Chapter 6
Boostrap Resampling

Peter Hall

6.1 Introduction to Four Bootstrap Papers

6.1.1 Introduction and Summary

In this short article we discuss four of Peter Bickel's seminal papers on theory and methodology for the bootstrap. We address the context of the work as well as its contributions and influence. The work began at the dawn of research on Efron's bootstrap. In fact, Bickel and his co-authors were often the first to lay down the directions that others would follow when attempting to discover the strengths, and occasional weaknesses, of bootstrap methods.

Peter Bickel made major contributions to the development of bootstrap methods, particularly by delineating the range of circumstances where the bootstrap is effective. That topic is addressed in the first, second and fourth papers treated here. Looking back over this work, much of it done 25–30 years ago, it quickly becomes clear just how effectively these papers defined the most appropriate directions for future research.

We shall discuss the papers in chronological order, and pay particular attention to the contributions made by Bickel and Freedman (1981), since this was the first article to demonstrate the effectiveness of bootstrap methods in many cases, as well as to raise concerns about them in other situations. The results that we shall introduce in Sect. 6.1.2, when considering the work of Bickel and Freedman (1981), will be used frequently in later sections, especially Sect. 6.1.5.

The paper by Bickel and Freedman (1984), which we shall discuss in Sect. 6.1.3, pointed to challenges experienced by the bootstrap in the context of stratified

P. Hall (✉)
Department of Mathematics and Statistics, The University of Melbourne, Melbourne, VIC, Australia
e-mail: halpstat@ms.unimelb.edu.au

J. Fan et al. (eds.), *Selected Works of Peter J. Bickel*, Selected Works in Probability and Statistics 13, DOI 10.1007/978-1-4614-5544-8_6,
© Springer Science+Business Media New York 2013

sampling. This is ironic, not least because some of the earliest developments of what, today, are called bootstrap methods, involved sampling problems; see, for example, Jones (1956), Shiue (1960), Gurney (1963) and McCarthy (1966, 1969).

Section 6.1.4 will treat the work of Bickel and Yahav (1988), which contributed very significantly to methodology for efficient simulation, at a time when the interest in this area was particularly high. Bickel et al. (1997), which we shall discuss in Sect. 6.1.5, developed deep and widely applicable theory for the m-out-of-n bootstrap. The authors showed that their approach overcame consistency problems inherent in the conventional n-out-of-n bootstrap, and gave rates of convergence applicable to a large class of problems.

6.1.2 *Laying Foundations for the Bootstrap*

Thirty years ago, when Efron's (1979) bootstrap method was in its infancy, there was considerable interest in the extent to which it successfully accomplished its goal of estimating parameters, variances, distributions etc. As Bickel and Freedman (1981) noted, Efron's paper "gives a series of examples in which [the bootstrap] principle works, and establishes the validity of the approach for a general class of statistics when the sample space is finite." Bickel and Freedman (1981) set out to assess the bootstrap's success in a much broader setting than this.

In the early 1980s, saying that the bootstrap "works" meant that bootstrap methods gave consistent estimators, and in this sense were competitive with more conventional methods, for example those based on asymptotic analysis. Within about 5 years the goals had changed; it had been established that bootstrap methods "work" in a very wide variety of circumstances, and, although there were counterexamples to this general rule, by the mid 1980s the task had become largely one of comparing the effectiveness of the bootstrap relative to more conventional techniques. But in 1981 the extent to which the bootstrap was consistent was still largely unknown. Bickel and Freedman (1981) contributed mightily to the process of discovery there.

In particular, Bickel and Freedman (1981) were the first to establish rigorously that bootstrap methodology is consistent in a wide range of settings. The impact of their paper was dramatic. It provided motivation for exploring the bootstrap more deeply in a great many settings, and furnished some of the mathematical tools for that development. In the same year, in fact in the preceding paper in the *Annals*, Singh (1981) explored second-order properties of the bootstrap. However, Bickel and Freedman (1980) also took up that challenge at a particularly early stage.

As a prelude to describing the results of Bickel and Freedman (1981) we give some notation. Let $\mathcal{X}_n = X_1, \ldots, X_n$ denote a sample of n independent observations from a given univariate distribution with finite variance σ^2, write $\bar{X}_n = n^{-1} \sum_i X_i$ for the sample mean, and define

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 ,$$

the bootstrap estimator of σ^2. Let $\chi_m^* = \{X_1^*, \ldots, X_m^*\}$ denote a resample of size m drawn by sampling randomly, with replacement, from χ, and put $\bar{X}_m^* = m^{-1} \sum_{i \le m} X_i^*$. Bickel and Freedman's (1981) first result was that, in the case of m-resamples, the m-resample bootstrap version of $\hat{\sigma}_n^2$, i.e.

$$\hat{\sigma}_m^{*2} = \frac{1}{m} \sum_{i=1}^{m} (X_i^* - \bar{X}_m^*)^2 ,$$

converges to σ^2 as both m and n increase, in the sense that, for each $\varepsilon > 0$,

$$P(|\hat{\sigma}_m^* - \sigma| > \varepsilon \,|\, \chi_n) \to 0 \tag{6.1}$$

with probability 1. Moreover, Bickel and Freedman (1981) showed that the conditional distribution of $m^{1/2}(\bar{X}_m^* - \bar{X}_n)$, given χ_n, converges to the normal $\mathrm{N}(0, \sigma^2)$ distribution. Taking $m = n$, the latter property can be restated as follows:

$$\begin{aligned} &\text{the probabilities } P\{n^{1/2}(\hat{\theta}^* - \hat{\theta}) \le \sigma x \,|\, \chi_n\} \text{ and } P\{n^{1/2}(\hat{\theta} - \theta) \le \sigma x\} \\ &\text{both converge to } \Phi(x), \text{ the former converging with probability 1,} \end{aligned} \tag{6.2}$$

where Φ denotes the standard normal distribution and, on the present occasion, $\theta = E(X_i)$, $\hat{\theta} = \bar{X}_n$ and $\hat{\theta}^* = \bar{X}_n^*$.

The second result established by Bickel and Freedman (1981) was a generalisation of this property to multivariate settings. Highlights of subsequent parts of the paper included its contributions to theory for the bootstrap in the context of functionals of a distribution function. For example, Bickel and Freedman (1981) considered von Mises functionals of a distribution function H, defined by

$$g(H) = \int\int \omega(x,y)\, dH(x)\, dH(y) ,$$

where the function ω of two variables is symmetric, in the sense that $\omega(x,y) = \omega(y,x)$, and where

$$\int\int \omega(x,y)^2\, dH(x)\, dH(y) + \int \omega(x,x)^2\, dH(x) < \infty . \tag{6.3}$$

If we take H to be either $\widehat{F}_n$, the empirical distribution function of the sample χ_n, or $\widehat{F}_n^*$, the version of $\widehat{F}_n$ computed from χ_n^*, then

$$g(\widehat{F}_n) = \frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \omega(X_{i_1}, X_{i_2}) , \quad g(\widehat{F}_n^*) = \frac{1}{n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \omega(X_{i_1}^*, X_{i_2}^*) .$$

Bickel and Freedman (1981) studied properties of this quantity. In particular they proved that if (6.3) holds with $H = F$, denoting the common distribution function of the X_is, then the distribution of $n^{1/2}\{g(\widehat{F}_n^*) - g(\widehat{F}_n)\}$, conditional on the data, is asymptotically normal $\mathrm{N}(0, \tau^2)$ where

$$\tau^2 = 4\left[\int\left\{\int \omega(x,y)\,dF(y)\right\}^2 dF(x) - g(F)^2\right].$$

This limit distribution is the same as that of $n^{1/2}\{g(\widehat{F}_n) - g(F)\}$, and so the above result of Bickel and Freedman (1981) confirms, in the context of von Mises functions of the empirical distribution function, that (6.2) holds once again, provided that σ there is replaced by τ and we redefine $\theta = g(F)$, $\hat{\theta} = g(\widehat{F}_n)$ and $\hat{\theta}_n^* = g(\widehat{F}_n^*)$. That is, the bootstrap correctly captures, once more, first-order asymptotic properties. Subsequent results of Bickel and Freedman (1981) also showed that the same property holds for the empirical process, and in particular that the process $n^{1/2}(\widehat{F}_n^* - \widehat{F}_n)$ has the same first-order asymptotic properties as $n^{1/2}(\widehat{F}_n - F)$. Bickel and Freedman (1981) also derived the analogue of this result for the quantile process.

Importantly, Bickel and Freedman (1981) addressed cases where the bootstrap fails to enjoy properties such as (6.2). In their Sect. 6 they gave two counterexamples, one involving U-statistics and the other, spacings between extreme order statistics, where the bootstrap fails to capture large-sample properties even to first order. In both settings the problems are attributable, at least in part, to failure of the bootstrap to correctly capture the relationships among very high-ranked, or very low-ranked, order statistics, and in that context we shall relate below some of the issues to which Bickel and Freedman's (1981) work pointed. This account will be given in detail because it is relevant to later sections.

Let $X_{(1)} < \ldots < X_{(n)}$ denote the ordered values in $\mathcal{X}_n$; we assume that the common distribution of the X_is is continuous, so that the probability of a tie equals zero. In this case the probability, conditional on $\mathcal{X}_n$, of the event $\mathcal{E}_n$ that the largest X_i, i.e. $X_{(n)}$, is in $\mathcal{X}_n^*$, equals 1 minus the conditional probability that $X_{(n)}$ is not contained in in $\mathcal{X}_n^*$. That is, it equals $1 - (1 - n^{-1})^n = 1 - e^{-1} + O(n^{-1})$. Therefore, as $n \to \infty$,

$$P(X_{(n)}^* = X_{(n)} \,|\, \mathcal{X}_n) = P(X_{(n)} \in \mathcal{X}_n^* \,|\, \mathcal{X}_n) \to 1 - e^{-1},$$

where the convergence is deterministic. Similarly, for each integer $k \geq 1$,

$$\pi_{nk} \equiv P(X_{(n)}^* = X_{(n-k)} \,|\, \mathcal{X}_n) \to \pi_k \equiv e^{-k}\left(1 - e^{-1}\right) \tag{6.4}$$

as $n \to \infty$; again the convergence is deterministic. Consequently the distribution of $X_{(n)}^*$, conditional on $\mathcal{X}_n$, is a mixture, and in particular is equal to $X_{(n-k)}$ with probability π_{nk}, for $k \geq 1$. Therefore:

given $\varepsilon>0$ and any metric, for example the Lévy metric, between distributions, we may choose $k = k(\varepsilon) \geq 1$ so large that the distribution of $X^*_{(n)}$, conditional on χ_n, is no more than ε from the discrete mixture $\sum_{0\leq j\leq k} X_{(n-j)} I_j$,where (a) exactly one of the random variables $I_1, I_2, \ldots$ is nonzero, (b) that variable takes the value 1, and (c) $P(I_k = 1) = \pi_k$ for $k \geq 0$. The upper bound of ε applies deterministically, in that it is valid with probability 1, in an unconditional sense. (6.5)

To indicate the implications of this property we note that, for many distributions F, there exist constants a_n and b_n, at least one of them diverging to infinity in absolute value as n increases; and a nonstationary stochastic process $\xi_1, \xi_2, \ldots$; such that, for each $k \geq 0$, the joint distribution of $(X_{(n)} - a_n)/b_n, \ldots, (X_{(n-k)} - a_n)/b_n$ converges to the distribution of $(\xi_1, \ldots, \xi_k)$. See, for example, Hall (1978). In view of (6.5) the distribution function of $(X^*_{(n)} - a_n)/b_n$, conditional on χ_n, converges to that of

$$Z = \sum_{j=0}^{\infty} \xi_j I_j \,,$$

where the sequence $I_1, I_2, \ldots$ is distributed as in (6.5) and is chosen to be independent of $\xi_1, \xi_2, \ldots$. In this notation,

$$P(X^*_{(n)} - a_n \leq b_n z \,|\, \chi_n) \to P(Z \leq z) \tag{6.6}$$

in probability, whenever z is a continuity point of the distribution of Z. On the other hand,

$$P(X_{(n)} - a_n \leq b_n z) \to P(\xi_1 \leq z) \,. \tag{6.7}$$

A comparison of (6.6) and (6.7) reveals that there is little opportunity for estimating consistently the distribution of $X_{(n)}$, using standard bootstrap methods. Bickel and Freedman (1981) first drew our attention to this failing of the conventional bootstrap. The issue was to be the object of considerable research for many years after the appearance of Bickel and Freedman's paper. Methodology for solving the problem, and ensuring consistency, was eventually developed and scrutinised; commonly the m-out-of-n bootstrap is used. See, for example, Swanepoel (1986), Bickel et al. (1997) and Bickel and Sakov (2008).

6.1.3 *The Bootstrap in Stratified Sampling*

Bickel and Freedman (1984) explored properties of the bootstrap in the case of stratified sampling from finite or infinite populations, and concluded that, with appropriate scaling, the bootstrap can give consistent distribution estimators in cases where asymptotic methods fail. However, without the proper scaling the bootstrap can be inconsistent.

The problem treated is that of estimating a linear combination,

$$\gamma = \sum_{j=1}^{p} c_j \mu_j , \tag{6.8}$$

of the means $\mu_1, \ldots, \mu_p$ of p populations $\Pi_1, \ldots, \Pi_p$ with corresponding distributions $F_1, \ldots, F_p$. The c_js are assumed known, and the μ_js are estimated from data. To construct estimators, a random sample $\chi(j) = \{X_{j1}, \ldots, X_{jn_j}\}$ is drawn from the jth population, and the sample mean $\underset{\sim}{X}(j) = n_j^{-1} \sum_i X_{ji}$ is computed in each case. Bickel and Freedman (1984) considered two different choices of c_j, valid in two respective cases: (a) if it is known that each $E(X_{ji}) = \mu$, not depending on j, and that the variance σ_j^2 of Π_j is proportional to r_j, say, then

$$c_j = \frac{n_j / r_j}{\sum_k (n_k / r_k)} ;$$

and (b) if the populations are finite, and in particular Π_j is of size N_j for $j = 1, \ldots, p$, then

$$c_j = \frac{N_j}{\sum_k N_k} .$$

In either case the estimator $\hat{\gamma}$ of γ reflects the definition of γ at (6.8):

$$\hat{\gamma} = \sum_{j=1}^{p} c_j \bar{X}(j) ,$$

where $\bar{X}(j)$ is the mean value of the data in $\chi(j)$.

In both cases Bickel and Freedman (1984) showed that, particularly if the sample sizes n_j are small, the bootstrap estimator of the distribution of $\hat{\gamma} - \gamma$ is not necessarily consistent, in the sense that the distribution estimator minus the true distribution may not converge to zero in probability. The asymptotic distribution of $\hat{\gamma} - \gamma$ is normal $\mathrm{N}(0, \tau_1^2)$, say; and the bootstrap estimator of that distribution, conditional on the data, is asymptotically normal $\mathrm{N}(0, \tau_2^2)$; but the ratio τ_1^2 / τ_2^2 does not always converge to 1. Bickel and Freedman (1984) demonstrated that this difficulty can be overcome by estimating scale externally to the bootstrap process, in effect incorporating a scale correction to set the bootstrap on the right path. Bickel and Freedman also suggested other, more ad hoc remedies.

These contributions added immeasurably to our knowledge of the bootstrap. Combined with the counterexamples given earlier by Bickel and Freedman (1981), those authors showed that the bootstrap was not a device that could be used naively in all cases, without careful consideration.

Some researchers, a little outside the statistics community, had felt that bootstrap resampling methods freed statisticians from influence by a mathematical "priesthood" which was "frank about viewing resampling as a frontal attack upon their own situations" (Simon 1992). To the contrary, the work of Bickel and Freedman (1981,

1984) showed that a mathematical understanding of the problem was fundamental to determining when, and how, to apply bootstrap methods successfully. They demonstrated that mathematical theory was able to provide considerable assistance to the introduction and development of practical bootstrap methods, and they provided that aid to statisticians and non-statisticians alike.

6.1.4 Efficient Bootstrap Simulation

By the mid to late 1980s the strengths and weaknesses of bootstrap methods were becoming more clear, especially the strengths. However, computers with power comparable to that of today's machines were not readily available at the time, and so efficient methods were required for computation. The work of Bickel and Yahav (1988) was an important contribution to that technology. It shared the limelight with other approaches to achieving computational efficiency, including the balanced bootstrap, which was a version for the bootstrap of Latin hypercube sampling and was proposed by Davison et al. (1986) (see also Graham et al. 1990); importance resampling, suggested by Davison (1988) and Johns (1988); the centring method, proposed by Efron (1990); and antithetic resampling, introduced by Hall (1990).

The main impediment to quick calculation for the bootstrap was the resampling step. In the 1980s, when for many of us computing power was in short supply, bootstrap practitioners nevertheless advocated thousands, rather than hundreds, of simulations for each sample. For example Efron (1988), writing for an audience of psychologists, argued that "It is not excessive to use 2,000 replications, as in this paper, though we might have stopped at 1,000." In fact, if the number of simulations, B, is chosen so that the nominal coverage level of a confidence interval can be expressed as $b/(B+1)$, where b is an integer, then the size of B has very little bearing on the coverage accuracy of the interval; (see Hall 1986). However, choosing B too small can result in overly variable Monte Carlo approximations to endpoints for bootstrap confidence intervals, and to critical points for bootstrap hypothesis tests.

It is instructive here to relate a story that G.S. Watson told me in 1988, the year in which Bickel and Yahav's paper was published. Throughout his professional life Watson was an enthusiast of the latest statistical methods, and the bootstrap was no exception. Shortly after the appearance of Efron's (1979) seminal paper he began to experiment with the percentile bootstrap technique. Not for Watson a tame problem involving a sample of scalar data; in what must have been one of the first applications of the bootstrap to spatial or spherical data, he used that technique to construct confidence regions for the mean direction derived from a sample of points on a sphere. He wrote a program that constructed bootstrap confidence regions, put the code onto a floppy disc, and passed the disc to a Princeton geophysicist to experiment with. This, he told the geophysicist, was the modern alternative to conventional confidence regions based on the von Mises-Fisher distribution. The latter regions, of course, took their shape from the mathematical form of the fitted distribution, with relatively little regard for any advice that the data might have to offer. What did the geophysicist think of the new approach?

In due course Watson received a reply, to the effect that the method was very interesting and remarkably flexible, adapting itself well to quite different datasets. But it had a basic flaw, the geophysicist said, that made it unattractive—every time he applied the code on the floppy disc to the same set of spherical data, he got a different answer! Watson, limited by the computational resources of the day, and by the relative complexity of computations on a sphere, had produced software that did only about $B = 40$ simulations each time the algorithm was implemented. Particularly with the extra degree of freedom that two dimensions provided for fluctuations, the results varied rather noticeably from one time-based simulation seed to another.

This tale defines the context of Bickel and Yahav's (1988) paper. Their goal was to develop algorithms for reducing the variability, and enhancing the accuracy in that sense, of Monte Carlo procedures for implementing the bootstrap. Their approach, a modification for the bootstrap of the technique of Richardson extrapolation (a classical tool in numerical analysis; see Jeffreys and Jeffreys 1988, p. 288), ran as follows. Let $\widehat{F}_n$ (not to be confused with the same notation, but having a different meaning, in Sect. 6.1.2) denote the data-based distribution function of interest, and let F_n be the quantity of which $\widehat{F}_n$ is an approximation. For example, $\widehat{F}_n(x)$ might equal $P(\hat{\theta}_n^* - \hat{\theta}_n \le x \,|\, \chi_n)$, where $\hat{\theta}_n$ denotes an estimator of a parameter θ, computed from a random sample χ_n of size n, in which case $\hat{\theta}_n^*$ would be the bootstrap version of $\hat{\theta}_n$. (In this example, $F_n(x) = P(\hat{\theta}_n - \theta \le x)$.) Instead of estimating $\widehat{F}_n$ directly, compute estimators of the distribution functions $\widehat{F}_{n_1}, \ldots, \widehat{F}_{n_r}$, where the sample sizes $n_1, \ldots, n_r$ are all smaller than n, and in fact so small that $n_1 + \ldots + n_r$ is markedly less than n. In some instances we may also know the limit F_∞ of F_n, or at least its form, $\widetilde{F}_\infty$ say, constructed by replacing any unknown quantities (for example, a variance) by estimators computed from χ_n. The quantities $\widehat{F}_{n_1}, \ldots, \widehat{F}_{n_r}$ and $\widetilde{F}_\infty$ are much less expensive, i.e. much faster, to compute than $\widehat{F}_n$, and so, by suitable "interpolation" from these functions, we can hope to get a very good approximation to $\widehat{F}_n$ without going to the expense of actually calculating the latter.

In general the cost of simulating, or equivalently the time taken to simulate, is approximately proportional to $C_n B$, where C_n depends only on n and increases with that quantity. Techniques for enhancing the performance of Monte Carlo methods can either directly produce greater accuracy for a given value of B (the balanced bootstrap has this property), or reduce the value of C_n and thereby allow a larger value of B (hence, greater accuracy from the viewpoint of reduced variability) for a given cost. Bickel and Yahav's (1988) method is of the latter type. By enabling a larger value of B it alleviates the problem encountered by Watson and his geophysicist friend.

Bickel and Yahav's (1988) technique is particularly widely applicable, and has the potential to improve efficiency more substantially than, say, the balanced bootstrap. Today, however, statisticians' demands for efficient bootstrap methods have been largely assuaged by the development of more powerful computers. In the last 15 years there have been very few new simulation algorithms tailored to the bootstrap. Philippe Toint's aphorism that "I would rather have today's algorithms on yesterday's computers, than vice versa," loses impact when an algorithm is to some

extent problem-specific, and its implementation requires skills that go beyond those needed to purchase a new, faster computer.

6.1.5 *The m-Out-of-n Bootstrap*

The m-out-of-n bootstrap is another example revealing that, in science, less is often more. Bickel and Freedman (1981, 1984) had shown that the standard bootstrap can fail, even at the level of statistical consistency, in a variety of settings; and, as we noted in Sect. 6.1.2, the m-out-of-n bootstrap, where m is an order of magnitude smaller than n, is often a remedy. Swanepoel (1986) was the first to suggest this method, which we shall define in the next paragraph. Bickel et al. (1997) made major contributions to the study of its theoretical properties. We shall give an example that provides further detail than we gave in Sect. 6.1.2 about the failure of the bootstrap in certain cases. Then we shall summarise briefly the contributions made by Bickel et al. (1997).

Consider drawing a resample $\chi_m^* = \{X_1^*, \ldots, X_m^*\}$, of size m, from the original dataset $\chi_n = \{X_1, \ldots, X_n\}$ of size n, and let $\hat{\theta} = \hat{\theta}_n$ denote the bootstrap estimator of θ computed from χ_n. In particular, if we can express θ as a functional, say $\theta(F)$, of the distribution function F of the data X_i, then

$$\hat{\theta}_n = \theta(\widehat{F}_n)\,, \tag{6.9}$$

where $\widehat{F}_n$ is the empirical distribution function computed from χ_n. Likewise we can define $\hat{\theta}_m^* = \theta(\widehat{F}_m^*)$, where $\widehat{F}_m^*$ is the empirical distribution function for χ_m^*. As we noted in Sect. 2, Bickel and Freedman (1981) showed that first-order properties of $\hat{\theta}_m^*$ are often robust against the value of m. In particular it is often the case that, for each $\varepsilon > 0$,

$$P(|\hat{\theta}_m^* - \hat{\theta}_n| > \varepsilon \,|\, \chi_n) \to 0\,, \quad P(|\hat{\theta}_n - \theta| > \varepsilon) \to 0 \tag{6.10}$$

as m and n diverge, where the first convergence is with probability 1. Compare (6.1). For example, (6.10) holds if θ is a moment, such as a mean or a variance, and if the sampling distribution has sufficiently many finite moments.

The definition (6.9) is conventionally used for a bootstrap estimator, and it does not necessarily involve simulation. For example, if $\theta = \int x dF(x)$ is a population mean then

$$\hat{\theta}_n = \int x d\widehat{F}_n(x) = \bar{X}\,, \quad \hat{\theta}_m^* = \int x d\widehat{F}_m^*(x) = \bar{X}^*$$

are the sample mean and resample mean, respectively. However, in a variety of other cases the most appropriate way of defining and computing $\hat{\theta}_n$ is in terms of the resample χ_n^*; that is, χ_m^* with $m = n$. Consider, for instance, the case where

$$\theta = P\left(X_{(n)} - X_{(n-1)} > X_{(n-1)} - X_{(n-2)}\right)\,, \tag{6.11}$$

in which, as in Sect. 6.1.2, we take $X_{(1)} < \ldots < X_{(n)}$ to be an ordering of the data in $\mathcal{X}_n$, assumed to have a common continuous distribution. For many sampling distributions, in particular distributions that lie in the domain of attraction of an extreme-value law, θ depends on n but converges to a strictly positive number as n increases.

In this example the bootstrap estimator, $\hat{\theta}_n$, of θ, based on a sample of size n, is defined by

$$\hat{\theta}_n = P\Big(X^*_{(n)} - X^*_{(n-1)} > X^*_{(n-1)} - X^*_{(n-2)} \;\Big|\; \mathcal{X}_n\Big), \tag{6.12}$$

where $X^*_{(1)} \le \ldots \le X^*_{(n)}$ are the ordered data in $\mathcal{X}^*_n$. Analogously, the bootstrap version, $\hat{\theta}^*_n$, of $\hat{\theta}_n$ is defined using the double bootstrap:

$$\hat{\theta}^*_n = P\Big(X^{**}_{(n)} - X^{**}_{(n-1)} > X^{**}_{(n-1)} - X^{**}_{(n-2)} \;\Big|\; \mathcal{X}^*_n\Big),$$

where $X^{**}_{(1)} \le \ldots \le X^{**}_{(n)}$ are the ordered data in $\mathcal{X}^{**}_n = \{X^{**}_1, \ldots, X^{**}_n\}$, drawn by sampling randomly, with replacement, from $\mathcal{X}^*_n$. However, for the reasons given in the paragraph containing (6.5), property (6.10) fails in this example, no matter how we choose m. (The m in (6.2) is different from the m for the m-out-of-n bootstrap.) The bootstrap fails to model accurately the relationships among large order statistics, to such an extent that, in the example characterised by (6.11), $\hat{\theta}_n$ does not converge to θ.

This problem evaporates if, in defining $\hat{\theta}_n$ at (6.12), we take the resample $\mathcal{X}^*_m$ to have size $m = m(n)$, where

$$m \to \infty \quad \text{and} \quad m/n \to 0 \tag{6.13}$$

as $n \to \infty$. That is, instead of (6.12) we define

$$\hat{\theta}_n = P\Big(X^*_{(m)} - X^*_{(m-1)} > X^*_{(m-1)} - X^*_{(m-2)} \;\Big|\; \mathcal{X}_n\Big), \tag{6.14}$$

where $X^*_1, \ldots, X^*_m$ are drawn by sampling randomly, with replacement, from $\mathcal{X}_n$. In this case, provided (6.5) holds, (6.2) is correct in a wide range of settings.

Deriving this result mathematically takes a little effort, but intuitively it is rather clear: By taking m to be of strictly smaller order than n we ensure that the probability that $X^*_{(m)}$ equals any given data value in $\mathcal{X}_n$, for example $X_{(n)}$, converges to zero, and so the difficulties raised in the paragraph containing (6.5) no longer apply. In particular, instead of (6.4) we have:

$$P(X^*_{(m-k)} = X_{(m-\ell)} \,|\, \mathcal{X}_n) \to 0$$

in probability, for each fixed, nonnegative integer k and ℓ, as $n \to \infty$. Further thought along the same lines indicates that the conditional distribution of $X^*_{(m)} - X^*_{(m-1)}$ should now, under mild assumptions, be a consistent estimator of the distribution of $X_{(n)} - X_{(n-1)}$.

Bickel et al. (1997) gave a sequence of four counter-examples illustrating cases where the bootstrap fails, and provided two examples of the success of the bootstrap. The first two counter-examples relate to extrema, and so are closely allied to the example considered above. The next two treat, respectively, hypothesis testing and improperly centred U and V statistics, and estimating nonsmooth functionals of the population distribution function. Bickel et al. (1997) then developed a deep, general theory which allowed them to construct accurate and insightful approximations to bootstrap statistics $\hat{\theta}_n$, such as that at (6.9), not just in that case but also when $\hat{\theta}_n$ is defined using the m-out-of-n bootstrap, as at (6.14). This enabled them to show that, in a large class of problems for which (6.13) holds, the m-out-of-n bootstrap overcomes consistency problems inherent in the conventional n-out-of-n approach, and also to derive rates of convergence.

A reliable way of choosing m empirically is of course necessary if the m-out-of-n bootstrap is to be widely adopted. In many cases this is still an open problem, although important contributions were made recently by Bickel and Sakov (2008).

References

Bickel PJ, Freedman DA (1980) On Edgeworth expansions and the bootstrap. Unpublished manuscript

Bickel PJ, Freedman DA (1981) Some asymptotic theory for the bootstrap. Ann Stat 9:1196–1217

Bickel PJ, Freedman DA (1984) Asymptotic normality and the bootstrap in stratified sampling. Ann Stat 12:470–482

Bickel P, Sakov A (2008) On the choice of m in the m out of n bootstrap and confidence bounds for extrema. Stat Sin 18:967–985

Bickel PJ, Yahav JA (1988) Richardson extrapolation and the bootstrap. J Am Stat Assoc 83: 387–393

Bickel PJ, Götze F, van Zwet WR (1997) Resampling fewer than n observations: gains, losses, and remedies for losses. Stat Sin 7:1–31

Davison AC (1988) Discussion of papers by D.V. Hinkley and by T.J. DiCiccio and J.P. Romano. J R Stat Soc Ser B 50:356–357

Davison AC, Hinkley DV, Schechtman E (1986) Saddlepoint approximations in resampling methods. Biometrika 75:417–431

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

Efron B (1988) Bootstrap confidence intervals: good or bad? (with discussion.) Psychol Bull 104:293–296

Efron B (1990) More efficient bootstrap computations. J Am Stat Assoc 85:79–89

Graham RL, Hinkley DV, John PWM, Shi S (1990) Balanced design of bootstrap simulations. J R Stat Soc Ser B 52:185–202

Gurney M (1963) The variance of the replication method for estimating variances for the CPS sample design. Memorandum, U.S. Bureau of the Census. Unpublished

Hall P (1978) Representations and limit theorems for extreme value distributions. J Appl Probab 15:639–644

Hall P (1986) On the number of bootstrap simulations required to construct a confidence interval. Ann Stat 14:1453–1462

Hall P (1990) Antithetic resampling for the bootstrap. Biometrika 76:713–724

Jeffreys Y, Jeffreys BS (1988) Methods of mathematical physics, 3rd edn. Cambridge University Press, Cambridge

Johns MV (1988) Importance sampling for bootstrap confidence intervals. J Am Stat Assoc 83:709–714
Jones HL (1956) Investigating the properties of a sample mean by employing random subsample means. J Am Stat Assoc 51:54–83
Mccarthy PJ (1966) Replication: an approach to the analysis of data from complex surveys. In: National Center for Health Statistics, Public Health Service (eds) Vital Health Statistics: Series 2. Public Health Service publication 1000, vol 14. U.S. Government Printing Office, Washington, DC
Mccarthy PJ (1969) Pseudo-replication: half samples. Rev Int Stat Inst 37:239–264
Shiue C-J (1960) Systematic sampling with multiple random starts. For Sci 6:42–50
Simon JC (1992) Barriers to adoption, and the future of resampling; resistances to using and teaching resampling. Unpublished
Singh K (1981) On the asymptotic accuracy of Efron's bootstrap. Ann Stat 9:1187–1195
Swanepoel JWH (1986) A note on proving that the (modified) bootstrap works. Commun Stat Ser A 15:3193–3203

The Annals of Statistics
1981, Vol. 9, No. 6, 1196-1217

SOME ASYMPTOTIC THEORY FOR THE BOOTSTRAP

BY PETER J. BICKEL[1] AND DAVID A. FREEDMAN[2]

University of California, Berkeley

Efron's "bootstrap" method of distribution approximation is shown to be asymptotically valid in a large number of situations, including t-statistics, the empirical and quantile processes, and von Mises functionals. Some counter-examples are also given, to show that the approximation does not always succeed.

