Mykhaylo DOROZHOVETS

Rzeszow University of Technology, Poland; National University "Lviv Polytechnic", Ukraine

The Estimation of Measurement Result and its Uncertainty from Random Observations Using the Comparison Method

***Abstract**. Proposed method based on the parallel comparison of the input observations (after ordering) with several sets of the reference observations, which are equal to the expected values of order statistics corresponded to the selected population probability density functions. The mathematical models of the determination of the best result and its standard uncertainty and also the results of investigation (by Monte Carlo method) of effectiveness of the proposed method are presented in this article.*

***Streszczenie.** Proponowana metoda bazuje na równoległym porównywaniu obserwacji wejściowych (po ich uporządkowaniu) z zestawem obserwacji referencyjnych, które są równe wartościom oczekiwanym statystyk pozycyjnych odpowiadających wybranym funkcjom gęstości rozkładu populacji. W artykule przedstawiono zależności pozwalające na wyznaczenie najlepszego wyniku oraz jego niepewności standardowej a także wyniki badań (metodą Monte Carlo) skuteczności proponowanej metody. (**Estymacja wyniku pomiaru i jego niepewności z losowych obserwacji metodą porównawczą**).*

Słowa kluczowe: obserwacje losowe i referencyjne, niepewność, porównanie.
Keywords: random and reference observations, uncertainty, comparison.

Introduction

The main problem of the measurement data (observations) processing is hidden in knowledge of the population probability density function (PDF). If such density is known, then the best estimation of measurement result (position parameter) and its standard uncertainty can be calculated correctly. International Guide ISO [1] recommends to calculate the measured result as the sample mean value $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and its standard uncertainty as $u_A(\bar{x}) = \sigma/\sqrt{n}$ (where *n* is a number of independent measurement observations $x_1, x_2, ..., x_n$, vector $\mathbf{X}^T = (x_1, x_2, ..., x_n)$, σ is standard deviation of observations.

But in the measurement practice serious inadequacies in calculation of the measurement result and its uncertainty may exist [2, 3]. Even after removal of all known systematic components from the raw observations, the corrected set of observations may not constitute a sample of purely random normal but rather distribution is very differ from normal. In these cases besides mean value another estimators of location parameter are often used: midrange, median, truncated and winsored means, L and M estimators etc [4]. Effectiveness of these estimators is strongly depended on sample PDF as well as on the properties of outliers.

For example, if PDF is uniform, then the best measurement results is so-called midrange, and its standard uncertainty is determined by the range R of sample and can be calculated after the known formula [5]: $u(x_M) = R/\sqrt{2(n+1)(n+2)}$. From this formula follows that uncertainty decreases practically proportional to the number of observations n on contrary to the $\sqrt{n}$ as in standard uncertainty of sample mean. However, midrange is very unstable if outliers appear.

Another example is corresponded with Laplace PDF (so called, double exponential). As it is well known, a sample median is the better location parameter in comparison to the mean value, because its standard uncertainty is less in approximately in $\sqrt{2}$ times from standard uncertainty of mean value [5]. The median is very stable to the present outliers.

If the PDF of the registered observations is a priori unknown, then usually the histogram method and proper statistical test (for example, based on χ^2 distribution) is used in such cases. But this method is just useful when the number of observations is about hundred and more. If the number of observations is limited (for example from 10 to 50) then the histogram come to unstable and the χ^2 test can be positive simultaneously for the few models of PDF and which is no reason to reject one of them. Therefore this way is not always useful.

The **purpose of this work** is the development and investigation method of processing random observations if their probability density distribution is unknown and number of observations is limited.

The bases of proposed method algorithm

In the metrology the highest accuracy of a measurement result can be obtained by the direct comparison method, i.e. the direct comparison between the measured and reference quantities.

