbins–Monro process has two useful properties. Firstly, if we know the optimum value of the step length constant, the variance of $U_{m+1}$ is equal to the Cramer–Rao lower bound to the variance of estimates of the upper limit (Wetherill, 1975) and, secondly, the process yields an asymptotically exact method under general conditions as bootstrap replications tend to infinity irrespective of the sample size. The second property is important because the number of bootstrap replications $b$ is limited only by computing power available, whereas we often have little control over the sample size. The double bootstrap described by Hinkley requires $b^2$ replications to achieve this; the above method requires just $2b$ ($b$ for each limit). The method has similarities to the iterative methods described by DiCiccio and Romano but has the advantage of known convergence properties.

In the presence of nuisance parameters, a 'conditional' interval may be obtained by carrying out the above search on the parameter of interest, while conditioning on the point estimates of nuisance parameters. With more than one parameter of interest, this might be applied to each parameter in turn, treating the others as nuisance parameters. The method may also prove useful when applied in conjunction with the methods described by DiCiccio and Romano.

The following contributions were received in writing after the meeting.

**Professor R. Beran** (University of California at Berkeley): I would like to add to Professor Hinkley's wide ranging review of bootstrap methods.

When constructing a bootstrap test, as in Section 5, the goal of correct asymptotic rejection probability under the null hypothesis usually imposes two requirements on the fitted model $\tilde{F}_H$:

(a) the possible values of $\tilde{F}_H$ should be distributions allowed by the null hypothesis;
(b) under the null hypothesis, $\tilde{F}_H$ should be a consistent estimate of the actual distribution.

The minimum distance estimate $\tilde{F}_H$ described by Hinkley has these properties very generally in the independent identically distributed case. Particularly in parametric models, much simpler consistent estimates $\tilde{F}_H$ are often also available.

Bootstrap tests neatly resolve several long-standing, analytically intractable problems in addition to those mentioned in Section 5, including finding asymptotically correct critical values for minimum distance test statistics, for test statistics which involve estimated parameters and for tests in multivariate analysis when normality is not assumed (Beran, 1986; Beran and Srivastava, 1985). When the limiting null distribution of the test statistic does not depend on unknown parameters, a bootstrap test is usually second order correct. For example, a parametric bootstrap version of the likelihood ratio test automatically accomplishes (to second order) the Bartlett adjustment to the chi-squared asymptotics. Similarly, Welch's approximate solution to the Behrens–Fisher problem is achieved by the parametric bootstrap critical vaue for that test statistic (Beran, 1988).

Outside the statistical literature, bootstrap methods have a prehistory. For example, Lampton *et al.* (1976) describe bootstrap-like simulations in a complex statistical problem arising in X-ray astronomy. Professional statisticians have greatly clarified the logical and mathematical bases for bootstrapping, have developed many new bootstrap methods and have reached a sound understanding of the role of bootstrap methods in analysing data.

I would also like to add a few remarks to Professor DiCiccio and Professor Romano's Sections 3 and 4.

In thinking about bootstrap confidence sets, it is useful to distinguish between two cases: case I, where the limit law $J(F)$ of the root does *not* depend on the unknown distribution $F$, and case II, where it does. In case I, the bootstrap confidence set $B_n$ is typically second order correct; it is asymptotically equivalent to the confidence set which refers the root to the estimate of the $(1-\alpha)$th quantile obtained from a two-term asymptotic expansion for $J_n(F)$. For example, a parametric bootstrap version of a likelihood ratio confidence set automatically achieves, up to second order, the Bartlett adjustment to the chi-squared asymptotics (cf. Cox (1987) with Beran (1988)). Case I also includes Studentized roots whose limit laws are standard normal or folded-over standard normal (e.g. the Behrens–Fisher problem).

In contrast, in case II the bootstrap confidence set $B_n$ is usually only first order correct; it is asymptotically equivalent to the confidence set which refers the root to the estimated $(1-\alpha)$th quantile of the limit law $J(F)$, because using additional terms from an estimated expansion for $J_n(F)$ to adjust the critical value does not affect order of coverage probability error in case II. A striking instance of this phenomenon, in a situation where the limit law is not normal, occurs in the analysis of confidence cones by Ducharme *et al.* (1985).

Prepivoting transforms a case II root into a preferable case I root, because the limit law of the new root $R_{n1}$ is uniform on (0, 1). Prepivoting a case I root is beneficial as well, because the distribution of $R_{n1}$ then also depends less strongly on $F$. More precisely, any dependence on $F$ is pushed into higher order terms of the expansion for the distribution of $R_{n1}$.

**Dr Peter Hall** (Australian National University, Canberra): The statistical community will be most grateful for the careful and thorough job which Hinkley and DiCiccio and Romano have done to consolidate our knowledge of the bootstrap. The last few years have led us to a good appreciation of theory for the bootstrap, and these two papers do much to put that work into perspective.

While we now have a good understanding of the bootstrap, our knowledge of the overall problem of nonparametric confidence interval construction is still very rudimentary. To indicate some of the avenues up which statisticians have scarcely glanced, let us pose a theoretical question whose solution has obvious practical implications. For simplicity I shall assume that the parameter of interest is a univariate mean, and that a major attribute of the confidence interval is coverage accuracy. A modified version of my problem admits interval length as a major issue, and that leads to consideration of nonparametric likelihood-based intervals.

Assume that the underlying density $f$ vanishes outside a given finite interval $(a, b)$, so that all moments exist. Given a random $n$ sample from this distribution we wish to construct a nonparametric, $\alpha$-level, two-sided confidence interval $I \equiv [\hat{c}, \hat{d}]$ for the unknown mean $\mu_f$. How accurately can we construct the interval? That is, how close to $\alpha$ can the coverage probability be?

