IMEKO World Congress
Fundamental and Applied Metrology
September 6–11, 2009, Lisbon, Portugal

THE BEST MEASURAND ESTIMATORS OF TRAPEZOIDAL PDF

Warsza Zygmunt Lech [1], *Galovska Maryna* [2]

[1] Polish Society of Metrology, Warsaw, Poland, zlw@wp.pl
[2] National Technical University of Ukraine "Kyiv Polytechnic Institute", galmv@ukr.net

Abstract - The basic estimators of trapezoidal probability distributions of the measured data are considered and approaches for theirs evaluation are proposed.. For symmetrical trapezoidal PDF of straight as well curved sides, using Monte-Carlo method of simulation, the standard deviation (SD) of some chosen 1- and 2-component estimators are evaluated It is established that in the ratio of upper and bottom bases of trapezoidal PDF in the range from 1 to 0,35 the most accurate is the mid-range value. Below this range smaller is the standard deviation (SD) of the mean value. The best for the whole family of the symmetric linear trapezium PDF 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 in the international guide ISO GUM [1]**.** Approaches of investigation could be effectively applied for some other models of PDF.

Keywords: estimators of probability density function, trapezoid 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) [2] than by Normal distribution as the range of their values is commonly limited. The mean value is the most effective measurand estimator of the *n*-element sample of Normal distribution. Its standard deviation (SD) is defined in GUM [1] as the uncertainty type A. For samples modelled by Normal, Uniform and Laplace (double-exponential) PDF distributions, it is presented in the paper [3] and then in [4] of 15th IMEKO TC4 Symposium, how to regard the data autocorrelation and which estimator has the smallest standard deviation (SD) to be chosen as the best one.

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

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 single models. It is assumed that measurement data do not contain unknown systematic errors and are not self-correlated.

The estimator of the distribution parameter should meet the requirements of solvency, sufficiency, efficiency and be unbiased [5]. First of all, efficiency of estimators is researched.

2. THE SYSTEMATIZATION OF APPROACHES TO THE BEST ESTIMATOR CHOOSING

Systematization of the main approaches to a problem of the effective estimation consists of:

– Monte-Carlo simulation of empirical distribution function and its testing;
– Resampling methods;
– Shape coefficient application method;
– Based on goodness-of-fit test and information about estimators for certain population.

The Monte-Carlo technique based on method of inverse function could be implemented by the next ways:

– evaluation of analytical expression for inverse empirical cumulative function and simulation from uniformly distributed sample by method of inverse function;
– the mixture algorithm is carried out simulation on each of intervals for grouped data (histogram).

Resampling methods, that are so popular in statistics recent years, include jackknife method and different techniques of bootstrap method.

The next approach based on dependency of estimators SD ratio on the kurtosis value [6-8].

The last approach is prevalent in metrological practice. It is provided by theoretical models of population distributions, and theirs known features. For example, a more effective estimator of measurand of Uniform samples is mid-range and for Laplace sample - median, respectively [3, 4]. Using one of goodness-of-fit tests (Kolmogorov–Smirnov , Cram´er–von Mises, Chi-Square and other tests) we make decision about the best estimation choice.

All considered methods could be applied in software packages for measurement data processing and measurement uncertainty evaluation, that are popular last years.

3. SINGLE COMPONENT ESTIMATORS

Let's check up which one of single-component estimators of PDF of particular samples: mean $\overline{X}$, 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 then 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 then sample mean.

4. THE BEST SINGLE COMPONENT ESTIMATORS of TRAPEZE DISTRIBUTIONS

4.1. Case of 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 estimators of the distribution centre of the triangular and uniform distributions are the sample mean and the mid-range respectively – see again Table 1.

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 [2];
- to use the inverse function method (derived in [10]).

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 β and number of observations n. Median SD is significantly larger and is not shown on fig .1.

The crossing of surfaces on β=0,35 is observed. It is independent from n. The same result is obtained by convolution of two rectangular distributions with the ratio of ranges about 2,05 (it corresponds to the above β value). This value of æ corresponds to kurtosis $E=1/æ^2=-0{,}805$. Dependence of E on ratio β of trapezium bases β is given on Fig. 2. Then we can find that E=-0805 corresponds to β=0,35.

