Resampling Fewer Than n Observations: Gains, Losses, and Remedies for Losses

By

P.J. Bickel[1]
Dept. of Statistics, UC Berkeley

F. Götze
Fakultät für Mathematik, Universität Bielefeld

W.R. van Zwet[1]
Dept. of Mathematics, University of Leiden

[1] Research supported by SFB 343 “Diskrete Strukturen in der Mathematik”, Universität Bielefeld

1. Introduction

Over the last 10-15 years Efron's nonparametric bootstrap has become a general tool for setting confidence regions, prediction, estimating misclassification probabilities, and other standard exercises of inference when the methodology is complex. Its theoretical justification is based largely on asymptotic arguments for its consistency or optimality. A number of examples have been addressed over the years in which the bootstrap fails asymptotically. Practical anecdotal experience seems to support theory in the sense that the bootstrap generally gives reasonable answers but can bomb.

In a recent paper Politis and Romano (1993), following Wu (1990), and independently Götze (1993) showed that what we call the m out of n without replacement bootstrap with $m = o(n)$ works to first order both in the situations where the bootstrap works and those known so far where it does not. In this paper we

- Investigate this phenomenon further and show the extent to which it is shared by the m out of n with replacement bootstrap which we show has some advantages. The m out of n with replacement bootstrap with $m = o(n)$ has been known to work in most of the examples of bootstrap failure.
- Study the price in efficiency when the usual bootstrap works that is paid for protection in case it doesn't and show how, by the use of extrapolation the price can be avoided.
- Support some of our theory with simulations.

With the exception of theorem 2 and corollary 1 we do not present essentially new results.

We suppose throughout that we observe $X_1, \ldots, X_n$ taking values in $\boldsymbol{X} = R^p$ (or more generally a separable metric space.) i.i.d. according to $F \in \mathcal{F}_0$. We are interested in the distribution of a symmetric function of $X_1, \ldots, X_n$; $T_n(X_1, \ldots, X_n, F) \equiv T_n(\hat{F}_n, F)$ where $\hat{F}_n$ is defined to be the empirical distribution of the data. More specifically we wish to estimate a parameter which we denote $\theta_n(F)$, of the distribution of $T_n(\hat{F}_n, F)$, which we denote by $\mathcal{L}_n(F)$. We will usually think of θ_n as real valued, for instance, the variance of $\sqrt{n}$ median $(X_1, \ldots, X_n)$ or the 95% quantile of the distribution of $\sqrt{n}(\bar{X} - E_F(X_1))$.

Suppose θ_n is defined naturally not just on $\mathcal{F}_0$ but on $\mathcal{F}$ which is large enough to contain all discrete distributions. It is then natural to estimate F by the nonparametric maximum likelihood estimate, $\hat{F}_n$, and hence $\theta_n(F)$ by $\theta_n(\hat{F}_n)$. This is Efron's (ideal) nonparametric (np) bootstrap. Since $\theta_n(F) \equiv \gamma(\mathcal{L}_n(F))$ and, in the cases we consider, computation of γ is straightforward the real issue is estimation of $\mathcal{L}_n(F)$. Efron's (ideal) bootstrap is to estimate $\mathcal{L}_n(F)$ by the distribution of $T_n(X_1^*, \ldots, X_n^*, \hat{F}_n)$ where, given $X_1, \ldots, X_n$, the X_i are i.i.d. $\hat{F}_n$, i.e. the bootstrap distribution of T_n. In practice, the bootstrap distribution is itself estimated by Monte Carlo or more sophisticated resampling schemes, (see deCiccio and Romano (1988) and Hinkley (1988)). We will not enter into this question further.

Theoretical analyses of the bootstrap and its properties necessarily rely on asymptotic theory, as $n \to \infty$ coupled with simulations. We follow this path and impose a minimum requirement, stability of θ_n,

(1.1) $$\theta_n(F) \to \theta(F)$$

for $F \in \mathcal{F}_0$, as $n \to \infty$.

We illustrate our point of view with a number of examples.

Examples of bootstrap success

Example 1: Suppose $\sigma^2(F) \equiv \mathrm{Var}_F(X_1) < \infty$ for all $F \in \mathcal{F}_0$.

a) Let $T_n(\hat{F}_n, F) \equiv \sqrt{n}(\bar{X} - E_F X_1)$. For the percentile bootstrap we are interested in $\theta_n(F) \equiv P_F[T_n(\hat{F}_n, F) \le t]$. Evidently $\theta(F) = \Phi\left(\frac{t}{\sigma(F)}\right)$. In fact, we want to estimate the quantiles of the distribution of $T_n(\hat{F}_n, F)$. If $\theta_n(F)$ is the $1-\alpha$ quantile then $\theta(F) = \sigma(F) z_{1-\alpha}$ where z is the Gaussian quantile

b) Let $T_n(\hat{F}_n, F) = \sqrt{n}(\bar{X} - E_F X_1)/s$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. If $\theta_n(F) \equiv P_F(T_n(\hat{F}_n, F) \le t]$ then, $\theta(F) = \Phi(t)$, independent of F.

Example 2: Suppose F has unique median $m(F)$, density $f(m(F)) > 0$ for all $F \in \mathcal{F}_0$ and $\theta_n(F) = \mathrm{Var}_F(\sqrt{n}$ median $(X_1, \ldots, X_n))$. Then $\theta(F) = [4f^2(m(F))]^{-1}$.

Note that, whereas θ_n is defined for all empirical distributions F in both examples $\theta(F)$ is not defined for such distributions in the second. Nevertheless, it is well known, see Efron (1979), that the nonparametric bootstrap is consistent in both examples in the sense that $\theta_n(\hat{F}_n) \xrightarrow{P} \theta(F)$ for $F \in \mathcal{F}_0$.

Examples of bootstrap failure

Trouble arises for the nonparametric bootstrap if θ is discontinuous, for example, $\theta_n(F) = 1(X_i = X_j$ for some $i \ne j)$ which leads to $\theta(F) = 1$ (F discrete). Although this type of breakdown may appear negligible in practice much subtler breakdowns of this type can be constructed. We give a striking example at the end of this section. More serious and common, we believe, are situations of the type successfully dealt with in example 2 where $\mathcal{F}$ is much larger than $\mathcal{F}_0$ and θ is undefined on much of $\mathcal{F}$. We give several examples of this type of failure first.

Example 3: Distribution of the minimum (Bickel, Freedman (1981)). Suppose $X_1 \ge 0$ and that all $F \in \mathcal{F}_0$ have a density F continuous and positive at 0. Let

$$T_n(\hat{F}_n) = n \min(X_1, \ldots, X_n) \ \text{ and } \ \theta_n(F) = P_F[T_n(\hat{F}_n) \ge t] = (1 - F(\frac{t}{n}))^n.$$

Then, $\theta(F) = e^{-f(0)t}$.

The (np) bootstrap fails in this case. In fact, let $N_n^*(t) = \sum_{i=1}^n 1(X_i^* < \frac{t}{n})$ where the X_i^* are a bootstrap sample from $\hat{F}_n$. Then, since $\{n\hat{F}_n(\frac{1}{n}) : t \ge 0\}$ converges weakly to a homogenous Poisson process $\{N(t) : t \ge 0\}$ with parameter $f(0)$, $\{N_n^*(\cdot)\}$ converges weakly (marginally) to a process $M(N(\cdot))$ where M is a standard Poisson process independent of N. Thus, if, in Efron's notation, we use P^* to denote conditional probability given $\hat{F}_n$ and let $\hat{F}_n^*$, be the empirical d.f. of $X_1^*, \ldots, X_n^*$ then

$$P^*[T_n(\hat{F}_n^*) \geq t] = P^*[N_n^*(t) = 0]$$

converges weakly to the random variable $P[M(N(t)) = 0|N] = e^{-N(t)}$ rather than to the desired constant $e^{-f(0)t}$.

This example is generalizable.

