https://doi.org/10.15407/ujpe67.5.346

P. KOSOBUTSKYY

Lviv Polytechnic National University
(Lviv 79012, Ukraine; e-mail: petkosob@gmail.com)

STATISTICAL ANALYSIS OF NORMALLY DISTRIBUTED DATA WITH A LIMITED SCATTERING INTERVAL OF VALUES CONVERTED BY DIRECT $g(x) = x^2$; cos x; a^x AND INVERSE FUNCTIONS

The work is devoted to the theoretical analysis of the correct application of the model of a continuous normally distributed random variable in the substantiation of the so-called error transfer formulas in the problem of a statistical processing of experimental data. Attention is paid to the role of limiting the scattering interval of values of a random variable subjected to nonlinear direct $g(X)$ transformation by elementary functions X^2; a^X and $\cos X$, as well as the inverse $g^{-1}(X) = \sqrt{X}$, $\arccos X$, $\log_a X$ to them. The regularities of the statistical averaging of the data obtained by the disorder of the Taylor transform functions are studied. To confirm the validity of the obtained results, the method of quadratic functional optimization is used.

K e y w o r d s: normal distribution, mathematical expectation, variance, random variables, calculation and transfer of errors, transformation of a random variable with elementary functions.

1. Introduction

It is known that the algorithm of processing of experimental data necessarily includes the estimation of statistical error Δx reconstructed curve or error transfer [1] as a direct problem of error theory. Its essence is that the given errors of the argument, get an estimate of the variance D_X and the standard deviation (root mean square) of the transformation function $\varphi(X)$ random variable (RV). Establishing the variance of the errors of the arguments for a given variance of the error of the function refers to the inverse problem. It is also known that if the RV X becomes a function $\varphi(X)$, then the approximate relations are true [2, 3]

$$m_{\varphi(X)} \cong \varphi(m_X), \quad D_{\varphi(X)} \cong \left(\frac{d\varphi}{dx}\right)^2\bigg|_{x=m_X} D_X^2. \qquad (1)$$

As follows from (1), the mathematical expectation m_X that is more resistant to transformation, so a more informative assessment Δx conversion function $\varphi(X)$, is the standard deviation (root mean square) $\sigma_X = +\sqrt{D_X}$.

Propagation errors of transformations are especially relevant for a normally $N_X(m_X, \sigma_X)$ distributed RV X with density function [1]:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\left(\frac{x-m_X}{\sqrt{2}\sigma_X}\right)^2}. \qquad (2)$$

This is due to the fact that distribution (2) is the limit to which all other distributions go, if the number of experimental measurements under the same conditions increases. Function (2) is nonlinear and has a maximum at a point m_X more resistant to transformation, so a more informative assessment m_X, therefore for calculation m_X, σ_X often apply an approximate schedule in the Taylor power series in the range, the boundaries of which are determined with sufficient accuracy by the rule of "three sigma": $m_X \pm 3\sigma_X$.

Algorithms for calculating errors by approximate methods are mostly difficult to implement in practice, so the analytical relationships, the so-called "error propagation formulas (EPF)" are of interest. Such formulas were proposed by the author [5–7] for elementary functions of direct $g(X) = X^2$; $\cos X$; a^X and inverse $g^{-1}(X) = \sqrt{X}$; $\arccos X$; $\log_a X$ transformations of a normally distributed RVX. Despite the fact that critical remarks were made about the algorithm of substantiation of EPF [8, 9], the appearance of works [5, 4] indicates the need to return to the study of the relevant statistical problem, which is the subject of this work. The conclusions obtained in the work are tested by the method of one-dimensional optimization of the quadratic variance functional D_X.

