20th IMEKO TC4 International Symposium and
18th International Workshop on ADC Modelling and Testing
Research on Electric and Electronic Measurement for the Economic Upturn
Benevento, Italy, September 15-17, 2014

Midrange as estimator of measured value for samples from population of uniform and Flatten-Gaussian distributions

Zygmunt L. Warsza[1], Stefan Kubisa[2]

[1] Industrial Research Institute for Automation and Measurements (PIAP), Poland, e-mail: zlw@op.pl
[2] West Pomeranian University of Technology, Faculty of Electrical Engineering, Poland, e-mail: kubisa@zut.edu.pl

***Abstract* – In this paper the statistical properties of midrange as estimator of measured value, for the samples of varying number of observations taken from a population of uniform distribution, were examined by the Monte Carlo simulation. The midrange of such samples had a smaller standard deviation than the mean value recommended by the Guide GUM (Fig. 1). A distribution similar to Student's t-distribution and an expanded uncertainty were also calculated for such samples (chapters 3 and 4). In chapter 5 it was found that for samples from the general population of Flatten-Gaussian distribution, with increasing share of the normal distribution, the advantage of midrange quickly decreased. Considerations have been illustrated by figures. Final conclusions have been enclosed.**

I. INTRODUCTION

The metrologically correct result of a measurement should contain the most probable value of a measurand together with an assessment of its accuracy. It should be determined in a widely accepted uniform manner. Seven international organizations recommend the procedure described in the guide known as the acronym GUM [1]. It assumes that observations are independent and come from a normally distributed population and their number in the measuring sample is large enough. In the laboratory measurements these assumptions are typically fulfilled and the uncertainties are determined for two or three significant digits. A description of the instrument accuracy by the worse case of limited errors is also used.

The statistical approach to calculate the unknown systematic errors and the error of final result, nearly similar as in Guide GUM [1], was known 40 years earlier in Poland, from S. Trzetrzewinski PhD thesis in 1951 at Gdansk Technical University [4]. But in the GUM this approach is presented widely, using another terminology (e.g. instead of the most probable final error – the concept of uncertainty) and recommendations of GUM are now internationally sanctioned.

Measurements and processing of the measurement data carried on in science, industry and many other fields commonly use now electronic and computers. Some of them do not fulfill the assumptions of GUM. The distribution of measured values, or components of a random signal is often better modeled by Non-Gaussian distributions. There are also distortions (random, continuous or intermittent) – so called outliers. Sometimes there is a need to make statistical evaluations from the samples of low number of elements. Since the mid-twentieth century the new statistical tools were developed, such as robust and resampling methods, to analyze these issues.

The statistical properties of samples from a population of uniform distribution and few different Flatten-Gaussian distribution in varying degree) will be examined in detail by using Monte Carlo simulation.

II. BASIC EQUATIONS

The classic approach of the measurement uncertainty calculation is in [1] and [2]. It is based on an assumption that the randomness of the observed N values of x is the source of their origin from the general population with normal distribution. After elimination of known systematical errors from measurement data, the best estimate of the measured value is determined as the arithmetic mean of data of the empirical sample:

$$x_{\mathrm{cl}} = \frac{1}{N}\cdot\sum_{n=1}^{N} x_n = \frac{1}{\nu+1}\cdot\sum_{n=1}^{\nu+1} x_n \qquad (1)$$

wherein $\nu = N - 1$ is the number of degrees of freedom.

The estimator s_{cl} of the standard deviation of the mean value x_{cl} is

$$s_{\mathrm{cl}} = \sqrt{\frac{\sum_{n=1}^{N}\left(x_{\mathrm{cl}} - x_n\right)^2}{N\cdot(N-1)}} = \sqrt{\frac{\sum_{n=1}^{\nu+1}\left(x_{\mathrm{cl}} - x_n\right)^2}{(\nu+1)\cdot\nu}} \qquad (2)$$

ISBN-14: 978-92-990073-2-7

This deviation of the sample, determined by statistical method is named in [1] as a standard uncertainty $u_A\ (x)$. In addition, based on the knowledge of the observer, a standard uncertainty $u_B(x)$ is estimated. Then the combined standard uncertainty is calculated. Considering the expansion coefficient k_p for the confidence level p, or using Monte Carlo method [2], the expanded uncertainty is $U\ (x) = k_p{\cdot}u_C\ (x)$.

