Measurement Techniques, Vol. 50, No. 6, 2007

STATISTICAL MODELING AS AN EFFECTIVE INSTRUMENT FOR INVESTIGATING THE DISTRIBUTION LAWS OF FUNCTIONS OF RANDOM QUANTITIES

B. Yu. Lemeshko and D. V. Ogurtsov

UDC 519.245:006.91.001

The probability distribution laws of different functions of random quantities, which obey different one-dimensional distribution laws, are investigated using specially developed software employing statistical modeling methods. The effectiveness of the procedure for investigating probability laws is demonstrated.

Key words*: statistical modeling, distribution law of a function of random quantities.*

The problem of determining the probability characteristics of a quantity Y, which is not measurable directly, using multiple measurements of accessible quantities $X_1, X_2, ..., X_k$, is quite often solved in metrology, if the function

$$Y = \varphi(X_1, X_2, ..., X_k)$$

is known or, in vector form $Y = \varphi(\bar{X})$, and the joint probability distribution of the input variables $X_1, X_2, ..., X_k$ is obtained from the results of a statistical analysis.

The classical approach [1, 2] to determining the probability distribution law of a function of a system of random quantities presupposes a knowledge of the joint probability density $f(x_1, x_2, ..., x_k)$ of the system of random quantities $X_1, X_2, ..., X_k$. However, an analytical solution using the classical approach can only be obtained for certain special cases of $Y = \varphi(\bar{X})$ and $f(x_1, x_2, ..., x_k)$ [2].

As a consequence of this, to determine the probability characteristics of the output variable of the model $Y = \varphi(\bar{X})$ when the input variables $X_1, X_2, ..., X_k$ are uncorrelated, linearization of the model

$$Y \approx \varphi(\bar{M}) + (\bar{X} - \bar{M})^{\mathrm{T}} \nabla\varphi(\bar{M}), \tag{1}$$

where $\bar{M}$ is the vector of the mathematical expectations of $\bar{X}$ and $\nabla\varphi(\cdot)$ is the gradient of the function, is recommended in [3].

This approach enables one to determine the corresponding characteristics of a random quantity Y from the distribution laws of the input variables $X_1, X_2, ..., X_k$ or their numerical characteristics fairly simply. Unfortunately, this approach also turns out to be effective in relatively rare cases when the function $\varphi(\bar{X})$ is close to linear.

In [4], using the example of the function $Y = X_1/X_2$, a difference in the solutions obtained using the classical approach and the linearization method is demonstrated and the unacceptably large errors, to which the use of the latter method leads, are emphasized.

Nevertheless, the linearization method is widely used in practice, including in data-measuring systems, which make use of indirect measurements. For example, this approach is used in [5] to investigate the metrological characteristics of mul-

Novosibirsk State Technical University; e-mail: lemeshko@fpm.ami.nstu.ru. Translated from Metrologiya No. 5, pp. 3–13, 2007. Original article submitted February 28, 2007.

TABLE 1. Results of a Check of the Goodness of Fit of a Sample of the Quantity $Y = X_1/X_2$ with the Standard Cauchy Distribution in the Case $X_1, X_2 \in N(0,1)$

Criterion	Value of the statistics	Level of significance achieved
Pearson χ^2 for $k = 15$	12.5470	0.5625
Kolmogorov	0.6087	0.8526
Mises ω^2	0.0521	0.8636
Anderson–Darling Ω^2	0.3416	0.9040

TABLE 2. Results of a Check of the Goodness of Fit of a Sample of the Quantity $Y = X_1/X_2$ with a Cauchy Distribution in the Case $X_1 \in N(0, 4)$, $X_2 \in N(0, 0.3)$

Criterion	Value of the statistics	Level of significance achieved
Pearson χ^2 for $k = 15$	10.7130	0.7084
Kolmogorov	0.7169	0.6829
Mises ω^2	0.0530	0.8577
Anderson–Darling Ω^2	0.3202	0.9222

tichannel data-measuring systems with multiplicative interaction of the channels, which, as a result of linearization, is replaced by additive interaction. We would expect that the use of this method in the present situation [5] should lead to understated estimates of the measurement error.

Hence, an analytical solution using the classical approach cannot be obtained in the majority of practical situations, while linearization leads to inadequate solutions. The purpose of the present paper is to draw the attention of metrologists to the effectiveness of the Monte Carlo method for investigating probability laws, and its capabilities when constructing probability models of functions of random quantities, and to refine the probability characteristics of the errors of indirect measurements. Unfortunately, the Monte Carlo method is undeservedly rarely used to investigate probability laws in Russian publications.

