ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

*Wiesław Wagner**

DISTRIBUTION OF LINEAR COMBINATION THE SAMPLE MEAN AND THE SAMPLE MEDIAN

Abstract. In the work there is examined the estimator of linear combination of arithmetic mean and median from a random sample of a random variable in the symmetrical distribution. The coefficients of combinations are determined according to the criterion of minimization of variances. Properties of the estimator are expressed by its density function and the given result from simulation research for the uniform distribution.

Key words: symmetrical distribution, arithmetic mean, median, estimator of linear combination, density function, Monte Carlo simulation.

I. INTRODUCTION

Arithmetic mean and median are universally applied unbiased estimators of the expected value of random variable of symmetrical distribution. Both these estimators are unbiased, but they have different variances (e.g. Lehmann 1990). Each of the estimators behaves in a different way for given probability distributions.

Instead of considering each of the mentioned estimators in the problems of estimation and verification of hypotheses, it is worth applying the complex estimator being the linear combination of the mentioned estimators. It has much higher efficiency in the sense of minimization of variance than the estimators of arithmetic mean and median. For the indicated complex estimator there is determined the probability distribution of a given density function belonging to the class of trimmed normal distributions.

II. SIMULATION RESEARCH

Let us assume that there is carried out a random experiment consisting in drawing $N = 1000$ times of $n = 15$ element sample from population of the uniform distribution $J(0,1)$. For drawing random numbers there was used the

* Professor, University of Information Technology and Management in Rzeszów.

function *LOS* in EXCEL calculation sheet program. For each sample there was determined arithmetic mean and median. For each of the mentioned numerical characteristics there was executed the histogram of size with *8* class ranges of the length *0,1* within the scope from *0,1* to *0,9* (figure 1 and 2).

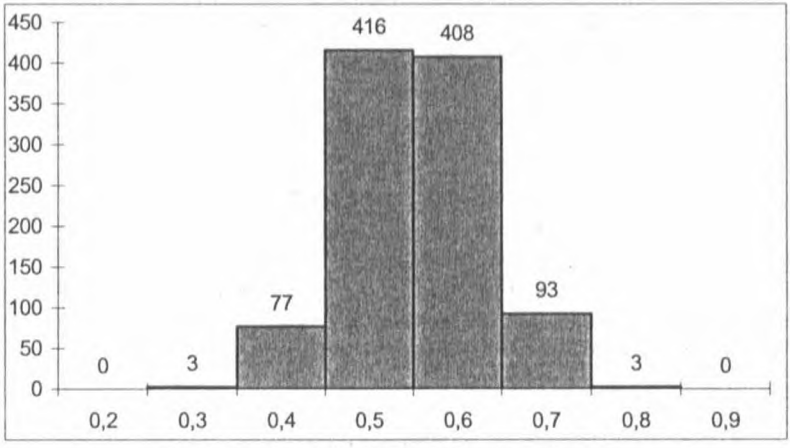

Fig. 1. Empirical distribution for arithmetic means
Source: Own elaboration.

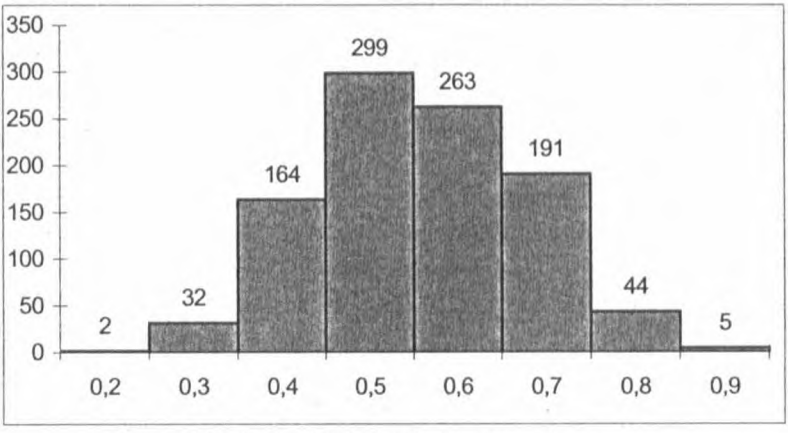

Fig. 2. Empirical distribution for medians
Source: Own elaboration.

