A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 3(302), 2014

Grzegorz Kończak[*]

ON THE RESAMPLING METHOD IN SAMPLE MEDIAN ESTIMATION

Abstract. Bootstrap is one of the resampling statistical methods. This method was proposed by B. Efron. The main idea of bootstrap is to treat the original sample of values as a stand-in for the population and to resample with replacement from it repeatedly. Bootstrap allows estimation of the sampling distribution of almost any statistics using only very simple methods. This paper presents a modification of a resampling procedure based on bootstrap sampling. The proposal leads to sampling from population with density function f(x), where f(x) is estimated based on the kernel estimation. The properties of the method were analyzed in the median estimation in Monte Carlo study.

The proposal could be useful for the parameters estimation in the case of a small sample. This method could be used in quality control procedures such as control charts or in the acceptance sampling.

Keywords: bootstrap, kernel estimation, small sample.

I. INTRODUCTION

Bootstrap was introduced by Bradley Efron in 1979 (B. Efron, 1979). The main idea of bootstrap is to treat the original sample of values as a stand-in for the population and to resample with replacement from it repeatedly. Bootstrap allows estimation of the sampling distribution of almost any statistic using only very simple methods. In the case of small samples the resampling is based only on few elements. K. Pruska (2007) uses bootstrap and jackknife methods for the estimation bias and variance of a sample median and concludes that these methods do not give good results. Another look at the resampling method is presented in the paper. This proposal of resampling is based on the kernel density estimation. It leads to sampling from the population with density function f(x), where f(x) is estimated based on the kernel estimation. This method could be used in monitoring the production processes especially of small samples. G.J. Janacek and S.E. Meikle (1997) proposed control charts based on

[*]Associate Professor, Department of Statistics, Katowice University of Economics, grzegorz.konczak@ue.katowice.pl.

a sample median. They suggested monitoring a median characteristic for non-normal random variables. The results of the proposed method have been compared with the bootstrap method in the Monte Carlo study.

II. CLASSICAL BOOTSTRAP PROCEDURE

Bootstrap resampling is usually used for hypothesis testing, variance estimation or construct the confidence intervals. This method is recommended for the following situations:

– theoretical distribution of a statistic is complicated or unknown,
– analytical calculation could not be used,
– the size of the sample is insufficient for statistical inference,
– small sample is available and power calculation is expected.

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n taken from the distribution F. and $x_1, x_2, \ldots, x_n$ be the realization of this sample. Let X_B be the random variable for which the probability distribution function has the following form (K. Pruska, 2007)

$$P(X_B = x_i) = \frac{1}{n} \text{ for } i = 1, 2, \ldots, n \tag{1}$$

The above distribution is called the bootstrap distribution. Let $T_n = T_n(X_1, X_2, ..., X_n)$ be an estimator of parameter θ of the population. The bootstrap estimation of the parameter θ leads to the generate n-element sequences of the pseudorandom numbers from the bootstrap distribution. Let N be a number of these sequences. The sequences are called bootstrap samples and can be denoted as $X^*_{1k}, X^*_{2k}, ..., X^*_{nk}$ ($k = 1, 2, \ldots, N$). The realization of the bootstrap sample can be denoted as $x^*_{1k}, x^*_{2k}, ..., x^*_{nk}$. The bootstrap estimator of parameter θ has the following form:

$$T_B = \frac{1}{N}\sum_{k=1}^{N} T^*_{nk} \tag{2}$$

where

$$T^*_{nk} = T_n(X^*_{1k}, X^*_{2k}, ..., X^*_{nk})$$

The standard error of estimator T_n could be estimated by (B. Efron, R. Tibshirani, 1993)

$$D(T_n) \approx \frac{1}{N-1}\sum_{k=1}^{N}(T_{nk}^{*} - T_B)^2 \tag{3}$$

The main idea of the bootstrap is sampling from the original sample. The modification based on the sampling from the estimated distribution will be considered.