To investigate the distribution laws of functions of random quantities, we have developed software which enables a sample of such functions to be modeled. An interface enables arbitrary functions of a system of independent (as yet) random quantities, having different one-dimensional distribution laws to be specified.

We will consider several examples which demonstrate the accuracy of statistical modeling and its effectiveness when investigating the behavior of the distribution laws of functions of random quantities.

Obviously, the distribution of the function Y depends considerably on the form of the laws which the random quantities X_i obey, and on the region in which they are defined. Moreover, the function Y of X_i, having some form of distribution law, can be described by very different models of probability laws depending on the parameters of the laws describing the random quantities X_i. We will show this using the function $Y = X_1/X_2$ when X_i obey normal laws.

Example 1. $Y = X_1/X_2$, where $X_1, X_2 \in N(0, 1)$ and are independent. The theoretical distribution law of Y is the standard Cauchy distribution with density $f(y) = [\pi(1 + y^2)]^{-1}$, $y \in (-\infty, +\infty)$. In Table 1, we show the results of a check of the agreement between the modeled sample of the quantity Y and the Cauchy distribution. In this and in the other cases, the volumes of modeled samples amounted to 10000 values. In the case of the Pearson χ^2 criterion, asymptotically optimum grouping is used which ensures the maximum power of relatively close competing hypotheses [6]. In Table 1, we show values of the statistics of the goodness of fit tests employed, calculated from the sample and the level of significance achieved for each criterion [6, 7]. The achieved level of significance is the probability $P\{S > S^*\}$, where S^* is the value of the statis-

TABLE 3. Results of a Check of the Goodness of Fit of a Sample of the Quantity $Y = X_1/X_2$ with the Distribution Obtained from Formula (2), in the Case When $X_1 \in N(1, 1)$ and $X_2 \in N(1, 1)$

Criterion	Value of the statistics	Level of significance achieved
Pearson χ^2 for $k = 15$	17.4060	0.2351
Kolmogorov	0.8085	0.5302
Mises ω^2	0.1462	0.4012
Anderson–Darling Ω^2	0.9025	0.4126

TABLE 4. Results of a Check of the Goodness of Fit of a Sample of the Quantity $Y = X_1/X_2$ with a Normal Distribution in the Case $X_1 \in N(1, 1)$, $X_2 \in N(10, 1)$

Criterion	Value of the statistics	Level of significance achieved
Pearson χ^2 for $k = 15$	23.2720	0.0255
Kolmogorov	0.6501	0.3991
Mises ω^2	0.0870	0.1564
Anderson–Darling Ω^2	0.6545	0.0827

tics S of the corresponding criterion calculated from the sample. The hypothesis of the goodness of fit of the empirical distribution with the theoretical one using the corresponding criterion is turned down if $P\{S > S^*\} < \alpha$, where α is a specified error probability of the first kind. In this case, the level of significance achieved using all the criteria indicates very good agreement between the empirical distribution obtained by modeling and the Cauchy distribution.

Example 2. $Y = X_1/X_2$, where $X_1 \in N(0, 4)$, $X_2 \in N(0, 0.3)$ and are independent. The theoretical distribution law of Y is the Cauchy distribution with density $f(y) = 1.2/\pi(16 + 0.09y^2)$, $y \in (-\infty, +\infty)$. The results of a goodness of fit check of the modeled sample with a Cauchy distribution are presented in Table 2.

Example 3. $Y = X_1/X_2$, where $X_1 \in N(a, 1)$ and $X_2 \in N(b, 1)$ are independent. When $a = b = 1$, the distribution law of Y is not a Cauchy distribution. An estimate of the parameters of the Cauchy density $f(y) = \theta_1/\pi(\theta_1^2 + (y - \theta_2)^2)$ using the modeled sample gives maximum-likelihood estimates of the scale parameters $\theta_1 = 0.7895$ and the shift $\theta_2 = 0.6150$. The estimate of the shift parameter is identical with the median empirical distribution. A check of the goodness of fit of the empirical distribution obtained by modeling with the Cauchy distribution law deviates with respect to all the criteria. This is a consequence of the fact that the actual distribution of the quantity Y in this case has become explicitly asymmetric.

In the general case, the distribution density of a particular Y when $X_1 \in N(a, 1)$ and $X_2 \in N(b, 1)$ can be distributed in the form [8, 9]:

$$f(y) = \frac{\exp(-(a^2 + b^2)/2)}{\pi(1 + y^2)}\left\{1 + \sqrt{2\pi}q\exp(q^2/2)[\Phi(q) - \Phi(0)]\right\}, \qquad (2)$$

where $q = (b + ay)/\sqrt{1 + y^2}$; and $\Phi(z)$ is the distribution function of the standard normal law.

The results of a check of the goodness of fit of the modeled sample with distribution (2) for $a = b = 1$ are shown in Table 3.