From the presented graphs one may draw the conclusions:

a) arithmetic means are concentrated, first of all, in the ranges $(0,4,\ 0,5)$ and $(0,5, 0,6)$ which contain *824* means, i.e. in the interval *0,2* there are concentrated *82,4* % means,

b) arithmetic means are good estimators of the expected value *0,5* of the distribution $J(0,1)$,

c) medians in the figure 2 show a big dispersion and are located mainly in the ranges from $(0,3, 0,4)$, ..., $(0,6, 0,7)$ containing in total *907* medians, i.e. in the interval of the length *0,4* there is contained *90,7* % medians,

d) a median is not a good estimator for the expected value in the distribution $J(0,1)$.

For indication of similarity between class sizes $\{(f_{1j}, f_{2j}) : j = 1,2,...,k\}$ of two distributive rows with k class ranges there is proposed the measure

$$MP = 1 - \frac{1}{2000} \sum_{j=1}^{k} | f_{1j} - f_{2j} |,$$

which assumes values from the range $\langle 0, 1 \rangle$. For the presented distributive rows $MP = 0,738$, i.e. they are of little similarity.

The results for *1000* samples were also used for indication how there behave mean of means (mean, mean), median of means (median, mean), mean of medians (means, medians) and median of medians (median, median). There was also determined the number pf cases when for a given sample the mean was greater than the median and the % of them was determined. Adequate results are presented in the setting-up:

N	500		750		1000	
Statistics	Mean	Median	Mean	Median	Mean	Median
Mean	0,4995	0,4962	0,5004	0,4974	0,5026	0,5014
Median	0,5049	0,4931	0,5046	0,4963	0,5065	0,5005
>	231		357		475	
%	46,2		47,6		47,5	

Very close to the number *0,5* there are the cases (mean, mean) for $N=500$ and $N=750$ and (median, median) for $N=1000$ or the two dimensional sample (means, medians) of the size $N=1000$ formed from the simulation, there was executed the correlation plot (figure 3) in which there is also presented the regression dependence of means on medians and their coefficient of linear correlation.

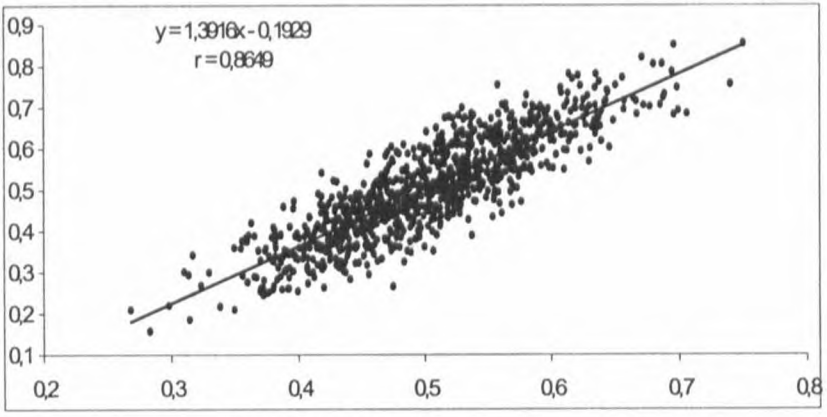

Fig. 3. Correlation plot of means and medias
Source: Own elaboration

The principal cloud of points is located at the rectangle $(0,4, 0,6) \times (0,3, 0,7)$ which contains *809* points, i.e. *80,9* % of all examined ones. Correlation between the examined numerical characteristics is high and it is $r=0,865$.

We will come back to the results of simulation research in the final part of the work.

III. DENOTATIONS

Further we assume the following denotations:

- X – random variable of continuous type of symmetrical probability distribution determined in the set R,
- $X_1, X_2, ..., X_n$ – simple sample of n random variables collected from the general population of distribution X,
- $f(x)$ – density function, continuous and positive in point 0 and symmetrical around the straight line $x = 0$,
- μ - parameter of position of distribution of general variable X,
- $f(x-\mu)$ – density function with the parameter of position μ,
- $\overline{X}, \widetilde{X}$ – arithmetic mean and median from sample $X_1, X_2, ..., X_n$,
- $v^2 = \int_{-\infty}^{\infty} x^2 f(x)dx$ - normal moment of 2nd rank,

➢ $\tau = \int_{-\infty}^{\infty} |x| f(x)dx$ - absolute normal moment of 1st rank.

