AMERICAN JOURNAL OF MATHEMATICAL AND MANAGEMENT SCIENCES

REVISITING QUALMS ABOUT BOOTSTRAP CONFIDENCE INTERVALS

MICHAEL R. CHERNICK
Lankenau Institute for Medical Research
100 Lancaster Avenue
Wynnewood, PA 19096
(chernickm@mlhs.org)

ROBERT A. LABUDDE
Least Cost Formulations, Ltd.
824 Timberlake Drive
Virginia Beach, VA 23464
(ral@lcfltd.com)

SYNOPTIC ABSTRACT

In a rather important paper Schenker (1985) used a particular chi-square distribution for a sample variance to show that the percentile method bootstrap and even the BC bootstrap break down for very practical sample sizes. This caused Efron (1987) to devise the BCa method in a very well-known paper. This paper revisits the issues surrounding bootstrap confidence intervals by looking at a particularly difficult problem, estimating variances in a nonparametric setting. We show by simulation methods that for certain heavy-tailed and skewed sampling distributions for the observed data, the convergence of even the second order accurate bootstrap methods BCa, and ABC, must be slow because even at a sample size of nSize=200 the confidence level is not close to the advertised and correct asymptotic level (e.g. for Student's t with 5 degrees of freedom the methods compared are between 4 and 5% below their nominal levels). At nSize =25, all of the methods provide true confidence levels that are at least 5% below their nominal confidence levels. We illustrate this by using confidence levels of 50%, 75% and 90% among others. To investigate more deeply the convergence properties, we took nSize=1000 or more (i.e. number of observations in the original data set). To adequately show the pattern of convergence, the standard deviation of the Monte Carlo approximation of the proportion needs to be around 0.005. But to achieve this requires close to nRepl =10,000 iterations (i.e. Monte Carlo replications of the bootstrap estimates) of the simulated results! For the high order bootstraps at nSize=1000 and nRepl =10,000 computations become prohibitively intensive even on a fast modern computer! It is interesting that for these simulations the first order percentile method bootstrap did nearly as well as, and sometimes better than, the higher order bootstraps.

Key Words and Phrases: Bootstrap confidence intervals, BC, BCa, ABC, second order bootstrap, iterated bootstrap, Efron's percentile method.

1. INTRODUCTION

Bootstrap confidence intervals were developed in the early 1980s after Efron's introduction of the bootstrap, Efron (1979). Efron's first suggested method was the percentile method which is simple and intuitive but found to converge too slowly in many cases. Adjustments using bias correction and an acceleration constant led to BC, and BCa respectively and the more easily computed ABC (a cheap man's BCa) described in DiCiccio and Romano (1988) (see also Efron and Tibshirani (1993) Chapter 22). The BCa intervals were introduced in Efron (1987) after the criticism of BC in Schenker (1985).
The rates of convergence were developed by Hall using Edgeworth and Cornish-Fisher expansions in a series of papers. This is covered thoroughly in Hall (1992). The faster converging methods (approaching the asymptotic confidence level more quickly) are called second order accurate. These include BCa, the iterated bootstrap and bootstrap t. Those that converge more slowly are first order accurate. In Efron and Tibshirani (1986), the authors illustrate nicely the hierarchy of the accuracy of the estimates and the assumptions required for them to work well. However, in the case of estimating the variance from an unspecified distribution the small sample coverage for all the methods can be severely below the nominal confidence level.

Our investigation through simulation confirms the asymptotic results but sometimes shows a different order of preference in the small sample size setting ($10 \leq n \leq 50$). Simulation studies addressing the small sample properties of specific bootstrap estimates occur sporadically in the literature. The two books Shao and Tu (1995) and Manly (2006) are two examples where such simulations are included. R code for lognormal distribution is presented in the Appendix.

1.1 Background

Let $X_1^*, X_2^*, \ldots, X_n^*$ be a bootstrap sample, that is a sample of size n taken with replacement from the empirical distribution F_n . Let θ be a parameter of the distribution that can be expressed as a functional of the distribution F with estimate based on the data $X_1, X_2, \ldots, X_n$. and denoted as $\theta(X_1, X_2, \ldots, X_n)$ for the functional applied to F_n. The bootstrap, estimates the variance of $\theta(X_1, X_2, \ldots, X_n)$ by computing or approximating the variance of $\theta^*=\theta(X_1^*, X_2^*, \ldots, X_n^*)$ a bootstrap sample estimate of θ. Sometimes, the bootstrap estimate can be obtained analytically, as with the case of the standard deviation for a median or mean (see Efron and Tibshirani [1993] p.14). However, in most cases the bootstrap estimate is approximated by Monte Carlo.

If nReal bootstrap resamples θ^* are drawn, the nReal θ^*'s provide a resampling distribution which estimates the sample distribution of θs to obtain standard deviations, bias or any other property of the distribution for the estimate of the

parameter θ. In our example θ is the variance and we will be interested in confidence intervals for θ by bootstrap. Note that there is variability among bootstrap samples because we sample with replacement. So, in a particular bootstrap sample, some observations will appear two or more times and others not at all. For estimating standard deviations, nReal is recommended to be 200 (Efron and Tibshirani [1993] p. 52), although some suggest that even larger values may be required. For confidence interval estimation or hypothesis testing at least 1000 bootstrap replications are recommended. Efron and Tibshirani (1993) indicate on page 52 that nReal must be considerably higher for confidence intervals. Although they do not explicitly suggest 1000 iterations it can be inferred that they believe 1000 are needed as their examples for the percentile method on page 172. Also since the percentile method involves estimating percentiles of a distribution their interpretation of Figure 19.3, page 276 is that it suggests that 500 to 1000 replications are needed when estimating the upper percentiles.

The simplest way to generate approximate nonparametric confidence intervals by the bootstrap is by taking the appropriate percentiles of the bootstrap samples, i.e. from the nReal realizations of bootstrap samples. For example a two-sided approximate 95% confidence interval for a parameter θ would be obtained as the interval from the 2.5 to the 97.5 percentile of the distribution of bootstrap samples. This method is called Efron's percentile method.