It is possible to make a conclusion, that at a ratio $\beta > 0{,}35$ an effective estimator is the sample mid-range; for $\beta < 0{,}35$ – the sample mean. The result remained stable for various sample volumes from 20 to 10 000. Novitzky and Zograph in their original book [6] show dependence of estimator (mid-range) efficiency on a type of distribution. Variances of estimators are equal when counter-kurtosis æ =0,675.

a)

b)

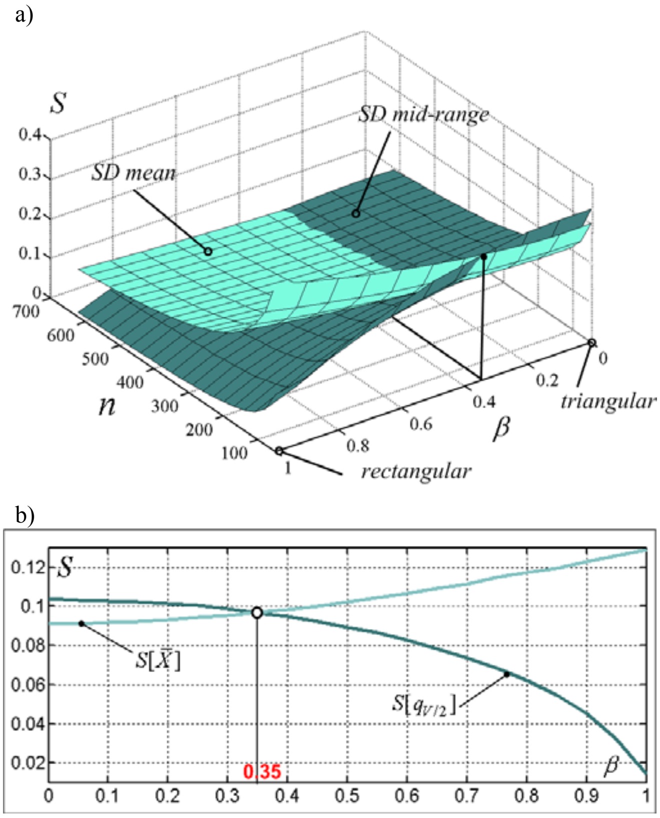


Fig. 1. Dependences of sample mean and midrange standard deviations S on ratio β of linear trapeze bases and of sample size n

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 best estimator
	S_{mean}	$S_{midrange}$	S_{med}		
Normal	**0,010**	0,220	0,013	sample mean	$u_A = S_x/\sqrt{n}$ [1]
Uniform	0,006	**1,4·10^-4**	0,010	mid-range	$\frac{V}{\sqrt{2}\,(n-1)}\sqrt{\frac{n+1}{n+2}}$ [3],[4]
Double-exponential	0,007	0,870	**7·10^-5**	median	$S_x/\sqrt{2n}$ [3],[4]
Triangular	**0,0040**	0,0045	0,0049	sample mean	$S_x/\sqrt{n}$ [3], [6]
Arcsine	0,067	**5·10^-5**	0,146	mid-range	$S_x \cdot \sqrt{5\pi^4/n^2}$ [6]

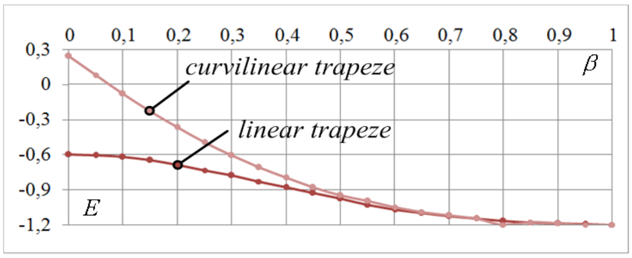


Fig. 2. Coefficient of kurtosis E as function of ratio of bases β

4.2. Case of 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.

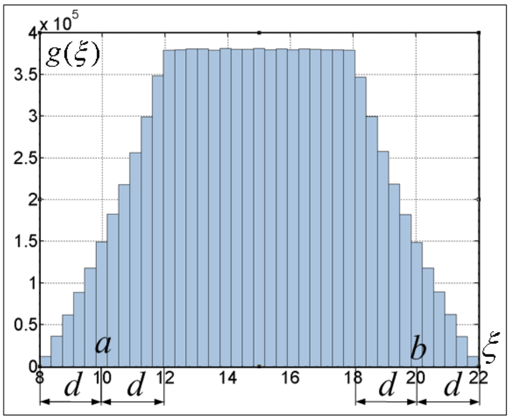


Fig. 3. Example of curvilinear trapezoid PDF

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 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:

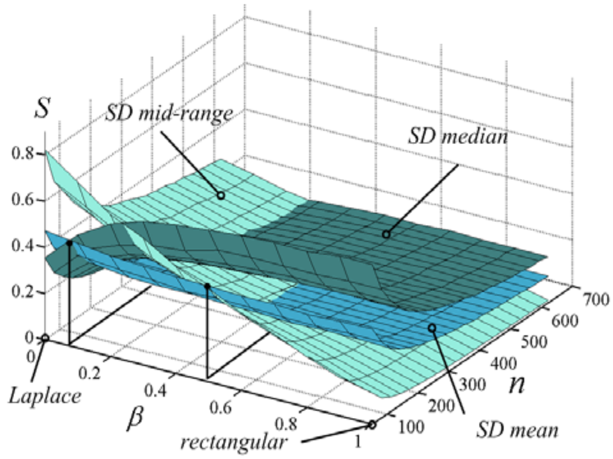
$$d = \frac{(b-a)}{2}\cdot\frac{(1-\beta_c)}{(1+\beta_c)} = 0{,}3\frac{(b-a)}{2} \Rightarrow \beta = 0{,}54$$

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

a)



b)

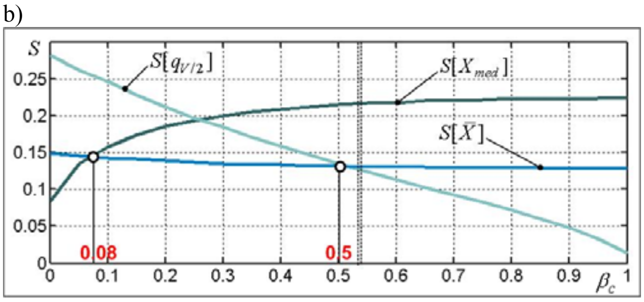


Fig. 4. a) Dependences of single component estimators on a ratio of bases β and on sample size n of the curvilinear trapezoidal PDF, b) visualization of crossing points

5. MULTI-COMPONENT ESTIMATORS OF TRAPEZE DISTRIBUTION

5.1. Two- and three-component estimators based on kurtosis value

Zakharov and Stephen in [7, 8] offer the 3-component estimator of measurand value for non-Gaussian PDFs

$$\hat{X} = k_1\bar{X} + k_2 q_{V/2} + k_3 X_{med} \qquad (1)$$

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, when

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

$$k_1 = -1{,}05E + 1{,}22\,,\; k_2 = -0{,}05E - 0{,}22\,,\; k_3 = 0\,.$$

Our modelling shows that proposed estimator is biased. So, it is not consistent with the requirements of the best estimator. For unbiased estimator the sum of all three coefficients must be equal to 1. From MC investigation:

$$k_1 = -1{,}05E + 1{,}22\,,\; k_2 = 1{,}05E - 0{,}22\,,\; k_3 = 0\,. \qquad (2)$$

Standard deviations SD of estimators for linear trapezes of different β are presented on Fig.5.

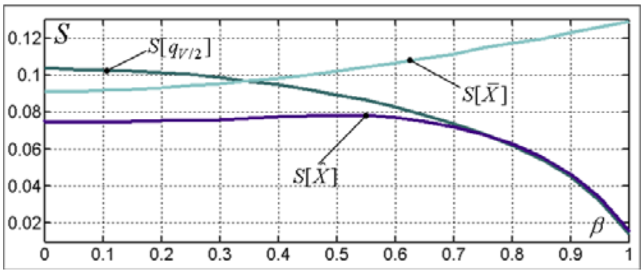


Fig. 5. Dependences of standard deviations for different statistics on a ratio of bases (linear trapeze)