In supplement 1 of GUM [2] Monte Carlo (MC) method is recommended as the most universal, based on elementary mathematical relationships, possible to apply for the highly nonlinear measurement functions, as well as in unusual cases, for example, asymmetrical distributions, as in [5]. If it is known that the observations come from different general population and the probability distribution is also given, it is better to use an approach called here: special.

Cramer, in his the excellent timeless monograph [3] for samples from a population with uniform distribution demonstrated analytically that a midrange is better estimator of a measurand than a mean due to having the smaller standard deviation. This is confirmed by numerical examples in [7] - [9] and the distributions of the three estimators of the samples with high cardinality N obtained by the MC method – Figure 1 [9]. Basic parameters of the sample are presented in the Table 1.

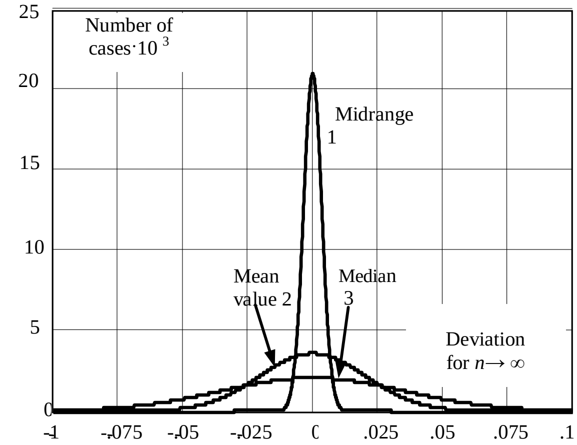

Fig. 1. Histograms of estimators of measurand value for samples from population of rectangular distribution simulated by $200{\times}2^{20}$ *random numbers: 1 – midrange; 2 – mean value; 3 – median.*

Table 1. Statistical parameters of the sample of uniform pdf.

Range of the sample	$V = x_{n\max} - x_{n\min}$	
Midrange:	$x_V = \dfrac{x_{n\max} + x_{n\min}}{2}$	(3)
Standard deviation of midrange x_V:	$s_V = \dfrac{V}{\sqrt{2}} \cdot \sqrt{\dfrac{N+1}{(N-1)^2(N+2)}} = \dfrac{V}{\sqrt{2}\cdot\nu} \cdot \sqrt{\dfrac{\nu+2}{\nu+3}}$	(4)

For observations from a population with uniform distribution the best estimate of the measured value is midrange of the sample x_V (3), and the estimate of the standard deviation is its empirical deviation s_V (4).

The paper presents the results of the properties of these estimators in function of degrees of freedom of sample $\nu = N - 1$. The simulations were carried out using the MC method. Observations were simulated with pseudo-random numbers from a population with a standard deviation $\sigma = 1$ for the number $M = 2 \times 10^5$ of simulations and the numbers of degrees of freedom $\nu = 1, 2, 3, 4, 5, 7, 10, 16, 32, 63, 125, 250, 500, 1000$. The tests were carried out for the case where the population of random numbers (from which the observations come from) has clear and uniform distribution and for several cases where this population has uniform distribution contaminated by normal distribution.

III. OBSERVATIONS FROM A POPULATION WITH UNIFORM DISTRIBUTION

In general case the estimator of the mean value x_{cl} (1) in classic approach and the estimator of midrange x_V (3) as special-estimator and their standard deviations respectively s_{cl} and s_V have different values. A comparison of deviations from formulas (2) and (4) is shown in Fig 2.

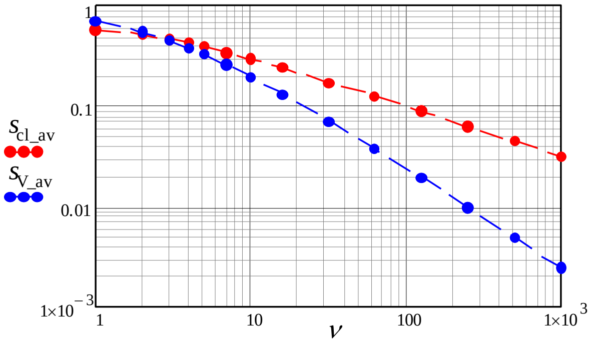

Fig. 2. Standard deviations of mean and of midrange as functions of number ν of degree of freedom.