In this paper we will use a second level of Monte Carlo to approximate the true confidence level of the bootstrap estimates. To distinguish the two levels of iteration from each other and the original sample size of the data we shall refer to the original sample size as nSize, the number of Monte Carlo replicate samples as nRepl and the number of Monte Carlo realizations for the bootstrap estimates as nReal.

1.2 Properties of Confidence Intervals

As discussed in Efron (1987) and Efron and Tibshirani (1993) confidence intervals are **accurate** or **nearly exact** when the stated confidence level for the intervals is approximately the long run probability that the random interval contains the "true" value of the parameter. Accurate confidence intervals are said to be **correct** if they approach asymptotically the endpoint of the ideal or shortest length confidence intervals possible for the given confidence level. For parametric families of distributions such optimal intervals sometimes exist. The rate of this convergence is called the order of correctness. In some situations exact confidence intervals can be constructed. But for the nonparametric bootstrap the population distribution is not assumed to have a parametric form and hence except in special cases like the median only intervals that approximate their nominal confidence level can be constructed (see Bahadur and Savage [1965] for a theorem related to the lack of exactness for nonparametric tests and hence also confidence intervals) . For some methods the rate of convergence is faster than for others and the small sample

properties for various population distributions have not commonly been studied. The BCa method and the iterated bootstrap (double bootstrap or bootstrap calibration) are examples of methods for constructing bootstrap confidence intervals that are closer to being exact (accurate) and correct than the percentile method in many circumstances. See Chernick (2007) pp. 57-65 for details on these methods.

Table 1 (see Appendix) shows the assumptions required for various bootstrap confidence intervals to be appropriate (asymptotically accurate and correct). The subscript h is used for "hat" to indicate that a sample estimate is used.

Results given in Table 1 depend on the following assumptions (1) the asymptotic results apply, (2) distributions for the parameter estimate (in this case the sample variance) are not heavy-tailed or highly skewed. Under the assumptions for correctness in column 4 of Table 1 (1) and (2) are satisfied. For small to moderate sample sizes these properties may not apply as (1) and (2) may not be satisfied. In this paper we show for heavy-tailed distributions with finite third moments that BCa may not be the most accurate bootstrap estimate unless the sample size is very large. This also has implications on the comparison of the difference between two variances by use of bootstrap confidence intervals. Advice on which method to use is also given in Carpenter and Bithell (2000). But for cases like those studied here this advice may not apply.

2. RESULTS

2.1 Definitions of Parameters and Methods

The key parameters of the simulations are:

nSize: The sample size of the originating data which is to be bootstrapped.

nReal: The number of bootstrap resamples used to estimate the bootstrap resampling distribution(The number of possible unique resamples is always no more than $nSize^{nSize}$).

nRepl: The number of Monte Carlo replications of the entire experiment, based on generating new samples of size nSize from the underlying assumed distribution, in order to estimate coverage accuracy and other errors of the bootstrap methodology.

We report on the following bootstrap methods:

Normal-t: A parametric Student-t confidence interval, with center point the sample variance and the standard error of variance estimated from that of the resampling distribution. This differs from a parametric normal bootstrap in that the percentile of the t distribution with n-1 degrees of freedom is used instead of the standard normal percentile. In Hesterberg's bootstrap chapter (Hesterberg et al. [2003]) it is referred to as the bootstrap-t but that confuses it with an earlier bootstrap with that name that we are not studying in this paper.

EP: Efron percentile interval, with endpoints the plug-in quantiles of the resampling distribution.
BC: The Efron bias-corrected interval. Simulations have shown that the BC method is, as expected, virtually identical in estimates to the BCa interval with the acceleration a = 0 (i.e., adjustment for median bias only).
BCa: The Efron bias-corrected-and-accelerated interval, with median bias correction and skew correction via a jackknife estimate of the (biased) coefficient of skewness from the original sample.
ABC: The Efron-DiCiccio approximate bootstrap confidence interval.

Simulations reported here were performed over an extensive time period. So they were different somewhat in experimental design. The first studies were simple and smaller scale, and the latest were very extensive and large scale. Along the way more extensive computational procedures were developed. The appendix supplies R code which will generate coverage accuracies for a given distribution (exemplified by the ln[Normal(0,1)] case), but this code runs extremely slowly for large-scale problems. Consequently compiled code in PowerBasic was developed for all methods except the ABC. This code allowed very large scale simulations to be carried out within the 24 hr time budgeted for each case. The studies are presented next in the chronological order they were carried out.

2.2 Gamma (2, 3) Distribution

In this example the sample size nSize is 50. 1000(nReal) bootstrap samples of size 50 (nSize) were generated by Monte Carlo. We considered both 1000 and 10,000 Monte Carlo replications (nRepl) for each case. This provides an example of a highly skewed population distribution. In this case we compare Efron's Percentile (EP). ABC and BCa.

The mean and variance of a Gamma (2, 3) Distribution are 6 and 18 respectively. A simple random sample was generated and gave a sample mean of 5.05 and variance of 11.5. This indicates the high variability present in small samples and illustrates why bootstrap estimates could have problems in small samples. Simulation results for EP, ABC and BCa are presented in Table 2.

Table 2: Gamma (2, 3) Distribution with Sample Size nSize=50

No. of Monte Carlo Replications (nRepl)	Nominal Confidence Level	EP (Efron's Percentile Method)	ABC	BCa
1000	95%	86%	88%	88.7%
10,000	95%	84.4%	88%	88.6%

2.3 Student's t Distribution with 5 Degrees of Freedom

The sample size nSize is chosen to be 50, 200, 400, 800, 1600 and 3200. 1000 (nReal) bootstrap samples of size 50, 200, 400, 800, 1600 and 3200 were generated by Monte Carlo to look at convergence. We considered both 1000 and 10,000 bootstrap realizations (nReal) for each case with nSize=50 (1000 only for values of nSize greater than 400). This provides an example of a heavy-tailed symmetric distribution with finite second moments for the original sample. We compare Efron's Percentile Method, ABC and BCa. The larger sample size is chosen to see how things change as the sample size gets large. We expect the coverage probability to converge to the nominal confidence level as nSize gets large.

