

# EFFECTIVE MEASURAND ESTIMATORS FOR SAMPLES OF TRAPEZOIDAL PDFs

Submitted 5<sup>th</sup> December 2010; accepted 22<sup>nd</sup> March 2011

Zygmunt Lech Warsza

## Abstract:

*This paper is final overview of investigations on the accuracy of basic estimators of trapezoidal probability distribution samples of the measured data. For symmetrical trapezoidal PDF of straight as well concaved sides, using Monte-Carlo method of simulation, the standard deviation (SD) of linear 1- and 2-component estimators are evaluated. Approaches for theirs evaluation are proposed. It is established that in the ratio of upper and bottom bases of trapezoidal PDF in the range from 1 to 0,35 the mid-range value has smaller standard deviation (SD) than the mean value and median. It is find then for the whole family of the symmetric linear trapezoidal PDF more accurate than above single element estimators are two-component (2C) estimators as the linear form of the mean and mid-range values of the sample. Their coefficients are found, properties discussed and formulas of SD are given. The new simplified 2C-estimator of equal coefficients is also proposed. These estimators successfully extend estimation of the measurand value as the sample mean and description of its accuracy by the uncertainty type A recommended by the international guides of uncertainty evaluation in measurement GUM-2008 [1], EA-4/02 [2] and by Handbook NASA [3]. Approaches of described below investigations could be effectively applied also for other models of convoluted PDF-s.*

**Keywords:** *estimators of probability density function, trapezoidal PDF, mid-range, uncertainty evaluation*

## 1. Introduction

Random components of measurement data can be in many cases more accurately modelled by non-Gaussian probability density distribution function (PDF) than by Normal distribution as the range of data random dispersion is commonly limited in reality. The mean value as the most effective measurand estimator of the  $n$ -element sample of Normal distribution is also used for other distributions. Its standard deviation (SD) is defined in GUM [1] as the uncertainty type A.

For data processing it is very important to choose an effective estimator of the centre coordinate of PDF, i.e. estimator of the smallest SD, as not proper evaluation entails incorrect assessment of the measurement accuracy.

For samples modelled by Normal, Uniform and Laplace (double-exponential) PDF distributions, it is presented in the paper [4] of 15<sup>th</sup> IMEKO TC4 Symposium in Iasi Romania, how to regard the data autocorrelation and which estimator has the smallest standard deviation (SD) to be chosen as the better accurate for any of them.

E. g. more effective estimator than mean value of measurand of Uniform samples is mid-range and for Laplace sample – median, respectively. Using one of goodness-of-fit tests (Kolmogorov–Smirnov, Cramér–von Mises, Chi-Square and other tests) we make decision about the estimation choice.

The main purpose of this work is the expansion of opportunities for choosing the best single or a few component estimators of empirical data modelled by more complex non-Gaussian distributions than the above models. It is assumed that treated measurement data do not contain unknown systematic errors and are not self-correlated. The estimator of the distribution parameter should meet also requirements of solvency, sufficiency, efficiency and be unbiased. First of all, efficiency of estimators is researched.

## 2. Single component estimators

Let's check up which one of single-component estimators of PDF of particular samples: mean  $\bar{X}$ , mid-range  $q_{V/2}$  or median  $X_{med}$ , satisfies the requirement of efficiency, i.e. has the least-possible sum of the square dispersion, denotes a minimum standard deviation in comparison with other estimators. Similarly, it is possible to receive results for other basic non-Gaussian distributions. In columns 3–5 of Tab. 1 values of standard deviations of three estimators of a few basic distribution models of empirical data (for demonstration of difference order only) are presented.

Standard deviation of the best single component estimator of the particular non-Gaussian distribution is significantly less than of other estimators even if difference between their values, e.g. between midrange and mean, is small. This is the cause to search for estimators better than the sample mean.

## 3. The best single component estimators of trapeze distributions

### 3.1. Linear trapeze

It is important to consider the problem of choice of an effective estimator for composition of simple distributions. In the measurement systems practically all analogue signals now are digitalised, and then uniform distributions are very common in these systems. So, with convolution of two different uniform distributions we get PDF as a symmetrical trapezoid of linear sides, from triangular to the uniform distribution as its boundary cases. The effective single component estimators of the centre of the triangular and uniform distributions are the sample mean and the mid-range respectively – see again Table 1.