This sort of question is perhaps best posed in a minimax setting. To remove pathological intervals from consideration we ask that the length of $I$ be of order $n^{-1/2+\varepsilon}$ for each $\varepsilon > 0$. Therefore, select a class $\boldsymbol{C}$ of $f$s supported on $(a, b)$, insist that

$$\inf_{f \in \boldsymbol{C}} P_f(\hat{d} - \hat{c} \leqslant n^{-1/2+\varepsilon}) \to 1 \quad \text{for all } \varepsilon > 0 \tag{1}$$

and ask how small we can make

$$s_n(\boldsymbol{C}) \equiv \inf^* \sup_{f \in \boldsymbol{C}} |\alpha - P_f(\mu_f \in I)|, \tag{2}$$

where inf* denotes the infimum over all nonparametric constructions of $I$ with given nominal level $\alpha$.

Presumably $\boldsymbol{C}$ is determined by a smoothness condition on its elements, and the convergence rate of $s_n(\boldsymbol{C})$ to zero depends on that condition. It is relatively easy to derive upper bounds to $s_n(\boldsymbol{C})$ for given smoothness classes, but more difficult to obtain lower bounds. The type of smoothness assumption imposed appears linked to the smoothness condition needed to assert existence of an Edgeworth expansion. Of course, it is of interest to know whether taking $I$ to be a bootstrap-related interval confers any optimality properties of the type suggested by this problem. A more detailed analysis may require conditions (1) or (2) to be modified.

**Scott Koslow and David W. Stewart** (University of Southern California, Los Angeles): Both papers raise intriguing points and suggest a variety of directions for future research. We discuss only one of those directions: conditional bootstrap methods.

As we have been engaged in this area for some time, we believe that the conditional bootstrap offers one potential solution to problems involving parameter estimation of observations arising from mixed distributions. This may occur when longitudinal data are collected at the individual level. Here, it may be unreasonable to assume that parameters estimated from an aggregation of the observations are consistent with parameters obtained at the individual level. If segments of individuals exist, it also may be unreasonable to assume that all the observations arise from the same underlying distribution. An example of such a situation involves scanner panel data that are now routinely collected in many retail outlets.

A conditional bootstrap approach to estimation may be applied in scanner data since the distribution of one parameter of interest, say the number of runs of identical purchases, is conditional, in part, on the total number of different brands purchased in a given time. Hinkley suggests that the direct determination of such conditional distributions via the bootstrap is not possible without an explicit model or a knowledge of the parameter on which conditioning occurs. While this is strictly true, we have found an alternative approach to this problem. Our work employs the bootstrap to determine a null conditional distribution, i.e. the chance distribution that would be expected given known characteristics of the data. Such a null distribution then allows us to ascertain whether the number of runs is significantly different from what might be expected by chance given that the individual has purchased $n$ brands.

Although we cite but one example, the conditional null distribution can be computed whenever the distribution of one parameter may be constrained by the distribution of another observable characteristic of the data. While the computation of a null conditional distribution does not directly solve the problem of estimating a conditional distribution, it does offer an opportunity to test whether such conditionality should be of concern.

**Professor Robert J. Tibshirani** (University of Toronto): The power of the bootstrap method stems from the fact that (in Hinkley's notation)

(a) $R_t(F, \tilde{F})$ can be arbitrarily complicated and hence not amenable to theoretical approximations,
(b) it is *automatic* and
(c) it is *exact*.

It is automatic because, having chosen $\tilde{F}$ and $R_t(F, \tilde{F})$, no special calculations are needed for specific problems, as in say an Edgeworth expansion. It is exact because, given an infinite number of Monte Carlo samples, we can compute the distribution of $R_t(F, \tilde{F})$ exactly.

For this reason, I was somewhat disappointed in Professor Hinkley's emphasis on methods for efficient computation and theoretical approximation. In particular, balanced sampling would seem to require

too much special effort in setting up the simulation (beyond first-order balance) and an empirical saddlepoint introduces an extra fixed source of error in the computation, namely the difference between the true distribution of $R_t(F, \tilde{F})$ and its saddlepoint approximation based on $F$. Furthermore, Feuerverger (1988) has shown that the saddlepoint approximation of $\sqrt{n}(\bar{X} - \mu)$ has an error of order $O_p(n^{-1/2})$, no better than a normal approximation. The same holds for the bootstrap distribution of $\sqrt{n}(\bar{X} - \mu)$ (Hartigan, 1986), but we can achieve an error of $O_p(n^{-1})$ by bootstrapping a Studentized pivot.

The 'bootstrap partial likelihood' problem (Hinkley's Section 8) is extremely interesting. Here is another (simple-minded) approach. Given some second-order correct confidence procedure with $\alpha$ end point $\theta[\alpha]$, define $-2$ log-lik$(\theta[\alpha])$ to be the so-called 'confidence distribution' for $\theta$. In detail, $-2$ log–lik$(\theta[\alpha]) = -2$ log-lik$[\theta([1-\alpha])] \equiv \chi^2_1(0.5) - \chi^2_1(1-\alpha)$, for all $\alpha \in [0, 0.5]$, $\chi^2_1(t)$ being the $t$th percentile of the $\chi^2_1$ distribution. This is similar to Hall's method, but simpler. If $\theta[\alpha]$ comes from the $BC_a$ procedure, this can be justified as being equivalent to the likelihood for the variance-stabilized parameter $h(\theta)$ based on $h(\hat{\theta})$. I tried this on the problem of Section 8 and the resulting likelihood was very close to those in Hinkley's Fig. 2. Any comments on this suggestion would be welcomed.