Example 4: Extrema. (Athreya and Fukuchi (1994), Deheuvels, Mason, Shorack (1993))

Suppose $F \in \mathcal{F}_0$ are in the domain of attraction of an extreme value distribution. That is: for some constants $A_n(F)$, $B_n(F)$,

$$n(1-F)(A_n(F) + B_n(F)x) \rightarrow H(x, F)$$

where H is necessarily one of the classical three types (David (1981), p.259). $e^{-\beta x}1(\beta x \geq 0)$, $\alpha x^{-\beta}1(x \geq 0)$, $\alpha(-x)^{\beta}1(x \leq 0)$, for $\alpha, \beta \neq 0$. Then,

$$\theta_n(f) \equiv P[(\max(X_1, \ldots, X_n) - A_n(F))/B_n(F) \leq t] \rightarrow e^{-H(t,F)} \equiv \theta(F). \tag{1.2}$$

We may without loss of generality take $A_n(F)$ to be the median of the distribution of $\max(X_1, \ldots, X_n)$ and $B_n(F)$ its interquartile range. This forces α, β in H to be independent of F. It is easy to see as before that

$$n(1 - \hat{F}_n(A_n(F) + tB_n(F))) \xrightarrow{w} N(t) \tag{1.3}$$

where N is an inhomogeneous Poisson process with parameter $H(t, F)$. Hence if $T_n(\hat{F}_n, F) = (\max(X_1, \ldots, X_n) - A_n(F))/B_n(F)$ then

$$P^*[T_n(\hat{F}_n^*, F) \leq t] \overset{w}{\Rightarrow} e^{-N(t)}. \tag{1.4}$$

It follows that the np bootstrap is inconsistent. If it were consistent, then

$$P^*[T_n(\hat{F}_n^*, \hat{F}_n) \leq t] \xrightarrow{P} e^{-H(t,F)} \tag{1.5}$$

for all t would imply that,

$$\frac{A_n(\hat{F}_n) - A_n(F)}{B_n(\hat{F}_n)} \xrightarrow{P} 0 \quad \text{and} \quad \frac{B_n(F)}{B_n(\hat{F}_n)} \xrightarrow{P} 1.$$

since given $\hat{F}_n$, $A_n(\hat{F}_n)$, $B_n(\hat{F}_n)$ are constants. Hence, $T_n(\hat{F}_n^*, F) = T_n(\hat{F}_n^*, \hat{F}_n) + o_p(1)$ and (1.4) and (1.5) contradict each other. Again, $\theta(F)$ is well defined for $F \in \mathcal{F}_0$ but not otherwise. Furthermore, small perturbations in F can lead to drastic changes in the nature of H. In fact, there are only 3 functions which are possible values of H given our choice of A_n, B_n so H is highly discontinuous in F if $\tilde{F}_0$ is as large as possible.

Example 5: Stable laws (Athreya (1987), Mammen (1992))

This is essentially equivalent to the previous example, see Loève (1978) section 2.3. Suppose that for $F \in \mathcal{F}_0$, there exists $A_n(F)$, $B_n(F)$ $\alpha(F)$, $\beta(F)$, such that,

(a) $nF(A_n(F)+B_n(F)x) \to L(x,F)$, $x<0$

(b) $n(1-F(A_n(F)+B_n(F)x)) \to L(x,F)$, $x>0$

(c) $\lim_{\epsilon\to 0}\lim_n n\int_{-\epsilon}^{\epsilon}(x-A_n(F))^2 B_n^{-2}(F)dF(x) = \beta^2(F)$

(d) $A_n(F) - n\int \frac{x}{1+x^2}dF(A_n(F)+B_n(F)x) \to \alpha(F)$

Then, if $T_n(\hat{F}_n,F) = B_n^{-1}(F)\sum_{i=1}^n X_i - A_n(F)$, $\mathcal{L}_F(T_n(\hat{F}_n,F))$ tends to a stable law with characteristic function e^{ψ} where,

$$\psi(u) = i\alpha u - \frac{\beta}{2}u^2 + \int_{R-\{0\}}(e^{iux} - 1 - \frac{iux}{1+x^2})dL(x)$$

see Loève (1978) pp. 343-350. The Lèvy functions L are necessarily of the form $\alpha|x|^{-\gamma}$. Again, the distribution of $T_n(\hat{F}_n^*,F)$, given F_n, has a random infinitely divisible limit with Lévy function $N(x)$ where N is a Poisson process with mean function $L(x)$. The bootstrap works if and only if $L\equiv 0$ as was shown by Mammen (1992). If $\theta_n(F)\equiv P_F[T_n(\hat{F}_n,F)\geq t]$, $A_n(F)$, $B_n(F)$ are defined as in example 4, we again have a well defined limit $\theta(F)$ for $F\in\mathcal{F}_0$ which is highly discontinuous. □

Example 6: Improperly centered U or V (Bretagnolle (1983)).

Let $\mathcal{F}_0 = \{F : F[-c,c]=1, E_F X_1 = 0\}$ and let $T_n(\hat{F}_n) = n\bar{X}^2 = n\int xy d\hat{F}_n(x)d\hat{F}_n(y)$. More generally, suppose $\psi : R^2 \to R$ is bounded and symmetric and let $\mathcal{F}_0 = \{F : \int\psi(x,y) = 0\ d\hat{F}_n(x)d\hat{F}_n(y)\}$ and $T_n(\hat{F}_n) = \frac{1}{n^2}\sum_{i,j}\psi(X_i,X_j)$. Then, it is easy to see that

$$T_n(\hat{F}_n) = \int\psi(x,y)dW_n^0(x)dW_n^0(y) \tag{1.6}$$

where $W_n^0(x)\equiv\sqrt{n}(\hat{F}_n(x)-F(x))$ and well known that

$$\theta_n(F)\equiv P_F[T_n(\hat{F}_n)\leq t] \to P[\int\psi(xy)dW^0(F(x))dW^0(F(y))\leq t]\equiv\theta(F)$$

where W^0 is a Brownian Bridge. On the other hand it is clear that,

$$\begin{aligned} T_n(\hat{F}_n^*) &= n\int\psi(x,y)d\hat{F}_n^*(x)d\hat{F}_n(y) \\ &= \int\psi(x,y)dW_n^*(x)dW_n^{0*}(y) \\ &\quad +2\int\psi(x,y)dW_n^0(x)dW_n^{0*}(y) \\ &\quad +\int\psi(x,y)dW_n^0(x)dW_n^0(y) \end{aligned} \tag{1.7}$$

where $W_n^{0*}(x)\equiv\sqrt{n}(\hat{F}_n^*(x)-\hat{F}_n(x))$. It readily follows that,

$$P^*[T_n(\hat{F}_n^*)\leq t] \overset{w}{\Rightarrow} P[\int\psi(x,y)dW^0(F(x))dW^0(F(y)) \tag{1.8}$$

$$+2\int \psi(x,y)dW^0(F(x))d\tilde{W}^0(F(y))$$
$$+\int \psi(x,y)d\tilde{W}^0(F(x))d\tilde{W}^0(F(y)) \le t|\tilde{W}^0]$$

where $\tilde{W}^0$, W^0 are independent Brownian Bridges.

This is again an instance where $\theta(F)$ is well defined for $F \in \mathcal{F}$ but $\theta_n(F)$ doesn't converge for $F \notin \mathcal{F}_0$. Alternatively, the special case where $\mathcal{F}_0 = \{F : E_F X_1 = 0\}$ may be seen as an instance of the bootstrap based on an inappropriate choice of estimate of F. The NPMLE for $\mathcal{F}_0$ is the empirical likelihood of Owen (1988) evaluated at $\mu = 0$. An equivalent simpler efficient estimate assigns mass $\frac{1}{n}$ to each residual $X_i - \bar{X}$, $i = 1, \ldots, n$. More generally this particular bootstrap "failure" arises when one seeks to estimate the distribution of test statistics under a semiparametric (restricted nonparametric) hypothesis and ignores the restrictions imposed by the hypothesis (see Mammen (1992) ch 2 for an example).

Example 7: Unsmooth function of $\bar{X}$ (Beran and Srivastava (1985) and Dümbgen (1993))

Let $\mathcal{F}_0 = \{F : E_F X_1^2 < \infty\}$ and

$$T_n(\hat{F}_n, F) = \sqrt{n}(h(\bar{X}) - h(\mu(F)))$$

when $\mu(F) = E_F X_1$. If h is differentiable the bootstrap distribution of T_n is, of course, consistent. But take $h(x) = |x|$, differentiable everywhere except at 0. It is easy to see then that if $\mu(F) \neq 0$, $\mathcal{L}_n(F) \to \mathcal{N}(0, \mathrm{Var}_F(X_1))$ but if $\mu(F) = 0$, $\mathcal{L}_n(F) \to |\mathcal{N}(0, \mathrm{Var}_F(X_1))|$. The reason is simply that $|\bar{X}| = \bar{X}1(\mu > 0) - \bar{X}1(\mu < 0) + |\bar{X}|1(\mu = 0)$ with probability tending to 1.

The bootstrap is consistent if $\mu \neq 0$ but not if $\mu = 0$. We can argue as follows. Under $\mu = 0$, $(\sqrt{n}(\bar{X}^* - \bar{X}), \sqrt{n}\bar{X})$ are asymptotically independent $\mathcal{N}(0, \sigma^2(F))$. Call these variable Z and Z'. Then, $\sqrt{n}(|\bar{X}^*| - |\bar{X}|) \overset{w}{\Rightarrow} |Z + Z'| - |Z'|$, a variable whose distribution is not the same as that of $|Z|$. The bootstrap distribution, as usual, converges (weakly) to the (random) conditional distribution of $|Z + Z'| - |Z'|$ given Z'. This phenomenon was first observed in a more realistic context by Beran and Srivastava. Dümbgen (1993) constructs similar reasonable though more complicated examples where the bootstrap distribution never converges. Note that, in this case $\theta_n(F)$ is well defined for all $F \in \mathcal{F}$ but is discontinuous at $\{F : \mu(F) = 0\}$.

When does the np bootstrap fail?

Our approach suggests that problems arise when:

1) θ_n is well defined on all of $\mathcal{F}$ and in particular for empirical distributions but $\theta(F)$ is defined only on the "model" $\mathcal{F}_0$.

2) θ is not smooth as a function of F on $\mathcal{F}_0$.