Table 1. Comparison of sufficiency of different estimators and expression of the standard uncertainty

Distribution	Standard deviations of sample estimators			The most effective estimator	Standard uncertainty of the most effective estimator
	$S_{mean}$	$S_{midrange}$	$S_{med}$		
Normal	<b>0,010</b>	0,220	0,013	sample mean	$u_A = S_x / \sqrt{n}$ [1]
Uniform	0,006	<b><math>1,4 \cdot 10^{-4}</math></b>	0,010	mid-range	$\frac{V}{\sqrt{2} (n-1)} \sqrt{\frac{n+1}{n+2}}$ [4]
Double-exponential	0,007	0,870	<b><math>7 \cdot 10^{-5}</math></b>	median	$S_x / \sqrt{2n}$ [4]
Triangular	<b>0,0040</b>	0,0045	0,0049	sample mean	$S_x / \sqrt{n}$ [3] - [5]
Arcsine	0,067	<b><math>5 \cdot 10^{-5}</math></b>	0,146	mid-range	$S_x \cdot \sqrt{5\pi^4} / n^2$ [5]

The aim of the following research is to find a position of border separating trapezoids of better mid-range or mean values. There are two ways to obtain the trapezoid in MC simulations :

- to generate two uniform distributions and theirs sum [9];
- to use the inverse function method (derived in [9] for trapezoid).

Both techniques were tested. Samples from population with trapezoidal distribution with  $\beta = a/b$  ratio of their shorter upper  $a$  and longer bottom  $b$  basis are simulated and stable results are obtained. Obviously  $\beta \in (0; 1)$  was taken. Fig. 1 shows how standard deviations of mean and mid-range are changed with a ratio  $\beta$  and number of ob-

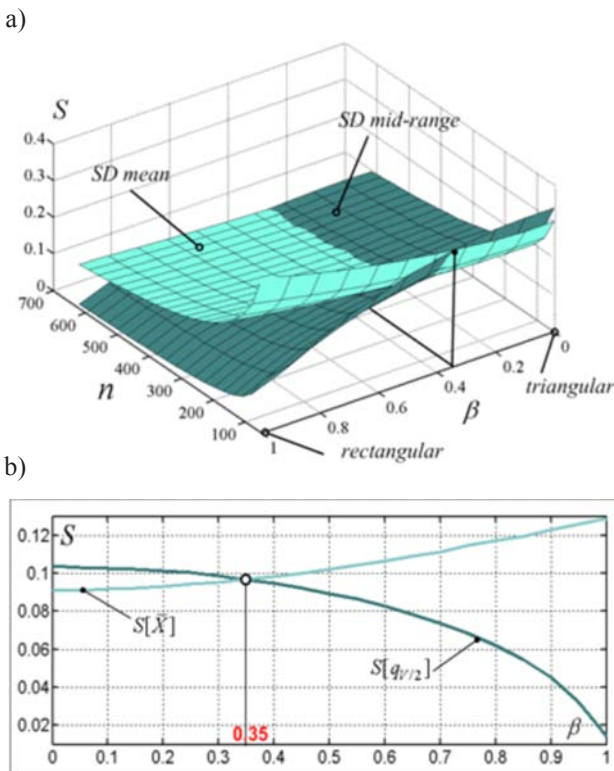


Fig. 1. Efficiency of single component sample estimators of  $Trap(a,b)$  distributions [10]: a. Dependences of sample mean and midrange standard deviations  $S$  on ratio  $\beta$  of linear trapeze bases and of sample size  $n$ , b. cut-set of  $S$  surfaces for  $n=const.=400$

servations  $n$  [11]. Median SD is significantly larger and is not shown on fig 1.

Border value  $\alpha$  of ratio  $\beta$  has been found for us also analytically by young mathematician P. Endovitskyi from TU Kiev. He obtained

$$\alpha = \frac{-10 + 3\pi + \sqrt{9\pi^2 - 72\pi + 140}}{14 - 3\pi - \sqrt{9\pi^2 - 72\pi + 140}} \approx 0,3546 \quad (1)$$

Novitzky and Zograph in their original book [6] show dependence of estimator (mid-range) efficiency on a type of distribution. Topographical classification of distributions is also offered and dependence of the estimator efficiency on the counter-kurtosis  $\alpha$  is presented. Variances of estimators are equal when  $\alpha = 0,675$ . This value of  $\alpha$  corresponds to kurtosis  $E = 1/\alpha^2 = -0,805$ . For Normal PDF  $E = 3$ .