Professor DiCiccio and Professor Romano give a detailed but very clear tour through the complicated jungle of bootstrap confidence intervals. I find the new procedure of Section 2.3 intriguing in its notational compactness. While this area is very fruitful for theoreticians, we must not lose sight of the practical issues, specifically

(a) bootstrap confidence intervals are primarily useful for nonparametric problems: parametric problems, for which exact and likelihood intervals are available, act as a test-bed and
(b) while we are all guilty of bootstrapping the mean in theoretical studies, it would be more prudent to bootstrap less sensitive functionals (like medians) in real problems (see Tibshirani and Wasserman (1987) for a discussion of this).

Finally, a question for any of the authors: is there evidence that nonparametric bootstrap methods are more effective than flexible parametric modelling, when both are applicable?

**Professor D. M. Titterington** (University of Glasgow): I should like to make some comments, mainly related to Professor Hinkley's paper.

First, I have a couple of brief remarks.

(a) In Section 3 it is remarked that the discreteness of $\tilde{F}$ renders difficult the theoretical work on the saddlepoint method. Would the use of a smoothed version of $\tilde{F}$ lead to any meaningful easing of this difficulty?
(b) The dog-leg regression problem of Section 6 generates a likelihood ratio test statistic with non-standard null distribution. An equivalent problem arises in testing for the number of components in a mixture, for which the bootstrap procedure is described by McLachlan (1987). I would have been interested to see how far the empirical distribution of $T^*_H$ really is from that of a $\chi^2$ distribution. In Figs 1 and 2 of McLachlan (1987) corresponding empirical distributions *appear* to be, intriguingly, not too far away from a $\chi^2$ shape.

My other comments concern the problem of a non-pure significance test. In the pure case, Monte Carlo tests can sometimes be derived that have exact significance levels associated with them. In the non-pure case (testing a parametric distributional hypothesis with unspecified values for the parameters, for instance), corresponding Monte Carlo tests are not so 'exact'. How close they are to being exact seems to be an open problem.

I have recently been looking at another aproach to this which uses a two-sample test statistic to test such a distributional hypothesis. In the terminology of Section 5, suppose that the null hypothesis is $H$ and that $\tilde{F}_H$ is obtained from the data. Two independent bootstrap samples are generated, one (possibly parametric) from $\tilde{F}_H$ and one from $\tilde{F}$. Suppose that these samples are denoted by $\tilde{F}^*_H$ and $\tilde{F}^*$ respectively. A two-sample test statistic (of the null hypothesis that two distributions are identical) is then evaluated with $\tilde{F}^*_H$ and $\tilde{F}^*$ as arguments and is referred to the relevant, standard rejection region. Although the theory of the method is not yet established, the procedure seems to work well in the one or two problems that I have looked at so far. An important point is that, if the two-sample test is one based on ranks or runs, it is crucial to generate $\tilde{F}^*$ as a smoothed bootstrap sample, to avoid ties.

**Mr R. J. Verrall and Professor H. P. Wynn** (City University, London): We should like to expand a little on the method of Ogbonmwan and Wynn (1988) referred to by Professor Hinkley. The basic

technique is to resample from a set $S(\theta)$ of $y^*$ vectors generated from $y_\theta = g_\theta(x)$ where $x = (x_1, \ldots, x_n)$ is the data set. For each $y^*$ we compute a statistic $T(y^*)$. The set of all $T(y^*)$ has an empirical CDF $\hat{F}_T(t \mid \theta)$ which depends on $\theta$. By simulating for a range of $\theta$ values we may obtain a cumulative likelihood in the usual way. We have suggested using a smoothed version of the density $\hat{f}_T(t \mid \theta)$.

For some models a different version may be derived by reconstructing alternative samples for the original data $x$. If $y^*$ is a member of $S(\theta)$ we may construct $x^* = \tilde{g}^{-1}(y^*)$. The tilde here denotes the fact that $\tilde{g}^{-1}$ may not be the precise inverse of $g$. This method is explained in more detail in Ogbonmwan *et al.* (1987) and is applied to time series where it is a natural procedure for autoregressive (AR($p$)) processes.

There is a close connection with nonparametric methods. Indeed the proper rerandomization likelihood is

$$\frac{1}{B} \operatorname{prob}(t(y^*) = t(y_\theta) \mid \theta)$$

where $B = \operatorname{card}(S(\theta))$. Nonparametric confidence regions are based on

$$P_\theta = \frac{1}{B} \operatorname{prob}(t(y^*) \leqslant t(y_\theta) \mid \theta),$$

the cumulative likelihood, which is then essentially the same as $\hat{F}_T(t \mid \theta)$. The outstanding problem is to invert statements like $P_\theta \geqslant 1 - \alpha$ to make statements about $\theta$ which depend only on $\theta$ and the randomization procedure which generates $S(\theta)$. The moral from this, in elementary terms, is to bootstrap the *true residuals* using stored values of $\theta$ rather than estimated residuals. If we do this the thin dividing line between nonparametrics and resampling is breached.

**Professor C. F. J. Wu** (University of Waterloo, Canada, and University of Wisconsin, USA): My comments are directed to four issues.

*Bootstrap for complex problems*

Despite its simplicity and versatility the bootstrap is not supported in complex situations by current theoretical advances. The bootstrap is readily applicable only if a complex problem can be described by a model driven by an exchangeable stochastic component. Otherwise the method is not automatic and cannot be used routinely. Modification of the independent identically distributed (IID) bootstrap by taking into account the nature of the problem's complexity seems necessary. Hinkley recognizes this in regression and suggests using stratification or local smoothing to obtain near exchangeability within stratum, requiring the assumption that the covariates $\{x_i\}$ are reasonably dense. Otherwise the set of $D^*$ close to $D$ is too small for inference. It may not even give good variance estimators. In contrast a weighted jackknife (Wu, 1986) gives consistent variance estimators for general heteroscedastic regression including GLIM. Can the delete-$d$ version of this jackknife be used for interval estimation? Another example is complex surveys, in which the independence assumption is often violated. For simple random sampling without replacement, increasing the sample size to $n/(1-f)$ gives consistent variance estimation but not $O(n^{-1/2})$ correction in the distribution approximation, since the bootstrap does not mimic without replacement sampling. Rao and Wu (1988) point out some serious problems with the current methods of adjusting sample sizes for more complex sampling plans and propose alternative procedures which are valid for variance estimation in general survey designs.