III. KERNEL ESTIMATION

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n taken from the population with the unknown density function $f(x)$. Let us assume that $\int_{-\infty}^{\infty} |f(x)dx|^2 < \infty$. To estimate the density $f(x)$ the kernel method can be used. The kernel density estimator of $f(x)$ can be written as follows (S.J. Sheather, 2004, Cz. Domański, K. Pruska, 2000)

$$f_n(x;a_n) = \frac{a_n}{n}\sum_{i=1}^{n} K[a_n(x - X_i)] \tag{4}$$

where $(a_n)_{n\in IN}$ is a sequence of positive numbers diverging to infinity and such that $\lim_{n\to\infty} \frac{a_n}{n} = 0$ and $K(x)$ is a kernel function satisfying the following conditions:

$$\int_{-\infty}^{\infty} |K(x)|^2 \, dx < \infty$$

$$K(x) = K(-x) \text{ for } x \in (-\infty, \infty)$$

$$\int_{-\infty}^{\infty} K(x)dx = 1$$

$$\sup_{-\infty<x<\infty} |K(x) \le A < \infty$$

$$\int_{-\infty}^{\infty} x^i K(x)dx = 0 \text{ for i = 1, 2, …, } s - 1$$

$$\int_{-\infty}^{\infty} x^s K(x)dx \neq 0$$

$$\int_{-\infty}^{\infty} x^s K(x)dx < \infty$$

where s is a fixed natural number. Let the sequence a_n be given by the formula $a_n = const = \frac{1}{h}$, where $h > 0$ is the smoothing parameter for $n = 1, 2, \ldots$. The kernel estimation could be written as follows

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left[\frac{x - X_i}{h}\right] \tag{5}$$

Various functions $K(x)$ could be used as a kernel. Cz. Domański and K. Pruska (2000) write of gaussian kernels, triangual kernels, rectangual kernels, Cauchy kernels and Epanecznikow kernels. The form of the estimated kernel strongly depends on the form of the $K(x)$. The kernel estimation for the fixed $n = 5$ element samples is shown in Fig 1.

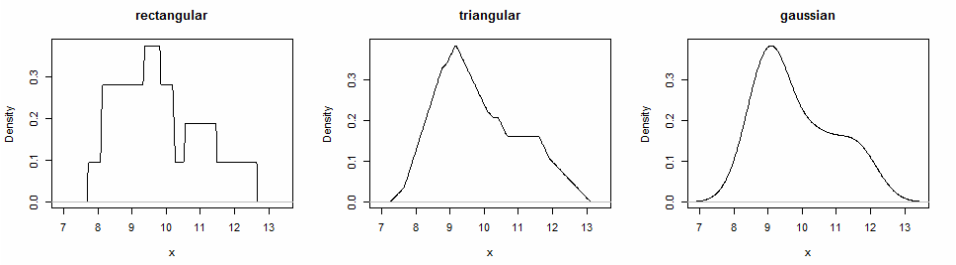

Figure 1. The kernel estimation for the fixed dataset ($n = 5$) for various kernel types (rectangular, triangular and gaussian)

The form of the estimator depends on the smoothing parameter (h). The kernel estimation for the fixed $n = 10$ element samples for various values of the smoothing parameters is shown in Fig 2.

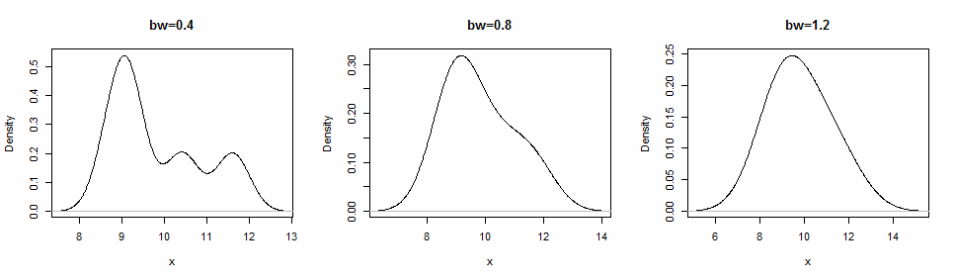

Figure 2. The kernel estimation for the fixed dataset (n = 10) for gaussian kernel and for various smoothing parameter (h = 0.4; 0.8 and 1.2)

