SILESIAN UNIVERSITY OF TECHNOLOGY PUBLISHING HOUSE

SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY 2021
ORGANIZATION AND MANAGEMENT SERIES NO. 153

COMPARISON OF JACKKNIFE AND BOOTSTRAP METHODS IN ESTIMATING CONFIDENCE INTERVALS

Łukasz SROKA

University of Economics, Katowice; lukasz.sroka@edu.uekat.pl, ORCID: 0000-0001-5721-2475

Purpose: The development of technology has allowed creating and using the new, more complex computational tools in static and econometrics in recent years. Since then, resampling methods has become more popular techniques in estimating statistics from small samples. The aim of the article is to present and to compare the bootstrap and the jackknife methods in estimation of interested statistics with explaining and illustrating the usefulness and limitation in the context of using in econometric.
Design/methodology/approach: To compare and present the methods, data of the length of bicycle paths divided into 371 polish counties from 2019 was received from Local Data Bank. From the data three samples were randomly selected and used as bootstrap and jackknife samples. Using the bootstrap and the jackknife simulations confidential intervals of the searching statistics with standard error were calculated. Results obtained for the methods were compared and described.
Research limitations/implications: An analysis of these methods will allow improving the efficiency and reducing the error in estimating confidence intervals for searching statistics.
Findings: As presented in the article, both the methods can be used to estimate mean, however, slightly better results are provided by the bootstrap. Furthermore, confidence intervals for confidence level at 95% created by these methods cover the population mean for each sample randomly selected from the population. To estimate standard deviation the better option is to choose the bootstrap method. Although, both confidence intervals for confidence at level 95% cover the population standard deviation, the bootstrap methods perform more accurate results with a smaller standard deviation.
Originality/value: It was proven that the bootstrap method is slightly better in estimation confidence intervals based on the skewed data in comparison with the jackknife method.

Keywords: jackknife, bootstrap, confidential intervals, resampling, simulations.

Category of the paper: research paper.

1. Introduction

The development of the technologies made it possible to use new, more powerful statistic tools in ways previously inconceivable. The new methods allow scientists do more realistic and accurate analysis. The bootstrap and the jackknife methods are parts of the technological revolution in statistics. These two methods can help to identify quantify uncertainty by calculating standard errors, confidence intervals and performing significance tests. The main advantage of the bootstrap and jackknife is that they require fewer assumptions than traditional methods and generated more accurate results (Hasenberg et al., 2003).

The bootstrap and the jackknife have been the object of statistic research since they were introduced. The results of the research were presented in the subject literature by B. Efron and C. Stein (1981), Beran and Ducharme (1991), Efron and Tibshirani (1993) or McIntosh (2016).

The aim of the article is to present and compare the bootstrap and the jackknife methods in estimation of interested statistics with explaining and illustrating the usefulness and limitation in the context of using econometric. The discussion in this article provides a mathematically detailed theory of the bootstrap and the jackknife with comparison these two methods using real data set.

The article is divided into two main parts – theoretical and empirical. The first theoretical part contains presentation of the bootstrap and the jackknife methods with their differences. The second empirical part presents comparison of the results of the usage of these two methods in estimation two statistics: mean and standard deviation from the samples of the populations. The conclusions of the empirical part are presented at the end of the article.

2. The bootstrap and the jackknife as simulation methods

2.1. The bootstrap method

The bootstrap as a method of simulation was presented in 1979 by B. Efron. It is widely use in estimating of the confidence intervals, approximation of estimator distributions or tests of the statistics. Using this simulation allows, among other things for obtaining the evaluation of the estimators variance or creating confidence intervals for population parameters. It also can be us to develop new statistics tests and estimation procedures. (Kończak, 2012, p. 108). The main advantage of the method is possibilities of statistical inference without knowing of the whole population. The bootstrap method prepares its estimates, only a sample from the observed population (Dunaj, 2017, p. 6).

