Statistica Sinica 1(1991), 477-491

BOOTSTRAP CALIBRATION FOR CONFIDENCE INTERVAL CONSTRUCTION AND SELECTION

Wei-Yin Loh

University of Wisconsin

Abstract: This article has three aims. First it is shown that, in simple situations, an exact procedure exists for implementing the calibration method of bootstrap interval construction. The procedure helps reveal the relationships between the calibration, bootstrap t and bootstrap root methods, and identifies settings in which all three methods yield the same result.

Secondly, a method of bootstrap interval construction is introduced which has the nice property of requiring only one level of bootstrap resampling, but which yields coverage error rates that are smaller than those obtained with the bootstrap t method. The method depends on the calibration of an Edgeworth-corrected confidence set, and its justification rests on Edgeworth expansions. Coverage error rates of order $O(n^{-3/2})$ and order $O(n^{-2})$ or smaller are obtained for one-sided and two-sided intervals respectively.

Because of the proliferation of numerous techniques for interval construction with different orders of asymptotic error rate, the practical problem of choosing among candidate intervals is becoming increasingly important. The third aim of this paper is to propose calibration as a method selection tool. It is shown that when the candidate intervals are derived from Edgeworth-corrected statistics, the calibration-selected interval possesses an asymptotic error rate equal to the best among them. Simulation results indicate that the finite-sample properties of the interval are also quite satisfactory.

Key words and phrases: Bootstrap root, confidence intervals, Edgeworth expansion, pivot.

1. Introduction

The subject of this article is construction of nonparametric confidence intervals with improved rates of convergence of coverage probabilities. Two basic approaches are well known. One employs Edgeworth expansions to accelerate the rate of convergence of a test statistic to asymptotic normality (Johnson (1978), Hall (1983), Withers (1983), Abramovitch and Singh (1985)). The other relies on bootstrapping. Specific bootstrap procedures include Efron's percentile, bias-corrected percentile, accelerated bias-corrected percentile, and bootstrap t methods (Efron (1982, 1987)), Beran's bootstrap root and prepivoting methods

(Beran (1987, 1988)), and the author's calibration method (Loh (1987, 1988)).

It is known (Abramovitch and Singh (1985), Hall (1986)) that when the Edgeworth expansion and bootstrap approaches are valid, the two can produce intervals with the same asymptotic error rates. Therefore in practice, bootstrap methods perform the Edgeworth corrections in a transparent but automatic fashion via simulation, while methods based on Edgeworth expansions implement the corrections analytically so that simulations are not required. Similarities exist among bootstrap methods themselves. Efron's bootstrap t method is a special case of Beran's bootstrap root method with the root chosen to be the Student t-statistic. And Beran's prepivoting method can be motivated by the idea of calibration (DiCiccio and Romano (1988)).

Theoretical comparisons of the asymptotic properties of Efron's methods (Hall (1983b)) have shown that for the estimation of smooth functions of vector means, the bootstrap t and accelerated bias-corrected percentile methods produce one-sided intervals with coverage error of order $O(n^{-1})$ compared to order $O(n^{-1/2})$ for the corresponding percentile and bias-corrected percentile intervals, as the sample size n increases. In the case of two-sided equal-tailed intervals, the coverage error is of order $O(n^{-1})$ for these four methods. On the other hand, Loh (1988) showed that a nonlinear approximation to the calibration method applied to the ordinary t-interval yields coverage errors of order $O(n^{-1})$ for one-sided intervals and order $O(n^{-2})$ for two-sided symmetric intervals. (An *equal-tailed* nominal $100(1-2\alpha)\%$ interval for θ takes the form $[\hat{\theta}-\hat{\xi}_1, \hat{\theta}+\hat{\xi}_2]$, where $\hat{\theta}$ is a point estimate of θ and the $\hat{\xi}_i$ are chosen such that $\Pr(\theta < \hat{\xi}_1)$ and $\Pr(\theta > \hat{\xi}_2)$ both converge to α as n increases. A *symmetric* interval at the same nominal level takes the form $[\hat{\theta}-\hat{\zeta}, \hat{\theta}+\hat{\zeta}]$, where $\hat{\zeta}$ is chosen such that $\Pr(|\hat{\theta}-\theta| > \hat{\zeta})$ converges to 2α. The reader is referred to Hall (1988a), for an analysis of the differences between equal-tailed and symmetric intervals.) It is shown in Section 2 that the same error rates may be obtained with Beran's bootstrap root and Efron's bootstrap t methods by appropriate choice of root; in fact, the calibration, bootstrap root, and bootstrap t intervals are *identical.* A practical method of performing the calibration exactly is described, which requires no more bootstrapping than the bootstrap root method.

Bootstrap intervals with even smaller coverage error rates are possible if bootstrapping is iterated. Their main problem is computational cost. When restricted to one level of bootstrapping, the best two-sided interval at this point (in terms of coverage error) seems to be the symmetric interval produced by either the bootstrap root or calibration methods. One criticism of the two-sided symmetric interval is that its end-points are always equidistant from $\hat{\theta}$ and therefore it does not reflect the skewness in the empirical cdf. The two-sided accelerated bias-corrected percentile interval has end-points that are not necessarily symmetric about $\hat{\theta}$, but it does not have as small a coverage error rate. A bootstrap procedurè that yields asymmetric two-sided intervals with error rates of order $O(n^{-2})$ is proposed in Section 3. It requires only one level of bootstrap sampling, and gives one-sided intervals with error rates of order $O(n^{-3/2})$.

