Calibrating Confidence Coefficients

WEI-YIN LOH*

Two approaches for dealing with the problem of poor coverage probabilities of certain standard confidence intervals are proposed. The first is a recommendation that the actual coverage be estimated directly from the data and its value reported in addition to the nominal level. This is achieved through a combination of computer simulation and density estimation. The asymptotic validity of the procedure is proved for a number of common situations. A classical example is the nonparametric estimation of the variance of a population using the normal-theory interval. Here it is shown that the estimated coverage probability consistently estimates the true coverage probability if the population distribution possesses a finite sixth moment.

The second approach is more traditional. It is a procedure for modifying an interval to yield improved coverage properties. Given a confidence interval, its estimated coverage probability obtained in the first approach is used to alter the nominal level of the interval. The interval with this modified nominal level is called a calibrated interval. In the case that the given interval is the normal-theory interval for the estimation of variance, the calibrated interval is proved to be asymptotically robust as long as sixth moments exist. As another application, the method is used to modify a bootstrap interval procedure for variance estimation. This leads to the derivation of a new bootstrap interval.

KEY WORDS: Bootstrap; Confidence level; Interval estimation; Kernel density estimation.

1. INTRODUCTION

Let $\theta = \theta(F)$ be a functional of a distribution F, and let I_n be a nominal $100\gamma\%$ confidence interval (CI) for θ based on a sample of size n. The word "nominal" indicates that the true coverage probability, γ_n say, of I_n may not be exactly γ. Usually, though, there is a class of F for which γ is a good approximation to γ_n, in the sense that $\gamma_n - \gamma \to 0$ as $n \to \infty$. For example, if θ is the mean of F and I_n the *normal-theory* t interval, it is well known that $\gamma_n \to \gamma$ provided only that F has a finite variance. Because the class of F with finite variance is rather large, the t interval is generally considered to be "robust." In contrast, the corresponding normal-theory interval for the variance σ^2 of F is nonrobust. In that case, $\gamma_n - \gamma \to 0$ iff the kurtosis of F is 3. For other F's the limiting difference can be quite large; for example, if F is the t_5 distribution, the nominal 90% CI for σ^2 has true confidence coefficient less than .60 in large samples—see Table 2 and Scheffé (1959, chap. 10).

The convergence of $\gamma_n - \gamma$ to zero is harder to ascertain for CI's constructed via more complicated procedures such as Efron's (1982) bootstrap method. In the latter, intervals are determined from bootstrap histograms of selected statistics and, except for certain classes of statistics (see, e.g., Abramovitch and Singh 1985; Beran 1982; Bickel and Freedman 1981; Singh 1981), the conditions under which $\gamma_n \to \gamma$ are not completely understood at the present time.

Even when $\gamma_n \to \gamma$ for each F in a class Ω, the convergence may not be uniform over Ω. Therefore, given a fixed n, I_n may be satisfactory for one F but not for another in Ω.

In view of these problems, I propose in this article a method of *estimating* γ_n directly from the data. The effectiveness of this proposal is demonstrated in some examples in Section 2. The method rests on the following argument. Since if F were known we would be able to say what γ_n is (by brute-force computer simulation if necessary), why not estimate F, by an estimator $\hat{F}_n$, say, from the data and then find out the probability $\hat{\gamma}_n$ that intervals from $\hat{F}_n$ will contain $\hat{\theta} = \theta(\hat{F}_n)$? (See Appendix B for a working definition of $\hat{\gamma}_n$. For example, when estimating the mean of F using the t interval, $\hat{\gamma}_n$ would be the proportion of those t intervals generated by samples from $\hat{F}_n$ that contain its mean $\hat{\theta}$.)

This idea is, strictly speaking, not new. It is just a new application of the bootstrap philosophy, in which the data is resampled for more information. What I hope to show is that it can lead to improvements over the use of γ as an estimator of γ_n. It is easy to see why $\hat{\gamma}_n$ should estimate γ_n consistently. Suppose that $C(\gamma^*)$ is a class of distributions containing F for which $\gamma_n \to \gamma^*$ as $n \to \infty$. If $\hat{F}_n$ eventually belongs to $C(\gamma^*)$ a.s., then we may also expect $\hat{\gamma}_n \to \gamma^*$ a.s.; that is, $\hat{\gamma}_n$ is a strongly consistent estimator of γ_n. The argument will be rigorous if it can be shown that $\gamma_n \to \gamma^*$ *uniformly* over $C(\gamma^*)$. Note that it is not necessary that $\gamma^* = \gamma$. This advantage will be clear in Example 4, where I apply the method to the normal-theory interval for σ^2 and show that $\hat{\gamma}_n - \gamma_n \to 0$ a.s. provided only that F has finite sixth moment.

Three more examples are given in Section 2. In Example 1, the estimation of the mean is considered. For the one-sided t interval, it is proved that, under moment conditions on F, $\hat{\gamma}_n - \gamma_n = O(n^{-1})$ a.s., whereas $\gamma_n - \gamma = O(n^{-1/2})$. For some bootstrap intervals, it can be proved only that both $\gamma_n - \gamma$ and $\hat{\gamma}_n - \gamma_n$ converge to zero as $n \to \infty$. The simulation results suggest, however, that the convergence rate for $\hat{\gamma}_n - \gamma_n$ may be faster.

Examples 2 and 3 illustrate two situations where, for all n, $\hat{\gamma}_n$ estimates γ_n *without error* for most reasonable $\hat{F}_n$. The interval in both cases is the bootstrap "percentile method" interval of Efron (1982). In Example 2, θ is the median. Here γ is known to be an excellent approximation for γ_n. In Example 3, θ is an endpoint of the support of F. Here $\gamma_n \equiv 0$ for all n, a totally unacceptable situation.

It is tempting to try to use this idea of "calibrating" γ with $\hat{\gamma}_n$ to construct new intervals whose true coverage

* Wei-Yin Loh is Assistant Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706. This work was supported in part by National Science Foundation Grants MCS 8300140 and DMS 8502303; the first grant also provided access to the CRAY supercomputer. The author is grateful to T. Hesterberg, E. L. Lehmann, C. F. J. Wu, the associate editor, and two referees for useful comments on various drafts of the manuscript, to B. Efron for pointing out an error in an early version of Table 1, to J. Gurland for helpful discussions on the mixture distribution in Section 2, and to Kathy Pedula for assistance with programming on the CRAY.

Journal of the American Statistical Association
March 1987, Vol. 82, No. 397, Theory and Methods

probabilities come closer to the desired nominal value. This is done in Section 3, leading to the introduction of calibrated intervals with possibly improved properties. In particular, a new method of constructing an interval from the bootstrap histogram is proposed and shown via simulation to be quite effective. Section 4 presents an application of this method to a bivariate data set.