Example 4. $Y = X_1/X_2$, where $X_1 \in N(1, 1)$ and $X_2 \in N(10, 1)$ and are independent. For a considerable excess of the absolute value of the shift parameter X_2 above the shift X_1, a good model for Y is a normal distribution. In Fig. 1, we show

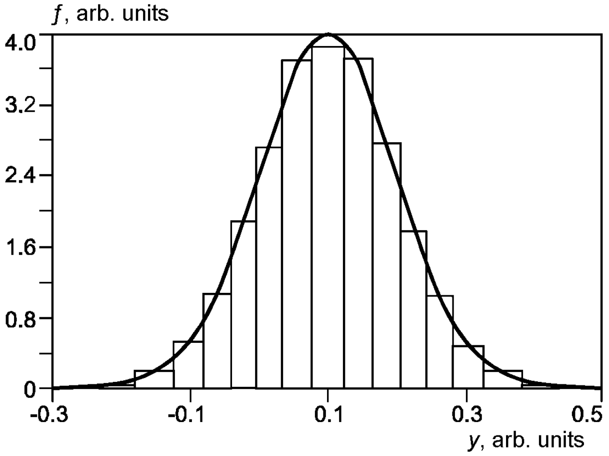

Fig. 1. Density and histogram of the distribution of $Y = X_1/X_2$ for $X_1 \in N(1, 1)$, $X_2 \in N(10, 1)$.

TABLE 5. Results of a Check of the Goodness of Fit of a Sample of the Quantity $Y = X_1/X_2$ with Distribution (2) in the Case When $X_1 \in N(1, 1)$, $X_2 \in N(10, 1)$

Criterion	Value of the statistics	Level of significance achieved
Pearson χ^2 for $k = 15$	17.0230	0.2549
Kolmogorov	0.9115	0.3770
Mises ω^2	0.1462	0.4012
Anderson–Darling Ω^2	1.0833	0.3163

TABLE 6. Results of a Check of the Goodness of Fit of a Sample of the Quantity $Y = X_1/X_2$ with Distribution (2) in the Case of When $X_1 \in N(10, 1)$, $X_2 \in N(1, 1)$

Criterion	Value of the statistics	Level of significance achieved
Pearson χ^2 for $k = 15$	14.9270	0.3831
Kolmogorov	0.8050	0.5359
Mises ω^2	0.1261	0.4710
Anderson–Darling Ω^2	0.92532	0.3989

a histogram, constructed for the empirical distribution with 15 intervals and an asymptotically optimal grouping [6, 10], and a normal distribution density with an estimate of the scale parameter $\theta_1 = 0.10051$ and of the shift parameter $\theta_2 = 0.10067$. In Table 4, we show the results of a check of the goodness of fit of the modeled sample with a normal distribution.

In Example 4 considered, the density of $Y = X_1/X_2$ has the form (2) with $a = 1$ and $b = 10$. The results of a check of the goodness of fit of the modeled sample with distribution (2) are shown in Table 5.

This example for the function $Y = X_1/X_2$ is a case when the use of linearization turns out to be legitimate. Linearization gives a normal distribution with a mathematical expectation of 0.1 and a variance of 0.0101, i.e., normal with a scale parameter $\theta_1 = 0.1$ and a shift parameter $\theta_2 = 0.100499$, which approximates to the true position of the objects.

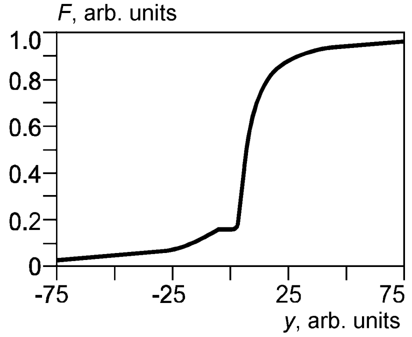

Fig. 2. Empirical distribution of $Y = X_1/X_2$ for $X_1 \in N(10, 1)$, $X_2 \in N(1, 1)$.