IV. ANOTHER WAY OF RESAMPLING

Let us consider a sample of size n taken from the continuous distribution $F(x)$ with the density function $f(x)$. The main idea of the bootstrap is sampling from the bootstrap distribution given by (1). Let us consider another way of resampling than the bootstrap method. Let us consider the n element samples taken from distribution with density $f_n(x)$, where $f_n(x)$ is the kernel estimation given by (5) of the density function $f(x)$. The procedure of sampling from distribution with known density $f_n(x)$ is an indirect method following S. Ulam and J. von Neuman (Y. Rubinstein, D.P. Kroese, 2008; R. Wieczorkowski, R. Zieliński, 1997). It is called an acceptance-rejection method and it can be done using the following steps:

1. Generate random value x from uniform distribution on the [a, b], where [a, b] is the domain of $f(x)$.
2. Generate random value y from the uniform distribution on [0, c] where $c = \max_{x\in[a,b]} f_n(x; a_n)$
3. If $y \le f(x)$ then return $z = x$. Otherwise return to step 1.

The bootstrap is based on sampling from the discrete distribution. The proposal is a two step procedure

– estimating the density based on the sample
– sampling from the estimated density

The procedures will be compared in the sample median estimation.

V. MEDIAN ESTIMATION – COMPARING TWO METHODS

The above described method of resampling was compared to the bootstrap in the series of computer simulations. The problem of the median estimation was considered. Let n be the size of a population sample. The sample median is the

observation on the $(n+1)/2$ position for the odd n in a nondecreasing sequence and the average of two observations with numbers $n/2$ and $n/2+1$ for the even n. There were 4 theoretical populations considered in the Monte Carlo study

a) Normal N(10, 1)
b) Log-normal LN (0, 1)
c) Beta B(2, 2)
d) Beta B(0.2, 0.2)
e) Exponential E(1)
f) Uniform[0, 10]

Table 1. Parameters of distributions used in Monte Carlo experiments

Distribution Simulation parameters	Parameters	Mean	Median
Normal $N(\mu,\sigma)$	$\mu = 10,\ \sigma = 1$	10	$\tilde{1}$
Log-normal $LN(\mu,\sigma)$	$\mu = 0,\ \sigma = 1$	$\sqrt{e}$	1
Beta $B(\alpha,\beta)$	$\alpha = 10, \beta = 1$	0.5	0.5
Beta $B(\alpha, \beta)$	$\alpha = 10, \beta = 1$	0.5	0.5
Exponential $E(\lambda)$	$\lambda = 1$	1	ln2
Uniform U[a, b]	a = 0, b = 10	5	5

Source: Own preparation.

The details of the analyzed populations are described in table 1. The graphical view of densities of these random variables is presented in Fig. 2.

The study included the following steps:

1. The sample of size n (n = 5, 10, 20) was taken from the considered distributions.
2. The median was estimated using the bootstrap method and the above described proposed methods.
3. Steps 1 and 2 were repeated N_{sim} = 1000 times.

The average bias and standard error of the analyzed estimators were calculated. The results of the Monte Carlo study are presented in table 2.