For the a priory known PDF of input observations for the determination of the location (L) and scale (V) parameters known method was proposed in [7, 8] can be used. Its essence is shown in figure 1. The main idea is based on direct comparison of the input observations (after sorting) $X_s^T = (x_{(1)}, x_{(2)},.., x_{(n)})$ with the expectation values of the order statistics (k = 1, 2, …, n):

$$\alpha_k = E[x_{(k)}] = \int_{-\infty}^{\infty} x_{(k)} \cdot p_{os}(x_{(k)}) dx_{(k)} . \tag{1}$$

where $p_{os}(x_{(k)})$ is a density distribution of the order statistic $x_{(k)}$ [6].

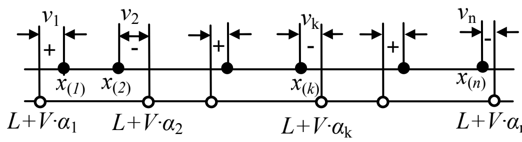

Fig. 1. The comparison of the ordered input observations $x_{(k)}$ and scaled (by L and V) the order statistics α_k

The search parameters L and V are calculated using the weighted least squares method (WLSM): due to minimization the sums of squares of the deviations $v_k = x_{(k)} - \left(L + \alpha_k \cdot V\right)$ of the input ordered observations and scaled expected values α_k by matrix formula [9]:

(2) $$\left(L,V\right)^T = \left(\mathbf{A}^T \mathbf{W} \cdot \mathbf{A}\right)^{-1} \mathbf{A}^T \cdot \mathbf{W} \cdot \mathbf{X}_s = \mathbf{REC} \cdot \mathbf{X}_s ,$$

where $\mathbf{A}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{pmatrix}$; $\mathbf{W} = \left[\mathbf{COV}\right]^{-1}$ is the weight matrix, which equals to the inverse covariance matrix $\mathbf{COV}$ of the order statistics (because however, the order statistics are mutually correlated [6]); and $\mathbf{REC}$ is so-called reconstructed matrix:

(3) $$\mathbf{REC} = \left(\mathbf{A}^T \mathbf{W} \cdot \mathbf{A}\right)^{-1} \mathbf{A}^T \cdot \mathbf{W} = \begin{pmatrix} g_1 & g_2 & g_3 & \cdots & g_{[(n+1)/2]} & \cdots & g_3 & g_2 & g_1 \\ -\gamma_1 & -\gamma_2 & -\gamma_3 & \cdots & 0 & \cdots & \gamma_3 & \gamma_2 & \gamma_1 \end{pmatrix},$$

The elements of the first row (g_i) of this matrix are the weighted coefficients (even-symmetrical from the middle) which are used to determine the location parameter (L) and the elements of the second row (γ_i) in (3) are the weighted coefficients (odd-symmetrical from the middle) which are used to determine dispersion parameter (V):

(4) $$L = \sum_{k=1}^{n} g_k \cdot x_{(k)} , \ V = \sum_{k=1}^{n} h_k \cdot x_{(k)})$$

The standard uncertainty $u_A(L)$ and $u_A(V)$ of these parameters are evaluated by the standard procedure of the weighed least-squares method, namely:

(5) $$u_A(L) = d_{0,0}\sqrt{S_R^2} , \quad u_A(V) = d_{1,1}\sqrt{S_R^2} ,$$

where S_R^2 is unbiased estimator of residuals weighted sums of squares of the deviations v_k; $D = \left(\mathbf{A}^T \cdot \mathbf{W} \cdot \mathbf{A}\right)^{-1} = \begin{pmatrix} d_{0,0}^2 & 0 \\ 0 & d_{1,1}^2 \end{pmatrix}$ - is a disperse matrix of 2×2 size, which in the case of symmetric PDF is diagonal, therefore the location and dispersion parameters are uncorrelated.

In [9] is shown that variation $var(L)$ is corresponded with variation of mean value $var(\bar{x})$ by the inequality:

(6) $$var(L) \le \frac{\sigma^2}{n} = var(\bar{x}),$$

and only if the PDF is normal then $L = \bar{x}$ and $var(L) = var(\bar{x}) = \sigma^2/n$.

From (6) follows that if observations PDF is differ from normal then method based on order statistics provides better results with less standard uncertainty than standard uncertainty of mean value.