Throughout this article, when I refer to bootstrap resampling I mean random sampling from the empirical cdf. Similarly, whenever I mention the percentile, bias-corrected percentile, and bootstrap t intervals, I mean the (unsmoothed) methods originally defined in Efron (1982).

2. FOUR EXAMPLES

2.1 Example 1: Estimating a Mean

Let θ be the mean of F and I_n the $100\gamma\%$ two-sided t interval, $\bar{X}_n \pm t_{n-1,1-\alpha} s_n n^{-1/2}$, where $\bar{X}_n$ and s_n^2 are the sample mean and variance, $t_{n,\alpha}$ is the 100α-percentile of the t distribution with n degrees of freedom, and $\gamma = 1 - 2\alpha$. The following theorem shows that $\hat{\gamma}_n$ is a better estimator of γ_n than γ for I_n, as well as its one-sided counterpart.

Theorem 1. Assume that F is continuous and has finite eighth moment. Let $\hat{F}_n$ be an estimator of F such that its first eight moments converge to those of F a.s. Then, for the one-sided t interval, $\gamma_n - \gamma = O(n^{-1/2})$ and $\hat{\gamma}_n - \gamma_n = O(n^{-1})$ a.s. If F has a finite tenth moment, then for the two-sided t interval, $\gamma_n - \gamma = O(n^{-1})$ and $\hat{\gamma}_n - \gamma_n = o(n^{-1})$ a.s.

Proof. The result for the one-sided t interval depends on the two-term Edgeworth expansion for the distribution of the t statistic, as follows:

$$\Pr(t \le x) = \Phi(x) + (6^{-1}\mu_3\sigma^{-3}n^{-1/2})(2x^2 + 1)\phi(x) + O(n^{-1}), \quad (2.1)$$

where σ^2 and μ_3 are the variance and third central moment of F, respectively, and $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cdf and density. [See, e.g., Hall (1983) or Abramovitch and Singh (1985). Chung (1946) demonstrated that the "$O(n^{-1})$" term is bounded by $Qn^{-1}h(x)$, where $h(x)$ is a function of x and Q is a constant depending only on the first eight moments of F.] Since $\mu_3(F) = 0$ when F is normal, it follows that $\gamma_n - \gamma = O(n^{-1/2})$. Applying the same expansion to $\hat{F}_n$ instead of F, we see that the distribution of t under $\hat{F}_n$ matches that under F up to and including the term in $n^{-1/2}$. Therefore, $\hat{\gamma}_n - \gamma_n = O(n^{-1})$ a.s.

For the two-sided t interval, the $n^{-1/2}$ term in the expansion for γ_n is missing, because the second term on the right side of (2.1) is an even function of x. Therefore, $\gamma_n - \gamma = O(n^{-1})$ in this case. By resorting to a three-term Edgeworth expansion, however, the same argument as above shows that now $\hat{\gamma}_n - \gamma_n = o(n^{-1})$ a.s. [A more careful analysis using the results in Chung (1946) indicates that the rate is $O(n^{-4/3})$.]

There are several other ways of constructing a CI for the mean. The most interesting of these attempt to provide asymmetric intervals (about $\bar{X}_n$), so as to reflect any skewness in F. Johnson (1978) proposed the *modified t* interval, $[\bar{X}_n + (6^{-1}\hat{\alpha}_{3,n}s_n^{-2}n^{-1})] \pm t_{n-1,1-\alpha}s_n n^{-1/2}$, where $\hat{\alpha}_{3,n}$ is the sample third central moment. A more recent and general technique is to construct CI's from a bootstrap histogram. In the case of the mean, this is a histogram of $\bar{X}_n^* = n^{-1}\sum_1^n X_i^*$, where $(X_1^*, X_2^*, \ldots, X_n^*)$ are iid observations from the empirical cdf that puts mass n^{-1} on each observation X_i $(i = 1, 2, \ldots, n)$. Efron (1982) gives a number of methods for setting CI's from this histogram. The *percentile method* prescribes as a nominal $100(1 - 2\alpha)\%$ CI the interval $[\theta_L, \theta_U]$, where θ_L and θ_U are the lower and upper α points of the histogram. The *bias-corrected percentile method* attempts to incorporate the skewness of F better by redistributing the probability unequally in the two tails of the histogram [see Efron (1982) for details]. Viewing $\bar{X}_n - \theta$ as an asymptotic pivot, it is also natural to consider the interval $[2\bar{X}_n - \theta_U, 2\bar{X}_n - \theta_L]$, which is the reflection of $[\theta_L, \theta_U]$ about $\bar{X}_n$. I will call this the *reflection method* in this article [see Loh (1984) and Efron (1979a, remark D) for arguments for and against this].

The next theorem gives sufficient conditions for $\gamma_n - \gamma$ and $\hat{\gamma}_n - \gamma_n$ to converge to zero a.s. for these bootstrap intervals. The proof is presented in Appendix A. Note that it is not necessary for the moments of $\hat{F}_n$ to converge to those of F. The convergence of $\gamma_n - \gamma$ to zero only is proved in Beran (1984) under more general conditions.

Theorem 2. Let F be any distribution with finite sixth moment, and let $\hat{F}_n$ be an estimator of F such that its first six moments converge a.s. Suppose that I_n is a bootstrap CI constructed from the percentile, bias-corrected percentile, or reflection methods. Then $\gamma_n - \gamma \to 0$ and $\hat{\gamma}_n - \gamma_n \to 0$ as $n \to \infty$ a.s.

Table 1 displays the results from a simulation experiment based on this example with $n = 10$ and $\gamma = .90$. Six interval procedures are compared: (a) two-sided t, (b) percentile method, (c) bias-corrected percentile method, (d) reflection method, (e) Johnson's t, and (f) *bootstrap t*. The bootstrap t was originally proposed in Efron (1982, sec. 10.10). It consists of applying the percentile method to the studentized form of the statistic. [The arguments in Hinkley and Wei (1984) and Abramovitch and Singh (1985) can be used to show that for the bootstrap t interval, typically $\gamma_n - \gamma = o(n^{-1/2})$ in the one-sided case and $\gamma_n - \gamma = o(n^{-1})$ in the two-sided case; the nominal level γ for this interval, therefore, matches the performance of $\hat{\gamma}_n$ as an estimator of γ_n for the t interval in Theorem 1.]