2. The Main Results

Figure 1 shows histograms of the distributions of the experimental values of the constant lattice a, the square of its value a^2 [5], the angle of the elementary lattice α, $\cos X$ [6], argument X in the functions e^x and natural logarithm [7]. We see from Figure 1 that, contrary to the assertions [5–7], the samples of experimental data used for the testing of EPF are not subject to the normal distribution. Figure 2 presents the results of one-dimensional optimization of the functional D_X as a function m_X for direct $g(X) = X^2$ and the inverse $g^{-1}(X) = \sqrt{X}$ transformations

$$D(x,\alpha) = \int_{-\infty}^{+\infty} (x-\alpha)^2 f(x)dx. \qquad (3)$$

The results presented in Figure 1 and Figure 2 are systematized in the form of Table 1 and Table 2. From them it can be concluded that if calculated by the optimization method and calculated by EPF [5–7], the values m_Y are mutually more or less consistent with each other, which does not contradict (1), then for standard deviation, there are differences.

3. Discussion

In [5–7], experimental data with a normal probability distribution were not used to test the EPF, which did not allow to reveal the controversial nature of the formal substitution of indices in the solutions of the biquadratic equation. Therefore, we substantiate the first two moments of transformation of the normally distributed RV X functions X^2, $\cos X$, $\exp X$ and inverse to them.

EPF [5–7], direct and inverse transformations were applied to the same RV X with parameters m_X, σ_X. Thus, in [5] to calculate the standard deviation of RV X with normal distribution, transformed by the function Y quadratic radical, the quadratic equation for variance was used $\sigma_Y^2 = \overline{Y^2} - \overline{Y}^2$. In this case, from the average [10]

$$\overline{X^4} = 3\sigma_X^4 + 6m_X^2\sigma_X^2 + m_X^4, \qquad (4)$$

in the range of values $X \in (-\infty \langle X \langle +\infty)$ argument X we have two equal solutions:

$$\sigma_X^2 = m_{X^2} \pm \sqrt{m_{X^2}^2 - \frac{1}{2}\sigma_{X^2}^2}. \qquad (5)$$

By definition $\sigma_X^2 = m_{X^2} - m_X^2$, then in (5) the solution with a minus sign taken and the scheme of substitution of indexes is applied to it

$$X^2 \to X,\ X \to \sqrt{X} \qquad (6)$$

resulting in EPF in the form of relations (5) in [5].

But under the condition of the problem, the set of values of a normally distributed of RV X is an unlimited interval $X \in (-\infty, +\infty)$ and for it the statistical averages and variance are calculated, while the set of values of the argument of the function $g^{-1}(X) = \sqrt{X}$ is limited by a semi-limited interval $[0 \leqslant X \langle +\infty)$. Therefore, the transformation algorithm (6) is incorrect. Indeed, integration of the function $\sqrt{X}$in the field of values

$$[0 \leqslant X,\ m_X,\ \sigma_X \langle +\infty) \qquad (7)$$

is accompanied by the appearance of a special non-analytical (tabulated) error function $\mathrm{erf}(\xi) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-\xi^2}d\xi$ [11, 12]. Then, due to the limitation of the range of values of RV $X \in (0, x)$, the factor of normalization C_X is equal to $C_X = \frac{2}{\left(1+\mathrm{erf}\left(\frac{m_X}{\sqrt{2}\sigma_X}\right)\right)}$, and the statistical averages will be equal to:

$$\overline{X} = m_X + \sigma_X\Lambda, \quad \overline{X^2} = m_X^2 + \sigma_X^2 + \sigma_X m_X\Lambda,$$
$$\Lambda = \frac{\sqrt{2/\pi}}{1 + \mathrm{erf}\left(\frac{m_X}{\sqrt{2}\sigma_X}\right)}\, e^{-m_X^2/2\sigma_X^2}. \qquad (8)$$