1. Introduction. Efron (1979) discusses a "bootstrap" method for setting confidence intervals and estimating significance levels. This method consists of approximating the distribution of a function of the observations and the underlying distribution, such as a pivot, by what Efron calls the bootstrap distribution of this quantity. This distribution is obtained by replacing the unknown distribution by the empirical distribution of the data in the definition of the statistical function, and then resampling the data to obtain a Monte Carlo distribution for the resulting random variable. This method would probably be used in practice only when the distributions could not be estimated analytically. However, it is of some interest to check that the bootstrap approximation is valid in situations which are simple enough to handle analytically. Efron gives a series of examples in which this principle works, and establishes the validity of the approach for a general class of statistics when the sample space is finite.

In Section 2 of the present paper, it will be shown that the bootstrap works for means, and hence for pivotal quantities of the familiar "t-statistic" sort; an extension to multi-dimensional data will be made. Section 3 deals with U-statistics and other von Mises functionals, and suggests the wide scope of the theory. Section 4 deals with the empirical process: one application is setting confidence bounds for the theoretical distribution function, even if the latter has a discrete component. In Section 5, the quantile process will be bootstrapped, leading to confidence intervals for quantiles. Trimmed means and Winsorized variances are also studied. In Section 6 some examples will be given where the bootstrap fails, for instance, when estimating θ from variables uniformly distributed over $[0, \theta]$.

Some of the problems discussed in this paper have been studied independently by Singh (1981).

2. Bootstrapping the mean. Let $X_1, X_2, \cdots, X_n$ be independent random variables with common distribution function F. Assume that F has finite mean μ and variance σ^2, both unknown. The conventional estimate for μ is the sample average, denoted here by μ_n. To analyze the sampling error in μ_n, it is customary to compute the sample standard deviation s_n, defined as

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_n)^2.$$

Received July 14, 1980; revised March 11, 1981.

[1] Research partially supported by Office of Naval Research, Contract N-00014-80-C-0163, and the Adolph and Mary Sprague Miller Foundation.

[2] Research partially supported by National Science Foundation Grant MCS-80-02535.

AMS 1970 *subject classifications.* Primary 62E20; secondary 62G05, 62G15.

Key words and phrases. Bootstrap, resampling, asymptotic theory.

By the Classical Central Limit Theorem, the distribution of the pivotal quantity

$$Q_n = \sqrt{n}(\mu_n - \mu)/s_n$$

tends weakly to $N(0, 1)$. So, in this situation, the asymptotics are known. However, there is some theoretical interest in seeing how the bootstrap would perform.

Let F_n be the empirical distribution of $X_1, \cdots, X_n$, putting mass $1/n$ on each X_i. The next step in the bootstrap method is to resample the data. Given $(X_1, \cdots, X_n)$, let $X_1^*, \cdots, X_m^*$ be conditionally independent, with common distribution F_n. We have allowed the resample size m to differ from the number n of data points, to estimate the distribution of the bootstrap pivotal quantity $Q_m^* = \sqrt{m}(\mu_m^* - \mu_n)/s_m^*$, where $\mu_m^* = (1/m)\sum_{i=1}^m X_i^*$ and $s_m^* = (1/m)\sum_{i=1}^m (X_i^* - \mu_m^*)^2$.

In the resampling, the n data points $X_1, \cdots, X_n$ are treated as a population, with distribution function F_n and mean μ_n; and μ_m^* is considered as an estimator of μ_n. First, take $m = n$. The idea is that the behavior of the bootstrap pivotal quantity Q_n^* mimics that of Q_n. Thus, the distribution of Q_n^* could be computed from the data and used to approximate the unknown sampling distribution of Q_n. Or even more directly, the bootstrap distribution of $\sqrt{n}(\mu_n^* - \mu_n)$ could be used to approximate the sampling distribution of $\sqrt{n}(\mu_n - \mu)$. Either approach would lead to confidence intervals for μ, and would be useful if the Central Limit Theorem were not available, and if the bootstrap approximation were valid.

Now take $m \neq n$. The resample size m does have some statistical import. For instance, a sample of size n can be bootstrapped to see what would happen with a sample of size n^2, or $\sqrt{n}$, or 10. Furthermore, with m and n free to vary separately, the second-moment condition in Theorem 2.1 becomes quite natural. If m goes to infinity first, then the conditional law of $\sqrt{m}(\mu_m^* - \mu_n)$ tends to normal, with mean 0 and variance s_n^2. As n tends to infinity, this converges if and only if s_n^2 does.

Mathematically, there is something rather delicate even about the present simple case, with $m = n$. Comparing the classical $\sqrt{n}(\mu_n - \mu)$ with the bootstrap $\sqrt{n}(\mu_n^* - \mu_n)$, the parameter μ is replaced by μ_n. But this change is of the critical order of magnitude, namely $1/\sqrt{n}$, and cannot be ignored. However, there is a second error: the X's have been replaced by X^*'s. In fact, these two errors cancel each other to a large extent. Our proof will make this idea precise, by showing that the distribution of the pivot does not change much if the empirical F_n is replaced by the theoretical F. The theorem is an asymptotic one, so the data $X_1, \cdots, X_n$ should be visualized as the beginning segment of an infinite series.

THEOREM 2.1. *Suppose $X_1, X_2, \cdots$ are independent, identically distributed, and have finite positive variance σ^2. Along almost all sample sequences $X_1, X_2, \cdots$, given $(X_1, \cdots, X_n)$, as n and m tend to ∞:*

(a) The conditional distribution of $\sqrt{m}(\mu_m^* - \mu_n)$ converges weakly to $N(0, \sigma^2)$.

(b) $s_m^* \to \sigma$ in conditional probability: that is, for ϵ positive,

$$P\{|s_m^* - \sigma| > \epsilon \,|\, X_1, \cdots, X_n\} \to 0 \text{ a.s.}$$

Relations (a) and (b) imply that the asymptotic distribution of the bootstrap pivot Q_n^* coincides with the classical one: convergence to the standard normal holds. There are several equivalent ways to prove these results. We choose an argument which is qualitative, but requires some machinery. Let Γ_2 be the set of distribution functions G satisfying $\int x^2\, dG(x) < \infty$, and introduce the following notion of convergence in Γ_2:

$$G_\alpha \Rightarrow G \quad \text{iff} \quad G_\alpha \to G \text{ weakly} \quad \text{and} \quad \int x^2\, dG_\alpha(x) \to \int x^2\, dG(x).$$

The strong law implies

$$F_n \Rightarrow F \quad \text{along almost all sample sequences.} \tag{2.1}$$

The conclusions of the theorem hold along any such sample sequence.

Our notion of convergence in Γ_2 is metrizable, for instance, by a "Mallows metric" d_2. The d_2-distance between G and H in Γ_2 is defined as follows: $d_2(G, H)^2$ is the infimum of $E\{(X - Y)^2\}$ over all joint distributions for the pair of random variables X and Y whose fixed marginal distributions are G and H respectively. This metric was introduced in Mallows (1972) and Tanaka (1973); it is related to the Vassershtein metrics of Dobrushin (1970), Major (1978), or Vallender (1973). For a detailed discussion of d_2, see Section 8 of the present paper.

Now let $Z_1(G), \cdots, Z_m(G)$ be independent random variables, with common distribution function G. Let $G^{(m)}$ be the distribution of

$$S_m(G) = m^{-1/2} \sum_{j=1}^m [Z_j(G) - E\{Z_j(G)\}].$$

If $G \in \Gamma_2$, so is $G^{(m)}$. By Lemma 3 of Mallows (1972),

$$d_2[G^{(m)}, H^{(m)}] \leqq d_2[G, H]. \tag{2.2}$$

Also see Lemma 8.7 below, and (8.2).

PROOF OF THEOREM 2.1, *Part a.* The bootstrap construction can be put into present notation as follows: conditionally, the law of $\sqrt{m}(\mu_m^* - \mu_n)$ is just $F_n^{(m)}$. But F_n is close to F in the d_2-metric on Γ_2, by (2.1). So $F_n^{(m)}$ is close to $F^{(m)}$ by (2.2). Now use the ordinary Central Limit Theorem on $F^{(m)}$.

Part b. This can be proved the same way. Let Γ_1 be the set of G's with $\int |x| G(dx) < \infty$, and define the metric d_1 on Γ_1 as the infimum of $E\{|X - Y|\}$ over all pairs of random variables X and Y with marginal distributions F and G respectively. Let $G^{(m)}$ be the distribution of $(1/m) \sum_{j=1}^m Z_j(G)$. The requisite analog of (2.2) is

$$d_1[G^{(m)}, H^{(m)}] \leqq d_1[G, H]. \tag{2.3}$$

For details on d_1, See Section 8, especially Lemma 8.6. □

The following generalization to higher dimensions may be of some interest. Let $\| \cdot \|$ denote length in R^k.

THEOREM 2.2. *Let $X_1, X_2, \cdots$ be independent, with common distribution in R^k. Suppose $E\{\|X_1\|^2\} < \infty$. Let F_n be the empirical distribution of $X_1, \cdots, X_n$. Given $X_1, \cdots, X_n$, let $X_1^*, \cdots, X_m^*$ be conditionally independent, with common distribution F_n. Along almost all sample sequences, as m and n tend to infinity:*

(*a*) *The conditional distribution of*

$$\sqrt{m}\left(\frac{1}{m}\sum_{j=1}^m X_j^* - \frac{1}{n}\sum_{i=1}^n X_i\right)$$

converges weakly to the k-dimensional normal distribution with mean 0, and variance-covariance matrix equal to the theoretical variance-covariance matrix of X_1.

(*b*) *The empirical variance-covariance matrix of $X_1^*, \cdots, X_m^*$ converges in conditional probability to the theoretical variance-covariance matrix of X_1.*

The requisite metrics are developed in Section 8. If, e.g., $E\{\|X_1\|^4\} < \infty$ then the estimated variance-covariance matrix can be bootstrapped in turn, and so on. We do not pursue this further.

Efron considers the possibility of resampling not from F_n, but from some other estimator, call it $\tilde{F}_n$, of F. The argument for Theorem 2.1 shows that this works too, provided $\tilde{F}_n \Rightarrow F$ in Γ_2, i.e., $\tilde{F}_n$ gets F almost right in the weak topology, and also gets the second moment almost right.

As a lead-in to the treatment of U-statistics in Section 3, fix a function h on $(-\infty, \infty)$ and let Γ_h be the set of distribution functions G satisfying

$$\int h^2(x)\, dG(x) < \infty.$$

Then the estimator $(1/n)\sum_{i=1}^n h(X_i)$ can be bootstrapped, provided the distribution of the X's is in Γ_h. The relevant notion of convergence seems to be this:

$$G_\alpha \Rightarrow G \text{ in } \Gamma_h \quad \text{iff} \quad \int h^2\, dG_\alpha \to \int h^2\, dG, \quad \text{and} \quad \int \theta(h)\, dG_\alpha \to \int \theta(h)\, dG$$

for all bounded continuous functions θ on the line. This just repeats the theorem, in a form more convenient for use in Section 3.

Let $\tilde{F}_n$ be an estimator of F. We continue to assume that $F \in \Gamma_h$. Consider bootstrapping $(1/n)\sum_{i=1}^n h(X_i)$, but resampling from $\tilde{F}_n$ rather than F_n. When will this be asymptotically right? What is needed is the analog of the strong law of large numbers,

$$\int v(x)\, d\tilde{F}_n(x) \to \int v(x)\, dF(x) \quad \text{a.s.} \tag{2.4}$$

whenever $\int |v(x)|\, dF(x) < \infty$. The exceptional null set may depend on v. In particular, suppose $\tilde{F}_n = F_{\hat{\theta}_n}$ where F_θ is some parametric model under consideration and $\hat{\theta}_n(X_1, \cdots, X_n)$ is an estimate of θ. Efron calls this the parametric bootstrap. Then (2.4) holds when $F = F_{\theta_0}$ if $\hat{\theta}_n$ is strongly consistent and the map $\theta \to \int v(x)\, dF_\theta(x)$ is continuous at θ_0 whenever $\int |v(x)|\, dF_{\theta_0}(x) < \infty$.

To close this section, we set our results in the general context introduced by Efron. He considers real valued functions $Z_n(\cdot, \cdot)$ on $Z^n \times \mathscr{F}$ where $\mathscr{F}$ is a set of probability distributions on R containing the "true" F and all distributions with finite support. The bootstrap works if the conditional distribution of $Z_n\{(X_1^*, \cdots, X_n^*), F_n\}$ is close to the distribution of $Z_n\{(X_1, \cdots, X_n), F\}$. We interpret this as follows: If the law of $Z_n\{(X_1, \cdots, X_n), F\}$ tends weakly to a limit as $n \to \infty$, then the conditional distribution of $Z_m \cdot \{(X_1^*, \cdots, X_m^*), F_n\}$ given $(X_1, \cdots, X_n)$ tends weakly to the same limit law with probability one as $m, n \to \infty$. Theorem 2.1 shows this for

$$Z_n\{(X_1, \cdots, X_n), F\} = n^{1/2}\left\{n^{-1}\sum_{i=1}^n X_i - \int x\, dF(x)\right\}.$$

The present notion of convergence is stronger than Efron's, who requires only that the conditional distributions converge weakly to the same limit law in probability. Efron has established convergence in his sense for the mean, when F has finite support.

3. Bootstrapping von Mises functionals. Suppose $X_1, \cdots, X_n$ are independent identically distributed p vectors. Many pivots of interest which have limiting normal distributions can be written in the form

$$\frac{n^{1/2}\{g(S_n/n) - g(\mu)\}}{v(T_n/n)}$$

where $g : R^k \to R$, $v : R^\ell \to R$,

$$S_n = \sum_{i=1}^n h(X_i), \tag{3.1}$$

$$T_n = \sum_{i=1}^n r(X_i), \tag{3.2}$$

$h : R^p \to R^k$, $r : R^p \to R^\ell$, and

$$\mu = Eh(X_1), \qquad \nu = Er(X_1). \tag{3.3}$$

The asymptotic theory for such things is, of course, based on linearization for the numerator

$$n^{1/2}\left\{g\left(\frac{S_n}{n}\right) - g(\mu)\right\} = \dot{g}(\mu)n^{1/2}\left(\frac{S_n}{n} - \mu\right)^T + o_p(1) \tag{3.4}$$

provided that $E \| h(X_1) \|^2 < \infty$, g has a total differential $\dot{g}_{1\times k}$ at μ, and for the denominator that v is continuous at ν in the sense

$$v\left(\frac{T_n}{n}\right) = v(\nu) + o_p(1). \tag{3.5}$$

The bootstrap commutes with smooth functions in exactly the same way. Let

$$\tilde{S}_n = \sum_{i=1}^n h(Y_i^*), \qquad \tilde{T}_n = \sum_{i=1}^n r(Y_i^*).$$

If $E \| h(X_1) \|^2 < \infty$ and $\dot{g}$ exists in a neighborhood of μ and is continuous at μ then,

$$n^{1/2}\left\{g\left(\frac{\tilde{S}_n}{n}\right) - g\left(\frac{S_n}{n}\right)\right\} = \dot{g}(\mu)n^{1/2}\left(\frac{S_n}{n} - \frac{\tilde{S}_n}{n}\right)^T + \Delta_n \tag{3.6}$$

where $\Delta_n \to 0$ in conditional probability and, of course, if v is continuous

$$v\left(\frac{\tilde{T}_n}{n}\right) \to v(\nu) \tag{3.7}$$

in conditional probability. The proof of (3.6) in a more general setting is given in Lemma 8.10 below.

Suppose now that g is a functional $g : \mathscr{F} \to R$ where $\mathscr{F}$ is a convex set of probability measures on R^m including all point masses and F. Suppose also that g is Gâteaux differentiable at F with derivative $\dot{g}(F)$ representable as an integral

$$\dot{g}(F)(G - F) = \frac{\partial}{\partial\epsilon} g(F + \epsilon(G - F))|_{\epsilon=0} = \int \psi(x, F)\, dG(x) \tag{3.8}$$

where necessarily

$$\int \psi(x, F)\, dF(x) = 0. \tag{3.9}$$

Such g are often called von Mises functionals. Asymptotic normality results in nonparametric statistics relate to quantities of the form $n^{1/2}\{g(F_n) - g(F)\}$ or asymptotically equivalent quantities. The result we usually want and get is that $n^{1/2}\{g(F_n) - g(F)\}$ and $n^{1/2} \int \psi(x, F)\, d(F_n - F)$ have the same $N(0, \int \psi^2(x, F)\, dF)$ limit law. As Reeds (1976) indicates, this reflects a general Taylor approximation

$$g(F_n) - g(F) = \dot{g}_F(F_n - F) + \Delta_n(F_n, F) \tag{3.10}$$

where

$$\Delta_n(F_n, F) = o_p(g_F(F_n - F)).$$

It is natural to hope that if we let G_n be the empirical d.f.of $X_1^*, \cdots, X_n^*$, then

$$g(G_n) - g(F_n) = \dot{g}_{F_n}(G_n - F_n) + \Delta_n(G_n, F_n),$$

where for almost all $X_1, X_2, \cdots$

$$n^{1/2}\Delta_n(G_n, F_n) \to 0 \tag{3.11}$$

in conditional probability, and thence that the conditional law of

$$n^{1/2}\dot{g}_{F_n}(G_n - F_n) = n^{-1/2}\sum_{i=1}^n \psi(X_i^*, F_n) \text{ tends to } N\left(0, \int \psi^2(x, F)\, dF(x)\right). \tag{3.12}$$

Simple conditions for the validity of (3.11) can be formulated using the theory of compact differentiation as in Reeds (1976). However, verification of these conditions in particular situations poses the same requirements for special arguments as in Reeds' verification of various examples of (3.10). Moreover, whereas convergence in law under F of $\int \psi(x, F)\, dF_n$ is immediate if $\int \psi^2(x, F)\, dF < \infty$, further continuity conditions on ψ as a function of F seem necessary to ensure that the conditional distributions of $\int \psi(x, F_n)\, dG_n$ tend weakly to $N(0, \int \psi^2(x, F)\, dF(x))$.

The simplest conditions sufficient to guarantee this behavior seem to be

$$\text{i)} \quad \int \psi^2(x, F)\, dF(x) < \infty.$$

$$\text{ii)} \quad \int (\psi(x, F_n) - \psi(x, F))^2\, dF_n \to 0 \text{ a.s.}$$

Condition (ii) implies that for almost all $X_1, X_2, \cdots$,

$$n^{-1/2} \sum_{i=1}^n \left[\psi(X_i^*, F_n) - \left\{ \psi(X_i^*, F) - \int \psi(x, F)\, dF_n \right\} \right] \to 0$$

in conditional probability, while condition (i) ensures the satisfactory behavior of $n^{-1/2}\Sigma\psi(X_i^*, F) - \int \psi(x, F)\, dF_n$. These conditions are exploited in Theorem 3.1 below.

We pursue these general considerations slightly in Section 8. Here we content ourselves with checking the bootstrap for the simplest nonlinear von Mises functionals

$$g(H) = \int\int \omega(x, y)\, dH(x)\, dH(y) \tag{3.13}$$

where $\omega(x, y) = \omega(y, x)$ and H is such that $g(H)$ is well defined. In particular,

$$g(F_n) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \omega(X_i, X_j).$$

A closely related statistic of interest is the U-statistic of order 2 defined by

$$g_n(F_n) = \binom{n}{2}^{-1} \sum_{i<j} \omega(X_i, X_j) = \frac{n}{n-1} g(F_n) - \frac{1}{n(n-1)} \sum_{i=1}^n \omega(X_i, X_i). \tag{3.14}$$

It is well known (von Mises, 1947) that if

$$\int \omega^2(x, y)\, dF(x)\, dF(y) < \infty \tag{3.15}$$

and

$$\int \omega^2(x, x)\, dF(x) < \infty, \tag{3.16}$$

then

$$n^{1/2}\{g(F_n) - g(F)\} \text{ tends weakly to } N(0, \sigma^2) \tag{3.17}$$

where

$$\sigma^2 = 4\left[\int \left\{ \int \omega(x, y)\, dF(y) \right\}^2 dF(x) - g^2(F) \right]. \tag{3.18}$$

This is in accord with (3.8) and (3.10), since in this case

$$\psi(x, F) = 2\left\{ \int \omega(x, y)\, dF(y) - g(F) \right\}. \tag{3.19}$$

THEOREM 3.1 *If* (3.15) *and* (3.16) *hold, and* g *is given by* (3.13) *and* σ^2 *by* (3.18), *then for almost all* $X_1, X_2, \cdots$, *given* $(X_1, \cdots, X_n)$,

$$n^{1/2}\{g(G_n) - g(F_n)\} \quad \text{converges weakly to} \quad N(0, \sigma^2).$$

PROOF. Define ψ and Δ_n as in (3.19) and (3.10). Then we will establish that (3.11) and (3.12) hold.

PROOF OF CLAIM (3.11). $\Delta_n(G_n, F_n) = \int\int \omega(x, y)\, d(G_n - F_n)(x)\, d(G_n - F_n)(y)$. By an inequality of von Mises (1947) (see also Hoeffding, 1948),

$$E\{\Delta_n^2(G_n, F_n) | X_1, \cdots, X_n\} \leqq n^{-2}\left\{C_1 \int\int \omega^2(x, y)\, dF_n\, dF_n + \frac{C_2}{n}\int \omega^2(x, x)\, dF_n\right\}.$$

where C_1 and C_2 are universal constants. Now

$$\int \omega^2(x, x)\, dF_n \to E\omega^2(X_1, X_1)$$

$$\int\int \omega^2(x, y)\, dF_n\, dF_n = \left(\frac{n}{n-1}\right)^2 \binom{n}{2}^{-1} \sum_{i<j} \omega^2(X_i, X_j) + n^{-2}\sum_i \omega^2(X_i, X_i) \to E\omega^2(X_1, X_2)$$

almost surely by the strong law of large numbers, as generalized to U-statistics (see Berk, 1966, page 56) and (3.11) follows.

PROOF OF CLAIM (3.12). As we noted earlier, it is enough to show that

$$\int \{\psi(x, F_n) - \psi(x, F)\}^2\, dF_n \to 0$$

with probability 1. But,

$$\begin{aligned}\int \{\psi(x, F_n) - \psi(x, F)\}^2\, dF_n(x) &= n^{-1}\sum_i \{\psi(X_i, F_n) - \psi(X_i, F)\}^2 \\ &= n^{-1}\sum_i \left\{n^{-1}\sum_j \omega(X_i, X_j) - \int \omega(X_i, y)\, dF(y)\right\}^2 \\ &= n^{-3}\sum_{i,j,k} \omega(X_i, X_j)\omega(X_i, X_k) \\ &\quad - 2n^{-2}\sum_{i,j} \omega(X_i, X_j)\int \omega(X_i, y)\, dF \\ &\quad + n^{-1}\sum_i \left\{\int \omega(X_i, y)\, dF\right\}^2.\end{aligned}$$

By an argument using a strong law of large numbers for U-statistics, these last three terms tend with probability 1 to

$$E\omega(X_1, X_2)\omega(X_1, X_3),\ -2E[\omega(X_1, X_2)E\{\omega(X_1, X_2)|X_2\}],\text{ and } E[E^2\{\omega(X_1, X_2)|X_2\}],$$

respectively. The sum of these numbers is 0 and claim (3.12) and the theorem follow. □

If $E\omega^2(X_1, X_2) < \infty$ and $E\omega^2(X_1, X_1) < \infty$, the conclusion of Theorem 3.1 clearly holds for the bootstrap distribution of the U-statistic $g_n(F_n)$ and, more generally, any convex combination of $g_n(F_n)$ and $n^{-1}\sum \omega(X_i, X_i)$ where the weight on $g_n(F_n)$ tends to 1. Failure of the conditions, however, can cause failure of the bootstrap (see Section 6).

As an example of the applicability of this result, it is valid to bootstrap the distribution of Wilcoxon's one sample statistic

$$\left\{\frac{n^{1/2}(n+1)}{2}\right\}^{-1} \sum_{i \leqq j} \{I(X_i + X_j > 0) - P(X_i + X_j > 0)\}$$

in order, for instance, to obtain approximations to its power.

Extensions of the theorem to the von Mises statistics corresponding to U-statistics of arbitrary order, vector U-statistics, U-statistics based on several samples, etc., is straightforward, provided, however, that the hypotheses appropriate to the von Mises statistics, as in Fillipova (1962), are kept.

Extending a remark made in Section 2, we can bootstrap U-statistics by resampling from a general $\{\tilde{F}_n\}$, provided that $\{\tilde{F}_n\}$ possesses a property analogous to the strong law of large numbers for U-statistics, viz.,

$$\int \cdots \int v(x_1, \cdots, x_k)\, dF_n(x_1) \ldots dF_n(x_k) \to \int \cdots \int v(x_1, \cdots, x_k)\, dF(x_1) \ldots dF(x_k) \text{ a.s.}$$

if $$\int |v(x_1, \cdots, x_k)|\, dF(x_1) \ldots dF(x_k) < \infty.$$

4. Bootstrapping the empirical process. The object of this section is to bootstrap the empirical process, (Theorem 4.1), and to obtain a fixed-width confidence band for the population distribution function which is valid even when the latter has a discrete component (Corollary 4.2). We first give two preliminary lemmas and then recall notions of weak convergence. Throughout this section, B is a Brownian bridge on $[0, 1]$. Theorem 3 of Komlos, Major and Tusnady (1975) implies the following result.

LEMMA 4.1 *There exist, on a sufficiently rich probability space, independent random variables $U_1, U_2, \ldots$ with common distribution uniform on $[0, 1]$, and a Brownian bridge B on $[0, 1]$ with the following property. Let H_m be the empirical distribution function of $U_1, \cdots, U_m$* and let

$$B_m(u) = m^{1/2}\{H_m(u) - u\} \quad \text{for } 0 \leqq u \leqq 1.$$

Then for some constant K_1, and $\epsilon_m = (\log m)/m^{1/2}$

$$P\{\|B_m - B\| \geqq K_1\epsilon_m\} \leqq K_1\epsilon_m.$$

To state the next result, which is an integrated form of Levy's modulus of continuity, let

$$\omega(\delta, f) = \sup\{|f(s) - f(t)| : |t - s| \leqq \delta\} \tag{4.1}$$

$$h(\delta) = \left(\delta \log \frac{1}{\delta}\right)^{1/2} \quad \text{for } 0 \leqq \delta \leqq \tfrac{1}{2}$$
$$= h(1/2) \quad \text{for } \delta \geqq \tfrac{1}{2} \tag{4.2}$$

LEMMA 4.2 *There is a constant K_2 such that $E\{\omega(\delta, B)\} \leqq K_2 h(\delta)$ for $0 < \delta \leqq \frac{1}{2}$.*

PROOF. Represent B as

$$B(u) = W(u) - uW(1) \quad \text{for} \quad 0 \leqq u \leqq 1,$$

where W is a Wiener process on $[0, \infty)$. Now

$$\omega(\delta, B) \leqq \omega(\delta, W) + \delta |W(1)|.$$

So it is enough to prove the lemma with W in place of B. Abbreviate

$$M_{k\delta} = \sup_s \{|W(s) - W(k\delta)| : k\delta \leqq s \leqq (k + 1)\delta\}.$$

Let K be the integer part of $1/\delta$. By the triangle inequality,

$$\omega(\delta, W) \leqq 3 \max_k \{M_{k\delta} : 0 \leqq k \leqq K\}.$$

Of course, the $M_{k\delta}$ are independent and identically distributed, so

$$E\{\omega(\delta, W)\} = \int_0^\infty P\{\omega(\delta, W) > x\}\, dx \leqq 3 \int_0^\infty [1 - \{1 - P(M_{o\delta} > x)\}^{K+1}]\, dx.$$

If $x < 2^{1/2}h(\delta)$, the integrand may be replaced by the trivial upper bound of 1. The integral over bigger x's is negligible for small δ; this may be seen by estimating the integrand as follows:

$$1 - (1-p)^{K+1} \leqq (K+1)p \qquad \text{for } 0 \leqq p \leqq 1$$

$$P\{M_{o\delta} > x\} \leqq 4(\delta/2\pi)^{1/2}x^{-1}e^{-x^2/2\delta}$$

and then making the change of variables $y = \delta^{-1/2}x$. □

Let D be the space of all real-valued functions f on $[-\infty, \infty]$, such that f vanishes continuously at $\pm\infty$, and is right continuous with left limits on $(-\infty, \infty)$. Give D the Skorokhod topology. Let Γ be the set of all distribution functions, in the sup norm. For $G \in \Gamma$, let $Z_1(G), \cdots, Z_m(G)$ be independent with common distribution G. Let G_m be the empirical distribution of $Z_1(G), \cdots, Z_m(G)$, and set

$$W_{Gm}(t) = \sqrt{m}[G_m(t) - G(t)] \qquad \text{for } -\infty < t < \infty, \tag{4.3}$$

extended to vanish at $\pm\infty$. Let $\psi_m(G)$ be the distribution of the process W_{Gm}. Thus, $\psi_m(G)$ is a probability measure on D. In this notation, the usual invariance principle states that $\psi_m(G)$ tends weakly to the law of $B(G)$ as $m \to \infty$, where B is the Brownian bridge, and $B(G)(t, \omega) = B\{G(t), \omega\}$.

The weak topology on the space of probability measures on D is metrized by a dual Lipschitz metric as follows. Let γ metrize the Skorokhod topology on D, and in addition satisfy

$$\gamma(f, g) \leqq \|f - g\| \wedge 1. \tag{4.4}$$

Here f and g are elements of D, i.e., function on $[-\infty, \infty]$, and $\|\cdot\|$ is the sup norm. Now

$$\rho(\pi, \pi') = \sup_\theta \left| \int_D \theta\tau d\pi - \int_D \theta\tau d\pi' \right| \tag{4.5}$$

where π and π' are probability measures on D, and θ runs through the functions on D which are uniformly bounded by 1 and satisfy the Lipschitz condition

$$|\theta(f) - \theta(g)| \leqq \gamma(f, g).$$

PROPOSITION 4.1. *There exists a universal constant C such that*

$$\rho[\psi_m(F), \psi_m(G)] \leqq C[\epsilon_m + h(\|F - G\|)],$$

where $\epsilon_m = m^{-1/2} \log m$ and h was defined in (4.2).

PROOF. Recall B_m from Lemma 4.1 Clearly, $\psi_m(F)$ and $\psi_m(G)$ are the probability distributions induced on D by $B_m(F)$ and $B_m(G)$ respectively. By the definition (4.5) of the dual Lipschitz metric ρ,

$$\rho[\psi_m(F), \psi_m(G)] \leqq \sup_\theta E\{|\theta[B_m(F)] - \theta[B_m(G)]|\} \leqq E\{\gamma[B_m(F), B_m(G)]\}.$$

Now (4.4) implies

$$E\{\gamma[B_m(F), B_m(G)]\} \leqq E\{\|B_m(F) - B_m(G)\| \wedge 1\} \tag{4.6}$$

Since $\|f - g\| \wedge 1$ is a metric, the triangle inequality implies

$$E\{\gamma[B_m(F), B_m(G])\} \leqq 2E\{\|B_m - B\| \wedge 1\} + E\{\omega(\|F - G\|, B)\}. \tag{4.7}$$

Now use Lemma 4.1 to estimate the first term on the right in (4.7):

$$E\{\|B_m - B\| \wedge 1\} \leqq K_1\epsilon_m + P\{\|B_m - B\| > K_1\epsilon_m\} \leqq 2K_1\epsilon_m.$$

The second term on the right in (4.7) can be estimated by Lemma 4.2. □

Return now to the setting of Section 2, but with no moment condition. There is a sample of size n from an unknown distribution function F, which is to be estimated by the empirical distribution function F_n. Given $X_1, \cdots, X_n$, let $X_1^*, \cdots, X_m^*$ be conditionally independent, with common distribution F_n. Let F_{nm} be the empirical distribution function of $X_1^*, \cdots, X_m^*$. And let

$$W_{nm}(t) = \sqrt{m}\{F_{nm}(t) - F_n(t)\} \qquad \text{for } -\infty < t < \infty, \tag{4.8}$$

extended to vanish at $\pm\infty$. The next result is the bootstrap analog of the invariance principle, which states that $\sqrt{n}\,(F_n - F)$ converges weakly to $B(F)$ as $n \to \infty$. No conditions are imposed on F; as usual, B is the Brownian bridge on [0, 1].