These are difficulties which already arise with nonparametric maximum likelihood estimation of parameters $\theta(F)$ e.g. empiricals don't have densities, but are much subtler in the more general bootstrap context since $\theta_n(F)$, as in example 3, can be very broadly defined with a narrowly defined limit. On the other hand, as example 2 illustrates, the bootstrap can work in situations where the limit is narrowly defined – but see Hall (1992) p. 319-320 for deficiencies of the (np) bootstrap to second order here.

3) In most instances where the bootstrap is inconsistent the limiting parameter $\theta(F)$ is not regularly estimable over $\mathcal{F}_0$ in the sense of Bickel, Klaassen, Ritov and Wellner (1993). However, example 6 shows that this statement has its exceptions.

4) The bootstrap is inconsistent when the statistic $T_n(\hat{F}_n, F)$ cannot at some level be identified with or approximated by a continuous function of the empirical process $\sqrt{n}(\hat{F}_n - F)$ viewed as an object weakly converging to a Gaussian limit in a suitable function space. This is suggested by the examples and has been made somewhat precise by Giné and Zinn (1989).

Possible remedies Putter and van Zwet (1993) show that if $\theta_n(F)$ is Hellinger continuous and there is a $\sqrt{n}$ Hellinger consistent estimate $\tilde{F}_n$ of F then bootstrapping from $\tilde{F}_n$ will work, i.e. $\theta_n(\tilde{F}_n)$ will be consistent except possibly for F in a "thin" set. Unfortunately, $\sqrt{n}$ Hellinger consistent estimates of F exist typically only for $\mathcal{F}_0$ regular parametric. Resampling from a smoothed version of $\hat{F}_n$, goes in this direction and indeed works for example 3 but not the others. Using the NPMLE for $\mathcal{F}_0 = \{F : E_F X_1 = 0\}$ in example 6 is another successful application of this notion.

A general approach of this type which we believe is worth investigating is to approximate $\mathcal{F}_0$ by a nested sequence of parametric models, (a sieve), $\{\mathcal{F}_{om}\}$, and use the M.L.E. $\tilde{\mathcal{F}}_{m(n)}$ for $\mathcal{F}_{om(n)}$, for a suitable sequence $m(n) \to \infty$. Unfortunately, as the Putter, van Zwet result suggests, the "right" choice of "sieve" will depend on the functional. For instance it is unclear what is appropriate for examples 4 and 5. An alternative approach which works in all of the above examples is to change not the argument of θ_n but θ_n itself. The changes we consider are the m out of n with replacement bootstrap, the $(n-m)$ out of n jackknife or $\binom{n}{m}$ bootstrap discussed by Wu (1990) and Politis and Romano (1993, 1994), and what we call the naive bootstrap below.

2. The bootstrap

Let h be a bounded real valued function defined on the range of T_n, for instance, $t \to 1$ $(t \leq t_0)$.

We view as our goal estimation of $\theta_n(F) \equiv E_F(h(T_n(\hat{F}_n, F)))$. More complicated functionals such as quantiles are governed by the same heuristics and results as those we detail below. Here are the procedures we discuss.

i) *The n/n bootstrap (The nonparametric bootstrap)*

$$\begin{aligned} B_n &\equiv \theta_n(\hat{F}_n) = E^* h(T_n(\hat{F}_n^*, \hat{F}_n)) \\ &= n^{-n} \sum_{(i_1,\ldots,i_n)} h(T_n(X_{i_1}, \ldots, X_{i_n}, \hat{F}_n)) \end{aligned}$$

ii) *The m/n bootstrap*

$$B_m \equiv \theta_m(\hat{F}_n) = n^{-m} \sum_{(i_1,\dots,i_m)} h(T_m(X_{i_1},\dots,X_{i_m},\hat{F}_n))$$

iii) *The $\binom{n}{m}$ bootstrap*

$$J_m = \binom{n}{m}^{-1} \sum_{i_1<\dots<i_m} h(T_m(X_{i_1},\dots,X_{i_m},\hat{F}_n))$$

iv) *The naive bootstrap*

Suppose $n = mk$. Define,

$$N_m \equiv k^{-1} \sum_{j=0}^{k-1} h(T_m(X_{jm+1},\dots,X_{(j+1)m},\hat{F}_n)).$$

For safety in practice one should start with a random permutation of the X_i

The naive bootstrap requires the least computation of any of the lot. Its obvious disadvantages are that it relies on an arbitrary partition of the sample and that since both m and k should be reasonably large, n has to be really substantial. The $\binom{n}{m}$ bootstrap coincides with the m/n with probability tending to 1 if $m = o(n^{1/2})$. Its advantage is that it never presents us with the ties which make resampling not look like sampling. As a consequence, as we note in theorem 1, it is consistent under really minimal conditions. On the other hand it is somewhat harder to implement by simulation and as we shall see below it is not as accurate.

Since $\theta_n(F) = \theta(F) + o(1)$, $B_{m(n)}$ and $J_{m(n)}$ where $m(n) \to \infty$ are as legitimate estimates of $\theta_n(F)$ as B_n is. A simple and remarkable result on $J_{m(n)}$ has been obtained by Politis and Romano (1993), generalizing Wu (1990). This result was also independently noted and generalized by Götze (1993). Here is a version of the Götze result.

Let $Q_{m,F}$ denote the distribution of $T_m(X_1,\dots,X_m,F)$, $Q^*_{m,n}$ the m/n bootstrap distribution of $T_m(X_1,\dots X_m,F)$, $\tilde{Q}_{m,n}$ the $\binom{n}{m}$ bootstrap distribution of $T_m(X_1,\dots,X_m,F)$ and $\hat{Q}_{m,n}$ the empirical distribution of
$\{T_m(X_{jm+1},\dots,X_{(j+1)m},\hat{F}_n\}$ $0 \le j \le k$ so that,

$$B_m = \int h(t)dQ^*_{m,n}(t) \tag{2.1}$$

$$J_m = \int h(t)d\tilde{Q}_{m,n}(t) \tag{2.2}$$

$$N_m = \int h(t)d\hat{Q}_{m,n}(t). \tag{2.3}$$

Introduce also the distributions $Q^*_{m,F}$, $\tilde{Q}_{m,F}$, $\hat{Q}_{m,F}$ given by substituting F for $\hat{F}_n$ on the right-hand of (ii)-(iv). Note that if $T_m(\hat{F}_n,F)$ does not depend on F then $Q^*_{m,n} = Q^*_{m,F}$, $\tilde{Q}_{m,n} = \tilde{Q}_{m,F}$, $\hat{Q}_{m,n} = \hat{Q}_{m,F}$.

Theorem 1: *Suppose* $\frac{m}{n} \to 0$, $m \to \infty$.

a) Then,

$$\int h(t) d\tilde{Q}_{m,F}(t) = \theta_m(F) + O_P((\frac{m}{n})^{\frac{1}{2}}).$$

If h *is continuous and*

$$T_m(X_1, \ldots, X_m, F) = T_m(X_1, \ldots, X_m, \hat{F}_n) + o_p(1) \tag{2.4}$$

then

$$J_m = \theta_m(F) + o_p(1) \tag{2.5}$$

If d *is the bounded Lipschitz metric on probability distributions on the range space of* T_m *then*

$$E_F d(\tilde{Q}_{m,F}, Q_{m,F}) = O(\frac{m}{n})^{1/2} \tag{2.6}$$

and if (2.4) holds

$$E_F d(\tilde{Q}_{m,n}, Q_{m,F}) = O(\frac{m}{n})^{\frac{1}{2}}. \tag{2.7}$$

b) The same claims are valid for $\hat{Q}_{m,F}$, $\hat{Q}_{m,n}$.

Proof: For (2.5) we may without loss of generality suppose T_m doesn't depend on F. Then, J_m is a U statistic with kernel $h(T_m(x_1, \ldots, x_m))$ and $E_F J_m = \theta_m(F)$ and (2.5) follows immediately. For (2.4) note that

$$E|J_m - \binom{n}{m}^{-1} \sum_{i_i < \ldots < i_m} h(T_m(X_{i_1}, \ldots, X_{i_m}, F))| \leq \tag{2.8}$$

$$E|h(T_m(X_1, \ldots, X_m, \hat{F}_n)) - h(T_m(X_1, \ldots, X_m, F))|$$

and (2.5) follows by bounded convergence. The proof of (2.6) and (2.7) is a standard exercise in the definition of the bounded Lipschitz metric. These results follows in the same way and even more easily for N_m. Note that if T_m doesn't depend on F, $E_F N_m = \theta_m(F)$ and,

$$\mathrm{Var}_F(N_m) = \frac{m}{n} \mathrm{Var}_F(h(T_m(X_1, \ldots, X_m))) > \mathrm{Var}_F(J_m) \quad \Box \tag{2.9}$$

For the m/n bootstrap define

$$\theta_{m,n}(F) \equiv n^{-m} \sum_{r=1}^{m} \sum \{E_F h(T_m(\underbrace{X_1, \ldots, X_1}_{i_1}, \underbrace{X_2, \ldots, X_2}_{i_2}, \ldots, \underbrace{X_r, \ldots, X_r}_{i_r}; F)) \binom{n}{r} \binom{m}{i_1, \ldots, i_r}$$

$$: i_1 + \ldots + i_r = m, i_j \geq 1\}$$

where the sum runs over all partitions of m.

