About Some Statistical Laws of the Processing of Experimental Data of Complex Systems

Petro Kosobutskyy[1], Mykhailo Lobur[1], Natalya Nestor[1], Alla Morgulis[2]

Lviv Polytechnic National University, Lviv 79012, Ukraine[1]
The City University of New York, USA[2]
petkosob@gmail.com; loburm@gmail.com

INTRODUCTION

The tasks of obtaining and processing experimental data of complex systems have always been relevant. Random misses, measurement errors, imperfections, limited mathematical models, and data processing methods can change the look of the data distribution and lead to improper use of algorithms, as is the case with Kalman filtering in control systems. More complicated methods for the identification of distribution laws require the study of quantum systems, a number of natural phenomena, environmental and biological processes, which are characterized by the presence of singularities and multimodality of distributions. Therefore, it is often not recommended to apply separate distribution laws to simulate probabilistic experimental data distributions, but a generalized distribution as a single statistical system, which known distributions include as individual partial cases.

On the other hand, it is known, that recommendation for the modelling of statistical distributions of experimental data is not to use specific/particular distribution laws, but the general distribution as the only unified statistical system, which includes all known distributions as individual partial cases. Characteristics of the so-called generalized normal distribution, which represents a superposition of the normal Gaussian distribution and exponential Laplace distribution, were investigated by the authors [1-4]. In [5] the generalized distribution, which reflects not only the variety of statistical distributions, but also takes away the need for numerous hypotheses on the statistics test, determines the affiliation of a random sample to a particular distribution law, was developed. Usage of general distributions is highly relevant in aeronautics [6], where reliable error estimation in determining the coordinates of aircraft and likely prediction of so-called rare events using ~~a~~ distribution tails is very important. In this paper, some laws such as $f_k(x) = a_k x^k e^{-p_k x^2}$,

$k = 0,1,2,3,4,..., X \in [0,+\infty)$ of the generalized distribution of Gaussian type are established.

THEORETICAL MODELLING AND DISCUSSION

The Gaussian distribution or the normal Gaussian distribution of e^{-x^2} type is widely used for statistical processing of the physical measurements results, since according to this law significant deviations from average values are unlikely, and, therefore, they can be neglected [7].

Moreover, it is assumed that the normal distribution of a random variable occurs when a significant number of random factors affects the physical system. Therefore, it is a limit for different types of distributions, which according to the central limit theorem, the sum of independent random variables with limited variance tends to the normal distribution.

The normal distribution is set by two parameters expected value and variance. However, Gaussian integrals, including integrals for the calculation of averages $\overline{X}$ and $\overline{X^2}$ can't be expressed through elementary functions. Special tabulated functions, such as gamma function, error function $erf(x)$, Laplace's function $\Phi\left(\frac{x-m_X}{\sigma_X}\right)$ [8] and justified recurrence relations [9,10], are used in calculations.

$$F_{2k+1}(x) = \frac{k!}{2}\left(1 - e^{-x^2}\sum_{s=1}^{k}\frac{x^{2s}}{s!}\right),$$
$$F_{2k}(x) = \frac{(2k-1)!!}{2^{k+1}}\sqrt{\pi}\cdot erf(x) - \frac{1}{2^k}xe^{-x^2}\sum_{j=1}^{k-1}\frac{(2k-1)!!2^k}{(2k+1)!!}x^{2k}, \quad (1)$$

where $F_k(x) = \int_0^x \xi^k e^{-\xi^2} d\xi$. For the first five values $k = 0,2,3,4$ the integrals (1) have the form:

$$\int_0^x x^0 e^{-p_{k=0}x^2} dx = \frac{1}{2p_{k=2}^{3/2}}\frac{1}{2}\sqrt{\frac{\pi}{p_{k=0}}}erf\left(\sqrt{p_{k=0}}x\right) = \frac{1}{2\sqrt{2\pi}\sigma_X}erf\left(\frac{x}{\sqrt{2\pi}\sigma_X}\right),$$