➢ *Theorem.* (Fisz (1967, s. 401), Lehmann (1983, s. 394), Serfling (1991), Samuel-Cahn (1994)). Let k_n be the sequence of such integers that $\frac{k_n}{n} = p + R_n$ $(0 < p < 1)$ z $\sqrt{n}R_n \to 0$ and let $X_1, X_2, ..., X_n$ constitute the sample of independent random variables of distribution F, for which $F(\xi_p) = p$ and density f is positive in ξ_p, then $\sqrt{n}(X_{k_n:n} - \xi_p) \sim N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right)$, where $X_{1:n}, X_{2:n}, ..., X_{n;n}$ are ordered statistics with $X_1, X_2, ..., X_n$.

➢ *Conclusion.* For $\tilde{x}$ there occurs $\sqrt{n}(X_{k_n:n} - \tilde{x}) \sim N\left(0, \frac{1}{4f^2(\tilde{x})}\right)$, i.e.

$$D^2(\sqrt{n}(\tilde{x} - \mu)) = D^2(\sqrt{n}(\tilde{x} - \mu)) = nD^2(\tilde{x}) = \frac{1}{4f^2(0)}.$$

➢ *Theorem* (Domilano and Puig 2004). The distribution of two-dimensional random variable $(\sqrt{n}(\bar{x} - \mu), \sqrt{n}(\tilde{x} - \mu))$ has an asymptotic two-dimensional normal distribution $\mathbf{z} \sim N_2(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{z} = \sqrt{n}\begin{bmatrix} \bar{x} - \mu \\ \tilde{x} - \mu \end{bmatrix}$ and

$$\mathbf{\Sigma} = \begin{bmatrix} v^2 & \frac{\tau}{2f(0)} \\ \frac{\tau}{2f(0)} & \frac{1}{4f^2(0)} \end{bmatrix}.$$

IV. ESTIMATOR OF LINEAR COMBINATION

A lot of authors (e.g. Chan and He, 1994, Samuel-Cahn 1994, Damilano and Puig 2004) dealt with examination of estimator $\overline{\mu} = w\bar{x} + (1-w)\tilde{x}$, where $w \in R$ .is weight which we select so that it will have the lowest variance.

Applying denotations presented in chapter 2 we have, for the presented estimator, the moments: expected value $E(\overline{\mu}) = 0$ and variance

$$D^2(\overline{\mu}) = w^2 D^2(\overline{x}) + (1-w)^2 D^2(\widetilde{x}) + 2w(1-w)Cov(\overline{x},\widetilde{x}) =$$

$$= w^2 \frac{v^2}{n} + (1-w)^2 \frac{1}{4nf^2(0)} + w(1-w)\frac{\tau}{nf(0)}.$$

In order to make this variance minimum, one should determine the derivative and, after equating to zero, solve the adequate equation, which leads to the derivative

$$\frac{\partial D^2(\overline{\mu})}{\partial w} = 2w\frac{v^2}{n} - (1-w)\frac{1}{2nf^2(0)} + (1-2w)\frac{\tau}{nf(0)}$$

and the adequate equation $4f^2(0)v^2 \cdot w - 1 + w + 2\tau f(0) - 4\tau f(0) \cdot w = 0$, and thus $w = \dfrac{1-2\tau f(0)}{4v^2 f^2(0) - 4\tau f(0) + 1}$.

Example 1. In the case $X \sim N(\mu,\sigma)$, we have $X - \mu \sim N(0,\sigma)$ and $f(0) = \dfrac{1}{\sigma\sqrt{2\pi}}$, $v^2 = \sigma^2$ and

$$\tau = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} |t| \exp\left\{-\frac{t^2}{2\sigma^2}\right\} dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{0}^{\infty} t \exp\left\{-\frac{t^2}{2\sigma^2}\right\} dt = \frac{\sigma}{\sqrt{2\pi}}.$$

The received values provide $w=1$. It means that in the case of normal i for big samples the mean is an effective estimator for parameter μ, and the median does not lead to raising the parameter's efficiency.

Now we will present the approach for determining estimator $\overline{\mu}$, without using the asymptotic properties of estimators:

- μ – parameter of position of random variable X,
- T_1, T_2 – unbiased estimators of parameter μ,
- $\sigma_i^2(\mu)$ – variances of estimators T_i, $i = 1, 2$,

➢ $\lambda^2 = \dfrac{\sigma_1^2(\mu)}{\sigma_2^2(\mu)} > 0$, measure of efficiency of relative variance of two estimators, where λ^2 does not depend on μ and, moreover, we assume that $0 < \lambda \leq 1$ where without loss of generality we assume that $\sigma_1^2(\mu) \leq \sigma_2^2(\mu)$,

➢ $\rho = \dfrac{Cov(T_1, T_2)}{\sigma_1(\mu) \cdot \sigma_2(\mu)}$ - measure of correlation of estimators T_1, T_2.