THEOREM 4.1. *Along almost all sample sequences, given* $(X_1, \cdots, X_n)$, *as* n *and* m *tend to infinity,* W_{nm} *converges weakly to* $B(F)$.

PROOF. This is almost immediate from Proposition 4.1. Conditionally, $W_{nm} = W_{F_n m}$ has the law $\psi_m(F_n)$, and $\|F_n - F\| \to 0$ a.s. by the Glivenko-Cantelli lemma, so $\psi_m(F_n)$, is nearly $\psi_m(F)$. The latter is almost the law of $B(F)$ by the ordinary invariance principle. Indeed, the argument shows that the ρ-distance between $\psi_m(F_n)$ and the law of $B(F)$ is at most a universal constant times $\epsilon_m + h(\|F_n - F\|)$. □

COROLLARY 4.1. *For almost all* $X_1, X_2, \cdots$, *given* $(X_1, \cdots, X_n)$, *as* n *and* m *tend to infinity,* $\|F_{nm} - F\|$ *tends to* 0 *in probability. Here,* F_{nm} *is the empirical distribution of the resampled data, as defined above.*

We now consider confidence bands for F which will be valid even when F has a discrete component.

COROLLARY 4.2. *Suppose* F *is nondegenerate. Fix* α *with* $0 < \alpha < 1$. *Choose* $c(F_n)$ *from the bootstrap distribution so that*

$$P\{n^{1/2} \sup_x |F_{nn}(x) - F_n(x)| \leqq c_n(F_n) \,|\, X_1, \cdots, X_n\} \to 1 - \alpha.$$

Then

$$P\{n^{1/2} \sup_x |F_n(x) - F(x)| \leqq c_n(F_n)\} \to 1 - \alpha.$$

PROOF. Indeed, $c_n(F_n)$ must converge to the $(1-\alpha)$-point of the law of $\sup_x |B(F(x))|$, which is continuous: see Lemma 8.11 below. So, $F_n \pm c_n(F_n)$ is the desired band.

Preliminary calculations suggest that the mapping $F \to \psi_m(F)$ is uniformly equicontinuous, in the sense that there is a function $q(t) \to 0$ as $t \to 0$, and for all m, F and G:

$$\rho[\psi_m(F), \psi_m(G)] \leqq q(\|F - G\|).$$

The argument rests on the following inequality, which may be of independent interest. Suppose F and G concentrate on [0, 1] and $\|F - G\| < \delta$. Then

$$\text{Lebesgue measure of } \{t{:}0 \leqq t \leqq 1 \quad \text{and} \quad |F^{-1}(t) - G^{-1}(t)| > \sqrt{\delta}\} < \sqrt{\delta}.$$

This is immediate from Chebychev's inequality; see (8.1).

Suppose the resampling is from another estimator $\tilde{F}_n$ for F. Bootstrapping may still be valid. Given $(X_1, \cdots, X_n)$, it can be shown that $W_{\tilde{F}_n m}$ tends weakly to $B(F)$ as m and n tend to ∞, provided $\tilde{F}_n \to F$ a.s. in the sup norm. Here $W_{\tilde{F}_n m}$ was defined in (4.3). This result can even be proven under the weaker hypothesis, that $\tilde{F}_n \to F$ a.s. in the Skorokhod topology.

5. The quantile process. Another interesting process in terms of which various statistics and pivots can be defined naturally is the quantile process Q_n which we define on (0, 1) by

$$Q_n(t) = n^{1/2}\{F_n^{-1}(t) - F^{-1}(t)\}$$

where the inverse of a distribution function H is given, in general, by

$$H^{-1}(t) = \inf\{x : H(x) \geqq t\}.$$

Our aim in this section is to justify the bootstrapping of this process. Applications which will be sketched briefly after the theorem include confidence intervals for the median and pivots based on trimmed means and Winsorized variances.

For convenience, throughout this section we use $\circ$ to denote composition. For example, $f \circ F^{-1}$ means $f(F^{-1})$.

It is well known (see Bickel, 1966, for example) that given $0 < t_0 \leqq t_1 < 1$, if

(5.1) F has continuous positive density f on R,

then

(5.2) Q_n tends weakly to $B/f \circ F^{-1}$ in the space of probability measures on $D[t_0, t_1]$.

Write G_n for F_{nn} as defined for (4.8) and let

$$Q_n = n^{1/2}(G_n^{-1} - F_n^{-1}).$$

THEOREM 5.1. *If* (5.1) *holds, then along almost all sample sequences* $X_1, X_2, \cdots$, *given* $(X_1, \cdots, X_n)$, Q_n *converges weakly to* $B/(f \circ F^{-1})$ *in the sense of weak convergence for probability measures on* $D[t_o, t_1]$.

PROOF. An equicontinuity argument does not work here since the behavior of the quantile process depends on the density of the limit distribution. This is also the reason we take $m = n$. We present a relatively ad hoc modification of an argument due to Pyke and Shorack (1968).

It is convenient to denote the sup norm in $D[t_o, t_1]$ by $\|\cdot\|$. Write

$$Q_n = n^{1/2}\frac{(F \circ G_n^{-1} - F \circ F_n^{-1})}{R_n},$$

where

$$R_n = \frac{F \circ G_n^{-1} - F \circ F_n^{-1}}{G_n^{-1} - F_n^{-1}}.$$

Continue by writing

$$\begin{aligned} n^{1/2}(F \circ G_n^{-1} - F \circ F_n^{-1}) = n^{1/2}[\{(F_n \circ G_n^{-1} - F \circ G_n^{-1}) - (F_n \circ F_n^{-1} - F \circ F_n^{-1})\} \\ + \{G_n \circ G_n^{-1} - F_n \circ G_n^{-1}\}] \\ - n^{1/2}(F_n \circ F_n^{-1} - G_n \circ G_n^{-1}). \end{aligned} \tag{5.3}$$

Let the probability space be rich enough to support the processes B_n and B of Lemma 4.1 as well as another pair $(\tilde{B}_n, \tilde{B})$ with the same distribution as (B_n, B) and independent of them.

We now represent $n^{1/2}(G_n - F_n)$ as $\tilde{B}_n \circ F_n$ and $n^{1/2}(F_n - F)$ as $B_n \circ F$ and call these processes $\tilde{W}_n$ and W_n respectively. Then we can write the right-hand side of (5.3) as

$$-\{(W_n \circ G_n^{-1} - W_n \circ F_n^{-1}) + \tilde{W}_n \circ G_n^{-1}\} - n^{1/2}\{(F_n \circ F_n^{-1} - I) - (G_n \circ G_n^{-1} - I)\}$$

where I is the identify. Therefore, to prove the theorem it is enough to show that the following five assertions, (5.4)–(5.8), hold for almost all $X_1, X_2, \cdots$.

$$\|F_n \circ F_n^{-1} - I\| = o(n^{-1/2}), \tag{5.4}$$

$$n^{1/2}\|G_n \circ G_n^{-1} - I\| \to 0 \tag{5.5}$$

in (conditional) probability,

$$\|R_n - f\circ F^{-1}\| \to 0 \tag{5.6}$$

in (conditional) probability,

$$-\tilde{W}_n\circ G_n^{-1} \text{ converges weakly to } B, \text{ on } [t_o, t_1] \tag{5.7}$$

$$\|W_n\circ G_n^{-1} - w_n\circ F_n^{-1}\| \to 0 \tag{5.8}$$

in (conditional) probability.

PROOF OF (5.4). F_n has jumps of size $1/n$ only.

PROOF OF (5.5). Bound (5.5) by

$$n^{1/2}\sup_x\{G_n(x+0) - G_n(x)\} \leqq \sup_x|\tilde{W}_n(x+0) - \tilde{W}_n(x)| + n^{-1/2}.$$

Since F is continuous and strictly increasing, so is F^{-1} and

$$\sup_x|\tilde{W}_n(x+0) - W_n(x)| = \sup|\tilde{W}_n\circ F^{-1}(x+0) - \tilde{W}_n\circ F^{-1}(x)|. \tag{5.9}$$

By Theorem 4.1, given $(X_1, \cdots, X_n)$, $\tilde{W}_n\circ F^{-1}$ converge weakly to B which is continuous. Therefore, the expression in (5.9) tends to 0 in conditional probability and (5.5) follows.

PROOF OF (5.6). By Corollary 4.1 since, by hypothesis, F^{-1} is continuous on (0, 1),

$$\|G_n^{-1} - F^{-1}\| \to 0 \tag{5.10}$$

in conditional probability, for almost all $X_1, X_2, \cdots$. Similarly, by the Glivenko-Cantelli Theorem, with probability 1,

$$\|F_n^{-1} - F^{-1}\| \to 0.$$

Claim (5.6) follows since the assumed continuity of F on R implies that F is uniformly differentiable on all compact subsets of R.

PROOF OF (5.7). By (5.10) and Theorem 4.1, given $(X_1, \cdots, X_n)$, the processes $(-\tilde{W}_n\circ F^{-1}, F\circ G_n^{-1})$ viewed as probability measures on $D[t_o, t_1] \times D[t_o, t_1]$ converge weakly to (B, I). By the continuity of the composition map $M:(f, g) \to f\circ g$ at all points of $C[0, 1] \times D[t_o, t_1]$, we have $-\tilde{W}_n\circ G_n^{-1}$ converging weakly to B and (5.7) is proven.

PROOF OF (5.8). We have to be careful here to control W_n with probability 1. Since $\|F\circ F_n^{-1} - F\circ G_n^{-1}\| \to 0$ in conditional probability and $W_n = B_n\circ F$, it is enough to check that if $\delta_n \to 0$,

$$\omega(\delta_n, B_n) \to 0 \text{ a.s.}$$

But this follows for instance from Komlos, Major and Tusnady (1975, Theorem 3). The theorem is proved.

REMARKS. (1) If $F^{-1}(0+) > -\infty$ and $F^{-1}(1) < \infty$ and f is continuous on $[F^{-1}(0+), F^{-1}(1)]$, the conclusion of the theorem holds in $D[F^{-1}(0+), F^{-1}(1)]$. For instance, if F is uniform on (0, 1), convergence holds in $D[0, 1]$. More generally, we may have one end of the support finite and the other infinite and have the appropriate theorem hold.

(2) Suppose $\{\tilde{F}_n\}$ is a general sequence of probability measures depending on $X_1, \cdots, X_n$ and G_n is the empirical d.f. of $Y_1, \cdots, Y_n$ which, given $(X_1, \cdots, X_n)$, are i.i.d. with common distribution F_n. We can give simple conditions for $\sqrt{n}\,(G_n^{-1} - \tilde{F}_n^{-1})$ to converge weakly, given $(X_1, \cdots, X_n)$ (as probability measures on $D([t_o, t_1])$ to $B/(f\circ F^{-1})$, provided

that we require the convergence to hold in probability as in Efron. All we need in addition to (5.1) is that (i) $n^{1/2}(\tilde{F}_n - F)$ converge weakly (as probability measures on D) to a limit with continuous sample functions, and (ii) $\sup_x |\tilde{F}_n(x+0) - \tilde{F}_n(x)| = o_p(n^{-1/2})$. Hence the parametric bookstrap works if, for example, $F = F_{\theta_0}$ satisfies (5.1) and $(\partial/\partial\theta) F_\theta|_{\theta_0}$ is continuous in x and $n^{1/2}(\hat{\theta}_n - \theta_0) = O_p(1)$.

Here are some applications which follow fairly easily from the theorem.

The median. Let m^* be the median of the X_i^* and m the median of the X_i.

PROPOSITION 5.1. *If F has a unique median μ and f has a positive derivative f continous in a neighborhood of μ, then along almost all sample sequences $X_1, X_2, \cdots$, given $(X_1, \cdots, X_n)$, $n^{1/2}(m^* - m)$ converges weakly to $N\left(0, \frac{1}{4f^2(\mu)}\right)$, the limit law of $n^{1/2}(m - \mu)$.*

By this result the quantiles of the bootstrap distribution of $n^{1/2}(m^* - m)$ can be used to set an approximate confidence interval for μ. An asymptotic pivot in which we estimate the density f and then scale can also be bootstrapped.

A more careful argument shows that Proposition 5.1 holds under the weakest natural conditions: μ is unique and F has positive derivative f at μ.

Quantile intervals. The usual interval for the population median is $[X_{(k)}, X_{(n-k+1)}]$ where $X_{(1)} < \cdots < X_{(n)}$ are the order statistics of the sample, and k is determined by the desired confidence coefficient through the relation

$$P\{X_{(j)} < \mu \leqq X_{(j+1)}\} = \binom{n}{j} 2^{-n}$$

valid for all continuous F.

Since $X_{(j)} = F_n^{-1}(j/k)$ is the j/k quantile of the law of X_1^*, given $(X_1, \cdots, X_n)$, the bootstrap principle leads us to believe

$$P\{X_{(k)} < M \leqq X_{(\ell)} \mid F_n\} \approx P\left\{F^{-1}\left(\frac{k}{n}\right) < m \leqq F^{-1}\left(\frac{\ell}{n}\right)\right\} \tag{5.11}$$

where $P(\cdot \mid F_n)$ is the conditional probability, given $(X_1, \cdots, X_n)$. Efron, by exact calculation, gets the unexpected approximation

$$P\{X_{(k)} < M \leqq X_{(\ell)} \mid F_n\} \approx P\{X_{(k)} < \mu \leqq X_{(\ell)}\}. \tag{5.12}$$

If we interpret $\approx$ as meaning that the difference of the two sides goes to 0 along almost all sample sequences, then both (5.11) and (5.12) can be established under the assumptions of Theorem 5.1.

Linear combinations of order statistics. Theorem 5.1 establishes the validity of the bootstrap for linear combinations of order statistics with nice weight functions concentrated on $[\alpha, 1-\alpha]$, $0 < \alpha < \frac{1}{2}$. That is,

$$n^{1/2}\left\{\int_\alpha^{1-\alpha} F_n^{-1}(t)\, d\Lambda_n(t) - \int_\alpha^{1-\alpha} F^{-1}(t)\, d\Lambda_n(t)\right\}$$

can be bootstrapped under condition (5.1) provided that $\Lambda_n \to \Lambda$ weakly. As a special case, if we take Λ_n to be the uniform distribution on $[\alpha, 1-\alpha]$, we see that the bootstrap provides confidence intervals for the center of symmetry of a symmetric distribution based on the α-trimmed mean. The bootstrap is also valid for estimates of the asymptotic variance of such linear combinations of order statistics and for pivots based on t-like statistics.

6. Counter-examples. In Sections 2 and 3 we checked the validity of the bootstrap for various functionals $R_n\{(X_1, \cdots, X_n); F_n\}$. Roughly, the bootstrap will work provided that

(6.1a) $R_n\{(Y_1, \cdots, Y_n); G\}$ tends weakly to a limit law $\mathscr{L}_G$ whenever $Y_1, \cdots, Y_n$ are i.i.d. with distribution G, for all G in a "neighborhood" of F into which F_n falls eventually with probability 1,

(6.1b) the convergence in (6.1a) is uniform on the neighborhood,

and

(6.1c) the function $G \to \mathscr{L}_G$ is continuous.

In the examples of this section, the bootstrap fails because uniformity does not hold on any usable neighborhoods.

Counter-example 1: a U-statistic. Let

$$R_n(Y_1, \cdots, Y_n; G) = n^{1/2}\left\{\binom{n}{2}^{-1} \sum_{i<j} [\omega(Y_i, Y_j) - \int \omega(x, y)\, dG(x)\, dG(y)]\right\} \tag{6.2}$$

a normalized centered U-statistic. As we have noted in the previous section, by a theorem of Hoeffding, if

$$\int \omega^2(x, y)\, dF(x)\, dF(y) < \infty, \tag{6.3}$$

then

(6.4) $R_n(X_1, \cdots, X_n; F)$ converges weakly to a $N(0, \sigma^2)$ random variable,

where σ^2 is given by (3.18).

To bootstrap the U-statistic, however, we have to assume not only (6.3) but also the von Mises condition

$$\int \omega(x, x)^2\, dF(x) < \infty \tag{6.5}$$

Absent this condition, the bootstrap can fail: indeed, $|R(X_1^*, \cdots, X_n^*; F_n)|$ can tend to ∞.

Suppose F is the uniform distribution on (0, 1) and write $\omega = \omega_1 + \omega_2$ where $\omega_1(x, y) = \omega(x, y)I(x \neq y)$. Let R_{n1}, R_{n2} be the U-statistics corresponding to ω_1, ω_2 respectively. Then $R_n = R_{n1} + R_{n2}$. If (6.3) holds, by Theorem 3.1, given $(X_1, \cdots, X_n)$, the conditional distribution of $R_{n1}(X_1^*, \cdots, X_n^*; F_n)$ tends weakly to $N(0, \sigma^2)$. An example will be given where $|R_{n2}(X_1^*, \cdots, X_n^*; F_n)|$ tends to ∞ in probability. Of course, $R_{n2}(X_1, \cdots, X_n; F) = 0$.

To develop this example, write

$$R_{n2}(X_1^*, \cdots, X_n^*; F_n) = \{n^{1/2}(n-1)\}^{-1} \sum_{i=1}^n \omega(X_i, X_i)\left\{\nu_{in}(\nu_{in} - 1) - \frac{n-1}{n}\right\}, \tag{6.6}$$

where

(6.7) ν_{in} is the number of j's with $1 \leqq j \leqq n$ and $X_j^* = X_i$.

Let $Z_i = \omega(X_i, X_i)$, $i = 1, \cdots, n$ and $Z_{(1)} \leqq \cdots \leqq Z_{(n)}$ be the corresponding order statistics. Take

$$\omega(x, x) = e^{1/x}.$$

We claim

(6.8) the conditional distribution of $\{n^{1/2}(n - 1/Z_{(n)}\}R_{n2}(X_1^*, \cdots, X_n^*; F_n)$ converges in probability to a limit law, namely the distribution of $\nu(\nu - 1) - 1$ where ν is a Poisson variable with mean 1.

Moreover

(6.9) $n^A/Z_{(n)}$ tends to 0 in probability as $n \to \infty$, for every positive A.
So R_{n2} does indeed dominate R_{n1}.

Our assertions about the behavior of R_n are proved as follows. Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics of $X_1, \cdots, X_n$. Then the distribution of

$$n^{-1}(\log Z_{(n)} - \log Z_{(n-1)}) = \frac{n(X_{(2)} - X_{(1)})}{(n^2 X_{(1)} X_{(2)})}$$

converges to a limit concentrating on $(0, \infty)$, since $nX_{(1)}$ and $n(X_{(2)} - X_{(1)})$ converge jointly in law to two independent exponentials. Therefore,

(6.10) $n^A Z_{(n-1)}/Z_{(n)}$ tends to 0 in probability, for any positive A.
Let I be the "antirank" of $Z_{(n)}$, defined by $Z_I = Z_{(n)}$. Then,

$$n^{1/2}(n-1)R_{n2}(X_1^*, \cdots, X_n^*; F_n)/Z_{(n)} = \nu_{In}(\nu_{In} - 1) + O_p\{n^2 Z_{(n-1)}/Z_{(n)}\},$$

since $\sum \nu_{in}(\nu_{in} - 1) \leqq n(n-1)$.

Now (6.8) follows: given $X_1, \cdots, X_n$, conditionally ν_{In} has a binomial distribution with n trials and success probability $1/n$, whose limit is Poisson with mean 1. The remainder is negligible, by (6.10).

The claim (6.9) follows by a previous argument, since $n^{-1} \log Z_{(n)} = (nU_{(1)})^{-1}$ converges in law.

Counter-example 2: the maximum and spacings. If F is uniform on $(0, \theta)$, the usual pivot for θ is $n(\theta - X_{(n)})/\theta$ which has a limiting standard exponential distribution. If we think of θ as the upper end point of the support of F then it is natural to bootstrap $(n(\theta - X_{(n)})/\theta$ by $n(X_{(n)} - X^*_{(n)})$, where $X^*_{(1)} \leqq \cdots \leqq X^*_{(n)}$ are the ordered X_i^*. This does not work. In fact,

$$P\{n(X_{(n)} - X^*_{(n)}) = 0 \mid F_n\} \to 1 - e^{-1} \doteq 0.63.$$

More generally, it is easy to see that for almost all $X_1, X_2, \cdots$,

$$P\{X^*_{(n)} < X_{(n-k+1)} \mid F_n\} \to e^{-k}, \qquad k = 1, \cdots.$$

Thus, with probability 1, the conditional distribution of $n(X_{(n)} - X^*_{(n)})/X_{(n)}$ does not have a weak limit: since $\limsup n(X_{(n)} - X_{(n-k+1)}) = \infty$, and $\liminf n(X_{(n)} - X_{(n-k+1)}) = 0$, a.s. for each k.

This unpleasant behavior cannot be mended by simple smoothing, e.g., replacing F_n by $\tilde{F}_n$ which puts mass $1/(n-1)$ uniformly into each interval $[X_{(n-k+1)}, X_{(n-k)}]$, for $k = 0, \cdots, n-2$. Nor does this behavior have much to do with the maximum. The conditional distributions of the spacings $n(X^*_{(k)} - X^*_{(k-1)})$ do not have weak limits, even though $n(X_{(k)} - X_{(k-1)})$ has an exponential limit.

The problem is the lack of uniformity in the convergence of F_n to F. Uniformity does hold for the parametric bootstrap, where F is estimated by $\hat{F}_n$, which is uniform on the interval $(0, X_{(n)})$. If $X_1^*, \cdots, X_n^*$ are a sample from $\hat{F}_n$, then

$$\mathcal{L}(X_1^*/X_{(n)}, \cdots, X_n^*/X_{(n)}) = \mathcal{L}(X_1/\theta, \cdots, X_n/\theta)$$

7. Other work. Freedman (1981) has pursued the use of the bootstrap for least squares estimates in regression models when the number of parameters is fixed, and arrived at results very similar to those obtained for means in the one-sample problem. Work is in progress at Berkeley on the behavior of other types of estimates in these models, as well as on the general theory of bootstrapping von Mises functionals in one-sample models.

The authors are studying the behavior of the bootstrap in regression models when the number of parameters is large as well as the sample size; also considered is the sampling of finite populations. An interesting new phenomenon surfaces: the bootstrap can work for

linear statistics based on large numbers of summands even though the normal approximation does not hold. On the other hand, the bootstrap fails quite generally when the number of parameters is too large.

8. Mathematical appendix. In Section 2, we used the Mallows metric d_2 and its cousin d_1. It may be helpful to give a fuller account of such metrics here. Let B be a separable Banach space with norm $\|\cdot\|$. The only present case of interest is finite-dimensional Euclidean space, in the Euclidean norm. Let $1 \leqq p < \infty$; only $p = 1$ or 2 are of present interest.[3]

Let $\Gamma_p = \Gamma_p(B)$ be the set of probabilities γ on the Borel σ-field of B, such that $\int \|x\|^p \gamma(dx) < \infty$. For α and β in Γ_p, let $d_p(\alpha, \beta)$ be the infimum of $E\{\|X - Y\|^p\}^{1/p}$ over pairs of B-valued random variables X and Y, where X has law α and Y has law β.

LEMMA 8.1. (a) *The infimum is attained.*
(b) *d_p is a metric on Γ_p.*

PROOF: *Claim* (*a*). Let X and Y be the coordinate functions on $B \times B$. Using weak compactness, it is easy to find a probability π on $B \times B$, such that $\pi X^{-1} = \alpha$, and $\pi Y^{-1} = \beta$, and $\int \|X - Y\|^p \, d\pi$ is minimal.

Claim (*b*). Only the triangle inequality presents any problem. Fix α, β and γ in Γ_p. Using the first claim, choose π on $B \times B$ so $[\int \|X - Y\|^p \, d\pi]^{1/p} = d_p(\alpha, \beta)$. Changing notation slightly, let Y and Z be the coordinates on another "plane" $B \times B$; find π' on this $B \times B$ so $[\int \|Y - Z\|^p \, d\pi']^{1/p} = d_p(\beta, \gamma)$. Now stitch the two planes together along the Y-axis into a 3-space $B \times B \times B$. More formally, let X, Y, Z be the coordinate functions on $B \times B \times B$. Define π^* on $B \times B \times B$ by the requirements:

- the π^*-law of Y is β;
- given Y, the variables X and Z are conditionally π^*-independent;
- the conditional π^*-law of X given $Y = y$ coincides with the conditional π-law of X given $Y = y$;
- the conditional π^*-law of Z given $Y = y$ coincides with the conditional π'-law of Z given $Y = y$.

In particular, the π^*-law of (X, Y) is π; the π^*-law of (Y, Z) *is* π'.

Minkowski's inequality can now be used, as follows:

$$
\begin{aligned}
d_p(\alpha, \gamma) &\leqq \left\{ \int \|X - Z\|^p \, d\pi^* \right\}^{1/p} \\
&\leqq \left\{ \int [\|X - Y\| + \|Y - Z\|]^p \, d\pi^* \right\}^{1/p} \\
&\leqq \left\{ \int \|X - Y\|^p \, d\pi^* \right\}^{1/p} + \left\{ \int \|Y - Z\|^p \, d\pi^* \right\}^{1/p} \\
&= \left\{ \int \|X - Y\|^p \, d\pi \right\}^{1/p} + \left\{ \int \|Y - Z\|^p \, d\pi' \right\}^{1/p} \\
&= d_p(\alpha, \beta) + d_p(\beta, \gamma)
\end{aligned}
$$

□

On the real line, Lemma 8.2 below gives a very convenient representation for d_p (see Major, 1978). In this case, the probabilities α and β are defined by their distribution functions F and G.

[3] The essential supremum corresponds to $p = \infty$ and can be handled analogously. The extension to Orlicz spaces might be useful: see Zaanen (1953) or Zygmund (1935).

LEMMA 8.2. *If B is the real line, with $\|x\| = |x|$, then*

$$d_p(F, G) = \left\{ \int_0^1 |F^{-1}(t) - G^{-1}(t)|^p \, dt \right\}^{1/p}$$

The case $p = 1$ is especially simple because

$$\int_0^1 |F^{-1}(t) - G^{-t}(t)| \, dt = \int_{-\infty}^{\infty} |F(t) - G(t)| \, dt. \tag{8.1}$$

Indeed, both sides of (8.1) represent the area between the graphs of F and G.

Return now to the general setting.

LEMMA 8.3. *Let $\alpha_n, \alpha \in \Gamma_p$. Then $d_p(\alpha_n, \alpha) \to 0$ as $n \to \infty$ is equivalent to each of the following.*

a) *$\alpha_n \to \alpha$ weakly and $\int \|x\|^p \alpha_n(dx) \to \int \|x\|^p \alpha(dx)$.*
b) *$\alpha_n \to \alpha$ weakly and $\|x\|^p$ is uniformly α_n-integrable.*
c) *$\int \phi \, d\alpha_n \to \int \phi \, d\alpha$ for every continuous ϕ such that $\phi(x) = 0(\|x\|^p)$ at infinity.*

PROOF. a) "Only if". Suppose $d_p(\alpha_n, \alpha) \to 0$. Let ξ_n have law α_n, and ζ have law α, and $E[\|\xi_n - \zeta\|^p]^{1/p} = d_p(\alpha_n, \alpha)$. Then

$$\left[\int \|x\|^p \alpha_n(dx)\right]^{1/p} - \left[\int \|x\|^p \alpha(dx)\right]^{1/p} = E\{\|\xi_n\|^p\}^{1/p} - E\{\|\xi\|^p\}^{1/p}$$

$$\leqq E\{\|\xi_n - \zeta\|^p\}^{1/p} \to 0$$

Likewise, if f is Lipschitz, that is $\|f(x) - f(y)\| \leqq K\|x - y\|$, then

$$\left| \int f(x)\alpha_n(dx) - \int f(x)\alpha(dy) \right| = |E\{f(\xi_n) - f(\zeta)\}| \leqq E\{|f(\xi_n) - f(\zeta)|\}$$

$$\leqq KE\{\|\xi_n - \zeta\|\} \leqq KE[\|\xi_n - \zeta\|^p]^{1/p} \to 0.$$

Then $\alpha_n \to \alpha$ weakly by a routine argument.

"If". Suppose $\alpha_n \to \alpha$ weakly and $\int \|x\|^p \alpha_n(dx) \to \int \|x\|^p \alpha(dx)$. A routine argument reduces the problem to the case where α_n and α concentrate on a fixed bounded set, using the condition on the norms; then the reduction to the case where α_n and α concentrate on a fixed compact set C is easy, using Prokhorov's theorem (Billingsley, 1968, page 37). Cover C by a finite disjoint union of sets C_i of diameter ϵ, with $\alpha(\partial C_i) = 0$, where ∂ represents the boundary. Choose $x_i \in C_i$. Replace α_n by $\tilde{\alpha}_n$, where $\tilde{\alpha}_n\{x_i\} = \alpha_n\{C_i\}$. Likewise for α. Clearly $d_p(\tilde{\alpha}_n, \alpha_n) \leqq \epsilon$ and $d_p(\tilde{\alpha}, \alpha) \leqq \epsilon$. But $d_p(\tilde{\alpha}_n, \tilde{\alpha}) \to 0$ by an easy direct argument. The rest is immediate. □

The argument for the "if" part of (a) is a variation on an argument for Vitali's theorem.

LEMMA 8.4. *Let X_i be independent B-valued random variables, with common distribution $\mu \in \Gamma_p$. Let μ_n be the empirical distribution of $X_1, \cdots, X_n$. Then $d_p(\mu_n, \mu) \to 0$ a.e.*

PROOF. Use Lemma 8.3 and the strong law. □

For B-valued random variables U and V, write $d_p(U, V)$ for the d_p-distance between the laws of U and V, assuming the latter are in Γ_p. The scaling properties of d_p are as follows:

$$d_p(aU, aV) = |a| \cdot d_p(U, V) \qquad \text{for any scalar } a \tag{8.2}$$

(8.3) $d_p(LU, LV) \leqq \|L\| \cdot d_p(U, V)$ for any linear operator L on B.

The next lemma involves two separable Banach spaces B and B', e.g., two finite-dimensional Euclidean spaces. Let $1 \leqq p, p' < \infty$.

LEMMA 8.5. *Suppose X_n is a B-valued random variable and $\|X_n\| \in L_p$; likewise for X; and $d_p(X_n, X) \to 0$. Let ϕ be a continuous function from B to B', and $\|\phi(x)\|^{p'} \leqq K\{1 + \|x\|^p\}$, where K is some constant. Then $d_{p'}[\phi(X_n), \phi(X)] \to 0$.*

PROOF. Use Lemma 8.3.

Can $d_{p'}[\phi(X_n), \phi(X)]$ be bounded above by some reasonable function of $d_p(X_n, X)$? Apparently not. Suppose $B = B'$ is the real line, $p = 2$ and $p' = 1$ and $\phi(x) = x^2$. Find real numbers x_n and y_n with $(x_n - y_n)^2 \to 0$ but $|x_n^2 - y_n^2| \to \infty$. Let $X_n = x_n$ and $Y_n = y_n$ a.s. Then $d_2(X_n, Y_n) \to 0$ but $d_1(X_n^2, Y_n^2) \to \infty$.

LEMMA 8.6. *Let U_j be independent; likewise for V_j; assume the laws are in Γ_p. Then*

$$d_p\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right) \leqq \sum_{j=1}^m d_p(U_j, V_j).$$

PROOF. In view of Lemma 8.1, assume without loss of generality that the pairs (U_j, V_j) are independent and

$$E\{\|U_j - V_j\|^p\}^{1/p} = d_p(U_j, V_j).$$

Now by Minkowski's inequality,

$$\begin{aligned} d_p\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right) &\leqq E\left\{\left\|\sum_{j=1}^m (U_j - V_j)\right\|^p\right\}^{1/p} \\ &\leqq \sum_{j=1}^m E\{\|U_j - V_j\|^p\}^{1/p} = \sum_{j=1}^m d_p(U_j, V_j). \quad \square \end{aligned}$$

In the presence of orthogonality, this result can be improved.

