Research Article

Double Bootstrap-t One-Sided Confidence Interval for Population Variance of Skewed Distributions

Wararit Panichkitkosolkul

Department of Mathematics and Statistics, Faculty of Science and Technology,
Thammasat University, Phathumthani, Thailand.
Corresponding author. E-mail address: wararit@mathstat.sci.tu.ac.th

Received February 4, 2013; Accepted May 7, 2013

Abstract

This paper proposes a double bootstrap-t one-sided confidence interval for population variance of skewed distributions. The upper endpoint and lower endpoint confidence intervals are studied. The one-sided confidence intervals based on the chi-square statistic, bootstrap-t method and double bootstrap-t method are compared via Monte Carlo simulations. The simulation results indicated that the coverage probabilities of bootstrap-t confidence interval can be increased by using double bootstrap resampling. The upper endpoint confidence interval using double bootstrap-t method predominates the other methods with respect to the coverage probability criteria. The performance of the proposed one-sided confidence interval is illustrated with an empirical example.

Key Words: Double Bootstrap-t; Confidence Interval; Variance; Skewed Distribution

Introduction

A confidence interval (CI) for a population parameter gives a bound computed from sample data containing the true value of the parameter with a specified confidence level. Confidence interval plays a significant role in statistical inference regarding the parameter. For confidence interval for population variance, well-known existing methods are based upon the chi-square statistic which is introduced by Pearson (1900). Based on the chi-square statistic, the upper endpoint and lower endpoint $(1-\alpha)100\%$ confidence intervals for σ^2 are (Cojbasic and Loncar, 2011)

$$UCI_{\chi^2} = \left(0\,, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha}}\right), \qquad (1)$$

and

$$LCI_{\chi^2} = \left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha}}\,, +\infty\right), \qquad (2)$$

where $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, $\chi^2_{n-1,\alpha}$ and $\chi^2_{n-1,1-\alpha}$ are the $(\alpha)100^{th}$ and $(1-\alpha)100^{th}$ percentiles of the central chi-square distribution with $n-1$ degrees of freedom. It is well-known that these upper endpoint and lower endpoint confidence intervals for σ^2 are constructed under the normal distribution. However, the underlying distribution is non-normal in some situations. Hence, it may be a skewed distribution. To deal with these situations, many researchers have proposed confidence interval for σ^2 of skewed distributions. For example, Bonett (2006) provided an approximate confidence interval for standard deviation and his proposed confidence interval is nearly exact under the normal distribution for small samples, and under the non-normal distributions for moderate samples. Cojbasic and Tomovic (2007) presented confidence intervals for the population variance and the difference in variances of two populations based on the ordinary t-statistics combined with the bootstrap method. In addition, Cojbasic and Loncar (2011) studied the coverage accuracy of one-sided bootstrap-t confidence intervals for the population variances combined with Hall's and Johnson's transformation. In some cases, we found that the above mentioned confidence intervals provided the coverage probability less than the nominal confidence interval. In this paper, we show that a slight modification of the usual bootstrap confidence interval can help to improve the accuracy of coverage probability. One approach of increasing the coverage probability of confidence interval is to use the double bootstrap described by Nankervis (2002, 2005), which can be thought of as bootstrapping the bootstrap (Scherer and Martin, 2005). Namely, the error in the coverage probability of bootstrap confidence intervals can be reduced by the use of double bootstrap confidence intervals (Nankervis, 2002).

The structure of this paper is as follows. The next section presents the bootstrap-t one-sided confidence interval for the variance, and the third section provides the details of the double bootstrap-t one-sided confidence interval for the variance. The fourth section presents the Monte Carlo simulation results. An empirical example is given in the fifth section, and the conclusions are in the sixth section.

Bootstrap-t One-Sided Confidence Interval for the Variance

The bootstrap introduced by Efron (1979) is a computer-based and resampling method for assigning measures of accuracy to statistical estimates (Efron and Tibshirani, 1993). For a sequence of independent and identically distributed (i.i.d.) random variables, the bootstrap procedure can be defined as follows (Tosasukul et al., 2009). Let $X_1, X_2, ..., X_n$ be independently and identically distributed random variables from some distribution with mean μ and variance σ^2. Let the random variables $\{X_j^*, 1 \le j \le m\}$ be the result from sampling m times from the population with replacement from the n observations $X_1, X_2, ..., X_n$. The random variables $\{X_j^*, 1 \le j \le m\}$ are called the bootstrap samples from original data $X_1, X_2, ..., X_n$. Let $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ be a sample variance. It is well-known that the pivotal quantity $(n-1)S^2/\sigma^2$ has central chi-square distribution with $n-1$ degrees of freedom (Bonett, 2006). A confidence interval for population variance can be constructed using aforementioned pivotal quantity. For large sample sizes, central chi-square distribution with $n-1$ degrees of freedom can be approximated by normal distribution with mean $n-1$ and variance $2(n-1)$ (Cojbasic and Tomovic, 2007). Therefore, the distribution of the standardized variable