Dependence of kurtosis differences  $E-3$  from Normal PDF on ratio of trapezium bases  $\beta$  are given on Fig. 2. Then we can find that  $E = -0,805$  corresponds to  $\beta = 0,35$ .

### 3.2. Curvilinear trapeze

In Table 1 of GUM Supplement 1 [2] the curvilinear trapezoidal of concave sides is given. This PDF model has the symbol  $CTrap(a,b,d)$ . It is proposed to be used when limits of upper  $a$  and lower  $b$  sides are inexactly given, i.e.  $a \pm d$  and  $b \pm d$ , where  $a, b$  and  $d$ , with  $d > 0$  and  $a + d < b - d$ , are specified. Histogram of these type simulated data is given on Fig. 3.

Fig. 4a, b shows how standard deviations of main estimators depend on the number  $n$  of observations in the sample and a ratio  $\beta_c = (a_2 - a_1 - 2d)/(b_2 - b_1 + 2d)$  of curvilinear-

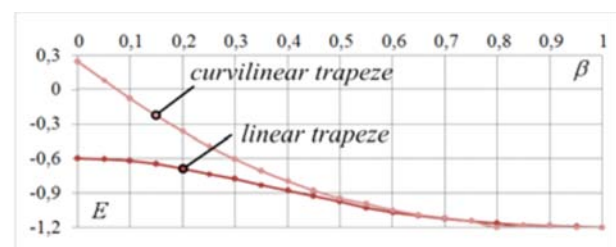


Fig. 2. Kurtosis differences  $(E-3)$  of trapezoid and Normal PDFs as function of ratio  $\beta$  of trapeze bases

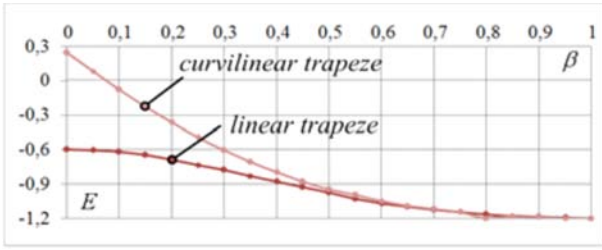


Fig. 3. Example of curvilinear trapezoid PDF

ear trapezoid basis. It is shown that median here is the best single component estimator if  $0 < \beta_c < 0,08$ ; mean – if  $0,08 < \beta_c < 0,5$  and mid-range if  $0,5 < \beta_c < 1$ . But, it should be taken into account, that in practice, uncertainty of uncertainty may be limited up to even 20–30%, then:

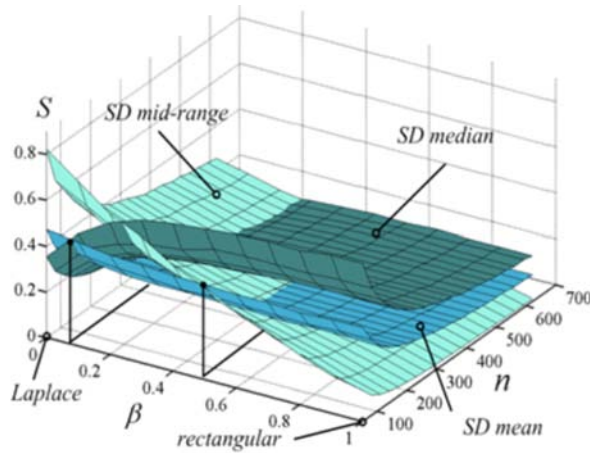
$$d = \frac{(b-a)}{2} \cdot \frac{(1-\beta_c)}{(1+\beta_c)} = 0,3 \frac{(b-a)}{2} \Rightarrow \beta = 0,54 \quad (2)$$

and could be decided that the mid-range may be applied as the most effective estimator to the border drawn in Fig. 4.

To increase accuracy of the measurement result other types of estimators, which contain a few components, may be also considered.