LEMMA 8.7. *Suppose B is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and $p = 2$. Suppose the U_j are independent, likewise for V_j; assume the laws are in Γ_2, and $E(U_j) = E(V_j)$. Then*

$$d_2\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right)^2 \leqq \sum_{j=1}^m d_2(U_j, V_j)^2.$$

PROOF. Make the same construction as in the previous lemma. Now $E\{\langle U_j - V_j, U_k - V_k \rangle\}$ is 0 or $d_2(U_j, V_j)^2$, according as $k \neq j$ or $k = j$. So

$$\begin{aligned} d_2\left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j\right)^2 &\leqq E\left\{\left\langle \sum_{j=1}^m (U_j - V_j), \sum_{j=1}^m (U_j - V_j) \right\rangle\right\} \\ &= \sum_{j=1}^m d_2(U_j, V_j)^2. \end{aligned}$$ $\square$

LEMMA 8.8. *Suppose B is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and $p = 2$. Let U and V be B-valued random variables, with $\|U\|$ and $\|V\|$ in L_2. Then*

$$d_2[U, V]^2 = d_2[U - E(U), V - E(V)]^2 + \|E(U) - E(V)\|^2.$$

PROOF. Write $a = E(U)$ and $b = E(V)$. Choose U and V so that $E(\|U - V\|^2) = d_2(U, V)^2$. Now

$$E\{\|(U - a) - (V - b)\|^2\} = E(\|U - V\|^2) - \|a - b\|^2$$

so

$$d_2(U - a, V - b)^2 \leqq d_2(U, V)^2 - \|a - b\|^2.$$

For the other inequality, choose U and V so that

$$E\{\|(U-a)-(V-b)\|^2\} = d_2(U-a, V-b)^2. \qquad \square$$

For simplicity, the next result will be given only for the line.

LEMMA 8.9. *Suppose B is the real line, $\|x\| = |x|$, and $p = 2$. Let d_2^1 be the corresponding Mallows metric. Let $U_1, \cdots, U_n$ be independent and identically distributed L_2-variables, and let U be the column vector $(U_1, \cdots, U_n)$. Let $V_1, \cdots, V_n$ and V be likewise. Suppose $E(U_i) = E(V_i)$. Let A be an $m \times n$ matrix of scalars. Now AU, AV are random vectors in R^m, equipped with the m-dimensional Euclidean norm. Write d_2^m for the corresponding d_2-metric. Then*

$$d_2^m(AU, AV)^2 \leqq \operatorname{trace}(AA') \cdot d_2^1(U_i, V_i)^2.$$

PROOF. As usual, suppose (U_i, V_i) are independent and $E\{(U_i - V_i)^2\}^{1/2} = d_2(U_i, V_i)$. Now

$$\begin{aligned} d_2(AU, AV)^2 &\leqq E\{\|AU - AV\|^2\} \\ &= E\{\operatorname{trace}[A(U-V)(U-V)'A']\} \\ &= \operatorname{trace}(AA') \cdot d_2^1(U_i, V_i)^2 \end{aligned}$$

because $E\{(U-V)(U-V)'\} = I_{n\times n} \cdot d_2^1(U_i, V_i)^2$, where $I_{n\times n}$ is the $n \times n$ identity matrix, and trace CD = trace DC, provided both matrix products make sense. $\square$

The next result expresses the idea that the bootstrap operation commutes with smooth functions. Let ϕ be a function from one separable Banach space B to another B'. Let $x_0 \in B$; most of the action will occur near x_0. Suppose that ϕ is continuously differentiable at x_0 in the following sense. For some $\delta_0 > 0$, if $\|x - x_0\| \leqq \delta_0$, then as real $h \to 0$,

$$\frac{\phi(x+hy) - \phi(x)}{h} \to \phi'(x)y \quad \text{weakly}$$

for all $y \in B$, where $\phi'(x)$ is a bounded linear mapping from B to B'. Assume too that if $\|x_n - x_0\| \to 0$ then $\|\phi'(x_n)y - \phi'(x_0)y\| \to 0$, uniformly on strongly compact y-sets. By the uniform boundedness principle, there is a positive $\delta_1 \leqq \delta$ such that $\|x - x_0\| \leqq \delta_1$ entails $\|\phi'(x)\| \leqq K$.

LEMMA 8.10. *Let X_n be a B-valued random variable and a_n a scalar tending to infinity, and $x_n \in B$ with $x_n \to x_0$. Suppose the law of $a_n(X_n - x_n)$ converges weakly to the law of W. Let ϕ be a smooth function from B to B', as above. Then the law of $a_n[\phi(X_n) - \phi(x_n)]$ converges weakly to the law of $\phi'(x_0)\,W$.*

PROOF. The argument is only sketched. Fix a bounded linear functional λ on B, an $x \in B$ with $\|x - x_0\| < \frac{1}{2}\delta_1$, a $y \in B$ with $\|y\| < \frac{1}{2}\delta_1$, and let t be real with $|t| \leqq 1$. Then

$$\frac{\partial}{\partial t}\lambda[\phi(x+ty)] = \lambda[\phi'(x+ty)y]. \tag{8.4}$$

The right hand side of (8.4) is a bounded function of t, so $t \to \lambda[\phi(x+ty)]$ is absolutely continuous, and

$$\lambda[\phi(x+ty)] = \lambda[\phi(x)] + \int_0^t \lambda[\phi'(x+uy)y]\,du. \tag{8.5}$$

Since (8.5) holds for all λ,

$$\phi(x+ty) = \phi(x) + \int_0^t \phi'(x+uy)y\,du \tag{8.6}$$

where $u \to \phi'(x+uy)y$ is strongly integrable by a direct argument. If n is large, $\|x_n - x_0\| < \frac{1}{2}\delta_1$; and $\|X_n - x_n\| < \frac{1}{2}\delta_1$ with overwhelming probability. Then, except for a set of

uniformly small probability, by substitution into (8.6),

$$a_n[\phi(X_n) - \phi(x_n)] = \int_0^1 \phi'[x_n + u(X_n - x_n)]\, a_n(X_n - x_n)\, du. \tag{8.7}$$

By Prokhorov's theorem, except on a set of uniformly small probability, $a_n(X_n - x_n) \in C$, a fixed large compact set. So, except for a set of uniformly small probability, the integrand on the right is uniformly close to $\phi'(x_0)a_n(X_n - x_n)$; this final approximation is even uniform in u. □

REMARK. The interaction of two standard terminologies is perhaps unfortunate: if b_n and $b \in B$, then $b_n \to b$ weakly means $\lambda(b_n) \to \lambda(b)$ for all bounded linear functionals λ on B. On the other hand, if W_n and W are B-valued random variables, the law of W_n converges weakly to the law of W iff $E\{\theta(W_n)\} \to E\{\theta(W)\}$ for all bounded functions θ on B which are continuous in the strong topology.

LEMMA 8.11. *If B is the Brownian bridge and T is a closed subset of* $[0, 1]$ *which contains points other than* 0 *and* 1, *then* $\sup_T |B(t)|$ *has a continuous distribution.*

Much more is probably true. The distribution of $\sup_T |B(t)|$ may well have a $\mathcal{C}^\infty$ density, and likewise for other diffusions. However, Lemma 8.11 is all we need for Corollary 4.2. To prove the lemma we need a couple of sub-lemmas. Recall that $B(\cdot)$ is a continuous Markov process.

LEMMA 8.11.1. *Let $\mathfrak{B}(t+)$ be the σ field in $\mathcal{C}[0, 1]$ of events which depend only on path behavior right after t* (Freedman, 1971, page 102). *Let P be the probability measure on $\mathcal{C}[0, 1]$ which makes the coordinate process a Brownian bridge. $\mathfrak{B}(t+)$ is trivial,* i.e., *if $A \in \mathfrak{B}(t+)$, then the conditional probability*

$$P(B \in A \mid B(t)) = 0 \quad \text{or} \quad 1$$

with probability 1.

PROOF. Given $B(t) = c$, the process $B(t + u)$ *for* $0 \leqq u \leqq 1 - t$ is Gaussian with the same joint distribution as

$$\sqrt{1-t}\, B\left(\frac{\tau}{1-t}\right) + c\frac{(1-t-\tau)}{1-t}.$$

By a remark of Doob (1949) this in turn has the same joint distributions as

$$\sqrt{1-t}\left(1 - \frac{u}{1-t}\right) W\left(\frac{u}{1-t-u}\right) + c\frac{(1-t-u)}{1-t}$$

where W is a Wiener process on $(0, \infty)$ and $W(0) = 0$. Lemma 8.11.1 follows from the Blumenthal $0 - 1$ law (see Freedman, 1971, page 106, for example).

LEMMA 8.11.2. *We can represent T as the union of two sets, T_{12} and $T - T_{12}$, such that every point in T_{12} may be approached by other points in T from both sides and $T - T_{12}$ is countable.*

PROOF. We can write $T = T_1 \cup T_2$ where T_1 is a closed perfect set and T_2 is countable (Hausdorff, 1957, page 159). Call a point of T_1 an endpoint if it can only be approached on one side by points in T_1. The set of endpoints, call it T_{11}, is clearly countable. Write $T_{12} = T_1 - T_{11}$.

PROOF OF LEMMA 8.11. Note that $\sup_T |B(t)|$ is actually a maximum since B is continuous and, moreover, that $\max_T |B(t)| > 0$ with probability 1 since T includes points other than $\{0, 1\}$. So what we need to prove is, for each $c > 0$,

$$P[\max_T |B(t)| = c] = 0.$$

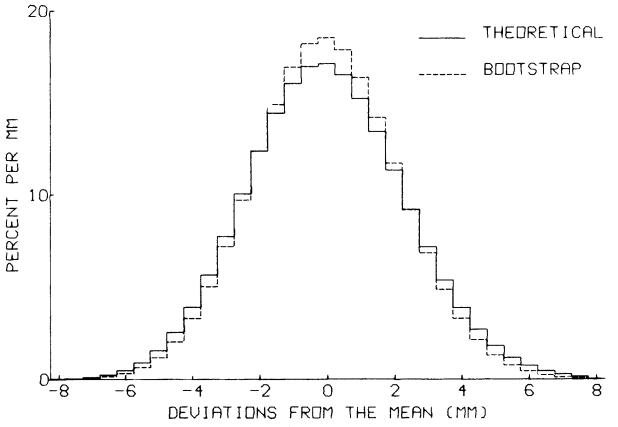

FIG. 1

A simulation, in which the bootstrap distribution is compared to the theoretical distribution.

We claim it is enough to show

$$P[\max_{T_{12}} |B(t)| = c, |B(t)| < c : t \in T - T_{12}] = 0 \tag{8.8}$$

since for $c > 0$,

$$\sum \{P[|B(t)| = c] : t \in T - T_{12}\} = 0. \tag{8.9}$$

Associate with each $t \in T_{12}$ in a measurable way a decreasing sequence $s_n(t) \downarrow t$, $s_n(t) \in T \; \forall \; n, t$. For example, take $s_n(t)$ to be the largest point in T which lies between t and $t + 1/n$. Now let σ be the first $t \in T$ such that $|B(t)| = c$ and $\sigma = 1$ otherwise. Then,

$$P[\max_{T_{12}} |B(t)| = c, |B(t)| < c, t \in T - T_{12}] \leqq P[\sigma \in T_{12}, |B(s_n(\sigma))| < |B(\sigma)| \text{ for large } n]. \tag{8.10}$$

But by Lemma 8.11.1, for any $t \in T_{12}$

$$P[|B(s_n(t))| < |B(t)| \text{ for large } n \,|\, B(t)] = 0 \text{ or } 1. \tag{8.11}$$

Since $t \in T_{12}$, $\liminf_n P[|B(s_n(t))| \geqq |c| \,|\, B(t) = c] > 0$ for any finite c and hence the probability in (8.11) is 0. By the strong Markov property the right-hand side of (8.10) is 0. Then (8.8) and the lemma follow. □

9. A simulation. To illustrate Theorem 1.1, a simulation was performed. The population consisted of the 6,672 Americans aged 18–79 in Cycle I of the Health Examination Survey.[4] The variable of interest was systolic blood pressure, with an average of 130.3 and a SD of 23.2 millimeters of mercury. The distribution had a longish right tail: the minimum was 73, the maximum 260, with skewness of 1.3 and kurtosis of 2.4.

A sample of 100 was drawn at random, with replacement. The sample average systolic blood pressure was 129.6 with a SD of 21.4. Consider these sample results from the point of view of a statistician who does not know the population figures, and has forgotten the "SD/$\sqrt{n}$" formula. Such a statistician could estimate the sampling error in the sample

[4] These 6,672 subjects were themselves a probability sample drawn from the American population. The data were provided by the National Center for Health Statistics.

average by the bootstrap principle (Theorem 1.1). The sampling error follows the theoretical sampling distribution of

$$\frac{X_1 + \cdots + X_{100}}{100} - \mu$$

where X_i is the blood pressure of the ith sample subject, and μ is the population average. This is approximated by the bootstrap distribution of

$$\frac{X_1^* + \cdots + X_{100}^*}{100} - \frac{X_1 + \cdots + X_{100}}{100},$$

where the X_i^* are drawn at random with replacement from $\{X_1, \cdots, X_{100}\}$, conditioning on these original X's.

Figure 1 compares the bootstrap distribution (dashed) with the theoretical distribution (solid). Both are rescaled convolutions, one of the population distribution, the other of the sample empirical distribution. These convolutions were computed exactly, using an algorithm based on the Fast Fourier Transform. As the figure shows, the bootstrap distribution follows the theoretical distribution rather closely.

Acknowledgment. We thank Persi Diaconis and Brad Efron for a number of helpful conversations.

REFERENCES

BERK, R. H. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37** 51–58.

BICKEL, P. J. (1966). Some contributions to the theory of order statistics. *Proceedings Fifth Berkeley Symp. Math. Statist. and Prob.* **1** 575–592.

BICKEL, P. J. and FREEDMAN, D. A. (1981). More on bootstrapping regression models. Technical report, Statistics Department, University of California, Berkeley.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

DOBRUSHIN, R. L. (1970). Describing a system of random variables by conditional distributions. *Theory Probab. Appl.* **15** 458–486 [especially Section 3].

DOOB, J. L. (1949). Heuristic approach to the Kolmogorov Smirnov theorems. *Ann. Math. Statist.* **20** 393–403.

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

FILLIPOVA, A. R. (1962). Mises' theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. *Theory Probab. Appl.* **7** 24–57.

FREEDMAN, D. A. (1971). *Brownian Motion and Diffusion.* Holden-Day, San Francisco.

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

HAUSDORFF, F. (1957). *Set Theory.* Chelsea, New York.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* **19** 293–325.

KOMLOS, J., MAJOR, P. and TUSNADY, G. (1975). An approximation of partial sums of independent r.v.'s and the sample d.f. I. *Z. Warsch. verw. Gebiete* **32** 111–131.

MAJOR, P. (1978). On the invariance principle for sums of independent, identically distributed random variables. *Jour. of Multivariate Anal.* **8** 487–501.

MALLOWS, C. L. (1972). A note on asymptotic joint normality. *Ann. Math. Statist.* **43** 508–515.

MISES, R. VON (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* **18** 309–348.

PYKE, R. and SHORACK, G. (1968). Weak convergence of a two-sample empirical process and a new approach to the Chernoff-Savage theorems. *Ann. Math. Statist.* **39** 755–771.

REEDS, J. (1976). On the definition of von Mises functionals. Thesis, Harvard University.

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

STIGLER, S. (1973). The asymptotic distribution of the trimmed mean. *Ann. Statist.* **6** 472–477.

TANAKA H. (1973). An inequality for a functional of probability distribution and its application to Kac's one-dimensional model of a Maxwellian gas. *Z. Warsch. verw. Gebiete* **27** 47–52.

VALLENDER, S. S. (1973). Calculation of the Vasershtein distance between probability distributions on the line. *Theory Probab. Appl.* **18** 784–786.

ZAANEN, A. C. (1953). *Linear Analysis.* Wiley, New York.

ZYGMUND, A. (1935). *Trigonometric Series,* (reprinted by Dover and by Chelsea, New York).

Statistics Department
University of California, Berkeley
BERKELEY, CALIFORNIA 94720

The Annals of Statistics
1984, Vol. 12, No. 2, 470–482

ASYMPTOTIC NORMALITY AND THE BOOTSTRAP IN STRATIFIED SAMPLING

BY P. J. BICKEL[1] AND D. A. FREEDMAN[2]

University of California, Berkeley

This paper is about the asymptotic distribution of linear combinations of stratum means in stratified sampling, with and without replacement. Both the number of strata and their size is arbitrary. Lindeberg conditions are shown to guarantee asymptotic normality and consistency of variance estimators. The same conditions also guarantee the validity of the bootstrap approximation for the distribution of the t-statistic. Via a bound on the Mallows distance, situations will be identified in which the bootstrap approximation works even though the normal approximation fails. Without proper scaling, the naive bootstrap fails.

1. Introduction. Consider the problem of estimating a linear combination $\gamma = \sum_{i=1}^{p} c_i \mu_i$ of the means $\mu_1, \cdots, \mu_p$ of p numerical populations $X_1, \cdots, X_p$ with corresponding distributions $F_1, \cdots, F_p$. For each $i = 1, \cdots, p$ there is a sample X_{ij} from population $\mathscr{X}_i$; the sample elements are indexed by $j = 1, \cdots, n_i$. Thus, n_i is the size of the sample from the ith population. Two situations will be discussed:

(a) The populations $\mathscr{X}_i$ are assumed arbitrary and the sampling is with replacement: X_{ij} for $j = 1, \cdots, n_i$ are identically distributed with common distribution F_i; all the X_{ij} are independent.

(b) The populations are assumed finite; $\mathscr{X}_i$ has known size N_i; sampling is without replacement and independent in i; in this case, F_i is uniform. Enumerate X_i as $\{x_{i1}, \cdots, x_{iN_i}\}$.

For simplicity, the populations are supposed univariate.

The natural unbiased estimate of γ is

$$\hat{\gamma} = \sum_{i=1}^{p} c_i X_{i\cdot}. \tag{1}$$

Here, the dot is the averaging operator.

Let τ_a^2 or τ_b^2 denote the variance of $\hat{\gamma}$ under sampling schemes (a) and (b) respectively. Let $\hat{\tau}_a^2$ or $\hat{\tau}_b^2$ be the customary unbiased variance estimates. Inference about γ can be based either on the normal approximation to the distribution of $(\hat{\gamma} - \gamma)/\hat{\tau}$ or on bootstrap approximations. This paper will discuss the validity of these approximations when the total sample size tends to ∞ in any way

Received February 1983; revised December 1983.

[1] This work was performed with the partial support of Office of Naval Research Contract N00014-80-C-0163. The hospitality of the Hebrew University, Jerusalem is also gratefully acknowledged.

[2] Research partially supported by National Science Foundation Grant MCS80-02535.

AMS 1980 *subject classification.* Primary 60F05; secondary 62E20.

Key words and phrases. Bootstrap, asymptotic normality, stratified sampling, standard errors.

whatsoever, e.g., many small samples or a few large samples or some combination thereof. More precisely: suppose p, the c_i, the populations, the N_i, and n_i all depend on an index ν such that $n(\nu) = n_1(\nu) + \cdots + n_p(\nu) \to \infty$ as $\nu \to \infty$. This index will be suppressed in the sequel.

Here are two examples.

(a) The X_{ij} are unbiased measurements of the same quantity μ, taken with p different instruments. So the precision of X_{ij}, viz.,

$$\sigma_i^2 = \int (x - \mu)^2 \, dF_i(x)$$

depends on i. If σ_i^2 is known to be proportional to r_i, then

$$\hat{\gamma} = \sum \frac{n_i}{r_i} X_{i\cdot} / \sum \frac{n_i}{r_i}$$

is the natural estimate of μ.

(b) In the classical stratified sampling model a population $\mathscr{X}$ of size N is broken up into disjoint strata $\mathscr{X}_1, \cdots, \mathscr{X}_p$ of sizes $N_1, \cdots, N_p$ respectively; $\sum_{i=1}^p N_i = N$. From stratum i the sample X_{ij} for $j = 1, \cdots, n_i$ is taken without replacement. Enumerate the ith stratum as $\{x_{i1}, \cdots, x_{iN_i}\}$. The population mean is

$$\gamma = \frac{1}{N} \sum_{i=1}^p \sum_{j=1}^{N_i} x_{ij} = \sum_{i=1}^p N_i x_{i\cdot} / N$$

and $\hat{\gamma} = \sum_{i=1}^p \mathrm{N}_i X_{i\cdot} / N$ is the usual estimate of γ.

We first take up the normal approximation in case (a). Suppose

$$\int x^2 \, dF_i < \infty \quad \text{and} \quad n_i \geq 2 \quad \text{for} \quad i = 1, \cdots, p. \tag{2}$$

Then

$$\tau_a^2 = \sum_{i=1}^p c_i^2 \sigma_i^2 / n_i \quad \text{where} \quad \sigma_i^2 = \operatorname{var} X_{ij}$$

and

$$\hat{\tau}_a^2 = \sum_{i=1}^p c_i^2 s_i^2 / n_i$$

where

$$s_i^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (X_{ij} - X_{i\cdot})^2.$$

Let

$$\begin{aligned} \phi(x, \varepsilon) &= x \quad \text{for} \quad |x| \geq \varepsilon \\ &= 0 \quad \text{otherwise} \\ \bar{\phi}(x, \varepsilon) &= x - \phi(x, \varepsilon). \end{aligned}$$

Suppose that for all $\varepsilon > 0$,

$$\tau_i^{-2} \sum_{i=1}^{p} n_i^{-1} c_i^2 E\{\phi^2(X_{ij} - \mu_i, \varepsilon n_i \tau_a |\, c_i\,|^{-1})\} \to 0. \tag{3}$$

By the Lindeberg-Feller theorem, $(\hat{\gamma} - \gamma)/\tau_a$ converges in law to $\mathcal{N}(0, 1)$, the standard normal distribution.

According to the first main theorem of this paper, conditions (2) and (3) are also sufficient to guarantee that $\hat{\tau}_a^2$ has the right limiting behavior. However, before giving a precise statement, it may be helpful to reformulate condition (3). Let $Y_{ij} = (X_{ij} - \mu_i)/\sigma_i$. Define the "variance weight" of the ith stratum by

$$w_i^2 = c_i^2 \sigma_i^2 / n_i \tau_a^2 = \text{var}\,\{c_i X_{i\cdot}/\tau_a\}.$$

Clearly,

$$\sum_{i=1}^{p} w_i^2 = 1.$$

Condition (3) can then be written

$$\sum_{i=1}^{p} E\{\phi^2(w_i Y_{ij}, \varepsilon \sqrt{n_i})\} \to 0 \quad \text{for all} \quad \varepsilon > 0. \tag{4}$$

THEOREM 1. *If* (2) *and* (4) *hold in case* (a), *then* $\hat{\tau}_a^2/\tau_a^2 \to 1$ *in probability.*

The proof is deferred.

COROLLARY. $(\hat{\gamma} - \gamma)/\hat{\tau}_a$ *tends to* $\mathcal{N}(0, 1)$ *in law.*

We consider next the bootstrap approximation in case (a); also see Babu and Singh (1983). For $i = 1, \cdots, p$, let $\hat{F}_i$ be the empirical distribution of X_{ij} for $j = 1, \cdots, n_i$. Take samples of size n_i with replacement from $\hat{F}_i$. That is, let $\{X_{ij}^*\}$ be conditionally independent given $\mathcal{F}$, the σ-field spanned by $\{X_{ij}\}$; let X_{ij}^* have common distribution $\hat{F}_i$ for $j = 1, \cdots, n_i$. Let

$$\hat{\gamma}^* = \sum_{i=1}^{p} c_i X_{i\cdot}^*, \quad s_i^{*2} = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (X_{ij}^* - X_{i\cdot}^*)^2$$

$$\hat{\tau}_a^{*2} = \sum_{i=1}^{p} c_i^2 s_i^{*2}/n_i, \quad \tilde{\tau}_a^2 = \sum_{i=1}^{p} c_i^2 (n_i - 1) s_i^2/n_i^2.$$

THEOREM 2. *If* (2) *and* (4) *hold in case* (a), *then the conditional distribution of* $(\hat{\gamma}^* - \hat{\gamma})/\tilde{\tau}_a$ *converges weakly to* $\mathcal{N}(0, 1)$ *in probability, and* $\hat{\tau}_a^*/\tilde{\tau}_a$ *converges to* 1 *in probability.*

The proof is deferred. The theorem points to a problem in using the bootstrap to make inferences: the scaling may go wrong. This is because $X_{i\cdot}^*$ has variance $(n_i - 1)s_i^2/n_i^2$, not s_i^2/n_i. To fix ideas, suppose there are many small strata: more particularly, that $n_i \le k$ for all i. Now

$$\tilde{\tau}_a^2 \le (k - 1)/k \cdot \hat{\tau}_a^2 \approx (k - 1)/k \cdot \tau_a^2.$$

The bootstrap distribution of $\hat{\gamma}^* - \hat{\gamma}$ has asymptotic scale $\tilde{\tau}_a$, while $\hat{\gamma} - \gamma$ has the scale τ_a.

We take up next the normal approximation in case (b). Suppose

$$2 \le n_i \le N_i - 1. \tag{5}$$

Then

$$\tau_b^2 = \sum_{i=1}^p c_i^2 \frac{\sigma_i^2}{n_i} \frac{(N_i - n_i)}{N_i - 1}$$

and

$$\hat{\tau}_b^2 = \sum_{i=1}^p c_i^2 \frac{s_i^2}{n_i} \frac{(N_i - n_i)}{N_i}.$$

To state the regularity condition, let v_i^2 be the "variance weight" in case (b): $v_i^2 = c_i^2 \sigma_i^2 (N_i - n_i)/n_i \tau_b^2 (N_i - 1) = \text{var}\{c_i X_{i\cdot}/\tau_b\}$. Let ρ_i be "the effective sample size:" $\rho_i = n_i(N_i - 1)/(N_i - n_i)$. Let $\mathscr{Y}_i = \{y_{i1}, \cdots, y_{iN_i}\}$ where $y_{ij} = (x_{ij} - \mu_{i\cdot})/\sigma_i$ and $\sigma_i^2 = N_i^{-1} \sum_{j=1}^{N_i} (x_{ij} - \mu_{i\cdot})^2$. So $Y_{ij} = (X_{ij} - \mu_i)/\sigma_i$ are sampled from $\mathscr{Y}_i$.

The condition is

$$\sum_{i=1}^p N_i^{-1} \sum_{j=1}^{N_i} \phi^2(v_i y_{ij}, \varepsilon\sqrt{\rho_i}) \to 0. \tag{6}$$

This may be compared with condition (4).

If $\sup_{1 \le i \le p} E \mid Y_i \mid^3$ is uniformly bounded independent of the hidden index ν, the Lindeberg conditions (4) and (6) are implied respectively by the natural conditions $\max_i w_i/\sqrt{n_i} \to 0$ or $\max_i v_i/\sqrt{\rho_i} \to 0$. Thus if the standardized populations have reasonably light tails, asymptotic normality holds if for each stratum the variance weight contribution is small or the stratum is heavily sampled.

THEOREM 3. *If* (5) *and* (6) *hold in case* (b), *then*

i) $$(\hat{\gamma} - \gamma)/\tau_b \to \mathscr{N}(0, 1) \quad \textit{in law}$$

and

ii) $$\hat{\tau}_b/\tau_b \to 1 \quad \textit{in probability.}$$

The proof is deferred.

COROLLARY. $(\hat{\gamma} - \gamma)/\hat{\tau}_b \to \mathscr{N}(0, 1)$ *in law.*

Finally, we consider the bootstrap in case (b). If $N_i/n_i = k_i$ an integer for each i, the natural bootstrap procedure was suggested by Gross (1980): given $\{X_{ij}\}$, to create populations $\hat{\mathscr{X}}_i$ consisting of k_i copies of each X_{ij} for $j = 1, \cdots, n_i$, then X_{ij}^* for $j = 1, \cdots, n_i$ are generated as a sample without replacement from $\hat{\mathscr{X}}_i$, the samples being independent for different $i = 1, \cdots, p$. In general, if $N_i = k_i n_i + r_i$ with $0 \le r_i < n_i$, form populations $\hat{\mathscr{X}}_{i0}$ and $\hat{\mathscr{X}}_{i1}$, where $\hat{\mathscr{X}}_{i0}$ consists of k_i

copies of each X_{ij}, for $j = 1, \cdots, n_i$; while $\mathscr{X}_{i1}$ consists of $k_i + 1$ copies. Let

$$\alpha_i = \left(1 - \frac{r_i}{n_i}\right)\left(1 - \frac{r_i}{N_i - 1}\right).$$

With probability α_i, let $(X_{i1}^*, \cdots, X_{in_i}^*)$ be a sample without replacement of size n_i from $\hat{\mathscr{X}}_{i0}$; with probability $1 - \alpha_i$, let $(X_{i1}^*, \cdots, X_{in_i}^*)$ be a sample without replacement of size n_i from $\hat{\mathscr{X}}_{i1}$. The virtue of this scheme is that both $\hat{\mathscr{X}}_{i0}$ and $\hat{\mathscr{X}}_{i1}$ have the same distribution $\hat{F}_i$ and

$$\operatorname{Var}(X_{i\cdot}^* \mid \{X_{ij}\}) = \frac{n_i - 1}{n_i^2} s_i^2 \left(\frac{N_i - n_i}{N_i - 1}\right).$$

The proof of the following theorem is similar to that of Theorem 2 and is omitted. Define $\hat{\gamma}^*$ as before, and $\hat{\tau}_b^{*2}$ by substituting X_{ij}^* for X_{ij} in $\hat{\tau}_b^2$.

THEOREM 4. *Let $\tilde{\tau}_b^2$ be the variance of $\hat{\gamma}^*$ given the data. Then, if* (5) *and* (6) *hold in case* (b), *the conditional distribution of $(\hat{\gamma}^* - \hat{\gamma})/\tilde{\tau}_b$ converges weakly to $\mathscr{N}(0, 1)$ and $\hat{\tau}_b^*/\tilde{\tau}_b \to 1$ in probability.*

The same inference problem arises as in the case of Theorem 2. The variance of $\hat{\gamma}^*$ given the data is an inconsistent estimate of the variance of $\hat{\gamma}$. We have side-stepped the issue by computing the scale externally to the bootstrap process. Other patches could be made: one is to rescale the elements of $\mathscr{X}_i$; another is to adjust the constants c_i. These fixes are all a bit *ad hoc*.

If γ stays bounded as $\nu \to \infty$, our results extend easily to pivots

$$\frac{g(\hat{\gamma}) - g(\gamma)}{g'(\gamma)\hat{\tau}_b}$$

where g is nonlinear continuously differentiable. The same issue as before arises a fortiori for nonlinear functions. Neither the variance of $g(\hat{\gamma}^*)$ given the data nor its natural approximation $[g'(\hat{\gamma})]^2\tilde{\tau}_b^2$ are consistent estimates of the asymptotic variance of $g(\hat{\gamma})$. A fix which works if $\sum_{i=1}^p |c_i\mu_i|$ stays bounded is as before to rescale the elements of $\mathscr{X}_i$ or the c_i before applying the bootstrap. Alternatives (the jackknife, linearization, BRR) are discussed in Krewski and Rao (1981). For the case of one stratum, Theorem 4 was derived independently by Chao and Lo (1983).

The bootstrap can work even when Theorem 4 fails but the circumstances are artificial. Suppose we have only one stratum and $N_1 - n_1 = k$ for all ν i.e., all but k members are sampled. Since $\sum_{j=1}^{N_1}(x_{1j} - \mu_1) = 0$, the pivot $(\hat{\gamma} - \gamma)/\tau_b$ is distributed as the standardized mean of a sample of size k taken without replacement from the population $\mathscr{Y}_1$. No matter how large N_1 is, if k is small and $\mathscr{Y}_1$ nonnormal, we would not expect the normal approximation to apply to $\hat{\gamma}$. To be specific let F_ν be the uniform distribution on $\mathscr{Y}_1$ and suppose

(7) $\qquad F_\nu$ converges to F in the Mallows d_2-metric,

i.e., $F_\nu \to F$ weakly and $\int x^2\, dF_\nu \to \int x^2\, dF$. Then $(\hat{\gamma} - \gamma)/\tau_b$ is distributed in the

limit as the standardized mean of k independent variables identically distributed according to F. On the other hand, since we have sampled nearly the whole population we expect the bootstrap to work.