Section 2 describes how calibration can be carried out without any approximation, in the context of estimating the mean of a distribution with the t-interval, and demonstrates its equivalence to the bootstrap root method. The new bootstrap recipe is derived and its asymptotic properties studied in Section 3 for one-term Edgeworth-corrected confidence bounds in the more general context of estimating a smooth function of vector means. Extensions to higher-order Edgeworth corrections are considered in Section 4.

Because of the proliferation of numerous techniques for interval construction with different orders of asymptotic error rate, the practical problem of choosing among candidate intervals is becoming increasingly important. A method of interval selection based on calibration is proposed in Section 5. It is shown for example, that when the candidate intervals are derived from Edgeworth-corrected statistics, the calibration-selected interval possesses an asymptotic error rate equal to the best among them. Simulation results indicate that the finite-sample properties of the interval are also quite satisfactory. Section 6 concludes the article with some remarks.

2. Calibration of the Normal-Theory Interval

Consider estimation of the mean, $\theta = \theta(F)$, of a distribution F using the nominal $100\alpha\%$ one-sided *normal-theory* interval

$$I^{(1)}_{\text{Norm}}(\alpha) = (-\infty,\ \hat{\theta} + n^{-1/2}\hat{\sigma}z_\alpha]$$

based on a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ of n observations from F, where $\hat{\sigma}^2$ is the usual unbiased estimate of variance and $z_\alpha = \Phi^{-1}(\alpha)$, with Φ the standard normal cdf. If F is a normal distribution and n is large, the coverage probability, $\pi(\alpha)$, of $I^{(1)}_{\text{Norm}}(\alpha)$ will be close to α. For other distributions, $\pi(\alpha)$ could be far from α. The idea of calibration is simply to replace α in $I^{(1)}_{\text{Norm}}(\alpha)$ with another value, α' say, such that the estimated coverage, $\hat{\pi}(\alpha')$, of the new interval $I^{(1)}_{\text{Norm}}(\alpha')$ satisfies $\hat{\pi}(\alpha') = \alpha$. In most applications the estimation will be carried out by bootstrap sampling via Monte Carlo simulation.

At first sight, this procedure appears impractical, as it suggests a potentially infinite search for α', with each candidate value requiring its own set of bootstrap samples. For this reason, solutions based on linear interpolation (Loh (1987)) and smooth nonlinear approximation (Loh (1988)) have been suggested. The latter was shown to give one-sided intervals with error rates as good as the bootstrap t and the accelerated bias-corrected percentile methods, and two-sided intervals with error rates an order of magnitude smaller than that of other equal-tailed intervals.

2.1. Exact calibration

In the present context α' can be obtained with just one set of B, say, bootstrap samples instead of infinitely many, as follows. For each $i = 1, \ldots, B$, let $\mathbf{X}_i^* = (X_{i1}^*, \ldots, X_{in}^*)$ denote the ith bootstrap sample from the empirical distribution. Let $\hat{\theta}_i^* = \hat{\theta}(\mathbf{X}_i^*)$, $\hat{\sigma}_i^* = \hat{\sigma}(\mathbf{X}_i^*)$, and denote the value of the t-statistic computed from the ith bootstrap sample by $t_i^* = n^{1/2}(\hat{\theta}_i^* - \hat{\theta})/\hat{\sigma}_i^*$. Finally let

$$\hat{\beta}_i = 1 - \Phi(t_i^*). \tag{1}$$

It is easy to see that the interval $I_{\text{Norm}}^{(1)*}(\alpha_0)$ based on the ith bootstrap sample will contain $\hat{\theta}$ if and only if $\alpha_0 \geq \hat{\beta}_i$ and that α' is just the α-quantile of the set $\{\hat{\beta}_1, \ldots, \hat{\beta}_B\}$.

In the case of the $100(1-2\alpha)\%$ two-sided interval

$$I_{\text{Norm}}^{(2)}(1-2\alpha) = [\hat{\theta} + n^{-1/2}\hat{\sigma}^{-1} z_\alpha,\ \hat{\theta} + n^{-1/2}\hat{\sigma}^{-1} z_{1-\alpha}],$$

define

$$\hat{\beta}_i = 1 - \Phi(|t_i^*|) \tag{2}$$

and take α' to be the 2α-quantile of $\{\hat{\beta}_1, \ldots, \hat{\beta}_B\}$.

2.2. Equivalence to the bootstrap root method

It is unnecessary to perform the inversions in (1) and (2), because $z_{\alpha'}$ instead of α' is the quantity needed in the computation of the desired interval. In the one-sided case, $z_{\alpha'}$ is the α-quantile of the set of t_i^*'s, and in the two-sided case, $z_{1-\alpha'}$ is the $(1-2\alpha)$-quantile of the set of $|t_i^*|$'s. Therefore the calibration method gives exactly the same intervals as Beran's bootstrap root method if the root is taken to be t and $|t|$ in the one-sided and two-sided case respectively. It now follows from known properties of the bootstrap t interval (Hall (1988b)) that the calibrated one- and two-sided intervals possess coverage errors of order $O(n^{-1})$ and order $O(n^{-2})$ respectively. These rates may be compared with rates of order $O(n^{-1/2})$ and order $O(n^{-1})$ for the one- and two-sided normal-theory interval.