*Bootstrap versus jackknife for variance estimation*

For the sample median the delete-1 jackknife variance estimator is inconsistent because it resamples from too few values. By using a delete-$d$ jackknife with $d$ depending on a smoothness measure of $\hat{\theta}$ (Shao and Wu, 1987), the support of resampled values is broadened to ensure consistency. Recently Shi (1987) showed that the delete-$d$ jackknife variance estimator with $d = \lambda n$, $0 < \lambda < 1$, is strongly consistent under *no* assumption on the moments or tails of $F$. The delete-$d$ jackknife does not resample from very extreme quantiles whereas the bootstrap does so with non-negligible probability. This explains the inconsistency of the bootstrap variance estimator for heavy-tailed distributions (Ghosh *et al.*, 1984).

*Non-uniformity in confidence estimation and testing*

Non-uniformity over $F$ of confidence intervals is shown by Bahadur and Savage (1956). For bootstrap intervals Loh (1988) recognizes the same problem and proposes a calibration method. Similarly the

bootstrap test $P$ value discussed by Hinkley has this problem. His $\tilde{P}_H$ value evaluated at $\tilde{F}_H$ may be too liberal. A corrective measure is to take the supremum of $\tilde{P}_H$ over a neighbourhood of $\tilde{F}_H$, which can be chosen using the frequentist approach (Loh, 1985) or the Bayesian approach.

*Confidence intervals: asymptotic theory and empirical evidence*

Most theoretical advances have been made in this area, primarily for simple models. Results on iterative bootstrap refinements are too good to be true in finite samples. However, for fixed $n$, there is a limit on the number of iterations yielding improvement. A more sensible asymptotic framework is to let both $n \to \infty$ and $j \to \infty$ with the latter at a slower rate depending on $n$. (Loh (1988) has recently refined his calibration method to avoid nested bootstrap calculations.) There is a gap between asymptotic theory and empirical results. An objective and extensive simulation study is called for.

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

**Professor David Hinkley:** In principle, bootstrap methods enrich our abilities to perform increasingly complex data analyses. The extent to which this becomes reliable practice depends on the kind of critical discussion which has taken place here, which helps to temper the often extravagant claims to which Dr Kent and Dr Davison obliquely refer. I hope that in the near future the Royal Statistical Society will arrange a corresponding discussion of bootstrap applications, wherein lies the real test. For the moment, however, we remain with the many interesting theoretical points raised here.

Although the bootstrap is usually applied by resampling from the discrete data distribution, it would often seem more natural to sample from a smooth estimate. Professor Silverman and Dr Young correctly point out that the utility of smoothing will depend on the type of statistic being considered. Further results in this direction have been obtained by Hall *et al.* (1988), their main application being quantile estimation. It would be interesting to see Professor Daniels's ingenious analysis extended to consideration of smooth bootstraps.

One illustration of the need to consider smoothing is the following example due to Taylor and Thompson (1986). Points $x$ are vectors representing random deviations of a missile from a point target, the quantity of operational interest being $\Pr(|X| \leqslant r) = \pi$. The usual bootstrap will give the absurd upper confidence limit zero for $\pi$ if none of the data points are within $r$ units of the target. There are various ways of introducing smoothness to deal with this difficulty, one being to smooth the data distribution.

Concern about computational efficiency is surely not quite the waste of time that Professor Tibshirani seems to suggest, although, if his point is that we should carefully weigh statistical and numerical errors, then I agree wholeheartedly. The theoretical improvements offered by first-order balance can be effected by careful algorithmic development, as shown by Gleason (1988). The importance sampling techniques explored by Johns (1986), and the similar technique mentioned by Dr Davison, can be genuine time savers, although the savings are not quite as spectacular as Dr Davison's numbers suggest when actual computation time is taken into account. Importance sampling can be easily adapted to reduce the heavy computational burden of the double bootstrap (Hinkley and Shi, 1988). One would hope that some of the special Monte Carlo techniques developed by Professor Adrian Smith and his co-workers for high dimensional likelihood integration will have useful analogues in resampling analysis.

Incidentally, Professor Tibshirani's remarks about saddlepoint approximation confuse numerical and statistical error, and in no way represent a case for using one minute of computer time when the same answer can be obtained in one second. The trick is to make the same numerical technique work when the statistical error is also controlled, and to do this requires a non-trivial generalization of the existing saddlepoint approximation—I hope that Professor Daniels will rise to this challenge! It would be interesting to know how much of the discrepancy calculated by Professor Daniels is due to discreteness and how much to lack of pivotality: the bootstrap refinements such as those described by DiCiccio and Romano, as well as the double-bootstrap methods, deal with the latter.

The double bootstrap arises quite naturally if we are considering use of the bootstrap to estimate the average statistical error of a bootstrap procedure. This simple but elegant concept, described to me by Dr Hall, should appeal to Dr Davison. A very simple bootstrap confidence limit method can be obtained by bootstrapping Efron's percentile method. For an upper $1-\alpha$ confidence limit we first determine that value $\gamma$ for which

$$\Pr(\Pr(T^{**} \leqslant T \mid \tilde{F}^{*}) \leqslant \gamma \mid \tilde{F}) = 1 - \alpha,$$

and then take as the upper limit the $\gamma$ quantile of the empirical distribution of $T^*$. Simple test cases for this method are the parametric bootstrap analyses for normal and exponential means. The theory and numerical implementation of the procedure will be described elsewhere.

