УДК 006.86 : 519.245

Z.L. Warsza[1], M.J. Korczycski[2]

[1] *Industrial Research Institute of Automation and Measurement (PIAP) Warsaw, Poland*
[2] *Technical University of Lodz, Dep. of Semiconductor and Optoelectronics Devices Lodz, Poland*

STATISTICAL PROPERTIES OF SKEWNESS AND KURTOSIS OF SMALL SAMPLES FROM NORMAL POPULATION

Statistics of skewness and kurtosis distributions and their basic parameters for a set of samples of certain small numbers of elements are find. These distributions were determined using the Monte Carlo method. The samples were repeatedly taken at random from a normally distributed population. Knowledge about statistics of skewness and kurtosis should allow to obtain a more reliable estimate of the standard deviation and the uncertainty of the measurand value estimator from samples of a small number of measurement observations, when range of their distribution is known.

Keywords: *handling of samples, skewness, kurtosis, data modeling.*

Introduction

In many cases in practice of measurements we can collect sample of only a small number of measurement observations for various reasons. A reason of a small measurement sample might be:

- The lack of a larger number of objects to be tested (eg. for the validation of the method used in only a few accredited laboratories).
- High cost of measurements or limited time of their execution.
- The inability to re-perform the measurements, eg. in the study of a distant terrain and in medicine.
- The limited number of collected data might be while object under investigation might be destroyed or its properties after test are changing irreversibly.

In all above mentioned cases, only a small sample might be available for further data handling.

According to GUM Guide [2], in determining the result and uncertainty of measurement any sample of the data is considered as it comes from a normally distributed population. The best estimator is regarded as a mean value of collected data, and the uncertainty of type A as a parameter, which characterize data scattering, basing on calculated of the variance of the sample.

Samples from a normally distributed population and other symmetric distributions can be asymmetric. The asymmetry of these samples increases with decreasing the number of elements in collected samples.

It was necessary to examine to what extent it can also be a significant knowledge of other statistical parameters of small and very small samples, including skewness and kurtosis of the cases where the type of distribution of the population is a priori not known. Here, as the first step of this research will appoint statistics of the skewness and kurtosis of small and very small samples taken at random from a normally distributed population. Monte Carlo simulation method was applied to analyze such small samples.

Pearson's skewness coefficients of population and sample

The probability density distribution (pdf) of the asymmetry of the right tail, with median, mean and

mode marked, as an example is presented in Fig. 1. For asymmetric distributions of the mean value μ, mode and median do not overlap with each other.

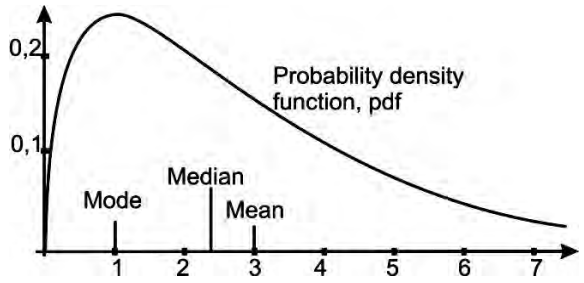


Fig. 1. Parameters of the right-asymmetric distribution function (pdf)

Unified description of various distributions allows asymmetry is given by Pearson's coefficient of skewness γ_1:

$$\gamma_1 = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{E\left[(X-\mu)^3\right]}{\left(E\left[(X-\mu)^2\right]\right)^{3/2}}, \quad (1)$$

where μ – mean value of distribution, σ – standard deviation of distribution, μ_3 – third central moment of distribution.

Then the Pearson's coefficient of skewness γ_1 given by Eq. (1) is based on the central moment of μ_3 and a standard deviation σ of the data population. The coefficient equals to zero for a symmetric distribution, and is positive for right-tailed asymmetry distributions (extended right tail). Also other non-classical skewness coefficients are used.

Pearson's skewness coefficient g_1 for a n-element sample of $x_i \subset X$ population is determined by this sample central moments: $m_2 = s^2$ and m_3 as estimators of μ_2, μ_3.

