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 235, 2010

*Aleksandra Baszczyńska**, *Dorota Pekasiewicz***

SELECTED METHODS OF INTERVAL ESTIMATION OF THE MEDIAN. THE ANALYSIS OF ACCURACY OF ESTIMATION

Abstract. The sample median is one of the estimators of population central tendency and can be used, when analyzed random variables have asymmetric distributions.

In the paper, selected methods of the interval estimation of a population median are presented. We generated different types of populations, then calculated confidence intervals for median and compared them by means of the obtained accuracy of estimation.

Key words: confidence interval, median, asymmetric distribution.

I. INTRODUCTION

Let us assume that we investigate a population with regard to random variable X with unknown continuous distribution with the median *Me*. Let $X_1,...,X_n$ be a simple random sample drawn from this population, $x_1,...,x_n$ be the realization of this sample, and $X_{(1)},...,X_{(n)}$ denote the original random sample after arrangement in increasing order.

The sample median, defined as $X_{\left(\frac{n+1}{2}\right)}$ for odd n and any number between $X_{\left(\frac{n}{2}\right)}$ and $X_{\left(\frac{n}{2}+1\right)}$ for even n, is one of the measures of location. When n is even, the most frequently used definition for the sample median is the following: $Me(n)=\frac{X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}}{2}$. It can be used as an estimate of the population central tendency, especially when analyzed random variables have asymmetric distributions or these distributions are heavy-tailed.

* Ph. D., Chair of Statistical Methods, University of Łódź.
** Ph. D., Chair of Statistical Methods, University of Łódź.

The median is an asymptotically unbiased and consistent estimator. But it is not effective even asymptotically. If random variable has a normal distribution, the median is a worse estimator of the population central tendency than arithmetic mean. But when the distribution of random variable is heavy tailed, median is a more stable estimator than arithmetic mean, which is very sensitive to outlying observations (cf.: Lehmann, (1991); Plucińska, Pluciński, (2000)).

In the paper some chosen nonparametric (without any assumption about the distribution of random variable) methods of estimating median using confidence intervals are presented. The first of these methods is based on order statistics and can be used for small and large samples. The second method is based on sample median and the limit theorem, so it can be applied for large samples. The last method is a bootstrap one.

II. CHOSEN METHODS OF ESTIMATING MEDIAN

The first of the analyzed methods of estimating median *Me* (method I) is based on order statistics from the sample $X_1,...,X_n$ (cf.: Domański, Pruska, Wagner, (1998)).

Let *K* denote a random variable whose values are equal to the number of observations in the sample smaller than the median. Random variable *K* has binomial distribution with probability distribution function (cf.: Domański, Pruska, (2000); Greń, (1987)):

$$P(K=k)=\binom{n}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k}=\binom{n}{k}\left(\frac{1}{2}\right)^n \text{ for } k=0,1,2,\ldots,n. \tag{1}$$

Let $T_r^{(n)}$ and $T_s^{(n)}$ denote statistics of order r and s, respectively $(1\le r<s\le n,\quad r,s\in N)$. They are determined on the basis of sample $X_1,...,X_n$.

The probability that the value of the median is in the interval $(T_r^{(n)},T_s^{(n)})$ is calculated by means of the formula:

$$P(T_r^{(n)}<Me<T_s^{(n)})=\sum_{k=r}^{s-1}P(K=k)=\left(\frac{1}{2}\right)^n\sum_{k=r}^{s-1}\binom{n}{k}. \tag{2}$$

The right side of this formula is the confidence coefficient. Properly, selecting the order r and s of statistics, we can get suitable (good) accuracy (with big value of confidence coefficient) of interval estimation of median *Me* of unknown distribution. The use of this method does not always result in obtaining symmetric intervals with respect to the sample median. Table 1 presents orders

of statistics used in estimating median for some sample sizes and values of confidence coefficients ($1-\alpha \geq 0.9$).