On the basis of $N - 1 = \nu$ pure samples of uniform distribution with a standard deviation $\sigma = 1$, in each simulation number m ($m = 1, \ldots M$), for each value ν, the values s_{cl_m} and s_{V_m} were calculated. Then, for each value of ν from M simulations average values s_{cl_av} and s_{V_av} were calculated. Figure 2 suggests a superiority of special estimators (3), (4) over the conventional estimators (1), (2), as for $\nu > 2$ is $s_{cl\_av} > s_{V\_av}$. For $\nu = 10^3$ a value of s_{V_av} is approximately 13 times less than the s_{cl_av}.

Comparing the estimates only by empirical deviations is not fully reliable from metrological point of view, because the expanded uncertainties are important. Only the expanded uncertainties U_A associated with the random scatter of observations were taken into consideration here. With the classical approach the expanded uncertainty $U_{\mathrm{Acl\_}m}$ is the product of deviations of the empirical $s_{\mathrm{cl\_}m}$ and expansion coefficient k_{cl} calculated by Student's t-distribution for $\nu = N - 1$ degrees of freedom and confidence level P:

$$U_{\mathrm{Acl\_}m} = k_{\mathrm{cl}} \cdot s_{\mathrm{cl\_}m} \quad . \qquad (5)$$

In a special approach one can express the expanded uncertainty of type A:

$$U_{\mathrm{AV\_}m} = k_{\mathrm{V}} \cdot s_{\mathrm{V\_}m} \, , \qquad (6)$$

where: k_{V} is the coverage factor, specially adapted to estimators (3) and (4).

Classic Student's t-variable is defined as:

$$t \overset{\mathrm{def}}{=} \frac{x_{\mathrm{cl}} - \mathrm{E}(x)}{s_{\mathrm{cl}}} = \frac{\Delta x_{\mathrm{cl}}}{s_{\mathrm{cl}}} \quad . \qquad (7)$$

where $E(x)$ is the measured (expected) value, known in simulation experiments, and Δx_{av} – error of estimate x_{cl}. Similarly, the variable t_{V} of quasi-Student distribution is defined as

$$t_{\mathrm{V}} \overset{\mathrm{def}}{=} \frac{x_{\mathrm{V}} - \mathrm{E}(x)}{s_{\mathrm{V}}} = \frac{\Delta x_{\mathrm{V}}}{s_{\mathrm{V}}} \, . \qquad (8)$$

This is equivalent to the Student's t-variable for a population of observations of the uniform distribution with estimators (3) and (4). The probability distribution of the variable t_{ν} can be called quasi-Student distribution.

The graphs of the coverage factor k_{V} calculated by MC method and the coverage factors k_{cl} for a confidence level p = 95% are shown in Fig. 3.

For following values ν_m the uncertainty $U_{\mathrm{Acl\_}}$(5) and $U_{\mathrm{AV\_}m}$ (6) M times were calculated and also their average values $U_{\mathrm{V\_av}}$ and $U_{\mathrm{cl\_av}}$ in the data sets of size M were found. The results as function of ν are given in Fig. 4.

These plots confirm the superiority of midrange estimators (3), (4) over the classical estimators of mean value (1), (2) for the numbers of degrees freedom ν greater than about 5. For $\nu = 10^3$ the value $U_{\mathrm{V\_av}}$ is about 11 times smaller than $U_{\mathrm{cl\_}av}$.

It is desirable to verify the MC simulation results. It may be done by testing the empirical probability of an event. It consists in verification if the estimate errors of the measured value (7), (8) are in the limits of the calculated uncertainty. For each of the $N = \nu + 1$ observations there is a need to calculate the number of successes and divide it by the number of M simulations. This quotient should have a value close to the postulated level of confidence p = 95%. The results of this verification are in Fig. 5.

They are fully satisfactory for a special approach – estimate (3) and (4). In contrast to the classical approach – estimate (1) and (2), they are unsatisfactory at small values of the number of observations $N = \nu + 1$.

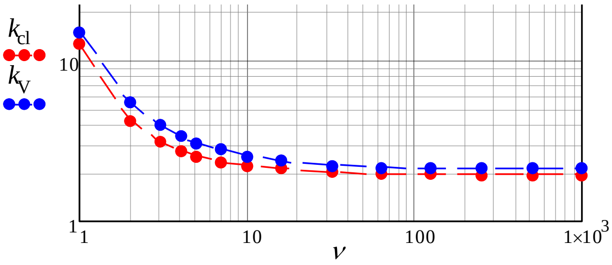

Fig. 3. Coverage factors as function of number of ν degrees of freedom, P = 0.95.