The distributions selected for the simulation are (a) normal, (b) uniform, (c) normal mixture, and (d) exponential. The particular normal mixture used is $\pi N(\mu_1, \sigma_1^2) + (1 - \pi)N(\mu_2, \sigma_2^2)$, with $\pi = .5504$, $\mu_1 = .3342$, $\mu_2 = -.4091$, $\sigma_1 = .2385$, and $\sigma_2 = 1.3603$. [As usual, $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2.] Lee and Gurland (1977) showed that this distribution is quite unfavorable for the one-sample t test when n is small. The estimate $\hat{F}_n$ used here is a data-based kernel density estimate. Appendix B describes the whole procedure in greater detail.

Table 1. Monte Carlo Estimates of γ_n, $E(\hat{\gamma}_n)$, and $sd(\hat{\gamma}_n)$ for Example 1 ($n = 10$, $\gamma = .90$)

Distribution	Method	γ_n	$E(\hat{\gamma}_n)$	$sd(\hat{\gamma}_n)$
Normal	*t* interval	.89	.897	.030
	Percentile	.83	.840	.036
	Bias-corrected	.83	.836	.037
	Reflection	.83	.836	.036
	Johnson *t*	.89	.898	.030
	Bootstrap *t*	.90	.897	.033
Uniform	*t* interval	.89	.898	.031
	Percentile	.86	.843	.039
	Bias-corrected	.85	.838	.039
	Reflection	.84	.834	.038
	Johnson *t*	.89	.900	.031
	Bootstrap *t*	.93	.905	.035
Mixture	*t* interval	.89	.874	.051
	Percentile	.76	.813	.055
	Bias-corrected	.73	.803	.059
	Reflection	.82	.816	.052
	Johnson *t*	.87	.873	.051
	Bootstrap *t*	.76	.860	.060
Exponential	*t* interval	.86	.876	.052
	Percentile	.79	.822	.053
	Bias-corrected	.79	.819	.054
	Reflection	.78	.812	.051
	Johnson *t*	.86	.877	.052
	Bootstrap *t*	.87	.887	.052
Maximum SE		±.02	±.002	±.004

NOTE: "±" quantities are maxima of estimated standard errors (SE's).

Table 1 indicates that, apart from the *t* and Johnson *t* intervals, the coverage probabilities of all of the bootstrap intervals can be quite poor. On the other hand, the accuracy of $\hat{\gamma}_n$ in predicting γ_n is quite good (except for the bootstrap *t* in the mixture distribution case). Note that the table does not show the Johnson *t* to be any better than the ordinary *t* interval.

I now give two situations where $\hat{\gamma}_n$ estimates γ_n without error.

2.2 Example 2: Estimating a Median

Let θ be the median of a continuous distribution F. Given the order statistics, $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, exact CI's for θ can be constructed by using the fact that, for any $1 \le k_1 < k_2 \le n$,

$$\Pr[X_{(k_1)} < \theta \le X_{(k_2)}] = \sum_{k=k_1}^{k_2} \binom{n}{k} (\tfrac{1}{2})^n \qquad (2.2)$$

for all F. Efron (1982) used this to demonstrate that the percentile method can be quite effective in setting intervals. From the bootstrap histogram of the sample median for odd n, this method yields an interval of the form $[X_{(k_1)}, X_{(k_2)}]$ with nominal (bootstrap) confidence level γ remarkably close to the γ_n given in (2.2). For example, if $n = 13$, $k_1 = 4$, and $k_2 = 10$, one gets $\gamma = .914$ and $\gamma_n = .908$. Now suppose that one did not know about (2.2) but constructed the bootstrap interval by using the percentile method. Because (2.2) is distribution free, my procedure would give $\hat{\gamma}_n \equiv \gamma_n$ regardless of which $\hat{F}_n$ is chosen, provided only that it is continuous.

2.3 Example 3: Estimating an Endpoint

Consider the estimation of the right endpoint θ of the support of a continuous distribution F, using the bootstrap percentile method. A natural quantity to bootstrap here is the largest order statistic $X_{(n)}$. Unfortunately, the bootstrap histogram of $X_{(n)}$ is of necessity to the left of θ. Because the percentile interval lies within the support of this histogram, it can never contain θ. Hence $\gamma_n \equiv 0$ for all γ. The fact that the latter holds for all continuous F, however, implies that we must also have $\hat{\gamma}_n \equiv 0$ if $\hat{F}_n$ is continuous. Thus $\hat{\gamma}_n \equiv \gamma_n$. Note that the same conclusions apply to the bias-corrected percentile method as well. The reflection method gives more sensible intervals, but they too may not be asymptotically consistent (see Loh 1984).

2.4 Example 4: Estimating a Variance

Let $(X_1, \ldots, X_n)$ be a random sample from F with variance σ^2. The $100(1 - 2\alpha)\%$ CI for σ^2 based on normal theory is

$$(n - 1)s_n^2/\chi^2_{n-1,1-\alpha} < \sigma^2 < (n - 1)s_n^2/\chi^2_{n-1,\alpha}, \qquad (2.3)$$

where s_n^2 is the unbiased estimate of variance and $\chi^2_{n,\alpha}$ is the 100α-percentile of the χ^2 distribution with n degrees of freedom. It is well known that this interval is sensitive to the kurtosis β of F. In fact (see Scheffé 1959, chap. 10),

$$\sqrt{(n - 1)/2}\{s_n^2\sigma^{-2} - 1\} \to N(0, B^2) \qquad \text{as } n \to \infty, \qquad (2.4)$$

where $B^2 = (\beta - 1)/2$. Hence the coverage γ_n of (2.3) tends to $1 - 2\Phi(B^{-1}z_\alpha)$, which equals $\gamma = 1 - 2\alpha$ only if $\beta = 3$. (Throughout this article, z_α refers to the 100α-percentile of the standard normal distribution.) The following theorem shows that, despite this, $\hat{\gamma}_n - \gamma_n \to 0$ a.s.

Theorem 3. Suppose that F has a finite sixth moment and $\hat{F}_n$ is an estimator of F such that its first six moments converge a.s., with the first four converging to those of F. Then, for the interval (2.3), $\hat{\gamma}_n - \gamma_n \to 0$ as $n \to \infty$ a.s.