Table 1. Orders of statistics used in estimating median and values of confidence coefficients for some sample sizes

Sample sizes	Order of statistic $T_r^{(n)}$	Order of statistic $T_s^{(n)}$	Confidence coefficient
30	11	20	0,901
	10	21	0,957
	9	24	0,991
50	19	31	0,909
	18	32	0,951
	17	36	0,991
70	29	43	0,904
	28	45	0,953
	25	47	0,991
100	43	60	0,905
	41	61	0,954
	37	63	0,991
120	50	69	0,912
	49	71	0,955
	46	75	0,992

Source: Authors' own calculations

The precision of estimation, defined as half of the length of the confidence interval, is the following:

$$d_I = 0{,}5(T_s^{(n)} - T_r^{(n)}). \tag{3}$$

The next analyzed method of interval estimation of a median (method II) ensures proper stability with respect to outliers, and results in obtaining symmetric intervals with respect to the sample median. It is rather simple to use in practice. Confidence interval for median constructed on the basis of the sample $X_1,...,X_n$ is the following:

$$P\left(Me(n) - t_\alpha s_{Me} < Me < Me(n) + t_\alpha s_{Me}\right) = 1-\alpha, \tag{4}$$

where s_{Me} is the standard error of the following form (cf.: Bloch, Gastwirth, (1968); Olive, (2005)):

$$s_{Me} = 0{,}5(T_{k_n}^{(n)} - T_{l_n}^{(n)}), \tag{5}$$

where

$$l_n = \begin{cases} [n/2] - \left[\sqrt{n/4}\right] & for \quad \sqrt{n/4} \in N \\ [n/2] - \left[\sqrt{n/4}\right] - 1 & for \quad \sqrt{n/4} \notin N \end{cases} \tag{6}$$

and

$$k_n = \begin{cases} n - [n/2] + \left[\sqrt{n/4}\right] & for \quad \sqrt{n/4} \in N \\ n - [n/2] + \left[\sqrt{n/4}\right] + 1 & for \quad \sqrt{n/4} \notin N \end{cases}, \tag{7}$$

where $[a]$ denotes the integral part of a.

The value t_α is from t-Student distribution with $Df = k_n - l_n - 1$ degrees of freedom ($Df \approx [\sqrt{n}]$). The number of the degrees of freedom is the same as in the case of the confidence interval for a trimmed mean (Olive, (2005)).

For the large samples confidence interval is the following:

$$P\left(Me(n) - u_\alpha s_{Me} < Me < Me(n) + u_\alpha s_{Me}\right) = 1 - \alpha. \tag{8}$$

Note that the above confidence intervals will give similar results for samples of size of about 900 or greater.

The precision of estimation for this method is the following:

$$d_{II} = 0{,}5(T_{k_n}^{(n)} - T_{l_n}^{(n)})t_\alpha. \tag{9}$$

Another method that can be used for estimating median by means of confidence interval is a bootstrap method (method III).

To derive the bootstrap evaluation of *Me* on the basis of simple sample $X_1,...,X_n$ we generate N ($N \geq 1000$) values $x_1^*, x_2^*,..., x_n^*$ from the bootstrap distribution $P\left(X_B = x_k\right) = \frac{1}{n}$, for k=1,..., n.

Values $x_1^*, x_2^*,..., x_n^*$ are the realizations of the bootstrap sample $X_1^*, X_2^*,..., X_n^*$. For N replications we get N sequences $x_{1,i}^*, x_{2,i}^*,..., x_{n,i}^*$, $i = 1,..., N,$ which are bootstrap samples values.

In bootstrap estimation based on percentiles, for every sequence $x_{1,i}^*, x_{2,i}^*,..., x_{n,i}^*$, $i = 1,..., N$, we compute value Me_i^B which is an evaluation of parameter Me on the basis of the *i*-th bootstrap sample. In this way we receive a sequence of values $Me_1^B, Me_2^B,..., Me_N^B$. Using this sequence we determine the percentiles of the order $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$, respectively. Confidence bootstrap interval for median *Me* based on percentiles with the confidence coefficient $1-\alpha$ is the interval of the following form (cf.: Domański, Pruska, (2000)):

$$P\left(Me_{\alpha/2}^B < Me < Me_{1-\alpha/2}^B\right) \approx 1-\alpha, \tag{10}$$

where $Me_{\alpha/2}^B$ and $Me_{1-\alpha/2}^B$ are the percentiles of order $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$, respectively.