The double-bootstrap approach can also be used to offset the potential difficulty with bootstrap significance tests, correctly noted by Professor Wu.

Dr Kent's doubts about the double-bootstrap plot are very perceptive. In general it would be useful to plot percentile or variance estimates obtained from $T^{**}$ against any potentially relevant characteristic, such as sample standard deviation $s^*$ when dealing with $T = \bar{X}$ or kurtosis when dealing with any second moment. In some cases, hidden effects can be diagnosed by noting excess variability in a plot such as Fig. 1, as outlined by Chapman and Hinkley (1986).

It will be interesting to see how well the various ideas for resampling significance tests work. I am indebted to Dr Young for reminding us about data corruption. Professor Titterington's suggestion is an interesting one, but it seems to contain a basic flaw. His test will tend to give a significant result simply because $\tilde{F}_H \neq \tilde{F}$, i.e. the null hypothesis of his test is false even when the null hypothesis $H$ of interest is true. To take a simple example, consider a parametric bootstrap two-sided test for a single mean in which $\tilde{F}$ is $N(\bar{x}, s^2)$ and $\tilde{F}_H$ is $N(0, s^2 + \bar{x}^2)$, corresponding to the hypothesis that the population mean is zero. Suppose that $\tilde{F}_H^*$ and $\tilde{F}^*$ are compared through their averages alone, using the usual two-sided two-sample $t$ test. If the nominal level of this latter test is 5%, then its overall error rate would be close to 10%. The conditional rejection rate ranges from 5% to 40% for values of $\bar{x}$ and $s^2$ which would not lead to rejection in the direct classical one-sample $t$ test.

The mention of empirical likelihood in Section 8 was perhaps too brief, and regrettably there was no discussion of this. Owen (1988a) has since extended the earlier work to vector parameters, and DiCiccio *et al.* (1988) have further pursued the extent to which empirical likelihood behaves like a classical likelihood function.

I was intrigued by Professor Wynn's outline of an alternative resampling likelihood. The randomization approach is necessarily of much more restricted legitimate applicability than the bootstrap. When randomization is valid, the implicit conditioning (Young, 1986) is helpful. An interesting non-trivial application of constrained randomization is described by Efron (1988).

Professor Tong is far too modest about his excellent start on applying bootstrap methods to non-linear time series. He might consider use of the double-bootstrap methods to offset the clear lack of pivots in several such problems, e.g. his example 3. The same advice is relevant in non-linear regression analysis, of course.

Nothing was said in the paper about jackknife methods, so I am grateful to Professor Wu for mentioning some of the work in this area. The key feature of the multiple-deletion jackknife was realized by Brillinger (1964), and there is some related material in Hinkley (1977). What the jackknife cannot give us is simple distribution estimates with second-order accuracy. It is a pity that Professor Wu does not give a suggestion about the choice of deletion set size $d$. The recent work by Graham *et al.* (1988) on balanced bootstraps suggests that $n$ balanced subsamples with $d \simeq \frac{1}{2}n$ may work well, the particular subsample design having elements in common with Wu's implementation of the jackknife.

The interesting applications mentioned by several authors suggest the considerable potential of bootstrap methods in testing, which may ultimately prove more important to applied statisticians than bootstrap confidence limit methods are.

The final question raised by Professor Tibshirani is important, with many facets. Perhaps the only satisfactory answer will come from discussions of bootstrap applications. I think that a fully integrated Bayesian aproach will always be best when all the relevant components are available. The kinds of applications which bootstrappers have in mind are different.

**Professor Thomas J. DiCiccio and Professor Joseph P. Romano:** It is clear from the contributed discussions that the bootstrap is a thriving area of theoretical research and practical application.

Dr Kent questions the possibility of defining second-order accuracy for nonparametric confidence intervals. Perhaps the following development of the bootstrap $t$ clarifies the definition. Let $x_n = (X_1, \ldots, X_n)$ be a sample of size $n$ from an unknown distribution $F$ on some arbitrary sample space. The problem is to construct a confidence interval for a functional $\theta = \theta(F)$ based on some estimator $\hat{\theta}_n$. Let $\hat{\sigma}_n^2$ denote an estimator of the variance of $n^{1/2}\hat{\theta}_n$ and set

$$J_n(x, F) = P_F\{n^{1/2}[\hat{\theta}_n - \theta(F)]/\hat{\sigma}_n \leqslant x\}.$$

Also, let

$$J_n^{-1}(\alpha, F) = \inf\{x : J_n(x, F) \geqslant \alpha\}$$

be an $\alpha$ quantile of the distribution $J_n(\cdot, F)$. Then, an exact upper $1-\alpha$ confidence limit for $\theta$ is

$$\hat{\theta}_{\mathrm{EX}}(1-\alpha) = \hat{\theta}_n - n^{-1/2}\hat{\sigma}_n J_n^{-1}(\alpha, F), \tag{3}$$

i.e. $P_F\{\theta(F) < \hat{\theta}_{\mathrm{EX}}(1-\alpha)\} \geqslant 1-\alpha$ and is exactly $1-\alpha$ provided that $J_n(x, F)$ is continuous and strictly increasing in $x$.

