Thailand Statistician
January 2009; 7(1) : 43-51
http://statassoc.or.th
Contributed paper

Dependent Bootstrap Confidence Intervals for a Population Mean

Jiraroj Tosasukul [a], Kamon Budsaba *[a], and Andrei Volodin [b]

[a] Department of Mathematics and Statistics, Thammasat University, Rangsit Center, Pathum Thani 12121, Thailand.

[b] Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, Canada S4S 0A2

***Author for correspondence**; e-mail: k_budsaba@hotmail.com

Received: 16 November 2008
Accepted: 18 December 2008.

Abstract

This study compares and analyzes the coverage probabilities and the average interval lengths of confidence interval for a population mean based on the dependent bootstrap procedure against those based on the independent bootstrap procedure. Both dependent and independent bootstrap confidence intervals for a population mean are computed by the Bootstrap-t, Percentile, and Modified Percentile methods. Simulations show that the coverage probabilities of the dependent bootstrap confidence intervals are similar to those of the independent bootstrap confidence intervals. The average interval lengths of the dependent bootstrap method are shorter for most situations. For both the independent and dependent bootstrap confidence intervals, the coverage probabilities increase and the average interval lengths decrease as the sample size *n* increase for Normal, Gamma, and Chi-square distributions, as well as three methods used in this work.

Keywords: average interval lengths, Bootstrap-t method, coverage probabilities, dependent bootstrap procedure, independent bootstrap procedure, Modified Percentile method, Percentile method.

1. Introduction

Efron [1] introduced the independent bootstrap as a tool to estimate the standard error of a statistic. An enormous amount of applied and theoretical research on the bootstrap technique has followed in the past two decades. The independent bootstrap is a general technique for estimating unknown quantities associated with statistical models and often used to find standard errors for estimators, confidence intervals for unknown parameters or *p* values for test statistics under a null hypothesis. Thus, the bootstrap is typically used to estimate quantities associated with the sampling distribution of estimators and test statistics.

Another bootstrap technique, introduced by Smith and Taylor [2], is called the dependent bootstrap. It resamples without replacement, which would reduce variation of estimators. This study aims to show the potential benefit of using the dependent bootstrap confidence intervals to extract coverage probabilities and average lengths of a population mean. The study compares the coverage probabilities and average lengths of several classes of bootstrap confidence intervals for population mean both independent and dependent bootstrap procedure.

The following sections explain methodology used in this study and present the results from simulations.

2. Methods

2.1 Independent bootstrap procedure

For a sequence of independent and identically distributed (i.i.d.) random variables, the independent bootstrap procedure can be defined as follows. Let the random variables $\{X^*_{n,j}, 1 \le j \le m\}$ be the results from sampling m times with replacement from the n observations $X_1, X_2 ..., X_n$. For each of the m selections, each X_i, where $1 \le i \le n$, has probability $\frac{1}{n}$ of being chosen. For each $n \ge 1$, the random variables $\{X^*_{n,j}, 1 \le j \le m\}$ is the so-called Efron [1] *independent bootstrap sample* from original data $X_1, ..., X_n$, with *bootstrap sample size* m . For example, let us draw a sample of size *n* (original data) from a population. From this sample, a random sample (*m*) from the original data is taken with a replacement selected from the *n* data values.

2.2 Dependent bootstrap procedure

For an arbitrary sequence of random variables, Smith and Taylor [2] defined the dependent bootstrap procedure as follows. Let $\{m, m \geq 1\}$ and $\{k, k \geq 1\}$ be two sequences of positive integers such that $m \leq nk$ for all $n \geq 1$. For a sample space (Ω) where $\omega \in \Omega$ and $n \geq 1$, the *dependent bootstrap* is defined to be the sample of size m, defined by $\{X'_{kn,j}, 1 \leq j \leq m\}$. This sample is drawn without replacement from the collection of nk items, which made up of k copies each of the sample observations ($X_1, \ldots, X_n$). For example, a sample of size *n* (original data) is taken from a population, where *k* is a number of copies. From this sample, a random sample (*m*) of the *k* copies of original data is drawn without a replacement.