Theorem 2: *Suppose* $m(n) \to \infty$ *and* $\frac{m}{n} \to 0$.

Under the conditions of theorem 1, B_m obeys the same conclusions provided that $\theta_n(F)$ is replaced by $\theta_{m,n}(F)$.

Corollary 1: *Let* $X_i^{(j)} \equiv (X_i, X_i, \ldots, X_i)_{1\times j}$ *and*

$$\delta_m(\frac{r}{m}) \equiv \max\{|E_F h(T_m(X_1^{(i_1)}, \ldots, X_r^{(i_r)}, F)) - \theta_m(F)| : \tag{2.10}$$

$$1 \le i_1 \le \ldots, \le i_r \le m, \quad i_1 + \ldots + i_r = m\},$$

and $\delta_m(x)$ *is defined by interpolation for all* $0 \le x \le 1$.

If $m = o(n)$ *and*

$$\delta_m(1 - xm^{-1/2}) \to 0$$

uniformly for $0 \le x \le M$ *for every* $M < \infty$ *then*

$$\theta_{m,n}(F) = \theta_m(F) + o(1). \tag{2.11}$$

Note that $\delta_m(1) = 0$, by definition, so that this condition is quite natural and we conjecture necessary.

The proof of theorem 2 and corollary 1 will be given in the appendix. There too we will check that, in all the examples we have discussed and some others, $J_{m(n)}$, $B_{m(n)}$, $N_{m(n)}$ are consistent for $m(n) \to \infty$, $\frac{m}{n} \to 0$.

Giné and Zinn (1989) have shown quite generally that when $\sqrt{n}(\hat{F}_n - F)$ is viewed as a member of a suitable Banach space $\mathcal{F}$ and,

a) $T_n(X_1, \ldots, X_n, F) = t(\sqrt{n}(\hat{F}_n - F))$ for t continuous

b) $\mathcal{F}$ is not too big

then B_n and $B_{m(n)}$ are consistent.

Praestgaard and Wellner (1993) extend these results to $J_{m(n)}$ with $m = o(n)$. Finally, under the Giné-Zinn conditions,

$$\|\sqrt{m}(\hat{F}_n - F)\| = (\frac{m}{n})\pi\sqrt{n}(\hat{F}_n - F)\| = O_P(\frac{m}{n})^{1/2} \tag{2.12}$$

if $m = o(n)$. Therefore,

$$t(\sqrt{m}(\hat{F}_m - \hat{F}_n)) = t(\sqrt{m}(\hat{F}_m - F)) + o_p(1) \tag{2.13}$$

and consistency of N_m if $m = o(n)$ follows from the original Giné-Zinn result.

3. Performance of B_m, J_m, and N_m as estimates of $\theta_n(F)$

As we have noted if we take $m(n) = o(n)$ then in all examples considered in which B_n is inconsistent, $J_{m(n)}$ $B_{m(n)}$, $N_{m(n)}$ are consistent. Two obvious questions are,

1) How do we choose $m(n)$?

2) Is there a price to be paid for using $J_{m(n)}$, $B_{m(n)}$, or $N_{m(n)}$ when B_n is consistent?

We shall turn to the first very difficult question in a forthcoming paper on diagnostics. The answer to the second is, in general, yes. To make this precise we take the point of view of Beran (1983) and assume that,

$$\theta_n(F) = \theta(F) + \theta'(F)n^{-1/2} + O(n^{-1}) \tag{3.1}$$

where $\theta(F)$ and $\theta'(F)$ are regularly estimable on $\mathcal{F}_0$ in the sense of Bickel, Klaassen, Ritov and Wellner (1993) and $O(n^{-1})$ is uniform on Hellinger compacts. There are a number of general theorems which lead to such expansions. See, for example, Bentkus, Götze and van Zwet (1994).

Then, under conditions which can be substantially relaxed, Beran has shown that $B_n = \theta_n(\hat{F}_n)$ is efficient as an estimate of $\theta(F) + \theta'(F)n^{-1/2}$. Under his conditions, and by essentially repeating his arguments, B_m as $m, n \to \infty$ is an efficient estimate of $\theta(F) + \theta'(F)m^{-1/2}$ and thus $\sqrt{n}(B_{m(n)} - \theta_n(F))$ has asymptotically the right variance but infinite bias $((\frac{n}{m})^{\frac{1}{2}} - 1)\theta'(F)$. Under the conditions of theorem 1, the bias behaviour of J_m, B_m, and N_m is the same so that the same efficiency 0 holds for $J_{m(n)}$ and $N_{m(n)}$.

It is true that the bootstrap is often used not for estimation but for setting confidence bounds. This is clearly the case for example 1b), the bootstrap t where $\theta(F)$ is known in advance. For example, Efron's percentile bootstrap uses the $(1-\alpha)$th quantile of the bootstrap distribution of $\bar{X}$ as a level $(1-\alpha)$ approximate upper confidence bound for μ. As is well known by now – see Hall (1993), for example, this estimate although, when suitably normalized, efficiently estimating the $(1-\alpha)$th quantile of the distribution of $\sqrt{n}(\bar{X} - \mu)$ does not improve to order $n^{-1/2}$ over the coverage probability of the usual Gaussian based $\bar{X} + z_{1-\alpha}\frac{s}{\sqrt{n}}$. However, the confidence bounds based on the bootstrap distribution of the t statistic $\sqrt{n}(\bar{X} - \mu(F))/s$ get the coverage probability correct to order $n^{-1/2}$. Unfortunately, this advantage is lost if one were to use the $1-\alpha$ quantile of the bootstrap distribution of $T_m(\hat{F}_m, F) = \sqrt{m}(\bar{X}_m - \mu(F))/s_m$ where $\bar{X}_m$ and s_m^2 are the mean and usual estimate of variance based on a sample of size m. The reason is that, in this case, the distribution function,

$$Q^*_{m,n}(t) = \Phi(t) - m^{-1/2}c(\hat{F}_n)\varphi(t)H_2(t) + O_P(m^{-1}) \tag{3.2}$$

rather than the needed,

$$Q^*_{n,n}(t) = \Phi(t) - n^{-1/2}c(\hat{F}_n)\varphi(t)H_2(t) + O_P(n^{-1}).$$

The error committed is of order $m^{-1/2}$.

The situation for $J_{m(n)}$ and $N_{m(n)}$ which function under minimal conditions, is even worse as we discuss in the next section.

4. Remedying the deficiencies of $B_{m(n)}$ when B_n is correct: Extrapolation

In Bickel and Yahav (1988), motivated by considerations of computational economy, we considered situations in which θ_n has an expansion of the form (3.1) and proposed using B_m at $m = n_0$ and $m = n_1$, $n_0 < n_1 << n$ to produce estimates of θ_n which behave like B_n.

Suppose that, as can be shown for a wide range of situations, if $m \to \infty$,

$$B_m = \theta_m(\hat{F}_n) = \theta(\hat{F}_n) + \theta'(\hat{F}_n)m^{-1/2} + O_P(m^{-1}). \tag{4.1}$$

(We sketch the argument for a special case.) Then, if $n_1 > n_0 \to \infty$

$$\theta'(\hat{F}_n) = (B_{n_0} - B_{n_1})(n_0^{-1/2} - n_1^{-1/2}) + O_P(n_0^{-1/2}) \tag{4.2}$$

$$\theta(\hat{F}_n) = \frac{n_0^{-1/2} B_{n_1} - n_1^{-1/2} B_{n_0}}{n_0^{-1/2} - n_1^{-1/2}} + O_P(n_0^{-1}) \tag{4.3}$$

and hence a reasonable estimate of B_n is,

$$B_{n_0,n_1} \equiv \frac{n_0^{-1/2} B_{n_1} - n_1^{-1/2} B_{n_0}}{n_0^{-1/2} - n_1^{-1/2}} + \frac{(B_{n_0} - B_{n_1})}{n_0^{-1/2} - n_1^{-1/2}} n^{-1/2}.$$

The estimate B_{n_0,n_1} has error which is $O_P(n_0^{-1})$ rather than the $O_P(n_1^{-1/2})$ of B_{n_1}. In particular, if $n^{-1/2} n_0 \to \infty$, B_{n_0,n_1} equals B_n to order $o_p(n^{-1/2})$. Since $\theta(F) + \frac{\theta'(F)}{n^{1/2}}$ is only estimable at rate $n^{-1/2}$, unless $\theta(F)$ is independent of F, this is much as we can expect. Assorted variations can be played on this theme depending on what we know or assume about θ_n. If, as in the case where T_n is a t statistic, the leading term $\theta(F)$ in (3.1) is $\equiv \theta_0$ independent of F, estimation of $\theta(F)$ is unnecessary and we need only one value of $m = n_0$. We are led to a simple form of estimate.

$$\hat{\theta}_{n_0} = (1 - (\frac{n_0}{n})^{1/2})\theta_0 + (\frac{n_0}{n})^{\frac{1}{2}} B_{n_0}. \tag{4.4}$$

This kind of interpolation is used to improve theoretically the behaviour of B_{m_0} as an estimate of a parameter of a stable distribution by Hall and Jing (1993) though we argue below that the improvement is somewhat illusory.