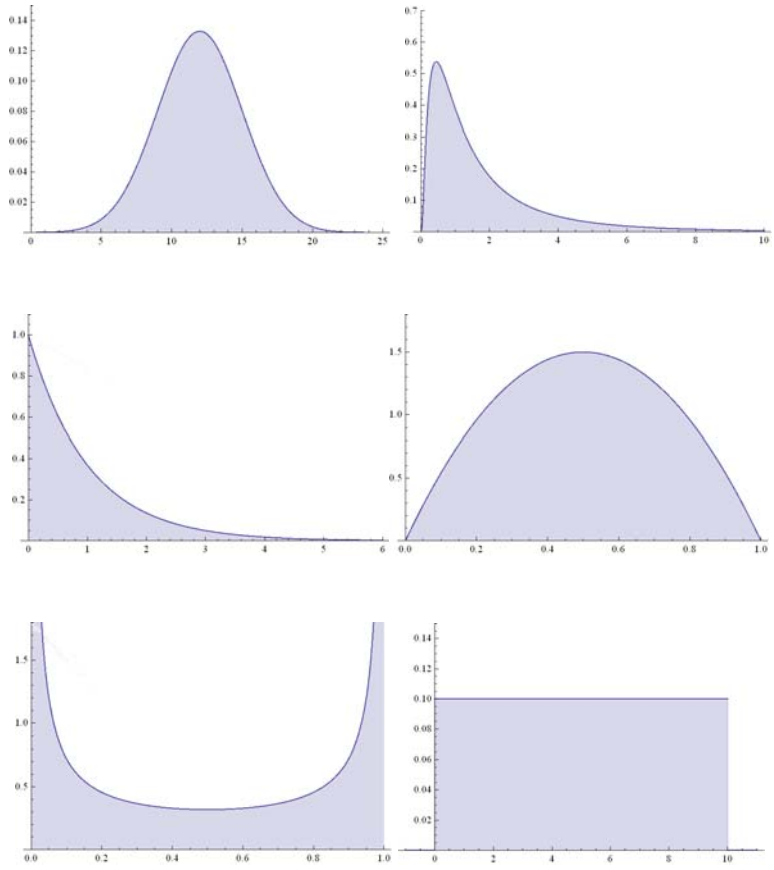

Figure 2. Densities of distributions analyzed in the Monte Carlo study (normal, log-normal, exponential, beta, beta and uniform)

Source: Own preparation in Mathematica.

Sample median estimators for the right skew distributions (log-normal distribution and exponential distribution) are biased. The bias is smaller for the bootstrap method. The standard error of the median estimation in the proposed methods is significantly smaller than in the bootstrap estimation in cases of beta(0.2, 0.2) and uniform (0, 10) distributions.

Table 2. Bias estimation and the standard error of the estimator

Distribution	Sample size n	Method			
		Bootstrap	Kernel		
			Rectangular	Triangle	Gaussian
Normal $N(\tilde{1}, 1)$	5	0.011	0.027	–0.001	0.016
		0.800	0.751	0.758	0.722
	10	0.021	0.000	0.019	0.027
		0.523	0.529	0.516	0.530
	20	0.006	0.006	0.000	0.011
		0.372	0.380	0.378	0.370
Log-normal $LN(0, 1)$	5	0.379	0.413	0.423	0.432
		1.393	1.336	1.652	1.512
	10	0.121	0.256	0.256	0.240
		0.652	0.720	0.685	0.706
	20	0.089	0.186	0.184	0.186
		0.447	0.431	0.470	0.446
Beta B(2,2)	5	0.004	0.007	0.000	–0.002
		0.179	0.176	0.173	0.173
	10	0.000	–0.002	0.000	–0.007
		0.134	0.125	0.127	0.128
	20	–0.005	–0.006	–0.005	–0.004
		0.096	0.095	0.093	0.092
Beta B(0.2,0.2)	5	–0.010	0.007	0.000	0.005
		0.396	0.347	0.351	0.355
	10	–0.019	–0.001	–0.004	0.001
		0.339	0.264	0.273	0.273
	20	0.000	0.003	–0.004	–0.003
		0.298	0.218	0.224	0.232
Exponential $E(1)$	5	0.199	0.247	0.218	0.243
		0.792	0.797	0.690	0.722
	10	0.085	0.163	0.166	0.153
		0.466	0.459	0.473	0.464
	20	0.059	0.107	0.111	0.104
		0.317	0.314	0.321	0.319
Uniform U[0, 10]	5	–0.082	–0.098	0.027	–0.067
		2.455	2.278	2.199	2.331
	10	–0.088	0.027	–0.034	0.007
		1.820	1.693	1.670	1.722
	20	0.021	0.047	0.032	–0.018
		1.458	1.288	1.258	1.284

Bootstrap was originally introduced to estimate the variance of complex estimators. Table 3 presents the average estimated variance of median estimators calculated using formula (3). The column 'Simulation' presents estimated variance of a sample median estimator based on N_{sim} = 1000 samples replications. Estimated values of a variance median estimator by the bootstrap method and three analyzed proposals are presented in the last 4 columns of table 3.