In general, we say that a proposed upper $1-\alpha$ confidence limit, $\hat{\theta}_{\mathrm{U}}$ (or $\hat{\theta}_{\mathrm{U}}(1-\alpha)$ to show the dependence on $\alpha$) is second order correct if

$$\hat{\theta}_{\mathrm{U}}(1-\alpha) - \hat{\theta}_{\mathrm{EX}}(1-\alpha) = O_P(n^{-3/2}). \tag{4}$$

It typically then follows that

$$P_{\mathrm{F}}\{\theta(F) \leqslant \hat{\theta}_{\mathrm{U}}\} = 1 - \alpha + O(n^{-1}). \tag{5}$$

The 'exact' upper $1-\alpha$ confidence limit $\hat{\theta}_{\mathrm{EX}}(1-\alpha)$ given by equation (3) is usually unknown because $F$ is unknown. The bootstrap $t$ method, introduced by Efron (1981), estimates $J_n(x, F)$ by $J_n(x, \hat{F}_n)$ and then estimates the $\alpha$ quantile of $J_n(x, F)$ by the corresponding $\alpha$ quantile of $J_n(x, \hat{F}_n)$, i.e. the upper $1-\alpha$ bootstrap $t$ confidence limit for $\theta(F)$ is defined to be

$$\hat{\theta}_{\mathrm{BT}}(1-\alpha) = \hat{\theta}_n - n^{-1/2}\hat{\sigma}_n J_n^{-1}(\alpha, \hat{F}_n). \tag{6}$$

Hall (1988) shows in the 'smooth function' model that $\hat{\theta}_{\mathrm{BT}}$ is second order accurate in the sense that equations (4) and (5) hold for the choice $\hat{\theta}_{\mathrm{U}} = \hat{\theta}_{\mathrm{BT}}$.

Dr Kent considers the parametric model where observations are normally distributed with unknown mean $\theta$ and unknown variance $(\sigma + a\theta)^2$. If $a$ is unknown, the unknown mean and variance parameters vary freely and the exact solution is the $t$ interval. The bootstrap $t$ solution to nonparametric problems also removes the effect of not knowing the variance. As Professor Beran points out, the bootstrap $t$ is second order accurate because the 'root' $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$ is typically asymptotically standard normal so that its asymptotic distribution does not depend on any unknown parameters. In the correlation example discussed by Dr Kent, the asymptotic distribution depends on unknown parameters, and so one way to improve on the construction of the interval is to apply the bootstrap $t$ with an appropriate estimate of standard error. Alternatively, we may apply Professor Beran's prepivoting operation.

As Dr Kent observes, the accuracy of bootstrap methods depends on the behaviour of the statistic or 'root' whose distribution is being estimated. The less dependent the distribution is on $F$ (in some 'neighbourhood' of the true $F$), the more reliable is the bootstrap based on substitution of $F$ by the empirical distribution $\hat{F}_n$. Certain methods rely explicitly or implicitly on reducing a nonparametric problem to a parametric problem by considering the behaviour of the statistic or root under a one-dimensional family of distributions containing $\hat{F}_n$. Such methods include nonparametric tilting, the $\mathrm{BC}_a$ method and Owen's (1988b) empirical likelihood. If such reductions are achieved through 'least favourable' families, then corrections to more simple-minded procedures such as the percentile method are non-trivial and second-order accuracy is achievable; see DiCiccio and Romano (1988a, b). However, for one-sided confidence limits, the empirical likelihood method is not in general second order accurate, though its construction relies on a least favourable family generated by maximizing out nuisance parameters; see DiCiccio *et al.* (1988). However, bootstrap calibration of the empirical likelihood method does produce second-order accurate intervals.

The issue of smoothing the empirical distribution before resampling, raised by Professor Silverman and Dr Young, deserves further investigation. The proper amount of smoothing is a delicate question, as illustrated in example 3.4. In this example and others (such as constructing a confidence interval for the mode), the naive amount of smoothing results in bootstrap confidence intervals that do not even have the correct level asymptotically. In Silverman's (1981) interesting approach to testing for unimodality, some theoretical justification is needed to apply his bootstrap procedure safely.

Silverman and Young (1987) attempt a more general approach to the problem of smoothing and obtain an interesting result for linear functionals. Though many statistical functionals are approximately linear, it seems that the proper amount of smoothing cannot be simply determined on the basis of such an approximation, since smoothing will typically have a secondary effect on the quality of the resulting confidence procedure because second-order properties of the functional reflected in the error of such a

linear approximation play a role. Improving the second-order asymptotic properties of any procedure can be extremely beneficial for finite sample sizes, so that the question of how much smoothing is worthwhile needs to be addressed. Perhaps more important, for smooth functionals such as the correlation where the bootstrap $t$ is second order accurate, the benefits of smoothing need to be made more explicit. For example, it is doubtful that the order of error, typically $n^{-1}$ for the bootstrap $t$ with no smoothing, can be improved with smoothing. Substantial gains, however, are possible when dealing with less smooth functionals of a distribution, especially when local properties of the distribution play a prominent role. As a start, in Hall *et al.* (1988), the bootstrap estimator of the variance of a quantile estimator has a convergence rate of $n^{-1/4}$ if no smoothing is done, but can be improved to $n^{-1/2+\varepsilon}$ for any $\varepsilon > 0$, if smoothing is applied, and the precise optimal amount of smoothing can be made explicit.

Professor Titterington addresses the problem of the accuracy of bootstrap tests when parameters need to be estimated under the null hypothesis to specify the appropriate resampling mechanism. Some asymptotic results have been obtained by Beran (1986, 1988) and Romano (1988a). For example, it has been shown that bootstrap goodness-of-fit tests and minimum distance tests are asymptotically valid. The power of resampling methods is quite evident in such problems since the asymptotic distributions of such minimum distance test statistics often do not possess a known analytical form. Even if they did, the asymptotic distribution would depend on nuisance parameters as well; also see Romano (1988b).