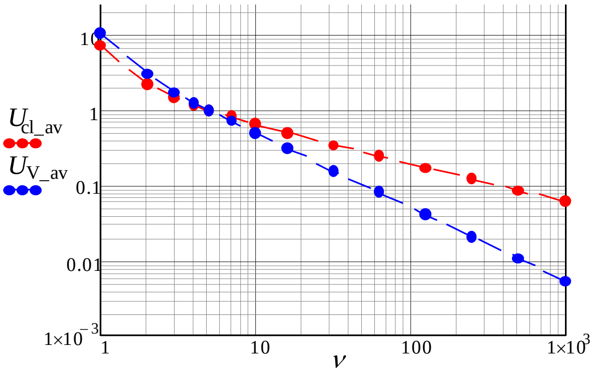

Fig. 4. Average expanded uncertainties as function of number of ν degree of freedom.

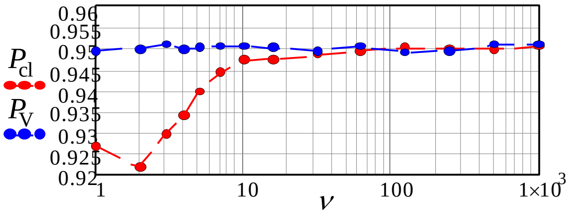

Fig. 5. Empirical probabilities as a function of number of ν degree of freedom.

IV. EXAMPLE

Let us find the expected value of the resistance R and its expanded uncertainty of a population of resistors with nominal value R_N = 100 Ω. It can be assumed that the values of resistance R in the population have a uniform distribution based on the information that this population is the result of:

– selection in production from a population of resistors of the value of R of a continuous distribution of considerable width, or

– choosing the specially calibrated resistors taken from the population with a continuous, very wide distribution of values R, e.g. by step by step method [6],

To determine the expected value of resistance in the population and its uncertainty, the sample of size N is randomly collected and each resistance value is measured. Table 2 shows the results of Monte Carlo simulations of such procedure.

Four variants are completed in the table: A – sample of size $N_A = 2$ ($\nu_A = 1$), B – sample of size $N_B = 4$ ($\nu_B = 3$), C – sample of size $N_C = 8$ ($\nu_C = 7$), D – sample of size $N_D = 17$ ($\nu_D = 16$). The calculations of estimates of: the measurand value (as expected value of resistance in the population), its standard deviation and the expanded uncertainty for a confidence level of $p = 95\%$ were performed for each of the samples. Two methods of calculations were used: the classical method (estimate = average value) by formulas (1) and (2), and a special method (estimate = midrange) by formulas (3) and (4). It was also assumed that the uncertainty of the measuring equipment used for measurement of resistances R is negligible.

Table 2. Example.

No	Sample			
R_i	A	B	C	D
1.	99.925	100.006	100.089	100.015
2.	100.057	99.929	100.016	100.025
3.	–	100.046	99.951	99.910
4.	–	99.906	99.970	99.952
5.	–	–	100.059	100.079
6.	–	–	99.914	100.046
7.	–	–	100.018	99.979
8.	–	–	99.940	100.081
9.	–	–	–	99.978
10.	–	–	–	99.971
11.	–	–	–	100.049
12.	–	–	–	100.048
13.	–	–	–	99.940
14.	–	–	–	100.036
15.	–	–	–	99.974
16.	–	–	–	99.922
17.	–	–	–	99.941
x_{av}	99.991	99.972	99.995	99.997
x_V	99.991	99.976	100.001	99.995
s_{cl}	**0.0660**	0.0327	0.0216	0.0132
s_V	0.0808	**0.0301**	**0.0168**	**0.00736**
k_{cl}	12.71	3.18	2.36	2.12
k_V	15.31	3.98	2.79	2.41
U_{cl}	**0.84**	**0.10**	0.051	0.028
U_V	1.2	0.12	**0.047**	**0.018**

The results of calculations are presented in the lower rows of Table 2. The best results are given in bold fonts. As expected, the average value x_{cl} has a lower value of the expanded uncertainty only for a very small sample sizes: $N_A = 2$ and $N_B = 4$ – variants A and B. In other cases, the midrange x_V is better as it has less uncertainty.