2.3 The Bootstrap-t confidence interval method

Let X_b^*, where $1 \leq b \leq B$, be the b^{th} bootstrap sample. Let $X_1^*, X_2^*, \ldots, X_B^*$ be the B bootstrap samples. The standardized bootstrap sample mean can be calculated as:

$$t_b^* = \frac{\overline{X}_b^* - \overline{X}_0}{\hat{s}^{*(b)}}$$

where the original sample mean $\left(\overline{X}_0\right)$ is computed by $\overline{X}_0 = \frac{1}{n}\sum_{i=1}^{n} X_i$, *i*=1, 2, …,*n,*

The original sample standard deviation $\left(S_0\right)$ is computed by $S_0 = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}_0\right)^2}{n-1}}$

The bootstrap sample mean $\left(\overline{X}_b^*\right)$ is computed by $\overline{X}_b^* = \frac{1}{m}\sum_{j=1}^{m} X_{jb}^*$

The bootstrap sample standard deviation $\left(S_b^*\right)$ of the b^{th} bootstrap sample is computed by $S_b^* = \sqrt{\frac{\sum_{j=1}^{m}\left(X_j^* - \overline{X}_b^*\right)^2}{m-1}}$, where *j*=1, 2, …, *m* and b =1, 2, …, B. The estimated standard error of $\overline{X}_b^*$ is $\hat{s}^{*(b)}$.

When resampling using the independent bootstrap method, $\hat{s}^{*(b)} = \frac{S_b^*}{\sqrt{m}}$, where as in the dependent bootstrap method, $\hat{s}^{*(b)} = S_b^* \sqrt{\frac{kn-1}{m(kn-m)}}$. Thus, the Bootstrap-t $(1-\alpha)$ 100% confidence interval for a population mean is

$$\left(\overline{X}_0 - t^*_{(1-\frac{\alpha}{2})} \frac{S_0}{\sqrt{n}}, \overline{X}_0 - t^*_{(\frac{\alpha}{2})} \frac{S_0}{\sqrt{n}} \right)$$

where t_r^* is the *rB*th ordered value in the list of the *B* standardized bootstrap sample means, $0 < r < 1$ by Smith and Taylor [2].

2.4 The Percentile confidence interval method

The Percentile $(1-\alpha)$ 100% confidence interval for a population mean is

$$\left(\overline{X}^*_{(\frac{\alpha}{2})}, \overline{X}^*_{(1-\frac{\alpha}{2})} \right)$$

where $\overline{X}_r^*$ is the *rB*th ordered value on the list of the *B* bootstrap sample means, $0 < r < 1$ by Efron and Tibshirani [3].

2.5 The Modified Percentile confidence interval method

The Modified Percentile $(1-\alpha)$ 100% confidence interval for a population mean is defined by Smith and Taylor [3] as:

$$\left(\overline{X}^*_{(\frac{p^*}{2})}, \overline{X}^*_{(1-\frac{p^*}{2})} \right)$$

where $\overline{X}_r^*$ is the *rB*th ordered value on the list of the *B* bootstrap sample means, and $0 < r < 1$. $\frac{p^*}{2}$ is found such that $1 - \Phi\left(Z_{\frac{p^*}{2}}\right) = \frac{p^*}{2}$

$Z_{\frac{p^*}{2}} = \theta\sqrt{\frac{kn-1}{kn-m}} Z_{\frac{\alpha}{2}} + (1-\theta) Z_{\frac{\alpha}{2}}$, for $0 < \theta < 1$. $Z_{\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})^{th}$ percentile of the standard normal. And *k* is the number of copies from original data without replacement.

In the simulations of this study, 1,500 samples of size *n*=20, 40, and 100 were generated from three distributions, each having the coefficient of variation (CV) $=1/\sqrt{2}$: Normal with mean 4 and variance 8, Beta with α =1.5 and β =7.5, Gamma with α =2 and β =2. For each original sample with size *n*, the traditional normal theory 90% confidence interval for a population mean was calculated, and the coverage probability and the length of the 1,500 intervals was computed.

Next, for each original sample, the 90% dependent bootstrap confidence interval for a population mean was created by drawing 2,000 dependent bootstrap samples of size *m*=*n,* and *k*, the replication factors, was equal to 2, 4, 6, 8, 10, 12, 20, and 30. From these 2,000 dependent bootstrap samples, the Bootstrap-t, Percentile and Modified Percentile confidence intervals for a population mean were then obtained. The same procedure was performed using the independent bootstrap method.

3. Results

The results of this study were categorized according to the original samples distribution. For each distribution, the estimated coverage probabilities and average interval lengths are presented in Tables 1-6.

Table 1. Coverage probabilities of mean for Normal distribution at 90% confidence level.