If we apply (4.5) to construct a bootstrap confidence bound the coverage probability is correct to order $n^{-1/2}$ but the error is $O_P((n_0 n)^{-1/2})$ rather than $O_P(n^{-1})$ as with B_n.

In a paper in preparation, we give some general results on asymptotic expansions for smooth functionals of the empirical process for sampling without replacement from a finite population. The essential character of these expansions, if $\frac{m}{n} = o(1)$, is

$$J_m = \theta(\hat{F}_n) + m^{-1/2}\theta'(\hat{F}_n) + \frac{m}{n}\gamma_m + O_P(m^{-1} + (\frac{m}{n})^2) \tag{4.5}$$

where γ_m is $O_P(1)$ and independent of m. The $\frac{m}{n}$ terms essentially come from the finite population correction to the variance and higher order cumulants of means of samples from a finite population. They reflect the obvious fact that if $\frac{m}{n} \to \lambda > 0$, J_m is, in general, incorrect even to first order. For instance, the variance of the $\tilde{Q}_{m,n}$ distribution corresponding to $\sqrt{m}(\bar{X} - \mu(F))$ is $\frac{1}{n}\sum(X_i - \bar{X})^2(1 - \frac{m-1}{n-1})$ which converges to $\sigma^2(F)(1-\lambda)$

if $\frac{m}{n} \to \lambda$. What this means is that if expansions (3.1), (4.1) and (4.5) are valid, then using $J_{m(n)}$ again gives efficiency 0 compared to B_n. Worse is that (4.2) with J_{m_0}, J_{n_1} replacing B_{n_0}, B_{n_1} will not work since the $\frac{n_1}{n}$ terms remain and make a contribution larger than $n^{-1/2}$ if $\frac{n_0}{n^{1/2}} \to \infty$. Essentially it is necessary to estimate the coefficient of $\frac{m}{n}$ and remove the contribution of this term also while keeping the three required values of m: $n_0 < n_1 < n_2$ such that the error $O(\frac{1}{n_0} + (\frac{n_2}{n})^2)$ is $o(n^{-1/2})$. This essentially means that n_0, n_1, n_2 have order larger than $n^{1/2}$ and smaller that $n^{3/4}$.

This effect persists if we seek to use an extrapolation of J_m for the t statistic. The coefficient of $\frac{m}{n}$ as well as $m^{-1/2}$ needs to be estimated. An alternative here and perhaps more generally is to modify the t statistic being bootstrapped and extrapolated. Thus $T_m(X_1, \ldots, X_m, F) \equiv \sqrt{m}\frac{(\bar{X}_m - \mu(F))}{\hat{\sigma}(1-\frac{m-1}{n-1})^{\frac{1}{2}}}$ leads to an expansion for J_m of the form,

$$J_m = \Phi(t) + \theta'(\hat{F}_n)m^{-1/2} + O_P(m^{-1} + (\frac{m}{n})^{1/2}), \tag{4.6}$$

and we again get correct coverage to order $n^{-1/2}$ by fitting the $m^{-1/2}$ term's coefficient, weighting it by $n^{-1/2} - m^{-1/2}$ and adding it to J_m.

If we know, as we sometimes at least suspect in symmetric cases, that $\theta(F) = 0$, we should appropriately extrapolate linearly in m^{-1} rather than $m^{-1/2}$.

The naive bootstrap situation is less satisfactory. Under (4.1), the coefficient of $\frac{1}{\sqrt{m}}$ is asymptotically constant. Put another way, the asymptotic correlation of B_m, $B_{\lambda m}$ as $m, n \to \infty$ for fixed $\lambda > 0$ is 1. This is also true for J_m under (4.5). However, consider N_m and N_{2m} (say) if $T_m = \sqrt{m}(\bar{X}_m - \mu(F))$. Let h be continuously boundedly differentiable, $n = 2km$. Then

$$\operatorname{cov}(N_m, N_{2m}) = \frac{1}{k}\operatorname{cov}(h(m^{-1/2}(\sum_{j=1}^{m}(X_j - \bar{X})))h(2m)^{-1/2}\sum_{j=1}^{2m}(X_j - \bar{X})). \tag{4.7}$$

Thus, by the central limit theorem,

$$\operatorname{corr}(N_m, N_{2m}) \to \frac{1}{2}\frac{\operatorname{cov}}{\operatorname{var}(Z_1)}(h(Z_1), h\frac{(Z_1 + Z_2)}{\sqrt{2}} \tag{4.8}$$

where Z_1, Z_2 are independent Gaussian $\mathcal{N}(0, \sigma^2(F))$ and $\sigma^2(F) = \operatorname{Var}_F(X_1)$. More generally, viewed as a process in m for fixed n. N_m centered and normalized is converging weakly to a non degenerate process. Thus, extrapolation doesn't make sense for N_m.

Two questions naturally present themselves.

(a) How do these games play out in practice rather than theory?

(b) If the expansions (4.1) and (4.5) are invalid beyond the 0th order, the usual situation, as we shall see, when the np bootstrap is inconsistent, what price do we pay theoretically for extrapolation?

Simulations giving limited encouragement in response to question (a) are given in Bickel and Yahav (1988). We give some further evidence in section 6. We turn to question (b) in the next section.

5. Behaviour of the smaller resample schemes when B_n is inconsistent, and alternatives

The class of situations in which B_n does not work is too poorly defined for us to come to definitive conclusions. But consideration of the examples suggests the following,

A. When, as in example 6, $\theta(F)$, $\theta'(F)$ are well defined and regularly estimable on $\mathcal{F}_0$ we should still be able to use extrapolation (suitably applied) to B_m and possibly to J_m to produce efficient estimates of $\theta_n(F)$.

B. When, as in all our other examples of inconsistency, $\theta(F)$ is not regularly estimable on $\mathcal{F}_0$ extrapolation shouldn't improve over the behaviour of B_{n_0}, B_{n_1}

C. If n_0, n_1 are comparable extrapolation shouldn't do particularly worse either.

D. A closer analysis of T_n and the goals of the bootstrap may, in these "irregular" cases, be used to obtain procedures which should do better than the m/n or $\binom{n}{m}$ or extrapolation bootstraps.

The only one of these claims which can be made general is C.

Proposition: *Suppose*

$$B_{n_1} = \theta_n(F) + \Omega(\epsilon_n)$$
$$B_{n_0} = \theta_n(F) + \Omega(\epsilon_n) \tag{5.1}$$

Then, if $\frac{n_0}{n_1} \not\to 1$

$$B_{n_0,n_1} = \theta_n(F) + \Omega(\epsilon_n). \tag{5.2}$$

Proof: Evidently, $\frac{B_{n_0}+B_{n_1}}{2} = \theta_n(F) + \Omega(\epsilon_n)$. On the other hand,

$$\frac{B_{n_0} - B_{n_1}}{n_0^{-1/2} - n_1^{-1/2}}\left(\frac{1}{\sqrt{n}} - \frac{1}{2}\left(\frac{1}{\sqrt{n_0}} + \frac{1}{\sqrt{n_1}}\right)\right)$$

$$= \Omega(\epsilon_n)\left(\sqrt{\frac{n_0}{n}} + \Omega(1)\right)$$

and the proposition follows. □

We illustrate the other three claims in going through the examples.

Example 3: Here

$$\theta_n(F) = e^{f(0)t}(1 + n^{-1}f'(\theta))\frac{t^2}{2}) + O(n^{-2}) \tag{5.3}$$

which is of the form (4.1). But the functional $\theta(F)$ is not regular and only estimable at rate $n^{-1/3}$ if one puts a first order Lipschitz condition on $F \in \mathcal{F}_0$. On the other hand,

$$\begin{aligned} \log B_m &= m\log(1 - \hat{F}_n(\frac{t}{m})) = m\log(1 - (\hat{F}_n(\frac{t}{m}) - \hat{F}_n(0))) \\ &= -m(F(\frac{t}{m}) - F(0)) - \frac{m}{\sqrt{n}}\sqrt{n}(\hat{F}_n(\frac{t}{m}) - F(\frac{t}{m})) \end{aligned} \tag{5.4}$$

$$+O_P(m(\hat{F}_n(\frac{t}{m}) - F(\frac{t}{m}))^2)$$

$$= \quad tf(0) + \Omega(\frac{1}{m}) + \Omega_P(\sqrt{\frac{m}{n}}) + O_P(\frac{1}{n})$$

where Ω, Ω_P indicates order of. As Politis and Romano point out, $m = \Omega(n^{1/3})$ yields the optimal rate $n^{-1/3}$ (under f Lipschitz). Extrapolation doesn't help because the $\sqrt{\frac{m}{n}}$ term is not of the form $\gamma_n\sqrt{\frac{m}{n}}$ where γ_n is independent of m. On the contrary, as a process in m, $\sqrt{mn}(\hat{F}_n(\frac{t}{m}) - F(\frac{t}{m}))$ behaves likes the sample path of a stationary Gaussian process. So conclusion B follows in this case.