One should select the best weighed estimator $T(w) = wT_1 + (1-w)T_2$, for $w \in R$ according to the criterion of the lowest variance, i.e. $D^2(T(w)) = \min$. We act analogically to the earlier signalized course of procedure:

$\Rightarrow$ we determine variance

$$D^2(T(w)) = w^2 D^2(T_1) + 2w(1-w)Cov(T_1, T_2) - (1-w)^2 D^2(T_1)$$

$$= D^2(T_2)\left[w^2\lambda^2 + 2w(1-w)\frac{Cov(T_1, T_2)}{D^2(T_2)} + (1-w)^2 \right] =$$

$$= \sigma_2^2(\theta)[w^2\lambda^2 + 2w(1-w)\rho\lambda + (1-w)^2],$$

$\Rightarrow$ we apply the necessary condition of existence of the extremum $\dfrac{\partial D^2(T(w))}{\partial w} = 0$, which leads to the equation $\lambda^2 w^2 + \rho\lambda(1-2w) - 1 + w = 0$,

$\Rightarrow$ $w^* = \dfrac{1-\rho\lambda}{\lambda^2 - 2\rho\lambda + 1}$ – determined weight,

$\Rightarrow$ for the presented solution the variance $D^2(T(w^*))$ is

$$D^2(T(w^*))$$

$$= \sigma_2^2(\theta)\left\{ \frac{(1-\rho\lambda)^2}{(1-2\rho\lambda+\lambda^2)^2} \cdot \lambda^2 + 2\frac{(1-\rho\lambda)(\lambda^2-\rho\lambda)^2}{(1-2\rho\lambda+\lambda^2)^2} + \frac{(\lambda^2-\rho\lambda)^2}{(1-2\rho\lambda+\lambda^2)^2} \right\}$$

$$= B\{[(1-\rho\lambda)\lambda + (\lambda^2-\rho\lambda)\rho]^2 + (\lambda^2-\rho\lambda)^2(1-\rho^2)\},$$

$$= B \cdot (1-\rho^2)[\lambda^2(1-\rho^2) + (\lambda^2-\rho\lambda)^2] =$$

$$B(1-\rho^2)\lambda^2(1-2\rho\lambda+\lambda^2) =$$

$$B(1-\rho^2)\lambda^2(1-2\rho\lambda+\lambda^2)=\frac{\sigma_2^2(\theta)}{1-2\rho\lambda+\lambda^2}\cdot(1-\rho^2)\lambda^2=\sigma_1^2(\theta)\frac{1-\rho^2}{1-2\rho\lambda+\lambda^2}$$

where $B=\dfrac{\sigma_2^2(\theta)}{(1-2\rho\lambda+\lambda^2)^2}$.

Example 2. We accept assumptions of random variable X as in the example 1. Let data be unbiased estimators of parameter $\mu: T_1=\overline{X}-$ the arithmetic mean from the sample, $T_2=\widetilde{X}-$the median from the sample. Their variances are $D^2(\overline{X})=\dfrac{\sigma^2}{n}$ and $D^2(\widetilde{X})=\dfrac{\pi\sigma^2}{2n}$, thus $\lambda^2=\dfrac{2}{\pi}=0{,}6366$ and $\lambda=0{,}7979$. After executed substitutions we have

$$w^*=\frac{1-\rho\sqrt{\dfrac{2}{\pi}}}{1-2\rho\sqrt{\dfrac{2}{\pi}+\dfrac{2}{\pi}}}=\frac{1-0{,}798\rho}{1{,}6366-1{,}5958\rho},$$

and at $\rho=0$, $w^*=0{,}61102$.

It means that in the case of small samples collected from population of normal distribution, the effective estimator of parameter μ will be determined from $\overline{\mu}=0{,}611\overline{x}+0{,}389\widetilde{x}$.

The joint distribution of estimators of mean and median for symmetrical distributions is given by the theorem.

Theorem (Domilano and Puig 2004). For symmetrical distributions for which the estimator of the parameter of position μ is in the form $\overline{\mu}=w\overline{x}+(1-w)\widetilde{x}$, have density

$$f(x;\mu,\sigma,\theta)=\frac{\varphi(\theta)}{2\sigma(1-\Phi(\theta))}\exp\left(-\theta\frac{|x-\mu|}{\sigma}-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where $\varphi(\theta)$, $\Phi(\theta)$ are density and distribution function $N(0,1)$, $\mu \in R, \sigma \in R_+$ are parameters of position and scale, and $\theta \in R$ of shape, and moreover $w = w(\theta) = \dfrac{1-\Phi(\theta)}{1-\Phi(\theta)+\theta\varphi(\theta)}$.