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

$$k_1 = \begin{cases} 0,\ if\ \ E \le -1.15; \\ 1,05 \cdot E + 1,22,\ if\ -1,15 < E \le -0,2; \\ 1,\ if\ \ E \ge -0,2; \end{cases}$$

$$k_2 = 1 - k_1 = \begin{cases} 1,\ if\ \ E \le -1.15; \\ -1,05 \cdot E - 0,22,\ if\ -1,15 < E \le -0,2. \\ 0,\ if\ \ E \ge -0,2; \end{cases} \quad (3)$$

$$k_3 = 0.$$

The results of application (2) with above coefficients are given in fig. 6. One can see, that in a short interval $\beta_c \in [0;0,8]$ the best estimator is median, that does not appear in approach proposed in [7] and [8].

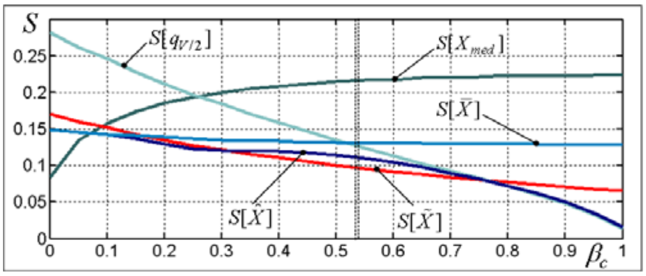


Fig. 6. Dependences of standard deviations for different statistics on a ratio of bases (curvilinear trapeze)

5.2. *Two-component estimators for trapeze distributions*

Let's broadly investigate problems for a whole family of linear trapezoids and find more effective unbiased two-component estimator. It could be expressed by

$$X_{eff} = k_1 \bar{X} + (1 - k_1) q_{V/2}\,, \quad (4)$$

This estimator was analyzed by changing k_1 in (4) for different trapeziums from rectangular to triangular shapes and values of k_1 corresponding to minimum SD were obtained. Results with the uncertainty under 10% are as follows:

$$k_1 = \begin{cases} -0,12\beta + 0,56,\ if\ 0 < \beta < 0.5 \\ 1 - \beta,\ if\ 0.5 < \beta < 1 \end{cases},$$

$$k_2 = 1 - k_1 = \begin{cases} 0,12\beta + 0,44,\ if\ 0 < \beta < 0.5 \\ \beta,\ if\ 0.5 < \beta < 1 \end{cases}. \quad (4a)$$

The results are stable even when n is changed from 10 up to 10000 for trapeziums with different ratio of their bases. Dependences of SD on k_1 for boundary cases of trapezium (triangular and rectangular PDF) are shown in Fig 7. Application of (4) gives the dependence on β as on Fig. 8.

Let us analyze simplified 2C-estimator based on two equal components

$$\tilde{X} = \frac{1}{2} \cdot \bar{X} + \frac{1}{2} \cdot q_{V/2}\,, \quad (5)$$

The results of its MC modelling for linear trapezoid are given also on Fig. 8. From these results one can see that simple 2C-estimator (3) is the best for a wide range of trapeziums ($0 < \beta < 0,75$).Values β>0,75 correspond to the ratio of ranges over 8. It means that one of uniform distributions is dominant. For this case mid-range is the best estimator (see Fig. 6).

The results of investigations, if formula (5) is applicable also for curvilinear trapezoid, are given in Fig. 7. One can see that it is not the most effective estimator for a full range. From analyzes of some cases of β >0,54 it is recommend to use in practice two given below formulas of the best estimator:

$$X_{eff} = \begin{cases} \frac{1}{2} \cdot q_{V/2} + \frac{1}{2} \bar{X}, & if \quad 0,54 < \beta < 0,8; \\ q_{V/2} & if \quad \beta > 0,8. \end{cases}$$

Additional investigations are needed for the full range of β.