The precision of estimation for this method is the following:

$$d_{III} = 0{,}5\left(Me_{1-\alpha/2}^B - Me_{\alpha/2}^B\right). \tag{11}$$

III. COMPARISON OF THE PRECISION OF INTERVAL ESTIMATION OF MEDIAN

To assess the interval estimation of a median the following simulation study was carried out. The populations of the following distributions were generated:

- exponential with parameters $\lambda = 0.5, 1, 2, 3, 4, 5, 6$,
- χ^2 with degrees of freedom $k = 1, 3, 4, 5, 6, 7, 8$
- mixture of distributions with the density function $f(x) = wf_1(x) + (1-w)f_2(x)$, where $f_1(x)$ is the density function of the normal distribution N(6,1); $f_2(x)$ the density function of the distribution N(0,1); $w \in (0,1)$.
- mixture of distributions with the density function $f(x) = wf_3(x) + (1-w)f_4(x)$, where $f_3(x)$ is the density function of the normal distribution N(6,4); $f_4(x)$ the density function of the distribution N(0,2); $w \in (0,1)$.

In the simulation study the populations considered were characterized as asymmetric (exponential and χ^2) or bimodal (mixtures of distributions). In

these situations the use of the median as the estimator of location parameter is reasonable.

From the generated populations, samples of sizes 30, 50, 70, 100 and 120 were chosen. For fixed confidence coefficients $1-\alpha = 0.9, 0.95, 0.99$ the confidence intervals for median mentioned earlier were computed. This procedure was repeated 10000 times. The precision of estimating median (d_I, d_{II}, d_{III}) was calculated. Moreover, confidence coefficient was estimated as proportion of the intervals which contain the true parameter of the population distribution (p_I, p_{II}, p_{III}). The obtained results for sample size 100, in the case of the exponential and χ^2 distribution (Table 2, Figure 1 and Figure 2) and mixture of distributions (Table 3 and Figure 3) with different values of parameter of distribution, for $(1-\alpha = 0.95)$ are presented below.

Table 2. Precisions of estimation and proportion of intervals which contain the true parameter of the population distribution for exponential and χ^2 distribution (sample size 100)

Type of the distribution of X	Parameter of distribution	Me	Method I		Method II		Method III	
			d_I	p_I	d_{II}	p_{II}	d_{III}	p_{III}
Exponential with parameter λ	0.5	0,345	0,104	0,951	0,103	0,920	0,098	0,962
	1	0,685	0,203	0,953	0,201	0,924	0,1938	0,949
	2	1,370	0,398	0,955	0,391	0,923	0,391	0,951
	3	2,072	0,605	0,956	0,601	0,920	0,588	0,955
	4	2,748	0,807	0,953	0,806	0,917	0,787	0,937
	5	3,469	1,035	0,959	1,027	0,928	0,973	0,945
	6	4,206	1,243	0,954	1,239	0,921	1,147	0,952
Chi-squared with k degree of freedom	1	0,456	0,226	0,957	0,219	0,919	0,208	0,955
	3	2,390	0,554	0,953	0,552	0,923	0,515	0,957
	4	3,324	0,644	0,955	0,648	0,919	0,621	0,956
	5	4,374	0,742	0,956	0,741	0,926	0,722	0,952
	6	5,387	0,832	0,955	0,836	0,927	0,856	0,947
	7	6,317	0,900	0,956	0,900	0,930	0,900	0,956
	8	7,356	0,969	0,956	0,968	0,931	0,942	0,941

Source: Authors' own calculations.

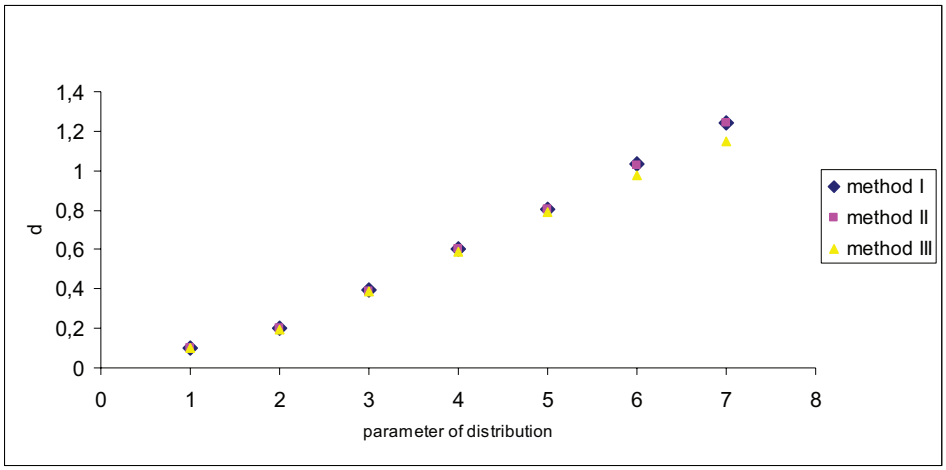

Figure 1. Precisions of estimation of median for exponential distribution (sample size 100)
Source: Authors' own calculations.