Description algorithm of the propose method

Method based on order statistics is useful only if the true PDF $p_{true}(x)$ of input observations is a priory known precisely. If the true PDF $p_{true}(x)$ is a priori unknown or is strongly differ from theoretical model then procedure of the determination location and disperse parameters $(L,V)^T$ based on order statistics must be changed and additionally extended. The problem of unknown $p_{true}(x)$ can be solved by comparison of the input observations with the set $j = 1, 2, ..., J$ series of the expected values $\alpha_{k,j}$ which are corresponded to the proper $j = 1, 2, ..., J$ models of the PDF $p_j(x)$. After comparison the most suitable model of $p_m(x)$ from these set of PDF will be selected.

The simplified block-scheme of the algorithm that realized the proposed method of the observations processing is presented in figure 2. The main steps of algorithm are as follows [10, 11]:

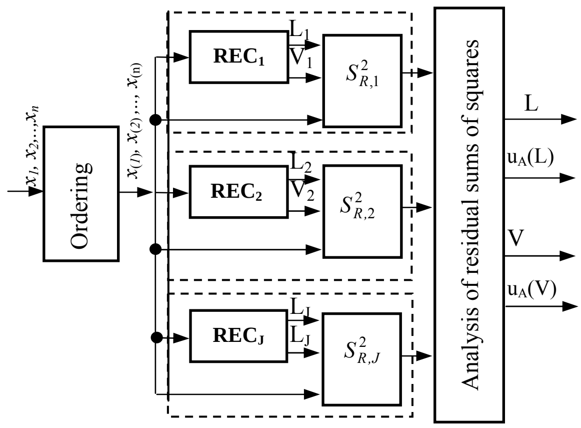

Fig. 2. Block-scheme of the proposed method algorithm of the location and disperse parameters evaluation

1: Ordering of the input observations: X_s^T ;

2: Calculation of the sets values of location and dispersion parameters (L_j, V_j) (after formulas (2) – (4));

3: Calculation of the unbiased estimator of the residual sums of squares $S_{R,j}^2$ as follows:

(7) $$S_{R,j}^2 = \frac{\left(\mathbf{X}_s - \mathbf{A}_j \cdot \left(L_j, V_j\right)^T\right)^T \cdot \mathbf{W}_j \cdot \left(\mathbf{X}_s - \mathbf{A}_j \cdot \left(L_j, V_j\right)^T\right)}{n-2};$$

4: Analysis of residual sums of squares and searches the best values L and V and also their uncertainties.

If the input observations x_k are described by PDF $p_j(x)$ then the expected values $\alpha_{k,j}$ of order statistics are the ideal position (without random fluctuation) of the these sorted observations $x_{(k)}$. Therefore they are named as reference (or ideal) observations. In figure 3 the expected values of $n = 19$ order statistics calculated from normalized ($m = 0, \sigma = 1$) Laplace, normal, uniform and arcsine PDF are shown.

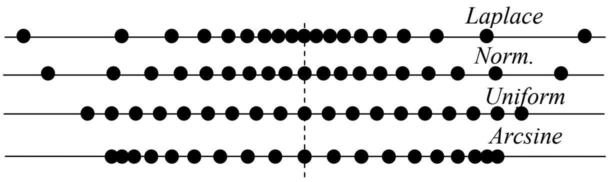

Fig. 3. The expected values α_k of order statistics, $n = 19$ (reference or ideal ordered observations)

Analysis of residual sums of squares

The most important is the last step of proposed algorithm. The values of the residual sums of squares (RSS) $S_{R,j}^2$ are different for the used different models of the PDF $p_1(x), p_2(x),..., p_J(x)$. There are two main reasons of such situation. The first one is caused by the random deviations $\upsilon_{k,j}$ ($k = 1,2,...,n;\ j = 1,2\ ,..., J$) of the ordered observations $\upsilon_{k,j} = \left(x_{(k)} - L_j\right)/V_j$ from their expected (reference) values $\alpha_{k,j}$. The random component $S_{R,j,rnd}^2$ is the first part of $S_{R,j}^2$. The second one is caused by systematic deviations $\varepsilon_{k,j}$ ($k = 1,2,...,n;\ j = 1,2\ ,..., J$) of ordered observations from their expected (reference) values that take place if used model $p_j(x)$ of PDF is different from the true PDF $p_{true}(x)$ of input observations (for example due to the influence of noise). The systematic component $S_{R,j,syst}^2$ is the second part of $S_{R,j}^2$. For example in (8) only the systematic residual sums of squares between reference observations (n = 19) corresponded to the 4 typical normalized model of PDF: Laplace, normal, uniform and arcsine are presented:

