The Annals of Statistics
2000, Vol. 28, No. 3, 756–778

FINITE SAMPLE NONPARAMETRIC INFERENCE AND LARGE SAMPLE EFFICIENCY

BY JOSEPH P. ROMANO[1] AND MICHAEL WOLF[2]

Stanford University and Universidad Carlos III de Madrid

Given a sample $X_1, \dots, X_n$ from a distribution F, the problem of constructing nonparametric confidence intervals for the mean $\mu(F)$ is considered. Unlike bootstrap procedures or those based on normal approximations, we insist on any procedure being truly nonparametric in the sense that the probability that the confidence interval contains $\mu(F)$ based on a sample of size n from F be at least $1-\alpha$ for all F and all n. Bahadur and Savage proved it is impossible to find an effective (or bounded) confidence interval for $\mu(F)$ without some restrictions. Thus, we assume that F is supported on a known compact set, which we take to be [0, 1]. In this setting, an asymptotic efficiency result is obtained that gives a lower bound on the size of any conservative interval. We then provide a construction of an interval that meets our finite sample requirement on level, yet has an asymptotic efficiency property. Thus, the price to be paid for using fully nonparametric procedures when considering the trade-off between exact inference statements and asymptotic efficiency is negligible. Much of what is accomplished for the mean generalizes to other settings as well.

1. Introduction. Suppose $X_1, \dots, X_n$ are i.i.d. according to a distribution F on the line. Consider the problem of constructing a level $1-\alpha$ confidence interval for $\mu(F)$, the mean of F. The distribution F is assumed to belong to a large class $\mathbf{F}$ of distributions. Clearly, $\mathbf{F}$ must be restricted somewhat since we are assuming $\mu(F)$ exists. If I_n is a random interval (or set), define the coverage level over $\mathbf{F}$ to be

$$\inf\{P_F\{\mu(F) \in I_n\}\colon F \in \mathbf{F}\}.$$

In fact, even if we assume $\mathbf{F}$ consists of all distributions F having finite moments of all orders, Bahadur and Savage (1956) proved the negative result that it is impossible to construct an effective confidence interval for $\mu(F)$, where the term "effective confidence interval," as utilized by Bahadur and Savage, refers to one which is not too big. In particular, if I_n is a random interval (depending on $X_1, \dots, X_n$) such that, even for one F, the probability under F that I_n is a bounded set is 1, then the coverage level over $\mathbf{F}$ is zero. Hence, no bounded interval can serve as a level $1-\alpha$ confidence interval while satisfying the level constraint for all F.

Nevertheless, the aforementioned result has not deterred the search for valid inference procedures, especially since Efron's (1979) discovery of the

Received May 1998; revised February 2000.

[1] Supported by NSF Grant DMS-97-04487.

[2] Supported by the Spanish Dirección General de Enseñanza Superior, PB98-0025.

AMS 1991 *subject classifications*. Primary 62G15; secondary 62G20.

Key words and phrases. Confidence intervals, efficiency, nonparametrics.

bootstrap, as well as methods based on Edgeworth expansions, likelihood, and other resampling refinements. Indeed, there are several methods yielding intervals I_n of nominal level $1-\alpha$ satisfying, for fixed F,

$$\big|P_F\{\mu(F) \in I_n\} - (1-\alpha)\big| = O(n^{-p}), \tag{1.1}$$

for some $p > 0$. In fact, $p = 1$ for intervals whose coverage error is of the same order as that provided by the normal approximation, $p = 2$ for second-order accurate intervals such as the bootstrap t-interval, and p can even be larger by bootstrap iteration (under assumptions to ensure the validity of Edgeworth expansions); these properties are derived in Hall (1992). Unfortunately, all these intervals have the property that their coverage level over $\mathbf{F}$ is zero.

The technical reason why these methods can misbehave so badly yet still satify (1.1) is that the convergence result in (1.1) holds for each fixed F and is not uniform over F. Therefore, a question worthy of investigation is to find methods that are appropriately uniform, and indeed, some results are obtained in Hall and Jing (1995), by imposing restrictions on $\mathbf{F}$. Here, however, we insist on considering only procedures whose coverage level over a large $\mathbf{F}$ is the nominal level. Uniform convergence to the nominal level over $\mathbf{F}$, while preferable to pointwise convergence for each fixed F, is not strong enough for finite sample inferential purposes; we aim for a coverage level $1-\alpha$ for any finite sample size rather than in the limit only. Intervals possessing this finite sample validity property that the parameter is contained in the interval with probability at least $1-\alpha$ for all F (and n) are called conservative. Because of the Bahadur and Savage result, we do need to make some restriction in order to construct conservative intervals that are even just bounded. The assumption imposed then is that the unknown F has support in a fixed known compact set, which we take to be [0, 1]; otherwise, F is arbitrary (and need not be continuous, for example).

The main problem considered in this paper is then the following. Let $\mathbf{F_0}$ be the class of all distributions on [0, 1]. The goal is to construct a confidence interval I_n for $\mu(F)$ that satisfies the level constraint for all F in $\mathbf{F_0}$, but that is also not only bounded but efficient in some sense. First, among confidence intervals I_n whose coverage level over $\mathbf{F_0}$ is $1-\alpha$, can we find an optimality result giving a lower bound on how large the length of I_n can be? Second, can we actually construct an optimal interval? The answer to the first question is not that surprising, as intervals based on the normal approximation serve as our gold standard. Thus, if $I_n = [L_n, U_n]$ and $D_n = n^{1/2}(U_n - L_n)/2$, then the standard asymptotic interval $\overline{X}_n \pm z_{1-\alpha/2}s_n/n^{1/2}$ (where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and s_n is the sample standard deviation) satisfies $D_n \to z_{1-\alpha/2}\sigma(F)$ in probability under F, where $\sigma^2(F)$ is the variance of F. Not surprisingly, in a nonparametric setting, this asymptotic constant $z_{1-\alpha/2}\sigma(F)$ is in some sense the best attainable.

Of course, the normal theory intervals are not fully nonparametric as their coverage level over $\mathbf{F_0}$ is also zero. Such is the case for bootstrap intervals as well, as made clear in Romano (1989). To appreciate why, consider $F_n \in \mathbf{F_0}$ with F_n the distribution that assigns mass p_n to 1 and $1-p_n$ to 0. Fix any θ

in $(\alpha, 1)$ and choose p_n small enough (but not 0) so that $(1 - p_n)^n \geq \theta$. Then, under F_n, a sample of size n will be a sample of all zeroes with probability at least θ. Since the resampled or bootstrap data sets will all be degenerate as well, the resulting bootstrap confidence interval will not include $\mu(F_n) = p_n$. Thus, with probability at least θ, the resulting bootstrap confidence interval will not cover the true mean, and hence the coverage probability is no bigger than $1 - \theta < 1 - \alpha$. In fact, since θ is arbitrary in the argument, the coverage level over $\mathbf{F_0}$ is zero.

A variety of conservative methods are presented in Bickel (1992), including some justification for Stringer's (1963) widely used proposal. However, none of the methods presented there are both conservative in level and efficient. We believe our proposal is the first one that is both nonparametric (or conservative) and efficient.

The paper is organized as follows. In Section 2, we derive a result which can be viewed as an asymptotic efficiency result for conservative confidence intervals for the mean. In essence, the result says that we cannot do better, in terms of length of the interval, than an interval based on a normal approximation. The novel aspect of the result is that its proof draws upon local asymptotic minimax estimation theory by looking at an appropriate least favorable parametric submodel, but intervals are not compared in the usual minimax way of worst case behavior over shrinking neighborhoods. In Section 3, we review a proposal of Anderson (1967), which leads to genuinely nonparametric (conservative) intervals whose length is of the right order, but the constant is too big. His proposal is generalized, though efficiency is not obtained. The reason for considering Anderson's approach is because it makes use of the fact that we can construct genuinely nonparametric confidence bands for the c.d.f. F by the usual Kolmogorov–Smirnov bands. Anderson's procedure uses these bands in a primary way. The efficient construction we present in Section 4 also makes use of the Kolmogorov–Smirnov bands, but somehow the construction relies on these bands in a more secondary way so that efficiency can ensue. A simulation study is presented in Section 5. In Section 6, we present some conclusions, variations, generalizations and directions for future work.

2. An asymptotic lower bound.

In this section, it is more convenient to index the probability distribution generating the data by the measure P (as opposed to the c.d.f. F). Suppose $X_1, \dots, X_n$ are i.i.d. according to a probability P, concentrated on $[0, 1]$; otherwise, nothing else is assumed about P. Inference focuses on the mean $\mu(P)$. The question considered in this section is the following. Among confidence intervals for the mean with guaranteed coverage, what is an asymptotic lower bound for the length of the interval? The following theorem answers the question. In fact, much more is shown. Specifically, the interval must be centered (to an appropriate order) at the sample mean, in order to be efficient. Moreover, an efficient interval, to order $o_P(n^{-1/2})$ behaves like an interval provided by the usual normal approximation, namely, $\overline{X}_n \pm \sigma(P) z_{1-\alpha/2}/n^{1/2}$. However, the normal approximation interval is not conservative, and so can not qualify to be efficient. In the statement of the theorem and in the proof, P^n refers to the product measure of n i.i.d. observations from P.