$$Z = \frac{\frac{(n-1)S^2}{\sigma^2} - (n-1)}{\sqrt{2(n-1)}} = \frac{S^2 - \sigma^2}{\sqrt{\mathrm{var}(S^2)}} \qquad (3)$$

converges to standardize normal distribution as n increases to infinity. One-sided bootstrap-t confidence interval for σ^2 is calculated based on the statistic

$$T = \frac{S^2 - \sigma^2}{\sqrt{\widehat{\mathrm{var}}(S^2)}}, \tag{4}$$

where $\widehat{\mathrm{var}}(S^2)$ is a consistent estimator of the variance of S^2. Casella and Berger (2001, pp.257) have shown the estimator of $\mathrm{var}(S^2)$ for non-normal distribution such that

$$\widehat{\mathrm{var}}(S^2) = \frac{1}{n}\left(\hat{\mu}_4 - \frac{n-3}{n-1}S^4\right) \qquad \text{and}$$

$$\hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4.$$

After re-sampling B bootstrap samples, in each bootstrap sample we compute the value of the following statistic

$$T^* = \frac{S^{*2} - S^2}{\sqrt{\widehat{\mathrm{var}}(S^{*2})}}, \tag{5}$$

where S^{*2} is a bootstrap replication of statistic S^2, $\widehat{\mathrm{var}}(S^{*2}) = \frac{1}{n}\left(\hat{\mu}_4^* - \frac{n-3}{n-1}S^{*4}\right)$ and $\hat{\mu}_4^* = \frac{1}{m}\sum_{i=1}^{m}(X_i^* - \bar{X}^*)^4$. The upper endpoint and lower endpoint $(1-\alpha)100\%$ bootstrap-t confidence intervals for σ^2 are

$$UCI_B = \left(0\,,\, S^2 + \hat{t}^*_{(1-\alpha)}\sqrt{\widehat{\mathrm{var}}(S^2)}\right), \tag{6}$$

and $$LCI_B = \left(S^2 + \hat{t}^*_{(\alpha)}\sqrt{\widehat{\mathrm{var}}(S^2)}\,,\, +\infty\right), \tag{7}$$

where $\hat{t}^*_{(\alpha)}$ and $\hat{t}^*_{(1-\alpha)}$ are the $(\alpha)100^{th}$ and $(1-\alpha)100^{th}$ percentiles of T^* shown in Eq. (5).

Double Bootstrap-t One-Sided Confidence Interval for the Variance

The details of double bootstrap-t one-sided confidence interval are as follows. For each of B bootstrap replications, the first-level bootstrap samples $\{X_j^*, 1 \le j \le m\}$ are first drawn from the original data. Next, the second-level bootstrap samples $\{X_j^{**}, 1 \le j \le m\}$ are drawn from the first-level bootstrap samples. The statistic ** based on the second-level bootstrap samples is computed as follows

$$T^{**} = \frac{S^{**2} - S^2}{\sqrt{\widehat{\mathrm{var}}(S^{**2})}}, \tag{8}$$

where S^{**2} is a standard deviation of the second-level bootstrap samples $\{X_j^{**}\}$, $\widehat{\mathrm{var}}(S^{**2}) = \frac{1}{n}\left(\hat{\mu}_4^{**} - \frac{n-3}{n-1}S^{**4}\right)$ and $\hat{\mu}_4^{**} = \frac{1}{m}\sum_{i=1}^{m}(X_i^{**} - \bar{X}^{**})^4$. Therefore, the upper endpoint and lower endpoint $(1-\alpha)100\%$ double bootstrap-t confidence intervals for σ^2 are

$$UCI_{DB} = \left(0\,,\, S^2 + \hat{t}^{**}_{(1-\alpha)}\sqrt{\widehat{\mathrm{var}}(S^2)}\right), \tag{9}$$

$$LCI_{DB} = \left(S^2 + \hat{t}^{**}_{(\alpha)}\sqrt{\widehat{\mathrm{var}}(S^2)}\,,\, +\infty\right), \tag{10}$$

where $\hat{t}^{**}_{(\alpha)}$ and $\hat{t}^{**}_{(1-\alpha)}$ are the $(\alpha)100^{th}$ and $(1-\alpha)100^{th}$ percentiles of ** given in Eq. (8).