Proof. Let $\mu = EX_1$ and $Y_i = \sigma^{-2}(X_i - \mu)^2$ ($i = 1, 2, \ldots, n$). Because the distribution of the left side of (2.4) is asymptotically equivalent to that of $(n/2)^{1/2}(\overline{Y}_n - 1)$, where $\overline{Y}_n$ denotes the mean of $\{Y_1, \ldots, Y_n\}$, it suffices to consider the limiting probability of the event $A(y) = \{n^{1/2}(\overline{Y}_n - 1) \le y\}$ under F and $\hat{F}_n$. The Berry–Esséen theorem implies that

$$\sup_y |P_F\{A(y)\} - \Phi\{y(\beta - 1)^{-1/2}\}| \le K\rho(F)(\beta - 1)^{-3/2}n^{-1/2}, \qquad (2.5)$$

where $\rho(F) = E|Y_1 - 1|^3 = E_F|\sigma^{-2}(X_1 - \mu)^2 - 1|^3$ and K is a universal constant. Applying the same result to $\hat{F}_n$ gives

$$\sup_y |P_{\hat{F}_n}\{A(y)\} - \Phi\{yb_n^{-1/2}\}| \le K\rho(\hat{F}_n)b_n^{-3/2}n^{-1/2}, \qquad (2.6)$$

where $b_n = \text{var}[\hat{\sigma}_n^{-2}(W - \hat{\mu}_n)^2]$, W has distribution $\hat{F}_n$, and $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are the mean and variance of W. The assumptions stated imply that $b_n \to (\beta - 1)$ and $\rho(\hat{F}_n) \to \rho_\infty$ (say) as

Table 2. Monte Carlo Estimates of γ_n, $E(\hat{\gamma}_n)$, and $sd(\hat{\gamma}_n)$ for (2.3) ($\gamma = .90$)

	$n = 25$			$n = 50$			
Distribution	γ_n	$E(\hat{\gamma}_n)$	$sd(\hat{\gamma}_n)$	γ_n	$E(\hat{\gamma}_n)$	$sd(\hat{\gamma}_n)$	$\lim_{n\to\infty} \gamma_n$
Normal	.90	.905	.053	.90	.905	.045	.900
Beta(.6825, 2)	.89	.891	.085	.91	.901	.061	.900
Uniform	.99	.953	.031	.99	.969	.024	.991
Exponential	.64	.765	.154	.64	.706	.149	.588
t_5	.71	.844	.108	.68	.811	.124	.588
Maximum SE	±.02	±.007	±.006	±.02	±.007	±.006	

NOTE: "±" quantities refer to the maximum estimated SE's. The SE's for the uniform distribution are less than half the maximum in each case.

$n \to \infty$ a.s. Here ρ_∞ may depend on the particular sequence $(X_1, X_2, \ldots)$. I conclude from (2.5) and (2.6) that $\hat{\gamma}_n - \gamma_n \to 0$ a.s.

Table 2 gives the results of a simulation experiment for five distributions, with $\gamma = .90$ and $n = 25$ and 50. The theoretical values of $\lim \gamma_n$ are also reported. The parameters of the beta distribution are chosen so that it has kurtosis equal to 3. The convergence of $\hat{\gamma}_n - \gamma_n$ to zero is seen to be quite good for the normal, beta, and uniform distributions, but slower for the exponential distribution. Note that the t_5 distribution is not covered by Theorem 3. Again, the details of the simulation can be found in Appendix B.

3. CALIBRATED INTERVALS

The preceding examples suggest that, given an interval procedure I_n for estimating θ, $\hat{\gamma}_n$ can be a more accurate estimate of γ_n than its nominal level. When this is the case, it is natural to ask whether one can use the information in $\hat{\gamma}_n$ to construct a better interval, I_n^* say, for θ, that is, one for which $\gamma_n(I_n^*)$ is closer to the desired level than $\gamma_n(I_n)$. This section proposes two methods and applies them to the problem of estimating the variance of F.

3.1 Calibrated Normal-Theory Interval

Suppose that in Example 4, we want a 90% confidence interval for the variance σ^2. Further suppose that, upon using the CI (2.3) with $\alpha = .05$ (so that $\gamma = .90$), we find $\hat{\gamma}_n = .70$. It is tempting now to *increase* γ (e.g., to .95) and recompute $\hat{\gamma}_n$ for the updated interval to see if $\hat{\gamma}_n$ is closer to .90. One might even imagine iterating this process (i.e., changing γ continuously) until $\hat{\gamma}_n$ is exactly .90. The final value of γ that results in this is then put back into (2.3) to obtain a modified interval. I will call this the *calibrated normal-theory* (CNT) *interval*.

The interesting question is what effect this process of calibration has on the coverage properties of the modified interval. The following theorem gives conditions on F for which γ_n (CNT) is consistent. It suffices to state and prove the result for the one-sided interval.

Theorem 4. Assume that F is continuous and has finite sixth moment. Let $\hat{F}_n$ be a continuous estimator of F such that its first six moments converge to those of F a.s. Let s_n^{*2} denote the sample variance based on a sample of size n from $\hat{F}_n$, and let $\hat{\sigma}^2$ denote the variance of $\hat{F}_n$. Finally, let $I_n = [\hat{k}s_n^2, \infty)$ be the calibrated interval for σ^2, where $\hat{k} = \hat{k}(X_1, X_2, \ldots, X_n)$ is chosen so that $\Pr_{\hat{F}_n}(\hat{k}s_n^{*2} < \hat{\sigma}^2) = \gamma$. (In this expression, $\hat{k}$ and $\hat{\sigma}^2$ are fixed given $\hat{F}_n$.) Then $\Pr_F(\sigma^2 \in I_n) \to \gamma$ as $n \to \infty$.

Proof. Let $\alpha_r(F) = E(X/\sigma)^r$ denote the standardized rth moment of F and $\hat{\alpha}_r = \alpha_r(\hat{F}_n)$. Hsu (1945) showed that

$$\sup_x \left|\Pr_F\{n^{1/2}(s_n^2\sigma^{-2} - 1)(\alpha_4 - 1)^{-1/2} \le x\} - \Phi(x)\right| \le An^{-1/2}\{\alpha_6(\alpha_4 - 1 - \alpha_3^2)^{-1}\}^{3/2},$$

for some universal constant A. It follows that

$$\begin{aligned}\gamma &= \Pr_{\hat{F}_n}(\hat{k}s_n^{*2} < \hat{\sigma}^2)\\ &= \Pr_{\hat{F}_n}\{n^{1/2}(s_n^{*2}\hat{\sigma}^{-2} - 1)(\hat{\alpha}_4 - 1)^{-1/2}\\ &\qquad < n^{1/2}\hat{k}^{-1}(1 - \hat{k})(\hat{\alpha}_4 - 1)^{-1/2}\}\\ &= \Phi\{n^{1/2}\hat{k}^{-1}(1 - \hat{k})(\alpha_4 - 1)^{-1/2}\} + O_P(n^{-1/2}) \text{ a.s.}\end{aligned}$$

