Some Asymptotic Theory for the Bootstrap

Peter J. Bickel; David A. Freedman

Annals of Statistics, Volume 9, Issue 6 (Nov., 1981), 1196-1217.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28198111%299%3A6%3C1196%3AASATFTB%3E2.0.CO%3B2-R

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,*

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

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

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

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

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

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

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

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

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

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

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,

$$\int \{\psi(x, F_n) - \psi(x, F)\}^2\, dF_n(x) = n^{-1} \sum_i \{\psi(X_i, F_n) - \psi(X_i, F)\}^2$$

$$= n^{-1} \sum_i \left\{n^{-1} \sum_j \omega(X_i, X_j) - \int \omega(X_i, y)\, dF(y)\right\}^2$$

$$= n^{-3} \sum_{i,j,k} \omega(X_i, X_j)\omega(X_i, X_k)$$

$$- 2n^{-2} \sum_{i,j} \omega(X_i, X_j) \int \omega(X_i, y)\, dF$$

$$+ n^{-1} \sum_i \left\{\int \omega(X_i, y)\, dF\right\}^2.$$

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

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

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

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

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

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

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

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

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, $F = F_{\theta_o}$ satisfies (5.1) and $(\partial/\partial\theta) F_\theta|_{\theta_o}$ is continuous in x and $n^{1/2}(\hat{\theta}_n - \theta_o) = O_p(1)$.

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

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

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

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

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 $[\alpha, 1 - \alpha]$, $0 < \alpha < ½$. 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 $X_{(1)}^* \leqslant \cdots \leqslant X_{(n)}^*$ 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 lim sup $n(X_{(n)} - X_{(n-k+1)}) = \infty$, and lim inf $n(X_{(n)} - X_{(n-k+1)}) = 0$, 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

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

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

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

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*

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

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

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

because $E\{(U-V)(U-V)^t\} = I_{n\times n} \cdot d_2^1(U_i, V_i)^2$, 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 $\mathbb{C}^\infty$ 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* $\mathbb{C}[0, 1]$ *of events which depend only on path behavior right after t* (Freedman, 1971, page 102). *Let P be the probability measure on* $\mathbb{C}[0, 1]$ *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 $c > 0$,

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

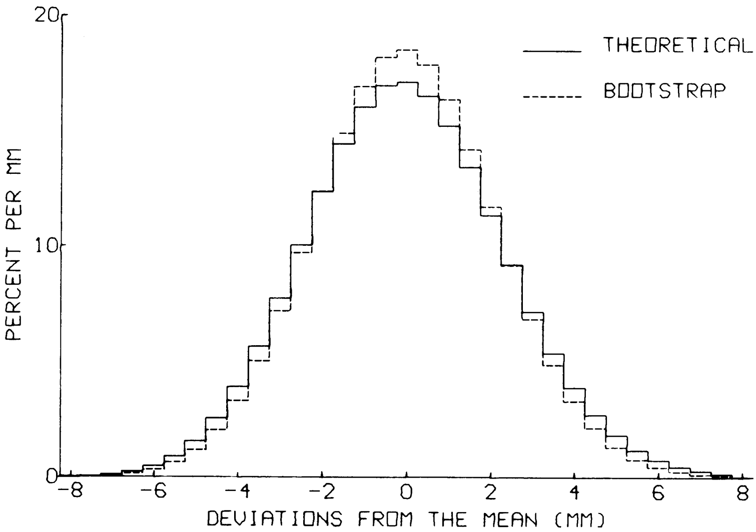

FIG. 1

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,

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

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