V. STATISTICAL PROPERTIES OF SAMPLES FROM A POPULATION OF FLAT-NORMAL PDF

The midrange value x_V of the samples from a population with uniform distribution depends only on the two external, minimal and maximal observations. Then the midrange value and its uncertainty is strongly influenced by data outliers. It can be eliminated as for Gaussian samples, according to Grubbs criteria. The second way is to calculate x_V for several external pairs of observations in the sample and to discard the outlier score.

In the measuring system can also occur samples from a population, which is a convolution of uniform distribution with another distribution. If the second one can be approximated by a normal distribution, the flatten-normal distribution is obtained. The calculation of the mean value and its uncertainty of the sample from such distribution by use a number of conventional methods is described in [14]. However, by MC method it will be tested whether a midrange of the samples from this distribution has similar properties to those of the uniform distribution.

The distribution of a flatten-normal population can be characterized by the degree of participation λ of normal distribution. It means that if the population standard deviation $\sigma = 1$, the standard deviation of the normal distribution component is $\sigma_N = \lambda$, and for the main component of the uniform distribution $\sigma_J = \sqrt{1-\sigma_N^2}$. Fig. 6 shows the plots of the plane-normal distribution in four different levels of λ. Thus, for $\lambda = 5\%$ the component with uniform distribution is characterized by standard deviation of $\sigma_J \approx 99.87\%$. Plots of the empirical deviations and expanded uncertainties with contribution of the normal distribution $\lambda = 5\%$ are not differ significantly from charts for a uniform distribution in Fig. 2 and Fig. 4.

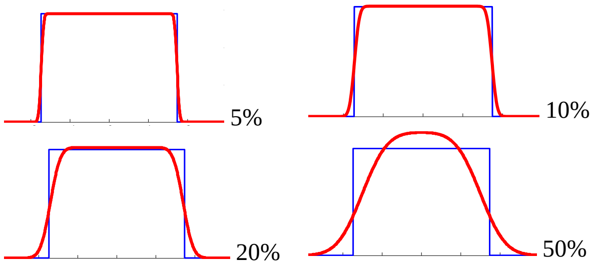

Fig. 6. Flatten-Gaussian distributions of different λ and uniform pdf.

In contrast, the empirical probability plots shown in Fig. 7 differ significantly from those shown in Fig. 4. Too small probabilities P_V for larger numbers of observations $\nu = n + 1$ indicate the need to extend the coverage factors k_V for flat-normal distribution.

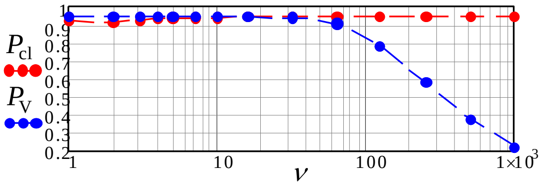

Fig. 7. Empirical probability for λ = 5% as function of number ν of degree of freedom.

Using the MC method the coefficients k_V for a confidence level p = 95% are calculated for the values

λ = 0%, 5%, 10%, 20%, 50% of the degree of participation of normal distribution. The results of this calculations are presented in the Tab. 3 and some of them in Fig. 8.

Table 3. Coverage factor k_{V} (p = 95%) as function of number ν of degree of freedom for some values of λ.

The number of degree of freedom ν	λ compactness of the normal distribution in %				
	λ = 0%	λ = 5%	λ = 10%	λ = 20%	λ =50%
1	15.31	15.54	15.01	14.40	11.94
2	5.51	5.47	5.42	5.31	4.66
3	3.98	3.95	3.96	3.91	3.67
4	3.42	3.41	3.41	3.40	3.34
5	3.09	3.09	3.10	3.12	3.23
7	2.79	2.79	2.81	2.88	3.19
10	2.57	2.59	2.62	2.74	3.36
16	2.41	2.42	2.52	2.83	3.96
32	2.24	2.39	2.71	3.55	5.63
63	2.18	2.61	3.48	5.13	8.68
125	2.14	3.45	5.23	8.20	14.23
250	2.13	5.31	8.60	13.93	24.33
500	2.14	8.79	14.79	24.30	42.57
1×10^3	2.13	15.39	26.44	43.61	76.29