3. Calibration of One-Term Edgeworth-Corrected Confidence Sets for Functions of Vector Means

Because calibration reduces the coverage error rate of an interval, we may apply it on an interval which has an initial error rate better than that of the classical normal-theory interval. A convenient candidate is that based on the one-term Edgeworth-corrected interval. In the case of estimating the mean, this is known as the Johnson t-interval (Johnson (1978))

$$I_{\mathrm{John}}(\alpha) = (-\infty, \hat{\theta}_J(\alpha)],$$

where

$$\hat{\theta}_J(\alpha) = \hat{\theta} + n^{-1/2}\hat{\sigma}\{z_\alpha + n^{-1/2}\hat{\lambda}(2z_\alpha^2 + 1)/6\}$$

and where $\hat{\lambda} = n^{-1}\hat{\sigma}^{-3}\sum(X_i - \overline{X})^3$ is the sample standardized skewness. The one- and two-sided equal-tailed Johnson intervals have coverage errors of the same order, namely $O(n^{-1})$. We show in this section that the calibrated one-term Edgeworth-corrected one- and two-sided equal-tailed intervals typically have coverage errors of order $O(n^{-3/2})$ and order $O(n^{-2})$ respectively. Another advantage of the interval $[\hat{\theta}_J(\alpha), \hat{\theta}_J(1-\alpha)]$ is that it is not necessarily symmetrical about $\hat{\theta}$.

3.1. Edgeworth expansions

To analyze the method more generally, suppose the data consist of a random sample of d-vectors $Y_1, \ldots, Y_n$. Let $\mu = E(Y_i)$ and $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$, and suppose that the parameter θ is related to μ through the relation $\theta = f(\mu)$ for some differentiable function f. Let $\hat{\theta} = f(\overline{Y})$ and suppose it has asymptotic variance $n^{-1}\sigma^2$, where $\sigma^2 = g(\mu)$ for another differentiable function g. Let $\hat{\sigma}^2 = g(\overline{Y})$ be the estimate of σ^2. This framework encompasses the estimation of means, variances, correlations, and functions of them.

Assume that the studentized statistic $n^{1/2}\hat{\sigma}^{-1}(\hat{\theta} - \theta)$ possesses the Edgeworth expansion

$$\Pr\{n^{1/2}\hat{\sigma}^{-1}(\hat{\theta} - \theta) \le x\} = \Phi(x) + \sum_{i=1}^{\nu} n^{-i/2} q_i(x)\phi(x) + O\big(n^{-(\nu+1)/2}\big)$$

for some $\nu \ge 2$, where $\Phi'(x) = \phi(x)$, and q_i is a polynomial of degree $3i - 1$ such that odd/even indexed polynomials are even/odd functions, respectively (Hall (1988b)). Suppose that a given confidence bound $\hat{\theta}(\alpha)$ admits an expansion of the form

$$\hat{\theta}(\alpha) = \hat{\theta} + n^{-1/2}\hat{\sigma}\Big\{z_\alpha + \sum_{i=1}^{3} n^{-i/2}\hat{s}_i(z_\alpha)\Big\} + O_p(n^{-5/2}),$$

where s_1 and s_3 are even polynomials and s_2 is an odd polynomial all of whose coefficients are functions of population moments, and where $\hat{s}_i$ is defined to be s_i with population moments replaced by sample moments. It is known (Hall (1988b), section 4.5) that under these assumptions, the interval $(-\infty, \hat{\theta}(\alpha)]$ has coverage probability $\pi(\alpha)$ given by

$$\pi(\alpha) = \alpha + n^{-1/2}\{s_1(z_\alpha) - q_1(z_\alpha)\}\phi(z_\alpha) - n^{-1}\left[\frac{1}{2}s_1^2(z_\alpha)z_\alpha + s_1(z_\alpha)\{q_1'(z_\alpha) - q_1(z_\alpha)z_\alpha\} - q_2(z_\alpha) - s_2(z_\alpha) + u_\alpha z_\alpha\right]\phi(z_\alpha) + n^{-3/2}r(z_\alpha)\phi(z_\alpha) + O(n^{-2}). \tag{3}$$

Here r is an even polynomial and u_α (independent of n) is defined by

$$E[n^{1/2}\hat{\sigma}^{-1}(\hat{\theta} - \theta) \times n^{1/2}\{\hat{s}_1(z_\alpha) - s_1(z_\alpha)\}] = u_\alpha + O(n^{-1}).$$

Therefore the two-sided interval $[\hat{\theta}(\alpha), \hat{\theta}(1-\alpha)]$ has coverage probability

$$\pi(1-\alpha) - \pi(\alpha) = 1 - 2\alpha - 2n^{-1}\left[\frac{1}{2}s_1^2(z_{1-\alpha})z_{1-\alpha} + s_1(z_{1-\alpha})\{q_1'(z_{1-\alpha}) - q_1(z_{1-\alpha})z_{1-\alpha}\} - q_2(z_{1-\alpha}) - s_2(z_{1-\alpha}) + u_{1-\alpha}z_{1-\alpha}\right]\phi(z_{1-\alpha}) + O(n^{-2}).$$