A skewness coefficient of sample g_1 is given by formula

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{3/2}}. \quad (2)$$

The skewness coefficient g_1 according (2) is biased. Formula for the unbiased skewness [1] is given below

$$g = \frac{\sqrt{n(n-1)}}{n-2} g_1. \quad (3)$$

The software supporting statistical calculation for skewness coefficient of the sample use a bit different form then given by (2) i (3) [3], like given by (4)

$$SKE = n\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3 \Big/ \left((n-1)(n-2)\, s^3\right) \quad (4)$$

and use also a standardised skewness coefficient

$$SSKE = SKE\sqrt{n/6}. \quad (5)$$

For symmetric population and n > 150 SSKE coefficient is of Normal distribution [3].

The differences between the values of the coefficient of skewness given by different definitions are not significant but for the very small samples should be considered.

Standard deviation of the skewness coefficient

Variation of the skewness coefficient of *g* for a sample of n elements from Normal population, according to [2] is

$$D(g) = \frac{6n(n-1)}{(n-2)(n+1)(n+3)}. \quad (6)$$

To estimate variance of D(g) of small samples the Smirnov formula [4] is also used:

$$D(g) = \frac{6(n-2)}{(n+1)(n+3)} = \frac{6}{n}\left[1 - \frac{12}{2n+7} + O\left(\frac{1}{n^3}\right)\right], \quad (7)$$

where $O(\cdot)$ is a residua part $1/n^3$.

Equation (7) refers to samples of n > 25 elements. For larger number of n a residua part in (7) becomes neglected and variance is tending to: $D(g) \to 6/n$.

Kurtosis of small samples

Kurtosis of the particular population is given by the ratio: μ_4/σ^4. It serves as a measure of flattening (slenderness) of distribution, which indicates the concentration of its data. For a Normal (Gaussian) distribution kurtosis is equal to 3.

Kurtosis of the sample is a ratio of its moments: m_4, m_2, so:

$$\text{Kurtosis} = \frac{m_4}{m_2^2} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{(n-1)\, s^4} \quad (8)$$

To compare other distributions with Normal distribution, the excess kurtosis coefficient of the population defined as: K = Kurtosis – 3 is used, and comparison of samples, for $n \geq 4$ the formula from [3] given below is applied

$$K(n) = \frac{n(n+1)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)s^4}. \quad (9)$$