Confidence interval method	*n*	Independent Bootstrap	Dependent Bootstrap							
			k=2	*k*=4	*k*=6	*k*=8	*k*=10	*k*=12	*k*=20	*k*=30
Bootstrap-t	20	0.89600	0.60200	0.79067	0.83400	0.85733	0.86600	0.87200	0.88200	0.88467
	40	0.90067	0.59800	0.79733	0.84133	0.86533	0.87067	0.87867	0.89067	0.89200
	100	0.89667	0.60400	0.79267	0.83600	0.84667	0.85733	0.86867	0.88067	0.88667
Percentile	20	0.86800	0.73800	0.82600	0.84267	0.85467	0.85667	0.86000	0.86400	0.86600
	40	0.88867	0.75800	0.83867	0.86267	0.86933	0.87733	0.87867	0.88267	0.88800
	100	0.89533	0.76867	0.84067	0.85733	0.87133	0.87467	0.87667	0.88067	0.88733
Modified Percentile	20	0.86800	0.85333	0.86533	0.86933	0.87200	0.87400	0.87133	0.87200	0.87067
	40	0.88867	0.87667	0.88667	0.89000	0.88800	0.89067	0.88867	0.89133	0.89200
	100	0.89533	0.87533	0.88267	0.89133	0.89000	0.88933	0.89400	0.89133	0.89400

Table 2. Average interval lengths of mean for Normal distribution at 90% confidence level.

Confidence interval method	*n*	Independent Bootstrap	Dependent Bootstrap							
			k=2	*k*=4	*k*=6	*k*=8	*k*=10	*k*=12	*k*=20	*k*=30
Bootstrap-t	20	2.16833	1.08966	1.62955	1.80668	1.90028	1.95198	1.98844	2.05869	2.09646
	40	1.49083	0.74842	1.12071	1.24374	1.30511	1.34145	1.3675	1.41767	1.4426
	100	0.93386	0.46760	0.70066	0.77777	0.81664	0.84027	0.85519	0.88673	0.90152
Percentile	20	1.99019	1.42771	1.73577	1.82286	1.86882	1.89373	1.90982	1.94068	1.95847
	40	1.43393	1.02086	1.24574	1.31136	1.34372	1.36049	1.37413	1.39914	1.41057
	100	0.92009	0.65190	0.79707	0.83969	0.86072	0.87289	0.88053	0.89635	0.90324
Modified Percentile	20	1.99019	1.90236	1.95602	1.97046	1.98018	1.98141	1.98597	1.98731	1.99296
	40	1.43393	1.36922	1.41041	1.42272	1.42385	1.42845	1.4291	1.43718	1.43571
	100	0.92009	0.88372	0.90338	0.91152	0.91551	0.91654	0.91904	0.92093	0.91947

Table 3. Coverage probabilities of mean for Beta distribution at 90% confidence level.

Confidence interval method	n	Independent Bootstrap	Dependent Bootstrap							
			k=2	k=4	k=6	k=8	k=10	k=12	k=20	k=30
Bootstrap-t	20	0.89267	0.59467	0.79333	0.83867	0.85800	0.86533	0.86933	0.88133	0.88333
	40	0.89733	0.57800	0.78200	0.81933	0.84133	0.85133	0.85867	0.86933	0.88000
	100	0.90267	0.58333	0.78467	0.83133	0.85067	0.86267	0.87400	0.88800	0.89667
Percentile	20	0.86533	0.72533	0.81533	0.83600	0.84333	0.84867	0.85267	0.85733	0.86000
	40	0.87467	0.73000	0.81733	0.84333	0.85333	0.85667	0.86200	0.86400	0.86933
	100	0.90000	0.74800	0.83733	0.86133	0.87467	0.88133	0.88067	0.88867	0.88933
Modified Percentile	20	0.86533	0.84600	0.85733	0.86000	0.86067	0.86000	0.86267	0.86467	0.86200
	40	0.87467	0.85733	0.86733	0.87467	0.86800	0.87467	0.87200	0.87267	0.87333
	100	0.90000	0.88067	0.88667	0.89400	0.89533	0.89667	0.89600	0.89667	0.89733

Table 4. Average interval lengths of mean for Beta distribution at 90% confidence level.

Confidence interval method	n	Independent Bootstrap	Dependent Bootstrap							
			k=2	k=4	k=6	k=8	k=10	k=12	k=20	k=30
Bootstrap-t	20	0.09342	0.04652	0.06996	0.07780	0.08166	0.08404	0.08566	0.08876	0.09048
	40	0.06358	0.03177	0.04766	0.05294	0.05570	0.05720	0.05833	0.06036	0.06146
	100	0.03914	0.01957	0.02934	0.03260	0.03427	0.03521	0.03586	0.03717	0.03784
Percentile	20	0.08269	0.05942	0.07218	0.07592	0.07767	0.07873	0.07943	0.08076	0.08148
	40	0.06009	0.04281	0.05223	0.05496	0.05636	0.05708	0.05766	0.05856	0.05908
	100	0.03829	0.02714	0.03319	0.03496	0.03585	0.03634	0.03665	0.03730	0.03765
Modified Percentile	20	0.08269	0.07884	0.08125	0.08199	0.08229	0.08236	0.08257	0.08268	0.08294
	40	0.06009	0.05732	0.05913	0.05964	0.05973	0.05989	0.05996	0.06015	0.06014
	100	0.03829	0.03676	0.03762	0.03795	0.03813	0.03816	0.03824	0.03831	0.03832

Table 5. Coverage probabilities of mean for Gamma distribution at 90% confidence level.