THEOREM 2.1. *Let $I_n = [L_n, U_n]$ be a sequence of intervals (being measurable functions of $X_1, \dots, X_n$) satisfying*

$$P^n\{\mu(P) \in I_n\} \ge 1 - \alpha$$

for all n and all P. Define $D_n = D_n(I_n) = n^{1/2}(U_n - L_n)/2$. Assume D_n is asymptotically concentrated on $[0, a(P)]$; that is, for every $\varepsilon > 0$, $P^n\{D_n < a(P) + \varepsilon\} \to 1$ as $n \to \infty$. Then:

(i) $a(P) \ge z_{1-\alpha/2}\sigma(P)$.

(ii) *If I_n is an interval such that the lower bound $z_{1-\alpha/2}\sigma(P)$ is attained, then I_n is centered at $\overline{X}_n$ in the sense*

$$n^{1/2}\left[\frac{(U_n + L_n)}{2} - \overline{X}_n\right] \to 0$$

in P^n-probability.

(iii) *If I_n is an interval such that the lower bound $z_{1-\alpha/2}\sigma(P)$ is attained, then $D_n \to z_{1-\alpha/2}\sigma(P)$ in P^n-probability, and so*

$$I_n = \overline{X}_n \pm \frac{\sigma(P) z_{1-\alpha/2}}{n^{1/2}} + o_P(n^{-1/2}).$$

PROOF. Fix $P = P_0$ having mean $\mu_0 = \mu(P_0)$, standard deviation $\sigma_0 = \sigma(P_0)$ and density f (with respect to some σ-finite measure dx). Introduce the parametric submodel $\{P^n_{n,\theta}\}$, where $P^n_{n,\theta}$ is the n-fold product measure of $P_{n,\theta}$, and $P_{n,\theta}$ has density

$$f_{n,\theta} = f(x)\left[1 + \frac{\theta(x - \mu_0)}{n^{1/2}}\right]$$

for $0 \le x \le 1$. Note that for any fixed real number θ, $f_{n,\theta}$ defines a density as soon as $n \ge \theta^2$. Further note that $f_{n,\theta}$ has mean

$$\mu(P_{n,\theta}) = \mu_0 + \theta n^{-1/2} \int_0^1 x(x - \mu_0) f(x)\, dx = \mu_0 + \frac{\theta\sigma_0^2}{n^{1/2}}.$$

By the usual examination of loglikelihood ratios [see Proposition 2.3 of Millar (1983)], it is immediate that the experiments $\{P^n_{n,\theta}\}$ converge to $\{P_\theta\}$, where P_θ is the Gaussian measure on the line with mean $\theta\sigma_0$ and variance 1. Indeed, the loglikelihood ratio at θ versus $\theta = 0$ is simply

$$\log\left[\prod_{i=1}^n f_{n,\theta}(X_i) \Big/ \prod_{i=1}^n f_{n,0}(X_i)\right] = \sum_{i=1}^n \log\left[1 + \frac{\theta(X_i - \mu_0)}{n^{1/2}}\right].$$

Under P_0^n (by a Taylor expansion and the law of large numbers), this loglikelihood ratio is asymptotic to $\theta n^{1/2}(\overline{X}_n - \mu_0) - \theta^2\sigma_0^2/2$, which of course is asymptotically normal with mean $-\theta^2\sigma_0^2/2$ and variance $\theta^2\sigma_0^2$ (implying contiguity). Moreoever, for this parametric submodel, it is then clear that the sample mean $\overline{X}_n$ is then locally asymptotically minimax for the mean.

Now, let l be any bounded subconvex loss function. Then, by the asymptotic minimax theorem [see Millar (1983), page 146],

$$\begin{aligned}\lim_{c\uparrow\infty}\lim_{n\to\infty}\inf_{T_n}\sup_{|\theta|\le c}\int l\big[n^{1/2}(T_n-\mu(P_{n,\theta}))\big]\,dP^n_{n,\theta}\\ =\lim_{c\uparrow\infty}\lim_{n\to\infty}\inf_{T_n}\sup_{|\theta|\le c}\int l\big[\sigma_0^2(T_n-\theta)\big]\,dP^n_{n,\theta}=\inf_{T}\sup_{\theta}\int l\big[\sigma_0^2(T-\theta)\big]\,dP_\theta,\end{aligned}$$

where the infimum over T_n and T refer to the infimum over all estimators available for experiments $P^n_{n,\theta}$ and P_θ, respectively. However, if Q_θ denotes the Gaussian distribution with mean θ and variance 1, the last expression can be evaluated as

$$\inf_{T}\sup_{\theta}\int l\big[\sigma_0(\sigma_0 T-\sigma_0\theta)\big]\,dP_\theta=\inf_{T}\sup_{\theta}\int l\big[\sigma_0(T-\theta)\big]\,dQ_\theta=El(\sigma_0 Z),$$

where Z is a standard normal variable. In particular, let $l=l_d$ be the (subconvex) loss function satisfying $l_d(T-\theta)=1$ if $|T-\theta|>d$ and is zero otherwise. The previous result then says

$$\lim_{c\uparrow\infty}\lim_{n\to\infty}\inf_{T_n}\sup_{|\theta|\le c}P^n_{n,\theta}\Big\{n^{1/2}|T_n-\mu(P_{\theta,n})|>d\Big\}=P\bigg\{|Z|\ge\frac{d}{\sigma_0}\bigg\}\tag{2.1}$$

for any d. Now, given the interval $I_n=[L_n,U_n]$, let $\hat\mu_n$ be the estimator $(L_n+U_n)/2$. Then, by the hypothesis on I_n, we get

$$P^n\big\{n^{1/2}|\hat\mu_n-\mu(P)|\le D_n\big\}\ge 1-\alpha.$$

Fix $a=a(P)$. We now claim that, for any P and any $\varepsilon>0$,

$$P^n\big\{n^{1/2}|\hat\mu_n-\mu(P)|\le a+\varepsilon\big\}\ge(1-\alpha)P^n\{D_n\le a+\varepsilon\}.\tag{2.2}$$

To prove (2.2),

$$\begin{aligned}&P^n\big\{n^{1/2}|\hat\mu_n-\mu(P)|\le a+\varepsilon\big\}\\ &\quad\ge P^n\big\{n^{1/2}|\hat\mu_n-\mu(P)|\le a+\varepsilon|D_n\le a+\varepsilon\big\}P^n\big\{D_n\le a+\varepsilon\big\}\\ &\quad\ge P^n\big\{n^{1/2}|\hat\mu_n-\mu(P)|\le D_n\big\}P^n\big\{D_n\le a+\varepsilon\big\}\ge(1-\alpha)P^n\big\{D_n\le a+\varepsilon\big\},\end{aligned}$$

and (2.2) follows. Now, under P_0, $P^n_0\{D_n\le a+\varepsilon\}\to 1$, by assumption. But, for any sequence $\{\theta_n\}$ with $|\theta_n|\le c$, $\{P^n_{n,\theta_n}\}$ is contiguous to $\{P^n_0\}$. It then follows that $P^n_{n,\theta_n}\{D_n\le a+\varepsilon\}\to 1$ as well. Hence, by (2.2),

$$\liminf_{n\to\infty}P^n_{n,\theta_n}\big\{n^{1/2}|\hat\mu_n-\mu(P_{n,\theta_n})|\le a+\varepsilon\big\}\ge 1-\alpha.$$

Therefore (by a subsequence argument), for any $\varepsilon>0$ and any fixed c, we have

$$\limsup_{n\to\infty}\sup_{|\theta|\le c}P^n_{n,\theta}\big\{n^{1/2}|\hat\mu_n-\mu(P_{n,\theta})|>a+\varepsilon\big\}\le\alpha.\tag{2.3}$$

Now, to prove the result $a \geq z_{1-\alpha/2}\sigma_0$, assume the opposite and let $\Delta = z_{1-\alpha/2}\sigma_0 - a$. Choose ε so that $(\Delta - \varepsilon)/\sigma_0 \equiv y > 0$. Then, choose δ small enough so that

$$P\{|Z| > z_{1-\alpha/2} - y\} - \delta > \alpha.$$

Finally, choose c large enough so that the result (2.1) with $d = a + \varepsilon$ gives

$$\lim_{n\to\infty} \inf_{T_n} \sup_{|\theta|\leq c} P^n_{n,\theta}\left\{n^{1/2}|T_n - \mu(P_{n,\theta})| > a + \varepsilon\right\} \geq P\left\{|Z| \geq \frac{a+\varepsilon}{\sigma_0}\right\} - \delta. \tag{2.4}$$

But the right-hand side of (2.4) is greater than or equal to

$$P\left\{|Z| \geq \frac{z_{1-\alpha/2}\sigma_0 - \Delta + \varepsilon}{\sigma_0}\right\} - \delta = P\{|Z| \geq z_{1-\alpha/2} - y\} - \delta > \alpha,$$

by choice of δ. This yields a contradiction because by (2.3), the estimator $\hat{\mu}_n$ would have a smaller asymptotic minimax risk than the bound of α which is valid for any estimator T_n given by (2.4).