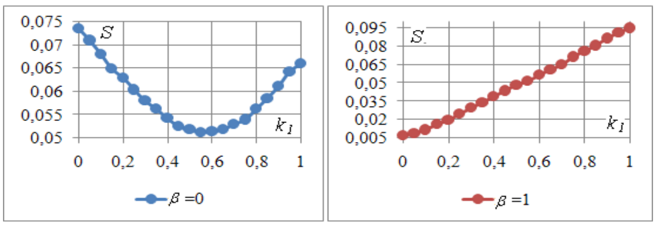


Fig. 7. Dependences of standard deviations on k_1

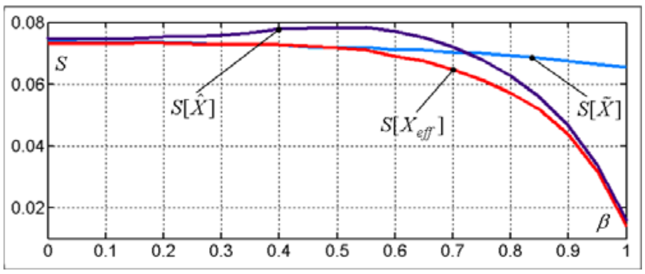


Fig. 8. Dependences standard deviations of the different statistics' on a ratio of bases

Variance of the best estimator should be minimum. Lets try to find analytically its value of k_1. From (4) it is:

$$S^2[X_{eff}] = (k_1 \cdot S[\bar{x}])^2 + ((1-k_1)\cdot S[q_{V/2}])^2,$$

where $S[\bar{X}] = \sigma_x / \sqrt{n}$, σ_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: $$k_1 = \frac{S^2[q_{V/2}]}{S^2[\bar{x}] + S^2[q_{V/2}]}, \qquad (6)$$

For triangular distribution (β =0) [4] and

$$S^2[q_{V/2}] = \frac{3\cdot(4-\pi)}{2n}\sigma_X^2, \qquad k_1 = \frac{3(4-\pi)}{2+3(4-\pi)} \approx 0{,}56.$$

It coincides with results of the earlier MC simulation.

For trapezoid with $\beta = 0{,}35$ we find that $S[\bar{X}] = S[q_{V/2}]$,

and from (6): $k_1 = \frac{\sigma_X^2}{n} \Big/ \frac{2\sigma_X^2}{n} = 0{,}5$.

It coincides with (4a).

For rectangular distribution (β =1):

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

If $n \to \infty$, $k_1 \to 0$. For $n = 30$, $k_1 = 0{,}04$ and n=10: $k_1 = 0{,}1$, so these results are not very far from the above k_1=0 for $n \to \infty$.

6. UNCERTAINTY EVALUATION

6.1. Theoretical background

Because of correlation between $X_{\min}$ and $X_{\max}$ conjoint density function has to be found here. Standard deviation SD of two-component estimator (4) is

$$S[X_{eff}] = \sqrt{k_1^2 S^2[\bar{X}] + (1-k_1)^2 S^2[q_{V/2}] + 2\rho k_1(1-k_1) S[\bar{X}] S[q_{V/2}]}$$

where: $S[\bar{X}] = \frac{S_x}{\sqrt{n}}$; $S^2[q_{V/2}] = \frac{V^2(1-\beta^2)}{16}\cdot\frac{n}{(n+1)(n+2)}$.

In $S[q_{V/2}]$ it was taken to account that $S[X_{\min}] \approx S[X_{\max}]$. The recommendations on correlation coefficient values are in Table 2.

For simplified two-component estimator (5) of k_1=k_2=0,5 proposed in this paper standard uncertainty is:

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

If $\rho \to 0$

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

Table 2. The values of correlation coefficients

n	[100; 200)	[200; 300)	[300; 500)	$n \to \infty$
ρ	0,25	0,2	0,15	$\rho \to 0$

The difference between analytical results and modeling is less than 5%.

6.2. Measurand value and uncertainty calculations

It has to be illustrated below by the numerical example. 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 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).