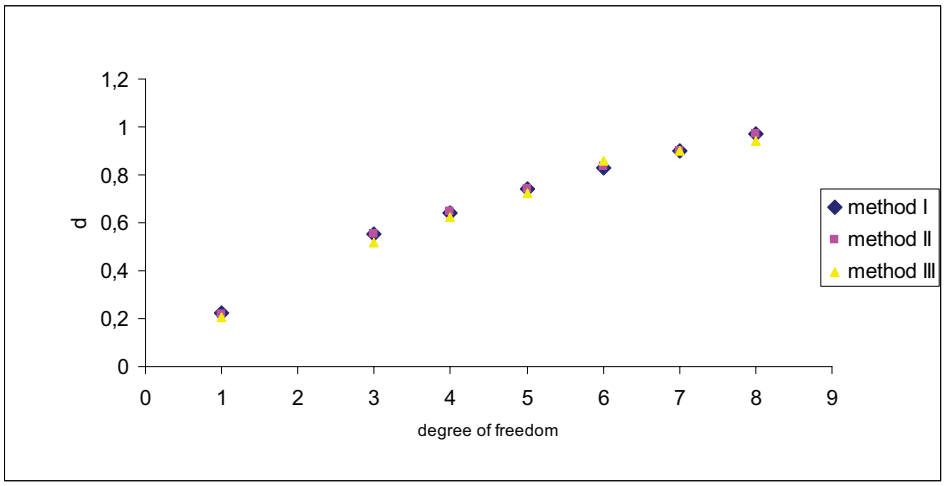

Figure 2. Precisions of estimation of median for χ^2 distribution (sample size 100)
Source: Authors' own calculations

Table 3. Precisions of estimation and proportion of intervals which contain the true parameter of the population distribution for mixture of normal distributions (sample size 100)

Type of distribution of X	w	Me	Method I		Method II		Method III	
			d_I	p_I	d_{II}	p_{II}	d_{III}	p_{III}
Mixture N(6,1) and N(0,1)	0,9	5,405	0,229	0,955	0,258	0,944	0,221	0,950
	0,8	4,806	0,209	0,956	0,234	0,948	0,200	0,957
	0,7	4,203	0,192	0,956	0,215	0,948	0,184	0,945
	0,6	3,600	0,182	0,958	0,203	0,952	0,174	0,946
	0,5	3,004	0,179	0,957	0,200	0,951	0,171	0,952
	0,4	2,410	0,182	0,955	0,206	0,946	0,175	0,945
	0,3	1,806	0,193	0,951	0,217	0,945	0,185	0,945
	0,2	1,206	0,209	0,954	0,234	0,948	0,202	0,955
	0,1	0,600	0,231	0,954	0,259	0,946	0,222	0,949
Mixture N(6,4) and N(0,2)	0,9	5,410	0,910	0,953	1,025	0,948	0,883	0,949
	0,8	4,818	0,816	0,955	0,918	0,943	0,786	0,952
	0,7	4,224	0,725	0,956	0,813	0,946	0,697	0,963
	0,6	3,612	0,638	0,956	0,715	0,949	0,611	0,941
	0,5	3,007	0,565	0,956	0,634	0,948	0,542	0,947
	0,4	2,400	0,505	0,958	0,563	0,951	0,485	0,941
	0,3	1,822	0,465	0,953	0,521	0,950	0,445	0,953
	0,2	1,219	0,452	0,955	0,511	0,942	0,436	0,954
	0,1	0,607	0,468	0,955	0,524	0,949	0,452	0,954

Source: Authors' own calculations.