This implies that $\hat{k} - 1 = 1 + z_\gamma n^{-1/2}(\alpha_4 - 1)^{1/2} + o_P(n^{-1/2})$ a.s. Hence

$$\begin{aligned}&\Pr_F(\sigma^2 \in I_n)\\ &\quad = \Pr_F(\hat{k}s_n^2 < \sigma^2)\\ &\quad = \Pr_F\{\sigma^{-2}s_n^2 < 1 + z_\gamma n^{-1/2}(\alpha_4 - 1)^{1/2} + o_P(n^{-1/2})\}\\ &\quad = \Pr_F\{n^{1/2}(\sigma^{-2}s_n^2 - 1)(\alpha_4 - 1)^{-1/2}\\ &\qquad < z_\gamma + o_P(1)\} \to \gamma \text{ a.s.}\end{aligned}$$

In general, it would be impractical to iterate the calibration process until $\hat{\gamma}_n$ converges to the desired level. I have found that often a *one-step* calibration *plus* linear interpolation is enough. To illustrate, suppose that we want a CNT interval with desired coefficient γ_0. First find $\hat{\gamma}_n$ for the interval (2.3) with $\gamma = \gamma_0$. Then set

$$\begin{aligned}\gamma_1 &= \gamma_0^2\hat{\gamma}_n^{-1} && \text{if } \hat{\gamma}_n \ge \gamma_0,\\ &= \gamma_0 + (1 - \gamma_0)(\gamma_0 - \hat{\gamma}_n)(1 - \hat{\gamma}_n)^{-1} && \text{if } \hat{\gamma}_n < \gamma_0.\end{aligned} \tag{3.1}$$

That is, the point (γ_1, γ_0) is gotten by linearly interpolating between $(\gamma_0, \hat{\gamma}_n)$ and either (0, 0) or (1, 1) depending on whether $\hat{\gamma}_n \ge$ or $< \gamma_0$. (For example, in the hypothetical case discussed in the beginning of this section, if $\gamma_0 = .90$

and $\hat{\gamma}_n = .70$, we will set $\gamma_1 = .9667$.) The CNT interval is then given by (2.3) with $\gamma = \gamma_1$.

It should be noted that although the CNT interval has been defined specifically for the estimation of variance, the basic definition is quite general. It includes, for example, the calibrated version of any interval of the form $\hat{\theta} \pm z_\alpha \hat{SE}(\hat{\theta})$, where $\hat{\theta}$ is any estimator of a parameter θ and $\hat{SE}(\hat{\theta})$ is any estimate of the standard error of $\hat{\theta}$ (such as a jackknife estimate).

3.2 A New Bootstrap Interval

Because our algorithm for constructing calibrated intervals is completely general, it can be applied to calibrate bootstrap intervals as well. For example, the undercoverage exhibited by the bootstrap intervals in Table 1 may, hopefully, be corrected via calibration. Since calibration is itself a form of bootstrapping, calibrated bootstrap intervals may also properly be called *iterated bootstrap* intervals.

Instead of examining the effect of calibrating any of the bootstrap intervals included in Table 1, I propose here a new bootstrap procedure designed to take full advantage of the calibration idea. Recall that, given a bootstrap histogram and a chosen value of γ, the reflection method prescribes the interval $I_n = [2\hat{\theta} - \theta_U, 2\hat{\theta} - \theta_L]$, where θ_L and θ_U are the lower and upper $(1 - \gamma)/2$-points of the histogram. The object here is to retain $100\gamma\%$ of the histogram mass. Unless the histogram is symmetric, there is no a priori reason for treating the tails symmetrically. Let $[\theta_L^*, \theta_U^*]$ be the *shortest* interval containing $100\gamma\%$ of the histogram. The corresponding reflection interval $I_n^* = [2\hat{\theta} - \theta_U^*, 2\hat{\theta} - \theta_L^*]$ would thus be shorter than I_n. If I_n undercovers θ, the interval I_n^* will only make the problem worse. If we do not stop here, however, but calibrate I_n^*, we may be able to overcome the undercoverage somewhat and *simultaneously* obtain a relatively short CI. I propose as a new bootstrap interval, therefore, the result of calibrating I_n^* and will refer to this as a *calibrated shortest reflection* (CSR) interval.

3.3 A Monte Carlo Study

To examine further the problem set out in Example 4, a Monte Carlo experiment was performed for $n = 20$. The nominal level chosen is $\gamma = .90$, and four distributions (all standardized so that $\sigma^2 = 1$) are used: (a) normal, (b) t_5, (c) uniform, and (d) exponential. The competing intervals are NT—normal-theory interval (2.3), CNT—calibrated NT interval, JK—jackknife interval based on s_n^2, JKL—jackknife interval based on $\log(s_n^2)$, PER—bootstrap percentile method, BCP—bias-corrected percentile method, BST—bootstrap t based on s_n^2, BSTL—bootstrap t based on $\log(s_n^2)$, PVT—Schenker's (1985) pivotal method, and CSR—calibrated shortest reflection interval.

The JK interval is $s_n^2 \pm t_{n-1,.95}\hat{SD}$, where $\hat{SD}$ is the jackknife estimate of standard error of s_n^2. The JKL interval is the jackknife interval for $\log(\sigma^2)$ based on $\log(s_n^2)$, subsequently exponentiated to recover the interval for σ^2. [Miller (1968) showed that jackknifing $\log(\sigma^2)$ is both powerful and robust for testing variances in the two-sample problem.] The PVT interval has the form $[s_n^4/\theta_U, s_n^4/\theta_L]$, where $[\theta_L, \theta_U]$ is the PER interval. (The PVT interval is obtained by treating $s_n^2\sigma^{-2}$ as a pivotal quantity and bootstrapping it.) The CNT interval uses (3.1). BST and BSTL are bootstrap t versions of JK and JKL, respectively.