The bootstrap distribution is obtained by estimation of independent samples created by sampling with replacement from the original dataset. Let F denote the distribution of an individual observation E. Let $G_n(u, F)$ denote the distribution of the estimator $\breve{\theta}$ (Hansen, 2021). That is:

$$G_n(u, F) = Pr(\breve{\theta} \leq u \mid F) \quad (1)$$

The G_n distribution is written as a function of n and F since they influence on the distribution of $\breve{\theta}$ If the F distribution is known, determination of the distribution G_n is not the issue. In the real data there are two main obstacles which do not allow us to determine the F distribution:

- The calculation of $G_n(u, F)$ is impossible, except some exceptions such as a normal regression model.
- The distribution of the individual observations from F population is not known (Hansen, 2001).

The bootstrap simulation is able to omit one of the obstacles by using empirical distribution function (EDF) F_n to estimate F. EDF is the simplest nonparametric estimator of the joint distribution of the observations. When F is replaced by F_n in $G_n(u, F)$ it is possible to receive the bootstrap estimator of the distribution of $\breve{\theta}$:

$$G_n^*(u) = G_n(u, F_n) \quad (2)$$

$G_n^*(u)$ is estimated by simulation. The simulation from F_n is sampling with replacement from the original data. Applying the estimation formula for T_n it is possible to receive draws from the distribution $G_n^*(u)$. By making a large number of the draws any feature of G_n^* can be estimated (Hansen, 2021, pp. 263-270).

The procedure of bootstrap is as follows:

1. Resample – creating new samples, called bootstrap samples by sampling with replacement from the original random sample. Each sample has the same size as the original random sample.
2. Calculating a bootstrap distribution – calculation the statistics for each obtained sample. The distribution of these samples is called the bootstrap distribution.
3. Using the bootstrap distribution – the received distribution from the bootstrap samples gives information about shape, center and spread of the samples distribution of the statistics (Hasenberg et al., 2003).

Let estimate the parameter θ from population. To receive the estimation of the parameter estimator $\breve{\theta}$ can be used. It is important to obtain an estimator of variance for the estimator as well. From the randomly selected n-element sample, bootstrap samples are drawn by N fold sampling with replacement. The subsequent bootstrap samples are labelled as presented below (Kończak, 2012, pp. 110-111):

$(\dot{x}_1^{(1)}, \dot{x}_2^{(1)}, \ldots, \dot{x}_n^{(1)})$ – the first bootstrap sample,
$(\dot{x}_1^{(2)}, \dot{x}_2^{(2)}, \ldots, \dot{x}_n^{(2)})$ – the second bootstrap sample,
…
$(\dot{x}_1^{(N)}, \dot{x}_2^{(N)}, \ldots, \dot{x}_n^{(1N)})$ – the N bootstrap sample.

For each bootstrap sample the value of $\breve{\theta}^{(i)}$ statistic (i = 1, 2, …, N) is obtained by using the same formula as for $\breve{\theta}$ statistic. The value of the statistic is determined on the basis of the particular bootstrap sample. In the next step, the value of bootstrap estimator is calculated as following:

$$\breve{\theta}_B = \frac{1}{N}\sum_{i=1}^{N} \breve{\theta}^{(i)} \tag{3}$$

And the variation of the estimation is calculated as presented below (Efron, Tibshirani, 1993):

$$\breve{V}_1(\breve{\theta}) = \frac{1}{N-1}\sum_{i=1}^{N} \left(\breve{\theta}^{(i)} - \breve{\theta}_B\right)^2 \tag{4}$$

The bootstrap method allows determining the confidence intervals for unknown value of parameter θ. The confidence intervals for the unknown parameter can be designated by limit theorem:

$$(\breve{\theta} - t^{*\left(1-\frac{\alpha}{2}\right)} D(\breve{\theta}); \breve{\theta} - t^{*(\alpha/2)} D(\breve{\theta})) \tag{5}$$

Where $t^{*\left(1-\frac{\alpha}{2}\right)}$ and $t^{*(\alpha/2)}$ are the percentiles of the empirical distribution for order $1-\frac{\alpha}{2}$ and $\frac{\alpha}{2}$ of statistic $t* = \frac{\breve{\theta}^* - \breve{\theta}}{\widetilde{D(\breve{\theta}^*)}}$.