Monte Carlo Simulation Results

The following Monte Carlo experiment compares the performance of one-sided confidence intervals for the variance of skewed distributions. The simulation study was conducted using the open source statistical package R (Ihaka and Gentleman

1996) to estimate the coverage probability of one-sided confidence interval. We chose to use some of the probability density functions of Cojbasic and Loncar (2011) in the simulation study. For each probability density function, we generated ten thousand random samples from Weibull, Exponential and Lognormal distribution and used 2,000 bootstrap samples. The different sample sizes ($n = 10, 20, 50, 100$) are considered.

Table 1 illustrates the results of the estimated coverage probabilities of 95% lower endpoint confidence intervals while the estimated coverage probabilities of 95% upper endpoint confidence intervals are shown in Table 2. We begin with the results for the lower endpoint confidence intervals (Table 1). The chi-square method provides estimated coverage probabilities of the lower endpoint confidence intervals close to the nominal confidence level 0.95 when the coefficients of skewness are equal to 0 and 0.62. For example, when the underlying distribution is Weibull distribution with shape parameter 2, the estimated coverage probabilities of 95% lower endpoint confidence interval attained by the chi-square method are 0.9446, 0.9446, 0.9398 and 0.9349 for $n = 10, 20, 50$ and 100, respectively. In addition, the bootstrap-t method provides the estimated coverage probabilities close to the nominal confidence level 0.95 when the coefficients of skewness are equal to 0.62 and 2. The estimated coverage probabilities of lower endpoint confidence interval by using double bootstrap-t method are close to one as skewness coefficient gets larger.

Next, the upper endpoint confidence intervals are considered (Table 2). The estimated coverage probabilities of upper endpoint confidence interval by using both chi-square and bootstrap-t method get reasonably close to the nominal confidence level 0.95 for low skewness is low (coefficients of skewness are equal to 0 and 0.62). Furthermore, the double bootstrap-t method provides the estimated coverage probabilities more than those of other methods. However, all methods have poor estimated coverage probabilities of upper endpoint confidence interval for medium and high skewness (coefficients of skewness are equal to 6.18 and 23.73). For instance, the estimated coverage probabilities of 95% upper endpoint confidence interval for Lognormal with $\sigma^2 = 2$ and $n = 20$ are 0.2872, 0.6047 and 0.7687 by chi-square, bootstrap-t and double bootstrap-t methods, respectively. Additionally, all estimated coverage probabilities tend to increase as sample size gets larger. The above results indicate that the upper endpoint confidence interval using double bootstrap-t method dominates the other approaches for almost all situations except low skewness.

An Empirical Example

To illustrate an empirical example of one-sided confidence intervals for population variance of skewed distributions that have been presented within the previous section, we have used the real environmental data. Sulfur dioxide (SO_2) contents of air in micrograms per cubic meter for forty U.S. cities were collected from U.S. government publications. The data were obtained from 1969 to 1971 (Source: http://lib.stat.cmu.edu/DASL). The histogram, density plot, box plot and normal QQ plot of SO_2 contents are displayed in Figure 1. It indicates that the distribution of SO_2 contents was positively skewed. The 95% lower and upper endpoint confidence intervals for the variance are constructed. As shown in Table 3, the lower endpoint confidence intervals computed via double bootstrap-t method provides the widest length as compared to those obtained from chi-square and bootstrap-t method. It is corresponding with the Monte Carlo studies that the double

Table 1 The estimated coverage probabilities of 95% lower endpoint confidence interval for the variance of standard normal distribution and skewed distributions.