3.2. One-sided intervals

3.2.1. Theoretical results

Let $\hat{\theta}_1(\alpha)$ denote the one-term Edgeworth-corrected nominal $100\alpha\%$ upper confidence bound

$$\hat{\theta}_1(\alpha) = \hat{\theta} + n^{-1/2}\hat{\sigma}\{z_\alpha + n^{-1/2}\hat{q}_1(z_\alpha)\}.$$

According to (3), its coverage probability $\pi_1(\alpha)$ has the expansion

$$\pi_1(\alpha) = \alpha - n^{-1}\{q_{21}(z_\alpha) + u_\alpha z_\alpha\}\phi(z_\alpha) + O(n^{-3/2}), \tag{4}$$

where $q_{21}(x) = q_1(x)q_1'(x) - (1/2)xq_1^2(x) - q_2(x)$. Let α' be the calibrated value of α such that $\hat{\pi}_1(\alpha') = \alpha$. From (4)

$$\alpha = \alpha' - n^{-1}\{\hat{q}_{21}(z_{\alpha'}) + \hat{u}_{\alpha'}z_{\alpha'}\}\phi(z_{\alpha'}) + O_p(n^{-3/2})$$

and so we have through a Taylor expansion

$$\alpha' = \alpha + n^{-1}\{\hat{q}_{21}(z_\alpha) + \hat{u}_\alpha z_\alpha\}\phi(z_\alpha) + O_p(n^{-3/2}). \tag{5}$$

Therefore

$$\begin{aligned} z_{\alpha'} &= z_\alpha + (\alpha' - \alpha)(dz_\alpha/d\alpha) + \cdots \\ &= z_\alpha + n^{-1}\{\hat{q}_{21}(z_\alpha) + \hat{u}_\alpha z_\alpha\} + O_p(n^{-3/2}) \end{aligned}$$

and the 100α% calibrated confidence bound is given by

$$\hat{\theta}_1(\alpha') = \hat{\theta} + n^{-1/2}\hat{\sigma}[z_\alpha + n^{-1/2}\hat{q}_1(z_\alpha) + n^{-1}\{\hat{q}_{21}(z_\alpha) + \hat{u}_\alpha z_\alpha\}] + O_p(n^{-3/2}).$$

Equation (3) shows that the coverage probability $\pi_1(\alpha')$ of the interval $(-\infty, \hat{\theta}_1(\alpha')]$ converges to α at a rate of order $O(n^{-3/2})$.

3.2.2. Exact calibration

In practice, α' is obtained as follows. For the ith bootstrap sample, compute $\hat{\theta}_i^*$ and $\hat{\sigma}_i$, and let $x = x_i^{(1)}, x_i^{(2)}$ $(x_i^{(1)} \le x_i^{(2)})$ be the solutions of the quadratic equation

$$\hat{\theta}_i^* + n^{-1/2}\hat{\sigma}_i\{x + n^{-1/2}\hat{q}_1^{(i)}(x)\} = \hat{\theta},$$

where $\hat{q}_1^{(i)}(x)$ is $\hat{q}_1(x)$ computed from the ith bootstrap sample. Then, depending on the sign of the coefficient of the x^2-term in $\hat{q}_1^{(i)}(x)$, the solution set of the inequality

$$\hat{\theta}_i^* + n^{-1/2}\hat{\sigma}_i\{x + n^{-1/2}\hat{q}_1^{(i)}(x)\} \ge \hat{\theta} \tag{6}$$

is either

$$\{x \le x_i^{(1)}\} \cup \{x \ge x_i^{(2)}\} \tag{7}$$

or

$$\{x_i^{(1)} \le x \le x_i^{(2)}\}. \tag{8}$$

If (7) occurs, define $c_i = 1$, $y_i^{(1)} = -1$, $y_i^{(2)} = 1$. Otherwise, if (8) is realized, define $c_i = 0$, $y_i^{(1)} = 1$, $y_i^{(2)} = -1$. In either case, let $C = \sum_{i=1}^{B} c_i$ and define the function

$$h(x) = B^{-1}\left[C + \sum_{i=1}^{B}\{y_i^{(1)}I(x \ge x_i^{(1)}) + y_i^{(2)}I(x \ge x_i^{(2)})\}\right].$$

It is clear that for any x, $h(x)$ is the proportion of i's for which inequality (6) holds. Let $(a_1 \le \cdots \le a_{2B})$ denote the set of ordered $\{x_i^{(1)}, x_i^{(2)}; i = 1, \ldots, B\}$ and define the calibrated level as

$$\alpha' = \min\{h(a_j) : j = 1, \ldots, 2B\}.$$

3.2.3. Approximate calibration

An easier way to obtain the same asymptotic coverage error rate is to use a nonlinear approximation of α' as follows. Given α, obtain the bootstrap estimate of $\pi_1(\alpha)$:

$$\hat{\pi}_1(\alpha) = \alpha - n^{-1}\{\hat{q}_{21}(z_\alpha) + \hat{u}_\alpha z_\alpha\}\phi(z_\alpha) + O_p(n^{-3/2}).$$

Let ψ be a strictly increasing function on the unit interval with unbounded range and first derivative ψ', and let δ denote the *excess* of the bootstrap coverage over the nominal level on the ψ-scale. That is, let

$$\begin{aligned}\delta &= \psi\{\hat{\pi}_1(\alpha)\} - \psi(\alpha)\\ &= -n^{-1}\{\hat{q}_{21}(z_\alpha) + \hat{u}_\alpha z_\alpha\}\phi(z_\alpha)\psi'(\alpha) + O_p(n^{-3/2}).\end{aligned}$$