$$\int_0^x x^1 e^{-p_{k=1}x^2} dx = \frac{1}{2p^{k=1}}\left(1 - e^{-p_{k=1}x^2}\right),$$

$$\int_0^x x^2 e^{-p_{k=2}x^2}dx = \frac{1}{2p_{k=2}^{3/2}}\left(\frac{\sqrt{\pi}}{2}erf\left(\sqrt{p_{k=2}}x\right)-\sqrt{p_{k=2}}xe^{-p_{k=2}x^2}\right), \quad (2)$$

$$\int_0^x x^3 e^{-p_{k=3}x^2}dx = \frac{1}{2p_{k=3}^{2}}\left(1-(1+p_{k=3}x^2)e^{-p_{k=3}x^2}\right),$$

$$\int_0^x x^4 e^{-p_{k=4}x^2}dx = \frac{1}{p_{k=4}^{5/2}}\left(\frac{3\sqrt{\pi}}{4}erf\left(\sqrt{p_{k=4}}x\right)-\left(\frac{3}{2}+p_{k=4}x^2\right)\sqrt{p_{k=4}}xe^{-p_{k=4}x^2}\right),$$

where $x\rangle 0,\quad p_k = 1/2\sigma_k{}^2\rangle 0$. Consequently, one-parameter distributions of random variables with probability densities of the Gaussian type can be introduced into consideration :

$$f_k(x) = a_k x^k e^{-p_k x^2}, k = 0,1,2,3,4,..., X \in [0,+\infty), \quad (3)$$

where $\sigma_{k=0,1,2,3,...}$ is the scale parameter. For the first five value k, σ_k equals:

$$a_{k=0} = \frac{1}{\sqrt{2\pi}\sigma_X}, a_{k=1} = \frac{1}{\sigma_{k=1}^{2}}, a_{k=2} = \frac{\sqrt{2/\pi}}{\sigma_{k=2}^{3}}, a_{k=3} = \frac{1}{2\sigma_{k=3}^{4}}, a_{k=4} = \frac{\sqrt{2/\pi}}{3\sigma_{k=4}^{5}}, (4)$$

for which the integral probability distribution is described by functions:

$$F_k(x) = \begin{cases} k=0: \dfrac{1}{\sqrt{2\pi}\sigma_k^{2}}\int_0^x x^0 e^{-p_k x^2} = erf\left(\sqrt{p_k}x\right) \\ k=1: \dfrac{1}{\sigma_k}\int_0^x x^1 e^{-p_k x^2} = \left(1-e^{-p_k x^2}\right) \\ k=2: \sqrt{\dfrac{2}{\pi}}\dfrac{1}{\sigma_k^{3}}\int_0^x x^2 e^{-p_k x^2} = erf\left(\sqrt{p_k}x\right)-2x\sqrt{\dfrac{p_k}{\pi}}e^{-p_k x^2} \\ k=3: \dfrac{1}{2\sigma_k^{4}}\int_0^x x^3 e^{-p_k x^2} = \left(1-(1+p_k x^2)e^{-p_k x^2}\right) \\ k=4: \sqrt{\dfrac{2}{\pi}}\dfrac{1}{3\sigma_k^{5}}\int_0^x x^4 e^{-p_k x^2}dx = \left(erf\left(\sqrt{p_k}x\right)-\left(\dfrac{2}{3}+\dfrac{4}{3}p_k x^2\right)\sqrt{\dfrac{p_k}{\pi}}xe^{-p_k x^2}\right) \\ .... \end{cases} \quad (5)$$

Functions (5) satisfy the basic requirements for the distributions of integral probabilities:

$\lim_{x_i \to 0} F(x_i) = 0; \quad \lim_{x_i \to +\infty} F(x_i) = 1; \quad F(x_i) \geq$ *for all* $x_i; \quad F(x_i) \geq F(x_j), \quad$ *if* $\quad x_i \rangle x_j$