To prove (ii), the result (2.3) implies that $\hat{\mu}_n$ is almost locally asymptotic minimax, which would be immediate if (2.3) were true with ε replaced by 0. The result would then follow by Theorem 4.1 of Hajek (1972). Since we cannot just replace ε by 0 in (2.3), we sketch an argument based on the proof of (4.3) in Hajek (1972). For notational simplicity, assume $\sigma_0 = 1$. Assume $n^{1/2}(\hat{\mu}_n - \overline{X}_n)$ fails to converge to 0 in P-probability. Then, there exists an $\bar{\varepsilon} > 0$ such that for all large m,

$$P^m_m\left(m^{1/2}|\hat{\mu}_m - \overline{X}_m| > \bar{\varepsilon}\right) > \bar{\varepsilon}.$$

If Z denotes an observation from the model $\{P_\theta\}$, then by Lemmas 3.1 and 3.2 of Hajek (1972), this would imply the existence of an estimator sequence $\xi_m(Z, U)$ (possibly depending on an auxiliary randomization U, independent of Z), satisfying

$$P^m_m\{|\xi_m(Z, U) - Z| > \bar{\varepsilon}\} > \bar{\varepsilon}$$

for all m large enough, and $\xi_m(Z, U) - \theta$ is constructed to have the same asymptotic properties as $m^{1/2}(\hat{\mu}_m - \mu(P_{\theta,m}))$; see (4.8) of Hajek. But by Lemma 2.1 of Hajek, there exists $\beta > 0$ depending only on $\bar{\varepsilon}$ so that the maximum risk of $\xi_m(Z, U)$ for the loss function $l_{z_{1-\alpha/2}+\varepsilon}$ is at least $\alpha + \beta$. But, the maximum risk of $\xi_m(Z, U)$ for the loss function $l_{z_{1-\alpha/2}+\varepsilon}$ is asymptotically [by our (2.3)] less than or equal to α, which is within $(2/\pi)^{1/2}\varepsilon$ of the best obtainable rate, since

$$|\alpha - P\{|Z| \geq z_{1-\alpha/2} + \varepsilon\}| \leq (2/\pi)^{1/2}\varepsilon.$$

Hence, by choosing ε small enough so that $(2/\pi)^{1/2}\varepsilon < \beta$, we get a contradiction.

Finally, to prove (iii), writing μ for $\mu(P)$,

$$\begin{aligned}1-\alpha \ge P\{L_n \le \mu \le U_n\} = P\{&n^{1/2}(\overline{X}_n - U_n) \le n^{1/2}(\overline{X}_n - \mu) \\ &\le n^{1/2}(\overline{X}_n - L_n)\}.\end{aligned} \tag{2.5}$$

But, by result (ii), $n^{1/2}(\overline{X}_n - L_n) = D_n + o_P(1)$. Therefore, invoking the asymptotic normality of $n^{1/2}(\overline{X}_n - \mu)$, the probability on the right side of (2.5) is

$$P\{-D_n \le n^{1/2}(\overline{X}_n - \mu) \le D_n\} + o_P(1).$$

But, the fact that D_n is asymptotically concentrated on $[0, z_{1-\alpha/2}\sigma(P)]$ and the asymptotic normality of $n^{1/2}(\overline{X}_n - \mu)$ with asymptotic variance $\sigma^2(P)$ forces $D_n \to z_{1-\alpha/2}\sigma(P)$ in P^n-probability. □

REMARK 2.1. If a conservative interval I_n is constructed such that the lower bound $a(P) = z_{1-\alpha/2}\sigma(P)$ is obtained for every fixed P, then we will call I_n efficient; otherwise, we will call the interval inefficient.

REMARK 2.2. Surprisingly, there is not a vast literature on efficiency theory in the construction of confidence intervals. Beran and Millar (1985) is a notable exception. They develop an asymptotic theory for confidence sets in a decision theoretic local asymptotic minimax framework. More specifically, a loss function is introduced to measure the performance of an interval I_n for a real-valued parameter θ (and more generally for abstract parameter sets) as follows. If μ is the true parameter and interval I_n is used, the loss is $g(n^{1/2}\sup_{y\in I_n}|y-\mu|)$; here, g is an increasing function on the positive reals. Evidently, the loss penalizes if the interval is too wide or if it is miscentered. Now, once a loss function is introduced, different procedures (satisfying the constraint on level) can be compared by comparing their risk functions, by looking at the maximum risk over shrinking neighborhoods as in the local asymptotic minimax theory for estimation problems. Note that our simple result (while indeed making use of local asymptotic minimax ideas) does not compare procedures by looking at the maximum risk over shrinking neighborhoods. The assumption that D_n is asymptotically concentrated on $[0, a]$ need only hold under a single law P, and then the conclusion is that the asymptotic constant a must be no less than $z_{1-\alpha/2}\sigma(P)$, under P. Issues like superefficiency do not come into play once we impose the coverage constraint, which must hold for all P for I_n to even be considered. Moreover, the optimality result we achieve is based directly on the length of the interval. In essence, however, our result is an asymptotic admissibility and asymptotic minimax result of sorts. Indeed, one can construct an interval I_n whose standardized half-length D_n is strictly less than the bound $z_{1-\alpha/2}\sigma(P)$ with probability tending to some number $p \in (0, 1)$. The number p can be positive but it cannot be 1 (by the theorem), so that shorter intervals are possible but only at the expense of intervals that must then be bigger some of the time. That is, our result can be viewed as a minimax result because it says that no interval (sequence) can have a half-length that is asymptotically concentrated on a *smaller* set. Furthermore, if an interval achieves the lower bound on the length in the sense of part (i) of Theorem 2.1, then it *already* must be appropriately centered in the sense of (ii) of the theorem.

REMARK 2.3. The theorem generalizes to parameters $\theta(P)$ other than the mean. If the functional $\theta(\cdot)$ is appropriately differentiable, then, in a nonparametric setting, efficient intervals must behave like

$$\theta(\hat{P}_n) \pm z_{1-\alpha/2}\tau(P)/n^{1/2} + o_P(n^{-1/2}),$$

where $\hat{P}_n$ is the empirical measure and $\tau^2(P)$ is the asymptotic variance of $n^{1/2}\theta(\hat{P}_n)$ under P.

3. Inefficient methods with guaranteed coverage. The possibility of finding confidence intervals for the mean with guaranteed coverage, but which are not too big, seems plausible given the following construction, due to Anderson (1967) [and later rediscovered by Breth, Maritz and Williams (1978)]. At this point, we switch notation and index the probability distribution generating the sample $X_1, \dots, X_n$ by the cumulative distribution function (c.d.f.) F. Then, $\mu(F) \equiv E_F(X_i)$. Here, all we are assuming is that $F \in \mathbf{F_0}$, the class of all c.d.f.'s supported on [0, 1]. Let $\hat{F}_n$ be the empirical c.d.f. For c.d.f.'s F and G, let the sup (Kolmogorov) distance d_K be defined by

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

Let $\hat{R}_{n,1-\alpha}$ be the Kolmogorov–Smirnov uniform confidence band for F of nominal level $1 - \alpha$ defined by

$$\hat{R}_{n,1-\alpha} = \{F \in \mathbf{F_0}\colon n^{1/2} d_K(\hat{F}_n, F) \le c_n(1-\alpha)\}, \tag{3.1}$$

where $c_n(1-\alpha)$ is the $1-\alpha$ quantile of the distribution of $n^{1/2}d_K(\hat{F}_n, F)$ under F when F is any continuous distribution. Note that, for any F (discrete or otherwise),

$$P_F\{F \in \hat{R}_{n,1-\alpha}\} \ge 1 - \alpha;$$

the inequality is an equality iff F is continuous. This leads to a nominal level $1 - \alpha$ confidence interval $I_{n,0}$ for $\mu(F)$ as follows. In words, the value μ is included in $I_{n,0}$ if there is some distribution F in $\hat{R}_{n,1-\alpha}$ that has mean μ. Then, the event $\{F \in \hat{R}_{n,1-\alpha}\}$ implies $\{\mu(F) \in I_{n,0}\}$ and so

$$P_F\{\mu(F) \in I_{n,0}\} \ge 1 - \alpha.$$

In fact, $I_{n,0} = \overline{X}_n \pm O_P(n^{-1/2})$, which follows from the following simple proposition.

PROPOSITION 3.1. *Let $m_k(F) = E_F(X_i^k)$. If F and G are in $\mathbf{F_0}$ and $d_K(F, G) \le \varepsilon$, then $|m_k(F) - m_k(G)| \le \varepsilon$.*

PROOF. By integration by parts,

$$\begin{aligned}|m_k(F) - m_k(G)| &= \left|\int_0^1 x^k d(F-G)(x)\right| \\ &= \left|\int_0^1 (F(x)-G(x))kx^{k-1}\,dx\right| \le \varepsilon k\int_0^1 x^{k-1}\,dx = \varepsilon.\end{aligned}$$

It follows immediately from Proposition 3.1 that the interval $I_{n,0}$ is contained in the interval $\overline{X}_n \pm c_n(1-\alpha)/n^{1/2}$. The claim that $I_{n,0} = \overline{X}_n \pm O_P(n^{-1/2})$ follows since $c_n(1-\alpha) \to c(1-\alpha)$, where $c(1-\alpha)$ is the upper $1-\alpha$ quantile of the distribution of $\sup_{0<t<1}|B(t)|$, where $B(\cdot)$ is a Brownian bridge process.