Table 3: Student's t Distribution with 5 Degrees of Freedom: Results for 90% Confidence Intervals

Sample size (nSize)	No. of Bootstrap realizations (nReal)	EP	ABC	BCa
50	1000 (10,000)	76.7% (75,3%)	80.4% (78.2%)	82.4%(79.3%)
200	1000	86.1%	85%	85.2%
400	1000 (10,000)	86.0% (85.2%)	85.1% (85.1%)	85.7% (85.7%)
800	1000	86.3%	86.1%	Not done *
1,600	1000	87.5%	87.6%	Not done *
3,200	1000	88.2%	87.2%	Not done*

*** BCa is a very computer –intensive method and the time to complete the calculations were considered prohibitive for nSize of 800 and above.**

From the above table we conclude (1) there is a large bias for all methods in small sample sizes and (2) with a sample size of 3,200 we see the confidence level converging to the nominal 90% from below.

2.4 Uniform (0, 1) Distribution

For this distribution we simulate two cases:

(1) The sample size nSize is 25, and 1000 bootstrap samples (nReal) of nSize 25 were generated by Monte Carlo. We considered 1000 Monte Carlo replications and also 64,000 replications (nRepl) for each case. This provides an example of a short-tailed distribution. In this case we compare the Normal-t with Efron's Percentile Method.

(2) The sample size is varied from 10 to 100 and we compare Normal-t, Efron's Percentile Method, ABC and BCa with sample size varied. For both (1) and (2), we used confidence levels of 50%, 60%, 70%, 80%, 90%, 95% and 99%.

Table 4: Uniform (0, 1) Distribution: Results for various confidence intervals: (1) Results: nSize=25, nReal=1000.

Confidence Level	Normal-t	EP
50%	49.5%	49%
60%	60.4%	57.3%
70%	67.2%	68%
80%	78.4%	77.9%
90%	87.6%	86.8%
95%	92.7%	92%
99%	97.4%	96.5%

Table 5: Uniform (0, 1) Distribution: Results for 90% confidence intervals: (2) Results: Sample Size (nSize), Bootstrap samples (nReal) and the number of Monte Carlo samples generated (nRepl) are varied and the Asymptotic Confidence Level is 90%

Sample Size (nSize)	nRepl	nReal	Normal-t	EP	ABC *	BCa *
10	64,000	16,000	86.42%	84.31%	81.65%	83.35%
20	64,000	16,000	88.89%	88.11%	88.35%	88.28%
25	64,000	16,000	89.21%	88.66%	88.15%	87.95%
30	64,000	16,000	89.41%	88.98%	88.98%	88.53%
40	64,000	16,000	89.69%	89.36%	88.30%	88.58%
50	64,000	16,000	90.17%	89.86%	89.95%	90.40%
70	64,000	16,000	89.91%	89.70%	Not done	Not done
100	64,000	16,000	90.11%	89.97%	Not done	Not done

* nRep and nReal are 4,000 each for ABC and BCa because of computation time required and n=70 and 100 are omitted for the same reason

We observe that EP and Normal-t were the best in general for this distribution so in (1) we only compared these two and found that the coverage is most accurate at 50% and 60%. For 80% and on up the actual coverage is considerably below the nominal coverage, which is due to the small sample size of 25. In (2) we see for confidence level of 90%, the sample size needs to be 50 or higher for the accuracy to be good. Again, we see that EP and Normal-t are slightly better especially when n is 20 or less. Also, in almost every case for each estimation procedure, we find that the true coverage is less than the nominal coverage. There are a few exceptions where the coverage is slightly higher than nominal.

2.5 Normal (0, 1) Distribution

In this example nSize is 25. 16,000 samples (nReal) of size 25 were generated by Monte Carlo. This is the same example used in Schenker (1985) as the sampling distribution for the observed data but Schenker did not do such an extensive set of simulations and his work preceded the development of BCa and ABC. Also, Schenker only looked at three sample sizes 20, 35 and 100 and did not look at changing coverage probabilities. We considered 64,000 Monte Carlo replications for each case and compare Normal-t, Efron's Percentile Method, BC, ABC and BCa. We used confidence levels of 50%, 60%, 70% and 80% for the Normal-t approximation, Efron's Percentile Method, and BC. We used confidence levels of 90%, 95% and 99% for Normal-t, Efron's Percentile Method, BC, ABC and BCa. This was chosen as a case where Normal-t and Efron's Percentile Method were thought to be the best based on Table 1 (from Efron and Tibshirani [1986]).

We note that for percentiles of 70% and below EP is closer to the nominal coverage than BC. For 80% and higher the two methods are close to the same with BC being slightly better. As a referee pointed out, Schenker looked at 90% coverage but none of the lower percentiles. So it was not possible to notice this change that takes place at 80% nominal coverage. Schenker did show through use of the Wilson-Hilferty transformation that normalizes chi-square random variables that the assumptions underlying the fitness of the percentile method and the BC method do not hold for N(0,1) data. If the reader is interested in the Wilson-Hilferty and other approximations to chi square distributions based on percentiles of the standard normal distribution see Chernick and Murthy (1983). The original contribution of Wilson and Hilferty is found in Wilson and Hilferty (1931).

Table 6: Normal (0, 1) Distribution: Results: for 50% nominal coverage.