The standardised Kurtosis coefficient also is used as

$$SK(n) = K \cdot \left(\frac{24}{n}\right)^{-1/2} \quad (10)$$

Monte Carlo modeling of skewness and kurtosis of small samples

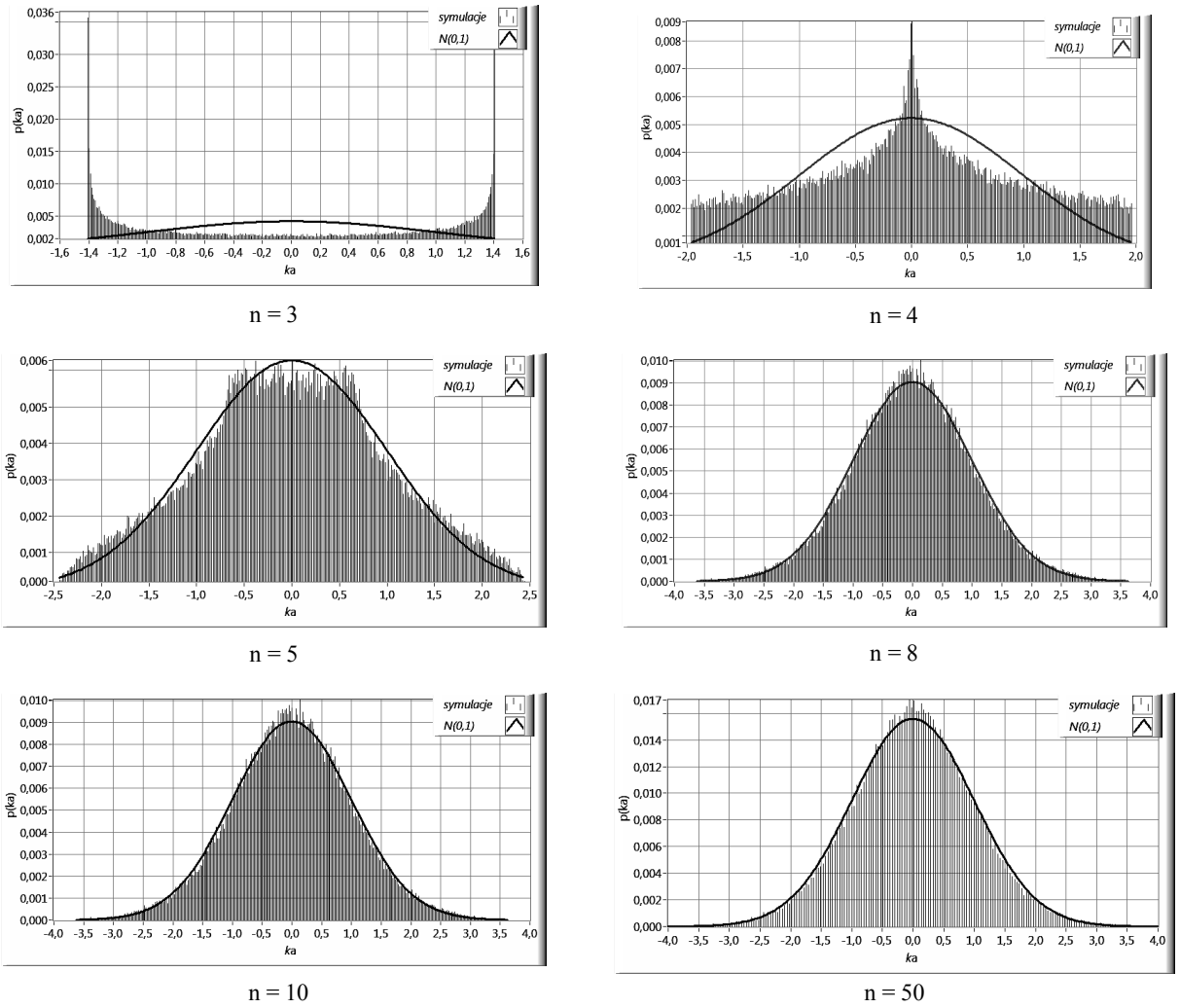
The authors noticed that consideration related to distribution of skewness coefficient g and excess kurtosis coefficient K of very small samples of number of elements below 25 (n < 25) might of worth of analysis. Such samples are analysed using Monte Carlo method. Skewness coefficient g and excess of kurtosis coefficient K for samples of 3< n < 25 elements are calculated and results in function of n are presented.

To determine distributions of the Pearson skewness coefficient for samples with n elements drawn from the population of X with normal distribution (Gaussian) was used unencumbered value of $p(x_i)$ the standard deviation of the sample. From (2)

$$g = \frac{m_3}{s^3} = \sum_{i=1}^{n}(x_i - \overline{x})^3 \Bigg/ \left(n\left(\sqrt{\sum_{i=1}^{n}\frac{(x_i - \overline{x})^2}{n-1}} \right)^3 \right), \quad (11)$$

where m_3 –the sample third central moment, s –sample unbiased standard deviation.

The samples of number of elements n = (3, 4, 5, ...) were extracted from the Normal distribution population and 100 000 such samples were analysed. In Fig. 2 selected results are presented. The shape of larger samples, eg. 50, is very close to Gaussian distribution.



n = 3 n = 4

n = 5 n = 8

n = 10 n = 50

Fig. 2. Some distributions of skewness coefficient $k_a \equiv g$ for samples of low number elements n

For samples of n ≥ 3 from the Normal distribution the mean of modulus of skewness coefficient $\overline{|g|}$ and its variance $D\overline{|g|}$ and standard deviation $s\overline{|g|} = \sqrt{D\overline{|g|}}$ were calculated.

Results are presented in Fig 3 as relations vs. the number of sample elements n. In addition the standard deviation $s(g) = \sqrt{D(g)}$ from the Smirnov formula (7) is also derived. Mean value of modulus of skewness $\overline{|g|}$ and its standard deviation s(g) for the small n-element sample has a maximum for ca. n = 6 and then is declining when n decreases.

Using the Monte Carlo simulation with sets of 100 000 of n-element samples from a normally distributed population, the excess of kurtosis coefficient K were analyzed.

Some histograms of them are in Fig. 4, 5 and the variance of their mean value and other statistical parameters as a function of n are presented in Fig. 6.