According to considered approaches, ratio of these components could be found by modelling and selection

a)



b)

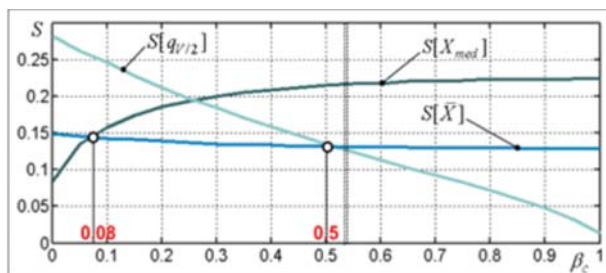


Fig. 4. Efficiency of single component estimators of CTrap(a,b,d) distribution: a) Dependences of SD on a ratio of bases  $\beta$  and on sample size  $n$  of the curvilinear trapezoidal PDF, b) visualization of crossing points for  $n=const=400$

of best values, or by known analytical equations. These equations are derived by numeric methods too and they based on shape coefficients or parameters of the distribution model.

#### 4. Multi-component estimators of trapeze distribution

##### 4.1. Three- and two-component estimators based on kurtosis E value

Zakharov and Stephen in [7, 8] considered for non-Gaussian symmetrical PDF the linear 3-component (3C) estimator of measurand value:

$$\hat{X} = k_1 \bar{X} + k_2 q_{V/2} + k_3 X_{med} \quad (3)$$

as the efficient estimate of the expectation. Coefficients  $k_1, k_2$  and  $k_3$  depend on the kurtosis  $E$  of the distribution of observation results.

For linear trapezoids of  $E \in (-1,15; -0,2)$  only two such coefficients are enough [7]:

$$k_1 = -1,05E + 1,22, \quad k_2 = -0,05E - 0,22, \quad k_3 = 0 \quad (4)$$

Modelling shows that such proposed estimator is biased [10] and it is not consistent with requirements of the effective estimator. For unbiased estimator the sum of all three coefficients must be equal to 1. From MC investigations [10, 11]

$$k_1 = -1,05E + 1,22, \quad k_2 = 1,05E - 0,22, \quad k_3 = 0 \quad (5)$$

Standard deviations of  $\bar{X}, q_{V/2}$  and 2C estimator  $\hat{X}$  corrected due (2) for linear trapezes of different  $\beta$  are given in Fig. 5.

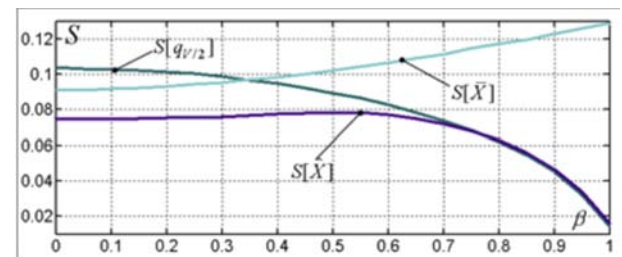


Fig. 5. Dependences of standard deviations for different statistics on a ratio of bases (linear trapeze)

In case of the curvilinear trapeze its kurtosis  $E \in (-1,2; 0,2)$  and following coefficients have been obtained

$$k_1 = \begin{cases} 0, & \text{if } E \leq -1,15; \\ 1,05 \cdot E + 1,22, & \text{if } -1,15 < E \leq -0,2; \\ 1, & \text{if } E \geq -0,2; \end{cases}$$

$$k_2 = 1 - k_1 = \begin{cases} 1, & \text{if } E \leq -1,15; \\ -1,05 \cdot E - 0,22, & \text{if } -1,15 < E \leq -0,2. \\ 0, & \text{if } E \geq -0,2; \end{cases} \quad (6)$$

$$k_3 = 0.$$





where:

$$S[\bar{X}] = \sigma_x / \sqrt{n}, \sigma_x$$

is the SD of whole population.

Coefficient  $k_1$  could be find from

$$\frac{\partial S^2[X_{eff}]}{\partial k_1} = 0$$

After calculations:

$$k_1 = \frac{S^2[q_{V/2}]}{S^2[\bar{x}] + S^2[q_{V/2}]}, \quad (13)$$

### 5.2. Particular cases

**For triangular distribution ( $\beta=0$ )** [5], [6]:

$$S^2[q_{V/2}] = \frac{3 \cdot (4 - \pi)}{2n} \sigma_x^2,$$

$$k_1 = \frac{3(4 - \pi)}{2 + 3(4 - \pi)} \approx 0,56.$$

It coincides with results of the earlier MC simulation.