Confidence interval method	*n*	Independent Bootstrap	Dependent Bootstrap							
			k=2	*k*=4	*k*=6	*k*=8	*k*=10	*k*=12	*k*=20	*k*=30
Bootstrap-t	20	0.89533	0.61330	0.7866	0.83200	0.85200	0.86460	0.86530	0.88200	0.88460
	40	0.89200	0.61400	0.79000	0.82667	0.84533	0.85600	0.86000	0.87600	0.88133
	100	0.89667	0.59200	0.78533	0.83133	0.84467	0.85467	0.85933	0.87867	0.88867
Percentile	20	0.85730	0.72000	0.80270	0.82330	0.83270	0.84070	0.84330	0.84870	0.85000
	40	0.86467	0.74667	0.82533	0.84133	0.84933	0.85200	0.85200	0.86200	0.86333
	100	0.88600	0.75267	0.83733	0.85400	0.86267	0.86867	0.87133	0.87800	0.88000
Modified Percentile	20	0.85733	0.83600	0.84733	0.85333	0.85467	0.85200	0.85133	0.85733	0.85533
	40	0.86467	0.85000	0.85933	0.86333	0.86200	0.86400	0.86533	0.86867	0.86533
	100	0.88600	0.87000	0.88000	0.88267	0.88667	0.88733	0.88267	0.88667	0.88533

Table 6. Average interval lengths of mean for Gamma distribution at 90% confidence level.

Confidence interval method	*n*	Independent Bootstrap	Dependent Bootstrap							
			k=2	*k*=4	*k*=6	*k*=8	*k*=10	*k*=12	*k*=20	*k*=30
Bootstrap-t	20	2.26430	1.13110	1.69560	1.88530	1.98290	2.03720	2.07630	2.15060	2.18860
	40	1.53155	0.76647	1.14798	1.27454	1.34117	1.37881	1.40299	1.45509	1.47939
	100	0.94613	0.47342	0.7093	0.78771	0.82821	0.85211	0.86721	0.89961	0.91436
Percentile	20	1.95344	1.40421	1.70417	1.79186	1.83418	1.85940	1.87676	1.90561	1.92211
	40	1.42657	1.01807	1.24011	1.30506	1.33882	1.35615	1.36711	1.39126	1.40297
	100	0.91913	0.65244	0.79713	0.83963	0.86068	0.87336	0.88060	0.89703	0.90389
Modified Percentile	20	1.95344	1.85289	1.91753	1.93476	1.94305	1.94412	1.95056	1.95096	1.95644
	40	1.42657	1.35901	1.40370	1.41557	1.41873	1.42348	1.42188	1.42908	1.42785
	100	0.91913	0.88283	0.90351	0.91140	0.91520	0.91732	0.91885	0.92125	0.92009

4. Conclusions

The simulation results of each original sample distribution agreed with one another. This dues to the non-parametric nature of the independent and dependent bootstrap procedures. Therefore no assumption is necessary for the population distribution.

The Modified Percentile confidence interval method with θ=0.85 of dependent bootstrap procedure gives higher coverage probabilities than other methods and are close to the confidence coefficient 0.90 for all distributions and sample size *n*.

For both the independent and dependent bootstrap confidence intervals, the coverage probabilities increase and are close to the confidence coefficient 0.90. The average interval lengths decrease as the sample size *n* increase for all distributions and methods.

Additionally, the coverage probabilities of the dependent bootstrap confidence intervals vary with replication factors or *k* copies. When *k* is large, the coverage probabilities of the dependent bootstrap confidence intervals are similar to the independent bootstrap confidence intervals for all distributions. This is because as *k* approaches infinity, the dependent bootstrap, which drawing sample of size *m* without replacement from the collection of *nk* items, reduces to the independent bootstrap with drawing sample of size *m* with replacement from *n* items. Moreover, the average interval lengths of the dependent bootstrap confidence intervals with three methods differ only at the first or the third decimal place. And the dependent bootstrap confidence intervals give shortest average interval lengths for all distributions.

References

[1] Efron, B. Bootstrap methods: Another look at the jackknife. Ann. Stat. 1979;7:1-26.

[2] Smith, W.D. and Taylor, R.L. Dependent bootstrap confidence intervals. Selected Proceedings of the Symposium on Inference for Stochastic Processes (Athens, GA, 2000). IMS Lecture Notes -Monograph Series. 2001;37: 91-107.

[3] Efron, B. and Tibshirani, R.J. An introduction to the bootstrap. Chapman & Hall United States of America. 1993.