Unlike the normal ($k = 0$) distribution where the most probable value of a random variable X coincides with the expected value $m_{k=0} = m_X$, distributions of type (2), where $k \rangle 0$, expected value m_k is shifted relative to the most probable value $x_{i,k}$,as a value of a coordinate $x_{\max,k}$, of an extremum of a function $x^k \exp\left(-px^2\right)$, which is defined as the solution of the differential equation $\frac{d}{dx}\left(x^k e^{-p_k x^2}\right) = 0$ i.e. the solution of equation $k \cdot x_{\max,k}{}^k = x_{\max,k}{}^2 / \sigma_k^2$, where the coordinate of the extremum (mode) will be equal to: $x_{\max,k} = \sqrt{k}\sigma_k$.

The mean $\overline{X_k}$ of the random variable X with a semibounded $X \in [0,+\infty)$ variance interval distributed by law (2) is equal to:

$\overline{X_k} == \frac{\Gamma\left(\frac{k+2}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)}\sqrt{2}\sigma_k$. The mean $\overline{X_k}$ coincides with the most probable value $x_{i,k}$ only for normal $N(\sigma_X, 0)$ distribution. If $k \rangle 0$, then the mean value moves to the region of larger values by value

$\Delta G_k = \sigma_k \left(\Gamma\left(\frac{k+2}{2}\right) \Big/ \Gamma\left(\frac{k+1}{2}\right)\sqrt{2} - \sqrt{k} \right)$ .The mean value $\overline{X_k{}^2}$ of the random variable X with a semibounded $X \in [0,+\infty)$ variance interval distributed by law (2) is equal to:

$$\overline{X_k{}^2} = \frac{a_k \int_0^\infty x^{k+2} e^{-p_k x^2} dx}{a_k \int_0^\infty x^k e^{-p_k x^2} dx} = \frac{\Gamma\left(\frac{k+3}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)} 2\sigma_k^2. \qquad (6)$$

Formula (6) makes it possible to formulate a new equation for the variance of statistically independent variables:

$$D_k = \overline{X_k^{\,2}} - \left(\overline{X_k}\right)^2 = 2\sigma_k^2 \left\{ \frac{\Gamma\left(\frac{k+3}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)} - \left[\frac{\Gamma\left(\frac{k+2}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)} \right]^2 \right\}. \qquad (7)$$

For the first five values $k = 0,1,2,3,4$ the variance is equal to:

$$D_k = 2\sigma_k^2 \left\{ \frac{\Gamma\left(\frac{k+3}{2}\right)}{\Gamma\left(\frac{k+1}{2}\right)} - \frac{\Gamma\left(\frac{k+2}{2}\right)^2}{\Gamma\left(\frac{k+1}{2}\right)^2} \right\} = \begin{cases} \sigma_{k=0}^2 \left(1 - \frac{2}{\pi}\right) \\ \sigma_{k=1}^2 \left(2 - \frac{\pi}{2}\right) \\ \sigma_{k=2}^2 \left(3 - \frac{2^3}{\pi}\right) \\ \sigma_{k=3}^2 \left(4 - \frac{3^2\pi}{2}\right) \\ \sigma_{k=4}^2 \left(5 - \frac{2^7}{3^2\pi}\right) \end{cases}. \qquad (8)$$

It folows from the relations (7) and (8), that in systems with dispersion, mean $\overline{X_k} \neq 0$, because $D_k \neq 0$, since with the shifting the start of the origin the classical normal distribution $N(\sigma_X, m_X)$ linearly converts to the standard form $N(\sigma_X, 0)$. But linear transformation of the random variable does not change the nature of its distribution, including the legitimacy of inequality $\overline{X_k} \neq 0$ in systems with dispersion.