Conclusions

For the collection of small samples from a normally distributed population, mean value of the coefficient of skewness deviates significantly from zero, and the kurtosis – from the value of 3 for this population. The skewness is a maximum for the number n of elements of the sample about 6, and excess of kurtosis - n approx. 20. Then, with increasing n, both parameters slowly decreases to 0 and to 3 for the population.

Shapes of distributions of both parameter histograms for small n are also very different from Gaussian pdf.

The skewness and kurtosis are not yet taken into account in determining whether the value of the sample mean as an estimator of the measurement result and its uncertainty as a measure of assessing the accuracy (precision) of the measurement result.

For samples from a population of distributions other than normal, e.g. uniform, trapeze and triangular The skewness and kurtosis functions versus the number of sample elements n should be also different than for a Normal pdf. the average value may be not the best estimator of the measurement result [7 – 9]. Comparison of some skewness statistical parameters of n-element samples from a normal, uniform and triangular distributions is given in [9].

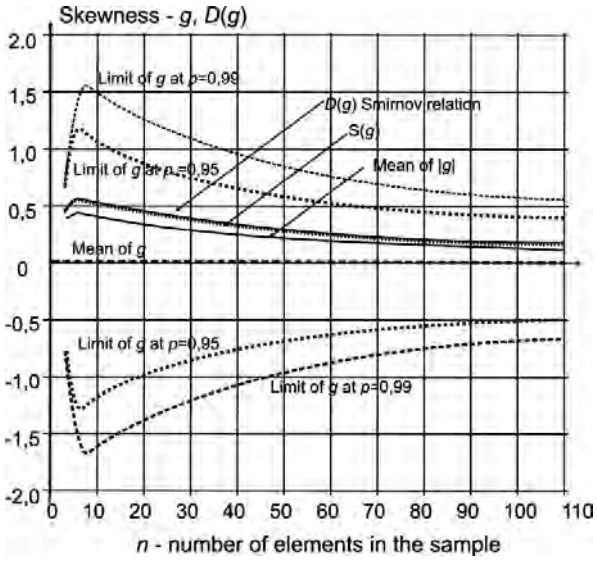
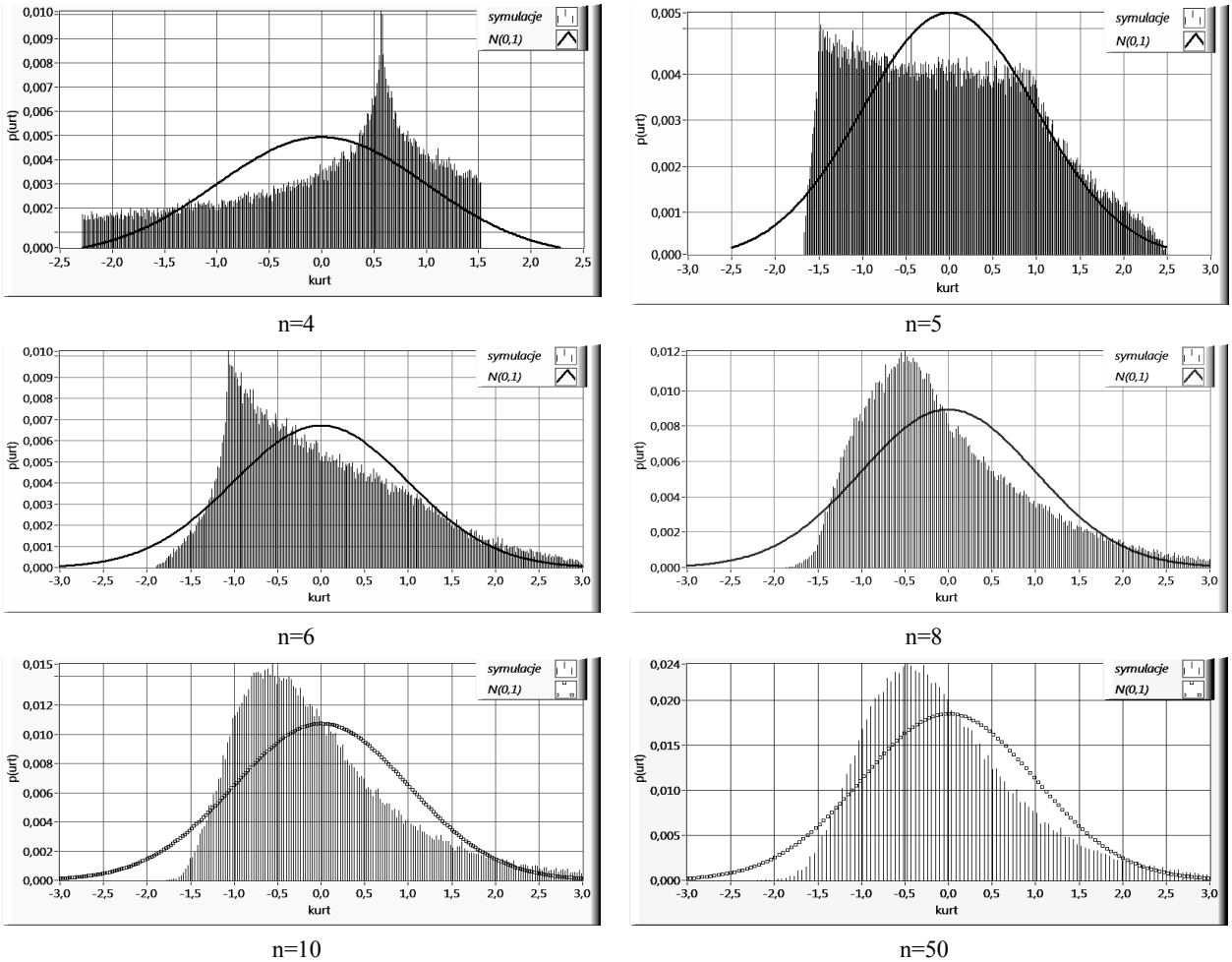