We thank Professor Wu for his remarks, particularly for his update on jackknife techniques. He reminds us of the need to extend bootstrap theory to more complex and interesting situations, such as those discussed by Professor Tong. On the issue of uniformity, not much is known, and the challenging problem raised by Dr Hall is most relevant. Sometimes, the bootstrap does behave well uniformly, and corrections to the bootstrap may be unnecessary. For example, reconsider example 3.2, the problem of constructing a confidence band for a distribution function $F$ on the line. For $\varepsilon > 0$, let $\mathbf{F}(\varepsilon)$ be the class of distributions on the line with no atoms having probability greater than $1 - \varepsilon$. Then, for example, $\mathbf{F}(\varepsilon)$ includes all continuous distributions and those discrete distributions that are essentially not degenerate so that an outcome from $F$ cannot be predicted correctly with probability greater than $1 - \varepsilon$. Let $g_n(1 - \alpha, F)$ be the actual coverage probability of the nominal $1 - \alpha$ bootstrap confidence set based on a sample of size $n$ from $F$. Then, for any $\varepsilon > 0$,

$$\sup\{|g_n(1-\alpha, F) - (1-\alpha)| : F \in \mathbf{F}(\varepsilon)\} \to 0$$

as $n \to \infty$. The result is proved in Romano (1987). A variation of example 3.5 shows this result cannot be extended to $\varepsilon = 0$.

Dr Garthwaite and Dr Buckland present an interesting simulation approach to the construction of confidence limits, based on the Robbins–Monro process. We look forward to comparing it with the method given by equation (2.4). Our proposed method, now called the automatic percentile method, has been extended to parametric problems with nuisance parameters and nonparametric problems and is second order accurate in general; see DiCiccio and Romano (1988a, b). Garthwaite and Buckland's approach to extending their method relies on conditioning on estimates of nuisance parameters. We do not believe that this will result in second-order accuracy, unless the parameters are parameterized orthogonally.

Professor Tibshirani claims the bootstrap to be exact. A clarification of this point is necessary. In our notation, the thrust of the bootstrap method of forming confidence intervals is to approximate the distribution of a statistic or root, $J_n(F)$, by $J_n(\hat{G})$ for some estimate $\hat{G}$ of $F$. Usually, $J_n(\hat{G})$ cannot be calculated exactly, but can be approximated by Monte Carlo methods. Moreover, it can be approximated to any desired degree of accuracy given sufficient Monte Carlo repetitions. In this sense, Professor Tibshirani describes the bootstrap to be exact. Unfortunately, the error in approximating $J_n(F)$ by $J_n(\hat{G})$ is the main component of error in the bootstrap method, unless the initial choice of root is a pivot. Hence, it should be understood that the bootstrap method does not in general yield exact inference, though we can often be satisfied with second-order correctness.

Professor Tibshirani's final question is intriguing. Though the question is vague in its present form, we do not believe the bootstrap to be the only solution to any inferential problem. If, indeed, a flexible parametric approach is sensible, one would hopefully compare results.

A very nice presentation of bootstrap iteration is given in the recent work of Hall and Martin (1988). The important problem of deciding when to stop iterating, as raised by Professor Davison, has yet to be properly addressed. An advantage of being able to correct a confidence set by a bootstrap calibration method is that we can simultaneously estimate the coverage error of the initial procedure; see, for example, section 3.3 of Beran (1988). The importance of efficient algorithms to implement bootstrap

procedures is more clear with iterative bootstrap methods, and the contributions by Professor Davison and of Professor Daniels may be helpful. Also, see Efron (1988).

Professor Davison has raised the issue of conditioning. In certain parametric contexts, the application of simulation methods to likelihood ratio statistics, as mentioned by Professor Beran, may take relevant sets into account. Cox (1980) and McCullagh (1984) have shown that these statistics can in certain cases account for natural ancillaries to second order. However, in nonparametric situations, conditioning is less clearly understood, and this important topic deserves further attention. We await with interest a fuller report of the approach taken by Professor Koslow and Professor Stewart.

## REFERENCES IN THE DISCUSSION


Bahadur, R. R. and Savage, L. J. (1956) The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.*, **27**, 1115–1122.

Beran, R. (1986) Simulated power functions. *Ann. Statist.*, **14**, 151–173.

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

———(1988) Prepivoting test statistics: a bootstrap view of asymptotic refinements. *J. Amer. Statist. Ass.*, **83**, in the press.

Beran, R. and Srivastava, M. (1985) Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.*, **13**, 95–115.

Brillinger, D. R. (1964) The asymptotic behaviour of Tukey's general method of setting approximate confidence limits (the jackknife) when applied to maximum likelihood estimates. *Rev. Inst. Int. Statist.*, **32**, 202–206.

Chan, W. S. and Tong, H. (1984) On tests for non-linearity in time series analysis. *J. Forecast.*, **5**, 217–228.

———(1988) On a likelihood ratio test for thresholds in time series. *IMS Bull.*, **17**, No. 2, Mar.–Apr., Abstr.

Chapman, P. L. and Hinkley, D. V. (1986) The double bootstrap, pivots and confidence limits. *Report 26*. Center for Statistical Sciences, University of Texas at Austin.

Cox, D. R. (1980) Local ancillarity. *Biometrika*, **67**, 273–278.

———(1987) Discussion on Better bootstrap confidence intervals *J. Amer. Statist. Ass.*, **82**, 190.

DiCiccio, T. J., Hall, P. and Romano, J. P. (1988) Comparison of parametric and empirical likelihood functions. *Technical Report 291*. Department of Statistics, Stanford University.

DiCiccio, T. J. and Romano, J. P. (1988a) On parametric bootstrap procedures for second-order accurate confidence limits. *Technical Report 293*. Department of Statistics, Stanford University.

———(1988b) Nonparametric confidence limits by resampling methods and least favorable families. *Technical Report 295*. Department of Statistics, Stanford University.

Ducharme, G. R., Jhun, M., Romano, J. P. and Truong, K. N. (1985) Bootstrap confidence cones for directional data. *Biometrika*, **72**, 637–645.

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

———(1981) Non-parametric standard errors and confidence intervals. *Can. J. Statist.*, **9**, 139–172.