In this statistic $D(\breve{\theta})$ is the standard deviation of $\breve{\theta}$ estimator and $\widetilde{D(\breve{\theta}^*)}$is the estimation of the estimator.

2.2. The jackknife method

The jackknife as a method was described in 1949 by M.H. Quenouille. At the beginning, the method was used as a procedure for correcting bias. In 1956 J. Turkey adapted the jackknife to construct a confidence limit for a large class of estimator. The method is similar to the bootstrap, however the main difference is that the jackknife estimator of a parameter is found by systematically leaving out each observation from a dataset and calculating the estimate and then finding the average of these calculations (McIntosh, 2016).

Let Y_i,…,Y_n denote a sample of independent and identically distributed random variables, $\breve{\theta}$ denotes an estimator of the parameterθ based on a sample of *n*. Let $\breve{\theta}^{(-i)}$ be the corresponding estimator based on the sample size *(g-1)/h,* where the *i-th* group of size h has been excluded (Miller, 1974).

$$\widetilde{\theta}_i = g\breve{\theta} - (g-1)\breve{\theta}^{(-i)} \tag{6}$$

where i = (1, 2, …, g). The estimator:

$$\breve{\theta} = \frac{1}{g}\sum_{i=1}^{q} \breve{\theta}^i = g\breve{\theta} - (g-1)\frac{1}{g}\sum_{i=1}^{g} \breve{\theta}^{(-i)} \tag{7}$$

excludes the order *1/n* term from a bias of the form

$$E(\breve{\theta}) = \theta + a_{1/n} + 0(\frac{1}{n^2}) \tag{8}$$

Turkey (1958) presented g values as widely used, approximately, independent and identically distributed random variables. The statistic:

$$\frac{\sqrt{g(\breve{\theta} - \theta)}}{(\frac{1}{1-g}\sum_{i=1}^{g}(\breve{\theta}_i - \breve{\theta}_i)^2)^2} \tag{9}$$

has an approximate *t* distribution with *g-1* degrees of freedom and constitutes a pivotal statistic for proper interval estimation.

In this article, the jackknife method with excluding one observation from the data set is described, however, in literature occurs the other method with removing more than only one observation (Kamiński, 2010).

The procedure of jackknife with excluding one observation is simpler than bootstrap procedure:

Excluding – from the data set one observation is omitted sequentially. The new data set with the excluded observation is called a jackknife sample.

Calculating the statistic – the interested statistic is calculated from each jackknife sample.

Finding average value – of all statistics from the second step the average value is calculated. The value is the jackknife estimation of the interested statistic.

The jackknife samples can be described as presented below:

$(\dot{x}_2^{(1)}, \dot{x}_3^{(1)}, \ldots, \dot{x}_n^{(1)})$ – the first jackknife sample,

$(\dot{x}_1^{(2)}, \dot{x}_3^{(2)}, \ldots, \dot{x}_n^{(2)})$ – the second jackknife sample,

…

$(\dot{x}_1^{(N)}, \dot{x}_2^{(N)}, \ldots, \dot{x}_{n-1}^{(1N)})$ – the N jackknife sample.

An example of using jackknife method in estimating median and variance is described as the following: let $\breve{\theta}$ denotes any estimator of a vector-valued parameter θ which is a function of a random sample size n. Let $V(\breve{\theta})$=var($\breve{\theta}$) be a variation of θ and $\breve{\theta}_{(-i)}$ denotes the leave-one-out estimators which are computed using the formula for $\breve{\theta}$ except the deleted observation *i*. Turkey's jackknife estimator for $V_{\breve{\theta}}$ is described as a scale of the sample variance of the leave-one-out estimators and presented below (Hansen, 2001):

$$V(\breve{\theta}) = \frac{N-1}{N}\sum_{i=1}^{N}(\breve{\theta}^{(-i)} - \bar{\theta})^2 \tag{10}$$

where $\bar{\theta}$ is the sample mean of the leave-one-out estimator:

$$\bar{\theta} = \frac{1}{N}\sum_{i=1}^{N}\breve{\theta}^{(-i)} \tag{11}$$

In the jackknife method the confidence intervals are determined in the same way as in the bootstrap method.