The given density function is a composition of Laplace's distribution and normal distribution. The first one refers to the criterion of determining the median from random sample based on absolute deviations $\min_a \sum_{i=1}^{n} |x_i - a|$, and the second one refers to the criterion of determining the arithmetic mean from squares of deviations $\min_a \sum_{i=1}^{n} (x_i - a)^2$. In particular when $\theta = 0$, then the given density is the density of normal distribution $N(\mu, \sigma)$. The shape of density depends on the value of parameter θ, which is shown in figure 4, for $\mu = 0, \sigma = 1$ and different values θ

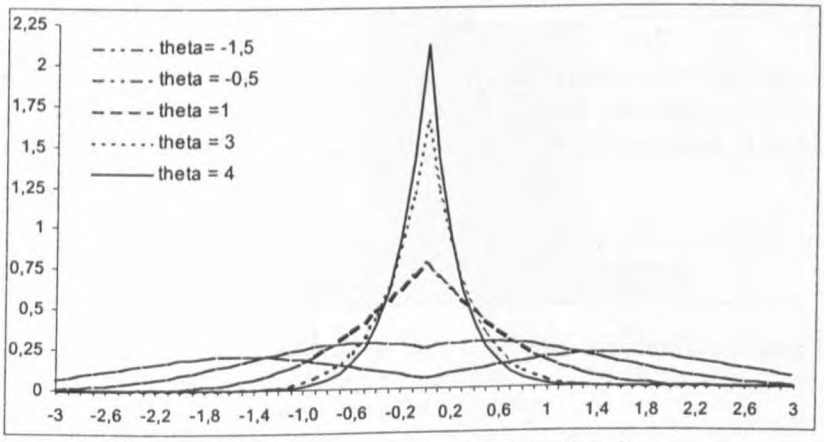

Fig. 4. Curves of density function $f(x;0,1,\theta)$

Source: Own elaboration

Behaviour of weights $w(\theta)$ depending on the value of parameter θ is shown in the setting-up:

theta	–3	–2,5	–2	–1.5	–1	–0,5	0	0,5	1	1,5	2	2,5	3	3,5
w	1,013	1,046	1,124	1,263	1,404	1,342	1,000	0,637	0,396	0,256	0,174	0,124	0,092	0,071

Weights are highest for $\theta = -1$ and ascending leftwards from this point and quite quickly descending rightwards from this point.

V. COMBINED ESTIMATOR FOR UNIFORM DISTRIBUTION

In chapter 2 we presented the results of simulation for the uniform distribution. Here we will present analytical results being a linear combination of mean and median from random sample for random variable X of distribution $J(0,1)$. For this purpose we present the successive results and facts referring to the mentioned estimators:

- expected value $\mu = E(X) = \frac{1}{2}$ and variance $\sigma^2 = D^2(X) = \frac{1}{12}$,
- $\overline{X}, \widetilde{X}$ - mean and median from random sample $X_1, X_2, ..., X_n$ from population of distribution $J(0,1)$,
- expected value and variance for mean – $E(\overline{X}) = \mu = \frac{1}{2}$,

$$D^2(\widetilde{X}) = \frac{\sigma}{n} = \frac{1}{12n},$$

- expected value and variance for median is determined from position statistics $(m+1)$-th, assuming, without loss of generality, that size of the sample is odd $n = 2m + 1$, and which has beta distribution with parameters $p = m+1$ i $q = m+1$, i.e.:

$$E(\widetilde{X}) = \frac{p}{p+q} = \frac{1}{2}, \quad D^2(\widetilde{X}) = \frac{pq}{(p+q)^2(p+q+1)} = \frac{1}{4(n+2)},$$

- measure of relative efficiency - $\lambda^2 = \frac{D^2(\overline{X})}{D^2(\widetilde{X})} = \frac{n+2}{3n} \to \frac{1}{3}$, when $n \to \infty$,
- covariance (Samuel-Cahn 1994) - $Cov(\overline{X}, \widetilde{X}) = \frac{n+1}{8n(n+2)}$,
- correlation coefficient -

$$\rho = \frac{Cov(\overline{X}, \widetilde{X})}{D(\overline{X})D(\widetilde{X})} = \frac{n+1}{8n(n+2)} \cdot \sqrt{48n(n+2)} = \frac{\sqrt{3}(n+1)}{2\sqrt{n(n+2)}} \to \frac{\sqrt{3}}{2} = 0{,}866,$$

when $n \to \infty$,

➢ at determined magnitudes the value for the weight coefficient is

$$w^* = \frac{1-\rho\lambda}{1-2\rho\lambda+\lambda^2} = \frac{1-\frac{\sqrt{3}}{2}\cdot\frac{1}{\sqrt{3}}}{1-2\frac{\sqrt{3}}{2}\cdot\frac{1}{\sqrt{3}}+\frac{1}{3}} = \frac{3}{2}.$$