Variance D_X is a fundamental concept of statistics [9]. Variance binds first and second moments and characterizes the intensity of fluctuations, therefore the case $D_X = 0$ has no physical meaning*). When

*) In probability theory, variance is a measure of dispersion from the mean, whereas in mathematical statistics, it characterizes the degree of dispersion of the quantitative values of the statistical sample relatively to the average – expected value of the square of the deviation of random variable from expected value.

the second *moment* of distribution characterizes dispersion relatively to the origin, variance - relatively to the average value. In addition, variance has a dimension of the square of the random variable, so second additional variable σ_X is introduced which has the same dimension as the variable x. From the physical point of view, variance for a deterministic value is absent, but in fact, in the same time, the second moment is not equal to zero. Indeed, the deterministic value is located at a certain distance from the origin, which
is not equal to zero, so the second moment is also not equal to zero.

For statistically independent random variables, the equation for variance can be represented as:

$$\sigma_k^2 + \overline{X_k}^2 = \overline{X_k^2}. \qquad (9)$$

It is known from mathematics that the relation (8) is a semicircle equation, if the dispersion region of a random mathematical expectation m_X is unlimited $-\infty\langle m_X\langle+\infty$, or a quaternary equation of the circle, if the dispersion region of a random variable is truncated $0 \le m_X\langle+\infty$, with the radius $\sqrt{\overline{X^2}}$ the center at the origin. Therefore, in two-dimensional probability space, dynamic random variables form, generally an ellipse equal probabilities and the equation (8) is convenient to rewrite in its canonical form:

$$\left(\frac{\sigma_k}{\sqrt{\overline{X_k^2}}}\right)^2 + \left(\frac{\overline{X}_k}{\sqrt{\overline{X_k^2}}}\right)^2 = 1 \quad \text{or} \quad \left(\frac{\sqrt{D_k}}{\sqrt{\overline{X_k^2}}}\right)^2 + \left(\frac{\sqrt{\left(\overline{X_k}\right)^2}}{\sqrt{\overline{X_k^2}}}\right)^2 = 1, \qquad (10)$$

where for the distributions of type (2), relation (generalized Gaussian of ratio [4]).

CONSLUSION

Mean values $\overline{X_{k=3}}$ and $\overline{X_{k=3}^2}$ and dispersion equation of a random variable X with distributions of probability density of Gaussian type $f_k(x) = a_k x^k e^{-p_k x^2}$, $k = 0,1,2,3,4,..., X \in [0,+\infty)$ are expressed through elementary functions of a dispersion parameter σ_k. Unlike the standard $N(\sigma_X,0)$ distribution, moda of the distributions $f_k(x) = a_k x^k e^{-p_k x^2}$ is a

function of a scale $\sigma_{k=0,1,2,3,...}$, which allows more effectively configure the theoretical model for the experimental data using one-parameter optimization.

REFERENCES

[1] M.K. Varanasi, B. Aazhang, "Parametric generalized Gaussian density estimation". *Journal of the Acoustical Society of America* **86** (4): 1404–141,. October 1989.

[2] S.Nadarajah, "A generalized normal distribution". *Journal of Applied Statistics* **32** (7): 685–694, September 2005.

[3] S. Fabian, G. Sebastian; B. Matthias, "Characterization of the p-Generalized Normal Distribution". Journal of Multivariate Analysis. 100 (5): 817–820, 2009.

[4] D. Armando; G.González-Farías,R.Graciela, M.Ramón, "A practical procedure to estimate the shape parameter in the generalized Gaussian distribution". *http://www.cimat.mx/reportes/enlinea/I-01-18_eng.pdf*; *www.cimat.mx.*

[5] V.V. Nashhitojy, "Methods of statistical analysis on the basis of generalized distributions". Minsk: Vedas, 2001. - 168 p.

[6] A unified Framework for Collision Risk Modeling in Support of the Manual on Airspace Planning Methodology for the Determination of Separation Minima. (Doc 9689).-*Montreal. ICAO, Cir319-AN/181*,2009.

[7] J. K. Patel, C. B. Read, "*Handbook of the Normal Distribution"*. New York: Dekker, 1982.

[8] E. Jahnke, F. Emdw, F. Losch, "*Tables of Higher Functions"*. McGraw-Hill, New York,1960.

[9] A. Hald, «*Statistical Theory with Engineering Applications"*. New York-London, 1952.