2.3. Differences between the bootstrap and the jackknife

There are a few differences between the bootstrap and the jackknife methods. The jackknife is an older method which is less computationally expensive, while the bootstrap is more computationally expensive but gives more precision with estimation of the parameters. In addition, bootstrap is conceptually simpler and produces less standard error than the jackknife. One of the advantages of the jackknife is that the method performs the same results every time, while bootstrap gives different results in every run. Therefore there are some conditions where each of the methods can be used. The jackknife is better for estimation of the confidence interval for pairwise agreement measures and it is more suitable for small data, however bootstrap achieves better results for data with skewed distribution (Nguyenova, 2020).

3. Results

To prepare the comparison between the bootstrap and the jackknife in estimation of the interested statistics (mean and standard deviation) and determination of the confidence intervals real data of length of bicycle paths from polish counties were obtained. From the population three samples were randomly selected. The first sample contains 25% observations from the population, the second 50% and the third 75%. In addition, 5000 bootstrap samples were created from each of the selected samples to prepare bootstrap estimations. The results of these two methods were described and compared to select the best method to estimate interested statistics.

As presented in the previous part of the article to prepare the bootstrap and the jackknife simulations a sample data from the population is needed. Because one of the aims of the article is comparing results of estimation of these two methods real data of length of bicycle paths from polish counties were obtained. The data contains information about the length of bicycle paths divided into 371 polish counties in 2019. The set was received from Local Data Bank. During analysis one outlier was detected and removed from the analysis. The outlying concern county of the capital city of Warsaw was removed. The county is an anomaly in the data – over two times higher than the next highest value.

Figure 1 presents the distribution of the bicycle paths lengths while the table 1 statistics for the data.

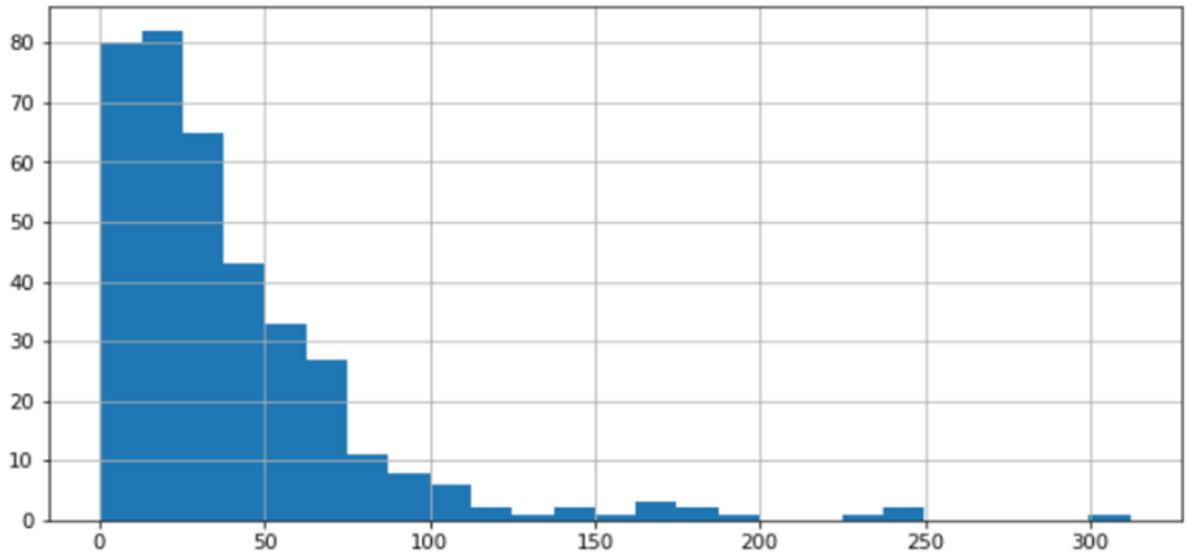

Figure 1. Distribution of the population of the bicycle paths lengths Source: Author's own elaboration based on: Local Data Bank.