(8) $$S_{R,j,k,syst}^2 = \begin{bmatrix} & L & N & U & As \\ L & 0 & 0{,}0159 & 0{,}0693 & 0{,}1179 \\ N & 0{,}0159 & 0 & 0{,}0197 & 0{,}0503 \\ U & 0{,}0693 & 0{,}0197 & 0 & 0{,}0077 \\ As & 0{,}1179 & 0{,}0503 & 0{,}0077 & 0 \end{bmatrix}.$$

From (8) follows, that with an increase of the distance (after contra-kurtosis) between appropriate PDF the systematic component $S_{R,j,syst}^2$ also increase. In real practice both random and systematic components are presented in RSS and (7) can be presented as:

(9) $$S_{R,j}^2 = S_{R,j,rnd}^2 + S_{R,j,syst}^2\ .$$

In figure 4 the typical dependence of the calculated values of the RSS versus appropriate models $p_1(x), p_2(x),..., p_J(x)$ of PDF are presented.

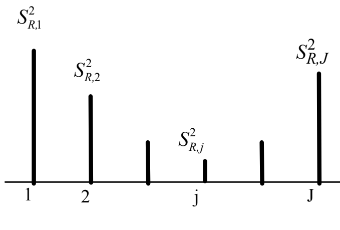

Fig. 4. RSS dependently appropriate models $p_1(x), p_2(x),..., p_J(x)$ of PDF

When one of the used models $p_1(x), p_2(x),..., p_J(x)$ (for example $p_j(x)$) is coincident to the true PDF $p_{true}(x)$ of input observations, then one of the RSS $S_{R,j}^2$ takes the smallest value (fig. 4), because in this case systematic part $S_{R,j,syst}^2$ of RSS is absent in (9). Due to these circumstances the simplest method of parameter location (best measurement result) calculation lies in determination of the number $j = 1, 2,..., J$ if the residual sum of the weighted squares (7) takes the smallest value (Algorithm 1):

(10) $$j = number\left\{ \underset{j=1,J}{MIN}\left[S_{R,1}^2,\ S_{R,2}^2,..., S_{R,j}^2,\ S_{R,J}^2 \right] \right\}.$$

In another case if the $p_{true}(x)$ of input observations is absent between used models $p_1(x), p_2(x),..., p_J(x)$, then systematic part $S_{R,j,syst}^2$ in (9) cannot be neglected. In this situation the weighted estimators of the L and V can be used (Algorithm 2):

(11) $$L = \sum_{j=1}^{J} \varsigma_j L_j\ ; \quad V = \sum_{j=1}^{J} \varsigma_j V_j\ ,$$

where values of weighted coefficients ς_j are inversely to values of appropriate $S_{R,j}^2$:

(12) $$\varsigma_j = \xi_j \Big/ \sum_{j=1}^{J} \xi_j = \frac{1}{S_{R,j}^2} \Big/ \sum_{j=1}^{J} \left(\frac{1}{S_{R,j}^2} \right).$$

For the evaluation of standard uncertainties $u_A(L)$ and $u_A(V)$ of parameters calculated by (12) the procedure of uncertainty propagation of indirect measurements can be used.

Investigation of the efficiency of the proposed method by a Monte Carlo method

Analytical investigation of efficiency of the proposed method for the arbitrary PDF of input sample and reference models $p_1(x), p_2(x),..., p_J(x)$ is very difficult. Therefore this efficiency can be investigated by a Monte Carlo method.