**For trapezoid with  $\beta = 0,35$**  we find that, and from (6):

$$k_1 = \frac{\sigma_x^2}{n} / \frac{2\sigma_x^2}{n} = 0,5.$$

It coincides with (10).

**For rectangular distribution ( $\beta=1$ ):**

$$k_1 = \frac{\frac{3\sigma_x^2}{2(n+1)(n+2)}}{\frac{\sigma_x^2}{n} + \frac{3\sigma_x^2}{2(n+1)(n+2)}} = \frac{3n}{2n^2 + 9n + 2}.$$

If  $n \rightarrow \infty$ ,  $k_1 \rightarrow 0$ . For  $n = 30$ ,  $k_1 = 0,04$  and  $n=10$ :  $k_1=0$ , so these results are not very far from the above  $k_1=0$  for  $n \rightarrow \infty$ .

Dependences of SD on  $k_1$  for boundary cases of trapezium shape (triangular and rectangular PDF) are shown in Fig 8.

It is natural that the triangular distribution is not exactly like Normal PDF, but an intermediate one, between Uniform and Normal. So it's the best estimator consisting also of both components.

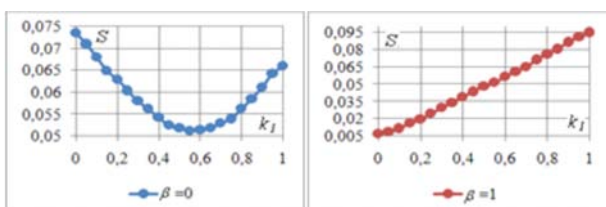


Fig. 8. Dependences of standard deviations on  $k_1$

### For simplified two-component estimator of (6)

The standard uncertainty (equivalent to  $u_A$  in GUM) is:

$$u_A = S[X_{eff}] = \frac{1}{2} \sqrt{S^2[\bar{X}] + S^2[q_{V/2}] + 2\rho S[\bar{X}]S[q_{V/2}]} = \frac{1}{2} \sqrt{\frac{S_x^2}{n} + \frac{V^2(1-\beta^2)}{16} \cdot \frac{n}{(n+1)(n+2)} + 2\rho \frac{S_x V \sqrt{(1-\beta^2)}}{\sqrt{(n+1)(n+2)}}}. \quad (14)$$

If  $\rho \rightarrow 0$

$$u_A = S[X_{eff}] = \frac{1}{2} \sqrt{S^2[\bar{X}] + S^2[q_{V/2}]} \quad (14a)$$

As standard deviation of the proposed estimator is used the standard uncertainty, we should give expressions for coverage factor  $k(P)$  to expanded uncertainty calculation.

The equation for the large sample size is [11, 12]:

$$k(P) = \sqrt{\frac{6}{1+\beta^2}} \left(1 - \sqrt{(1-P) \cdot (1-\beta^2)}\right) \quad (15)$$

### 6. Numerical Example

Considerations has to be illustrated below by the numerical example of measurand value and uncertainty calculations. Data values of the sample size  $n=200$  obtained in simulated experiment are shown in Fig. 9. As no other information is available then should be presume that this observations are not autocorrelated and cleaned before from systematic errors.

Let's find the measurement result as the best estimator of measurand value, its standard and expanded uncertainties. The proper PDF model of this sample has to be chosen.

Sample observations are arranged into 15 groups (Fig. 10).

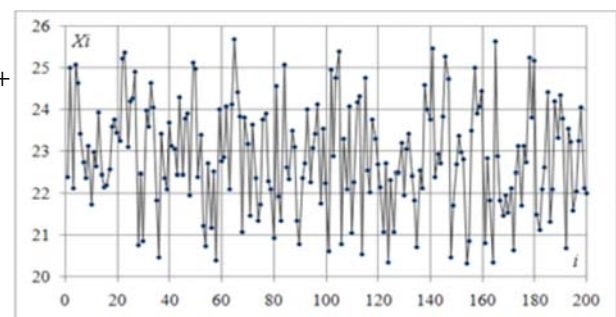


Fig. 9. Values of sample observations.