Table 1.
Statistics of the population of the bicycle paths lengths

Statistic	**Result (in meters)**
Mean	40,15
Standard deviation	40,14
Min	0,20
Max	312,10
Percentile 0,25	15,00
Percentile 0,50	29,60
Percentile 0,75	312,10

Source: own calculations based on Local Data Bank.

As presented in figure 1 and table 1 the population of the length of the bicycle path has the right-hand distribution with the mean equal to 40,15 meters and the standard deviation of 40,14 meters. The population contains a lot of outliers which is very common in real data sets. The huge difference between min and max values is related to the fact that not all polish counties have an extensive network of bicycle paths.

Table 2 presents the statistics of the three randomly selected samples from the population. As described in the methodology, the samples are used to create the bootstrap and the jackknife simulations and are necessary to estimate the interested statistics. Each of the samples has a different number of observations. The first has 25% observations of the total population (n = 93), the second has 50% observations of the total population (n = 186), and the third has 75% observations of the total population (n = 279).

Table 2.
The statistics for the samples of the bicycle paths lengths sample size

Statistic	Result for 25% sample n = 93	Result for 50% sample n = 186	Result for 75% sample n = 279
Mean	46,23	43,38	38,12
Standard deviation	49,23	42,10	36,70
Min	2,20	0,20	0,20
Max	248.50	312,10	312,10
Percentile 0,25	12,90	17,23	14,85
Percentile 0,50	33,30	32,45	29,40
Percentile 0,75	56,30	55,38	53,20

Source: own calculations based on Local Data Bank.

Below, the tables from 3 to 8 contain the results obtained using the bootstrap and the jackknife methods. The point estimations of mean and standard deviation are presented in tables 3, 5 and 7. Also standard error of the estimators was included in these tables. Confidence intervals for confidence level 95% were described in the tables 4, 6 and 8.

Table 3.
The results for 25% sample of the population for the bootstrap and the jackknife simulations

Statistic	Bootstrap		Jackknife	
	Estimation	Standard error of the estimation	Estimation	Standard error of the estimation
Mean	44,98	5,68	44,96	5,78
Standard deviation	54,43	9,06	56,54	9,56

Source: own calculations based on Local Data Bank.

Table 4.
The confidential intervals for 25% sample of the population n for the bootstrap and the jackknife simulations

Statistic	Bootstrap		Jackknife	
	2,5%	97,5%	2,5%	97,5%
Mean	34,35	56,86	33,64	56,28
Standard deviation	35,86	71,04	37,80	75,28

Source: own study based on Local Data Bank.

The results provided by the two methods are very similar, however, the jackknife makes slightly better estimation for the population mean, and the bootstrap estimates the population standard deviation closer to population result. Additionally, a lower level of standard error for the both estimations occurs in the bootstrap. It can be assumed that the means and standard deviations calculated based on bootstrap samples have a lower variance, therefore, the estimations are more stable. Both the methods cover population mean and standard deviation by their confidence intervals for confidence level equal to 95%.

Table 5.
The confidential intervals for 50% sample of the population for the bootstrap and the jackknife simulations

Statistic	Bootstrap		Jackknife	
	Estimation	Standard error of the estimation	Estimation	Standard error of the estimation
Mean	43,37	3,09	43,38	3,09
Standard deviation	41,40	5,95	42,55	6,24

Source: own study based on Local Data Bank.

Table 6.
The confidential intervals for 25% sample of the population for the bootstrap and the jackknife simulations

Statistic	Bootstrap		Jackknife	
	2,5%	97,5%	2,5%	97,5%
Mean	37,67	49,75	37,33	49,42
Standard deviation	30,21	53,07	30,32	54,79

Source: own study based on Local Data Bank.

Also for the second sample both methods estimate mean almost the same. In addition, standard error of the estimation is the same for the bootstrap and the jackknife methods. The main difference occurs in estimation of standard deviation. The methods overestimated population standard deviation, however, results provided by bootstrap are closer to the real standard deviation than the jackknife propose. Confidence intervals created by the two methods cover the population mean and standard deviation at the 95% level of confidence.