Further let α^* be the approximate value of α' given by

$$\begin{aligned}\alpha^* &= \psi^{-1}\{\psi(\alpha) - \delta\}\\ &= \psi^{-1}[\psi(\alpha) + n^{-1}\{\hat{q}_{21}(z_\alpha) + \hat{u}_\alpha z_\alpha\}\phi(z_\alpha)\psi'(\alpha)] + O_p(n^{-3/2})\\ &= \alpha + n^{-1}\{\hat{q}_{21}(z_\alpha) + \hat{u}_\alpha z_\alpha\}\phi(z_\alpha) + O_p(n^{-3/2}).\end{aligned}$$

Because of the similarity between this expansion and that in (5), it may be verified that the calibrated confidence bound $\hat{\theta}_1(\alpha^*)$ has coverage probability $\pi_1(\alpha^*) = \alpha + O(n^{-3/2})$.

3.3. Two-sided intervals

The coverage probability of the nominal $100(1-2\alpha)\%$ interval $[\hat{\theta}_1(\alpha), \hat{\theta}_1(1-\alpha)]$ possesses the expansion

$$\pi_1(1-\alpha) - \pi_1(\alpha) = 1 - 2\alpha - 2n^{-1}\{q_{21}(z_{1-\alpha}) + u_{1-\alpha}z_{1-\alpha}\}\phi(z_{1-\alpha}) + O(n^{-2}),$$

because q_{21} is an odd polynomial and $u_\alpha = u_{1-\alpha}$. It is more difficult to obtain an exact calibrated value of α in this case. Instead, we propose an approximation and show that it suffices.

The bootstrap coverage probability has the expansion

$$\hat{\pi}_1(1-\alpha) - \hat{\pi}_1(\alpha) = 1 - 2\alpha - 2n^{-1}\{\hat{q}_{21}(z_{1-\alpha}) + \hat{u}_{1-\alpha}z_{1-\alpha}\}\phi(z_{1-\alpha}) + O_p(n^{-2}).$$

Let

$$\begin{aligned}\delta &= \psi\{\hat{\pi}_1(1-\alpha) - \hat{\pi}_1(\alpha)\} - \psi(1-2\alpha)\\ &= -2n^{-1}\{\hat{q}_{21}(z_{1-\alpha}) + \hat{u}_{1-\alpha}z_{1-\alpha}\}\phi(z_{1-\alpha})\psi'(1-2\alpha) + O_p(n^{-2}),\end{aligned}$$

and define the approximate calibrated value

$$\begin{aligned}1-2\alpha^* &= \psi^{-1}\{\psi(1-2\alpha)-\delta\}\\ &= 1-2\alpha+2n^{-1}\{\hat{q}_{21}(z_{1-\alpha})+\hat{u}_{1-\alpha}z_{1-\alpha}\}\phi(z_{1-\alpha})+O_p(n^{-2}).\end{aligned}$$

Then

$$\begin{aligned}\alpha^* &= \alpha-n^{-1}\{\hat{q}_{21}(z_{1-\alpha})+\hat{u}_{1-\alpha}z_{1-\alpha}\}\phi(z_{1-\alpha})+O_p(n^{-2})\\ &= \alpha+n^{-1}\{\hat{q}_{21}(z_{\alpha})+\hat{u}_{\alpha}z_{\alpha}\}\phi(z_{\alpha})+O_p(n^{-2})\end{aligned}$$

and

$$z_{\alpha^*} = z_\alpha + n^{-1}\{\hat{q}_{21}(z_\alpha)+\hat{u}_\alpha z_\alpha\} + O_p(n^{-2}),$$

which implies that

$$\begin{aligned}\hat{\theta}_1(\alpha^*) = \hat{\theta} + n^{-1/2}\hat{\sigma}\big[z_\alpha + n^{-1/2}\hat{q}_1(z_\alpha) + n^{-1}\{\hat{q}_{21}(z_\alpha)+\hat{u}_\alpha z_\alpha\}\\ + n^{-3/2}\hat{q}_1'(z_\alpha)\{\hat{q}_{21}(z_\alpha)+\hat{u}_\alpha z_\alpha\}\big] + O_p(n^{-5/2}).\end{aligned}$$

It follows from (3) that the calibrated interval $[\hat{\theta}_1(\alpha^*), \hat{\theta}_1(1-\alpha^*)]$ has coverage probability

$$\pi_1(1-\alpha^*) - \pi_1(\alpha^*) = 1-2\alpha+O(n^{-2}).$$

3.4. A small simulation experiment

A simulation experiment was performed to examine the relevance of the asymptotic results for finite sample sizes. Three sample sizes ($n = 10, 25, 50$) were used with each of two distributions (normal and exponential) to compare the following intervals for the one-sample mean problem: (i) normal-theory, (ii) Johnson t, (iii) bootstrap t, (iv) approximate calibrated normal-theory, and (v) approximate calibrated Johnson t. The ψ-function used in the approximate calibration was the inverse normal cumulative distribution function. The value of α was 0.05 in all experiments, giving nominal 95% one-sided and 90% two-sided intervals. Five thousand Monte Carlo iterations were employed in each experiment, with each iteration utilizing five hundred bootstraps. This gave estimated standard errors of about 0.005 for the coverage probabilities. In the case of two-sided intervals, estimated average interval lengths were also computed.