For the previous calculation of reconstructed matrix **REC** and other matrixes the 8 reference models of PDF were used. The first 6 models are based on common exponential density distribution of order r [12]: r = 0,5; r = 1 (Laplace); r = 1,5; r = 2 (normal); r = 4; r = ∞ (uniform). Another two reference models of PDF are arcsine and distribution of sampled in random moments the periodical pulses of exponential sides.

Input observations being investigated are formed using 7 models of PDF (Laplace, normal, convolution of normal and uniform (ratio variance 1:1), triangle, convolution of 2 uniform (width ratio 1:2), uniform and arcsine) with the mean value μ = *5* and standard deviation σ = 0,2. The number of realizations is M=10^5.

For the every $m = 1,2,..., M = 10^5$ realization of each $i = 1,2,...,7$ input observations, for all $j = 1, 2\ ,..., 8$ reference set observations the best estimators

$L_{j,i,m} = \sum_{k=1}^{n} g_{j,i,k} x_{(k),i,m}$ of location and $V_{j,i,m} = \sum_{k=1}^{n} \gamma_{j,i,k} x_{(k),i,m}$ scale parameters and their uncertainties $u_A(L_{j,i,m})$, $u_A(V_{j,i,m})$ are determined. And then the errors $\Delta_{j,i,m} = L_{j,i,m} - \mu$ of the location parameter are calculated and also mean values $\overline{\Delta}_{j,i} = \frac{1}{M}\sum_{m=1}^{M}\Delta_{j,i,m}$, experimental standard deviations $s_{\Delta,j,i} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\Delta_{j,i,m} - \overline{\Delta}_{j,i}\right)^2}$ and maximum of absolute value $\Delta_{j,i,\max} = \max\left|\Delta_{j,i,m}\right|$ of these errors are determined. Using these errors the histograms were built as well as the frequencies (estimators of probabilities) of correct identification of the input sample PDF and acceptance of other nearest PDF.

Results analysis

Some of the obtained results are presented in the next figures. The histograms of the reconstruction errors of the measurement result (location parameter) obtained for *n* = 19 for the Laplace, normal, uniform and arcsine PDF of input observations are shown in fig. 5. The dependencies of the determined mean standard uncertainty of the location parameters from number of observation for these PDF are shown in fig. 6. For the purpose to compare with traditional method the theoretical values of standard deviation of the mean values $s_{\bar{x}} = \sigma/\sqrt{n}$ are shown in this figure by dash line.

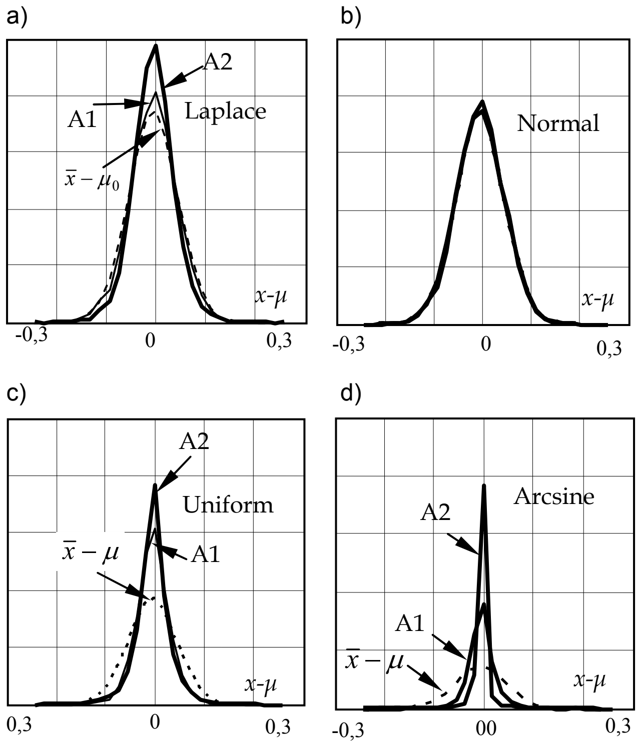

Fig. 5. Histograms of errors sample location parameter (σ = 0,2, *n* = 19, A1 – first algorithm, A2 – second algorithm)

From these histograms (in figure 5) follows that if the distribution of input observations becomes close to the theoretical model of distribution, then errors in the determination of the location parameter decrease (decrease its standard uncertainty), and vice versa. And also from next fig. 6 follows that if the distribution of input observations differs from normal, then the proposed method provides more accurate result with the smaller uncertainty $u_A(L)$ in comparison with the accuracy of mean value – standard uncertainty $\sigma/\sqrt{n}$ (dash curve in fig. 6). This property relates to both algorithms (A1, A2).