In fact, it can be argued that if F is uniform on $(0,1)$

$$I_{n,0} = \overline{X}_n \pm c(1-\alpha)/n^{1/2} + o_P(n^{-1/2}).$$

The constant $c(1-\alpha)$ is too big and should be compared with $z_{1-\alpha/2}\sigma(F)$. For example, with $\alpha = 0.05$, $c(0.95) \doteq 1.36$, while $z_{1-\alpha/2}\sigma(F) \doteq 1.96\sigma(F) \le 0.98$ for all F. In particular, when F is the uniform distribution on (0, 1), the asymptotically best constant is $1.96/(12)^{1/2} \doteq 0.57$. The ratio $1.36/0.57 \doteq 2.4$ measures the inefficiency of Anderson's procedure. Indeed, a sample of approximately $(2.4)^2 n = 5.76n$ is needed when using Anderson's procedure to be as efficient as an efficient procedure based on a sample of size n.

Anderson's method, with the hope of improved efficiency, is generalized in Gasko (1991) (who treats the one-sided case) as follows. If $\underline{x}_n = (X_1, \dots, X_n)$ is a sample of size n from F, let $X_{i:\,n}$ denote the ith order statistic. Define $\pi_i = \pi_i(\underline{x}_n, F) = F(X_{i+1:\,n}) - F(X_{i:\,n})$ for $i = 1, \dots, n-1$; also, let $\pi_0 = F(X_{1:\,n})$ and $\pi_n = 1 - F(X_{n:\,n})$. Then, $\underline{\pi} = \underline{\pi}(\underline{x}_n, F)$ is a random point in the $n+1$ simplex Σ_n in $\mathbf{R}^{n+1}$. Note that the distribution of $\underline{\pi}(\underline{x}_n, F)$ under F is the same for all continuous F.

Now, let $S_{n,1-\alpha} \subset \Sigma_n$ be a region of probability content at least $1-\alpha$, meaning

$$P_F\{\underline{\pi}(\underline{x}_n, F) \in S_{n,1-\alpha}\} \ge 1-\alpha \tag{3.2}$$

for all F. Then, the region $S_{n,1-\alpha}$ induces a confidence set $C_{n,1-\alpha}$ for F by including F in $C_{n,1-\alpha}$ if $\underline{\pi}(\underline{x}_n, F) \in S_{n,1-\alpha}$. Clearly,

$$P_F\{F \in C_{n,1-\alpha}\} \ge 1-\alpha,$$

and so $C_{n,1-\alpha}$ is a level $1-\alpha$ confidence set for F.

We can now proceed as before to construct an interval I_n for the mean. That is, a value μ belongs in the interval I_n if there is some distribution $F \in C_{n,1-\alpha}$ with mean μ. Then, $P_F\{\mu(F) \in I_n\} \ge 1-\alpha$. However, regardless of the choice of region $S_{n,\alpha}$ the resulting interval I_n is not efficient.

PROPOSITION 3.2. *Any region $S_{n,1-\alpha}$ satisfying* (3.2) *leads to a conservative $1-\alpha$ confidence interval I_n for $\mu(F)$. However, no choice of $S_{n,1-\alpha}$ leads to efficiency.*

PROOF. Fix (a sequence of regions) $S_{n,1-\alpha}$ satisfying (3.2), and let I_n denote the induced confidence interval for $\mu(F)$. Assume efficiency holds; then, by Theorem 2.1(iii), we must have

$$I_n = \overline{X}_n \pm \frac{z_{1-\alpha/2}\sigma(F)}{n^{1/2}} + O_P(n^{-1/2})$$

for all F. Now, $S_{n,1-\alpha}$ also induced a level $1-\alpha$ confidence interval, M_n, for the parameter $m(F) = E_F(X_i^2)$ by a similar prescription: a value m is included in M_n if there is an F in $C_{n,1-\alpha}$ with $m(F) = m$. By construction, we in fact have that $(\mu(F), m(F)) \in (I_n, M_n)$ with probability at least $1-\alpha$ for all F.

We now argue that the interval for $m(F)$ based on data $\underline{x}_n$ is the same as the interval for $\mu(F)$ based on data $\underline{y}_n$, where $\underline{y}_n = (Y_1, \dots, Y_n)$ and $Y_i = X_i^2$. To appreciate why this is so, if the interval M_n for $m(F)$ based on $\underline{x}_n$ includes a value m, then there is an F in $C_{n,1-\alpha}$ with $m(F) = m$ and $\underline{\pi}(\underline{x}_n, F) \in S_{n,1-\alpha}$. But then, $\underline{\pi}(\underline{y}_n, \widetilde{F}) \in S_{n,1-\alpha}$, where $\widetilde{F}(\cdot) = F(\cdot^{1/2})$ is the distribution of $Y_i = X_i^2$ if X_i has c.d.f. F. This follows because

$$F(X_{i:\,n}) = F(Y_{i:\,n}^{1/2}) = \widetilde{F}(Y_{i:\,n}).$$

Hence, based on data $\underline{y}_n$, the value

$$\int y\, d\widetilde{F}(y) = \int x^2\, dF(x) = m$$

is included in the confidence interval for the mean. Conversely, if the interval for the mean based on $\underline{y}_n$ includes m, so does the interval for the second moment based on $\underline{x}_n$. Therefore, the hypothesis that I_n is efficient entails M_n is efficient and so,

$$M_n = \overline{Y}_n \pm \frac{z_{1-\alpha/2}\sigma(\widetilde{F})}{n^{1/2}} + o_P(n^{-1/2}).$$

We are now in a position to arrive at a contradiction. By the bivariate central limit theorem,

$$\begin{aligned} 1-\alpha &\le P_F\{\mu(F) \in I_n,\ m(F) \in M_n\} \\ &= P\biggl\{ n^{1/2}|\overline{X}_n - \mu(F)| \le z_{1-\alpha/2}\sigma(F),\ n^{1/2} \\ &\qquad \times \left| n^{-1}\sum_{i=1}^n X_i^2 - m(F)\right| \le z_{1-\alpha/2}\sigma(\widetilde{F})\biggr\} + o_P(1). \end{aligned}$$

Now, the last expression converges to

$$P\{|Z_1| \le z_{1-\alpha/2}\sigma(F),\ |Z_2| \le z_{1-\alpha/2}\sigma(\widetilde{F})\},$$

where (Z_1, Z_2) is bivariate normal, mean 0, $\mathrm{Var}(Z_1) = \sigma^2(F)$, $\mathrm{Var}(Z_2) = \sigma^2(\widetilde{F})$, and covariance $\mathrm{Cov}(X_1, X_1^2)$. Hence,

$$\begin{aligned}1-\alpha &\le P\{|Z_1| \le z_{1-\alpha/2}\sigma(F),\ |Z_2| \le z_{1-\alpha/2}\sigma(\widetilde{F})\}\\ &= P\{|Z_1| \le z_{1-\alpha/2}\sigma(F)\} - P\{|Z_1| \le z_{1-\alpha/2}\sigma(F),\ |Z_2| > z_{1-\alpha/2}\sigma(\widetilde{F})\}\\ &= (1-\alpha) - P\{|Z_1| \le z_{1-\alpha/2}\sigma(F),\ |Z_2| \ge z_{1-\alpha/2}\sigma(\widetilde{F})\}.\end{aligned}$$

This last probability can only be 0 if Z_1 and Z_2 are perfectly correlated, which means that X_1 and X_1^2 must be linearly dependent with probability 1. Clearly, this is not the case when X_1 is uniform on (0, 1), and so a contradiction is obtained.

REMARK 3.1. An alternative way to generalize Anderson's idea is just to change the metric. The same basic argument in the proof of Proposition 3.2 would show inefficiency.

REMARK 3.2. While no particular region $S_{n,1-\alpha}$ leads to efficiency, Gasko (1991) argues that improvements in efficiency at a particular $F_0 \in \mathbf{F_0}$ are possible, and that perhaps F_0 can be chosen adaptively. Some simulations show some improvement over Anderson's method, but no proof of efficiency is supplied (and it seems doubtful that one exists).

4. An efficient construction for the mean. Let $X_1, \dots, X_n$ be i.i.d. with c.d.f. F, mean $\mu(F)$, variance $\sigma^2(F)$, and set $\rho(F) \equiv E_F[|X_i - \mu(F)|^3]$. The unknown c.d.f. F is assumed to be in $\mathbf{F_0}$. Again, the goal is to construct a confidence I_n for $\mu(F)$ that contains $\mu(F)$ with probability at least $1-\alpha$ for every F; in addition, the interval I_n must be efficient. In particular, the interval must satisfy the square root of the sample size multiplied by the length of the interval tends in probability to $2z_{1-\alpha/2}\sigma(F)$, for every $F \in \mathbf{F_0}$.

Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $J_n(F)$ be the distribution of $n^{1/2}(\overline{X}_n - \mu(F))$ under F, with corresponding c.d.f.,

$$J_n(x, F) = P_F\{n^{1/2}(\overline{X}_n - \mu(F)) \le x\}.$$

Also, let

$$d_n(\alpha, F) \equiv \inf\{x\colon J_n(x, F) \ge \alpha\}.$$

Note that $J_n(d_n(\alpha, F), F) \ge \alpha$ and $J_n(d_n(\alpha, F)^-, F) \le \alpha$. Suppose β_n is a sequence of numbers in [0, 1] converging to 0. Let $\hat{R}_{1,1-\beta_n}$ be defined by (3.1) with α there replaced by β_n. Define

$$\hat{d}_{n,U}\left(1-\frac{\alpha}{2}\right) \equiv \sup\left\{d_n\left(1-\frac{\alpha}{2}+\beta_n, F\right)\colon F \in \hat{R}_{n,1-\beta_n}\right\} \tag{4.1}$$

and

$$\hat{d}_{n,L}\left(\frac{\alpha}{2}\right) \equiv \inf\left\{d_n\left(\frac{\alpha}{2}-\beta_n, F\right)\colon F \in \hat{R}_{n,1-\beta_n}\right\}. \tag{4.2}$$

The interval we first propose is defined by

$$I_{n,1} = \left\{\mu: \hat{d}_{n,L}\left(\frac{\alpha}{2}\right) \le n^{1/2}(\overline{X}_n - \mu) \le \hat{d}_{n,U}\left(1 - \frac{\alpha}{2}\right)\right\}. \tag{4.3}$$

This construction leads to an interval which, in fact, is determined by two one-sided conservative level $1 - \alpha/2$ intervals; as such, we are tacitly assuming $\alpha \le 0.5$. At this point, it is certainly not clear how to compute $I_{n,1}$ because of the sup in the definition (4.1), which is a sup over an infinite-dimensional set. For now, we postpone this computational issue and present the following result. Note, however, that in the course of analyzing $I_{n,1}$, we will derive further more conservative intervals (which are efficient, too), but which are directly computable.

THEOREM 4.1. *For each n and all $F \in \mathbf{F_0}$,*

$$P_F\{\mu(F) \in I_{n,1}\} \ge 1 - \alpha.$$

PROOF.

$$\begin{aligned} P_F\{\mu(F) \in I_{n,1}\} &\ge P_F\left\{\hat{d}_{n,L}\left(\frac{\alpha}{2}\right) \le n^{1/2}(\overline{X}_n - \mu(F))\right.\\ &\qquad\qquad \left.\le \hat{d}_{n,U}\left(1 - \frac{\alpha}{2}\right) \bigcap F \in \hat{R}_{n,1-\beta_n}\right\}\\ &\ge P_F\left\{d_n\left(\frac{\alpha}{2} - \beta_n, F\right) \le n^{1/2}(\overline{X}_n - \mu(F))\right.\\ &\qquad\qquad \left.\le d_n\left(1 - \frac{\alpha}{2} + \beta_n, F\right) \bigcap F \in \hat{R}_{n,1-\beta_n}\right\}\\ &\ge P_F\left\{d_n\left(\frac{\alpha}{2} - \beta_n, F\right) \le n^{1/2}(\overline{X}_n - \mu(F))\right.\\ &\qquad\qquad \left.\le d_n\left(1 - \frac{\alpha}{2} + \beta_n, F\right)\right\} - P_F\{F \notin \hat{R}_{n,1-\beta_n}\right\}, \end{aligned} \tag{4.4}$$

where we have used the trivial inequality $P(A \cap B) \ge P(A) - P(B^c)$. However, (4.4) is bounded below by

$$(1 - \alpha + \beta_n) - P_F\{F \notin \hat{R}_{n,1-\beta_n}\} \ge 1 - \alpha$$

and the proof is complete. □

REMARK 4.1. Setting $\beta_n = 0$ in the construction of $I_{n,1}$ would lead to a conservative interval, but not one which is efficient.

REMARK 4.2. The above construction is related to the constructions presented in Loh (1985) and Silvapulle (1996), both of whom consider parametric testing problems. Note, however, that Loh's construction in essence replaces the value $d_n(1-\alpha/2+\beta_n, F)$ in the definition (4.1) by $d_n(1-\alpha/2, F)$, which would not lead to a conservative interval. The goal here is not just to produce conservative intervals, but efficient ones in a nonparametric setting.

THEOREM 4.2. *Suppose* $\beta_n > 0$ *and* $\beta_n \to 0$ *in such a way that* $\log(\beta_n)/n \to 0$ *as* $n \to \infty$. *Then,* $\hat{d}_{n,U}(1-\alpha/2) \to \sigma(F)z_{1-\alpha/2}$ *in probability under* F; *similarly,* $\hat{d}_{n,L}(\alpha/2) \to \sigma(F)z_{\alpha/2}$ *in probability under* F. *Therefore,* $I_{n,1}$ *is efficient in the sense of Theorem* 2.1.

In order to prove Theorem 4.2, we seek an upper bound for $\hat{d}_{n,U}(1-\alpha/2)$ which is analytically tractable and which tends to $\sigma(F)z_{1-\alpha/2}$ in probability under F. At the same time, we constructively derive an upper bound which may be computed explicity.

First, let

$$\Delta_n(F) = C_{\mathrm{BE}} n^{-1/2}\rho(F)\sigma^{-3}(F), \tag{4.5}$$

where C_{BE} is the smallest known universal constant valid in the Berry–Esseen Theorem. Then, let $\delta_{n,\alpha}(F)$ be defined as follows:

$$\delta_{n,\alpha}(F) = \inf\{\delta > 0\colon \Phi(z_{1-\alpha/2+\beta_n}(1+\delta)) - \Phi(z_{1-\alpha/2+\beta_n}) \ge \Delta_n(F)\}. \tag{4.6}$$

In (4.6), $\Phi(\cdot)$ denotes the standard normal c.d.f.

PROPOSITION 4.1. *The following bound holds*:

$$d_n\left(1 - \frac{\alpha}{2} + \beta_n, F\right) \le z_{1-\alpha/2+\beta_n}(1+\delta_{n,\alpha}(F))\sigma(F). \tag{4.7}$$

Therefore,

$$\hat{d}_{n,U}\left(1-\frac{\alpha}{2}\right) \le z_{1-\alpha/2+\beta_n}\sup\{(1+\delta_{n,\alpha}(F))\sigma(F)\colon F \in \hat{R}_{n,1-\beta_n}\} \tag{4.8}$$

and

$$\hat{d}_{n,L}\left(\frac{\alpha}{2}\right) \ge z_{\alpha/2-\beta_n}\inf\{(1+\delta_{n,\alpha}(F))\sigma(F)\colon F \in \hat{R}_{n,1-\beta_n}\}. \tag{4.9}$$

PROOF OF PROPOSITION 4.1. Let $r \equiv r_{n,\alpha}(F) = 1 + \delta_{n,\alpha}(F)$. By the Berry–Esseen theorem,

$$|J_n(z_{1-\alpha/2+\beta_n}\sigma(F)r, F) - \Phi(z_{1-\alpha/2+\beta_n}r)| \le \Delta_n.$$

By definition of $\delta_{n,\alpha}(F)$ (and hence r),

$$\Phi(z_{1-\alpha/2+\beta_n}r) - \left(1 - \frac{\alpha}{2} + \beta_n\right) \ge \Delta_n.$$

Therefore, by the triangle inequality,

$$J_n(z_{1-\alpha/2+\beta_n}\sigma(F)r, F) - \left(1 - \frac{\alpha}{2} + \beta_n\right) \ge 0,$$

and (4.7) follows, as then do (4.8) and (4.9).

Now, in order to use the bounds (4.8) and (4.9), we first need to understand the behavior of $\delta_{n,\alpha}(F)$. Indeed, $\delta_{n,\alpha}(F)$ is order $n^{-1/2}$ in probability; in fact, the following is true. Of course, $\phi(\cdot)$ denotes the standard normal density.

LEMMA 4.1. *For all n large enough,*

$$\delta_{n,\alpha}(F) \le \Delta_n(F)[\phi^2(z_{1-\alpha/2+\beta_n}) - 2(2\pi e)^{-1/2}\Delta_n(F)]^{-1/2}/z_{1-\alpha/2+\beta_n};$$

the bound holds for n satisfying $\phi^2(z_{1-\alpha/2+\beta_n}) > 2(2\pi e)^{-1/2}\Delta_n(F)$, which holds for all large n.