The following observations are evident from the results given in Table 1.

1. The incorrect use of the z_α factor instead of t_α in the normal-theory interval is obvious in the case of the normal distribution with small n.

Table 1. Estimated coverage probabilities of intervals; simulation standard errors about 0.005; average interval lengths in parentheses; ψ-function was inverse normal cdf.

Interval type	Sample size	Method: Normal theory	Johnson t	Bootstrap t or exact calib. t	Approx. calib. normal-theory	Approx. calib. Johnson t
Normal distribution						
95%	$n = 10$	.937	.933	.951	.947	.940
left-closed	$n = 25$	.949	.949	.953	.953	.952
right-open	$n = 50$	.946	.947	.952	.948	.947
95%	$n = 10$	.938	.938	.956	.948	.945
left-open	$n = 25$	.943	.942	.948	.948	.947
right-closed	$n = 50$	.945	.945	.947	.945	.944
90%	$n = 10$	.875(1.02)	.871(1.02)	.914(1.18)	.904(1.13)	.891(1.10)
2-sided	$n = 25$	.891(0.65)	.891(0.65)	.906(0.68)	.906(0.68)	.899(0.67)
	$n = 50$	.891(0.46)	.892(0.46)	.896(0.47)	.896(0.47)	.895(0.47)
Exponential distribution						
95%	$n = 10$	.976	.954	.992	.968	.944
left-closed	$n = 25$	.973	.947	.989	.951	.944
right-open	$n = 50$	.974	.951	.989	.949	.948
95%	$n = 10$	.844	.872	.827	.889	.901
left-open	$n = 25$	.886	.919	.859	.927	.936
right-closed	$n = 50$	.911	.933	.888	.936	.944
90%	$n = 10$	.820(0.95)	.826(0.95)	.879(1.44)	.859(1.13)	.853(1.09)
2-sided	$n = 25$	.859(0.64)	.866(0.64)	.888(0.73)	.882(0.69)	.886(0.68)
	$n = 50$	.885(0.46)	.885(0.46)	.896(0.48)	.895(0.48)	.895(0.47)

2. The faster rate of convergence of the one-sided Johnson t intervals over that of the corresponding one-sided normal-theory intervals is apparent in the case of the exponential distribution. (The two-sided normal-theory and Johnson t intervals have coverage error rates of the same order.)
3. When the distribution is normal, the results for the bootstrap t, approximate calibrated normal-theory, and approximate calibrated Johnson t are similar; they are indistinguishable (within simulation error) from the results for the non-bootstrap intervals for $n = 25$ or 50, and are marginally superior to the latter only for $n = 10$.

   When the distribution is exponential, however, the approximate calibrated one-sided intervals appear to be quite a bit better than the one-sided bootstrap t. The superiority of the approximate calibrated Johnson t may be due to its faster rate of convergence. The superiority of the one-sided approximate calibrated normal-theory over the one-sided bootstrap t (which is the same as the exact calibrated normal-theory) may be because use of the ψ-function produced a smoother bootstrap.

4. The better coverage of the two-sided bootstrap t (compared to its one-sided relatives) is easily explained: its coverage probability has the expansion (Hall (1988b), p. 948)

$$\begin{aligned}&\pi_{\mathrm{STUD}}(1-\alpha)-\pi_{\mathrm{STUD}}(\alpha)\\&\quad=1-2\alpha-(1/3)n^{-1}(\kappa-3\lambda^2/2)z_{1-\alpha}(2z_{1-\alpha}^2+1)\phi(z_{1-\alpha})+O(n^{-3/2}),\end{aligned}$$

where λ and κ are the standardized skewness and kurtosis. The $O(n^{-1})$ term vanishes for the exponential distribution because $\kappa-3\lambda^2/2=0$. Therefore in this special case the two-sided bootstrap t interval incurs an error rate of order $O(n^{-3/2})$ instead of order $O(n^{-1})$.

5. Somewhat surprisingly, the bootstrap t method had the longest average interval length for small samples. Whether or not this is an artifact of the choice of distributions remains to be seen.

4. Higher-Order Edgeworth Corrections

The ability of the approximate calibration technique to improve the rate of convergence of coverage probabilities is not restricted to one-term Edgeworth corrections, but works more generally. We briefly explain why this is so in this section, for the one-sided case.

It is known that under weak conditions (Withers (1983), p. 585) there exist polynomials $s_i(x)$, $i=1,2,\ldots$, with coefficients that are functions of moments such that for each $\nu\geq 1$ the Edgeworth-corrected statistic

$$\hat\theta_\nu(\alpha)=\hat\theta+n^{-1/2}\hat\sigma\Big\{z_\alpha+\sum_{i=1}^{\nu}n^{-i/2}\hat s_i(z_\alpha)\Big\} \tag{9}$$

possesses the expansion

$$\pi_\nu(\alpha)=\Pr\{\hat\theta_\nu\leq\theta\}=\alpha-n^{-\nu/2}\phi(z_\alpha)s_{\nu+1}(z_\alpha)+O(n^{-(\nu+1)/2}).$$