Sample size nSize	nRepl	nReal	Normal-t	EP	BC
10	64,000	16,000	40.32%	41.01%	36.45%
20	64,000	16,000	44.36%	44.67%	42.75%
25	64,000	16,000	45.67%	45.76%	44.48%
30	64,000	16,000	46.48%	46.48%	45.50%
40	64,000	16,000	47.11%	47.06%	46.07%
50	64,000	4000	47.47%	47.56%	46.86%
100	64,000	4000	48.80%	48.72%	48.39%

Results: for 60% nominal coverage

Sample size nSize	nRepl	nReal	Normal-t	EP	BC
10	64,000	16,000	49.42%	49.18%	45.94%
20	64,000	16,000	53.87%	53.64%	52.15%
25	64,000	6,000	55.21%	54.93%	53.83%
30	64,000	16,000	56.18%	55.89%	55.02%
40	64,000	16,000	56.80%	56.56%	55.88%
50	64,000	4000	57.16%	56.94%	56.54%
100	64,000	4000	58.73%	58.55%	58.21%

Results: for 70% nominal coverage

Sample size nSize	nRepl	nReal	Normal-t	EP	BC
10	64,000	16,000	58.86%	57.33%	55.28%
20	64,000	16,000	63.54%	62.72%	62.00%
25	64,000	16,000	64.80%	64.10%	63.54%
30	64,000	16,000	65.96%	65.42%	64.54%
40	64,000	16,000	66.41%	65.94%	65.81%
50	64,000	4000	66.88%	66.40%	66.49%
100	64,000	4000	68.47%	68.15%	68.10%

Results: for 80% nominal coverage

Sample size nSize	nRepl	nReal	Normal-t	EP	BC
10	64,000	16,000	68.26%	65.86%	64.68%
20	64,000	16,000	73.20%	71.94%	72.05%
25	64,000	16,000	74.39%	73.39%	73.39%
30	64,000	16,000	75.28%	74.42%	74.83%
40	64,000	16,000	76.15%	75.47%	75.94%
50	64,000	4000	76.78%	76.24%	76.23%
100	64,000	4000	78.31%	78.02%	78.20%

Results: for 90% nominal coverage

Sample size nSize	nRepl	nReal	Normal-t	EP	BC	ABC *	BCa *
10	64,000	16,000	77.84%	74.83%	75.28%	73.98%	77.51%
20	64,000	16,000	82.83%	81.64%	82.56%	82.39%	83.20%
25	64,000	16,000	83.99%	83.03%	84.05%	84.02%	84.67%
30	64,000	16,000	85.09%	84.31%	85.02%	84.68%	84.95%
40	64,000	16,000	86.02%	85.55%	86.33%	86.13%	86.26%
50	64,000	4000	86.36%	86.01%	86.65%	87.54%	87.53%
100	64,000	4000	88.18%	87.94%	88.31%	88.48%	88.63%

Results: for 95% nominal coverage

Sample size nSize	nRepl	nReal	Normal-t	EP	BC	ABC *	BCa *
10	64,000	16,000	83.03%	80.13%	81.44%	79.90%	83.74%
20	64,000	16,000	87.75%	86.96%	88.35%	88.17%	89.68%
25	64,000	16,000	88.94%	88.43%	89.83%	90.03%	90.72%
30	64,000	16,000	89.79%	89.50%	90.63%	90.49%	90.94%
40	64,000	16,000	91.16%	90.96%	91.83%	91.85%	92.16%
50	64,000	4000	91.35%	91.32%	92.15%	92.59%	92.91%
100	64,000	4000	93.09%	93.09%	93.55%	93.96%	93.93%

Results: for 99% nominal coverage

Sample size nSize	nRepl	nReal	Normal-t	EP	BC	ABC *	BCa *
10	64,000	16,000	86.21%	87.86%	85.23%	88.78%	86.21%
20	64,000	16,000	92.72%	94.05%	94.68%	95.17%	92.72%
25	64,000	16,000	94.17%	95.31%	95.71%	95.99%	94.17%
30	64,000	16,000	95.08%	95.99%	96.38%	96.61%	95.08%
40	64,000	16,000	96.09%	96.87%	97.15%	97.39%	96.09%
50	64,000	4000	96.54%	97.16%	97.76%	97.84%	96.54%
100	64,000	4000	97.76%	98.12%	98.33%	98.41%	97.76%

***BCa and ABC have nRep = 16,000 and nReal = 2,000 due to the computational complexity that causes a very long running time.**

From the preceding table we conclude that all methods over-estimate coverage by 10 - 12% when nSize = 10. When nSize =100 these methods over-estimate their asymptotic values by approximately 2%. For confidence levels from 50% to 80%

the order of preference in accuracy is, Normal-t, EP and BC regardless of sample size. For confidence levels from 90% to 99% the order of preference is BCa, ABC, BC and EP. For any given nSize the results do not change much as the asymptotic coverage probability changes. Coverage improves by about 10% as the sample size increases from 10 to 100.

2.6 ln[Normal (0,1)] Distribution

In this example we considered nSize = 25. 64,000 (nRepl) samples of size 25 were generated by Monte Carlo. We considered 16,000 bootstrap iterations for each case and compare Normal-t, Efron's Percentile Method, BC, ABC and BCa. We used confidence levels of 50%, 60%, 70% 80% 90% 95% and 99% for the comparisons. This is an example of a highly skewed distribution for the original data.

Table 7: ln[Normal (0,1)] Distribution: Results for nominal coverage ranging 50% to 99% and nSize=25.

Nominal Confidence level	nRepl	nReal	Normal-t	EP	BC	ABC	BCa
50%	64,000	16,000	24.91%	32.42%	21.86%	24.34%	21.99%
60%	64,000	16,000	31.45%	35.87%	35.32%	30.08%	35.95%
70%	64,000	16,000	38.91%	38.91%	41.49%	36.38%	43.07%
80%	64,000	16,000	44.84%	43.74%	46.70%	43.90%	48.71%
90%	64,000	16,000	50.32%	50.11%	52.52%	53.03%	56.43%
95%	64,000	16,000	53.83%	53.06%	56.60%	59.09%	61.66%
99%	64,000	16,000	60.05%	59.00%	61.68%	65.43%	67.29%

Table 8. ln[Normal(0,1)] Distribution: Results for 50% nominal coverage.