Fig. 3. Statistical parameters of skewness g as function of small number n elements of samples extracted from Gauss population and ranges of its dispersions at level of confidence p = 0,95 and 0,99 respectively

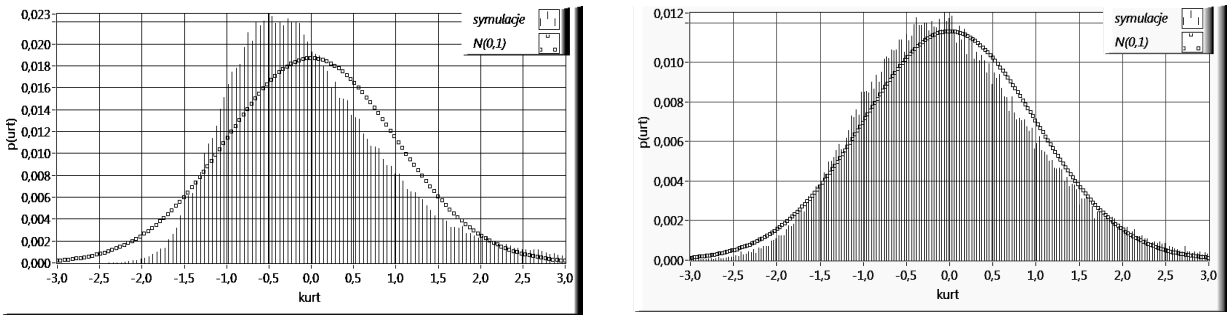


n=4 n=5

n=6 n=8

n=10 n=50

Fig. 4. Histograms of kurtosis excess K-3 of n-element samples from Normal population (n=6, 8, 10, 50)



n=100 n=1000

Fig. 5. Histograms of kurtosis excess K-3 of n-element samples from Normal population (n=100, 1000)

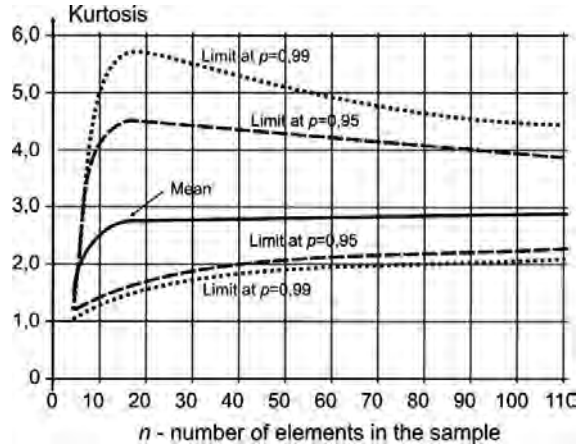


Fig. 6. Statistical parameters of kurtosis K as function of number of elements n of sample from Normal population

It is intended to reproduce the data of small samples by resampling method and examine whether the skewness and kurtosis of these data enable a more precise estimate the measured value and its expanded uncertainty, e.g. by Monte Carlo method [8].