Distribution	Skewness coefficient	Sample size	Method		
			Chi-square	Bootstrap-t	Double bootstrap-t
Standard normal	0	10	0.9506	0.8588	0.8965
		20	0.9490	0.8942	0.9414
		50	0.9511	0.9203	0.9687
		100	0.9508	0.9328	0.9789
Weibull with shape parameter 2	0.62	10	0.9446	0.9455	0.9755
		20	0.9446	0.9541	0.9858
		50	0.9398	0.9512	0.9899
		100	0.9349	0.9440	0.9900
Exponential with mean 1	2	10	0.8805	0.9720	0.9877
		20	0.8608	0.9777	0.9969
		50	0.8416	0.9670	0.9975
		100	0.8253	0.9610	0.9950
Lognormal with $\sigma^2 = 1$	6.18	10	0.8918	0.9964	0.9991
		20	0.8593	0.9953	0.9997
		50	0.8200	0.9900	0.9998
		100	0.7909	0.9852	0.9998
Lognormal with $\sigma^2 = 2$	23.73	10	0.9221	0.9990	0.9994
		20	0.8959	0.9991	1.0000
		50	0.8642	0.9970	1.0000
		100	0.8327	0.9945	0.9999

Table 2 The estimated coverage probabilities of 95% upper endpoint confidence interval for the variance of standard normal distribution and skewed distributions.

Distribution	Skewness coefficient	Sample size	Method		
			Chi-square	Bootstrap-t	Double bootstrap-t
Standard normal	0	10	0.9460	0.9467	0.9940
		20	0.9502	0.9503	0.9917
		50	0.9476	0.9466	0.9910
		100	0.9550	0.9561	0.9914
Weibull with shape parameter 2	0.62	10	0.9492	0.9350	0.9911
		20	0.9455	0.9274	0.9837
		50	0.9438	0.9344	0.9829
		100	0.9385	0.9404	0.9837
Exponential with mean 1	2	10	0.7981	0.8259	0.9398
		20	0.7803	0.8505	0.9324
		50	0.7793	0.8939	0.9577
		100	0.7841	0.9118	0.9677
Lognormal with $\sigma^2 = 1$	6.18	10	0.5340	0.7038	0.8432
		20	0.4946	0.7233	0.8439
		50	0.4973	0.7728	0.8711
		100	0.5048	0.7977	0.8923
Lognormal with $\sigma^2 = 2$	23.73	10	0.2960	0.5979	0.7410
		20	0.2872	0.6047	0.7687
		50	0.2888	0.6395	0.7792
		100	0.3029	0.6727	0.7928

Conclusions

A double bootstrap-t one-sided confidence interval for population variance of skewed distributions has proposed in this paper. The study was carried out to compare the performance of a proposed confidence interval with the existing confidence intervals. Three one-sided confidence intervals are considered: the one-sided confidence interval based on chi-square statistic, the bootstrap-t one-sided confidence interval and the double bootstrap-t one-sided confidence interval. Based on simulation studies, the double bootstrap-t one-sided confidence interval provides good coverage probability for the upper endpoint confidence interval. On the other hand, the double bootstrap resampling can also improve the accuracy of the upper endpoint confidence interval for population variance of skewed distributions. The behind bootstrap-t method provides the estimated coverage probabilities more than those of other methods. Therefore, the double bootstrap-t method is not suitable for this case. However, the length of the upper endpoint confidence interval computed by double bootstrap-t method is shorter than other confidence intervals.

Table 3 The 95% lower and upper endpoint confidence intervals for the variance of SO_2 contents.

Method	Lower endpoint confidence interval	Upper endpoint confidence interval
Chi-square	[0 , 831.33]	[395.24 , ∞]
Bootstrap-t	[0 , 1396.97]	[298.09 , ∞]
Double bootstrap-t	[0 , 1802.90]	[204.10 , ∞]