This implies that

$$\hat\pi_\nu(\alpha)=\alpha-n^{-\nu/2}\phi(z_\alpha)\hat s_{\nu+1}(z_\alpha)+O_p(n^{-(\nu+1)/2}).$$

Let

$$\begin{aligned}\delta&=\psi\{\hat\pi_\nu(\alpha)\}-\psi(\alpha)\\&=-n^{-\nu/2}\phi(z_\alpha)\hat s_{\nu+1}(z_\alpha)\psi'(z_\alpha)+O_p(n^{-(\nu+1)/2})\end{aligned}$$

and define the approximate value

$$\begin{aligned}\alpha^* &= \psi^{-1}\{\psi(\alpha) - \delta\} \\ &= \psi^{-1}\{\psi(\alpha) + n^{-\nu/2}\phi(z_\alpha)\hat{s}_{\nu+1}(z_\alpha)\psi'(\alpha) + O_p(n^{-(\nu+1)/2})\} \\ &= \alpha + n^{-\nu/2}\phi(z_\alpha)\hat{s}_{\nu+1}(z_\alpha) + O_p(n^{-(\nu+1)/2}).\end{aligned}$$

Then

$$\begin{aligned}\hat{\theta}_\nu(\alpha^*) &= \hat{\theta} + n^{-1/2}\hat{\sigma}\Big\{z_{\alpha^*} + \sum_{i=1}^{\nu} n^{-i/2}\hat{s}_i(z_{\alpha^*})\Big\} \\ &= \hat{\theta} + n^{-1/2}\hat{\sigma}\Big\{z_\alpha + \sum_{i=1}^{\nu} n^{-i/2}\hat{s}_i(z_\alpha)\Big\} + O_p(n^{-(\nu+2)/2}).\end{aligned}$$

It follows that the coverage probability $\pi_\nu^*(\alpha)$ of the interval $(-\infty, \hat{\theta}_\nu(\alpha^*)]$ has the expansion

$$\pi_\nu^*(\alpha) = \alpha - n^{-(\nu+1)/2}\phi(z_\alpha)s_{\nu+2}(z_\alpha) + O(n^{-(\nu+2)/2}),$$

which implies that the error is one order of magnitude smaller than the uncalibrated interval.

5. Calibration for Interval Selection

Suppose that $\{I_{n,\nu_1}(\alpha), \ldots, I_{n,\nu_k}(\alpha)\}$ is a set of confidence intervals for a parameter θ such that for each $i = 1, \ldots, k$, the coverage probability $\pi_{\nu_i}(\alpha)$ of $I_{n,\nu_i}(\alpha)$ converges to α at rate of order $O(n^{-\nu_i/2})$ as $n \to \infty$. Further suppose that $\nu_1 < \cdots < \nu_k$. The interval with the best convergence rate is clearly $I_{n,\nu_k}(\alpha)$. Because the coverage error for a given sample size n typically depends on other quantities such as population moments, the interval with the smallest error is not necessarily $I_{n,\nu_k}(\alpha)$. Empirical evidence given in Hall (1983) for the estimation of a nonparametric mean shows for example that intervals based on high-order Edgeworth expansions may be unstable for small n, causing overcorrection and hence increase of coverage error with increase in the number of Edgeworth terms.

In view of the fact that $\pi_{\nu_i}(\alpha)$ may be estimated by the bootstrap estimate $\hat{\pi}_{\nu_i}(\alpha)$ for each i, it is natural to consider the following procedure for choosing among the intervals: Select the interval $I_{n,\hat{\nu}}(\alpha)$, where $\hat{\nu}$ is the value of ν_i such that $|\hat{\pi}_{\nu_i}(\alpha) - \alpha|$ is minimized for $i = 1, \ldots, k$. It is not difficult to show that this "calibration-selected" interval has a coverage error rate equal to that of the best interval $I_{n,\nu_k}(\alpha)$. We will illustrate this for the special case when

$$I_{n,\nu_i}(\alpha) = (-\infty, \hat{\theta}_{\nu_i}(\alpha)], \quad i = 1, 2, \ldots, k,$$

where $\hat{\theta}_\nu(\alpha)$ is given in (9). Then

$$\hat{\pi}_{\nu_i}(\alpha) = \alpha - n^{-\nu_i/2}\phi(z_\alpha)\hat{s}_{\nu_i+1}(z_\alpha) + O_p(n^{-(\nu_i+1)/2})$$
$$= \alpha - n^{-\nu_i/2}\phi(z_\alpha)s_{\nu_i+1}(z_\alpha) + o(n^{-\nu_i/2}) \quad \text{a.s.},$$

and hence the coverage error of $I_{n,\hat{\nu}}(\alpha)$ converges to zero at the same rate as that of $I_{n,\nu_k}(\alpha)$.

Table 2 contains the results from a simulation experiment on the finite-sample performance of the calibration-selected interval for estimating the population mean. The interval I_{n,ν_1} is the uncorrected (normal-theory) interval and I_{n,ν_2} and I_{n,ν_3} are the first two Edgeworth-corrected intervals, respectively. The two distributions employed are the normal and exponential. All the intervals are left-open, right-closed 95% intervals (see Hall (1983), Section 3, for the formulas). Each simulation employed five thousand replications, with $I_{n,\hat{\nu}}$ computed from five hundred bootstraps in each replicate. This gave simulation standard errors of the coverage probabilities around 0.005. The results show that the coverage probability of the calibration-selected interval is always within one simulation standard error of that of the best interval.