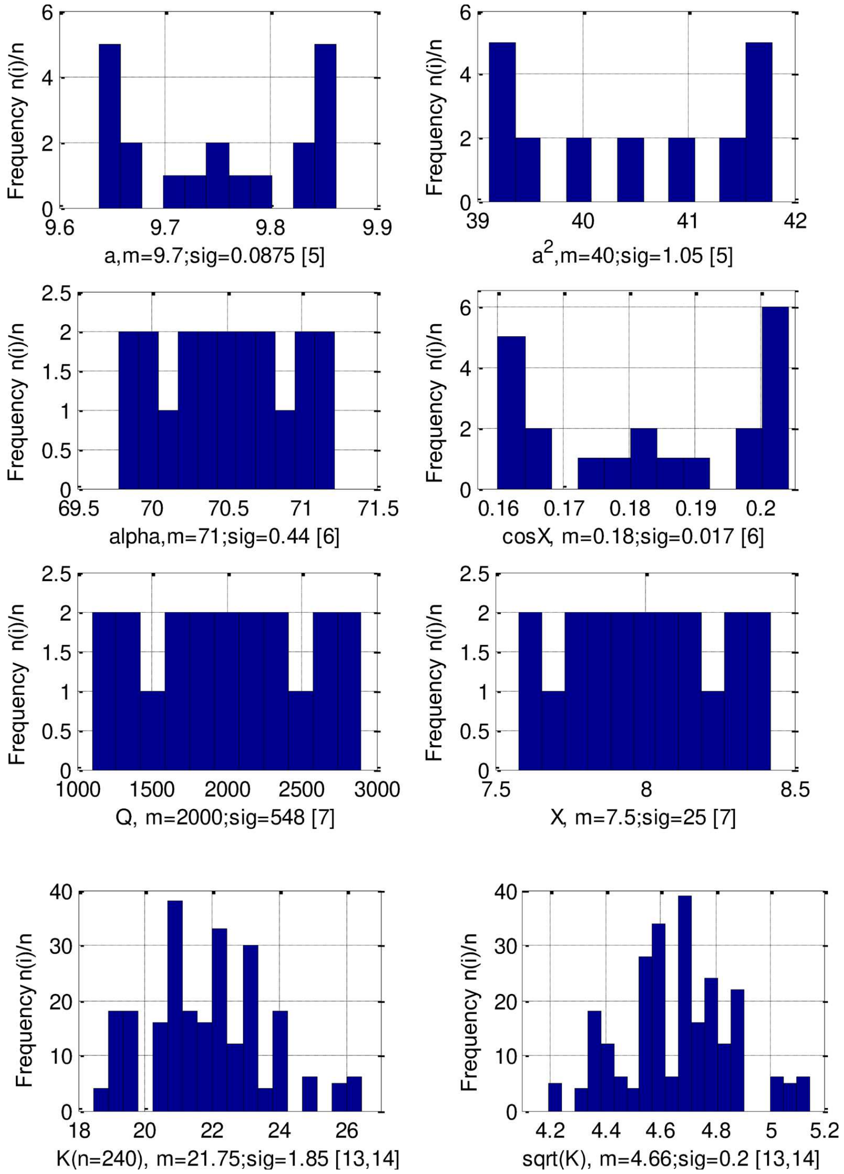


Fig. 1. Histograms of distributions of experimental data of a constant lattice a, a^2 [5], the angle of an elementary lattice $\alpha, \cos X$ [6], and values of arguments for functions e^x and natural logarithm [7]

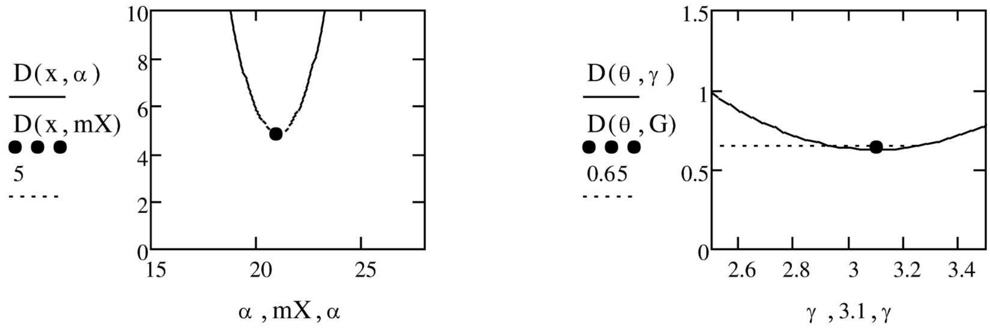
Table 1. **Calculated by optimization of variance and FPP [5] values m_Y, σ_Y, by functions Y of RV X transformations**