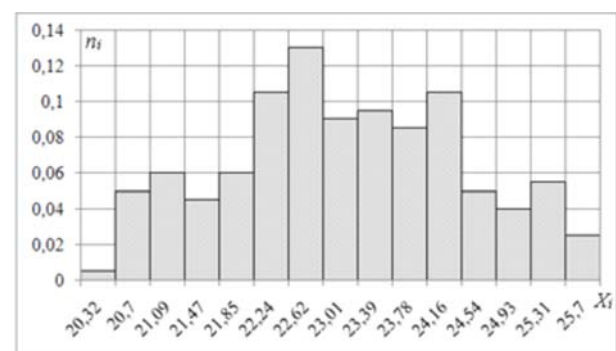


Fig. 10. Histogram of data relative frequencies

Hypothesis about compliance with three different theoretical distributions are verified by  $\chi^2$  test.

Number of freedom is 11. Compliance with Uniform and Normal distributions is not fulfilled, but with linear trapezoidal distribution is accepted at significance level 0,05, because:

$$\chi^2 = 17,3 < \chi_{11, 0,05}^2 = 19,7 .$$

The trapezoid PDF model of 5.38 and 1.79 bases are found. Its parameter  $\beta=1/3$ . As the best estimator of the measurand value is used (4). Values of distribution parameters are:

$$\bar{X} = 22,873 , q_{V/2} = 23,010 , \tilde{X} = 22,942 .$$

Sample standard deviation:

$$S_X = 1,309 .$$

Standard deviation of the mean:

$$S[\bar{X}] = S_X / \sqrt{n} = 0,0926 .$$

Standard deviation of the mid-range:

$$S[q_{V/2}] = \frac{V}{4} \cdot \sqrt{\frac{n \cdot (1 - \beta^2)}{(n+1)(n+2)}} = 0,089 .$$

Standard uncertainty of the 2-component estimator is

$$u_A = \frac{1}{2} \sqrt{S^2[\bar{X}] + S^2[q_{V/2}] + 2 \cdot 0,2 \cdot S[\bar{X}] \cdot S[q_{V/2}]} = 0,0703 .$$

$$= 0,0703 .$$

The value of uncertainty for estimator (5) does not differ significantly from above.

Distributions of  $q_{V/2}$  and  $\tilde{X}$  for trapeze pdf are unknown but expected to be smoother than Normal one. For these estimators is taken the same coverage factor as for normal pdf, i.e.:  $K(P=0,95)=1,96$ .

For coverage probability  $P$  expanded uncertainty is:

$$U(P) = K(P) \cdot u_A \quad (16)$$

Results are put together in Table 3.

Table 3. Representations of the measurement result and accuracy

	By standard uncertainty	by expanded uncertainty
$\bar{X}$	$X = 22,87; u_A = 0,09$	$X = (22,87 \pm 0,19), P = 0,95$ $X \in (22,68; 23,06), P = 0,95$
$q_{V/2}$	$X = 23,01; u_A = 0,09$	$X = (23,01 \pm 0,18), P = 0,95$ $X \in (22,83; 23,19), P = 0,95$
$\tilde{X}$	$X = 22,94; u_A = 0,07$	$X = (23,01 \pm 0,14), P = 0,95$ $X \in (22,87; 23,15), P = 0,95$

The most accurate is the last one – simplified 2C estimator  $\tilde{X}$ . Values of each estimator are lying in the expanded uncertainty ranges of two others.

## 7. Final conclusions

- It is very important to choose the most accurate, i.e. effective estimator at data processing for correct estimation of the measurand uncertainty corresponding to  $u_A$  (type A).
- For samples of distributions modelled by trapezoid, the best single-component estimator depends on its shape. If it is nearer to rectangular ( $1 \geq \beta \geq 0,35$ ) then the best effective estimator of measurand is the mid-range. Below  $\beta=0,35$  up to  $\beta=0$  of the triangle distribution, the sample mean is better.
- The 2-component estimator as the linear form of above two estimators is better for samples of trapezium PDF.
- For the broad range of trapezium shapes ( $0,75 \geq \beta \geq 0$ ) the simplified form of this double component estimator of equal both coefficients  $k_1 = k_2 = 0,5$  is proposed and may be used with sufficiently good accuracy acceptable in practice.
- For a number of sample observations  $n \geq 10$  all coefficients are practically independent from  $n$ . For smaller size  $n < 10$  individual modelling is needed for trapezium PDF.
- All conclusions are positively tested by MC simulations and also by several numerical examples.
- Estimators of trapezoidal distributions given in this work could be applied not only in measurement practice and for extending of GUM, NIST and NASA recommendations [1] – [3], [9] but also in the statistics, when trapezoidal models are also used [8].