Examples 4 and 5: A major difficulty here is defining $\mathcal{F}_0$ narrowly enough. If $\mathcal{F}_0$ in these examples is all distributions in the domain of attraction of stable laws or extreme value distributions it is easy to see that $\theta_n(F)$ can converge to $\theta(F)$ arbitrarily slowly. This is even true in example 1 if we remove the Lipschitz condition on f. By putting on conditions as in example 1, it is possible to obtain rates. Hall and Jing (1993) specify a possible family for example 5 in which $B_n = \Omega(n^{-\frac{1}{\alpha}})$ where α is the index of the stable law and α and the scales of the (assumed symmetric) stable distribution are not regularly estimable but for which rates such as $n^{-2/5}$ or a little better are possible. The expansions for $\theta_n(F)$ are not in powers of $n^{-1/2}$ and the expansion for B_n is even more complex. It seems evident that extrapolation doesn't help. Hall and Ji's theoretical results and simulations show that $B_{m(n)}$ though consistent, if $m(n)/n \to 0$, is a very poor estimate of $\theta_n(F)$. They obtain at least theoretically superior results. by using interpolation between B_m and the, "known up to α", value of $\theta(F)$. However, the conditions defining $\mathcal{F}_0$ which permit them to deduce the order of B_n are uncheckable so that this improvement appears illusory.

Example 7: The discontinuity of $\theta(F)$ at $\mu(F) = 0$ under any reasonable specification of $\mathcal{F}_0$ makes it clear that extrapolation cannot succeed. The discontinuity in $\theta(F)$ persists even if we assume $\mathcal{F}_0 = \{\mathcal{N}(\mu, 1) : \mu \in R\}$ and use the parametric bootstrap. In the parametric case it is possible to obtain constant level confidence bounds by inverting the tests for $H : |\mu| = |\mu_0|$ vs $K : |\mu| > |\mu_0|$ using the noncentral χ^2_1 distribution of $(\sqrt{n}\bar{X})^2$. Asymptotically conservative confidence bounds can be constructed in the nonparametric case by forming a bootstrap confidence interval for $\mu(F)$ using $\bar{X}$ and then taking the image of this interval into $\mu \to |\mu|$. So this example illustrates points B and D.

Example 6: This provides our support for claim A. We give the argument for $T_n(\hat{F}_n) = n\bar{X}^2$. Using a standard application of the approach of Bhattacharya and Ghosh (1978) to the Edgeworth expansions of $m^{1/2}(\bar{X}^*_m - \bar{X})$ obtained by Singh (1981) for example we get if $m = o(n)$, $E_F X_1 = 0$, $E_F X_1^4 < \infty$, and X_1 obeys Cramér's condition that

$$P^*[[\sqrt{m}\bar{X}^*_m]^2 \leq t] = \tag{5.5}$$

$$= 2\Phi(\frac{\sqrt{t}}{\hat{\sigma}}) - 1 + B_n(t)\frac{m}{n} + A_n(t)\sqrt{\frac{m}{n}} + O_P(\frac{1}{m}) + O_P(\frac{m}{n})$$

where A_n and B_n are $O_P(1)$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. (There is no $m^{-1/2}$ term because of symmetry.) If $m^2 \to \infty$, $\frac{m^{\frac{3}{2}}}{n} \to 0$ the remainder terms are $o_p(n^{-1/2})$. Therefore, if by extrapolation we remove the $\frac{m}{n}$ and $\sqrt{\frac{m}{n}}$ terms we obtain an estimate equivalent to $2\Phi(\sqrt{t}/\hat{\sigma}) - 1$ which is efficient as an estimate of $\theta(F) = 2\Phi(\frac{\sqrt{t}}{\sigma(F)}) - 1$ where $\sigma^2(F) = \mathrm{Var}_F X_1$. So a general extrapolation prescription is to remove any terms in $\sqrt{\frac{m}{n}}$ and $\frac{m}{n}$ and add the coefficient of any term in $\frac{1}{\sqrt{m}}$ weighted by $\frac{1}{\sqrt{n}}$. It is easy to see that such a prescription applies to J_m also. Unfortunately, it hardly seems worthwhile and in practice it is probably not good to estimate so many parameters using the $\{B_m\}$ or $\{J_m\}$ plot for m small. Analysis of the present situation of course reveals that the same effect can be achieved easily by resampling from the residuals $X_i - \bar{X}$. Or in the general U or V statistic case, bootstrapping not $T_m(\hat{F}_n, F) \equiv n\int \psi(x,y)d\hat{F}_n(x)d\hat{F}_n(y)$ but rather $n\int \psi(x,y)d(\hat{F}_n - F)(x)d(\hat{F}_n - F)(y)$ is the right thing to do. So this example illustrates both A and D. $\Box$

6. Simulations and Conclusions

The simulation algorithms were written and carried out by Adele Cutler and Jiming Jiang. We are very grateful. Two situations were simulated, one already studied in Bickel and Yahav (1988) where the bootstrap is consistent (essentially example 1) the other (essentially example 3) where the bootstrap is inconsistent.

Example 1:

We considered samples from $F = \chi_1^2$.

Sample sizes: $n = 50, 100, 400$

Bootstrap sample sizes: $B = 500$

Simulation size: $N = 2000$

Statistics:

i) $T_m = \sqrt{m}(\sqrt{\bar{X}_m} - \sqrt{\mu(F)})$

ii) $T_m = \sqrt{m}\frac{(\bar{X}_m - \mu(F))}{s_m}$

where

$$s_m^2 = \frac{1}{m-}\sum_{i=1}^m (X_i - \bar{X}_m)^2.$$

Parameters of resampling distributions: $G_m^{-1}(.1)$, $G_m^{-1}(.9)$ where G_m is the distribution of T_m under the appropriate resampling scheme. We use B, J, N to distinguish the schemes. These parameters were used to form upper and lower "90%" confidence bounds for $\theta \equiv \sqrt{\mu(F)}$ as in

i):

$$\bar{\theta}_{mB} = \sqrt{\bar{X}_n} - \frac{1}{\sqrt{n}}G_{mB}^{-1}(.1) \tag{6.1}$$

for the "90%" upper confidence bound based on the m/n bootstrap.
ii):

$$\bar{\theta}_{mB} = ((\bar{X}_n - \frac{s_n}{\sqrt{n}} G_{mB}^{-1}(.1))_+)^{1/2} \tag{6.2}$$

where G_{mB} now corresponds to the t statistic. $\underline{\theta}_{mB}$, $\underline{\theta}_{mJ}$, $\underline{\theta}_{mN}$ are defined similarly. We also specify 2 subsample sizes $n_0 < n_1$ for the extrapolation bounds, $\underline{\theta}_{n_0,n_1}$ $\bar{\theta}_{n_0,n_1}$. These are defined in (i) for example by,

$$\bar{\theta}_{n_0,n_1} = \sqrt{\bar{X}_n} - \frac{1}{\sqrt{n}}\{\frac{(G_{n_0B}^{-1}(.1) + G_{n_1B}^{-1}(.1))}{2} \tag{6.3}$$

$$+(n^{-1/2} - \frac{1}{2}(n_0^{-1/2} + n_1^{-1/2}))(G_{n_0B}^{-1}(.1) - G_{n_1B}^{-1}(.1)/(n_0^{-1/2} - n_1^{-1/2}).$$

We consider roughly, $n_0 = 2\sqrt{n}$, $n_1 = 4\sqrt{n}$ and specifically, the triples (n, n_0, n_1) : $(50, 15, 30), (100, 20, 40)$ and $(400, 40, 80)$. We table,

$CP \equiv$ Coverage probability

$RMSE = \sqrt{E(\text{Bound} - \text{Parameter})^2}$

This is done for the B_{m_1}, B_n, and B_{n_0,n_1} bounds, based on the t statistics. The percentile bootstrap, as in Bickel-Yahav (1988), has CP and $RMSE$ for B_n, B_{n_0,n_1} and B_{n_1} agreeing to the accuracy of the Monte Carlo and we omit these tables. We also give some results for J_m and N_m which indicate their shortcomings as well as some additional results for the B bounds under different distributions and sample sizes.

Example 3:

We consider samples from $F = \chi_2^2 =$ exponential with scale 2. Samples and simulation sizes are as above.

Statistic: $T_m = m(\min(X_1, \dots, X_m) - F^{-1}(0))$

Parameters of resampling distribution: $H_m^{-1}(.1)$, $H_m^{-1}(.9)$ where H_m is the distributions of T_m under the appropriate resampling scheme. We use B, J, N as before and put a confidence bound on $\theta = F^{-1}(0)$ as in,

$$\bar{\theta}_m = \max(X_1, \dots, X_n) - \frac{1}{n} G_{mB}^{-1}(.9).$$

The extrapolation bound is defined as

$$\underline{\theta}_{n_0,n_1} = \min(X_1, \dots, X_n) - \frac{1}{n}\frac{(G_{n_0B}^{-1}(.9) + G_{n,B}^{-1}(.9))}{2}$$

$$+(n^{-1} - \frac{(n_0^{-1} + n_1^{-1})}{2})(G_{n_0B}^{-1}(.9) - G_{n_1B}^{-1}(.9))(n_0^{-1} - n_1^{-1}).$$

The upper bounds turn out to coincide at $\min_i X_i$ in our simulations and are conservative but uninformative. We give representative results for these bounds and also partial results for N_m.