The results are shown in Table 3. The values for $E(L)$ generally refer to estimates of the expected lengths of the intervals truncated at zero. The only exception is for the BSTL interval at the t_5 and exponential distributions. The BSTL interval seems to be extremely unstable in these two situations—estimates of $E(L)$ are many orders of magnitude larger than for the other methods, and the associated estimates of standard errors did not seem to decrease with increase in the number of Monte Carlo replications. I conjecture that $E(L)$ is infinite for the BSTL interval at these two distributions when $n = 20$. Therefore, instead of expected length, estimates of *median* length are reported (in parentheses) in the table. The JKL, PVT, and BST intervals also appear to be quite unstable for the sample size studied, though not as much as the BSTL. [The relative instability of jackknife intervals in other problems has also been observed in Efron (1982, p. 15) and Wu (in press).] On the other hand, the PER and BCP intervals tend to be too short and hence undercover σ^2. There is some indication that the CNT interval is trying to set right the miscoverage of the NT interval, although not as much as one would like. Except for the exponential

Table 3. Estimates of γ_n and E(L) for Estimating σ^2 ($\gamma = .90$; $n = 20$)

Distribution		NT	CNT	JK	JKL	PER	BCP	PVT	BST	BSTL	CSR
Normal	γ_n	.90	.89	.86	.90	.81	.80	.84	.88	.88	.83
	E(L)	1.25	1.22	1.07	1.26	.91	.93	1.27	1.46	1.71[a]	1.00
t_5	γ_n	.76	.78	.76	.85	.71	.72	.79	.85	.85	.87
	E(L)	1.25	1.41	1.38	2.4[b]	1.17	1.21	2.1[c]	3.2[b]	(1.4)	1.62
Uniform	γ_n	.99	.96	.90	.91	.86	.85	.85	.88	.90	.87
	E(L)	1.25	1.08	.77	.82	.68	.67	.88	.82	.77	.74
Exponential	γ_n	.64	.69	.72	.80	.68	.69	.72	.83	.84	.71
	E(L)	1.25	1.56	1.56	3.5[b]	1.33	1.37	3.2[c]	5.5[d]	(2.6)	1.43

NOTE: Median lengths are given in parentheses. Unless otherwise stated, maximum SE's for $\gamma_n \approx .01$ and maximum SE's for $E(L) \approx .02$.

[a] SE = .06.
[b] SE = .2.
[c] SE = .1.
[d] SE = .3.

distribution case, in which no method appears satisfactory in terms of both coverage probability and interval length, the JK and CSR intervals appear quite reasonable for the other distributions. The high computational cost of the CSR method is somewhat offset by its ability to provide asymmetric intervals (a property that the JK method lacks). Further details of the simulation are given in Appendix B.

4. A BIVARIATE EXAMPLE: THE LAW SCHOOL DATA

The preceding section demonstrated that the CSR method can produce intervals that are fairly short as well as have quite satisfactory coverage probabilities. One advantage of any bootstrap method is the potential for constructing asymmetric intervals (about the point estimate). We will examine this property of the CSR method by applying it to a real bivariate data set. The exercise will also illustrate how the method can be extended to multidimensional data.

The data, given in Efron (1979b, 1982), consist of the average LSAT and GPA scores for 15 American law schools. The problem is to construct a 68% CI for the correlation coefficient ρ. The sample correlation is $\hat{\rho} = .776$. To apply the CSR method, we use the variable kernel algorithm of Breiman, Meisel, and Purcell (1977) with a normal kernel to estimate first the true bivariate density. [See Devroye (1985) for some large sample properties of this density estimator.] Figure 1 shows a contour plot of the estimate superimposed on the 15 data points. The estimate is unimodal, has a little ridge running northeast-southwest, and has correlation coefficient .344. (The difference between this correlation and $\hat{\rho}$ is an indication of the amount of smoothing produced by the variable kernel estimate.)

Because only one set of data is involved, we can afford to be a little more elaborate in calibrating the shortest length interval. Instead of using just one calibration as in (3.1), two shortest length reflection intervals were calibrated, with nominal levels 68% and 90%, respectively.

Table 4. 68% Confidence Intervals for ρ

Method	Interval	Length
Normal-theory	$(\hat{\rho} - .16, \hat{\rho} + .09)$	.25
Percentile	$(\hat{\rho} - .12, \hat{\rho} + .13)$	.25
Bias-corrected percentile	$(\hat{\rho} - .17, \hat{\rho} + .10)$	.27
Bootstrap $t(\hat{\rho})$	$(\hat{\rho} - .19, \hat{\rho} + .15)$	.34
Bootstrap $t(\text{arctanh } \hat{\rho})$	$(\hat{\rho} - .42, \hat{\rho} + .09)$	.51
CSR	$(\hat{\rho} - .16, \hat{\rho} + .11)$	.27

The calibration was carried out with 1,000 replicate samples drawn from the density estimate. For each replicate sample, a bootstrap histogram for the sample correlation was constructed, using another 1,000 bootstrap samples. The values of $\hat{\gamma}_n$ thus obtained were, respectively, .615 and .772. Linear interpolation gave $\hat{\gamma} = .771$ as the adjusted nominal level.

The resulting CSR interval is shown in Table 4, together with the corresponding intervals based on normal theory and other bootstrap methods. The two bootstrap t intervals are based on the t statistics computed from $\hat{\rho}$ and Fisher's transformation arctanh($\hat{\rho}$), respectively, with the corresponding jackknife estimates of standard error used for studentization. Efron (1982, p. 83) noted that for this data, the bias-corrected percentile interval is more similar to the normal-theory interval than the uncorrected percentile interval, the latter being too symmetric. In this respect the CSR interval is in qualitative agreement with the former two. Its length is also not much different. In contrast, both of the bootstrap t intervals appear to be conservative. [Efron (1982, p. 88) observed that the bootstrap t seems to be specific to translation problems and its application to the correlation coefficient gives poor results.]

5. CONCLUDING REMARKS

The ideal confidence interval is one for which (a) its true coverage probability γ_n is close to the nominal level γ, and (b) this property holds uniformly for as many distributions as possible, at least for large enough n. These

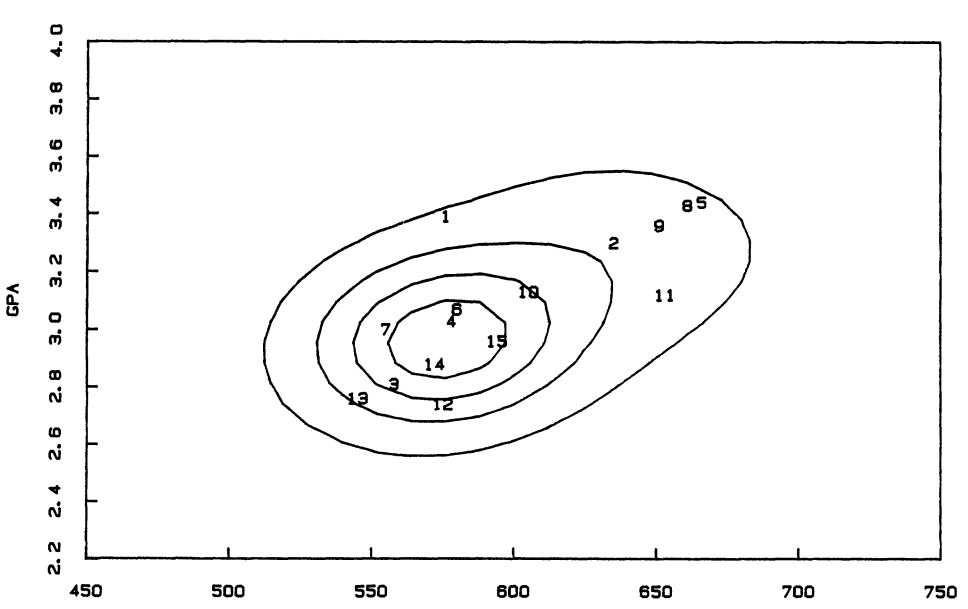

Figure 1. Contour Plot of Density Estimate.

twin goals may be called "accuracy" and "robustness of validity," respectively. Unfortunately, it is well known that, except for certain problems admitting nonparametric solutions (such as estimating the median), the two goals are often incompatible (Bahadur and Savage 1956).