THEOREM 5. *If* (7) *holds, the conditional distribution of* $(\hat{\gamma}^* - \hat{\gamma})/\tilde{\tau}_b$ *converges weakly in probability to the same limit as that of the unconditional distribution of* $(\hat{\gamma} - \gamma)/\tau_b$. *Moreover,* $\hat{\tau}_b/\tau_b$ *and* $\hat{\tau}_b^*/\tilde{\tau}_b$ *both tend to* 1 *in probability.*

We can extend this result somewhat by replacing (7) with a compactness-in-d_2 condition on $\{F_\nu\}$

$$\lim \sup_{m\to\infty} \lim \sup_\nu N_1^{-1} \sum_{j=1}^{N_1} \phi^2(v_1 y_{1j}, m) = 0.$$

This condition is evidently weaker than (6) for $p = 1$. The conclusion now is that the d_2-distance between the conditional distribution of $(\hat{\gamma}^* - \hat{\gamma})/\hat{\tau}_b^*$ and the unconditional distribution of $(\hat{\gamma} - \gamma)/\hat{\tau}_b$ tends in probability to 0. A further extension to an arbitrary number of strata which includes both Theorems 4 and 5 is also possible but not worthwhile.

2. Some lemmas. Recall the truncation operator ϕ from Section 1.

LEMMA 1. a) $$\left| \phi\left(\frac{1}{k} \sum_{i=1}^k y_i, \varepsilon\right) \right| \le \sum_{i=1}^k |\phi(y_i, \varepsilon/k)|; \quad \textit{equivalently,}$$

$$|\phi(\textstyle\sum_1^k y_i, \varepsilon)| \le k \sum_{i=1}^k |\phi(y_i, \varepsilon/k^2)|$$

b) *Let* $Y_1, Y_2, \cdots$ *be independent and identically distributed. Then*

$$E\left\{\phi^2\left(\frac{1}{k} \sum_{i=1}^k Y_i, \varepsilon\right)\right\} \le k^2 E\{\phi^2(Y_i, \varepsilon/k)\}.$$

PROOF. Claim a). As is easily verified,

$$\left| \phi\left(\frac{1}{k} \sum_{i=1}^k y_i, \varepsilon\right) \right| \le \phi\left(\frac{1}{k} \sum_{i=1}^k |y_i|, \varepsilon\right).$$

Without loss of generality, suppose all $y_i \ge 0$. Let $y_{(k)}$ be the largest y_i. If $y_{(k)} < \varepsilon/k$, both sides of the inequality vanish. If $y_{(k)} \ge \varepsilon/k$, the left side is the average of the y_i, or zero; the right side is at least the maximum $y_{(k)}$.

Claim b) follows by the Cauchy-Schwarz inequality. □

LEMMA 2. *Let* $(X'_1, \cdots, X'_n)$ *and* $(X_1, \cdots, X_n)$ *be distributed respectively as samples with and without replacement from a finite population. Then*

$$E\{\phi^2(\textstyle\sum_{i=1}^n X_i, \varepsilon)\} \le E\{\phi^2(\sum_{i=1}^n X'_i, \tfrac{1}{2}\varepsilon)\}.$$

PROOF. By a theorem of Hoeffding (1963), if g is convex, then

$$E\{g(\textstyle\sum X_i)\} \le E\{g(\sum X'_i)\}.$$

Let

$$g(x, \varepsilon) = x^2 \qquad \text{for} \quad |x| \geq \varepsilon$$
$$= 2\varepsilon|x| - \varepsilon^2 \qquad \text{for} \quad \tfrac{1}{2}\varepsilon \leq |x| \leq \varepsilon$$
$$= 0 \qquad \text{otherwise.}$$

Then g is convex and

$$\phi^2(x, \varepsilon) \leq g(x, \varepsilon) \leq \phi^2(x, \tfrac{1}{2}\varepsilon).$$

So

$$E\{\phi^2(\sum X_i, \varepsilon)\} \leq E\{g(\sum X_i, \varepsilon)\} \leq E\{g(\sum X_i', \varepsilon)\} \leq E\{\phi^2(\sum X_i', \tfrac{1}{2}\varepsilon)\}. \quad \square$$

The next result involves the Mallows metric d_2; see Mallows (1972) or Bickel and Freedman (1981).

LEMMA 3. *Let $\mathscr{X}$ and $\mathscr{Y}$ be two finite populations of real numbers, of the same size N. Let F and G be the uniform distributions on $\mathscr{X}$ and $\mathscr{Y}$. Suppose F and G have the same means. Let $X_1, \cdots, X_n$ be a sample of size n, drawn at random without replacement from $\mathscr{X}$; let $F_{(n)}$ be the law of $X_1 + \cdots + X_n$. Likewise for $Y_1, \cdots, Y_n$ and $G_{(n)}$. Then*

$$d_2[F_{(n)}, G_{(n)}]^2 \leq \frac{n(N-n)}{N-1} d_2(F, G)^2.$$

PROOF. Enumerate $\mathscr{X}$ as $x_1 \leq x_2 \leq \cdots \leq x_N$ and $\mathscr{Y}$ as $y_1 \leq y_2 \leq \cdots \leq y_N$. Then

$$(1/N) \sum_{i=1}^N (x_i - y_i)^2 = d_2(F, G)^2.$$

This follows from Bickel and Freedman (1981, Lemmas 8.2 and 8.3). Let $Z = \{1, \cdots, N\}$. Let $Z_1, \cdots, Z_n$ be a sample of size n, drawn at random without replacement from Z. Set $X_i = x_{Z_i}$ and $Y_i = y_{Z_i}$. Now

$$d_2[F_{(n)}, G_{(n)}]^2 \leq E\{[\textstyle\sum_{i=1}^n (X_i - Y_i)]^2\} = \dfrac{n(N-n)}{N-1} E[(X_i - Y_i)^2]$$

$$= \frac{n(N-n)}{N-1} d_2(F, G)^2. \quad \square$$

Here is an easy generalization of Lemma 3.

LEMMA 4. *For $i = 0, 1$ let $\mathscr{X}_i = \{x_{i1}, \cdots, \mathscr{X}_{iN_i}\}$ be finite populations and F_i the associated uniform distributions on $\mathscr{X}_i$. Let F_{ni} be the distribution of $\sum_{j=1}^n X_j$ when $X_1, \cdots, X_n$ is a sample without replacement from $\mathscr{X}_i$. Let $n \leq N_0 \leq N_1$. If J is a subset of $\{1, \cdots, N_1\}$, let F_{1J} be the uniform distribution on $\{x_{1j}: j \in J\}$.*

Then,

$$d_2(F_{n0}, F_{n1})^2 \leq \frac{n(N_0 - n)}{N_0 - 1} \frac{1}{\binom{N_1}{N_0}} \sum_J \{d_2(F_0, F_{1J})^2 : |J| = N_0\}.$$

LEMMA 5. *For $\nu \geq 1$ let $\mathscr{X}_\nu$ be a finite population of size N_ν, F_ν the uniform distribution on $\mathscr{X}_\nu$, $X_1, \cdots, X_{n_\nu}$, a sample without replacement from $\mathscr{X}_\nu$, $\hat{F}_\nu$ the empirical* df *of the sample. If for some F, $d_2(F_\nu, F) \to 0$ as $\nu \to \infty$ and $n_\nu \to \infty$ then $d_2^2(\hat{F}_\nu, F) \to 0$ in probability.*

PROOF. If g is continuous and bounded

$$E \int g(x)\, d\hat{F}_\nu(x) = \int g(x)\, dF_\nu(x) \to \int g(x)\, dF(x),$$

$$\operatorname{Var}\left(\int g(x)\, d\hat{F}_\nu(x) \right) \to 0.$$

So,

$$\int g(x)\, d\hat{F}_\nu(x) \to \int g(x)\, dF(x) \tag{8}$$

in probability. Moreover,

$$\limsup_\nu E \int \phi(x, M)^2\, d\hat{F}_\nu(x) = \int \phi(x, M)^2\, dF(x)$$

by Lemma 8.3c) of Bickel and Freedman (1981). Since we can make $\int \phi(x, M)^2\, dF(x)$ small for M large we conclude that (8) holds for $g(x) = x^2$ also and the lemma follows. □

3. Proving the theorems in case (a).

PROOF OF THEOREM 1. Recall the variance weights w_i from Section 1. As is easily verified, $\hat{\tau}_a^2/\tau_a^2 = 1 + \xi - \zeta$, where

$$\xi = \sum_{i=1}^p w_i^2 (n_i - 1)^{-1} \sum_{j=1}^{n_i} (Y_{ij}^2 - 1) \tag{9a}$$

$$\zeta = \sum_{i=1}^p w_i^2 (n_i - 1)^{-1} (n_i Y_{i\cdot}^2 - 1). \tag{9b}$$

To prove the theorem, it is enough to show that ξ and ζ are both small. But $\xi = \xi_1 + \xi_2$, where

$$\xi_1 = \sum_{i=1}^p (n_i - 1)^{-1} \sum_{j=1}^{n_i} [\bar{\phi}^2(w_i Y_{ij}, \varepsilon\sqrt{n_i}) - E\{\bar{\phi}^2(w_i Y_{ij}, \varepsilon\sqrt{n_i})\}] \tag{10a}$$

$$\xi_2 = \sum_{i=1}^p (n_i - 1)^{-1} \sum_{j=1}^{n_i} [\phi^2(w_i Y_{ij}, \varepsilon\sqrt{n_i}) - E\{\phi^2(w_i Y_{ij}, \varepsilon\sqrt{n_i})\}]. \tag{10b}$$

Now

$$E(\xi_1^2) = \operatorname{var} \xi_1 = \sum_{i=1}^{p}(n_i - 1)^{-2} \sum_{j=1}^{n_i} \operatorname{var}\{\bar{\phi}^2(w_i Y_{ij}, \varepsilon\sqrt{n_i})\}$$
$$\leq \sum_{i=1}^{p} (n_i - 1)^{-2} n_i E\{\bar{\phi}^4(w_i Y_{ij}, \varepsilon\sqrt{n_i})\}$$
$$\leq \varepsilon^2 \sum_{i=1}^{p} (n_i - 1)^{-2} n_i^2 E\{\bar{\phi}^2(w_i Y_{ij}, \varepsilon\sqrt{n_i})\}$$
$$\leq \varepsilon^2 \sum_{i=1}^{p} (n_i - 1)^{-2} n_i^2 w_i^2 E\{Y_{ij}^2\}$$
$$\leq 4\varepsilon^2 \sum_{i=1}^{p} w_i^2 = 4\varepsilon^2.$$

On the other hand, $E\{|\,\xi_2\,|\} \to 0$ for each $\varepsilon > 0$, by (4). This disposes of ξ.

The term ζ in (9b) can be decomposed according to whether $n_i > M$ or $n_i \leq M$. Since

$$\sum_i \{(n_i - 1)^{-1} w_i^2 \colon n_i \geq M + 1\} \leq M^{-1}$$

and $E\{n_i Y_{i\cdot}^2\} = 1$, the strata i with $n_i \geq M + 1$ are negligible. For the i with $n_i \leq M$, $\zeta = \zeta_1 + \zeta_2$ where

$$\zeta_1 = \sum_i \frac{n_i}{n_i - 1} [\bar{\phi}^2(w_i Y_{i\cdot}, \varepsilon\sqrt{n_i}) - E\{\bar{\phi}^2(w_i Y_{i\cdot}, \varepsilon\sqrt{n_i})\}] \tag{11a}$$

$$\zeta_2 = \sum_i \frac{n_i}{n_i - 1} [\phi^2(w_i Y_{i\cdot}, \varepsilon\sqrt{n_i}) - E\{\phi^2(w_i Y_{i\cdot}, \varepsilon\sqrt{n_i})\}]. \tag{11b}$$

The sums need be extended only over i with $2 \leq n_i \leq M$. Now whatever n_i may be, as for ξ_1,

$$E\{\zeta_1^2\} \leq 4\varepsilon^2 \tag{12}$$

is small. Next,

$$E\{|\,\zeta_2\,|\} \leq 2 \sum_i \frac{n_i}{n_i - 1} E\{\phi^2(w_i Y_{i\cdot}, \varepsilon\sqrt{n_i})\}$$
$$\leq 4M^2 \sum_i E\{\phi^2(w_i Y_{ij}, \varepsilon\sqrt{n_i}/M)\} \tag{13}$$

because $2 \leq n_i \leq M$; see Lemma 1. So ζ_2 is small too, by condition (4). □

PROOF OF THEOREM 2. The Lindeberg condition is applied, given $\mathscr{F}$. It is enough to check that for every $\varepsilon > 0$,

$$\tilde{\tau}_a^{-2} \sum_{i=1}^{p} n_i^{-1} c_i^2 E\{\phi^2(X_{ij}^* - X_{i\cdot}^*, \varepsilon n_i \tilde{\tau}_a\,|\,c_i\,|^{-1})\,|\,\mathscr{F}\} \to 0 \tag{14}$$

in probability, where $\tilde{\tau}_a^2 = \sum_{i=1}^{p} c_i^2 (n_i - 1) s_i^2 / n_i^2$ is the conditional variance of $\hat{\gamma}^*$ given $\mathscr{F}$. For then, Theorem 1 can be applied to X_{ij}^*.

Since $n_i \geq 2$,

$$\tfrac{1}{2}\hat{\tau}_a \leq \tilde{\tau}_a \leq \hat{\tau}_a. \tag{15}$$

Thus $\hat{\tau}_a$ and hence τ_a may be substituted in (14) for $\tilde{\tau}_a$. So (14) reduces to

$$\tau_a^{-2} \sum_{i=1}^{p} c_i^2 n_i^{-2} \sum_{j=1}^{n_i} \phi^2(X_{ij} - X_{i\cdot}, \varepsilon n_i \tau_a |\, c_i\,|^{-1}) \to 0$$

in probability. This in turn reduces to

$$\sum_{i=1}^{p} n_i^{-1} \sum_{j=1}^{n_i} \phi^2[w_i(X_{ij} - X_{i\cdot})/\sigma_i, \varepsilon\sqrt{n_i}] \to 0 \tag{16}$$

in probability.

Now $(X_{ij} - X_{i\cdot})/\sigma_i = Y_{ij} - Y_{i\cdot}$. Use Lemma 1a) with $k = 2$ to see that (16) follows from (17) and (18):

$$\sum_{i=1}^{p} n_i^{-1} \sum_{j=1}^{n_i} \phi^2(w_i Y_{ij}, \tfrac{1}{4}\varepsilon\sqrt{n_i}) \to 0 \quad \text{in probability} \tag{17}$$

and

$$\sum_{i=1}^{p} \phi^2(w_i Y_{i\cdot}, \tfrac{1}{4}\varepsilon\sqrt{n_i}) \to 0 \quad \text{in probability.} \tag{18}$$

Clearly, (17) follows from (4). We bound the expected value of the left side of (18). Take first those i with $n_i \le M$. In view of Lemma 1b), the sum over such i is bounded above by

$$M^2 \sum_i E\{\phi^2(w_i Y_{ij}, \tfrac{1}{4}\varepsilon\sqrt{n_i}/M)\}$$

which tends to zero by condition (4). Take next those i with $n_i > M$. The sum over such i is bounded above by

$$\sum_i E\{(w_i Y_{i\cdot})^2\} = \sum_i w_i^2 n_i^{-1} < M^{-1} \sum_i w_i^2 \le M^{-1}$$

which is small for M large.

That $\tilde{\tau}_a^*/\tilde{\tau}_a \to 1$ follows from Theorem 1. □

REMARKS. (i) The Lindeberg-Feller theorem can be supplemented by direct bounds generalizing those of Berry-Esseen; see Petrov (1975, Theorem 3, page 111 or Theorem 8, page 118). These bounds may give estimates on the discrepancy between the bootstrap distribution and the true distribution.

(ii) The difference between the distribution of $(\hat{\gamma} - \gamma)/\tau_a$ and the bootstrap distribution of $(\hat{\gamma}^* - \hat{\gamma})/\tilde{\tau}_a$ can be estimated using the Mallows metric as in equation (2.2) of Bickel and Freedman (1981). The condition needed to push this through is stronger than (4).

(iii) The results can be extended in an obvious way to vector X_{ij}, and under further conditions to nonlinear statistics such as $\sum_{i=1}^{p} [g_i(X_{i\cdot}) - g_i(\mu_i)]$; this covers ratio estimates.

4. Proving the theorems in case (b)

PROOF OF THEOREM 3. The Lindeberg-Feller theorem does not apply to give us i) directly here, since the X_{ij} are dependent for fixed i; however, essentially

the same ideas can be used. The proof we give is a bit complicated; an alternative but we believe no simpler approach is given by Dvoretzky (1971). Our argument is by cases, and the focus is on asymptotic normality. Without loss of generality, assume $\mu_i \equiv 0$, $c_i \equiv 1$. In outline, the argument is as follows.

CASE 1. There is only one stratum, and $n \leq \frac{1}{2}N$; we drop the unnecessary stratum subscript i. Then ρ^2 is of order n, and asymptotic normality follows from Erdös-Renyi (1959). Also see Rosén (1967), Dvoretzky (1971).

CASE 2. There is only one stratum, and $n > \frac{1}{2}N$. Apply Case 1 to the "co-sample" consisting of the objects not in the sample.

CASE 3. The number of strata is bounded; no variance weight tends to zero. Case 1 or Case 2 applies to each stratum individually.

CASE 4. There are many strata, each of small variance weight; in each stratum, $n_i \leq \frac{1}{2}N_i$. Then $\hat{\gamma}/\tau_b$ is the sum of p independent u.a.n. summands: $\text{var}\{X_i./\tau_b\} = v_i^2$ being uniformly small by assumption. We must verify the Lindeberg condition on $X_i./\tau_b$, and do so by an indirect argument. Let X'_{ij} be sampled with replacement from $\mathscr{X}_i$. And let

$$\hat{\gamma}' = \sum_{i=1}^{p} \frac{1}{n_i} \sum_{j=1}^{n_i} X'_{ij}.$$

Since $n_i \leq \frac{1}{2}N_i$, the variance weights v_i^2 and w_i^2 are of the same order, as are the total variances τ_a^2 and τ_b^2. In particular, condition (6) implies (4). Thus, the Lindeberg condition holds for the individual summands in $\hat{\gamma}'/\tau_a$, viz., $X'_{ij}/n_i\tau_a$, and asymptotic normality of $\hat{\gamma}'$ follows. By the converse to Lindeberg's theorem, his condition holds for the stratum averages $(1/n_i)\sum_{j=1}^{n_i} X'_{ij}/\tau_a$. Hence, by Lemma 2, the condition holds for the stratum averages taken without replacement, viz., $(1/n_i)\sum_{j=1}^{n_i} X_{ij}/\tau_b$. Now a second application of the direct Lindeberg theorem gives asymptotic normality of $\hat{\gamma}$.

CASE 5. There are many strata, each of small variance weight; on each stratum, $n_i > \frac{1}{2}N_i$. Apply Case 4 to the co-samples.

CASE 6. There are many strata, each of small variance weight. Consider two groups of strata: in the first, $n_i \leq \frac{1}{2}N_i$; in the second, $n_i > \frac{1}{2}N_i$. Case 4 applies to the first group, Case 5 to the second. (One of the two groups may be negligible.)

The general case. We combine cases 3 and 6. Let

$$J_k(\nu) = \left\{ i : v_i \geq \frac{1}{k} \right\}; \quad V_k(\nu) = \sum \{v_i^2 : i \in J_k(\nu)\}$$

where dependence on the hidden index is made explicit. Given any subsequence of $\{\nu\}$ we can extract a subsequence $\{\nu_r\}$ such that for all k, as $r \to \infty$, $V_k(\nu_r)$ tends

to a finite limit V_k. If $V_k = 0$ for all k, there must be $k_r \to \infty$ such that $V_{k_r}(\nu_r) \to 0$. Hence, as $r \to \infty$,

$$\sum\{X_i./\tau_b\colon i \in J_{k_r}(\nu_r)\} \to 0 \quad \text{in probability.} \tag{19}$$

But, $\max\{v_i\colon i \notin J_{k_r}(\nu_r)\} \le 1/k_r \to 0$. So we can apply case 6 to get that

$$\sum\{X_i./\tau_b\colon i \notin J_{k_r}(\nu_r)\} \quad \text{is asymptotically} \quad N(0, 1). \tag{20}$$

Combining (19) and (20), we get

$$\sum X_i./\tau_b \quad \text{is asymptotically} \quad N(0, 1), \quad \text{as} \quad r \to \infty. \tag{21}$$

On the other hand, suppose $V_k > 0$ for some k. Since $J_k(\nu_r)$ has at most k^2 members, we can apply case 3 to see that for all k, as $r \to \infty$,

$$\sum\{X_i./\tau_b\colon i \in J_k(\nu_r)\} \quad \text{is asymptotically} \quad N(0, V_k).$$

By a standard argument, there are $k_r \to \infty$ such that

$$\sum\{X_i./\tau_b\colon i \in J_{k_r}(\nu_r)\} \quad \text{is asymptotically} \quad N(0, \sup_k V_k). \tag{22}$$

Applying case 6 as above,

$$\sum\{X_i./\tau_b\colon i \notin J_{k_r}(\nu_r)\} \quad \text{is asymptotically} \quad N(0, 1 - \sup_k V_k). \tag{23}$$

Combining (22) and (23) we obtain (21) in this case also. Part (i) of the theorem follows by a standard compactness argument. The proof of (ii) follows the pattern of that of Theorem 1 and is omitted. □

PROOF OF THEOREM 5. We simplify the argument by supposing n_1 divides N_1 so we can use the naive bootstrap. (The general argument uses Lemma 4.) Moreover, without loss of generality let $\mu_1 = 0$, $\sigma_1 = 1$. Since $p = 1$ we want to compare the distribution of the standardized mean of a sample of size n_1 from the population $\mathscr{Y}_1$ and the distribution of the standardized mean of a sample of size n_1 from the population composed of N_1/n_1 copies of the standardized sample: $(X_{ij} - \hat{\mu}_1)/\hat{\sigma}_1$, $1 \le j \le n_1$, where $\hat{\mu}_1$ is the sample mean and $\hat{\sigma}_1$ is sample standard deviation. So by Lemma 3,

$$d_2^2\left\{\mathscr{L}\left(\frac{\hat{\gamma} - \gamma}{\tau_b}\right), \mathscr{L}\left(\frac{\hat{\gamma}^* - \hat{\gamma}}{\hat{\tau}_b}\right) \,\middle|\, X_{1j},\ 1 \le j \le n_1\right\} \le d_2^2\{F_\nu, \hat{F}_\nu, \hat{F}_\nu(\hat{\sigma}_1 x + \hat{\mu}_1)\}.$$

By Lemma 5, $d_2^2(F_\nu, \hat{F}_\nu)$, $\hat{\mu}$, and $\hat{\sigma}_1 - 1$ all tend in probability to 0 as $\nu \to \infty$. A truncation argument of the type we have already used shows that $\hat{\tau}_b/\tau_b$ and $\hat{\tau}_b^*/\hat{\tau}_b$ both tend in probability to 1. The theorem follows. □

REFERENCES

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

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

BICKEL, P. J. and KRIEGER, A. (1983). Using the bootstrap to set confidence bands for a distribution function: Monte Carlo and some theory. In preparation.

CHAO, M. T. and LO, S. H. (1983). A bootstrap method for finite population. Preprint.

DVORETZKY, A. (1971). Asymptotic normality for sums of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Probab.* II 513. (Ed: Le Cam, Neyman, Scott). Univ. of Calif. Press, Berkeley.

ERDÖS, P. and RENYI, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hung. Acad. Sci.* **4** 49–61.

GROSS, S. (1980). Median estimation in sample surveys. Paper presented at 1980 Amer. Statist. Assoc. meeting.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

KREWSKI, D. and RAO, J. N. K. (1981). Inference from stratified samples: properties of linearization, jackknife, and balanced repeated replication. *Ann. Statist.* **9** 1010–1019.

MALLOWS, C. (1972). A note on asymptotic joint normality. *Ann. Math. Statist.* **43** 508–515.

PETROV, V. (1975). *Sums of Independent Random Variables.* Springer, Berlin.

ROSÉN, B. (1967). On the central limit theorem for sums of dependent random variables. *Z. Wahrsch. verw. Gebiete* **7** 48–82.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

Richardson Extrapolation and the Bootstrap

PETER J. BICKEL and JOSEPH A. YAHAV*

Simulation methods [particularly Efron's (1979) bootstrap] are being applied more and more frequently in statistical inference. Given data $(X_1, \ldots, X_n)$ distributed according to P, which belongs to a hypothesized model $\mathbf{P}$, the basic goal is to estimate the distribution $\mathbf{L}_P$ of a function $T_n(X_1, \ldots, X_n, P)$. The bootstrap presupposes the existence of an estimate $\hat{P}(X_1, \ldots, X_n)$ and involves estimating $\mathbf{L}_P$ by the distribution $\mathbf{L}_n^*$ of $T_n(X_1^*, \ldots, X_n^*, \hat{P})$, where $(X_1^*, \ldots, X_n^*)$ is distributed according to $\hat{P}$. The method is of particular interest when L_n^*, though known in principle, can realistically only be computed by simulation. Such computation can be expensive if n is large and T_n is complex (e.g., see the multivariate goodness-of-fit tests of Beran and Millar 1986). Even when bootstrap application to a single data set is not excessively expensive, Monte Carlo studies of the bootstrap are another matter. We propose a method based on the classical ideas of Richardson extrapolation for reducing the computational cost inherent in bootstrap simulations and Monte Carlo studies of the bootstrap, by performing simulations for statistics based on two smaller sample sizes. We study theoretically which ratio of the two small sample sizes is apt to give best results. We show how our method works for approximating the χ^2, t, and smoothed binomial distributions, and for setting bootstrap percentile confidence intervals for the variance of a normal distribution with a mean of 0.

KEY WORDS: Cost of computation; Edgeworth expansion; Approximation.

1. INTRODUCTION

Let L_n^* be the bootstrap distribution of $T_n(X_1, \ldots, X_n, P)$. With knowledge of particular features of L_n^*, various devices such as importance sampling can reduce the number r of Monte Carlo replications needed to compute (or rather estimate) L_n^* closely. The total computation cost for a simulation is proportional to $c(n)r$, where $c(n)$, the cost of computing T_n, usually rises at least linearly with n (and often faster). In this article we explore a way of reducing $c(n)$ rather than r. Suppose that T_n is univariate, and let F_n^* be the distribution function of L_n^*. For most T_n of interest, it is either known or plausibly conjectured that F_n^* tends to a limit A_0 in probability

$$F_n^*(x) = A_0(x) + o_p(1), \tag{1.1}$$

for all x and often uniformly in x as well. Examples include the usual pivots for parameters $\theta(F)$ when $X_1, \ldots, X_n$ are iid F and $\hat{P} \leftrightarrow \hat{F}$ is the empirical distribution. Thus if $T_n = \sqrt{n}(\theta(\hat{F}) - \theta(F))$, then $A_0 = \mathbf{N}(0, \sigma^2(F))$, under mild conditions; if $T_n = \sqrt{n}[(\theta(\hat{F}) - \theta(F))/\sigma(\hat{F})]$, then $A_0 = \mathbf{N}(0, 1)$. The value A_0 can also be known to exist but not be readily computable. For example, let $T_n = \sqrt{n}\sup_x|\hat{F}(x) - F(x)|$, with F possibly discrete (see Bickel and Freedman 1981). Furthermore, an asymptotic expansion in powers of $n^{-1/2}$ is known to be true in some cases and reasonably conjectured in many others. That is,

$$F_n^*(x) = A_0(x) + \sum_{j=1}^{k} n^{-j/2}A_j(x) + O_P(n^{[-(k+1)]/2}). \tag{1.2}$$

The most important special cases arise when A_0 is normal and the expansion (1.2) is of Edgeworth type. Such expansions appear in the bootstrap context in works by Singh (1981), Bickel and Freedman (1981), and Abramovitch and Singh (1985), among others. Expansions for the distributions F_n of $T_n(X_1, \ldots, X_n)$ under fixed F have been studied extensively (e.g., see Bhattacharya and Ranga Rao 1976).

In this context, we propose to calculate $F_{n_1}, \ldots, F_{n_{k+1}}$, where

$$n_1 + \cdots + n_{k+1} = b \ll n. \tag{1.3}$$

We use the F_{n_j} to approximate F_n. This procedure is classically used in numerical analysis (where it is called Richardson extrapolation) to approximate F_∞. Our application of these ideas differs, in that

1. We are interested in F_n, not F_∞.
2. F_∞ is sometimes known, as in the Edgeworth case, and can be used to improve the approximation.
3. We are interested in the design problem of selecting the n_j, subject to the budget constraint (1.3).

Using our method in the bootstrap context involves simply putting $*$ on the F_{n_j} and F_n. In Section 2 we develop the method in detail and give explicit solutions to three formulations of the design problem for $k = 1$. Finally, in Section 3 we test the method on approximations of known F_n, as well as some bootstrap examples. The results are very encouraging.

2. EXTRAPOLATION

Throughout this section IK refers to Isaacson and Keller (1966). Write $t = n^{-1/2}$ $(0 < t \leq 1)$. Given a sequence of distribution functions $F_n \triangleq G_t$, write

$$G_t = P_t + \Delta_t, \qquad P_t = A_0 + \sum_{j=1}^{k} t^j A_j. \tag{2.1}$$

The argument in the functions G_t and A_j plays no role in our discussion and is omitted. We calculate $G_{t_0}, \ldots,$

* Peter J. Bickel is Professor, Department of Statistics, University of California, Berkeley, CA 94720. Joseph A. Yahav is Professor, Department of Statistics, Hebrew University, Jerusalem, Israel. This work was partially supported by Office of Naval Research Contract N00014-80-C0163. The authors thank Persi Diaconis for a reference to Kuipers and Niederreiter (1974), which they used to obtain a considerable simplification of the original proof of the theorem in the Appendix, and Adele Cutler, for the programming of the simulations and other calculations in Section 3.

© 1988 American Statistical Association
Journal of the American Statistical Association
June 1988, Vol. 83, No. 402, Theory and Methods

G_{t_k} $(t < t_0 < \cdots < t_k)$. If $\Delta_t = 0$ for $t, t_0, \ldots, t_k$ we obtain G_t perfectly from the G_{t_j} by using the Lagrange interpolating polynomial (IK, p. 188):

$$\hat{G}_t = \sum_{j=0}^{k} G_{t_j}\phi_{k,j}(t), \qquad \phi_{k,j}(t) = \prod_{i \neq j} [(t - t_i)/(t_j - t_i)]. \tag{2.2}$$

In particular for the only case we study in detail, $k = 1$,

$$\hat{G}_t = (t_1 - t_0)^{-1}[(t_1 - t)G_{t_0} + (t - t_0)G_{t_1}]. \tag{2.3}$$

We consider three classes for Δ, depending on a parameter M:

1. $\mathbf{D}_1 = \{\Delta: d^{k+1}\Delta_t/dt^{k+1}$ exists and $\sup_t|(d^{k+1}\Delta_t)/dt^{k+1}| \leq M\}$. Since Δ is only defined at the points $n^{-1/2}$ $(n = 1, 2, \ldots)$ we interpret $\Delta \in \mathbf{D}_1$ as applying to some smooth function agreeing with Δ at all points $n^{-1/2}$. The other two classes make no smoothness assumptions on Δ.
2. $\mathbf{D}_2 = \{\Delta: \sup_t t^{-(k+1)}|\Delta_t| \leq M\}$.
3. $\mathbf{D}_3 = \{\Delta: 0 \leq t^{-(k+1)}\Delta_t \leq M$ for all $t > 0$, or $-M \leq t^{-(k+1)}\Delta_t \leq 0$ for all $t > 0\}$.

For fixed $t, t_0, \ldots, t_k$ we define the error of approximation by

$$E_i(t, t_0, \ldots, t_k) = \sup\{|\hat{G}_t - G_t|: \Delta \in D_i\}, \quad 1 \leq i \leq 3.$$

We want to minimize E_i, subject to a fixed budget b, where

$$\sum_{j=0}^{k} t_j^{-2} = b. \tag{2.4}$$

If t_j satisfy (2.4) and $b \to \infty$, then $t_0 \to 0$.