But if the PDF of input observations is coincides with one of the used reference models of PDF then algorithm A1 provides better results (less standard deviation of location parameter) in comparison with algorithm A2. On the other hand, if the input sample distribution does not coincide with any reference model, the second algorithm A2 provides better result, because the first algorithm chooses the closest distribution for which the RSS is the smallest and the second algorithm takes an intermediate between the two neighboring distributions.

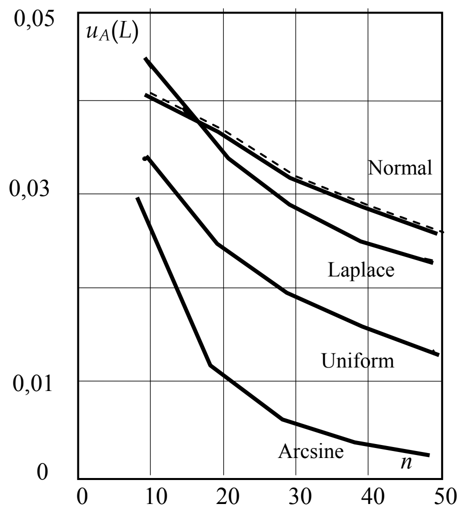

Fig. 6. The dependences of the standard uncertainty of the location parameter of the different PDF of input observations versus number of observation (n), by dash line standard uncertainty of mean value is shown, σ = 0,2

The very important parameter of the quality of the proposed algorithms is the probability of the identification of the true (or nearest of true) PDF of input observations. The dependencies of the frequency (estimator of probability, *P, %*) of the detection true PDF and also frequencies (P3, %) of identification one of 3 PDF: true and nearest left and right PDF versus number of observation *n* are presented in figure 7. From these results follows, that if the number of input observations *n* ≈ ≥ 15 then proposed algorithm correctly chooses the true model of PDF approximately in 50 % and more (fig. 7,a). In other approximately 40 % cases the nearest PDF (left and right after the value of contra-kurtosis) of true PDF are chosen by algorithms, and sum frequencies (P3, %) is more than 90% (fig. 7,b).

If (after the criterion of contra-kurtosis) two PDF are close to each other, then they are characterized by the close values of their position parameters and their uncertainties. Therefore, if the algorithm selects a PDF near to the true distribution (instead exactly the true distribution), it does not lead the catastrophic consequences, i.e., to completely incorrect result, but only some differ from it. If

PDF of input observations differs significantly from normal, then the obtained result will be the best in the most cases in comparison with the average value nevertheless on the not entirely correct (fig 6).

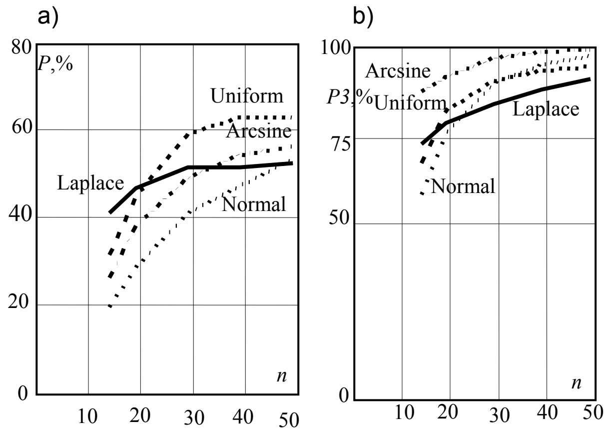

Fig. 7. Frequencies (in %) of identification of appropriate PDF for the distribution of input observations (a) and frequencies (in %) of identification sum of 3 PDF: true and nearest left and right PDF (b) versus number of observation (n)

Obviously, it is true that with the increase of number of input observations quality of choice of the correct distribution becomes better (fig. 7). For the increase of efficiency of proposed method in practice it is necessary to increase the number of models of PDF therefore they would be arranged on the contra excess axis more tightly.