Y	X	X^2	X^2 [5]	$\sqrt{X}$	$\sqrt{X}$ [5]
m_Y	21	445	446	4.65	4.58
σ_Y	2.24	89.5	94.2	0.285	0.06

The variance of the RV $X \in (0, x)$ with this transformation will be equal to:

$$D_X = \sigma_X^2 + \left(m_X - \overline{X}\right)^2 + \sigma_X (2\left(m_X - \overline{X}\right) - m_X)\Lambda. \tag{9}$$

Thus, in the limited interval (7), the statistical average $\overline{X^4} \neq 3\sigma_X^4 + 6m_X^2\sigma_X^2 + m_X^4$ and the problem



a *b*

Fig. 2. Illustration of determining the mean and variance of the functions of transformations $g(X) = X$ (*a*) and $g^{-1}(X) = \sqrt{X}$ (*b*) by the method of functional optimization (3)

Table 2. **Values $\frac{\sigma_Y}{m_Y}$ determined from Figure 2 and calculated from FPP [5–7]**

Y	X	X^2	$\sqrt{X}$	$\cos\left(X\frac{\pi}{180}\right)$	$\alpha\cos\left(X\frac{\pi}{180}\right)$	$e^{X/5}$	$\ln(X/5)$
m_Y	22	474	4.65 4.68 (7) [5]	0.93	1.18 1.183	83	1.466 1.535 (13) [7]
σ_Y	1.9	86.5	0.2 0.245 (7) [5]	0.013	0.042 0.001 (11) [6]	33.3	0.0868 0.0037 (13) [7]
$V = \frac{\sigma_Y}{m_Y}$	0.086	0.18	0.043 0.05	0.014	0.036 0.001	0.4	0.06 0.0024

(4)–(5) with the subsequent replacement of the indices (6) do not have an exact analytical solution.

In principle, the parameters $m_{\sqrt{X}}, \sigma_{\sqrt{X}}$ can be calculated using tabular integrals (2.3.15 (6)) [13–14]. But their use requires other special cylindrical functions. Therefore, we give another statistical justification for the formula for calculating the standard deviation of the $\sqrt{X}$ transformation, for which we take into account the equality $\overline{\left(\sqrt{X}\right)^2} = \overline{X}$:

$$D_{\sqrt{X}} = \overline{(\sqrt{X})^2} - (\overline{\sqrt{X}})^2 = m_X + \Lambda\sigma_X - (\overline{\sqrt{X}})^2. \quad (10)$$

According to the schedule (3), for the normal distribution the approximation is true $\overline{\sqrt{X}} \cong \sqrt{m_X} - \frac{1}{4}\frac{\sigma X^2}{mX^{3/2}}$, so

$$\sigma_{\sqrt{X}} = m_X + \Lambda\sigma_X - \left(\sqrt{m_X} - \frac{1}{4}\frac{\sigma_X^2}{m_X^{3/2}}\right)^2. \quad (11)$$

For RV X with parameters $m_X \cong 21$ and $\sigma_X \cong 1.85$, standard deviation $\sigma_{\sqrt{X}} \cong 0.77$, while determined from the histogram in Figure 1, $\sigma_{\sqrt{X}} \cong 0.2$. Value $\sigma_{\sqrt{X}} \cong 0.2$ is the limit to which the standard deviation goes $\sigma_{\sqrt{X}}$ (6) at $m_X \to \infty$, since by the law of large numbers, the arithmetic mean converges in probability to its mathematical expectation.