We claim that

$$E_1 \sim \frac{M}{(k+1)!}\prod_{j=0}^{k}(t_j - t), \tag{2.5}$$

$$E_2 \sim M\left\{\sum_{j=0}^{k}|\phi_{k,j}(t)|t_j^{k+1} + t^{k+1}\right\}, \tag{2.6}$$

and

$$E_3 \sim M\left\{\left[\sum_{j=0}^{k}[\phi_{k,j}(t)]_+ t_j^{k+1}\right] \vee \left[\sum_{j=0}^{k}[\phi_{k,j}(t)]_- t_j^{k+1}\right] + t^{k+1}\right\}, \tag{2.7}$$

where $a_+ = a \vee 0$ and $a_- = -(a \wedge 0)$. To check (2.5), apply theorem 1 of IK (p. 90), which has

$$G_t - \hat{G}_t = [(k+1)!]^{-1}\prod_{i=0}^{k}(t - t_i)\frac{d^{k+1}G_t}{dt^{k+1}}(\xi), \tag{2.8}$$

where $t < \xi < t_k$. Note that $(d^{k+1}/dt^{k+1})P_t = 0$. To check (2.6) and (2.7), note that interpolation is linear, so $\hat{G}_t = \hat{P}_t + \hat{\Delta}_t$. Since $P_t = \hat{P}_t$, we have $G_t - \hat{G}_t = \Delta_t - \hat{\Delta}_t$; (2.6) and (2.7) follow from (2.2). From (2.5), E_1 is minimized subject to (2.3) as $b \to 0$ by

$$t_0 = \cdots = t_k = \sqrt{(k+1)/b}. \tag{2.9}$$

must be distinct. Nevertheless, if the error term Δ is sufficiently smooth, the n_j should be chosen as nearly equal to each other as possible.

This procedure is analogous to that of the "leave-one-out" jackknife process. This conclusion is clearly valid not just under (2.4), but under any reasonable symmetric-side condition on $t_0, \ldots, t_k$. If we suppose that $t = o(t_0)$, that is, the budget is much smaller than n, we can simplify (2.6) to

$$E_2 \sim M\left(\prod_{j=0}^{k} t_j\right)\sum_{j=0}^{k} t_j^k\left[\prod_{i<j}(t_j - t_i)\prod_{i>j}(t_i - t_j)\right]^{-1} \tag{2.10}$$

and (2.7) to

$$E_3 \sim M\left(\prod_{j=0}^{k} t_j\right)\sum_{j=0}^{k} t_j^k \times \min\left\{\left[\prod_{i \neq j}(t_j - t_i)\right]_+, \left[\prod_{i \neq j}(t_j - t_i)\right]_-\right\}. \tag{2.11}$$

Evidently, (2.10) is minimized asymptotically by $t_j^{-2} = \lambda_j^2 b$, where $\lambda_j > 0$,

$$\sum_{j=0}^{k}\lambda_j^2 = 1, \tag{2.12}$$

and $\lambda_0, \ldots, \lambda_k$ minimize

$$\left(\prod_{j=0}^{k}\lambda_j\right)^{-1}\sum_{j=0}^{k}\left[\lambda_j\prod_{i<j}(\lambda_i - \lambda_j)\prod_{i>j}(\lambda_j - \lambda_i)\right]^{-1}, \tag{2.13}$$

subject to (2.12). In principle, this minimization can be carried out for any k. The explicit solutions for the cases we are primarily concerned with, E_2 and E_3 for $k = 1$, are as follows (if we ignore the restriction that the $\lambda_j^2 b$ are integers): For E_2,

$$\lambda_0^2 = 1 - \lambda_1^2 = .89, \tag{2.14}$$

or more specifically $\lambda_0 = \cos[\frac{1}{2}(\sin^{-1}(1/\omega_0))]$, where $\omega_0 = (1 + \sqrt{5})/2 = 1.6180$ is the unique positive root of $\omega^3 - 2\omega - 1 = 0$. To see this note that for $k = 1$, (2.13) is simply $(\lambda_0\lambda_1)^{-1}(\lambda_0 - \lambda_1)^{-1}(\lambda_0 + \lambda_1)$. Substitute $\lambda_0 = \cos\theta$ to get the objective,

$$2(1 + \sin 2\theta)(\cos 2\theta \sin 2\theta)^{-1},$$

and then substitute $\sin 2\theta = (1 - v^2)^{1/2} = 1/\omega$. Similarly, for $k = 1$, $E_3 \sim M[t_0t_1^2/(t_1 - t_0)]$; a similar minimization gives

$$\lambda_0^2 = \tfrac{1}{2}(1 + (1/\sqrt{2})) = .85. \tag{2.15}$$

In all of these cases, $E_j = O(b^{-(k+1)/2})$.

We check our approach in the following examples of $\{F_n\}$, belonging to $\mathbf{D}_1$ and $\mathbf{D}_3$, respectively.

Example 1: The Gamma Family. Let F_n be the distribution of $(S_n - n)(2n)^{-1/2}$, where S_n has the χ_n^2 distri-

bution. Evidently, we can define G_t for $t > 0$ with

$$G_t(x) = \Gamma^{-1}(\nu^{-1})\lambda^{\nu}\int_0^{x_t} e^{-\lambda s}s^{\nu-1}\,ds \qquad (2.16)$$

where $x_t = x\nu^{1/2} + (\nu/\lambda)$, $\nu = 2t^{-2}$, and $\lambda = \frac{1}{2}$. Using standard Stirling expansions for Γ and its derivatives, it is easy to show that

$$G_t(x) = \frac{e^{-\nu}\nu^{\nu-1/2}}{\Gamma(\nu)}\int_{-\sqrt{\nu}}^{x} \times [(\exp(-u\nu^{-1/2}))(1 + u\nu^{-1/2})]^{\nu}(1 + u\nu^{-1/2})^{-1}\,du,$$

that $A_0 = \Phi$, the standard normal distribution, and that G_t has bounded derivatives of all orders in t. Thus $\Delta \in \mathbf{D}_1$ for all k. Evidently, taking $\lambda = \frac{1}{2}$ plays no role, and this observation applies to the standardized gamma family in general.

Example 2: The Binomial Distribution With Continuity Correction. Let F_n be the distribution of $(S_n - np)/(npq)^{1/2}$ convoluted with the uniform distribution on

$$[-1/(2\sqrt{npq}),\ 1/(2\sqrt{npq})],$$

where S_n has a binomial (n, p) distribution $q = 1 - p$ $(0 < p < 1)$. It is well known that F_n is of the form $F_n(x) = \Phi(x) + n^{-1/2}A_1(x) + O(n^{-1})$ (e.g., see Feller 1971, p. 540). But if we analyze the remainder term further, by theorem 23.1 of Bhattacharya and Ranga Rao (1971, p. 238) it is of the form

$$F_n(x) - \Phi(x) - n^{-1/2}A_1(x) = n^{-1}\left[\int_{-1/2}^{1/2} uS_1(np + x\sigma\sqrt{n} - u)\,du\right] \times P(x, \sigma) + o(n^{-1}), \qquad (2.17)$$

where $\sigma = (pq)^{1/2}$, $S_1(t) = t - \frac{1}{2}$ $(0 < t < 1)$, and $S_1(t + 1) = S_1(t)$. Check that

$$\int_{-1/2}^{1/2} uS_1(v - u)\,du = -\frac{S_1^2}{2}\left(x + \frac{1}{2}\right). \qquad (2.18)$$

Unless $x = 0$ and p is rational, the sequence $S_1(np + \sqrt{n}\sigma x + \frac{1}{2})$ is uniformly distributed modulo 1; that is, $\#\{h : S_1(np + \sqrt{n}\sigma x + \frac{1}{2}) \le t,\ n \le N\}/N \to t + \frac{1}{2}$ as $N \to \infty$ if $(-\frac{1}{2} < t < \frac{1}{2})$. A proof is given in the Appendix.

Thus as $n \to \infty$ the coefficient of n^{-1} in (2.17) ranges over an interval $[0, \frac{1}{8}]$ or $[-\frac{1}{8}, 0]$, and comes arbitrarily close to all values in the interval. Hence, $\{F_n\}$ belongs to $\mathbf{D}_3$ for $k = 1$.

Notes. In many examples (including the two we have discussed) A_0 is known. Then, if (2.1) holds for $k = r + 1$, we can improve our estimate using only k sample sizes and still have an error $O(b^{-(r+1)/2})$. We define $Q_t = (G_t - A_0)/t$ and use the estimate $G_t^* = A_0 + t\hat{Q}_t$, where $\hat{Q}_t$ is defined by (2.2), with $k = r$. In particular, for $r = 1$ the allocations (2.9), (2.14), and (2.15) give errors $O(b^{-3/2})$. In the next section we study this approximation by simulation as well.

In some cases such as F_n, the t distribution with n degrees of freedom, the series is in powers of n^{-1}. In this case it is easy to obtain the optimal choice of t_0/t_1 for $\mathbf{D}_2$ and $\mathbf{D}_3$, that is, for (2.4) replaced by $t_0^{-1} + t_1^{-1} = b$. We find for $\mathbf{D}_2$

$$n_j = \rho_j b,\quad \rho_1 = 1 - \rho_0,\quad \rho_0 = .5(1 + \sqrt{3}) = .79, \qquad (2.19)$$

and for $\mathbf{D}_3$

$$\rho_0 = .75. \qquad (2.20)$$

If (as is usually the case in applications) the A_j and t are unknown, it would seem safer to use the approximation for $t = n^{-1/2}$.

An undesirable feature of our approach is that no a posteriori estimate of the error actually incurred is available. If t_1 is small and $\Delta \in \mathbf{D}_1$, we can get an estimate by increasing our budget. We add $\tilde{t}^{-2} \ne t_j^{-2}$ $(j = 0, 1)$ units and calculate $G_{\tilde{t}}$. Now, by (2.8),

$$G_u - \hat{G}_u = \tfrac{1}{2}(d^2\Delta/dt^2)(\xi)(t_1 - t_0)^{-1}(u - t_0)(u - t_1), \qquad (2.21)$$

where $t < \xi < t_1$ for any $t \le u \le t_1$. If t_1 is small we expect the coefficient $d^2\Delta/dt^2$ in (2.21) to be stable, so we obtain

$$|G_t - \hat{G}_t| \propto |(t - t_0)(t - t_1)(s - t_0)^{-1}(s - t_1)^{-1}| \times |G_{\tilde{t}} - \hat{G}_{\tilde{t}}|. \qquad (2.22)$$

If $\Delta \in \mathbf{D}_2$ or $\mathbf{D}_3$, no realistic estimate of the error presents itself. Suppose, however (as may be seen in Ex. 2), that if $0 < \lambda_1 < \cdots < \lambda_k < 1$, $a_1, \ldots, a_k$ are real, $n \to \infty$, and $s_j = [\lambda_j n]^{-1/2}$, then

$$\#(\Delta_{s_j} \le s_j^2 a_j : 1 \le j \le k)/n \to \prod_{j=1}^{k} G(a_j). \qquad (2.23)$$

That is, $s_1^{-2}\Delta_{s_1}, \ldots, s_k^{-2}\Delta_{s_k}$ are asymptotically independently distributed with common distribution G. This is, of course, a poor approximation if λ_j and λ_{j+1} are too close and we cannot use (2.23) for design. But if we increase our budget we can calculate G at $l \ge 3$ points $t_0, t_1, \ldots, t_l$, with $l \ge 2$. If we assume (2.1), it is natural to consider the estimate

$$\hat{G}_t^l = \hat{A}_0^l + t\hat{A}_1^l, \qquad (2.24)$$

where $\hat{A}_0^l$ and $\hat{A}_1^l$ are the weighted fixed least squares estimates of A_0 and A_1,

$$\hat{A}_1^l = \sum_{i=0}^{l}(t_i - \bar{t})G_{t_i}\sigma_i^{-2}\Big/\sum_{i=0}^{l}(t_i - \bar{t})^2\sigma_i^{-2}, \qquad (2.25)$$

and

$$\hat{A}_0^l = \sum_{i=0}^{l} G_{t_i}\frac{\sigma_i^{-2}}{W} - \hat{A}_1^l\bar{t}, \qquad (2.26)$$

Table 1. Richardson Extrapolation for χ^2_n

n_0, n_1	Percentiles 10	90	95	99
$n = 50$				
True values	−1.2311	1.3167	1.7505	2.6154
Fisher approximation	−1.1995	1.3637	1.7802	2.5969
	(.0317)	(.0470)	(.0297)	(−.0185)
	$n_0 + n_1 = 15$			
1, 14	−1.2557	1.3572	1.7995	2.6686
	(−.0246)	(.0404)	(.0490)	(.0532)
3, 12	−1.2528	1.3417	1.7796	2.6446
	(−.0217)	(.0250)	(.0291)	(.0292)
4, 11	−1.2511	1.3391	1.7764	2.6410
	(−.0199)	(.0224)	(.0259)	(.0256)
6, 9	−1.2493	1.3367	1.7735	2.6377
	(−.0182)	(.0200)	(.0230)	(.0223)
	$n_0 + n_1 = 20$			
2, 18	−1.2481	1.3374	1.7747	2.6400
	(−.0169)	(.0207)	(.0242)	(.0246)
4, 16	−1.2448	1.3319	1.7680	2.6324
	(−.0137)	(.0152)	(.0175)	(.0170)
5, 15	−1.2439	1.3307	1.7665	2.6307
	(−.0127)	(.0139)	(.0160)	(.0153)
8, 12	−1.2424	1.3289	1.7644	2.6258
	(−.0113)	(.0122)	(.0139)	(.0131)
$n = 100$				
True values	−1.2475	1.3080	1.7212	2.5319
Fisher approximation	−1.2235	1.3397	1.7406	2.5176
	(.0239)	(.0317)	(.0193)	(−.0143)
	$n_0 + n_1 = 20$			
2, 18	−1.2740	1.3399	1.7585	2.5694
	(−.0265)	(.0319)	(.0373)	(.0374)
4, 16	−1.2687	1.3314	1.7481	2.5576
	(−.0212)	(.0234)	(.0268)	(.0257)
5, 15	−1.2671	1.3294	1.7457	2.5550
	(−.0197)	(.0214)	(.0245)	(.0231)
8, 12	−1.2649	1.3267	1.7424	2.5515
	(−.0174)	(.0187)	(.0212)	(.0196)
	$n_0 + n_1 = 30$			
3, 27	−1.2628	1.3252	1.7410	2.5508
	(−.0154)	(.0172)	(.0197)	(.0189)
6, 24	−1.2592	1.3205	1.7354	2.5448
	(−.0117)	(.0125)	(.0142)	(.0129)
7, 23	−1.2585	1.3198	1.7345	2.5439
	(−.0110)	(.0117)	(.0133)	(.0120)
12, 18	−1.2569	1.3180	1.7324	2.5418
	(−.0095)	(.0100)	(.0112)	(.0099)

where $\sigma_i = t_i^2$, $W = \sum_{i=0}^{l} \sigma_i^{-2}$, and $\hat{t} = \sum_{i=0}^{l} t_i \sigma_i^{-2}/W$. The error, $G_t - \hat{G}_t^l$, can be estimated by

$$\left\{ W^{-1} + t^2[\Sigma(t_i - \hat{t})^2\sigma_i^{-2}]^{-1} \times \sum_{i=0}^{l} (G_{t_i} - \hat{A}_0^l - \hat{A}_1^l t_i)^2 \sigma_i^{-2} \right\}^{1/2}. \quad (2.27)$$

The range of validity of the approximations (2.22) and (2.27) needs to be investigated by simulation.

3. COMPUTATION AND SIMULATION

In this section we study the actual performance of the approximations in the Section 2 examples. We also study the performance of the approximation for the Student-t distribution, where the expansion is in powers of $1/n$. Finally, we provide the results of a bootstrap simulation, where we compare the operating characteristics of confidence bounds based on a Richardson extrapolation approximation with those based on a full bootstrap.

χ^2_n Approximation. We computed the Richardson extrapolation for $[\chi^2_n(\alpha) - n]/(2n)^{1/2}$ ($\alpha = 10\%, 90\%, 95\%, 99\%$), where $\chi^2_n(\alpha)$ is the αth percentile of the χ^2_n distribution, and compared it with the Fisher square-root approximation applied to the quantiles:

$$[\chi^2_n(\alpha) - n]/\sqrt{2n} \propto Z(\alpha) + [Z^2(\alpha)]/2\sqrt{2n},$$

where $Z(\alpha)$ is the standard normal α percentile. We used $n = 50, 100$, $b = 15, 20, 30$, and $1 - \lambda = n_0/b = .1, .2, .25, .40$, where $n_0 < n_1$ and $n_0 + n_1 = b$. Note the following:

1. The approximation improves as b and n increase.
2. The allocation $\lambda = .6$ is best, as expected.
3. For $n_0 + n_1 = 15$, 20, and all λ, the Richardson extrapolation is essentially as good as Fisher's approximation for the .9 and .1 percentiles, and still gives the same two significant figures as Fisher's for the .95 and .99 percentiles.
4. For $n_0 + n_1 = 30$ it is better in all cases save one, where the results are virtually equivalent. The $\lambda = .6$ allocation seems to give nearly three significant figures (see Table 1).

Table 2. Richardson Extrapolation for χ^2_{50} Knowing the Limit

n_0, n_1	Percentiles 10	90	95	99
$n = 50$				
True values	−1.2311	1.3167	1.7505	2.6154
Fisher approximation	−1.1995	1.3637	1.7802	2.5969
	(.0317)	(.0470)	(.0297)	(−.0185)
	$n_0 + n_1 = 15$			
1, 14	−1.2289	1.3165	1.7510	2.6178
	(.0022)	(−.0002)	(.0005)	(.0024)
3, 12	−1.2306	1.3168	1.7510	2.6172
	(.0006)	(.0001)	(.0005)	(.0018)
4, 11	−1.2307	1.3168	1.7509	2.6171
	(.0004)	(.0001)	(.0005)	(.0017)
6, 9	−1.2308	1.3168	1.7509	2.6169
	(.0003)	(.0001)	(.0004)	(.0016)
	$n_0 + n_1 = 20$			
2, 18	−1.2305	1.3167	1.7509	2.6168
	(.0006)	(.0000)	(.0004)	(.0015)
4, 16	−1.2309	1.3168	1.7508	2.6166
	(.0002)	(.0001)	(.0003)	(.0012)
5, 15	−1.2309	1.3168	1.7508	2.6165
	(.0002)	(.0001)	(.0003)	(.0011)
8, 12	−1.2310	1.3168	1.7508	2.6164
	(.0001)	(.0001)	(.0003)	(.0010)
	$n_0 + n_1 = 30$			
3, 27	−1.2310	1.3167	1.7507	2.6161
	(.0002)	(.0000)	(.0002)	(.0007)
6, 24	−1.2311	1.3167	1.7506	2.6159
	(.0001)	(.0000)	(.0002)	(.0005)
7, 23	−1.2311	1.3167	1.7506	2.6159
	(.0001)	(.0000)	(.0001)	(.0005)
12, 18	−1.2311	1.3167	1.7506	2.6159
	(.0000)	(.0000)	(.0001)	(.0005)

Table 3. Richardson Extrapolation for the t Distribution

n_0, n_1	Percentiles 10	90	95	99
$n = 50$				
True values	−1.2987	1.2987	1.6759	2.4033
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(.0171)	(−.0171)	(−.0310)	(−.0770)
	$n_0 + n_1 = 15$			
3, 12	−1.2849	1.2849	1.6376	2.2099
	(.0138)	(−.0138)	(−.0383)	(−.1934)
4, 11	−1.2878	1.2878	1.6462	2.2595
	(.0110)	(−.0110)	(−.0298)	(−.1438)
6, 9	−1.2900	1.2900	1.6526	2.2947
	(.0087)	(−.0087)	(−.0233)	(−.1086)
	$n_0 + n_1 = 20$			
4, 16	−1.2922	1.2922	1.6584	2.3198
	(.0065)	(−.0065)	(−.0175)	(−.0835)
5, 15	−1.2933	1.2933	1.6614	2.3357
	(.0055)	(−.0055)	(−.0146)	(−.0676)
8, 12	−1.2945	1.2945	1.6649	2.3535
	(.0042)	(−.0042)	(−.0111)	(−.0498)
$n = 100$				
True values	−1.2901	1.2901	1.6602	2.3642
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(.0085)	(−.0085)	(−.0153)	(−.0379)
	$n_0 + n_1 = 20$			
4, 16	−1.2818	1.2818	1.6378	2.2577
	(.0083)	(−.0083)	(−.0224)	(−.1065)
5, 15	−1.2831	1.2831	1.6417	2.2785
	(.0070)	(−.0070)	(−.0185)	(−.0857)
8, 12	−1.2848	1.2848	1.6463	2.3018
	(.0053)	(−.0053)	(−.0139)	(−.0624)
	$n_0 + n_1 = 30$			
6, 24	−1.2869	1.2869	1.6520	2.3274
	(.0031)	(−.0031)	(−.0082)	(−.0368)
7, 23	−1.2873	1.2873	1.6530	2.3321
	(.0028)	(−.0028)	(−.0072)	(−.0321)
12, 18	−1.2880	1.2880	1.6550	2.3414
	(.0020)	(−.0020)	(−.0053)	(−.0228)

In Table 2 we exhibit the Richardson extrapolation results for the χ^2_n distribution, using the knowledge of the limit as $n \to \infty$ (see Sec. 2).That is, we use the expansion

$$[\chi^2_n(\alpha) - n]/\sqrt{2n} = Z(\alpha) + A_1(1/\sqrt{n}) + A_2\frac{1}{n} + o_P\left(\frac{1}{n}\right)$$

or

$$\sqrt{n}\,\{[\chi^2_n(\alpha) - n]/\sqrt{2n} - Z(\alpha)\} = A_1 + A_2(1/\sqrt{n}) + o_P(1/\sqrt{n}),$$

where $Z(\alpha)$ is the α percentile of the standard normal. A_1 and A_2 are estimated using $\chi^2_{n_0}$ and $\chi^2_{n_1}$. The results are extremely good for both $n = 50$ and $n = 100$ (omitted here). The extrapolation, even for $n_0 + n_1 = 15$ and $\lambda = .9$, gives three significant figures for all percentiles. For $n_0 + n_1 = 30$, it often gives five significant figures.

The Student-t distribution has an expansion in powers of $1/n$. The Richardson extrapolation (2.3) with $1/\sqrt{n}$ gave no improvement over the ordinary normal approximation, as expected. In Table 3 we present the Richardson extrapolation to the t distribution and compare these results with the normal approximation. We looked at the same values of n, b, λ, and α for approximation to $t_n(\alpha)$, the αth percentile of the t distribution with n degrees of freedom. For $\lambda = .6$ and $b = 30$, the approximation is valid to 3 significant figures for $n = 100$ in all but one case, and improves on the normal approximation.

Tables 4 and 5 give the Richardson extrapolation for the continuity-corrected binomial distribution. That is, we define

$$B_n(s) = \sum_{k=0}^{[s]} \binom{n}{k} p^k(1-p)^{n-k} + (s - [s])\binom{n}{[s]+1} p^{[s]+1}(1-p)^{n-1-[s]}$$

Table 4. Richardson Extrapolation for the Binomial Distribution With p = .2

n_0, n_1	Percentiles 10	90	95	99
$n = 50$				
True values	−1.2591	1.3125	1.7177	2.4900
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(−.0225)	(−.0309)	(−.0728)	(−.1637)
	$n_0 + n_1 = 15$			
1, 14	−1.2689	1.2591	1.6071	2.4969
	(−.0097)	(−.0533)	(−.1106)	(.0068)
3, 12	−1.2392	1.3702	1.6743	2.6561
	(.0199)	(.0577)	(−.0434)	(.1661)
4, 11	−1.1692	1.2861	1.5821	2.3349
	(.0900)	(−.0264)	(−.1356)	(−.1551)
6, 9	−1.1679	1.4169	1.6882	2.5377
	(.0913)	(.1044)	(−.0295)	(.0477)
	$n_0 + n_1 = 20$			
2, 18	−1.2182	1.3060	1.6984	2.4728
	(.0409)	(−.0065)	(−.0193)	(−.0172)
4, 16	−1.2751	1.2595	1.7357	2.4304
	(−.0160)	(−.0530)	(.0180)	(−.0597)
5, 15	−1.2724	1.3224	1.7362	2.5814
	(−.0133)	(.0099)	(.0185)	(.0914)
8, 12	−1.1082	1.2538	1.8704	2.8400
	(.1509)	(−.0587)	(.1527)	(.3500)
$n = 100$				
True values	−1.2733	1.3036	1.6922	2.4351
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(−.0083)	(−.0220)	(−.0473)	(−.1088)
	$n_0 + n_1 = 20$			
2, 18	−1.2111	1.2835	1.6845	2.4144
	(.0622)	(−.0202)	(−.0078)	(−.0207)
4, 16	−1.2699	1.2280	1.7121	2.3686
	(.0033)	(−.0757)	(.0199)	(−.0665)
5, 15	−1.2638	1.3054	1.7208	2.5622
	(.0094)	(.0018)	(.0286)	(.1271)
8, 12	−1.0651	1.2220	1.8840	2.8864
	(.2082)	(−.0816)	(.1918)	(.4513)
	$n_0 + n_1 = 30$			
3, 27	−1.2644	1.3313	1.6692	2.4688
	(.0089)	(.0277)	(−.0231)	(.0337)
6, 24	−1.2233	1.3226	1.6628	2.4172
	(.0500)	(.0190)	(−.0294)	(−.0179)
7, 23	−1.2386	1.2849	1.7134	2.3774
	(.0347)	(−.0187)	(.0211)	(−.0577)
12, 18	−1.1641	1.3332	1.4957	2.4281
	(.1092)	(.0296)	(−.1966)	(−.0070)

Table 5. Richardson Extrapolation for the Binomial Distribution With $p = .4$

n_0, n_1	Percentiles 10	90	95	99
$n = 50$				
True values	−1.2776	1.2882	1.6694	2.3720
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(−.0040)	(−.0066)	(−.0245)	(−.0457)
	$n_0 + n_1 = 15$			
1, 14	−1.2692	1.2656	1.6423	2.4690
	(.0084)	(−.0226)	(−.0271)	(.0971)
3, 12	−1.1991	1.3197	1.6443	2.3799
	(.0785)	(.0315)	(−.0252)	(.0079)
4, 11	−1.2562	1.1647	1.6847	2.2989
	(.0214)	(−.1235)	(.0153)	(−.0731)
6, 9	−1.2007	1.0239	1.8705	2.4680
	(.0770)	(−.2643)	(.2010)	(.0961)
	$n_0 + n_1 = 20$			
2, 18	−1.2344	1.2776	1.6229	2.4184
	(.0432)	(−.0105)	(−.0466)	(.0464)
4, 16	−1.2823	1.2781	1.6526	2.3583
	(−.0047)	(−.0101)	(−.0168)	(−.0137)
5, 15	−1.2702	1.2816	1.6411	2.3911
	(.0074)	(−.0066)	(−.0283)	(.0191)
8, 12	−1.3054	1.2832	1.7722	2.5967
	(−.0278)	(−.0049)	(.1027)	(.2247)
$n = 100$				
True values	−1.2811	1.2892	1.6619	2.3475
Normal approximation	−1.2816	1.2816	1.6449	2.3263
	(−.0005)	(−.0076)	(−.0170)	(−.0212)
	$n_0 + n_1 = 20$			
2, 18	−1.2167	1.2576	1.5951	2.4224
	(.0644)	(−.0316)	(−.0668)	(.0749)
4, 16	−1.2738	1.2589	1.6383	2.3349
	(.0073)	(−.0303)	(−.0236)	(−.0126)
5, 15	−1.2674	1.2767	1.6183	2.4029
	(.0138)	(−.0125)	(−.0436)	(.0554)
8, 12	−1.3107	1.2668	1.7943	2.6437
	(−.0295)	(−.0224)	(.1324)	(.2962)
	$n_0 + n_1 = 30$			
3, 27	−1.2317	1.2922	1.6644	2.3581
	(.0494)	(.0030)	(.0025)	(.0106)
6, 24	−1.2379	1.2613	1.6580	2.3519
	(.0432)	(−.0279)	(−.0039)	(.0043)
7, 23	−1.2601	1.3154	1.6403	2.3348
	(.0210)	(.0262)	(−.0216)	(−.0128)
12, 18	−1.2439	1.2762	1.6676	2.3576
	(.0373)	(−.0130)	(.0057)	(.0101)

and

$$Q_n(u) = B_n(np + u\sqrt{np(1-p)}).$$

We approximated the percentiles $Q_n^{-1}(\alpha)$ for n, b, and λ as before, with $p = .2$ and .4. Note that the $\lambda = .75$ allocation seems to work best, but differs little from $\lambda = .8$ and .9. On the other hand, $\lambda = .6$ is poorer. (This is in agreement with our theory for class $\mathbf{D}_3$.) For $p = .2$, $n = 50$, 100, and $b = 15$, 20, the $\lambda = .75$ allocation does as well as the normal. For $b = 20$, 30 it is better, typically giving an additional significant figure. For $p = .4$, it is generally poorer, though far from terrible. This is understandable, since for $p = .5$, $A_1 = 0$, and the extrapolation is adding noise to the normal approximation.

In Table 6 we show the results for the bootstrap experiment. The population is $\sigma^2\chi_1^2$, and we are interested in a confidence bound for σ. We study the unadjusted bootstrap, that is, the percentiles of the bootstrap distribution of $(\bar{X}_n)^{1/2}$, where $\bar{X}$ is the sample mean. For $n = 50$, 100, and 500 we took 500 samples of size n from χ_1^2. For each sample we took 1,000 bootstrap samples and computed the .05, .1, and .95 percentiles of the bootstrap distribution of $(\bar{X}_n)^{1/2}$ for sample size n_0, n_1, and n. We study the behavior of the 90% lower confidence bound and the 90% confidence interval, that is, the .1 percentile and the interval between the .95 and .05 percentiles. This is Efron's (1979) percentile method, which we do not endorse in practice but use as a simple example of the bootstrap.

For each n we count the number of times the population parameter falls inside the confidence set, out of the 500 samples. We compute the average and standard deviation of the rescaled lower bound, that is, $\sqrt{n}(1 - G_n^{*-1}(.1))$, and the rescaled interval, that is, $I_n^*(.9) = \sqrt{n}(G_n^{*-1}(.95) - G_n^{*-1}(.05))$, where $G_n^{*-1}(\alpha)$ is the α percentile of the bootstrap distribution of $\bar{X}^{1/2}$. Table 6 shows clearly that Richardson extrapolation is a good approximation to the full bootstrap and is not very sensitive to the allocation of n_0 and n_1. The last entry gives estimated computation times on Sun workstations at the University of California. The expected linear saving in the sample size is confirmed.

APPENDIX: THEORY FOR EXAMPLE 2

We establish the claim asserted in Example 2 in the form of a theorem.

Theorem. $[an + b\sqrt{n}]$ is uniformly distributed (ud) mod 1 unless $b = 0$ and a is rational.

Proof. We refer repeatedly to the text of Kuipers and Niederreiter (KN 1974). Suppose that a is irrational. Note that

$$a(n+1) + b\sqrt{n+1} - an - b\sqrt{n} = a + b0(n^{-1/2}) \to a,$$

as $n \to \infty$. By theorem 3.3 of KN, $an + b\sqrt{n}$ is ud mod 1.

If a is rational we apply the following lemma.

Lemma. Let b_n be a sequence such that $\{b_{sj+k}\}_{j\geq 1}$ is ud mod 1 for $s \neq 0$ $(0 \leq k \leq s)$. Then if a is rational, $a = r/s$ and $an + b_n$ is ud mod 1.

Proof. Check Weyl's criterion (KN). Let $n = ms$. Then

$$\begin{aligned}\left|\frac{1}{n}\sum_{l=1}^{n}\exp[2\pi ih(a_l + b_l)]\right| &= \left|\frac{1}{ms}\sum_{j=0}^{m-1}\sum_{k=0}^{s-1}\exp\{2\pi ih[r(k/s) + b_{sj+k}]\}\right| \\ &\leq \frac{1}{s}\sum_{k=0}^{s-1}\left|\frac{1}{m}\sum_{j=0}^{m-1}\exp(2\pi ihb_{sj+k})\right| \to 0, \qquad (A.1)\end{aligned}$$

as $m \to \infty$ by Weyl's criterion applied to $\{b_{sj+k}\}_{j\geq 1}$. If $n = ms + b$ $(0 < b < s)$, the difference from (A.1) is at most $b/ms \to 0$. The lemma follows by Weyl's criterion.