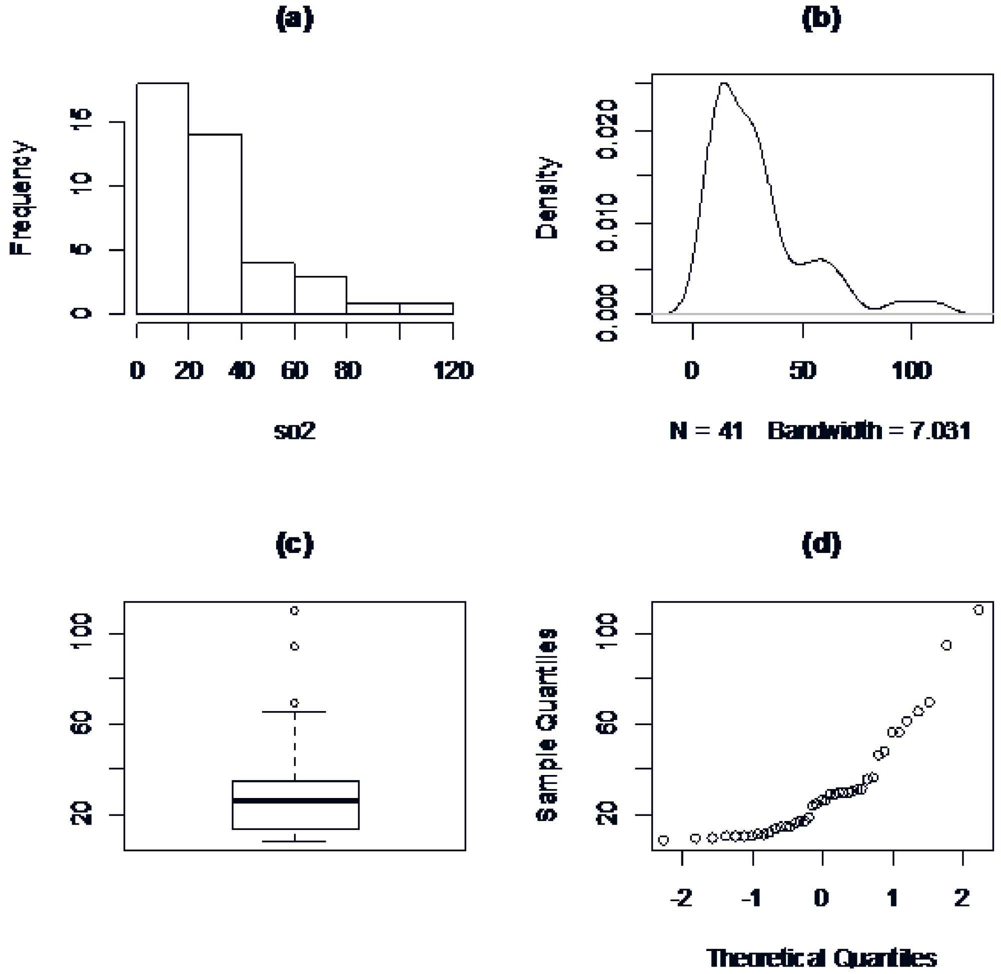

Figure 1 (a) Histogram (b) Density plot (c) Box plot and (d) Normal QQ plot of Sulfur dioxide(SO_2) contents of air for forty-one US cities.

reason is that the resulting double bootstrap confidence intervals have been shown to have a smaller order of error. For example, Hall (1986) has shown that, in general, the coverage rate of a 100(1-2α)% equal-tailed bootstrap confidence interval is corrected from 1-2α + $O(n^{-1})$ to 1-2α + $O(n^{-2})$ for a double bootstrap confidence interval. In addition, the coverage probability of double bootstrap-t lower endpoint confidence interval does not achieve exactly the nominal confidence level.

Acknowledgements

The authors thank the anonymous referees for their constructive suggestions and comments that resulted in an improved present of this paper.

References

Bonett, D. G. (2006) Approximate confidence interval for standard deviation of nonnormal distributions. *Computational Statistics & Data Analysis* 50(3): 775-782.

Casella, G. and Berger, R. L. (2001) *Statistical Inference*. Duxbury Press, Pacific Grove, pp.257.

Cojbasic, V. and Loncar, D. (2011) One-sided confidence intervals for population variances of skewed distribution. *Journal of Statistical Planning and Inference* 141(5): 1667-1672.

Cojbasic, V. and Tomovic, A. (2007) Nonparametric confidence intervals for population variances of one sample and the difference of variances of two samples. *Computational Statistics & Data Analysis* 51(12): 5562-5578.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7(1): 1-26.

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

Hall, P. (1986) On the bootstrap and confidence intervals. *Annals of Statistics* 14(4): 1431-1452.

Ihaka, R. and Gentleman, R. (1996) "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics* 5: 299-314.

Nankervis, J. C. (2002) Stopping rules for double bootstrap confidence intervals, [Online URL: www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.8480&rep=rep1&type=pdf] accessed on November 19, 2012.

Nankervis, J. C. (2005) Computational algorithms for double bootstrap confidence intervals. *Computational Statistics & Data Analysis* 49(2): 461-475.

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supported to have arisen from random sampling. *Philosophical Magazine* 50(5): 157-175.

Scherer, B. and Martin, R. D. (2005) *Introduction to Modern Portfolio Optimization with NUOPT and S-PLUS*. Springer, New York.

Tosasukul, J., Budsaba, K., and Volodin, A. (2009) Dependent bootstrap confidence intervals for a population mean. *Thailand Statistician* 7(1): 43-51.