The proof of the lemma will be deferred to the Appendix. Now, define $\hat{\Delta}_n$ and $\hat{\delta}_n$ to be

$$\hat{\Delta}_n = \sup\{\Delta_n(F)\colon F \in \hat{R}_{n,1-\beta_n}\} \tag{4.10}$$

and

$$\hat{\delta}_n = \inf\{\delta > 0\colon \Phi(z_{1-\alpha/2+\beta_n}(1+\delta)) - \Phi(z_{1-\alpha/2+\beta_n}) \ge \hat{\Delta}_n\}. \tag{4.11}$$

Also, define

$$\hat{\sigma}_{n,U} = \sup\{\sigma(F)\colon F \in \hat{R}_{n,1-\beta_n}\}. \tag{4.12}$$

From (4.8) and (4.9), this leads to the following bounds.

PROPOSITION 4.2.

$$\hat{d}_{n,U}\left(1-\frac{\alpha}{2}\right) \le z_{1-\alpha/2+\beta_n}(1+\hat{\delta}_n)\hat{\sigma}_{n,U} \tag{4.13}$$

$$\hat{d}_{n,L}\left(\frac{\alpha}{2}\right) \ge z_{\alpha/2-\beta_n}(1+\hat{\delta}_n)\hat{\sigma}_{n,U}. \tag{4.14}$$

Note, from Lemma 4.1, we have

$$\hat{\delta}_n \le \hat{\Delta}_n[\phi^2(z_{1-\alpha/2+\beta_n}) - 2(2\pi e)^{-1/2}\hat{\Delta}_n]^{-1/2}/z_{1-\alpha/2+\beta_n} \tag{4.15}$$

as soon as the term in brackets is nonnegative. So, to bound $\hat{\delta}_n$ in (4.13) and (4.14), it suffices to understand the behavior of $\hat{\Delta}_n$. But, using the (crude) inequality $\rho(F) \le \sigma^2(F)$ (for $F \in \mathbf{F_0}$), we get

$$\hat{\Delta}_n \le C_{\mathrm{BE}} n^{-1/2} \sup\{\sigma^{-1}(F)\colon F \in \hat{R}_{n,1-\beta_n}\} = C_{\mathrm{BE}} n^{-1/2}\hat{\sigma}_{n,L}^{-1}, \tag{4.16}$$

where $\hat{\sigma}_{n,L}$ is defined by

$$\hat{\sigma}_{n,L} = \inf\{\sigma(F)\colon F \in \hat{R}_{n,1-\beta_n}\}. \tag{4.17}$$

Therefore, to complete our series of successive approximations, it is only necessary to bound $\sigma(F)$ (in both directions) as F varies in $\hat{R}_{n,1-\beta_n}$. We will appeal to the following lemma.

LEMMA 4.2. *Fix F and G, c.d.f.'s in $\mathbf{F_0}$. If $d_K(F, G) \le \varepsilon$, then*

$$\left|\sigma^2(F) - \sigma^2(G)\right| \le 3\varepsilon.$$

Therefore,

$$|\sigma(F) - \sigma(G)| \le (3\varepsilon)^{1/2} \tag{4.18}$$

or

$$|\sigma(F) - \sigma(G)| \le 3\varepsilon/\sigma(F). \tag{4.19}$$

For the proof, apply Proposition 3.1.

Now, for $F \in \hat{R}_{n, 1-\beta_n}$, it follows from (4.19) [noting that we could have employed (4.18)] that

$$\left|\sigma(\hat{F}_n) - \sigma(F)\right| \le \frac{3c_n(1-\beta_n)}{n^{1/2}\sigma(\hat{F}_n)} \tag{4.20}$$

and so the following is true.

PROPOSITION 4.3.

$$\hat{\sigma}_{n,U} \le \sigma(\hat{F}_n) + \frac{3c_n(1-\beta_n)}{n^{1/2}\sigma(\hat{F}_n)}, \tag{4.21}$$

$$\hat{\sigma}_{n,L} \ge \sigma(\hat{F}_n) - \frac{3c_n(1-\beta_n)}{n^{1/2}\sigma(\hat{F}_n)}. \tag{4.22}$$

THEOREM 4.3. *Suppose β_n satisfies $\log(\beta_n)/n \to 0$ as $n \to \infty$. Let*

$$I_{n,2} = \overline{X}_n \pm n^{-1/2} z_{1-\alpha/2+\beta_n}(1+\hat{\delta}_n)\hat{\sigma}_{n,U}. \tag{4.23}$$

Then $I_{n,2}$ contains $I_{n,1}$ (hence is conservative) and $I_{n,2}$ (hence $I_{n,1}$) is efficient in the sense of Theorem 2.1.

PROOF OF THEOREMS 4.2 AND 4.3. It suffices to show $\hat{\delta}_n \to 0$ in probability under F and $\hat{\sigma}_{n,U} \to \sigma(F)$ in probability under F. First, to show $\hat{\sigma}_{n,U} \to \sigma(F)$ in probability, by the law of large numbers and (4.21), it suffices to show $c_n(1-\beta_n)/n^{1/2} \to 0$. But, by the Dvoretsky–Kiefer–Wolfowitz inequality, there is a universal constant C_{DKW} such that

$$P_F\left\{n^{1/2} d_K(\hat{F}_n, F) > t\right\} \le C_{\text{DKW}} \exp(-2t^2),$$

so that $c_n(1-\beta_n)$ satisfies

$$C_{\text{DKW}} \exp(-2c_n^2(1-\beta_n)) \ge \beta_n,$$

which implies

$$\frac{c_n^2(1-\beta_n)}{n} \le \frac{1}{2n} \log\left(\frac{C_{\text{DKW}}}{\beta_n}\right).$$

The right-hand side tends to 0 by our hypothesis on β_n. Thus, $\hat{\sigma}_{n,U} \to \sigma(F)$ in probability. By the same reasoning, $\hat{\sigma}_{n,L} \to \sigma(F)$ in probability. Therefore, by (4.16), $\hat{\Delta}_n \to 0$ in probability, and by (4.15), $\hat{\delta}_n \to 0$ in probability. The proof is complete. □

REMARK 4.3. Actually, the bounds used in the proof show much more. Specifically,

$$|\hat{\sigma}_{n,U} - \sigma(F)| = O_P\left(\frac{c_n(1-\beta_n)}{n^{1/2}}\right) = O_P\left(\left|\frac{\log(n)}{n}\right|^{1/2}\right)$$

if, for example, β_n satisfies $\beta_n = n^{-p}$ for any $p > 0$. Furthermore, $|z_{1-\alpha/2+\beta_n} - z_{1-\alpha/2}| = O(\beta_n)$. Also, by (4.16) and (4.20), $\hat{\Delta}_n = O_P[(\log(n)/n)^{1/2}]$, and so $\hat{\delta}_n$ is this same order, by Lemma 4.1. Hence, for $j = 1, 2$,

$$\begin{aligned} I_{n,j} = \overline{X}_n \pm n^{-1/2}\bigl(z_{1-\alpha/2} + O_P(\beta_n)\bigr)\bigl(1 + O_P[(\log(n)/n)^{1/2}]\bigr) \\ \times \left(\sigma(F) + O_P\left(\left|\frac{\log(n)}{n}\right|^{1/2}\right)\right) \end{aligned}$$

or

$$I_{n,j} = \overline{X}_n \pm n^{-1/2} z_{1-\alpha/2}\sigma(F) + O_P\left(\frac{\beta_n}{n^{1/2}} + \frac{[\log(\beta_n)]^{1/2}}{n}\right). \tag{4.24}$$

Taking $\beta_n = O(n^{-1/2})$ then yields the following corollary.

COROLLARY 4.1. *If $\beta_n = O(n^{-1/2})$, then*

$$I_{n,j} = \overline{X}_n \pm n^{1/2} z_{1-\alpha/2}\sigma(F) + O_P\left(\frac{\log(n)}{n}\right). \tag{4.25}$$

Of course, the intervals $I_{n,j}$ are conservative in level, but asymptotically satisfy

$$P_F\{\mu(F) \in I_{n,j}\} \to 1-\alpha$$

as $n \to \infty$, for any fixed nondegenerate F. In fact, the proof shows a much stronger statement. Let $\mathbf{F}_\tau$ be all distributions F on [0, 1] with $\sigma(F) \geq \tau$. Then, the arguments in this section show the following.

COROLLARY 4.2. *For any $\tau > 0$ and for $j = 1, 2$,*

$$\sup_{F\in\mathbf{F}_\tau}\bigl[P_F\{\mu(F) \in I_{n,j}\} - (1-\alpha)\bigr] \to 0$$

as $n \to \infty$.

REMARK 4.4. By now, the constructive approach used clearly shows efficient intervals can even be computed by hand. Indeed, employing even cruder approximations while still retaining efficiency, use Proposition 4.3 to get the following conservative and efficient interval:

$$I_{n,3} = \overline{X}_n \pm n^{-1/2} z_{1-\alpha/2+\beta_n} \left(\sigma(\hat{F}_n) + \frac{3c_n(1-\beta_n)}{n^{1/2}\sigma(\hat{F}_n)} \right)(1 + \hat{\delta}_{n,U}),$$

where $\hat{\delta}_{n,U}$ is obtained by replacing $\hat{\Delta}_n$ in (4.11) by

$$\hat{\Delta}_{n,U} = C_{\mathrm{BE}} n^{-1/2} \left[\sigma(\hat{F}_n) - \frac{3c_n(1-\beta_n)}{n^{1/2}\sigma(\hat{F}_n)} \right]^{-1};$$

tacitly, we are assuming the term in brackets is positive for such an approximation to be employed (which happens with probability tending to 1). The reasoning follows from inequalities (4.16) and (4.22).