Table 7.
The results for 75% sample of the population for the bootstrap and the jackknife simulations

Statistic	Bootstrap		Jackknife	
	Estimation	Standard error of the estimation	Estimation	Standard error of the estimation
Mean	38,16	2,20	38,12	2,20
Standard deviation	36,30	4,46	37,10	4,77

Source: own study based on Local Data Bank.

Table 8.
The confidential intervals for 75% sample of the population for the bootstrap and the jackknife simulations

Statistic	Bootstrap		Jackknife	
	2,5%	97,5%	2,5%	97,5%
Mean	33,99	42,60	33,81	42,42
Standard deviation	28,23	45,56	27,65	46,36

Source: own study based on Local Data Bank.

The same as for the previous results estimation of population mean provided by the bootstrap and the jackknife are very similar with standard error at the same level. As for estimation of standard deviation both of the methods underestimated the population standard deviation, however, the jackknife estimation is closer to the population statistic. In both cases

confidence intervals at 95% level of confidence cover the population standard deviation, the same as population mean.

4. Conclusions

In this article the bootstrap and the jackknife methods were discussed. Both the methods are very useful tools in statistics analyses. The main advantage of these techniques is the possibility of using them when the underlying distribution for the population is not known and the traditional formulas are difficult or impossible to apply. The paper focuses on describing both methods and comparing them in estimating selected statistics.

The theoretical part of the article presents the consideration about the two methods. The procedure of creating the bootstrap and the jackknife simulations were described. Also the use of these methods to estimate interested statistics with the standard error was presented. In the empirical part of the paper the bootstrap and the jackknife methods were the tools to estimate mean and standard deviation of the length of the bicycle paths.

As presented in the article, both the methods can be used to estimate mean, however, slightly better results are provided by the bootstrap. Furthermore, confidence intervals for confidence level at 95% created by these methods cover the population mean for each sample randomly selected from the population.

To estimate standard deviation the better option is to choose the bootstrap method. Although, both confidence intervals for confidence at level 95% cover the population standard deviation, the bootstrap methods perform more accurate results with a smaller standard deviation.

It has been proven that both methods are suitable for estimating the mean and standard deviation with the use of samples with different numbers of observations. Nevertheless, the solutions that will allow for the reduction of standard errors and the narrowing of the confidence intervals of the interested statistics should still be sought.

References

1. Beran, R., and Ducharme, G.R. (1991). *Asymptotic Theory for Bootstrap Methods in Statistics*. Montreal: Les Publications CRM.
2. Dunaj, J. (2017). *Bootstrap i jego zastosowania do analizy wrażliwości estymatora wartości zagrożonej,* https://ftims.pg.edu.pl/katedra-analizy-nieliniowej-i-statystyki/prace-dyplomowe.

3. Efron, B., Stein, C. (1981). *The jackknife in estimate of variance. The analyst of statistics Vol. 9, Iss. 3,* pp. 586-596.
4. Efron, B., Tibshirani, J.R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
5. Hansen, E.B. (2001). *Econometrics*. Wisconsin: University of Wisconsin, Department of Economics, pp. 253-270.
6. Hasenberg, T., Monaghan, S., Moore, S.D., Clipson, A., Epstein, R. (2003*). Bootstrap Methods and Permutation tests*. New York: W.H. Freeman and Company, pp. 11-13.
7. Kamiński, A. (2010). *Wykorzystanie algorytmów Bootstrap i Jacknife w estymacji parametrów regresji,* http://docplayer.pl/36440257-Wykorzystanie-algorytmow-bootstrap-i-jacknife-w-estymacji-parametrow-regresji.html.
8. Kończak, G. (2012) *Wprowadzenie do symulacji komputerowych*. Katowice: Wydawnictwo Uniwersytetu Ekonomicznego.
9. McIntosh, A. (2016). *The Jackknife Estimation Method,* https://arxiv.org/abs/1606.00497.
10. Miller, G.R. (1974). *The Jackknife – A Review*. Biometrika, *Vol. 61, Iss. 1*, pp. 1-15.
11. Nguyenova, L. (2020). *Bootstrapping vs. jackknife,* https://medium.com/@ lymielynn/ bootstrapping-vs-jackknife-d5172965207b.