Table 2. Coverage probabilities of three 95% Edgeworth-corrected and one calibration-selected one-sided intervals based on 500 bootstraps; simulation standard errors about 0.005.

Interval	Normal			Exponential		
	$n=10$	$n=25$	$n=50$	$n=10$	$n=25$	$n=50$
I_{n,ν_1}	0.939	0.944	0.944	0.847	0.890	0.907
I_{n,ν_2}	0.940	0.942	0.944	0.871	0.915	0.933
I_{n,ν_3}	0.932	0.944	0.945	0.864	0.919	0.938
$I_{n,\hat{\nu}}$	0.939	0.945	0.944	0.868	0.917	0.934

6. Concluding Remarks

1. We showed that the bootstrap root and calibration methods give identical answers when the root is the t-statistic and the interval to be calibrated is the normal-theory (or, equivalently, the t) interval. This equivalence breaks down in more complicated situations such as calibration of Edgeworth-corrected intervals where the necessary root is neither obvious nor easy to compute. Calibration, on the other hand, is a fairly straightforward and automatic procedure. Further, the bootstrap root method appears to be restricted to those problems that permit solutions in the form of t-intervals. The calibration method is more general. For example, it works just as well

(perfectly) in the classical problem of nonparametric estimation of a population median using an interval based on order statistics (Loh (1987)).

2. There is another method of applying an Edgeworth correction to a t-statistic that does not depend on z_α. In the case that θ is the mean of F and t is the Student t-statistic, Abramovitch and Singh (1985) obtained the one-term corrected statistic
$$t_1 = t + n^{-1/2}\hat{\lambda}(2t^2+1)/6 \tag{10}$$
where $\hat{\lambda}$ is the sample standardized skewness, and showed that under weak regularity conditions, $\Pr(t_1 \le x) = \Phi(x) + O(n^{-1})$. Although the bootstrap root method may be applied with t_1 as root to obtain a confidence set with coverage error converging to zero at the same rate as that obtained by calibrating the interval based on $\hat{\theta}_1(\alpha)$, the result is not necessarily a contiguous interval. This is because (10) is a quadratic function of t, and there is no guarantee that a contiguous interval for θ results from inverting confidence sets of the form $(-\infty, t_1]$ or $\{x_1 \le t_1 \le x_2\}$. On the other hand, calibrating an interval always produces an interval. (Abramovitch and Singh (1985) also give statistics with higher-order corrections than (10) that induce confidence sets with smaller error rates, but the preceding remark still applies since the confidence sets typically consist of several disjoint pieces.)
3. Although we have demonstrated the improvements attainable through calibration of Edgeworth-corrected intervals, it should be emphasized that the application of our method is quite independent of the availability of Edgeworth expansions. If such an expansion is already known (such as when the parameter to be estimated is a function of vector means), then the calibration method should of course take advantage of it. In those problems where Edgeworth expansions are unknown or would require much effort to derive, it may be sufficient to apply the calibration method to the normal-theory interval (with any convenient estimate of standard error, including the jackknife if necessary). The rate of convergence of the resulting interval will still be faster than the uncalibrated interval.
4. The use of calibration as a method of interval selection is another application of the bootstrap idea that holds promise as a potentially powerful and automatic procedure when more than one interval is available and the user is afraid of choosing the worst one for a given sample size. The calibration-selected interval as described here can provide a solution for this problem. Besides being natural, it retains the best asymptotic error rate among the intervals.

Acknowledgment

This research was partially supported by NSF Grants DMS 85-02303 and DMS 88-03271.

References

Abramovitch, L. and Singh, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap. *Ann. Statist.* **13**, 116–132.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74**, 457–468.

Beran, R. (1988). Balanced simultaneous confidence sets. *J. Amer. Statist. Assoc.* **83**, 679–686.

DiCiccio, T. J. and Romano, J. P. (1988). A review of bootstrap confidence intervals (with discussion). *J. Roy. Statist. Soc. Ser.B* **50**, 338–354.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *J. Amer. Statist. Assoc.* **82**, 171–200.

Hall, P. (1983). Inverting an Edgeworth expansion. *Ann. Statist.* **11**, 569–576.

Hall, P. (1986). On the bootstrap and confidence intervals. *Ann. Statist.* **14**, 1431–1452.

Hall, P. (1988a). On symmetric bootstrap confidence intervals. *J. Roy. Statist. Soc. Ser.B* **50**, 35–45.

Hall, P. (1988b). Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.* **16**, 927–985.

Johnson, N. J. (1978). Modified t tests and confidence intervals for asymmetrical populations. *J. Amer. Statist. Assoc.* **73**, 536–544.

Loh, W.-Y. (1987). Calibrating confidence coefficients. *J. Amer. Statist. Assoc.* **82**, 155–162.

Loh, W.-Y. (1988). Discussion of "Theoretical comparison of bootstrap confidence intervals," by P. Hall. *Ann. Statist.* **16**, 972–976.

Withers, C. S. (1983). Expansions for the distribution and quantiles of a regular functional of the empirical distribution with applications to nonparametric confidence intervals. *Ann. Statist.* **11**, 577–587.

Department of Statistics, University of Wisconsin, 1210 W. Dayton Street, Madison, WI 53706, U.S.A.

(Received February 1990; accepted January 1991)