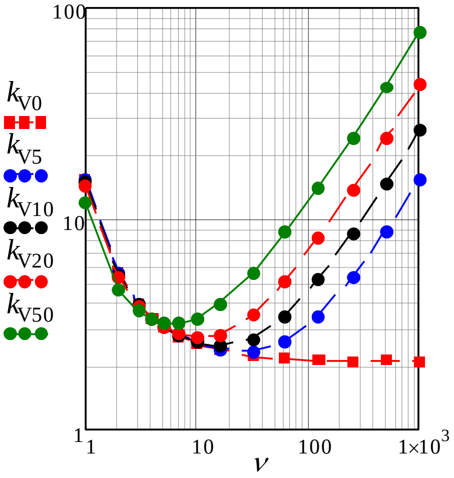

Fig. 8. Coverage factors (p = 0.95) of Flatten-Gaussian pdf of normal pdf level λ = 0%, 5%, 10%, 20%, 50% as function of number ν of degree of freedom.

With the increase of λ as degree of participation of the normal distribution in the flat-normal population the efficiency of special approach (for the midrange) compared to classical approach is decreasing. This is illustrated in Figure 9 by graphs of the average expanded uncertainties and of the control probabilities.

VI. REMARKS AND CONCLUSIONS

The results of the MC simulation were achieved with errors inversely proportional to the squareroot of M.

In particular, the error of probability calculation is binomial with a standard deviation:

$$\sigma_P = \sqrt{\frac{P\cdot(1-P)}{M}} \qquad (9)$$

For P = 0.95 and $M = 2\cdot10^5$ one can receive $\sigma_P \approx 5\cdot10^{-4}$. For large values of M the error of probability calculation approaches the normal distribution and error limit can be estimated with the range $3\cdot\sigma$ as approximately 0.15%. This validates irregularities of plots in Fig. 5 and Fig. 9.

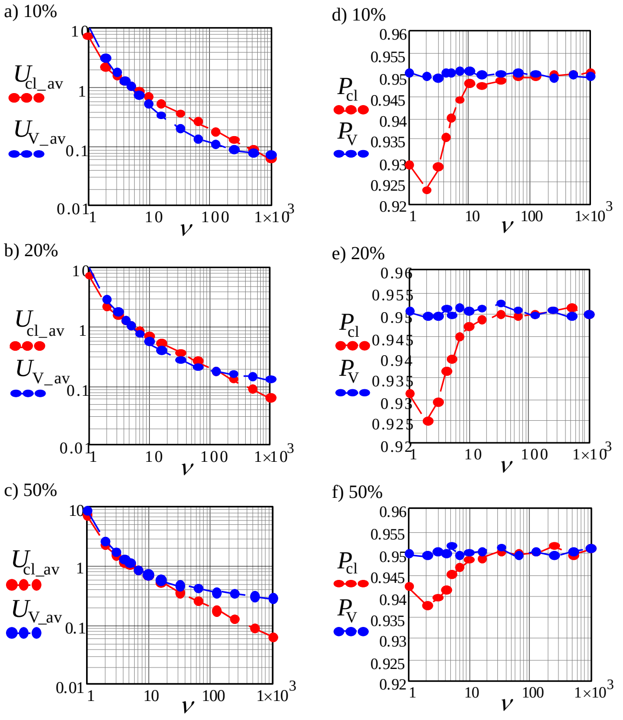

Fig. 9. Uncertainties $U_{\mathrm{cl\ av}}$, $U_{\mathrm{V\ av}}$ – a, b, c and control probabilities P_{cl}, P_{V} – d, e, f for levels of standard deviation of normal distribution λ = 10%, 20% and 50%.

For the uniform distribution the use of special approach (3) and (4) is effective when the number of degrees of freedom ν is greater than 5 – see Fig. 4. Then the average expanded uncertainty $U_{\mathrm{V\_av}}$, calculated according to a special approach using a special coverage factor, is less than the average uncertainty $U_{\mathrm{cl\_av}}$ calculated classically by the GUM recommendations. $U_{\mathrm{V\_av}}$ is less with greater number of degrees of freedom ν.

For the observations from a population with convoluted uniform distribution even with a low content of another distribution, for example, λ = 5% of the normal distribution, there is a need to increase the coverage factor (Fig. 8). It was assumed that this additional component has a normal distribution. Increasing the degree of participation of the normal distribution λ approach reduces the effectiveness of the special approach – see Fig. 9 a, b, c. For example, for λ = 20% (Fig. 9a) the effectiveness is only for the number of degrees of freedom ν from about 5 to about 100, and for

λ = 50% (Fig. 9c) the special approach is worse to the classic in the whole range of numbers of degrees of freedom. The degree of participation of 50% means that the standard deviation of additional component is 50% of the total standard deviation. The standard deviation of the main component is then $\sqrt{1-0{,}5^2} \approx 0{,}87 = 87\%$ of the standard deviation.