5. Some finite sample results. In this section, we examine finite sample properties of various confidence intervals via the use of some simulation studies. While the interval $I_{n,3}$ that was derived in the proof of Theorem 4.2 can be computed exactly in practice, it is, unfortunately, much too wide for a reasonable sample and the same is expected for $I_{n,2}$. We therefore restrict ourselves to an approximation of the 'ideal' interval $I_{n,1}$ that is computed as follows. We generate bootstrap distributions $\hat{F}^*_{n,k}$ by resampling with replacement from the observed data until we obtain K of those that lie within the Kolmogorov–Smirnov band $\hat{R}_{n,1-\beta_n}$; bootstrap distributions generated in this process that fall outside the band are discarded. In addition, we consider the lower and upper limit of the Kolmogorov–Smirnov band denoted by $\hat{F}_{n,1-\beta_n,lo}$ and $\hat{F}_{n,1-\beta_n,up}$, respectively, and given as

$$\hat{F}_{n,1-\beta_n,lo}(x) = \max\{0, \hat{F}_n(x) - n^{-1/2}c_n(1-\beta_n)\}$$

and

$$\hat{F}_{n,1-\beta_n,up}(x) = \min\{1, \hat{F}_n(x) + n^{-1/2}c_n(1-\beta_n)\}.$$

The approximations of (4.1) and (4.2) are then defined as the respective maximum and minimum over these $K + 2$ distributions only. Note here that the quantiles $d_n(1-\alpha/2+\beta_n, F)$ and $d_n(\alpha/2-\beta_n, F)$ can in general not be calculated analytically but have to be estimated by another round of bootstrapping (based on B resamples), say. We denote the resulting interval by $\hat{I}_{n,1}$. It is clear that $\hat{I}_{n,1}$ is no longer strictly conservative but the corresponding loss is expected to be very small. An alternative avenue in approximating (4.1) and (4.2) would be to compute the respective maximum and minimum of all distributions that make up an ε-net of $\hat{R}_{n,1-\beta_n}$ and to explicitly correct for the approximation error while retaining a conservative interval. However, we will leave this approach for future work. The interval $\hat{I}_{n,1}$ should be close enough to $I_{n,1}$ to give an idea of what price one is paying by insisting on the correct coverage level for all distributions.

In our simulations, we compare $\hat{I}_{n,1}$ to $I_{n,0}$ (Anderson's conservative interval), $\mathrm{Boot}_{n,P}$, $\mathrm{Boot}_{n,H}$, $\mathrm{Boot}_{n,S}$ (percentile, hybrid, and Studentized bootstrap based on B resamples, respectively), and CLT (the gold standard). Performance is measured by empirical coverage and mean length of nominal 90% and 95% confidence intervals. Note that all intervals are truncated at 0 and 1 if necessary. The sample sizes considered are $n = 10$ and $n = 30$ and the underlying distributions considered are uniform, triangle, inverted-triangle, skewed-triangle (see Figure 1 for their densities) and two-point having mass 0.95 at 0 and mass 0.05 at 1. Furthermore, for $\alpha = 0.1$, we use $\beta_n = 0.01$ and for $\alpha = 0.05$, we use $\beta_n = 0.005$. Since $\hat{I}_{n,1}$ is computationally very expensive, we had to restrict ourselves to $K = 150$ bootstrap distributions $\hat{F}^*_{n,k}$. Finally,

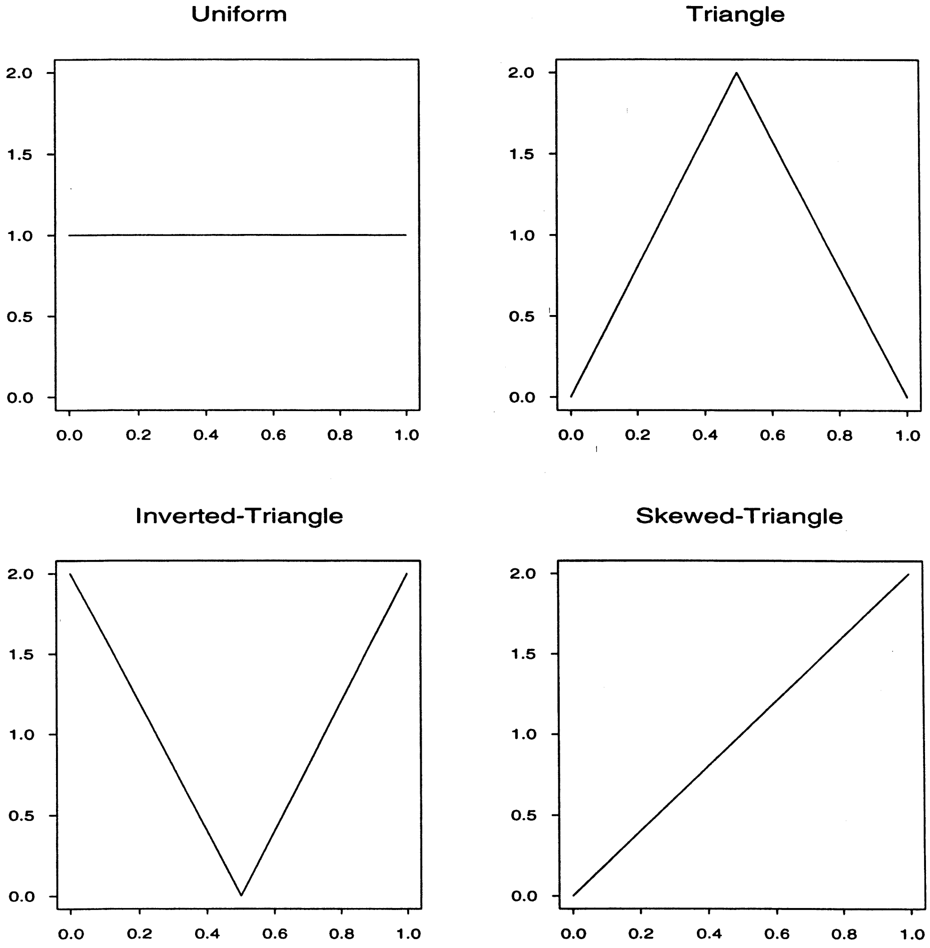

FIG. 1. *Densities of some of the distributions used in the simulations.*

TABLE 1
Estimated coverage probabilities of various confidence intervals based on 1000 *replications for each scenario*

			Uniform distribution				
n	**Target**	$I_{n,0}$	$\hat{I}_{n,1}$	$\mathbf{Boot}_{n,P}$	$\mathbf{Boot}_{n,H}$	$\mathbf{Boot}_{n,S}$	**CLT**
10	0.90	1.00	0.94	0.83	0.83	0.93	0.85
30	0.90	1.00	0.93	0.87	0.87	0.90	0.88
10	0.95	1.00	0.97	0.91	0.89	0.97	0.92
30	0.95	1.00	0.97	0.93	0.94	0.97	0.94
			Triangle distribution				
10	0.90	1.00	0.98	0.87	0.87	0.92	0.88
30	0.90	1.00	0.98	0.89	0.87	0.90	0.89
10	0.95	1.00	0.99	0.90	0.89	0.95	0.91
30	0.95	1.00	0.99	0.95	0.95	0.96	0.96
			Inverted-triangle distribution				
10	0.90	1.00	0.95	0.86	0.83	0.98	0.87
30	0.90	1.00	0.94	0.88	0.87	0.93	0.89
10	0.95	1.00	0.96	0.91	0.87	0.99	0.91
30	0.95	1.00	0.96	0.93	0.91	0.96	0.93
			Skewed-triangle distribution				
10	0.90	1.00	0.98	0.84	0.84	0.91	0.85
30	0.90	1.00	0.98	0.88	0.88	0.91	0.88
10	0.95	1.00	0.99	0.91	0.89	0.97	0.91
30	0.95	1.00	0.99	0.93	0.92	0.96	0.94
			Two-point distribution				
10	0.90	1.00	1.00	0.39	0.39	0.39	0.40
30	0.90	1.00	1.00	0.78	0.79	0.40	0.78
10	0.95	1.00	1.00	0.35	0.35	0.35	0.35
30	0.95	1.00	1.00	0.76	0.77	0.42	0.76

we employ $B = 1000$. The results are presented in Tables 1 and 2. Note that the column labelled "Target" in both tables refers to the nominal level. Thus, the column labelled "CLT" in Table 2 refers to the asymptotic lower bound in length as provided by the central limit theorem; that is, for a given n and level $1 - \alpha$, the column labelled "CLT" is $2\sigma(P)n^{-1/2}z_{1-\alpha/2}$.