To circumvent somewhat this difficulty, I propose in this article a new way of looking at the problem—namely, to estimate γ_n directly from the data and report it in addition to γ. The potential value of this approach is demonstrated in the examples, where we see that, besides improving accuracy, it can sometimes correct a γ that is totally wrong.

The proposed method is, of course, not foolproof. A counterexample is the estimation of an endpoint θ of F, where F is completely unspecified. Here any asymptotically valid interval must depend on some knowledge of the density of F near θ. Unless our estimator $\hat{F}_n$ is told this, the procedure cannot be expected to give good results all the time. (This remark does not contradict Example 3, since the percentile interval considered there is not asymptotically valid for any F.)

If we give up the requirement of uniform convergence of γ_n and ask only that $\gamma_n \to \gamma$ at each fixed F, then many methods are available for speeding up the convergence. Hall (1983), Hinkley and Wei (1984), and Abramovitch and Singh (1985), for example, gave methods based on inverting Edgeworth expansions. The calibrated intervals have the same aim. Unlike methods based on Edgeworth expansions, however, which require knowledge of the leading terms of the expansions and calculation of high-order moments (which may be unstable), the methods introduced in Section 3 are less demanding of mathematical expertise, since they are entirely based on simulation. Therefore, they might be easier to implement in practice, if a computer is available.

The calibrated intervals obviously require many more arithmetic operations to be performed than, say, the Hall intervals. In the case of the CSR interval, if B sets of pseudorandom samples are used to construct the bootstrap histogram and C sets of samples are used to calibrate each of these, then a total of BC sets of samples need to be generated and processed. In other words, if it takes one unit of computer time to calculate a Hall interval and B units to compute a percentile interval, then it would take BC units to obtain a CSR interval. The calibrated version of a standard (nonbootstrap) interval, for example, the CNT interval, on the other hand, requires only C units of computer time, because no bootstrap histogram is required. The appropriate values of B and C to use will depend on the problem, but with the greater availability of fast computers, the computational cost should be more affordable with time (see Efron 1979b).

The reader is referred to Loh (1985) for a discussion of similar issues in a hypothesis testing setting.

APPENDIX A: PROOF OF THEOREM 2

Only the proof for the percentile method is given, because similar proofs hold for the other two methods. The proof is broken into two lemmas.

Lemma 1. Let $(X_1, \ldots, X_n)$ be an iid sample from F with finite third moment. Let s_n^2 and r_n denote the sample variance and third absolute central moment, respectively. Then there is a continuous function $K(s_n, r_n)$ such that

$$|n^{1/2}(\theta_L - \overline{X}_n)s_n^{-1} - \Phi^{-1}(\alpha)| \le K(s_n, r_n)n^{-1/2} \text{ a.s.}$$

A similar result holds for θ_U.

Proof. Let Pr* denote probabilities under bootstrap resampling and $\overline{X}_n^*$ be a bootstrap mean. The Berry–Esséen theorem implies that

$$\begin{aligned}
&|\alpha - \Phi[n^{1/2}s_n^{-1}(\theta_L - \overline{X}_n)]| \\
&\quad = |\Pr^*\{n^{1/2}s_n^{-1}(\overline{X}_n^* - \overline{X}_n) \le n^{1/2}s_n^{-1}(\theta_L - \overline{X}_n)\} \\
&\qquad - \Phi[n^{1/2}s_n^{-1}(\theta_L - \overline{X}_n)]| \\
&\quad \le K_1 r_n s_n^{-3} n^{-1/2} \text{ a.s.},
\end{aligned}$$

where K_1 is a universal constant. Inverting this gives the result.

Lemma 2. Let I_n be a CI for the mean constructed from the percentile method, and suppose that F has a finite sixth moment. Then, for every $\varepsilon > 0$, there is a continuous function $C_F(\varepsilon)$, depending only on ε and the first six moments of F, such that $|\gamma_n - \gamma| \le \varepsilon + n^{-1/2}C_F(\varepsilon)$.

Proof. From Lemma 1, we have

$$\begin{aligned}
\Pr(\theta_L \ge \theta) &\le \Pr\{n^{1/2}s_n^{-1}(\overline{X}_n - \theta) \ge \Phi^{-1}(\alpha) - n^{-1/2}K(s_n, r_n)\} \\
&= \Pr\{n^{1/2}(\overline{X}_n - \theta)\sigma^{-1} \ge D(s_n, r_n)\},
\end{aligned}$$

where $D(s_n, r_n) = \sigma^{-1}s_n[\Phi^{-1}(\alpha) - n^{-1/2}K(s_n, r_n)]$ and σ^2 is the variance of F. Let $\varepsilon > 0$ and $K_1, \ldots, K_4$ denote constants depending only on ε and the first six moments of F. By Chebyshev's inequality, there exists K_1 such that the event $A = \{n^{1/2}|s_n^2 - \sigma^2| > K_1 \text{ or } |r_n - \rho| > K_1\}$ has probability less than ε. Here ρ denotes the third absolute central moment of F. By continuity, the minimum of $D(s_n, r_n)$ over the complement of A is bounded below by $M = \Phi^{-1}(\alpha)[1 + n^{-1/2}K_2(F)]$, for some K_2. Therefore,

$$\begin{aligned}
\Pr(\theta_L \ge \theta) &\le \varepsilon + \Pr\{n^{1/2}(\overline{X}_n - \theta)\sigma^{-1} \ge M\} \\
&\le \varepsilon + \alpha + n^{-1/2}K_3,
\end{aligned}$$

by the Berry–Esséen theorem. This and a corresponding result for θ_U imply that $|\gamma_n - \gamma| \le 2\varepsilon + n^{-1/2}K_4$.

Lemma 2 implies that $\gamma_n - \gamma \to 0$ as $n \to \infty$. The proof of Theorem 2 is completed by applying the same lemma to $\hat{F}_n$.