Let $b_n = b\sqrt{n}$. If $b > 0$, $b_{s(j+1)+k} - b_{sj+k}$ is decreasing to 0 in j, since $\sqrt{x}$ is concave. Moreover, $j(b_{s(j+1)+k} - b_{sj+k}) = \Omega(j^{1/2})$

Table 6. A Bootstrap Experiment

					Rescaled				
n	n_0	n_1	Lower-bound count	Interval count	Confidence-bound average	Confidence-bound SD	Average length	SD length	Time*
50	full	bootstrap	462	443	.83732	.007231	2.23093	.018770	1,603
50	2	18	455	439	.84251	.007652	2.26337	.019407	680
50	4	16	468	449	.83286	.007090	2.23915	.018084	680
100	full	bootstrap	457	445	.85825	.005957	2.25543	.015018	3,171
100	2	18	459	438	.85736	.006190	2.27469	.014191	688
100	4	16	472	446	.85685	.006639	2.26594	.014139	686
500	full	bootstrap	453	453	.89302	.003029	2.31675	.070418	15,754
500	5	45	454	448	.88568	.004092	2.31589	.009186	1,665
500	10	40	454	455	.89668	.004200	2.33916	.086705	1,666

NOTE: SD represents standard deviation.
* In central-processing-unit seconds.

$\to \infty$. By Fejer's theorem (KN, theorem 2.5), $\{b_{nj+k}\}$ is ud mod 1, and the theorem follows.

[*Received August 1986. Revised September 1987.*]

REFERENCES

Abramovitch, L., and Singh, K. (1985), "Edgeworth Corrected Pivotal Statistics and the Bootstrap," *The Annals of Statistics,* 13, 116–132.

Beran, R., and Millar, P. W. (1986), "Confidence Sets for a Multivariate Distribution," *The Annals of Statistics,* 14, 431–443.

Bhattacharya, R., and Ranga Rao, R. (1976), *Normal Approximation and Asymptotic Expansions,* New York: John Wiley.

Bickel, P. J., and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics,* 9, 1196–1217.

Efron, B. (1979), "Bootstrap Methods: Another Look at Jackknife," *The Annals of Statistics,* 7, 1–26.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications,* New York: John Wiley.

Isaacson, E., and Keller, H. B. (1966), *Analysis of Numerical Methods,* New York: John Wiley.

Kuipers, L., and Niederreiter, H. (1974), *Uniform Distribution or Sequences,* New York: John Wiley.

Singh, K. (1981), "On Asymptotic Accuracy of Efron's Bootstrap," *The Annals of Statistics,* 9, 1187–1195.

Statistica Sinica **7**(1997), 1-31

RESAMPLING FEWER THAN n OBSERVATIONS: GAINS, LOSSES, AND REMEDIES FOR LOSSES

P. J. Bickel, F. Götze and W. R. van Zwet

University of California, Berkeley,
University of Bielefeld and University of Leiden

Abstract: We discuss a number of resampling schemes in which $m = o(n)$ observations are resampled. We review nonparametric bootstrap failure and give results old and new on how the m out of n with replacement and without replacement bootstraps work. We extend work of Bickel and Yahav (1988) to show that m out of n bootstraps can be made second order correct, if the usual nonparametric bootstrap is correct and study how these extrapolation techniques work when the nonparametric bootstrap does not.

Key words and phrases: Asymptotic, bootstrap, nonparametric, parametric, testing.

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 (1994), following Wu (1990), and independently Götze (1993) showed that what we call the m out of n without replacement bootstrap with $m = o(n)$ typically works to first order both in the situations where the bootstrap works and where it does not.

The m out of n with replacement bootstrap with $m = o(n)$ has been known to work in all known realistic examples of bootstrap failure. In this paper,

- We show the large extent to which the Politis, Romano, Götze property is shared by the m out of n with replacement bootstrap and show that the latter has advantages.
- If the usual bootstrap works the m out of n bootstraps pay a price in efficiency. We show how, by the use of extrapolation the price can be avoided.

• We support some of our theory with simulations.

The structure of our paper is as follows. In Section 2 we review a series of examples of success and failure to first order (consistency) of (Efron's) nonparametric bootstrap (nonparametric). We try to isolate at least heuristically some causes of nonparametric bootstrap failure. Our framework here is somewhat novel. In Section 3 we formally introduce the m out of n with and without replacement bootstrap as well as what we call "sample splitting", and establish their first order properties restating the Politis-Romano-Götze result. We relate these approaches to smoothing methods. Section 4 establishes the deficiency of the m out of n bootstrap to higher order if the nonparametric bootstrap works to first order and Section 5 shows how to remedy this deficiency to second order by extrapolation. In Section 6 we study how the improvements of Section 5 behave when the nonparametric bootstrap doesn't work to first order. We present simulations in Section 7 and proofs of our new results in Section 8. The critical issue of choice of m and applications to testing will be addressed elsewhere.

2. Successes and Failure of the Bootstrap

We will limit our work to the i.i.d. case because the issues we discuss are clearest in this context. Extension to the stationary mixing case, as done for the m out of n without replacement bootstrap in Politis and Romano (1994), are possible but the study of higher order properties as in Sections 4 and 5 of our paper is more complicated.

We suppose throughout that we observe $X_1, \ldots, X_n$ taking values in $X = R^p$ (or more generally a separable metric space). i.i.d. according to $F \in \mathcal{F}_0$. We stress that $\mathcal{F}_0$ need not be and usually isn't the set of all possible distributions. In hypothesis testing applications, $\mathcal{F}_0$ is the hypothesized set, in looking at the distributions of extremes, $\mathcal{F}_0$ is the set of populations for which extremes have limiting distributions. 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 $T_n(\cdot, F)$ and hence 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, (NPMLE), $\hat{F}_n$, and hence $\theta_n(F)$ by the plug in $\theta_n(\hat{F}_n)$. This is Efron's (ideal) nonparametric 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 (1989) and Hikley (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 restrict analysis to $T_n(\hat{F}_n, F)$ which are asymptotically stable and nondegenerate on $\mathcal{F}_0$. That is, for all $F \in \mathcal{F}_0$, at least weakly

$$\begin{aligned} \mathcal{L}_n(F) &\to \mathcal{L}(F) \text{ non degenerate} \\ \theta_n(F) &\to \theta(F) \end{aligned} \tag{2.1}$$

as $n \to \infty$.

Using m out of n bootstraps or sample splitting implicitly changes our goal from estimating features of $\mathcal{L}_n(F)$ to features of $\mathcal{L}_m(F)$. This is obviously nonsensical without assuming that the laws converge.

Requiring non degeneracy of the limit law means that we have stabilized the scale of $T_n(\hat{F}_n, F)$. Any functional of $\mathcal{L}_n(F)$ is also a functional of the distribution of $\sigma_n T_n(\hat{F}_n, F)$ where $\sigma_n \to 0$ which also converges in law to point mass at 0. Yet this degenerate limit has no functional $\theta(F)$ of interest.

Finally, requiring that stability need occur only on $\mathcal{F}_0$ is also critical since failure to converge off $\mathcal{F}_0$ in a reasonable way is the first indicator of potential bootstrap failure.

2.1. When does the nonparametric bootstrap fail?

If θ_n does not depend on n, the bootstrap works, (is consistent on $\mathcal{F}_0$), if θ is continuous at all points of $\mathcal{F}_0$ with respect to weak convergence on $\mathcal{F}$. Conversely, the nonparametric bootstrap can fail if,

1. θ is not continuous on $\mathcal{F}_0$.
   An example we explore later is $\theta_n(F) = 1(F \text{ discrete})$ for which $\theta_n(\hat{F}_n)$ obviously fails if F is continuous.
   Dependence on n introduces new phenomena. In particular, here are two other reasons for failure we explore below.
2. θ_n is well defined on all of $\mathcal{F}$ but θ is defined on $\mathcal{F}_0$ only or exhibits wild discontinuities when viewed as a function on $\mathcal{F}$. This is the main point of examples 3-6.
3. $T_n(\hat{F}_n, F)$ is not expressible as or approximable on $\mathcal{F}_0$ by a continuous function of $\sqrt{n}(\hat{F}_n - F)$ viewed as an object weakly converging to a Gaussian limit in a suitable function space. (See Giné and Zinn (1989).) Example 7 illustrate this failure. Again this condition is a diagnostic and not necessary for failure as Example 6 shows.

We illustrate our framework and discuss prototypical examples of bootstrap success and failure.

2.2. Examples of bootstrap success

Example 1. Confidence intervals: 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 $\theta(F) = \Phi(\frac{t}{\sigma(F)})$. 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. It seems silly to be estimating a parameter whose value is known but, of course, interest now centers on $\theta'(F)$ the next higher order term in $\theta_n(F) = \Phi(t) + \frac{\theta'(F)}{\sqrt{n}} + O(n^{-1})$.

Example 2. Estimation of variance: Suppose F has unique median $m(F)$, continuous density $f(m(F)) > 0$, $E_F|X|^\delta < \infty$, some $\delta > 0$ for all $F \in \mathcal{F}_0$ and $\theta_n(F) = \mathrm{Var}_F(\sqrt{n}$ median $(X_1, \ldots, X_n))$. Then $\theta(F) = [4f^2(m(F))]^{-1}$ on $\mathcal{F}_0$.

Note that, whereas θ_n is defined for all empirical distributions F in both examples the limit $\theta(F)$ is 0 or ∞ 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 $F \in \mathcal{F}_0$.

2.3. Examples of bootstrap failure

Example 3. Confidence bounds for an extremum: This is a variation on Bickel Freedman (1981). Suppose that all $F \in \mathcal{F}_0$ have a density f continuous and positive at $F^{-1}(0) > -\infty$. It is natural to base confidence bounds for $F^{-1}(0)$ on the bootstrap distribution of

$$T_n(\hat{F}_n, F) = n(\min_i X_i - F^{-1}(0)).$$

Let

$$\theta_n(F) = P_F[T_n(\hat{F}_n, F) > t] = (1 - F(\frac{t}{n} + F^{-1}(0))^n.$$

Evidently $\theta_n(F) \to \theta(F) = \exp(-f(F^{-1}(0))t)$ on $\mathcal{F}_0$.

The nonparametric bootstrap fails. Let

$$N_n^*(t) = \sum_{i=1}^n 1(X_i^* \le \frac{t}{n} + X_{(1)}), t > 0,$$

where $X_{(1)} \equiv \min_i X_i$ and $1(A)$ is the indicator of A. Given $X_{(1)}$, $n\hat{F}_n(\frac{t}{n} + X_{(1)})$ is distributed as $1+$ binomial $(n-1, \frac{F(\frac{t}{n}+X_{(1)}) - F(X_{(1)})}{(1-F(X_{(1)}))})$ which converges weakly

to a Poisson $(f(F^{-1}(0))t)$ variable. More generally, $n\hat{F}_n(\frac{\cdot}{n} + X_{(1)})$ converges weakly conditionally to $1 + N(\cdot)$, where N is a homogeneous Poisson process with parameter $f(F^{-1}(0))$. It follows that $N_n^*(\cdot)$ converges weakly (marginally) to a process $M(1+N(\cdot))$ where M is a standard Poisson process independent of $N(\cdot)$. 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 $P^*[T_n(\hat{F}_n^*) > t] = P^*[N_n^*(t) = 0]$ converges weakly to the random variable $P[M(1+N(t)) = 0|N] = e^{-(N(t)+1)}$ rather than to the desired $\theta(F)$.

Example 4. Extrema for unbounded distributions: (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 $A_n(F)$, $B_n(F)$,

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

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 $\alpha, \beta \neq 0$. Let,

$$\theta_n(F) \equiv P[(\max(X_1, \ldots, X_n) - A_n(F))/B_n(F) \leq t] \to e^{-H(t,F)} \equiv \theta(F). \quad (2.2)$$

Particular choices of $A_n(F)$, for example, $F^{-1}(1 - \frac{1}{n})$ and $B_n(F)$ are of interest in inference. However, the bootstrap does not work. It is easy to see that

$$n(1 - \hat{F}_n(A_n(F) + tB_n(F))) \xrightarrow{w} N(t), \quad (2.3)$$

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

$$P^*[T_n(\hat{F}_n^*, F) \leq t] \xRightarrow{w} e^{-N(t)}. \quad (2.4)$$

It follows that the nonparametric bootstrap is inconsistent for this choice of A_n, B_n. If it were consistent, then

$$P^*[T_n(\hat{F}_n^*, \hat{F}_n) \leq t] \xrightarrow{P} e^{-H(t,F)} \quad (2.5)$$

for all t and (2.5) would imply that it is possible to find random A real and $B \neq 0$ such that $N(Bt + A) = H(t, F)$ with probability 1. But $H(t, F)$ is continuous except at 1 point. So (2.4) and (2.5) 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, so that θ is not continuous if $\mathcal{F}_0$ is as large as possible.

Essentially the same bootstrap failure arises when we consider estimating the mean of distributions in the domain of attraction of stable laws of index $1 < \alpha \leq 2$. (See Athreya (1987))

Example 5. Testing and improperly centered U and V statistics: (Bretagnolle (1983))

Let $\mathcal{F}_0 = \{F : F[-c,c] = 1, E_F X_1 = 0\}$ and let $T_n(\hat{F}_n) = n\bar{X}^2 = n\int xy d\hat{F}_n(x) d\hat{F}_n(y)$. This is a natural test statistic for $H : F \in \mathcal{F}_0$. Can one use the nonparametric bootstrap to find the critical value for this test statistic? Intuitively, $\hat{F}_n \notin \mathcal{F}_0$ and this procedure is rightly suspect. Nevertheless, in more complicated contexts, it is a mistake made in practice. David Freedman pointed us to Freedman et al. (1994) where the Bureau of the Census appears to have fallen into such a trap. (see Hall and Wilson (1991) for other examples.) The nonparametric bootstrap may, in general, not be used for testing as will be shown in a forthcoming paper.

In this example, due to Bretagnolle (1983), we focus on $\mathcal{F}_0$ for which a general U or V statistic T is degenerate and show that the nonparametric bootstrap doesn't work. More generally, suppose $\psi : R^2 \to R$ is bounded and symmetric and let $\mathcal{F}_0 = \{F : \int \psi(x,y) dF(x) = 0 \text{ for all } y\}$.

Then, it is easy to see that

$$T_n(\hat{F}_n) = \int \psi(x,y) dW_n^0(x) dW_n^0(y), \tag{2.6}$$

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

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

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

$$\begin{aligned} T_n(\hat{F}_n^*) &= n \int \psi(x,y) d\hat{F}_n^*(x) d\hat{F}_n(y) \\ &= \int \psi(x,y) dW_n^*(x) dW_n^{0*}(y) + 2\int \psi(x,y) dW_n^0(x) dW_n^{0*}(y) \\ &\quad + \int \psi(x,y) dW_n^0(x) dW_n^0(y), \end{aligned} \tag{2.7}$$

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

$$\begin{aligned} P^*[T_n(\hat{F}_n^*) \le t] \overset{w}{\Rightarrow} P\Big[&\int \psi(x,y) dW^0(F(x)) dW^0(F(y)) \\ &+2\int \psi(x,y) dW^0(F(x)) d\tilde{W}^0(F(y)) \\ &+ \int \psi(x,y) d\tilde{W}^0(F(x)) d\tilde{W}^0(F(y)) \le t | \tilde{W}^0\Big], \end{aligned} \tag{2.8}$$

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 $\theta_n(F)$ does not converge for $F \notin \mathcal{F}_0$

Example 6. Nondifferentiable functions of the empirical: (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 $\mu(F) \neq 0$, $\mathcal{L}_n(F) \to \mathcal{N}(0, \text{Var}_F(X_1))$ but if $\mu(F) = 0$, $\mathcal{L}_n(F) \to |\mathcal{N}(0, \text{Var}_F(X_1))|$.

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 variables 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 (1985). Dümbgen (1993) constructs similar reasonable though more complicated examples where the bootstrap distribution never converges. If we represent $T_n(\hat{F}_n, F) = \sqrt{n}(T(\hat{F}_n) - T(F))$ in these cases then there is no linear $\dot{T}(F)$ such that $\sqrt{n}(T(\hat{F}_n) - T(F)) \approx \sqrt{n}\dot{T}(F)(\hat{F}_n - F)$ which permits the argument of Bickel-Freedman (1981).

2.4. Possible remedies

Putter and van Zwet (1993) show that if $\theta_n(F)$ is continuous for every n on $\mathcal{F}$ and there is a 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.

If we review our examples of bootstrap failure, we can see that constructing suitable $\tilde{F}_n \in \mathcal{F}_0$ and consistent is often a remedy that works for all $F \in \mathcal{F}_0$ not simply the complement of a set of the second category. Thus in Example 3 taking $\tilde{F}_n$ to be $\hat{F}_n$ kernel smoothed with bandwidth $h_n \to 0$ if $nh_n^2 \to 0$ works. In the first and simplest case of Example 4 it is easy to see, Freedman (1981), that taking $\tilde{F}_n$ as the empirical distribution of $X_i - \bar{X}$, $1 \leq i \leq n$ which has mean 0 and thus belongs to $\mathcal{F}_0$ will work. The appropriate choice of $\tilde{F}_n$ in the other examples of bootstrap failure is less clear. For instance, Example 4 calls for $\tilde{F}_n$ with estimated tails of the right order but how to achieve this is not immediate.

A general approach which we believe is worth investigating is to approximate $\mathcal{F}_0$ by a nested sequence of parametric models, (a sieve), $\{\mathcal{F}_{0,m}\}$, and use the M.L.E. $\tilde{\mathcal{F}}_{m(n)}$ for $\mathcal{F}_{0,m(n)}$, for a suitable sequence $m(n) \to \infty$. See Shen and Wong (1994) for example.

The alternative approach we study is to change θ_n itself as well as possibly its argument. 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 (1994), and what we call sample splitting.

3. The m Out of n Bootstraps

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*)

Let,

$$B_n(F) = E^* h(T_n(\hat{F}_n^*, F)) = n^{-n} \sum_{(i_1,\ldots,i_n)} h(T_n(X_{i_1}, \ldots, X_{i_n}, F)).$$

Then, $B_n \equiv B_n(\hat{F}_n) = \theta_n(\hat{F})$ is the n/n bootstrap.

(ii) *The m/n bootstrap*

Let

$$B_{m,n}(F) \equiv n^{-m} \sum_{(i_1,\ldots,i_m)} h(T_m(X_{i_1}, \ldots, X_{i_m}, F)).$$

Then, $B_{m,n} \equiv B_{m,n}(\hat{F}_n) = \theta_m(\hat{F}_n)$ is the m/n bootstrap.

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

Let

$$J_{m,n}(F) = \binom{n}{m}^{-1} \sum_{i_1<\cdots<i_m} h(T_m(X_{i_1}, \ldots, X_{i_m}, F)).$$

Then, $J_{m,n} \equiv J_{m,n}(\hat{F}_n)$ is the $\binom{n}{m}$ bootstrap.

(iv) *Sample splitting*

Suppose $n = mk$. Define,

$$N_{m,n}(F) \equiv k^{-1} \sum_{j=0}^{k-1} h(T_m(X_{jm+1}, \ldots, X_{(j+1)m}, F))$$

and $N_{m,n} \equiv N_{m,n}(\hat{F}_n)$ as the sample splitting estimates. For safety in practice one should start with a random permutation of the X_i.

The motivation behind $B_{m(n),n}$ for $m(n) \to \infty$ is clear. Since, by (2.1), $\theta_{m(n)}(F) \to \theta(F)$, $\theta_{m(n)}(\hat{F}_n)$ has as good a rationale as $\theta_n(\hat{F}_n)$. To justify $J_{m,n}$ note that we can write $\theta_m(F) = \theta_m(\underbrace{F \times \cdots \times F}_{m})$ since it is a parameter of the

law of $T_m(X_1, \ldots, X_m, F)$. We now approximate $F \times \cdots \times F$ not by the m dimensional product measure $\underbrace{\hat{F}_n \times \cdots \times \hat{F}_n}_{m}$ but by sampling without replacement. Thus sample splitting is just k fold cross validation and represents a crude approximation to $\underbrace{F \times \cdots \times F}_{m}$.

The sample splitting method 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. This method and compromises between it and the $\binom{n}{m}$ bootstrap are studied in Blom (1976) for instance. The $\binom{n}{m}$ bootstrap differs from the m/n by $o_P(1)$ 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. We shall study both of these methods further, below, in terms of their accuracy.

A simple and remarkable result on $J_{m(n),n}$ has been obtained by Politis and Romano (1994), generalizing Wu (1990). This result was also independently noted and generalized by Götze (1993). Here is a version of the Götze result and its easy proof. Write J_m for $J_{m,n}$, B_m for $B_{m,n}$, N_m for $N_{m,n}$.

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

Then,

$$J_m(F) = \theta_m(F) + O_P((\frac{m}{n})^{\frac{1}{2}}). \tag{3.1}$$

If h is continuous and

$$T_m(X_1, \ldots, X_m, F) = T_m(X_1, \ldots, X_m, \hat{F}_n) + o_p(1) \tag{3.2}$$

then

$$J_m = \theta_m(F) + o_p(1). \tag{3.3}$$

Proof. Suppose T_m does not 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 (3.1) follows immediately. For (3.2) note that

$$\begin{aligned} & E_F|J_m - \binom{n}{m}^{-1} \sum_{i_1 < \cdots < i_m} h(T_m(X_{i_1}, \ldots, X_{i_m}, F))| \\ \leq\ & E_F|h(T_m(X_1, \ldots, X_m, \hat{F}_n)) - h(T_m(X_1, \ldots, X_m, F))| \end{aligned} \tag{3.4}$$

and (3.2) follows by bounded convergence. These results follows in the same way and even more easily for N_m. Note that if T_m does not depend on F, $E_F N_m = \theta_m(F)$ and,

$$\mathrm{Var}_F(N_m) = \frac{m}{n}\mathrm{Var}_F(h(T_m(X_1, \ldots, X_m))) > \mathrm{Var}_F(J_m). \tag{3.5}$$

Note. It may be shown, more generally under (3.2), that, for example, distances between the $\binom{n}{m}$ bootstrap distributions of $T_m(\hat{F}_m, F)$ and $\mathcal{L}_m(F)$ are also $O_P(m/n)^{1/2}$.

Let $X_j^{(i)} = (X_j, \ldots, X_j)_{1\times i}$

$$h_{i_1,\ldots,i_r}(X_1, \ldots, X_r) = \frac{1}{r!} \sum_{1\le j_1 \ne \cdots \ne j_r \le r} h(T_m(X_{j_1}^{(i_1)}, \ldots, X_{j_r}^{(i_r)}, F)), \tag{3.6}$$

for vectors $\boldsymbol{i} = (i_1, \ldots, i_r)$ in the index set

$$\Lambda_{r,m} = \{(i_1, \ldots, i_r) : 1 \le i_1 \le \cdots \le i_r \le m, i_1 + \cdots + i_r = m\}.$$

Then

$$B_{m,n}(F) = \sum_{r=1}^{m} \sum_{i\in\Lambda_{r,m}} \omega_{m,n}(\boldsymbol{i}) \frac{1}{\binom{n}{r}} \sum_{1\le j_1\le\cdots\le j_r\le m} h_i(X_{j_1}, \ldots, X_{j_r}, F), \tag{3.7}$$

where

$$\omega_{m,n}(\boldsymbol{i}) = \binom{n}{r}\binom{m}{i_1, \ldots, i_r}/n^m.$$

Let

$$\theta_{m,n}(F) = E_F B_{m,n}(F) = \sum_{r=1}^{m} \sum_{i\in\Lambda_{r,m}} \omega_{m,n}(i) E_F h_i(X_1, \ldots, X_r). \tag{3.8}$$

Finally, let

$$\delta_m(\frac{r}{m}) \equiv \max\{|E_F h_i(X_1, \ldots, X_r) - \theta_m(F)| : \boldsymbol{i} \in \Lambda_{r,m}\} \tag{3.9}$$

and define $\delta_m(x)$ by extrapolation on $[0, 1]$. Note that $\delta_m(1) = 0$.

Theorem 2. *Under the conditions of Theorem* 1

$$B_{m,n}(F) = \theta_{m,n}(F) + O_P(\frac{m}{n})^{\frac{1}{2}}. \tag{3.10}$$

If further,

$$\delta_m(1 - xm^{-1/2}) \to 0 \tag{3.11}$$

uniformly for $0 \le x \le M$, *all* $M < \infty$, *and* $m = o(n)$, *then*

$$\theta_{m,n}(F) = \theta_m(F) + o(1). \tag{3.12}$$

Finally if,

$$T_m(X_1^{(i_n)}, \ldots, X_r^{(i_r)}, F) = T_m(X_1^{(i_1)}, \ldots, X_r^{(i_r)}, \hat{F}_n) + o_P(1) \tag{3.13}$$

whenever $\boldsymbol{i} \in \Lambda_{r,m}, m \to \infty$ *and* $\max\{i_1, \ldots, i_r\} = O(m^{1/2})$ *then, if* $m \to \infty, m = o(n)$,

$$B_m = \theta_m(F) + o_p(1). \tag{3.14}$$

The proof of Theorem 2 will be given in the Appendix. There too we will show briefly that, in the examples we have discussed and some others, $J_{m(n)}$, $B_{m(n)}$, $N_{m(n)}$ are consistent for $m(n) \to \infty$, $\frac{m}{n} \to 0$.

According to Theorem 2, if T_n does not depend om F the m/n bootstrap works as well as the $\binom{n}{m}$ bootstrap if the value of T_m is not greatly affected by a number on the order of $\sqrt{m}$ ties in its argument. Some condition is needed. Consider $T_n(X_1, \ldots, X_n) = 1(X_i = X_j \text{ for some } i \neq j)$ and suppose F is continuous. The $\binom{n}{m}$ bootstrap gives $T_m = 0$ as it should. If $m \neq o(\sqrt{n})$ so that the $\binom{n}{m}$ and m/n bootstraps do not coincide asymptotically the m/n bootstrap gives $T_m = 1$ with positive probability. Finally, (3.13) is the natural extension of (3.2) and is as easy to verify in all our examples.

A number of other results are available for m out of n bootstraps.

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) extended these results to $J_{m(n)}$ with $m = o(n)$. Finally, under the Giné-Zinn conditions,

$$||\sqrt{m}(\hat{F}_n - F)|| = (\frac{m}{n})||\sqrt{n}(\hat{F}_n - F)|| = O_P(\frac{m}{n})^{1/2} \tag{3.15}$$

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

$$t(\sqrt{m}(\hat{F}_m - \hat{F}_n)) = t(\sqrt{m}(\hat{F}_m - F)) + o_p(1) \tag{3.16}$$

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

We close with a theorem on the parametric version of the m/n bootstrap which gives a stronger property than that of Theorem 1.

Let $\mathcal{F}_0 = \{F_\theta : \theta \in \Theta \subset R^p\}$ where Θ is open and the model is regular. That is, θ is identifiable, the F_θ have densities f_θ with respect to a σ finite μ and the map $\theta \to \sqrt{f_\theta}$ is continuously Hellinger differentiable with nonsingular derivative. By a result of LeCam (see Bickel, Klaassen, Ritov, Wellner (1993) for instance), there exists an estimate $\hat{\theta}_n$ such that, for all θ,

$$\int (f_{\hat{\theta}_n}^{1/2}(x) - f_\theta^{1/2}(x))^2 d\mu(x) = O_{P_\theta}(\frac{1}{n}). \tag{3.17}$$

Theorem 3. *Suppose $\mathcal{F}_0$ is as above. Let $F_\theta^m \equiv \underbrace{F_\theta \times \cdots \times F_\theta}_{m}$ and $\|\cdot\|$ denote the variational norm. Then*

$$\|F_{\hat{\theta}_n}^m - F_\theta^m\| = O_p((\frac{m}{n})^{1/2}). \tag{3.18}$$

Proof. This is consequence of the relations (LeCam (1986)).

$$\|F_{\theta_0}^m - F_{\theta_1}^m)\| \leq H(F_{\theta_0}^m, F_{\theta_1}^m)[(2 - H^2(F_{\theta_0}^m, F_{\theta_1}^m)], \tag{3.19}$$

where

$$H^2(F, G) = \frac{1}{2}\int(\sqrt{dF} - \sqrt{dG})^2 \tag{3.20}$$

and

$$H^2(F_{\theta_0}^m, F_{\theta_1}^m) = 1 - (\int \sqrt{f_{\theta_0} f_{\theta_1}} d\mu)^m = 1 - (1 - H^2(F_{\theta_0}, F))^m. \tag{3.21}$$

Substituting (3.21) into (3.20) and using (3.17) we obtain

$$\|F_{\hat{\theta}_n}^m - F_\theta^m\| = O_{P_\theta}(1 - \exp O_{P_\theta}(\frac{m}{n}))^{\frac{1}{2}}(1 + \exp O_{P_\theta}(\frac{m}{n})^{\frac{1}{2}}) = O_{P_\theta}(\frac{m}{n})^{\frac{1}{2}}. \tag{3.22}$$

This result is weaker than Theorem 1 since it refers only to the parametric bootstrap. It is stronger since even for $m = 1$, when sampling with and without replacement coincide, $\|\hat{F}_n - F_\theta\| = 1$ for all n if F_θ is continous.

4. 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 (1982) and assume that at least on $\mathcal{F}_0$,

$$\theta_n(F) = \theta(F) + \theta'(F)n^{-1/2} + O(n^{-1}), \tag{4.1}$$

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).

Somewhat more generally than Beran, we exhibit conditions under which $B_n = \theta_n(\hat{F}_n)$ is fully efficient as an estimate of $\theta_n(F)$ and show that the m out n bootstrap with $\frac{m}{n} \to 0$ has typically relative efficiency 0.

We formally state a theorem which applies to fairly general parameters θ_n. Suppose ρ is a metric on $\mathcal{F}_0$ such that

$$\rho(\hat{F}_n, F_0) = O_{P_{F_0}}(n^{-1/2}) \text{ for all } F_0 \in \mathcal{F}_0. \tag{4.2}$$

Further suppose

A. $\theta(F), \theta'(F)$ are ρ Fréchet differentiable in $\mathcal{F}$ at $F_0 \in \mathcal{F}_0$. That is,

$$\theta(F) = \theta(F_0) + \int \psi(x, F_0) dF(x) + o(\rho(F, F_0)) \tag{4.3}$$

for $\psi \in L_2^0(F_0) \equiv \{h : \int h^2(x) dF_0(x) < \infty, \int h(x) dF_0(x) = 0\}$ and θ' obeys a similar identity with ψ replaced by another function $\psi' \in L_2^0(F_0)$. Suppose further

B. The tangent space of $\mathcal{F}_0$ at F_0 as defined in Bickel et al. (1993) is $L_2^0(F_0)$ so that ψ and ψ' are the efficient influence functions of θ, θ'. Essentially, we require that in estimating F there is no advantage in knowing $F \in \mathcal{F}_0$.

Finally, we assume,

C. For all $M < \infty$,

$$\sup\{|\theta_m(F) - \theta(F) - \theta'(F) m^{-1/2}| : \rho(F, F_0) \le M_n^{-1/2}, F \in \mathcal{F}\} = O(m^{-1}) \tag{4.4}$$

a strengthened form of (4.1). Then,

Theorem 4. *Under regularity of θ, θ' and A and C at F_0,*

$$\begin{aligned} \theta_m(\hat{F}_n) \;\equiv\; & \theta(F_0) + \theta'(F_0) m^{-1/2} + \frac{1}{n} \sum_{i=1}^{n} (\psi(X_i, F_0) + \psi'(X_i, F_0) m^{-1/2}) \\ & + O(m^{-1}) + o_p(n^{-1/2}). \end{aligned} \tag{4.5}$$

If B also holds, $\theta_n(\hat{F}_n)$ is efficient. If in addition, $\theta'(F_0) \neq 0$, and $\frac{m}{n} \to 0$ the efficiency of $\theta_m(\hat{F}_n)$ is 0.

Proof. The expansions of $\theta(\hat{F}_n)\theta'(\hat{F}_n)$ are immediate by Fréchet differentiability and (4.5) follows by plugging these into (4.1). Since θ, θ' are assumed regular, ψ and ψ' are their efficient influence functions. Full efficiency of $\theta_n(\hat{F}_n)$ follows by general theory as given in Beran (1983) for special cases or by extending Theorem 2, p.63 of Bickel et al. (1993) in an obvious way. On the other hand, if $\theta'(F_0) \neq 0$, $\sqrt{n}(\theta_m(\hat{F}_n) - \theta_n(F_0))$ has asymptotic bias $(\sqrt{\frac{n}{m}} - 1)\theta'(F_0) + O(\frac{\sqrt{n}}{m}) = \sqrt{\frac{n}{m}}(1 + o(1))\theta'(F_0) \to \pm\infty$ and inefficiency follows.