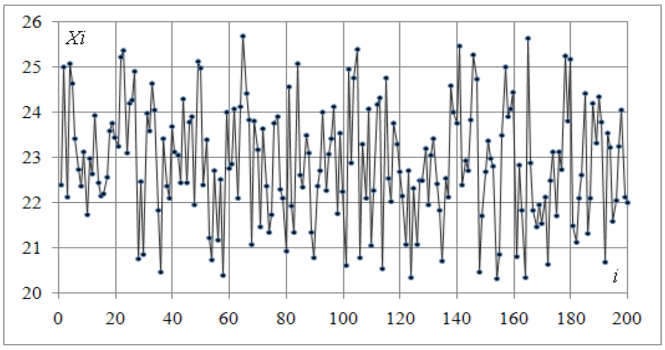


Fig. 9. Values of sample observations

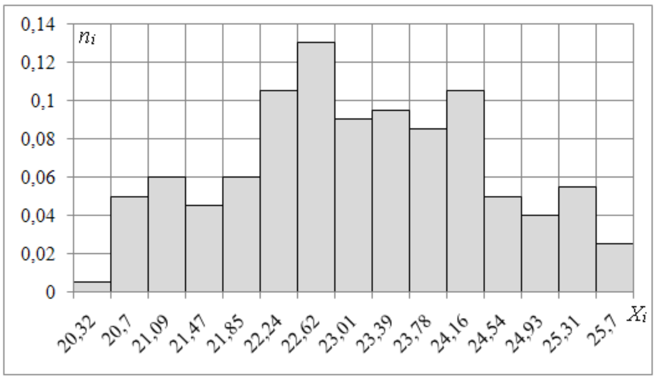


Fig. 10. Histogram of data relative frequencies

Hypothesis about compliance with three different theoretical distributions are verified by χ^2 test.

Parameter β of theoretical distribution could be chosen by the minimum value of χ^2 statistic or by least square method.

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:

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

The trapezoid PDF model of 5.38 and 1.79 bases are found. Its parameter β=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

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

The value of uncertainty for estimator (5a) does not differ significantly from above.
Coverage factor:. $K(P = 0,95) = 1,96$ For coverage probability P expanded uncertainty is:

$$U(P) = K(P) \cdot u_A\ .\ X = (23,01 \pm 0,18),\ P = 0,95$$

Results for all three models are put together in Table 3. We could use KS-test(Kolmogorov-Smirnov).

Table 3. Representations of the measurement result and accuracy

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

7. FINAL CONCLUSIONS

● It is very important to choose the most accurate, 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 β=0,35 up to β=0 of the triangle distribution the sample mean is better.

● The 2-component estimator as the linear form of above two estimators is the best 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 recommendations [1, 2, 11], but also in the statistics, when trapezoidal models are also used [9].

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 similar [12].

REFERENCES

[1] Guide to the expression of uncertainty in measurement (GUM) JCGM OIML 1993

[2] S*upplement 1 to the Guide to the expression of uncertainty in measurement* (*GUM*) *– Propagation of distributions using a Monte Carlo method,* Guide OIML G 1-101 Edition 2007 (E)

[3] Dorozhovets M., Warsza Z. L "Upgrading calculating methods of the uncertainty of measurement results in practice", *Przegląd Elektrotechniki* (*Electrical Review,* Poland,) no 1, 2007, pp.1-13 (in Polish)

[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", *Proceedings of 15th 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, 2nd ed., John Wiley & Sons, New-York, 1977

[6] Novickij P.V., Zograf I.A. "*Estimation of the measurement result errors*", Energoatomizdat, Leningrad,1985 (in Russian only)

[7] Zakharov I.P, Shtefan N.V.: "Minimization of uncertainties in measurements with repeated observations", *Proceedings of 10th IMEKO TC7 International Symposium*, Saint-Petersburg, June 2004, Tomsk Technical University, pp. 189 -192

[8] Zakharov I.P., Shtefan N.V., "Algorithms for reliable and effective estimation of type A uncertainty " *Measurement Techniques,* vol. 48, 5, 2005 p.427-437, www. Springer com. (transl. from Russian *Izmieritelnaja Tekhnika* no 2, 2005 p.9 -15)

[9] Van Dorp J.R., Kotz S., "Generalized Trapezoidal Distributions" *Metrika*, Vol. 58, Issue 1, July 2003

[10] Kacker R. N., Lawrence J. F., "Trapezoidal and triangular distributions for Type B evaluation of standard uncertainty" *Metrologia* 44 (2007), pp. 117–127

[11] EA-4/02 • *Expression of the Uncertainty of Measurement in Calibration,* EA European cooperation for Accreditation, December 1999, pp. 63-65

[12] Fotowicz P.,"Accuracy of convolution of rectangular and normal distributions approximation by trapezoidal distribution", *Pomiary Automatyka Robotyka* 5/2001 pp.9-11. (in Polish)