Literature list

1. Beniamin, J. R., Cornell, C. A.: Probability, Statistics and Decision ..., Mc-Graw Hill Inc. (1970).

2. Guide to the Expression of Uncertainty in Measurement GUM. BIPM JCGM 100 (2008)

3. Dobosz, M.: Statistical analysis of experimental results. Exit Warszawa (2004) - in Polish.

4. Bolshev, L. N., Smirnov, N. V.: Tables of mathematical statistics. Nauka (1983) - in Russ.

5. Warsza Z. L.: One component estimators of measurand value of data samples of some non-Gaussian PDF-s. PAK (Pomiary Automatyka Kontrola - Measurement Automation and Monitoring), no 1, p.101-104, (2011), -in Polish.

6. Warsza, Z. L.: Effective Measurand Estimators for Samples of Trapezoidal PDFs. JAMRIS (Journal of Automation Mobile Robotics &Intelligent Systems) vol. 6, n.1, p. 35 - 41, (2012).

7. Kubisa, S., Warsza, Z. L.: Mid-range as estimator of samples from uniform and flatten-Gaussian distribution. Proc. of 20th IMEKO TC4 Symposium, Benevento, Italy, Sept. 15 - 17, (2014) - Acta IMEKO.

8. Propagation of distributions using a Monte Carlo method, Suppl.1 to Guide to the Expression of Uncertainty in Measurement GUM. BIPM JCGM, G1 –101 (2007).

9. Warsza, Z. L., Korczynski M. J.: Statistics of skewness and kurtosis of small measurement samples from populations of normal and few other distributions. PAK (Pomiary Automatyka Kontrola - Measurement Automation and Monitoring), no 12, in print, (2014), - in Polish.

Надійшла до редколегії 12.12.2014

Рецензент: д-р техн. наук, проф. И.П. Захаров, Харківський національний університет радіоелектроніки, Харків.

СТАТИСТИЧЕСКИЕ СВОЙСТВА АСИММЕТРИИ И ЭКСЦЕССА МАЛЫХ ВЫБОРОК ИЗ НОРМАЛЬНО РАСПРЕДЕЛЕННОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

З.Л. Варша М.Дж. Корчинский

Найдены статистики распределений асимметрии и эксцесса и их основных параметров для выборок определенного малого числа элементов. Эти распределения были определены с помощью метода Монте-Карло. Выборки неоднократно извлекали случайным образом из нормально распределенной генеральной совокупности. Знания о статистике асимметрии и эксцесса должны позволить получить более достоверные оценки стандартного отклонения и неопределенности оцениваемого значения измеряемой величины по выборкам с небольшим числом наблюдений, когда диапазон их распределения известен.

Ключевые слова: *обработка выборок, асимметрия, эксцесс, моделирование данных.*

СТАТИСТИЧНІ ВЛАСТИВОСТІ АСИМЕТРІЇ І ЕКСЦЕСІВ МАЛИХ ВИБІРОК З НОРМАЛЬНО РОЗПОДІЛЕНОЇ ГЕНЕРАЛЬНОЇ СУКУПНОСТІ

З.Л. Варша, М.Дж. Корчинський

Знайдені статистики розподілів асиметрії та ексцесу та їх основних параметрів для вибірок певного малого числа елементів. Ці розподіли були визначені за допомогою методу Монте-Карло. Вибірки неодноразово витягували випадковим чином з нормально розподіленої генеральної сукупності. Знання про статистику асиметрії та ексцесу повинні дозволити одержати більш достовірні оцінки стандартного відхилення і невизначеності оцінюваного значення вимірюваної величини за вибірками з невеликим числом спостережень, коли діапазон їх розподілу відомий.

Ключові слова: *обробка вибірок, асиметрія, ексцес, моделювання даних*