Conclusion

1. The possibility of obtaining of the random observations parameters estimations close ones to the optimum under conditions of the absence of a priori information about their distribution is the positive feature of the proposed method.

2. From the theoretical point of view proposed method of measurement observations processing is based on their comparison with the set of sequences of the reference observations that are equal to the expectation of order statistics corresponded with the assigned PDF. But from the practical point of view the proposed method is realized by a very simple procedure: at the first the sorting of input observations and the next steps they are averaged by the set of previously calculated weight coefficients, corresponded with the assigned models of PDF.

3. From the obtained results it follows that if the number of the input observations is approximately $n \approx> 15$ then the proposed algorithm provides smaller value of the measurement result standard uncertainty in comparison with the standard uncertainty of the mean value, which is recommended by the Guide to the Expression of Uncertainty in Measurement. In ≈ 50 % cases the proposed method detects the true model of distribution correctly and in approximately in ≈ 90…95% and more cases the true distribution or nearest (after the contra-kurtosis) PDF to it are detected.

4. The effectiveness of the proposed algorithm grows with an increase of the deviation between of the real distribution of input observations and normal distribution.

5. The quality of proposed method can be improved by using of the large numbers of reference observations corresponding to the reference PDF that are closely located in the contra-kurtosis axle as well as by the more complicated procedure of residual sums of squares analysis using.

REFERENCES

[1] GUM Guide of the Expression of Uncertainty In Measurement. ISO, Switzerland (1995)

[2] Dorozhovets M., Some Problems of Calculating the Errors and Uncertainty of Measurement Results in Practice. *Folia Scientiarum Universitatis Technicae Resovientis* (*Rzeszow Uniwersity of Technology)*. *Elektrotechnics* N 233, (2006), T.29. 9-44 (In Polish)

[3] Dorozhovets M., Warsza Z. L., Influence of the inadequate choice of distribution parameters on the uncertainty evaluated by method type A. *Pomiary, Automatyka, Kontrola* (*Measurements, Automatics, Checking*), vol. 53(2007, nr 9 bis, 25-28 (In Polish)

[4] Lehmann, E., Theory of Point Estimation, John Wiley & Sons, New York (1983)

[5] Cramer H., Mathematical methods of statistics. Uppsala: Almqvist & Wiksells, (1945)

[6] Fisz M., Probability Theory and Mathematical Statistics. John Willey & Sons, London (1963)

[7] Lloyd E.H., Least-squares estimation on location and scale parameters using order statistics. *Biometrika*, 39 (1952), 39-88

[8] Downton F. A note of ordered least-squares estimation. *Biometrika*, 40 (1953), 457

[9] Kendal M.G., Stuart A. The Advanced Theory of Statistics. Charles Griffin & Co. Ltd. Third edition, London, (1973)

[10] Dorozhovets M., Investigation of the Test Samples Method, Used for the Evaluation of Measurement Result and its Uncertainty. In Proc. of International Conference on Precision Measurement (2008). TU Ilmenau, 91-92

[11] Dorozhovets M., Kochan O., Estimation of the best measurement result and its standard uncertainty by input observations processing using the method of reference samples based on order statistics. Proceedings of the 5-th IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications. (2009), Rende (Cosenza), Italy, 351-354

[12] Novitski P.V., Zograf I.A., Estimation of measurement result errors. Energoizdat. (1985) (In Russian).

***Autor**: Prof. dr hab. inż. Mykhaylo Dorozhovets, Politechnika Rzeszowska, Katedra Metrologii i Systemów Diagnostycznych, ul. Powstańców Warszawy 12, 35-959 Rzeszów, E-mail: michdor@prz.edu.pl.*