The classical approach is more efficient if there is a significant decrease of the uniform distribution in the plane-normal distribution (1), (2). However, the small numbers of degrees of freedom $\nu < 20$ give a bit too low level of control probability P_{cl} – Fig. 9 d,e,f.

In addition, it is worth to mention that the other simple distributions also have the single component estimators better than the mean value. For U distribution (arc sin) midrange also is better, for the Laplace distribution (two-exponential) the best is median [7] – [9].

By MC method there are also examined the families of trapezoidal distributions, linear one – Trap as a convolution of two different uniform distributions and CTrap of concave shape [10] – [13]. For the ratio β of two bases of the trapezium in the range of 1 - 0.6315 the midrange is a better estimator than the mean value as it has a smaller standard deviation. For the linear Trap and concave CTrap trapezoidal distributions two-component estimator: 0.5 (midrange + mean) is proposed [10] – [13]. It is more effective than any single-element estimator almost in the full (0; 1) range of the ratio β of trapezium basis.

REFERENCES

[1] Guide to the Expression of Uncertainty in Measurement (GUM). ISO/IEC/OIML/BIPM, first edition, 1992, last ed. BIPM JCGM 100 (2008).

[2] Guide to the Expression of Uncertainty in Measurement (GUM), OIML ed. 2008 Supplement Propagation of distributions using a Monte Carlo method, G 1 – 101, (2007).

[3] H. Cramer, "Mathematical Methods of Statistics", Stockholm Univ. (1946) Chapter 19.1.

[4] S. Trzetrzewiński ed., Drewnowski et all, "Pomiary Elektryczne" (Electrical Measurements), Chapter 2 of part I PWN Warszawa (1959), in Polish.

[5] S. Kubisa, S. Moskowicz, "A study on transitivity of Monte Carlo based evaluation of the confidence interval for a measurement result", PAK (Pomiary Automatyka Kontrola) no 6 (2007), pp. 3 –7.

[6] S. Kubisa, "Error distribution of a set of measuring instrument and an influence of step by step calibration procedure on the distribution", Metrologia i Systemy Pomiarowe. vol. V no 4 (1998), PWN, pp. 291 – 302.

[7] M. Dorozhovets, Z.L.Warsza, "Upgreading calculating methods of the uncertainty in measurements", Przegląd Elektrotechniczny - Electrical Review nr 1, (2007), pp. 1 – 13, in Polish.

[8] M. Dorozhovets, Z. L. Warsza, "Methods of upgrading the uncertainty Part 2 Elimination of the influence of autocorrelation of observations and choosing the adequate distribution", Proceedings of 15th IMEKO TC4 Symposium, part 1, Sept. (2007) Iasi Romania, pp. 199 - 204.

[9] Z. L. Warsza, M. Dorozhovets, "Type A uncertainty evaluation of autocorrelated observations and choosing the best estimators of data distribution", Proceedings of 18th National Symposium Metrology and Metrology Assurance, Sept. (2008), Sozopol Bulgaria, pp. 70 –78.

[10] Warsza Z. L., Galovska M., "About the best measurand estimators of trapezoidal probability distributions", Przegląd Elektrotechniczny - Electrical Review 5 (2009), pp. 86 -91.

[11] Z. L Warsza, M. Galovska, "The best measurand estimators of trapezoidal PDF", Proceedings of IMEKO World Congress "Fundamental and Applied Metrology" Lisbon, 2009, (CD pp. 2405 –10).

[12] M. Galovska, Z. L., Warsza, "The ways of effective estimation of measurand", PAKgoś (Pomiary Automatyka Komputery w gospodarce i ochronie środowiska), no 1 (2010), pp. 33 -41.

[13] Z. L. Warsza, "Effective Measurand Estimators for Samples of Trapezoidal PDFs", JAMRIS (Journal of Automation, Mobile Robotics and Intelligent Systems) vol. 6, no 1, (2012) pp. 35-41.

[14] P. Fotowicz, "Method of calculating the coverage interval based on Flatten-Gaussian distribution", Measurement, vol. 55 (2014) pp. 272 -275.