Example 1; t statistic B boostraps

Table 1:

		Coverage probabilities (CP)			*RMSE*		
		B	B1	BR	B	B1	BR
50							
	UB	.88	.90	.88	.19	.21	.19
	LB	.90	.90	.90	.15	.15	.15
100							
	UB	.90	.93	.89	.13	.14	.12
	LB	.91	.90	.91	.11	.10	.11
400							
	UB	.91	.94	.90	.06	.07	.06
	LB	.91	.90	.91	.05	.05	.05

Notes B1 corresponds to $m = n1$. BR corresponds to $n_0 = 2\sqrt{n}$, $n_1 = 4\sqrt{n}$ as specified.

Example 2; Min statistic bootstrap

Table 2:

		CP	*RMSE*
50			
	B	.75	.01
	B1	.78	.07
	BR	.70	.07
	B1S	.82	.07
	BRS	.80	.07
100			
	B	.75	.04
	B1	.82	.03
	BR	.76	.04
	B1S	.87	.03
	BRS	.86	.03
400			
	B	.75	.09
	B1	.86	.01
	BR	.83	.01

Notes: (1) B1 corresponds to $m = n1 = 4\sqrt{n}$. BR to $n_0 = 2\sqrt{n}$, $n_1 = 4\sqrt{n}$. B1S to $m = n_1 = 16$, BRS to $n_0 = 4$, $n_1 = 16$.

2) Extrapolation formula used $\frac{1}{m}$ rather than $\frac{1}{\sqrt{m}}$.

3) In all cases these are lower confidence bounds.

Table 3: The N bootstrap upper bound and its B competitor

Example 1

	CP		*RMSE*	
	N	BO	N	BO
50	.98	.99	.79	.43
100	.97	.98	.31	.20
400	.95	.95	.09	.07

Example 2

	CP		*RMSE*	
	N	BO	N	BO
50	.90	.87	.09	.06
100	.90	.88	.05	.03
400	.90	.89	.01	.06

Note: (1) In all cases $m = \frac{n}{10}$ and the UCB was calculated.

Table 4: The J bootstrap and competitor in example 1 (i)

		CP		*E(Length)*	
n	m	J	B	J	B
50	16	.82	.88	.07	.09
100	16	.86	.88	.04	.05
500	40	.88	.90	.01	.01

Note: These figures are for simulation sizes of $N = 500$ and for 90% percentile confidence intervals.

Conclusions: The conclusions we draw are limited by the range of our simulations. We opted for realistic sample sizes, of 50, 100 and a less realistic 400. For $n = 50$, 100 the subsample sizes $n_1 = 30$ (for $n = 50$) and 40 (for $n = 100$) are of the order $n/2$ rather than $o(n)$. For all sample sizes $n_0 = 2\sqrt{n}$ is not really "of larger order than $\sqrt{n}$". The simulations in fact show the asymptotic as very good when the bootstrap works even for relatively small sample sizes. The story when the bootstrap doesn't work is less clear.

When the bootstrap works (Example 1)

- BR and B are very close both in terms of CP, and RMSE even for $n = 50$ from table 1.

- B1's CP though sometimes better than B's consistently differs more from B's and its RMSE follows suit In particular, for UB in table 1, the RMSE of B1 is generally larger. LB exhibits less differences but this reflects that UB is governed by the behaviour of χ_1^2 at 0. In simulations we do not present we get similar sharper differences for LB when F is a heavy tailed distribution such as Pareto with $EX^5 = \infty$.

- The effects, however, are much smaller than we expected. This reflects that these are corrections to the coefficient of the $n^{-1/2}$ term in the expansion. Perhaps the most surprising aspect of these tables is how well B1 performs.

- From table 3 we see that because the m we are forced to by the level considered is small, CP for the N bounds can differ from the nominal level. But it is close to that of the corresponding BO bound. However, its RMSE is considerably worse. Nevertheless, it is clear that these easy, naive bounds advocated earnestly to us by John Hartigan and anonymous physicists are well worth studying further.

- The J_m bounds are inferior as table 4 shows. This reflects the presence of the finite population correction $\frac{m}{n}$, even though these bounds were considered for the more favorable sample size $m = 16$ for $n = 50$, 100 rather than $m = 30$, 40. Corrections such as those of Bertail (1994) or simply applying the finite population correction to s would probably bring performance up to that of B_{n_1}. But the added complication doesn't seem worthwhile.

When the bootstrap doesn't work (Example 2)

- From table 2, as expected, the CP of B for the lower confidence bound was poor for all n. For $n_0 = 2\sqrt{n}$, $n_1 = 4\sqrt{n}$, CP for B1 was constantly better than B for all n. BR is worse than B1 but improves with n and was nearly as good as B1 for $n = 400$. For small n_0, n_1 both B1 and BR do much better.

We did not give results for the upper confidence bound because the granularity of the bootstrap distribution of $\min X_1$ for these values of m and n made the upper confidence bound $= \min_i X_i$ and $CP = 1$ in all cases.

Evidently, n_0, n_1 play a critical role here. What apparently is happening is that for n_0, n_1 not sufficiently small compared with n extrapolation picks up the wrong slope and moves the not so good B1 bound even further towards the poor B bound.

The principal message of these simulations to us is that extrapolation of the B_m plot may carry risks not fully revealed by the asymptotics. The greatest value of the B_m plot may lie in the construction of diagnostics which we are now pursuing. On the other hand, if n_0 and n_1 are chosen in a reasonable fashion extrapolation on the $\sqrt{n}$ scale works well when the bootstrap does. Two notes based on simulations we do not present should be added to the optimism of Bickel, Yahav (1988) however. There may be risk if n_0 is really small compared to $\sqrt{n}$. We obtained poor results for BR for the t statistics for $n_0 = 4$ and 2. Thus $n_0 = 4$, $n_1 = 16$ gave the wrong slope to the extrapolation which tended to overshoot badly. Also, taking n_1 and n_0 close to each other, as the theory of the 1988 paper suggests is appropriate for statistics possessing high order expansions when the expansion coefficients are deterministic, gives poor results. It can also be seen theoretically that the sampling variability of the bootstrap for m of the order $\sqrt{n}$ makes this prescription unreasonable.

Appendix

Proof of Theorem 2: For $1 \le i_1 \le i_2 \le \ldots \le i_r \le m$, $i_1 + \ldots + i_r = m$, let

$$U(i_1, \ldots, i_r) = \frac{1}{\binom{n}{r}\binom{m}{i_1,\ldots,i_r}} \sum \{h(T_m(X_{j_1}^{(i_1)}, \ldots, X_{j_r}^{(i_r)}, F) : \quad 1 \le j_1 \ne j_2 \ne \ldots \ne j_r \le n\}$$

where $X_j^{(i)} = (X_j, \ldots, X_j)_{1\times i}$. Then, by arguing as for U statistics, (the special case: $r = m$, $i_1 = \ldots = i_m = 1$)

(a) $\mathrm{Var}U(i_1, \ldots, i_r) \le \|h\|_\infty^2 \frac{r}{n}$

On the other hand,

(b) $EU(i_1, \ldots, i_r) = Eh(T_m(X_1^{(i_1)}, X_2^{(i_2)}, \ldots, X_r^{(i_r)}))$

and

(c) $B_m = \sum\{w(i_1, \ldots, i_r)U(i_1, \ldots, i_r) : 1 \le i_1 \le i_2 \le \ldots \le i_r \le m,\ i_1 + \ldots + i_r = m\}$

where

$$w(i_1, \ldots, i_r) = \binom{n}{r}\binom{m}{i_1, \ldots, i_r}/n^m$$

and

(d) $\sum\{w(i_1, \ldots, i_r) : 1 \le i_1 \le \ldots \le i_r \le m,\ \sum_{j=1}^r i_j = m\} = 1$

Thus, by (c)

(e) $\mathrm{Var}^{\frac{1}{2}} B_m \le \sum\{w(i_1, \ldots, i_r)\mathrm{Var}^{\frac{1}{2}} U(i_1, \ldots, i_r)\}$:

$$\le \max \mathrm{Var}^{\frac{1}{2}} U(i_1, \ldots, i_r) \le \|h\|_\infty (\frac{m}{n})^{\frac{1}{2}}$$

by (a).

The theorem follows from (b) and (e). □

Proof of Corollary 1:

Since,

(a)

$$|\theta_{m,n}(F)-\theta(F)|=\ \sum_{r=1}^{m}\sum\{(E_F h(T_m(X_1^{(i_1)},\ldots,X_r^{(i_r)},F))-\theta(F))\frac{\binom{n}{r}\binom{m}{i_1,\ldots,i_r}}{n^m}: \\ i_j\geq 1,\ 1\leq j\leq r, \sum_{j=1}^{r} i_j = m\},$$

by theorem 2 we need only show that

$$\sum_{r=1}^{m}\delta_m(\frac{r}{m})P_{m,n}[R_m=r]\to 0$$

where

(b) $P_{m,n}[R_m{=}r]=\{\#$ of partitions of m into r ordered integers $\geq 1\}$ $\binom{n}{r}\binom{m}{i_1,\ldots,i_r}/n^m$.