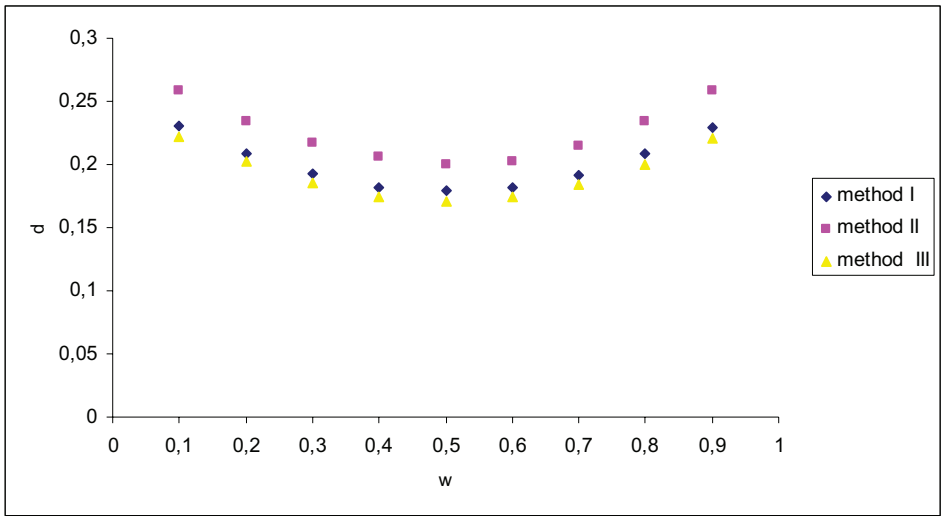

Figure 3. Precisions of estimation of median for mixture of normal distributions (sample size 100)

Source: Authors' own calculations.

IV. CONCLUSIONS

In case of the first group of experiments (asymmetric distributions) confidence intervals for median have the shortest lengths when the bootstrap method was used. Application of the first two methods resulted nearly in the same way. But in the case of method II probability that median belongs to the calculated interval was much smaller than fixed confidence coefficient. So, method III (bootstrap method) allows to get the best estimation of median (the biggest precision) for the fixed confidence coefficient.

In case of the second group of experiments (mixture of normal distributions) all of the obtained confidence intervals fulfilled the assumption of value of confidence coefficient. However method III appeared to be the most effective.

For others values of confidence coefficients the results were similar for determined classes of distributions. The results were similar also for larger sample sizes.

In case of sample sizes smaller than 100, method II is not satisfactory – results are much wore than in method I or III. The confidence intervals are longer and probability that median belongs to the calculated intervals is smaller than fixed confidence coefficient.

The application of method II, only for very big samples, allows to obtain confidence intervals for median with fixed confidence coefficient.

REFERENCES

Bloch D. A., Gastwirth J. L. (1968), *On a Simple Estimate of the Reciprocal of the Density Function*, The Annals of Mathematical Statistics, 39, 1083–1085

Domański Cz., Pruska K., Wagner W., (1998), *Wnioskowanie statystyczne przy nieklasycznych założeniach*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź

Domański Cz., Pruska K., (2000), N*ieklasyczne metody statystyczne*, Polskie Wydawnictwo Ekonomiczne, Warszawa

Greń J., (1987), *Statystyka matematyczna. Podręcznik programowany*, Polskie Wydawnictwo Naukowe, Warszawa

Lehmann E. L., (1991), *Teoria estymacji punktowej*, Wydawnictwo Naukowe PWN, Warszawa

Olive D. J., (2005), *A Simple Confidence Interval for the Median*, http://www.math.siu.edu/olive/ppmedci.pdf

Plucińska A., Pluciński E., (2000), *Probabilistyka. Rachunek prawdopodobieństwa. Statystyka matematyczna. Procesy stochastyczne*, Wydawnictwa Naukowo-Techniczne, Warszawa

Aleksandra Baszczyńska, *Dorota Pekasiewicz*

WYBRANE METODY ESTYMACJI PRZEDZIAŁOWEJ MEDIANY. ANALIZA DOKŁADNOŚCI OSZACOWAŃ

Mediana z próby jest jednym z estymatorów parametru położenia i może być stosowana do szacowania, między innymi, gdy analizowane zmienne losowe mają rozkłady asymetryczne.

W pracy przedstawione są wybrane metody estymacji przedziałowej mediany. W celu przeprowadzenia analizy efektywności rozważanych metod wygenerowano populacje o różnych typach rozkładów, a następnie wyznaczano przedziały ufności dla mediany i porównywano uzyskane dokładności oszacowań.