One could forecast that way to obtain two-component measurand estimators for samples modelled by convolution of other two distributions such as Uniform and Normal, Uniform and arcsine, etc. may be interesting.

## ACKNOWLEDGMENTS

Author wishes to express his many of thanks to Maryna Galovska M.Sc., now a scientific assistant at the Institute of Manufacturing Metrology (Institut für Produktionsmesstechnik), Technical University of Braunschweig Germany, which in the years 2007 - 2009 as PhD student in Ukrainian National Technical University „Kiev Polytechnic Institute“ on their own choice cooperated with the author on the above issues and gave a lot of her own initiative for obtaining the results of this work by Monte Carlo simulation [10-12]. Thanks to that a number of intuitive ideas of the author about trapezoidal distributions are checked and new one- and two-component estimators for these distributions are established.

## AUTHOR

Zygmunt Lech Warsza – Industrial Institute of Control and Measurement PIAP, Warsaw, Poland,  
E-mail: zlw@wp.pl.

## References

- [1] *Evaluation of measurement data - Guide to the expression of uncertainty in measurement (GUM)*, BIPM, JCGM 100, (Ed. 1993 –2008), and Supplement 1 Propagation of distributions using a Monte Carlo method. Guide OIML G1-101, 2007.
- [2] EA-4/02 • Expression of the Uncertainty of Measurement in Calibration, EA European Cooperation for Accreditation, December 1999, pp. 63-65.
- [3] *Measurement Uncertainty Analysis Principles and Methods, NASA Measurement Quality Assurance Handbook –Annex 3*, HDBK-8739.19-3, July 2010 Washington DC.
- [4] Dorozhovets M., Warsza Z., “Methods of upgrading the uncertainty of type A evaluation (2). Elimination of the influence of autocorrelation of observations and choosing the adequate distribution”. In: *Proceedings of 15<sup>th</sup> IMEKO TC4 Symposium*, Iasi, pp. 199-204.
- [5] Johnson N. L., Leone F. C., *Statistics and experimental design in engineering and physical sciences*, vol.1, 2<sup>nd</sup> ed., John Wiley & Sons, New-York, 1977.
- [6] Novickij P.V., Zograf I.A., *Ocenka pogreshnostej rezultatov izmerenii (Estimation of the measurement result errors)*, Energoatomizdat, Leningrad, 1985 (in Russian only).
- [7] Zakharov I.P., Shtefan N.V., ”Algorithms for reliable and effective estimation of type A uncertainty”, *Measurement Techniques*, vol. 48, 5, 2005, pp.427-437, www. Springer com. (transl. from *Izmeritel'naja Tekhnika* no 2, 2005 p. 9-15)
- [8] Van Dorp J.R., Kotz S., “Generalized Trapezoidal Distributions”, *Metrika*, vol. 58, Issue 1, July 2003.
- [9] Kacker R. N., Lawrence J. F., “Trapezoidal and triangular distributions for Type B evaluation of standard uncertainty” *Metrologia*, no. 44, 2007, pp. 117–127.
- [10] Warsza Z. L., Galovska M., “About the best measurand estimators of trapezoidal probability distributions”, *Przegląd Elektrotechniki (Electrical Review)*, no. 5, 2009, pp. 86–91.
- [11] Warsza Z. L., Galovska M., “The best measurand estimators of trapezoidal PDF”. In: *Proceedings of IMEKO World Congress Fundamental and Applied Metrology*, 2009, Lisbon, CD, pp. 2405–2410.
- [12] Galovska M., Warsza Z. L., The ways of effective estimation of measurand”, *PAKgoś (Pomiary Automatyka Komputery w gospodarce i ochronie środowiska)*, no.1, 2010, pp. 18-20.