Thus in the case of distribution $J(0,1)$ for the estimator of linear combination the weight coefficient will prefer the value of mean with positive weight *1,5* and negative weight *–0,5* for the median.

Coming back to the results of the simulation for the sample of size $n = 15$ presented in chapter 2, we have: $\lambda^2 = 0{,}3378$, $\rho = 0{,}8649$ and $w^* = 1{,}493$, i.e. these results slightly differ from the presented theoretical values. Histogram of value $\overline{\mu} = 1{,}5 \cdot \overline{x} - 0{,}5 \cdot \tilde{x}$ for *1000* samples of size $n=15$ is presented in figure 5.

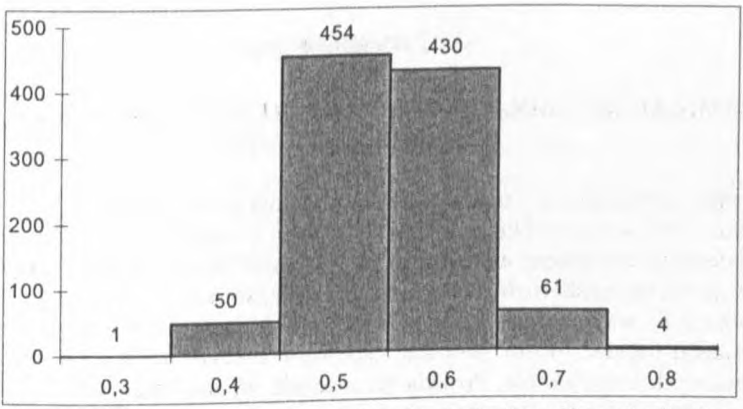

Fig. 5. Empirical distribution for the weight estimator in the uniform distribution
Source: Own elaboration.

Comparing sizes of histograms in figure 1 and 5 we received the measure of similarity
MP =0,939, and measures of entropy for assessment of uniformity of the distribution of class sizes in both cases were $E_1 = -\sum_{j=1}^{k} c_j \log_2 c_j = 1{,}7079$ and

$E_2 = 1{,}5448$, where $c_j = f_j / 1000$. The smaller measure of entropy indicates greater concentration of sizes around the value $0{,}5$, and at the same time it means that the value of the combined estimator is higher than for arithmetic mean.

REFERENCES

Chan Y. M., He X., (1994), *A simple and competitive estimator of location*, Statist. Probab. Lett. 19, 137–142.

Damilano G., Puig P., (2004), *Efficiency of a linear combination of median and the sample mean: the double truncated normal distribution*, Scandinavian Journal of Statistics, 31, No 4, 629–637.

Fisz M., (1967), *Rachunek prawdopodobieństwa i statystyka matematyczna*, PWN, Warszawa.

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

Samuel-Cahn E., (1994), *Combining unbiased estimators*, Amer. Statist., 48, 34–46.

Wiesław Wagner

ROZKŁAD KOMBINACJI LINIOWEJ ŚREDNIEJ ARYTMETYCZNEJ I MEDIANY Z PRÓBY

Średnia arytmetyczna i mediana są powszechnie stosowanymi estymatorami nieobciążonymi wartości oczekiwanej zmiennej losowej o rozkładzie symetrycznym. Oba te estymator są nieobciążone, ale mają różne wariancje. Każdy z estymatorów różnie się zachowuje dla zadanych rozkładów prawdopodobieństwa.

Zamiast rozważać każdy ze wspomnianych estymatorów w problemach estymacji i weryfikacji hipotez, warto stosować estymator złożony będący liniową kombinacją nadmienionych estymatorów. Posiada on znacznie wyższą efektywność w sensie minimalizacji wariancji, niż estymatory średniej arytmetycznej i mediany. Dla wskazanego estymatora złożonego określa się rozkład prawdopodobieństwa o zadanej funkcji gęstości, należący do klasy uciętych rozkładów normalnych.