Restrictions on the set of argument values also apply to the natural logarithm function $\ln X$ and $a\cos X$ transformation. In addition, in the case of power transformation $\overline{e^X}$, when calculating the averages $\overline{\exp X}$ and $\overline{(\exp X)^2}$, in [7] was incorrect in formula (13) on p. 737:

$$1 + \frac{\sigma_X^2}{m_X^2} m_{\ln X} = \frac{1}{2}\ln\left(\frac{m_X^4}{\sigma_X^2 + m_X^2}\right). \quad (12)$$

Its essence is that the average $\overline{\exp X}$ for normally distributed RV X is not equal to the integral

$$\overline{\exp X} \neq \frac{1}{\sqrt{2\pi}\sigma_X}\int_0^{+\infty} e^x\, e^{-\frac{(x-m_X)^2}{2\sigma_X^2}}\, dx. \quad (13)$$

In formulation (13), the subintegral expression is the product of two power functions, in one of which the power is not expressed in relative units, as is the case for the Gaussian exponent. Therefore, to ensure the correctness of the calculations, the normally distributed RV X must be expressed in relative units:

$$X \to \frac{X}{\sigma_X}, \tag{14}$$

where it is taken into account that $\sigma_X \neq 0$. Then the statistical average $\overline{\exp X}$ and $\overline{(\exp X)^2}$ in the interval $[0 \leqslant X \langle +\infty)$ will be equal to:

$$\overline{\exp\left(\frac{X}{\sigma_X}\right)} = \frac{C_X}{\sqrt{2\pi}\sigma_X} \int\limits_0^{+\infty} e^{\frac{x}{\sigma_X}}\, e^{-\frac{(x-m_X)^2}{2\sigma_X^2}}\, dx =$$
$$= \frac{C_X}{2}\left(1 + \operatorname{erf}\left(\frac{m_X + \sigma_X}{\sqrt{2}\sigma_X}\right)\right) \exp\left(\frac{m_X}{\sigma_X} + \frac{1}{2}\right), \tag{15}$$

$$\overline{\left[\exp\left(\frac{X}{\sigma_X}\right)\right]^2} = \frac{C_X}{\sqrt{2\pi}\sigma_X} \int\limits_0^{+\infty} e^{\frac{2x}{\sigma_X}}\, e^{-\frac{(x-m_X)^2}{2\sigma_X^2}}\, dx =$$
$$= \frac{C_X}{2}\left(1 + \operatorname{erf}\left(\frac{m_X + 2\sigma_X}{\sqrt{2}\sigma_X}\right)\right) \exp\left[2\left(\frac{m_X}{\sigma_X} + 1\right)\right]. \tag{16}$$

In addition, note that the simultaneous use of functions $\cos X$ and $a\cos X$ RV transformations also require an argument X in relative units, such as $X \to \to \frac{X}{\max(X)}$ given that the scope of the function argument changes $[-1, +1]$. In accordance with the standard for normally distributed RV, in this case it is possible to limit the approximation $\max(X) \cong \cong m_X + 3\sigma_X$.

A similar approach is valid in the case of the transformation of trigonometric functions of a normally distributed random variable. However, it should be borne in mind that the function $Y = \cos X$ at $0 \leqslant X \leqslant \pi$ mutually inverse function $Y = a\cos X$ in the sense in which power and logarithmic when mutually inverted $X > 0$ and the functions of squaring and taking the square root of $X \geqslant 0$. Then in the interval $[0; \pi]$, transformation $a\cos$ are numerically equal to the angle whose cosine is equal to X, then from the definition of functions $\cos X$ and $a\cos X$ it follows that $\cos(a\cos X) = X$ at $X \leqslant 1$ and $a\cos(\cos X) = X$ at $0 \leqslant X \leqslant \pi$.

4. Conclusions

This work is a continuation of the previous [8, 9] research conducted by the author, which further focuses on the problems that accompany the incorrect application of models of probability theory and mathematical statistics with a limited scattering interval for the statistical analysis of experimental data in physical systems with fluctuations. Other problems that arise in this case are described in detail in [15].