Expression (b) is easily recognized as the probability of getting $n-r$ empty cells when throwing n balls independently into m boxes without restrictions — see Feller (1968) p.19. Then it is well known or easily seen that

(c) $E_{m,n}(R_m)=n(1-(1-\frac{1}{n})^m)$

(d) $\mathrm{Var}_{m,n}(R_m)=n\{(1-\frac{1}{n})^m-(1-\frac{2}{n})^m\}+n^2\{(1-\frac{2}{n})^m-(1-\frac{1}{n})^{2m}\}$.

It is easy to check that, if $m=o(n)$

(e) $E_{m,n}(R_m)=m(1+O(\frac{m}{n}))$

(f) $\mathrm{Var}_{m,n}(R_m)=O(m)$

so that $\frac{R_m}{m}=1+O_P(m^{-1/2})$.

The corollary follows from (e) and (f). □

Note that this implies that the m/n bootstrap works if about $\sqrt{m}$ ties do not affect the value of T_m much.

Checking that J_m, B_m, N_m $m=o(n)$ works

The arguments we give for B_m also work for J_m only more easily since theorem 1 can be verified. It is usually easier to directly verify that, in all our examples, $Q^*_{m,n}$ converges weakly (in probability) to the limit Q_F of $Q_{m,F}$ and conclude that theorem 2 holds for all h continuous and bounded than to check the conditions of theorem 2 or corollary 1. However, we note that these conditions can be verified in all our examples for suitable h.

Example 1a): Verification is hardly needed since the result follows from Giné and Zinn (1989). We do sketch heuristically how one would argue for functionals considered in section 2 rather than quantiles. For J_m we need only check that (2.6) holds since $\sqrt{m}(\bar{X}-$

$\mu(F)) = o_p(1)$. For B_m we note heuristically that the distribution of $m^{-1/2}(i_1 X_1 + \ldots + i_r X_r)$ differs from that of $m^{-1/2}(X_1 + \ldots + X_m)$ by $O(\sum_{j=1}^r \frac{(i_j^2-1)}{m})$. If we maximize $\sum_{j=1}^r (i_j^2 - 1)$ subject to $\sum_{j=1}^r i_j = m$, $i_j \geq 1$ we obtain $\frac{2(m-r)}{m} + \frac{(m-r)^2}{m}$. Thus for suitable h, $\delta_m(x) = 2(1-x) + \frac{1}{\sqrt{m}}(1-x)^2$ and the hypothesis of corollary 1 holds.

Example 1b): The conditions of the theorems don't apply immediately here until we notice that,

$$P[\sqrt{n}\frac{(\bar{X} - \mu(F))}{s} \leq t] = P[\sqrt{n}(\bar{X} - \mu(F)) - st \leq 0]$$

and we can apply the previous arguments to $T_n(\hat{F}_n, F) \equiv \sqrt{n}(\bar{X} - \mu(F)) - st$.

Example 2: In example 2 the variance corresponds to $h(x) = x^2$ if $T_m(\hat{F}_m, F) = m^{1/2}(\text{med}(X_1, \ldots, X_m) - F^{-1}(\frac{1}{2}))$. An argument parallel to that in Efron (1979) works. However, let us again verify the condition of corollary 1 for bounded h and T_m as above. Note that,

$$(a) \qquad P[\text{med}(X_1^{(i_1)}, \ldots, X_r^{(i_r)}) \neq \text{med}(X_1^{(i_1)}, \ldots X_r^{(i_r - 1)}, X_{r+1})] \leq \frac{1}{r+1}.$$

Thus,

$$(b) \qquad P[\text{med}(X_1^{(i_1)}, \ldots, X_r^{(i_r)}) \neq \text{med}(X_1, \ldots, X_m)]$$

$$\leq \sum_{j=r+1}^{m} \frac{1}{j} \leq \log(\frac{m}{r}).$$

Hence for h bounded,

$$\delta_m(x) \leq \|h\|_\infty \log(\frac{1}{x}).$$

and we can apply corollary 1.

Example 3: Follows by checking (2.4) in theorem 1 and that theorem 2 applies for J_m by arguing as above for B_m, (see Politis and Romano (1993)). Alternatively, argue as in the example, noting that $\{m\hat{F}_n(\frac{t}{m}) : t \geq 0\}$ converges weakly to the function (degenerate process) $tf(0)$ since $\text{Var}(m\hat{F}_n(\frac{t}{m})) = \frac{m^2}{n} F(\frac{t}{m})(1 - F(\frac{t}{m})) \to 0$. Hence the bootstrap process, (given $\hat{F}_n$),

$$(c) \qquad N_{m,n}^*(t) \equiv \sum_{i=1}^{m} 1(X_i^* < \frac{t}{m})$$

converges to a Poisson process with parameter $f(0)$ and $P^*[N_{m,n}^*(t) = 0] \xrightarrow{P} e^{-f(0)t}$ as it should. This argument is given in Athreya and Fukushi (1994). $\square$

Example 4: The same arguments apply.

Example 5: We can argue as in example 3 above to check that the random Lévy function of the bootstrap distribution converges to the Lévy function (degenerate process) of the desired limit.

Example 6: Here again it is easier to apply the descriptions (1.6) and (1.7) and note that for $T_m(\hat{F}_m^*)$, $W_n^{0*}(x)$ is replaced by $W_n^{0*}(x)$ while W_n^0 is replaced by $\sqrt{\frac{m}{n}}W_n^0$ so that the last two terms in (1.7) are asymptotically negligible and the m/n bootstrap works here.

Example 7: Theorem 1 applied to $T_m(\hat{F}_m) = \sqrt{m}|\bar{X}_m|$ under $\mu(F) = 0$, $EX_1^2 < \infty$ is clearly valid. The condition of corollary 1 is subtler to apply but can be shown to work if h has a Taylor expansion to two terms. Again a direct argument that the m/n bootstrap works for $\sqrt{m}\bar{X}_m$ suffices. The same line may be taken with the Beran, Srivastava or Dümbgen examples.

References

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

Athreya, K.B. and Fukuchi, J. (1994). Bootstrapping extremes of I.I.D. random variables. Proceedings of Conference on Extreme Value Theory (NIST)

Bentkus, V., Götze, F. and van Zwet, W.R. (1994). An Edgeworth expansion for symmetric statistics. *Tech Report* Univ. of Bielefeld.

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

Bhattacharya, R. and Ghosh, J.K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6**, 434-451.

Bickel, P.J., Klaassen, C.K., Ritov, Y. and Wellner, J. (1993). Efficient and Adaptive Estimation in Semiparametric Models. Johns Hopkins University Press, Baltimore.

Bickel, P.J. and Yahav, Y. (1988). Richardson extrapolation and the bootstrap. *J. Amer. Statist. Assoc.*

Bretagnolle, J. (1983). Lois limites du bootstrap de certaines fonctionelles. *Ann. Inst. H. Poincaré, Sec. B* **19**, 281-296.

Deheuvels, P., Mason, D. and Shorack, G. (1993). Some results on the influence of extremes on the bootstrap. *Ann. Inst. H. Poincaré* **29**, 83-103.

DeCiccio, T.J. and Romano, J.P. (1989). The automatic percentile method: Accurate confidence limits in parametric models, *Can. J. Statist.* **17**, 155-169.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Technical Report* UC Berkeley.

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

Efron, B. and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman & Hall, London, New York.

Feller W. (1968). Probability Theory v1. J. Wiley, New York.

Giné, E. and Zinn, J. (1989). Necessary conditions for the bootstrap of the mean. *Ann. Statist.* **17**, 684-691.

Götze, F. (1993). *Bulletin* I.M.S.

Hall, P. (1992). The Bootstrap and Edgeworth Expansion. Springer Verlag, New York.

Hall, P. and Jing Bing, Y. (1993). Performance of boostrap for heavy tailed distributions. *Tech Report* A.N.U. Canberra.

Hinkley, D.V. (1988). Bootstrap methods (with discussion). *J. Royal. Statist. Soc. B* **50**, 321-337.

Loève, M. (1978). Probability theory v1. Springer Verlag, New York.

Mammen, E. (1992). When does bootstrap work? Springer Verlag, New York.

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

Politis, D.N. and Romano, J.P. (1992). A general theory for large sample confidence regions based on subsamples under minimal assumptions. Technical Report No. 399, Department of Statistics, Stanford University.

Praestgaard, J. and Wellner, J. (1993). Exchangeably weighted bootstrap of the general empirical process. *Ann. Probab.* **21**, 2053-2086.

Putter, H. and van Zwet, W.R. (1993). Consistency of plug in estimators with applications to the bootstrap. Submitted to *Ann. Statist.*

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

Wu, C.F.J. (1990). On the asymptotic properties of the jackknife histogram. *Ann. Statist.* **18**, 1438-1452.

Addresses of the authors:

P. Bickel, Dept. of Statistics, UC Berkeley, 367 Evans Hall, Berkeley Ca 94720, U.S.A.

F. Götze, Universität Bielefeld, Fakultät für Mathematik,Universitätsstr., D-33501 Bielefeld, Germany

W. R. van Zwet, Dept. of Mathematics, University of Leiden, PO Box 9512, 2300 RA Leiden, Netherlands