nSize	nRepl	nReal	Normal-t	EP	BC	ABC	BCa
10	64,000	16,000	17.86%	25.26%	20.11%	16.88%	18.37%
25	64,000	16,000	24.91%	32.42%	21.86%	24.34%	21.99%
50	16,000	16,000	29.69%	35.22%	26.49%	28.98%	27.01%
100	16,000	16,000	33.38%	37.74%	32.27%	32.74%	31.60%
250	16,000	16,000	36.56%	39.60%	36.36%	36.29%	35.58%
500	16,000	16,000	39.65%	42.03%	39.09%	38.46%	38.81%
1000	16,000	16,000	41.33%	43.18%	41.06%	*	40.72%
2000	16,000	16,000	43.23%	44.69%	42.48%	*	42.46%

*** Omitted calculations due to excessive computational time.**

Results for 60% coverage.

nSize	nRepl	nReal	Normal-t	EP	BC	ABC	BCa
10	64,000	16,000	22.68%	28.61%	26.52%	20.84%	25.78%
25	64,000	16,000	31.45%	35.87%	35.32%	30.08%	35.95%
50	16,000	16,000	37.02%	40.28%	39.38%	35.59%	40.24%
100	16,000	16000	41.51%	43.76%	43.19%	40.13%	43.55%
250	16,000	16,000	45.21%	46.80%	46.42%	44.42%	46.68%
500	16,000	16,000	48.37%	49.76%	48.98%	49.94%	49.28%
1000	16,000	16,000	50.74%	51.59%	50.99%	*	51.07%
2000	16,000	16,000	52.85%	53.64%	52.24%	*	52.13%

* Omitted calculations due to excessive computational time.

Results for 70% coverage.

nSize	nRepl	nReal	Normal-t	EP	BC	ABC	BCa
10	64,000	16,000	28.64%	30.59%	31.30%	25.32%	32.54%
25	64,000	16,000	38.91%	39.07%	41.49%	36.38%	43.07%
50	16,000	16,000	44.89%	44.54%	46.62%	42.96%	48.11%
100	16,000	16,000	49.74%	49.49%	50.75%	47.98%	52.49%
250	16,000	16,000	54.21%	54.10%	54.74%	53.12%	55.69%
500	16,000	16,000	57.83%	57.68%	57.94%	55.50%	58.63%
1000	16,000	16,000	60.36%	60.37%	59.97%	*	60.37%
2000	16,000	16,000	62.54%	62.38%	61.47%	*	61.67%

* Omitted calculations due to excessive computational time.

Results for 80% coverage.

nSize	nRepl	nReal	Normal-t	EP	BC	ABC	BCa
10	64,000	16,000	35.04%	33.76%	35.01%	30.51%	36.74%
25	64,000	16,000	44.84%	43.74%	46.70%	43.90%	48.71%
50	16,000	16,000	51.14%	50.51%	53.11%	51.49%	55.34%
100	16,000	16,000	56.48%	56.19%	58.61%	57.38%	60.54%
250	16,000	16,000	62.26%	62.14%	63.29%	63.06%	64.81%
500	16,000	16,000	66.24%	66.08%	67.10%	65.70%	68.11%
1000	16,000	16,000	69.31%	69.03%	69.35%	*	69.80%
2000	16,000	16,000	71.80%	71.40%	71.28%	*	71.58%

* Omitted calculations due to excessive computational time.

Results for 90% coverage.

nSize	nRepl	nReal	Normal-t	EP	BC	ABC	BCa
10	64,000	16,000	39.98%	37.12%	39.38%	37.11%	41.03%
25	64,000	16,000	50.32%	50.11%	52.52%	53.03%	56.13%
50	16,000	16,000	56.93%	57.39%	60.43%	62.04%	64.63%
100	16,000	16,000	62.93%	63.93%	66.71%	68.50%	70.27%
250	16,000	16,000	69.35%	70.56%	72.74%	74.41%	75.33%
500	16,000	16,000	74.13%	74.99%	76.63%	77.09%	78.69%
1000	16,000	16,000	77.63%	78.40%	79.59%	*	80.81%
2000	16,000	16,000	80.38%	80.85%	81.83%	*	82.36%

* Omitted calculations due to excessive computational time.

Results for 95% coverage.

nSize	nRepl	nReal	Normal-t	EP	BC	ABC	BCa
10	64,000	16,000	43.36%	39.53%	41.56%	39.21%	44.58%
25	64,000	16,000	53.83%	53.06%	56.60%	59.09%	61.66%
50	16,000	16,000	60.94%	61.66%	65.26%	69.02%	70.65%
100	16,000	16,000	67.13%	68.88%	71.26%	75.70%	76.82%
250	16,000	16,000	73.96%	75.94%	78.18%	81.83%	82.01%
500	16,000	16,000	78.59%	80.56%	82.58%	84.07%	85.26%
1000	16,000	16,000	82.23%	83.72%	85.23%	*	86.97%
2000	16,000	16,000	85.06%	86.48%	87.43%	*	88.85%

*** Omitted calculations due to excessive computational time.**

Results for 99% coverage.

nSize	nRepl	nReal	Normal-t	EP	BC	ABC	BCa
10	64,000	16,000	49.51%	42.84%	44.92%	34.14%	47.83%
25	64,000	16,000	60.05%	59.00%	61.68%	65.43%	67.29%
50	16,000	16,000	67.64%	68.90%	71.58%	77.44%	78.06%
100	16,000	16,000	74.11%	76.45%	78.99%	84.69%	84.82%
250	16,000	16,000	80.93%	83.71%	85.47%	90.16%	89.82%
500	16,000	16,000	85.53%	88.16%	89.41%	92.24%	92.68%
1000	16,000	16,000	88.49%	90.74%	91.77%	*	94.20%
2000	16,000	16,000	91.31%	93.13%	93.83%	*	95.49%