Inefficiency results of the same type or worse may be proved about J_m and N_m but require going back to $T_m(X_1, \ldots, X_m, F)$ since J_m and B_n are not related in a simple way. We pursue this only by way of Example 1. If $\theta_n(F) = \mathrm{Var}_F(\sqrt{n}(\bar{X} - \mu(F)) = \theta(F)$, $B_m = B_n$ but,

$$J_m = \sigma^2(\hat{F}_n)(1 - \frac{m-1}{n-1}). \tag{4.6}$$

Thus, since $\theta'(F) = 0$ here, B_m is efficient but J_m has efficiency 0 if $\frac{m}{\sqrt{n}} \to \infty$. N_m evidently behaves in the same way.

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 of 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 (1992)), 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 bsed on a sample of size m. The reason is that, in this case, the bootstrap distribution function is

$$\Phi(t) - m^{-1/2}c(\hat{F}_n)\varphi(t)H_2(t) + O_P(m^{-1}) \tag{4.7}$$

rather than the needed,

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

The error committed is of order $m^{-1/2}$. More general formal results can be stated but we do not pursue this.

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.

5. Remedying the Deficiencies of $B_{m(n)}$ when B_n is Correct: Extrapolation

In Bickel and Yahav (1988), motivated by considerations of computational economy, situations were considered in which θ_n has an expansion of the form (4.1) and it was 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. We sketch the argument for a special case.

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

$$B_m = \theta_m(\hat{F}_n) = \theta(\hat{F}_n) + \theta'(\hat{F}_n)m^{-1/2} + O_P(m^{-1}). \tag{5.1}$$

Then, if $n_1 > n_0 \to \infty$

$$\theta'(\hat{F}_n) = (B_{n_0} - B_{n_1})(n_0^{-1/2} - n_1^{-1/2})^{-1} + O_P(n_0^{-1/2}) \tag{5.2}$$

$$\theta(\hat{F}_n) = \frac{n_0^{-1/2}B_{n_1} - n_1^{-1/2}B_{n_0}}{n_0^{-1/2} - n_1^{-1/2}} + O_P(n_0^{-1}) \tag{5.3}$$

and hence a reasonable estimate of B_n is,

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

More formally,

Proposition. *Suppose* $\{\theta_m\}$ *obey* C *of Section* 4 *and* $n_0 n^{-1/2} \to \infty$. *Then,*

$$B_{n_0,n_1} = B_n + o_p(n^{-1/2}). \tag{5.4}$$

Hence, under the conditions of Theorem 3 B_{n_0,n_1} *is efficient for estimating* $\theta_n(F)$.

Proof. Under C, (5.4) holds. By construction,

$$\begin{aligned} B_{n_0,n_1} &= \theta(\hat{F}_n) + \theta'(\hat{F}_n)n^{-1/2} + O_P(n_0^{-1}) + O_P(n_0^{-1/2}n^{-1/2}) \\ &= \theta_n(\hat{F}_n) + O_P(n_0^{-1}) + O_P(n_0^{-1/2}n^{-1/2}) + O_P(n^{-1}) \\ &= \theta_n(\hat{F}_n) + O_P(n_0^{-1}) \end{aligned} \tag{5.5}$$

and (5.4) follows.

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 (4.1) 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, since ψ of Theorem 4 is 0,

$$\hat{\theta}_{n_0} = (1 - (\frac{n_0}{n})^{1/2})\theta_0 + (\frac{n_0}{n})^{1/2}B_{n_0}. \tag{5.6}$$

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 (5.4) to construct a bootstrap confidence bound we expect the coverage probability to be 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. We do not pursue a formal statement.

5.1. Extrapolation of J_m and N_m

We discuss extrapolation for J_m and N_m only in the context of the simplest Example 1, where the essential difficulties become apparent and we omit general theorems.

In work in progress, Götze and coworkers are developing expansions for general symmetric statistics under sampling from a finite population. These results will permit general statements of the same qualitative nature as in our discussion of Example 1. Consider $\theta_m(F) = P_F[\sqrt{m}(\bar{X}_m - \mu(F)) \le t]$. If $EX_1^4 < \infty$ and the X_i obey Cramér's condition, then

$$\theta_m(F) = \Phi(\frac{t}{\sigma(F)}) - K_3(F)\frac{\varphi}{6\sqrt{m}}(\frac{t}{\sigma(F)})H_2(\frac{t}{\sigma(F)}) + O(m^{-1}), \tag{5.7}$$

where $\sigma^2(F)$ and $K_3(F)$ are the second and third cumulants of F and $H_k(t) = \frac{(-1)^k}{\varphi(t)}\frac{d\varphi^k(t)}{dt^k}$. By Singh (1981), $B_m = \theta_m(\hat{F}_n)$ has the same expansion with F replaced by $\hat{F}_n$. However, by an easy extension of results of Robinson (1978) and Babu and Singh (1985),

$$J_m = \Phi(\frac{t}{\hat{K}_{2m}}) - \varphi(\frac{t}{\hat{K}_{2m}^{1/2}})\frac{\hat{K}_{3m}}{6m^{1/2}}H_2(\frac{t}{\hat{K}_{2m}^{1/2}}) + O_P(m^{-1}), \tag{5.8}$$

where

$$\hat{K}_{2m} = \sigma^2(\hat{F}_n)(1 - \frac{m-1}{n-1}) \tag{5.9}$$

$$\hat{K}_{3m} = K_3(\hat{F}_n)(1 - \frac{m-1}{n-1})(1 - \frac{2(m-1)}{n-2}). \tag{5.10}$$

The essential character of expansion (5.8), if $m/n = o(1)$, is

$$J_m = \theta(\hat{F}_n) + m^{-1/2}\theta'(\hat{F}_n) + \frac{m}{n}\gamma_n + O_P(m^{-1} + (\frac{m}{n})^2 + \frac{m^{\frac{1}{2}}}{n}), \tag{5.11}$$

where γ_n is $O_P(1)$ and independent of m. The m/n terms essentially come from the finite population correction to the variance and highter order cumulants of means of samples from a finite population. They reflect the obvious fact that if $m/n \to \lambda > 0$, J_m is, in general, incorrect even to first order. For instance, the variance of the $\binom{n}{m}$ bootstrap distribution corresponding to $\sqrt{m}(\bar{X} - \mu(F))$ is $1/n\sum(X_i - \bar{X})^2(1 - \frac{m-1}{n-1})$ which converges to $\sigma^2(F)(1-\lambda)$ if $m/n \to \lambda > 0$. What this means is that if expansions (4.1), (5.1) and (5.11) are valid, then using $J_{m(n)}$ again gives efficiency 0 compared to B_n. Worse is that (5.2) with J_{n_0}, J_{n_1} replacing B_{n_0}, B_{n_1} will not work since the n_1/n terms remain and make

a contribution larger than $n^{-1/2}$ if $n_1/n^{1/2} \to \infty$. Essentially it is necessary to estimate the coefficient of m/n and remove the contribution of this term at the same time 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 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 $T_m(X_1, \ldots, X_m, F) \equiv \sqrt{m}\frac{(\bar{X}_m - \mu(F))}{\hat{\sigma}(1-\frac{m-1}{n-1})^{1/2}}$ leads to an expansion for J_m of the form,

$$J_m = \Phi(t) + \theta'(\hat{F}_n)m^{-1/2} + O_P(m^{-1} + m/n), \tag{5.12}$$

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 sample splitting situation is less satisfactory in the same example. Under (5.1), the coefficient of $1/\sqrt{m}$ 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 (5.11). 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

$$\operatorname{Cov}(N_m, N_{2m}) = \frac{1}{k}\operatorname{Cov}\Big(h(m^{-1/2}(\sum_{j=1}^{m}(X_j - \bar{X}))), h((2m)^{-1/2}\sum_{j=1}^{2m}(X_j - \bar{X}))\Big). \tag{5.13}$$

Thus, by the central limit theorem,

$$\operatorname{Corr}(N_m, N_{2m}) \to \frac{1}{2}\frac{\operatorname{Cov}}{\operatorname{Var}(Z_1)}\Big(h(Z_1), h\frac{(Z_1 + Z_2)}{\sqrt{2}}\Big), \tag{5.14}$$

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 does not 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 (5.1) and (5.11) are invalid beyond the 0th order, the usual situation when the nonparametric 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 7. We now turn to question (b) in the next section.

6. Behaviour of the Smaller Resample Schemes When B_n is Inconsistent, and Presentation of 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 better 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 should not improve over the behaviour of B_{n_0}, B_{n_1}.

C. If n_0, n_1 are comparable extrapolation should not 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 1. *Suppose*

$$B_{n_1} - \theta_n(F) \asymp B_{n_0} - \theta_n(F), \tag{6.1}$$

where $\asymp$ *indicates that the ratio tends to* 1. *Then, if* $n_0/n_1 \not\to 1$

$$B_{n_0,n_1} - \theta_n(F) \asymp B_{n_0} - \theta_n(F). \tag{6.2}$$

Proof. Evidently, $\frac{B_{n_0}+B_{n_1}}{2} = \theta_n(F) + \Omega(\epsilon_n)$ where $\Omega(\epsilon_n)$ means that the exact order of the remainder is ϵ_n. On the other hand,

$$\frac{B_{n_0} - B_{n_1}}{n_0^{-1/2} - n_1^{-1/2}}\Big(\frac{1}{\sqrt{n}} - \frac{1}{2}\big(\frac{1}{\sqrt{n_0}} + \frac{1}{\sqrt{n_1}}\big)\Big) = \Omega(\epsilon_n)\Big(\sqrt{\frac{n_0}{n}} + \Omega(1)\Big)$$

and the proposition follows.

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

Example 3. Here, $F^{-1}(0) = 0$,

$$\theta_n(F) = e^{f(0)t}\Big(1 + n^{-1}f'(0)\frac{t^2}{2}\Big) + O(n^{-2}) \tag{6.3}$$

which is of the form (5.1). 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,

$$\begin{aligned}\log B_m &= m\log(1-\hat{F}_n(\frac{t}{m})) = m\log(1-(\hat{F}_n(\frac{t}{m})-\hat{F}_n(0)))\\ &= -m(F(\frac{t}{m})-F(0))-\frac{m}{\sqrt{n}}\sqrt{n}(\hat{F}_n(\frac{t}{m})-F(\frac{t}{m}))+O_P(m(\hat{F}_n(\frac{t}{m})-F(\frac{t}{m}))^2)\\ &= tf(0)+\Omega(\frac{1}{m})+\Omega_P(\sqrt{\frac{m}{n}})+O_P(\frac{1}{n}), \qquad (6.4)\end{aligned}$$

where as before Ω, Ω_p indicate exact order. As Politis and Romano (1994) point out, $m = \Omega(n^{1/3})$ yields the optimal rate $n^{-1/3}$ (under f Lipschitz). Extrapolation does not 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 like the sample path of a stationary Gaussian process. So conclusion B holds in this case.

Example 4. A major difficulty here is defining $\mathcal{F}_0$ narrowly enough so that it is meaningful to talk about expansions of $\theta_n(F)$, $B_n(F)$ etc. If $\mathcal{F}_0$ in these examples is 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 the stable law attraction domain estimation of the mean mentioned in Example 4 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 does not help. Hall and Jing's (1993) 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 the value of the stable law index 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 6. 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_1^2 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.

We shall discuss claims A and D in the context of Example 5 or rather its simplest case with $T_n(\hat{F}_n, F) = n\bar{X}^2$. We begin with,

Proposition 2. *Suppose $E_F X_1^4 < \infty$, $E_F X_1 = 0$, and F satisfies Cramer's condition. Then,*

$$B_m \equiv P^*[|\sqrt{m}\bar{X}^*|^2 \le t^2] = 2\Phi(\frac{t}{\hat{\sigma}}) - 1 - \frac{m\bar{X}^2}{\hat{\sigma}^3} t\varphi(\frac{t}{\hat{\sigma}}) - \frac{\hat{K}_3\bar{X}}{3\hat{\sigma}^4}\varphi H_3(\frac{t}{\hat{\sigma}}) + O_P(\frac{m}{n})^{3/2} + O_P(m^{-1}). \tag{6.5}$$

If $m = \Omega(n^{1/2})$ then

$$P^*[|\sqrt{m}\bar{X}^*|^2 \le t^2] = P_F[n\bar{X}^2 \le t] + O_P(n^{-1/4}) \tag{6.6}$$

and no better choice of $\{m(n)\}$ is possible. If $n_0 < n_1$, $n_0 n^{-1/2} \to \infty$, $n_1 = o(n^{3/4})$,

$$B^{n_0,n_1} \equiv B_{n_0} - n_0\{(B_{n_1} - B_{n_0})/(n_1 - n_0)\} = P_F[n\bar{X}^2 \le t] + O_P(n^{-1/2}). \tag{6.7}$$

Proof. We make a standard application of Singh (1981). If $\hat{\sigma}^2 \equiv \frac{1}{n}\sum(X_i - \bar{X})^2$, $\hat{K}_3 \equiv \frac{1}{n}\sum(X_i - \bar{X})^3$ we get, after some algebra and Edgeworth expansion,

$$P^*[\sqrt{m}\bar{X}^* \le t] = \Phi\Big(\frac{t - \sqrt{m}\bar{X}}{\hat{\sigma}}\Big) - \frac{1}{\sqrt{m}}\varphi\Big(\frac{t - \sqrt{m}\bar{X}}{\hat{\sigma}}\Big)\frac{\hat{K}_3}{6}H_2\Big(\frac{t - \sqrt{m}\bar{X}}{\hat{\sigma}}\Big) + O_p(m^{-1}).$$

After Taylor expansion in $\sqrt{m}\frac{\bar{X}}{\hat{\sigma}}$ we conclude,

$$P^*[m\bar{X}_m^{*2} \le t^2] = 2\Phi(\frac{t}{\hat{\sigma}}) - 1 + \frac{\varphi'}{2}(\frac{t}{\hat{\sigma}})m\bar{X}^2 - \frac{\hat{K}_3}{3\hat{\sigma}^4}[\varphi H_3](\frac{t}{\hat{\sigma}})\bar{X} + O_P(\frac{m}{n})^{3/2} + O_P(m^{-1}) \tag{6.8}$$

and (6.5) follows. Since $m\bar{X}^2 = \Omega_P(m/n)$, (6.6) follows. Finally, from (6.5), if $n_0 n^{-1/2}$, $n_1 n^{-1/2} \to \infty$

$$B_{n_0} - n_0\{(B_{n_1} - B_{n_0})/(n_1 - n_0)\} = 2\Phi(\frac{t}{\hat{\sigma}}) - 1 - \frac{K_3}{6}\varphi H_2(\frac{t}{\hat{\sigma}})\bar{X} + O_P(n^{-3/4}) + O_P(n^{-1/2}) + O_P(n^{-1/2}). \tag{6.9}$$

Since $\bar{X} = O_P(n^{-1/2})$, (6.7) follows.

Example 5. As we noted, the case $T_n(\hat{F}_n, F) = n\bar{X}^2$ is the prototype of the use of the m/n bootstrap for testing discussed in Bickel and Ren (1995). From (6.7) of proposition 2 it is clear that extrapolation helps. However, it is not true that B^{n_0,n_1} is efficient since it has an unnecessary component of variance $(\hat{K}_3/6)[\varphi H_2](\frac{t}{\hat{\sigma}})\bar{X}$ which is negligible only if $K_3(F) = 0$. On the other hand it is easy to see that efficient estimation can be achieved by resampling not the X_i but the residuals $X_i - \bar{X}$, that is, a consistent estimate of F belonging to $\mathcal{F}_0$. So this example illustrates both A and D. 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.

7. Simulations and Conclusions

The simulation algorithms were written and carried out by Adele Cutler and Jiming Jiang. 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.

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

Bootstrap sample size: $B = 500$

Simulation size: $N = 2000$

Distributions: Example 1: $F = \chi^2_1$; Example 3: $F = \chi^3_2$

Statistics:

Example 1(a) modified: $T_m^{(a)} = \sqrt{m}(\sqrt{\bar{X}_m} - \sqrt{\mu(F)})$

Example 1(b): $T_m^{(b)} = \sqrt{m}\frac{(\bar{X} - \mu(F))}{s_m}$ where $s_m^2 = \frac{1}{m-1}\sum_{i=1}^m (X_i - \bar{X}_m)^2$.

Example 3. $T_m^{(c)} = m(\min(X_1, \ldots, X_m) - F^{-1}(0))$

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 m/n, $\binom{n}{m}$ and sample splitting respectively.

In Example 1 the G_m^{-1} parameters were used to form upper and lower "90%" confidence bounds for $\theta \equiv \sqrt{\mu(F)}$. Thus, from $T_m^{(a)}$,

$$\bar{\theta}_{mB} = \sqrt{\bar{X}_n} - \frac{1}{\sqrt{n}} G_{mB}^{-1}(.1)) \tag{7.1}$$

for the "90%" upper confidence bound based on the m/n bootstrap and, from $T_m^{(b)}$,

$$\bar{\theta}_{mB} = ((\bar{X}_n - \frac{s_n}{\sqrt{n}} G_{mB}^{-1}(.1))_+)^{1/2}, \tag{7.2}$$

where G_{mB} now corresponds to the t statistic. $\underline{\theta}_{mB}$, is defined similarly. The $\bar{\underline{\theta}}_{mJ}$ bounds are defined with G_{mJ} replacing G_{mB}. The $\bar{\underline{\theta}}_{mN}$ bounds are considered only for the unambiguous case m divides n and α an integer multiple of m/n.

Thus if $m = n/10$, $G^{-1}_{mN}(.1)$ is simply the smallest of the 10 possible values $\{T_m(X_{jm+1}, \ldots, X_{(j+1)m}, \hat{F}_n), 0 \le j \le 9\}$.

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 for $T^{(a)}_m$, for example, by.

$$\bar{\theta}_{n_0,n_1} = \sqrt{\bar{X}_n} - \frac{1}{\sqrt{n}}\Big\{\frac{(G^{-1}_{n_0B}(.1) + G^{-1}_{n_1B}(.1))}{2} + (n^{-1/2} - \frac{1}{2}(n_0^{-1/2} + n_1^{-1/2}))(G^{-1}_{n_0B}(.1) - G^{-1}_{n_1B}(.1))/(n_0^{-1/2} - n_1^{-1/2})\Big\}. \quad (7.3)$$

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 $(400, 40, 80)$.

In Example 3, we similarly study the lower confidence bound on $\theta = F^{-1}(0)$ given by,

$$\underline{\theta}_m = \max(X_1, \ldots, X_n) - \frac{1}{n}G^{-1}_{mB}(.9). \quad (7.4)$$

and the extrapolation lower confidence bound

$$\underline{\theta}_{n_0,n_1} = \min(X_1, \ldots, X_n) - \frac{1}{n}\frac{(G^{-1}_{n_0B}(.9) + G^{-1}_{n_1B}(.9))}{2} + (n^{-1} - \frac{(n_0^{-1} + n_1^{-1})}{2})(G^{-1}_{n_0B}(.9) - G^{-1}_{n_1B}(.9))(n_0^{-1} - n_1^{-1}). \quad (7.5)$$

Note that we are using $1/m$ rather than $1/\sqrt{m}$ for extrapolation.

Measures of performance:

$CP \equiv$ Coverage probability, the actual probability under the situation simulated that the region prescribed by the confidence bound covers the true value of the parameter being estimated.

$$RMSE = \sqrt{E(\text{Bound}-\text{Actual quantile bound})^2}\ .$$

Here the actual quantile bound refers to what we would use if we knew the distribution of $T_n(X_1, \ldots, X_n, F)$. For example for $T^{(a)}_m$ we would replace $G^{-1}_{mB}(.1)$ in (7.1) for $F = \chi^2_1$ by the .1 quantile of the distribution of $\sqrt{n}(\sqrt{\frac{S_m}{m}} - 1)$ where S_m has a χ^2_m distribution, call it $G^{*-1}_m(.1)$. Thus, here,

$$MSE = \frac{1}{n}E(G^{-1}_{mB}(.1) - G^{*-1}_m(.1))^2.$$

We give in Table 1 results for the B_{n_1}, B_n and B_{n_0,n_1} bounds, based on $T^{(b)}_m$. The $T^{(a)}_m$ bootstrap, as in Bickel and 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 give the corresponding results for lower confidence bounds based on $T_m^{(c)}$ in Table 2. Table 3 presents results for sample splitting for $T_m^{(a)}$. Table 4 presents $T_m^{(a)}$ results for the $\binom{n}{m}$ bootstrap.

Table 1. The t bootstrap: Example 1(b) at 90% nominal level

	Coverage probabilities (CP)			*RMSE*		
n	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: (a) B1 corresponds to (6.2) or its LCB analogue for $m=n_1(n)=30, 40, 80$. Similarly B corresponds to $m = n$.
(b) BR corresponds to (6.3) or its LCB analogue with $(n_0, n_1) = (15, 30), (20, 40), (40, 80)$.

Table 2. The min statistic bootstrap: Example 3 at the nominal 90% level

n		*CP*	*RMSE*
50			
	B	.75	.01
	B1	.78	.07
	BR	.70	.07
	B1S	.82	.07
	BRS	.80	.07

n		*CP*	*RMSE*
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: (a) B corresponds to (6.4) with $m = n$, B1 with $m=n_1=30, 40, 80$, B1S with $m=n_1=16$.
(b) BR corresponds to (6.5) with $(n_0, n_1)=(15, 30), (20, 40), (40, 80)$, BRS with $(n_0, n_1)=(4, 16)$.

Table 3. Sample splitting in Example 1(a)

		CP		$RMSE$	
n		N	$B_{m(n)}$	N	$B_{m(n)}$
50					
	UB	.82	.86	.32	.18
	LB	.86	.91	.28	.16
100					
	UB	.86	.89	.30	.14
	LB	.84	.90	.26	.12
400					
	UB	.85	.89	.28	.08
	LB	.86	.91	.27	.09

Note: N here refers to $m = .1n$ and $\alpha = .1$.

Table 4. The $\binom{n}{m}$ bootstrap and the m/n bootstrap in Example 1(a)

	CP			E(Length)	
n	m	J	B	J	B
50	16	.82	.88	.07	.09
100	16	.86	.88	.04	.05
400	40	.88	.90	.01	.01

Note: These figures are for simulation sizes of $N = 500$ and for 90% confidence intervals. Thus, the end points of the intervals are given by (7.1) and its UCB counterpart for B and J but with .1 replaced by .05. Similarly, $[E(\text{Bound}-\text{Actual quantile bound})^2]^{1/2}$ is replaced by the expected length of the confidence interval.

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 asymptotics 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^2_1 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 sample splitting bounds differs from the nominal level. If $n \to \infty$, $m/n \to .1$ the coverage probability doesn't tend to .1 since the estimated quantile doesn't tend to the actual quantile and both CP and $RMSE$ behave badly compared to B_m. This naive method can be fixed up (see Blom (1976) for instance). However, its simplicity is lost and the $\binom{n}{m}$ or m/n bootstrap seem preferable.
- The $\binom{n}{m}$ bounds are inferior as Table 4 shows. This reflects the presence of the finite population correction 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 3)

- From Table 2, as expected, the CP of the n/n bootstrap 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. However, it is clear that the smaller m of B1S is better than all other choices.

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

A message of these simulations to us is that extrapolation of the B_m plot may carry risks not fully revealed by the asymptotics. 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.

The principal message we draw is that it is necessary to develop data driven methods of selection of m which lead to reasonable results over situations where both the bootstrap works and where it doesn't. Such methods are being pursued.

Acknowledgement

We are grateful to Jiming Jiang and Adele Cutler for essential programming, to John Rice for editorial comments, and to Kjell Doksum for the Blom reference. This research was supported by NATO Grant CRG 920650, Sonderforschungsbereich 343 Diskrete Strukturen der Mathematik, Bielefeld and NSA Grant MDA 904-94-H-2020.

Appendix

Proof of Theorem 2. For $\boldsymbol{i} = (i_1, \ldots, i_r) \in \Lambda_{r,m}$ let $U(\boldsymbol{i}) = \frac{1}{\binom{n}{r}} \sum\{h_i(X_{j_1}, \ldots, X_{j_r}, F) : 1 \le j_1 < \cdots < j_r \le n\}$. Then, since $h_{\boldsymbol{i}}$ as defined is symmetric in its arguments it is a U statistic and $\|h\|_\infty$ is an upper bound to its kernel. Hence

(a) $$\operatorname{Var}_F U(\boldsymbol{i}) \le \|h\|_\infty^2 \frac{r}{n}. \quad \text{On the other hand,}$$

(b) $$EU(\boldsymbol{i}) = E_F h_i(X_1, \ldots, X_r, F) \quad \text{and}$$

(c) $$B_{m,n}(F) = \sum_{r=1}^{m} \sum\{w_{m,n}(\boldsymbol{i}) U(\boldsymbol{i}) : \boldsymbol{i} \in \Lambda_{r,m}\} \text{ by (3.7). Thus, by (c),}$$

(d) $$\operatorname{Var}_F^{1/2} B_{m,n}(F) \le \sum_{r=1}^{m} \sum\{w_{m,n}(\boldsymbol{i}) \operatorname{Var}_F^{1/2} U(\boldsymbol{i}) : \boldsymbol{i} \in \Lambda_{r,m}\}$$
$$\le \max \operatorname{Var}_F^{1/2} U(\boldsymbol{i}) \le \|h\|_\infty (\frac{m}{n})^{1/2}$$

by (a). This completes the proof of (3.10).

The proof of (3.11) is more involved. By (3.8)

(e) $$|\theta_{m,n}(F) - \theta(F)| \le \sum_{r=1}^{m} \sum\{|E_F h_i(X_1, \ldots, X_r) - \theta_m(F)| w_{m,n}(\boldsymbol{i}) : \boldsymbol{i} \in \Lambda_{r,m}\}.$$

Let,

$$\text{(f)} \qquad P_{m,n}[R_m = r] = \sum\{w_{m,n}(\boldsymbol{i}) : \boldsymbol{i} \in \Lambda_{r,m}\}.$$

Expression (f) 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

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

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

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

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

$$\text{(j)} \qquad \text{Var}_{m,n}(R_m) = O(m)$$

so that,

$$\text{(k)} \qquad \frac{R_m}{m} = 1 + O_P(m^{-1/2}).$$

From (e),

$$\text{(l)} \qquad |\theta_{m,n}(F) - \theta(F)| \leq \sum_{r=1}^{m} \delta_m(\frac{r}{m}) P_{m,n}[R_m = r].$$

By (k), (l) and the dominated convergence theorem (3.12) follows from (3.11) and (k).

Finally, as in Theorem 1, we bound, as in (3.4),

$$\text{(m)} \qquad |B_{m,n}(F) - B_m(F)| \leq \sum_{r=1}^{m} \sum\{E_F|h_i(X_1, \ldots, X_r) - h_i(X_1, \ldots, X_r, \hat{F}_n)| : \boldsymbol{i} \in \Lambda_{r,m}\} w_{m,n}(\boldsymbol{i}),$$

where

$$\text{(n)} \qquad h_i(X_1, \ldots, X_r, \hat{F}_n) = \frac{1}{r!} \sum_{1 \leq j_1 \neq \cdots \neq j_r \leq r} h(T_m(X_{j_1}^{(i_1)}, \ldots, X_{j_r}^{(i_r)}, \hat{F}_n)).$$

Let R_m be distributed according to (f) and given $R_m = r$, let $(I_1, \ldots, I_r)$ be uniformly distributed on the set of partitions of m into r ordered integers, $I_1 \leq I_2 \leq \cdots \leq I_r$. Then, from (m) we can write

$$|B_{m,n}(F) - B_m(F)| \leq E\Delta(I_1, \ldots, I_{R_m}), \tag{o}$$

where $\|\Delta\|_\infty \leq \|h\|_\infty$. Further, by the continuity of h and (3.13), since $I_1 \leq \cdots \leq I_{R_m}$,

$$\Delta(I_1, \ldots, I_{R_m}) 1(I_{R_m} \leq \epsilon_m m) \xrightarrow{P} 0 \tag{p}$$

whenever $\epsilon_m = O(m^{-1/2})$. Now, $I_{R_m} > \epsilon_m m$,

$$m = \sum_{j=1}^{R_m} I_j \tag{q}$$

and $I_j \geq 1$ imply that,

$$m(1 - \epsilon_m) \geq \sum_{j=1}^{R_m - 1} I_j \geq (R_m - 1). \tag{r}$$

Thus,

$$P_{m,n}(I_{R_m} > \epsilon_m m) \leq P_{m,n}\left(\frac{R_m}{m} - 1 \leq -\epsilon_m + O(m^{-1})\right) \to 0 \tag{s}$$

if $\epsilon_m m^{1/2} \to \infty$. Combining (s), (k) and (p) we conclude that

$$E\Delta(I_1, \ldots, I_{R_m}) \to 0 \tag{t}$$

and hence (o) implies (3.14).

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 easier to directly verify that, in all our examples, the m/n bootstrap distribution of $T_n(\hat{F}_n, F)$ converges weakly (in probability) to its limit $\mathcal{L}(F)$ and conclude that Theorem 2 holds for all h continuous and bounded than to check the conditions of Theorem 2. Such verifications can be

found in the papers we cite. We sketch in what follows how the conditions of Theorem 1 and 2 can be applied.

Example 1. (a) We 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 note that the distribution of $m^{-1/2}(i_1X_1 + \cdots + i_rX_r)$ differs from that of $m^{-1/2}(X_1 + \cdots + X_m)$ 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 hypotheses of Theorem 2 hold.
(b) Note that,

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

and 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. Here is a direct argument for h bounded.

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

Thus,

$$\text{(b)} \qquad P[\text{med}(X_1^{(i_1)}, \ldots, X_r^{(i_r)}) \neq \text{med}(X_1, \ldots, X_m)] \leq \sum_{j=r+1}^{m} \frac{1}{j} \leq \log(\frac{m}{r}).$$

Hence for h bounded,

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

and we can apply Theorem 2.

Example 3. Follows by checking (3.2) in Theorem 1 and that Theorem 2 applies for J_m by arguing as above for B_m. Alternatively, argue as in Athreya and Fukushi (1994).

Arguments similar to those given so far can be applied to the other 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).

Babu, G. J. and Singh, K. (1985). Edgeworth expansions for sampling without replacement from finite populations. *J. Multivariate Anal.* **17**, 261-278.

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

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

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

Bertail, P. (1994). Second order properties of an extrapolated bootstrap without replacement. Submitted to *Bernoulli*.

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

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

Bickel, P. J. and Ren, J. J. (1995). The m out of n bootstrap and goodness of fit tests with double censored data. Robust Statistics, Data Analysis and Computer Intensive Methods Ed. H. Rieder Lecture Notes in Statistics, Springer-Verlag.

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, J. A. (1988). Richardson extrapolation and the bootstrap. J. *Amer. Statist. Assoc.* **83**, 387-393.

Blom, G. (1976). Some properties of incomplete U statistics. *Biometrika* **63**, 573-580.

Bretagnolle, J. (1981). Lois limites du bootstrap de certaines fonctionelles. *Ann. Inst. H. Poincaré, Ser.B* **19**, 281-296.

David, H. A. (1981). *Order Statistics.* 2nd edition, John Wiley, New York.

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, *Canad. J. Statist.* **17**, 155-169.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probab. Theory Related Fields* **95**, 125-140.

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. John Wiley, New York.

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

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 Wilson, S. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757-762.

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

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

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

Politis, D. N. and Romano, J. P. (1994). A general theory for large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22**, 2031-2050.

Praestgaard, J. and Wellner, J. (1993). Exchangeably weighted bootstraps 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.*

Robinson, J. (1978). An asymptotic expansion for samples from a finite population. *Ann. Statist.* **6**, 1005-1011.

Shen, X. and Wong, W. (1994). Convergence rates of sieve estimates. *Ann. Statist.* **22**, 580-615.

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.

Department of Statistics, University of California, Berkeley, 367 Evans Hall, #3860, Berkeley, CA 94720-3860, U. S. A.

Department of Mathematics, University of Bielefeld, Universitatsstrasse 4800, Bielefeld, Germany.

Department of Mathematics, University of Leiden, PO Box 9512 2300RA, Leiden, Netherlands.

(Received August 1995; accepted June 1996)