———(1982) The jackknife, the bootstrap, and other resampling plans. In *Regional Conference Series in Applied Mathematics*, No. 38, ch. 5. Philadelphia: Society for Industrial and Applied Mathematics.

———(1988a) Three examples of computer-intensive statistical inference. *Report 121*. Division of Biostatistics, Stanford University.

———(1988b) More efficient bootstrap computations. *Technical Report 124*. Division of Biostatistics, Stanford University.

Feuerverger, A. (1988) On the empirical saddlepoint. To be published.

Garthwaite, P. H. and Buckland, S. T. (1988) Generating Monte Carlo confidence intervals by the Robbins–Monro process. To be published.

Ghosh, M., Parr, W., Singh, K. and Babu, G. (1984) A note on bootstrapping the sample median. *Ann. Statist.*, **12**, 1130–1135.

Gleason, J. R. (1988) Algorithms for balanced bootstrap simulations. Unpublished. Department of Psychology, Syracuse University.

Graham, R. L., Hinkley, D. V., John, P. W. M. and Shi, S. (1987) Balanced design of bootstrap simulations. *Report 48*. Center for Statistical Sciences, University of Texas at Austin.

Hall, P. (1988) Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.*, to be published.

Hall, P., DiCiccio, T. J. and Romano, J. P. (1988) On smoothing and the bootstrap. *Technical Report 286*. Department of Statistics, Stanford University.

Hall, P. and Martin, M. (1988) On bootstrap resampling and iteration. Unpublished.

Hartigan, J. (1986) Discussion on The bootstrap method for assessing statistical accuracy. *Statist. Sci.*, **1**, 54–77.

Hinkley, D. V. (1977) Jackknife confidence limits using Student $t$ approximations. *Biometrika*, **64**, 21–28.

Hinkley, D. V. and Shi, S. (1988) Importance sampling and the double bootstrap. *Report 66*. Center for Statistical Science, University of Texas at Austin.

Johns, V. (1988) Importance sampling for bootstrap confidence intervals. *J. Amer. Statist. Ass.*, to be published.

Kendall, D. G. and Kendall, W. S. (1980) Alignments in two-dimensional random sets of points. *Adv. Appl. Probabil.*, **12**, 380–424.
Lampton, M., Margon, M. and Bowyer, S. (1976) Parameter estimation in X-ray astronomy. *Astrophys. J.*, **208**, 177–190.
Loh, W. Y. (1985) A new method for testing separate families of hypotheses. *J. Amer. Statist. Ass.*, **80**, 362–368.
———(1987) Calibrating confidence coefficients. *J. Amer. Statist. Ass.*, **82**, 155–162.
———(1988) Looking up the wrong tables correctly: discussion on Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.*, **16**, in the press.
McCullagh, P. (1984) Local sufficiency. *Biometrika*, **71**, 233–244.
McLachlan, G. J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.*, **36**, 318–324.
Muirhead, R. J. (1982) *Aspects of Multivariate Statistical Theory*, p. 159. New York: Wiley.
Ogbonmwan, S. M., Verrall, R. J. and Wynn, H. P. (1987) Resampling codes and regenerated likelihoods. *Lect. Notes Econ. Math. Syst.*, **297**, 114–119.
Ogbonmwan, S. M. and Wynn, H. P. (1988) Resampling generated likelihoods. In *Statistical Decision Theory and Related Topics IV* (eds S. S. Gupta and J. O. Berger), vol. 1, pp. 133–147. New York: Springer.
Owen, A. (1988a) Empirical likelihood ratio confidence regions. *Ann. Statist.*, **16**, in the press.
———(1988b) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.
Rao, J. N. K. and Wu, C. F. J. (1988) Resampling inference with complex survey data. *J. Amer. Statist. Ass.*, **83**, in the press.
Robbins, H. and Monro, S. (1951) A stochastic approximation method. *Ann. Math. Statist.*, **22**, 400–407.
Romano, J. P. (1987) Are bootstrap confidence procedures uniform in *P*? Unpublished. Department of Statistics, Stanford University.
———(1988a) A bootstrap revival of some nonparametric distance tests. *J. Amer. Statist. Ass.*, to be published.
———(1988b) Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.*, to be published.
Shao, J. and Wu, C. F. J. (1987) A general theory for jackknife variance estimation. *Ann. Statist.*, to be published.
Shi, Xiquan (1987) Some asymptotic results for jackknifing the sample quantiles. *Preprint*. Wuhan Institute of Hydraulic and Electrical Engineering, Wuhan.
Silverman, B. W. (1981) Using kernel density estimates to investigate multimodality. *J. R. Statist. Soc.* B, **43**, 97–99.
———(1986) *Density Estimation for Statistics and Data Analysis*, section 6.3. London: Chapman and Hall.
Silverman, B. and Young, A. (1987) The bootstrap: to smooth or not to smooth? *Biometrika*, **74**, 469–479.
Taylor, M. S. and Thompson, J. R. (1986) A data based algorithm for the generation of random vectors. *Comp. Statist. Data Anal.*, **4**, 93–101.
Tibshirani, R. and Wasserman, L. (1987) Sensitive parameters. *Technical Report*. Department of Statistics, University of Toronto.
Tong, H. (1983) Threshold models in non-linear time series analysis. *Lect. Notes Statist.*, **21**.
Wetherill, G. B. (1975) *Sequential Methods in Statistics*, 2nd edn. London: Chapman and Hall.
Wu, C. F. J. (1986) Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, **14**, 1261–1350.
Young, A. (1986) Conditioned data-based simulations: some examples from geometrical statistics. *Int. Statist. Rev.*, **54**, 1–13.
———(1988) A note on bootstrapping the correlation coefficient. *Biometrika*, **75**, 370–373.