From Table 8 we see that all methods over-estimate coverage (nominal coverage higher than actual coverage) by a factor of two when nSize=10. For nSize=100 these methods over-estimate coverage by 10% or more. Even for nSize=2000, coverage error is still 3.5% or more for the best method (BCa @ 99% nominal confidence). For nominal confidence levels from 50% to 60% the EP method generally performs best. For confidence levels of 60% or more the BCa method is generally best in accuracy (at 60% EP and BCa are virtually equal). As nSize becomes large, the differences among methods shrink in size. For nSize small (10 or less), the Normal-t method performs best, probably because nSize is too small for the generation of reasonable resampling distributions. Regardless of these comments, it should be noted that all methods have large coverage errors for nSize = 100 or less, and this does not improve much even for nSize as large as 2000,

Table 9. ln[Normal(0,1)] Distribution: Apparent asymptotic order vs. Confidence Level							
Nominal Confidence Level							
Method	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**99%**
Normal-t	-0.2952	-0.3122	-0.3278	-0.3415	-0.3367	-0.3352	-0.3797
EP	-0.2733	-0.3019	-0.3278	-0.3351	-0.3468	-0.3688	-0.4423
BC	-0.3064	-0.2673	-0.2776	-0.3064	-0.3493	-0.3740	-0.4504
BCa	-0.3023	-0.2555	-0.2631	-0.2901	-0.3286	-0.3697	-0.4823
ABC	-0.2592	-0.2709	-0.2734	-0.3019	-0.3377	-0.3816	-0.5065

Table 9 shows the slope of a straight line fit of the natural logarithm of the absolute value of coverage error from nominal vs. ln(nSize) for nSize = 50 to 2000 for each of the methods used. A "first-order" accurate method would be expected to have a slope of –0.5 in this fit, and a "second-order" accurate method, a slope of –1.0 (see Section 3.1 below). Instead, all methods are sub-first-order in accuracy, except for BCa and ABC for the 99% nominal confidence level. The order of accuracy increases with nominal confidence level. It is not clear why only sub-first-order accuracy is obtained, but it may relate to the non-existence of the moment generating function for this distribution. It is also not clear why coverage accuracy so strongly depends upon confidence level, although obviously low confidence levels correspond to the central portion of the distribution and high confidence level depends more heavily on the extremes.

3. SUMMARY, CONCLUSIONS AND REMARKS

The simulations addressed only the problem of confidence interval estimation of a population variance from a variety of distributions by a bootstrap procedure. Our comments and conclusions therefore only pertain to that situation. When bootstrapping other parameters the properties of the various bootstrap procedures can be very different. We considered only two-sided confidence intervals for the variance when the confidence coefficient ranged from 50% to 99%. The distributions that were simulated and reported on in this paper were the Gamma(2, 3) distribution, Student's t distribution with 5 degrees of freedom, the uniform distribution on the interval [0, 1], the standard normal distribution, and the lognormal distribution with the particular normal distribution taken to be the standard normal. Properties of the interval estimates that have a bearing on the results in order, bias-correction, the type of statistic used to estimate the parameter, connection of the asymptotic theory to the resampling distribution for the estimate of the parameter of interest, dependence of the methods on nReal and whether or not the sampling distribution of the parameter estimate is symmetric. These remarks are covered in the following subsections.

3.1 Order

Order of accuracy in terms of asymptotic error in powers of $1/\sqrt{n}$Size is typically defined in terms of the asymptotic error of an Edgeworth expansion of the sampling distribution, see Hall (1992) for a thorough discussion. An approximation which is correct up to second-order moments inclusive is "first-order" in accuracy, one which is correct up to third-order moments (i.e., skewness) is "second-order" in accuracy, and one which is correct on all four first moments is "third-order" in accuracy.

"First-order" in sample size denotes asymptotic error O [$1/\sqrt{n}$Size], where nSize is the sample size of the original data. The coverage accuracy of Normal-t, EP,

and BC methods are usually first-order in nSize. "Second-order" in sample size denotes asymptotic error O [1/nSize], due to correction for skewness. The coverage accuracy of the BCa, and ABC methods are second-order in nSize. The bias in the variance estimate is also second-order in nSize.

Also as shown by Hall (1992) the order can be different for one-sided intervals than it is for two-sided ones, due to cancellation in error between the two sides. Only two-sided intervals were considered here. Note that if the sampling distribution is symmetric then no skewness correction is needed. Hence, first-order methods such as Normal-t, EP and BC become second-order in accuracy due to the zero third-order central moment.

3.2 Bias Correction

The classical bias in the variance statistic is reintroduced with each level of bootstrap. The factor of nSize/(nSize -1) is second-order in nSize and becomes relatively unimportant as nSize increases to 25 or more. Consequently, estimating bias and correcting for it in the case of the variance was not found to have significant benefit with respect to confidence interval coverage accuracy. This is particularly true when the bias correction is estimated from resampling and not provided by theory.

3.3 Summary Statistics

It may be obvious, but it bears repeating that bootstrap resampling depends upon the statistic of interest (here the variance). Smoothness properties for the statistic are required for the existence of Edgeworth expansions which are needed to mathematically prove the order of the approximation. For very irregular statistics the ordinary bootstrap may not converge properly (e.g. the extreme order statistics).

3.4 Asymptotics

It is important to understand in comparing methods that the asymptotic limit differs between methods. For example, the EP method confidence interval coverage accuracy has the asymptotic rate of convergence based upon the resampling distribution of the statistic of interest (here, the estimate of variance) at the confidence level quantiles.

The BCa method, however, not only depends on the resampling quantiles of the statistic of interest, but also on the convergence of the coefficient of skewness estimate ("acceleration"), which depends upon the second and third moments, to the theoretical value. In addition, the percentiles of interest are not those associated directly with the confidence interval, but rather shifted percentiles derived from the BCa methodology. In the presence of skewness, one of the two percentiles will be more extreme than that for the nominal interval.

So what is not commonly discussed is this trade-off between the asymptotic order of the method and the size of the errors due to skewness estimation as well as the size of their coefficients in the asymptotic expansion. It may very well happen that for small nSize, the EP method could outperform in coverage accuracy the BCa method, even for a skewed resampling distribution. Inaccuracy in estimating higher order moments is well known, and explains the usual recommendation not to use Edgeworth expansions of higher than the fourth order (kurtosis-corrected).