APPENDIX B: COMPUTATIONAL DETAILS

The experiments in Examples 1 and 4 were based on 500 replications each. For each replication, a kernel estimate of the underlying density of F was first obtained. The normal kernel was used throughout, with bandwidth chosen via the data-based algorithm suggested in Scott, Tapia, and Thompson (1977) [see also Scott and Factor 1981, formulas (2.10) and (2.11)]. Starting with the sample range as the initial guess, 20 iterations of this algorithm were executed to arrive at the eventual bandwidth—I did this instead of using the Newton–Raphson procedure proposed by the original authors to avoid the possibility of convergence to zero. After the bandwidth was selected, $\hat{\theta} = \theta(\hat{F}_n)$ was computed and 100 samples of size n from $\hat{F}_n$ were drawn.

The fraction of these samples for which the corresponding intervals contained $\hat{\theta}$ gave an estimate of $\hat{\gamma}_n$. The average and standard deviation of these values of $\hat{\gamma}_n$ over the 500 replications in the outermost layer of the Monte Carlo provided the estimates of $\mathrm{E}(\hat{\gamma}_n)$ and $\mathrm{sd}(\hat{\gamma}_n)$ in Tables 1 and 2. For the intervals derived via bootstrap methods in Table 1, the bootstrap histograms of the sample mean were constructed from 100 bootstrap samples. This formed the third (innermost) layer of the Monte Carlo.

The results for Table 3 were obtained by using the following variance reduction technique. Let γ_n^{NT} and $E(L^{NT})$ denote the true coverage probability and expected length of the NT interval. For each distribution and all other methods, Monte Carlo estimates of $\gamma_n - \gamma_n^{NT}$ and $E(L - L^{NT})$ were obtained, using the same 2,000 replicates. [For the CNT interval, e.g., let i (NT) be 1 or 0 according to whether NT contains σ^2 or not, for each replicate sample. Define i (CNT) similarly, and let $W = i$ (CNT) $- i$ (NT). Then $\gamma_n - \gamma_n^{NT}$ is estimated by averaging W over the 2,000 replicates.] Because the NT interval is much quicker to compute, γ_n^{NT} was estimated separately via another Monte Carlo run, using 50,000 replicates. The estimates of γ_n reported in Table 3 are the sums of the estimates of γ_n^{NT} and $\gamma_n - \gamma_n^{NT}$, with estimates of standard errors adjusted accordingly. Estimates of $E(L)$ are obtained similarly, although $E(L^{NT})$ is calculated exactly. [$E(L^{NT}) = 1.25$ for all of the distributions, since $\sigma^2 = 1$.] Quite substantial reductions in variances were achieved (as much as one-half of what would have been obtained had a direct Monte Carlo been used). In the case of the BSTL interval, median(L) is estimated by the median of the Monte Carlo realizations of L.

All of the bootstrap intervals in Table 3 were based on 100 bootstrap replicates, and another 100 replicates were used for calibrating the CNT and CSR intervals. Again the Scott et al. (1977) algorithm was employed to estimate densities.

The results in Table 3 were computed on a CRAY supercomputer. The rest of the computations for this article were done on a VAX 11/750, using random number generators from the International Mathematical and Statistical Library.

[*Received June 1984. Revised May 1986.*]

REFERENCES

Abramovitch, L., and Singh, K. (1985), "Edgeworth Corrected Pivotal Statistics and the Bootstrap," *The Annals of Statistics*, 13, 116–132.

Bahadur, R. R., and Savage, L. J. (1956), "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical Statistics*, 27, 1115–1122.

Beran, R. (1982), "Estimated Sampling Distributions: The Bootstrap and Competitors," *The Annals of Statistics*, 10, 212–225.

——— (1984), "Bootstrap Methods in Statistics," *Jber. d. Dt. Math-Verein*, 86, 14–30.

Bickel, P. J., and Freedman, D. A. (1981), "Some Asymptotic Theory for the Bootstrap," *The Annals of Statistics*, 9, 1196–1217.

Breiman, L., Meisel, W., and Purcell, E. (1977), "Variable Kernel Estimates of Multivariate Densities," *Technometrics*, 19, 135–144.

Chung, K. L. (1946), "The Approximate Distribution of Student's Statistic," *Annals of Mathematical Statistics*, 17, 447–465.

Devroye, L. (1985), "A Note on the L_1 Consistency of Variable Kernel Estimates," *The Annals of Statistics*, 13, 1041–1049.

Efron, B. (1979a), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

——— (1979b), "Computers and the Theory of Statistics: Thinking the Unthinkable," *SIAM Review*, 21, 460–480.

——— (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, Philadelphia: Society for Industrial and Applied Mathematics.

Hall, P. (1983), "Inverting an Edgeworth Expansion," *The Annals of Statistics*, 11, 569–576.

Hinkley, D., and Wei, B.-C. (1984), "Improvements of Jackknife Confidence Limit Methods," *Biometrika*, 71, 331–340.

Hsu, P. L. (1945), "The Approximate Distributions of the Mean and Variance of a Sample of Independent Variables," *Annals of Mathematical Statistics*, 16, 1–29.

Johnson, N. J. (1978), "Modified t Tests and Confidence Intervals for Asymmetrical Populations," *Journal of the American Statistical Association*, 73, 536–544.

Lee, A. F. S., and Gurland, J. (1977), "One-Sample t-Test When Sampling From a Mixture of Normal Distributions," *The Annals of Statistics*, 5, 803–807.

Loh, Wei-Yin (1984), "Estimating an Endpoint of a Distribution With Resampling Methods," *The Annals of Statistics*, 12, 1543–1550.

——— (1985), "A New Method for Testing Separate Families of Hypotheses," *Journal of the American Statistical Association*, 80, 362–368.

Miller, R. G. (1968), "Jackknifing Variances," *Annals of Mathematical Statistics*, 39, 567–582.

Scheffé, H. (1959), *The Analysis of Variance*, New York: John Wiley.

Schenker, N. (1985), "Qualms About Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, 80, 360–361.

Scott, D. W., and Factor, L. E. (1981), "Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators," *Journal of the American Statistical Association*, 76, 9–15.

Scott, D. W., Tapia, R. A., and Thompson, J. R. (1977), "Kernel Density Estimation Revisited," *Journal of Nonlinear Analysis*, 1, 339–372.

Singh, K. (1981), "On the Asymptotic Accuracy of Efron's Bootstrap," *The Annals of Statistics*, 9, 1187–1195.

Wu, C. F. J. (in press), "Jackknife, Bootstrap and Other Resampling Inference in Regression" (with discussion), *The Annals of Statistics*.