As expected, our method is paying some price to achieve correct coverage level for all distributions. However, unlike Anderson's method, the price is not unreasonable (for the distributions considered). The estimated coverage probability of Anderson's interval is constantly equal to 1 for all scenarios considered. On the other hand, the interval $\hat{I}_{n,1}$ consistently overcovers, but estimated coverage is always below 1 except for the two-point distribution. Of the "standard" methods, all but the Studentized bootstrap generally undercover, even for $n = 30$. The Studentized bootstrap tends to work well except for the two-point distribution where it fails, as do the remaining "standard" intervals.

TABLE 2

Estimated mean lengths of various confidence intervals based on 1000 *replications for each scenario*

				Uniform distribution			
n	**Target**	$I_{n,0}$	$\hat{I}_{n,1}$	**Boot**$_{n,P}$	**Boot**$_{n,H}$	**Boot**$_{n,S}$	**CLT**
10	0.90	0.62	0.39	0.28	0.28	0.34	0.29
30	0.90	0.41	0.22	0.17	0.17	0.18	0.17
10	0.95	0.73	0.46	0.33	0.33	0.44	0.35
30	0.95	0.46	0.26	0.20	0.20	0.22	0.20
				Triangle distribution			
10	0.90	0.55	0.30	0.20	0.20	0.24	0.21
30	0.90	0.36	0.16	0.12	0.12	0.13	0.12
10	0.95	0.65	0.35	0.23	0.23	0.31	0.25
30	0.95	0.40	0.19	0.14	0.14	0.15	0.1446
				Inverted-triangle distribution			
10	0.90	0.67	0.47	0.34	0.34	0.43	0.36
30	0.90	0.43	0.26	0.21	0.21	0.22	0.21
10	0.95	0.78	0.53	0.41	0.41	0.58	0.43
30	0.95	0.49	0.30	0.25	0.25	0.26	0.25
				Skewed-triangle distribution			
10	0.90	0.57	0.36	0.23	0.23	0.29	0.24
30	0.90	0.37	0.20	0.14	0.14	0.15	0.14
10	0.95	0.68	0.40	0.27	0.27	0.37	0.29
30	0.95	0.42	0.23	0.16	0.16	0.18	0.17
				Two-point distribution			
10	0.90	0.45	0.35	0.13	0.10	0.06	0.12
30	0.90	0.28	0.20	0.11	0.09	0.08	0.11
10	0.95	0.54	0.35	0.12	0.09	0.05	0.11
30	0.95	0.32	0.22	0.12	0.09	0.08	0.11

Looking at estimated mean length, it turns out that our method is somewhat worse than the Studentized bootstrap but a big improvement over Anderson's method. In summary, in finite samples our method seems to provide a reasonable compromise between being too conservative (as with Anderson's method) and being too optimistic (as with the bootstrap, which typically undercovers and has no finite sample validity).

6. Conclusions and directions for future work. In this paper, we have demonstrated that it is indeed possible to construct confidence intervals for the mean that have a finite sample nonparametric validity and large sample efficiency, if it assumed the observations lie in a compact set. Our approach was constructive and made use of some crude inequalities (so there is, hopefully, room for improvement); nevertheless, we have proved existence of efficient intervals which can even be calculated by hand. Ultimately, we would not want to employ the most conservative of the intervals, $I_{n,3}$, and we seek better ways at approximating $I_{n,1}$.

There are obvious approaches based on simulation, such as the one employed in the previous section. It might be possible to account for the simulation error so that the guaranteed coverage property can be maintained. Given that the simulations look promising, future work will address this issue.

Our method can be generalized in a number of directions. First, the Kolmogorov–Smirnov distance played a role but only a minor one; however, it proves to be convenient at this time. Second, instead of basing the interval on the distribution of $\overline{X}_n - \mu(F)$, it may be better to look at a Studentized quantity with hopes of higher efficiency. Next, the basic interval $I_{n,1}$ defined in (4.3) applies to other parameters; simply, replace $\mu(F)$ by the parameter $\theta(F)$, $\overline{X}_n$ by $\hat{\theta}_n$ and let $J_n(F)$ be the distribution of $n^{1/2}(\hat{\theta}_n - \theta(F))$ under F. Extensions to two (or more) sample problems are immediate as well; individual confidence bands for the unknown laws are constructed and utilized appropriately. Finally, when the observations are no longer real-valued, we no longer have the convenience of distribution-free confidence bands for the unknown distribution. However, these bands really play a secondary role, and we can utilize exponential inequalities for the sup distance between the empirical measure and the true measure to get conservative bands which may be good enough.

As a final, more philosophical, point we advocate that the goal of constructing procedures that have finite sample validity requirements should be a primary consideration. There are many methods that enjoy excellent properties, such as (1.1) with $p = 2$. Rather than finding methods that satisfy (1.1) with $p = 3$, for example, we should only do so when we do not have to compromise on finite sample validity. Thus, now that we have intervals that are conservative and efficient, can we retain these two properties and still have (1.1) with p large?

APPENDIX

Proof of Lemma 4.1. For purposes of the proof, set $z = z_{1-\alpha/2+\beta_n}$, $\Delta = \Delta_n(F)$, $\delta = \delta_{n,\alpha}(F)$ and $r = 1 + \delta$. By Taylor's theorem,

$$\Phi(zr) - \Phi(z) = z(r-1)\phi(z) + \tfrac{1}{2}z^2(r-1)^2\phi'(z^*),$$

where z^* is between z and zr. Then $\phi'(z^*) \leq 0$ and $|\phi'(z)|$ is maximized when $z = 1$, which leads to

$$\Phi(zr) - \Phi(z) \geq z(r-1)\phi(z) - \tfrac{1}{2}z^2(r-1)^2(2\pi e)^{-1/2}.$$

So it suffices to choose r large enough so that the right-hand side of the last expression is greater than or equal to Δ; equivalently, it suffices to choose δ

large enough so that

$$-\tfrac{1}{2}z^2(2\pi e)^{-1/2}\delta^2 + z\phi(z)\delta - \Delta \geq 0.$$

Solving the quadratic leads to choosing δ so that

$$\delta \geq \frac{\phi(z) - [\phi^2(z) - c\Delta]^{1/2}}{zc/2},$$

where $c = 2(2\pi e)^{-1/2}$. [As in the statement of the lemma, we are assuming $\phi^2(z) \geq z\Delta$.] By expanding the function $f(\Delta) = (\phi^2(z) - c\Delta)^{1/2}$, so that

$$f(\Delta) = \phi(z) - \tfrac{1}{2}\Delta c\big[\phi^2(z) - c\Delta^*\big]^{-1/2}$$

for some Δ^* between 0 and Δ, it suffices to choose δ so that

$$\delta \geq \Delta\big[\phi^2(z) - c\Delta^*\big]^{-1/2}/z.$$

So, by monotonicity in Δ^*, it suffices to choose δ so

$$\delta \geq \Delta\big[\phi^2(z) - c\Delta\big]^{-1/2}/z.$$

Hence, the smallest δ that will work is no bigger than the right-hand side of this last expression.

REFERENCES

ANDERSON, T. (1967). Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function. *Bull. ISI* **43** 249–251.

BAHADUR, R. and SAVAGE, L. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.* **25** 1115–1122.

BERAN, R. and MILLAR, W. (1985). Asymptotic theory of confidence sets. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* **2** 865–887. Wadsworth, Belmont, CA.

BICKEL, P. (1992). Inference and auditing: the Stringer bound. *Internat. Statist. Rev.* **60** 197–209.

BRETH, M., MARITZ, J. and WILLIAMS, E. (1978). On distribution-free lower confidence limits for the mean of a nonnegative random variable. *Biometrika* **65** 529–534.

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

GASKO, M. (1991). A new distribution-free lower confidence bound for the mean of a positive random variable. Working Paper QS 92-006, San Jose State Univ. College of Business.

HAJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **1** 175–194. Univ. California Press, Berkeley.

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

HALL, P. and JING, B. (1995). Uniform coverage bounds for confidence intervals and Berry–Esseen theorems for Edgeworth expansion. *Ann. Statist.* **23** 363–375.

LOH, W. (1985). A new method for testing separate families of hypotheses. *J. Amer. Statist. Assoc.* **80** 362–368.

MILLAR, W. (1983). The minimax principle in asymptotic statistical theory. *Ecole d'Ete de St. Flour XI. Lecture Notes in Math.* **976** 76–265. Springer, New York.

ROMANO, J. (1989). Do bootstrap confidence procedures behave well uniformly in *P*? *Canad. J. Statist.* **17** 75–80.

SILVAPULLE, M. (1996). A test in the presence of nuisance parameters. *J. Amer. Statist. Assoc.* **91** 1690–1693.

STRINGER, K. (1963). Practical aspects of statistical sampling in auditing. In *Proceedings of Business and Economic Statistics Section*. Amer. Statist. Assoc., Alexandria, VA.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL
STANFORD, CALIFORNIA 94305-4065
E-MAIL: romano@stat.stanford.edu

DEPARTAMENTO DE ESTADISTICA
Y ELONUMETRICA
UNIVERSIDAD CARLOS III
DE MADRID
CALLE MADRID 126
28903 GETAFE
SPAIN
E-MAIL: mwolf@est-econ.uc3m.es