Skewness matters less near the center of a distribution and it has been shown in our simulations that the EP method can outperform BCa method even for nSize at 500 or higher if the confidence level is less than 70%.

3.5 Dependence on nReal

Higher-order methods, such as BCa, depending upon shifted percentiles must have accurate enough resampling distributions to allow accurate estimation of such quantiles.

As an example, the 95% confidence interval using the EP method requires 2.5% and 97.5% quantile estimates from the resampling distribution. The standard error of these quantiles is on the order of $\sqrt{m}$, where m is then number of realizations at or beyond the quantile. For the 2.5% quantile, the standard error then will be proportional to $\sqrt{[0.025\ nReal]}$, and the relative standard error is proportional to $1 / \sqrt{[0.025\ nReal]}$. If this factor is required to be, e.g., less than 20%, then we need $nReal > 25/0.025 = 1000$. This has been found in the simulations to be approximately the required minimum nReal for the EP method and this confidence interval.

Now suppose the BCa method is used, which shifts the 2.5% percentile to, say, the 0.1% percentile. For the BCa method, we would then require $nReal > 25/0.001 = 25000$, a large increase. So a clear computational cost of large nReal may be required to attain full accuracy with the BCa method.

For fixed nReal (e.g., 500), the BCa method may not be superior to the EP method. This may be one way of explaining Efron's recommendation for at least 1000 bootstrap samples when estimating confidence intervals.

It should be noted that, as mentioned above, for a statistic with a symmetric sampling distribution (e.g., mean based upon a normal or uniform distribution), the Normal-t,, EP, and BC methods all become second-order accurate, due to vanishing of the coefficient of skewness. Similarly, for slightly skewed distributions with 2-sided confidence intervals, the errors made on the two sides tend to cancel whereas EP and BC are only first order accurate for one-sided intervals.

3.6 Symmetric Sampling Distribution.

It should be noted that, as mentioned above, for a statistic with a symmetric sampling distribution (e.g., mean based upon a normal or uniform distribution), the Normal-t,, EP, and BC methods all become second-order accurate, due to vanishing of the coefficient of skewness. Similarly, for slightly skewed distributions with 2-sided confidence intervals, the errors made on the two sides tend to cancel whereas EP and BC are only first order accurate for one-sided intervals.

3.7 Acceptable Coverage

In our simulations for the variance, we find situations where all the methods have high coverage error. So even if BCa is superior to EP because of the second-order accuracy, the difference in accuracy pales in comparison to the coverage error for both methods.

3.8 Choice of Method

Generally, for confidence levels below 70%, the EP method gives the best coverage accuracy for any of the chosen distributions when estimating a variance. When confidence levels are at or above 70%, however, the number of realizations required to provide a good enough approximation to the bootstrap sampling distribution can be very large and time consuming particularly when BCa intervals are estimated using an interpreted statistical system such as R. For even modest sample sizes (e.g. nSize > 25) the ABC method is computationally much more efficient and has almost the same accuracy as BCa. This dependence on the confidence level may not have much practical importance since most confidence intervals use levels of 90%, 95%, 97.5% and 99% where the results are more consistent as to which is the better method.

ACKNOWLEDGMENT

We would like to thank a very diligent anonymous referee for carefully reading the original manuscript and offering many suggestions to improve the detail and clarity of the paper. We believe the article benefited greatly from changes we made based on the referee's suggestions. The referee also clarified several points regarding the contributions in Schenker's important work.

REFERENCES

Bahadur, R. and Savage, L. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.* **27,** 1115-1122.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* **19,** 1141-1164.

Chernick, M.R. (1999). *Bootstrap Methods: A Practitioner's Guide.* Wiley, New York.

Chernick, M.R. (2007). *Bootstrap Methods: A Guide for Practitioners and Researchers, 2nd Edition.* Wiley, New York.

Chernick, M. R. and Murthy, V. K. (1983). Chi-square percentiles: old and new approximations, with applications to sample size determination. *Amer. J. Math. and Manag. Sci.* **3,** 145-161.

DiCiccio, T. J. and Romano, J. P. (1988). A review of bootstrap confidence intervals (with discussion). *J. Roy. Statist. Soc. B* **50,** 338-370.

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

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrial and Applied Mathematics CBMS-NSF Regional Conference Series **38,** Philadelphia.

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

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science* **1,** 54-77.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

Hesterberg, T., Monaghan, S., Moore, D. S., Clipson, A. and Epstein, R. (2003) *Bootstrap Methods and Permutation Tests: Companion Chapter 18 to The Practice of Business Statistics.* W. H. Freeman and Company, New York.

Manly, B. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology, 3rd Edition.* Chapman & Hall/CRC, Boca Raton.

Shao, J. and Yu, D. (1995). *The Jackknife and Bootstrap.* Springer-Verlag, New York.

Schenker, N. (1985). Qualms about bootstrap confidence intervals. *J. Am. Statist. Assoc.* **80,** 360-361.

Wilson, E. B. and Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences U. S. A.*. **28,** 94-100.

Appendix

R code for Bootstrap Simulations of Variance Confidence Intervals for the Lognormal example
&
Table 1