1. D.J. Hudson. *Statistics. Lectures on Elementary Statistics and Probability* (CERN, 1964).
2. Z. Pawlowsky. *Wstep do Statystyki Matematycznej* (Statystyk, 1967).
3. I.M. Kodolov, S.T. Khudyakov. *Teoreticheskie osnovy verojtnostnyh metodov v inzhenerno-economicheskich zadachah. Funkcional'nye pereobrazovanij sluchajnyh velechyn i sluchajnye fynkcii* (MADI, 1985).
4. V.I. Romanenko, N.V. Kornilovska. On the accuracy of error propagation calculations by analytic formulas obtained for the inverse transformation. *Ukr. J. Phys.* **64** (3), 217 (2019).
5. G.G. Rode. Propagation of the measurement errors andmeasured means of physical quantities for the elementaryfunctions x^2 and $\sqrt{x}$ *Ukr. J. Phys.* **62**, 148 (2017).
6. G.G. Rode. Propagation of measurement errors and measured means of a physical quantity for the elementary functions $\cos(x)$ and arccos(x) *Ukr. J. Phys.* **61**, 345 (2016).
7. G.G. Rode. Propagation of the measurement error and the measured mean of a physical quantity for elementary functions a^x and $\log_a x$. *J. Phys.* **64** (5), 345 (2019).
8. P.S. Kosobutskyy. On the simulation of the mathematical expectation and variance of samples for gaussian-distributed random variables. *Ukr. J. Phys.* **62** (2), 827 (2017).
9. P.S. Kosobutsky. Analytical relations for the mathematical expectation and variance of a standard distributed random variable subjected to the $\sqrt{X}$ transformation. *Ukr. J. Phys.* **63**, 215 (2018).
10. I.I. Gorban. *Probability Theory and Mathematical Statistics for Scientists and Engineers* (NAS of Ukraine, IPPM-MiS, 2003).
11. E. Ng, M. Geller. A table o f integrals of the error functions. *J. Research of the National Bureau of Standerds-B. Math. Sci. B* **73** (1), 1 (1969).
12. From Web Resource: Table of Integrals. 2014 From http://integral-table.com
13. A.P. Prudnikov, Yu.A. Brychkov, O.I. Marichev. *Integraly i Rjady. Elementarnye Funkcii* (Nauka, 1981).
14. From Web Resource: http://mathworld.wolfram.com /GaussianIntegral.html
15. T. Lang. Twentl y errors even you can find in biomedical research articles. *Croatian Medical J.* **45** (4), 361 (2004).

Received 28.10.19

П.С. Кособуцький

СТАТИСТИЧНИЙ АНАЛІЗ НОРМАЛЬНО РОЗПОДІЛЕНИХ ДАНИХ ІЗ ОБМЕЖЕНИМ ІНТЕРВАЛОМ РОЗСІЯННЯ ЗНАЧЕНЬ, ПЕРЕТВОРЕНИХ ПРЯМИМИ $g(x) = x^2$; $\cos x$; a^x ТА ОБЕРНЕНИМИ ДО НИХ ФУНКЦІЯМИ

Робота присвячена теоретичному аналізу коректного застосування моделі неперервної нормально розподіленої випадкової величини при обґрунтуванні так званих формул перенесення похибок в задачі статистичного опрацювання експериментальних даних. Звернуто увагу на роль обмеження інтервалу розсіяння значень випадкової величини, підданої нелінійним прямим $g(X)$ перетворенням елементарними функціями X^2; a^X та $\cos X$, і оберненими до них $g^{-1}(X) = \sqrt{X}$, $\arccos X$, $\log_a X$. Досліджено закономірності статистичного усереднення даних, одержаних шляхом розкладу функцій перетворення в ряд Тейлора. Для підтвердження правомірності одержаних результатів використано метод оптимізації квадратичного функціонала.

Ключові слова: нормальний розподіл, математичне сподівання, дисперсія, перенесення похибок, перетворення випадкової змінної елементарними функціями.