Table 3. Variance of the median estimator – Monte Carlo results

Distribution	Sample size *n*	Method				
		Simulation	Bootstrap	Kernel		
				Rectangular	Triangle	Gaussian
Normal N(10, 1)	5	0.292	0.220	0.536	0.574	0.543
	10	0.129	0.113	0.282	0.270	0.272
	20	0.078	0.067	0.149	0.147	0.155
Log-normal LN(0, 1)	5	0.474	0.537	2.292	3.543	1.774
	10	0.159	0.160	0.510	0.418	0.502
	20	0.080	0.078	0.226	0.219	0.182
Beta B(2,2)	5	0.074	0.038	0.079	0.078	0.079
	10	0.046	0.028	0.043	0.043	0.045
	20	0.035	0.023	0.028	0.030	0.030
Beta B(0.2,0.2)	5	0.129	0.063	0.129	0.125	0.128
	10	0.082	0.049	0.068	0.071	0.074
	20	0.065	0.041	0.047	0.051	0.052
Exponential E(1)	5	0.227	0.184	0.456	0.481	0.518
	10	0.102	0.092	0.182	0.203	0.228
	20	0.049	0.042	0.102	0.098	0.102
Uniform U[0, 10]	5	3.671	2.269	5.166	5.120	5.643
	10	1.897	1.468	2.809	2.854	3.120
	20	1.168	0.921	1.687	1.794	1.710

The bootstrap estimation of variance of an estimator leads to best results for the normal, log-normal and exponential distributions. In the case of beta distribution the proposal leads to better variance estimation.

V. CONCLUDING REMARKS

Bootstrap is one of the most commonly used resampling techniques. This method treats the original sample of values as a stand-in for the population and results in resampling with replacement from it repeatedly. The modification of this method was proposed in the paper. The proposal is based on sampling from the estimated distribution. The Monte Carlo study was used to analyze the properties of the proposal. The results obtained due to modification are similar to the original bootstrap. In the case of two distributions, the standard error of an estimator was smaller in the proposed method than in bootstrap.

The proposal could be used in monitoring of the production processes. The method can be especially useful in monitoring the median in non-normal processes where the samples are small.

ACKNOWLEDGEMENTS

The research was supported by Polish National Science Centre grant DEC-2011/03/B/HS4/05630.

REFERENCES

Domański Cz., Pruska K. (2000) *Nieklasyczne metody statystyczne*, PWE Warszawa.

Efron B. *Bootstrap Methods: Another Look at the Jackknife*, Annals of Statistics 7, 1–26, 1979.

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

Janacek, G. J. and Meikle, S. E. (1997), *Control charts based on medians*. Journal of the Royal Statistical Society: Series D (The Statistician), 46: 19–31.

Pruska K. (2007) Estimations of Bias and Variance of Sample Median by Jackknife and Bootstrap Method, [in:] Acta Universitatis Lodziensis, Folia Oeconomica, 206. s. 67–78.

Rubinstein R.Y., Kroese D.P. (2008) *Simulation and the Monte Carlo Method*, John Wiley & Sons, Inc. New Jersey.

Sheather S.J. (2004) *Density Estimation*, Statistical Science vol. 19, no. 4, s. 588–597

Wieczorkowski R., Zieliński R. (1997) *Komputerowe generatory liczb losowych*, Wydawnictwa Naukowo-Techniczne, Warszawa.

Grzegorz Kończak

O ESTYMACJI MEDIANY METODĄ REPRÓBKOWANIA DLA MAŁYCH PRÓB

Najczęściej wykorzystywaną w badaniach statystycznych metodą repróbkowania jest bootstrap. Metoda ta prowadzi do wielokrotnego zwrotnego pobierania próbki losowej z próby pierwotnej. Zaletą tej metody jest fakt, że może być wykorzystana do wnioskowania o parametrach populacji nawet wówczas, gdy nie jest znany jej rozkład. W opracowaniu przedstawiono propozycję modyfikacji metody bootstrap. Repróbkowanie przeprowadza się z rozkładu, który otrzymuje się metodą estymacji jądrowej funkcji gęstości. Proponowana metoda została porównana z klasyczną metodą bootstrap z wykorzystaniem symulacji komputerowych. W badaniach porównawczych skoncentrowano się na estymacji parametrów populacji na podstawie małych prób.