R code for Bootstrap Simulations of Variance Confidence Intervals for the Lognormal example

```
#Bootstrap Method, LnNormal(0,1) Variance Distribution
#created: 11.04.07 by r.a. labudde from bootNormvaryN.r
#changes: 05.19.08 ral, fix up
library(boot)
trueMean = exp(0.5)
trueVar = exp(1)*(exp(1)-1)
p =c(.25, .75, .2, .8, .15, .85, .1, .9, .05, .95, .025, .975, .005, .995)
cl = c(.5,.6,.7,.8,.9,.95,.99)
varf = function (x,i) { var(x[i]) }
fabc = function (x, w) { sum(x^2*w)/sum(w)-(sum(x*w)/sum(w))^2 } #wgt average
(biased) variance
nRepl = 1000
nReal = 1000  #number realizations
for (N in c(10,25,50,100,250,500,1000)) {
 #calculate interval coverages
 nCl90 = rep(0,7) #array of counts for 7 types of intervals
 nCl95 = rep(0,7) #array of counts for 7 types of intervals
 nCl99 = rep(0,7) #array of counts for 7 types of intervals
 v = NULL #hold sample variance estimates
 bv = NULL #hold bootstrap mean variance estimates
 bias = NULL #hold bias estimates
 for (i in 1:nRepl) {
  s = rlnorm(N, 0, 1) #get N random lnnormal deviates
 v[i] = var(s) #get sample variance
   b = boot(s, varf, R=nReal) #bootstrap
   bv[i] = mean(b$t) #bootstrap mean variance
   bias[i] = bv[i] - v[i] #bias estimate
   bCl90 = boot.ci(b, conf=0.90)
   abcCl90 = abc.ci(s, fabc, conf=0.90)
   if (bCl90$normal[2]<= trueVar && bCl90$normal[3]>=trueVar) nCl90[1] =
nCl90[1] +1
```

```
      if (bCI90$basic[4]<= trueVar && bCI90$basic[5]>=trueVar) nCI90[2] =
    nCI90[2] +1
      if (bCI90$percent[4]<= trueVar && bCI90$percent[5]>=trueVar) nCI90[3]
    =nCI90[3] +1

  if (bCI90$bca[4]<= trueVar && bCI90$bca[5]>=trueVar) nCI90[4] =nCI90[4] + 1
  if (abcCI90[2]<= trueVar && abcCI90[3]>=trueVar) nCI90[5] =nCI90[5] + 1
#   bt = tilt.boot(s, varf, R=c(nReal,nReal,nReal), alpha=c(.05,.95))
#   if (bt$theta[1]<= trueVar && bt$theta[2]>=trueVar) nCI90[6] =nCI90[6] + 1
  bCI95 = boot.ci(b, conf=0.95)
  abcCI95 = abc.ci(s, fabc, conf=0.95)
  if (bCI95$normal[2]<= trueVar && bCI95$normal[3]>=trueVar) nCI95[1] = nCI95[1] +1
  if (bCI95$basic[4]<= trueVar && bCI95$basic[5]>=trueVar) nCI95[2] = nCI95[2] +1
  if (bCI95$percent[4]<= trueVar && bCI95$percent[5]>=trueVar) nCI95[3] =nCI95[3] +1
  if (bCI95$bca[4]<= trueVar && bCI95$bca[5]>=trueVar) nCI95[4] =nCI95[4] + 1
  if (abcCI95[2]<= trueVar && abcCI95[3]>=trueVar) nCI95[5] =nCI95[5] + 1
#   bt = tilt.boot(s, varf, R=c(nReal,nReal,nReal), alpha=c(.025,.975))
#   if (bt$theta[1]<= trueVar && bt$theta[2]>=trueVar) nCI95[6] =nCI95[6] + 1
  bCI99 = boot.ci(b, conf=0.99)
  abcCI99 = abc.ci(s, fabc, conf=0.99)
  if (bCI99$normal[2]<= trueVar && bCI99$normal[3]>=trueVar) nCI99[1] = nCI99[1] +1
  if (bCI99$basic[4]<= trueVar && bCI99$basic[5]>=trueVar) nCI99[2] = nCI99[2] +1
  if (bCI99$percent[4]<= trueVar && bCI99$percent[5]>=trueVar) nCI99[3] =nCI99[3] +1
  if (bCI99$bca[4]<= trueVar && bCI99$bca[5]>=trueVar) nCI99[4] =nCI99[4] + 1
  if (abcCI99[2]<= trueVar && abcCI99[3]>=trueVar) nCI99[5] =nCI99[5] + 1
#   bt = tilt.boot(s, varf, R=c(nReal,nReal,nReal), alpha=c(.005,.995))
#   if (bt$theta[1]<= trueVar && bt$theta[2]>=trueVar) nCI99[6] =nCI99[6] + 1
 }
 cat("nSize: ",N, " Sample: ",mean(v)," Boot: ",mean(bv)," Bias: ",mean(bias),"\n")
 cat("nSize: ",N, "CI90: ",nCI90/nRepl, "\n")
 cat("nSize: ",N, "CI95: ",nCI95/nRepl, "\n")
 cat("nSize: ",N, "CI99: ",nCI99/nRepl, "\n")
}
```

Table 1: Four Methods for Setting Approximate Confidence Intervals for a Real-Valued Parameter θ. (From Efron and Tibshirani (1986), with permission.)

Method	Abbreviation	α-Level Endpoint	Correct if
Standard Normal Approximation	$\theta_S[\alpha]$	$\theta_h + \sigma_h z^{(\alpha)}$	$\theta_h \approx N(\theta, \sigma^2)$ with σ constant
Percentile	$\theta_P[\alpha]$	$G_h^{-1}(\alpha)$	There exists a monotone transformation such that $\phi_h=g(\theta_h)$ where $\phi = g(\theta)$ and $\phi_h \approx N(\phi, \tau^2)$ and τ is constant
Bias-corrected	$\theta_{BC}[\alpha]$	$G_h^{-1}(\Phi\{2z_0 + z^{(\alpha)}\})$	There exists a monotone transformation such that $\phi_h=g(\theta_h)$ where $\phi = g(\theta)$ and $\phi_h \approx N(\phi - z_0\tau, \tau^2)$ and τ and z_0 are constant
BC_a	$\theta_{BCa}[\alpha]$	$G_h^{-1}(\Phi\{z_0 + [z_0 + z^{(\alpha)}]/[1 - a(z_0+z^{(\alpha)})]\})$	There exists a monotone transformation such that $\phi_h=g(\theta_h)$ where $\phi = g(\theta)$ and $\phi_h \approx N(\phi - z_0\tau_0, \tau_0^2)$ where $\tau_0 = 1 + a\phi$ and z_0 and a are constant. Note that "a" is estimated by the jackknife estimate of the coefficient of skewness μ_3/σ^3 divided by 6.