TABLE 7. Confidence Intervals for a Product of $k = \overline{2, 5}$ Normal Standard Random Quantities

k	90% interval	95% interval
1	–1.645; 1.645	–1.960; 1.960
2	–1.627; 1.603	–2.185; 2.167
3	–1.314; 1.356	–1.980; 2.057
4	–1.053; 1.116	–1.836; 1.934
5	–0.868; 0.749	–1.463; 1.296

TABLE 8. Confidence Intervals for a Product of $k = \overline{2, 5}$ Normal Random Quantities with Shift and Scale Parameters Equal to Unity

k	90% interval		95% interval	
	actual	linearized	actual	linearized
1	–0.645; 2.645	–0.645; 2.645	–0.960; 2.960	–0.960; 2.960
2	–1.191; 4.360	–1.326; 3.326	–1.828; 5.293	–1.772; 3.772
3	–1.870; 6.104	–1.849; 3.849	–2.850; 8.375	–2.395; 4.395
4	–2.419; 7.540	–2.290; 4.290	–4.158; 10.942	–2.920; 4.920
5	–3.157; 8.483	–2.678; 4.678	–5.620; 13.793	–3.383; 5.383

As the absolute value of the shift parameter X_2 increases with respect to the shift X_1, the distribution of Y approaches a normal distribution (with equality of the variances). As the variance of X_2 increases, the distribution of Y begins to deviate from a normal distribution. Under these conditions, as the variance of X_1 increases with respect to the variance of X_2, the distribution is well approximated by a normal distribution, and the use of linearization leads to a normal law with a more appreciable shift with respect to the true distribution law.

When the standard deviation of X_i is much less than its mathematical expectation and the distributions of X_i are close to normal, the distribution of $Y = X_1/X_2$ is well approximated by a normal law, and its linearization also gives good results.

Example 5. $Y = X_1/X_2$, where $X_1 \in N(10, 1)$, $X_2 \in N(1, 1)$ and are independent. The distribution density of Y in this case has the form of (2) for $a = 10$ and $b = 1$. The results of a check on the goodness of fit of the modeled sample with distri-

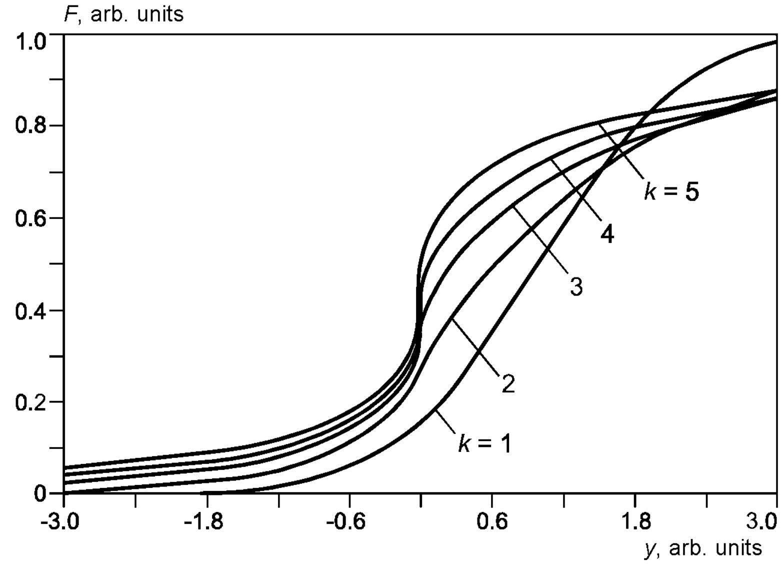

Fig. 3. Empirical distributions of products of $k = \overline{1, 5}$ normal quantities with shift and scale parameters equal to unity.

bution (2) are shown in Table 6. The form of the empirical distribution function, obtained by modeling, is represented in Fig. 2. It is obvious that it can be well described by a certain mixture of distributions, the analytical form of which differs from law (2).

In [5], the use of linearization is specified for the distribution law of the product of random quantities. We will determine the correctness of the use of linearization in this case.

Suppose $Y = \prod_{i=1}^{k} X_i$, where X_i are uncorrelated random quantities with mathematical expectation M_i and variance D_i. According to (1),

$$Y \approx \sum_{i=1}^{k} X_i \prod_{\substack{j=1 \\ j \neq i}}^{k} M_j - (k-1) \prod_{j=1}^{k} M_j ,$$

the mathematical expectation $E[Y] \approx \prod_{j=1}^{k} M_j$ and the variance

$$D[Y] \approx \sum_{i=1}^{k} D_i \left(\prod_{\substack{j=1 \\ j \neq i}}^{k} M_j \right)^2 .$$

The distributions of the products were investigated for different distribution laws of X_i.

When X_i belongs to standard normal laws for $k = \overline{2, 5}$, the use of linearization is impossible, since the variance turns out to be zero. The distributions of Y obtained by modeling are asymmetric laws with zero median. These distributions cannot be adequately described by any one parametric model of the law, but they can be sufficiently well approximated by mixtures of the form

$$\alpha \frac{\theta_3}{2\theta_2 \Gamma(1/\theta_3)} \exp\left\{ -\left(\frac{|y - \theta_1|}{\theta_2} \right)^{\theta_3} \right\} + (1-\alpha) \frac{1}{\theta_1} \exp\left\{ \frac{y - \theta_4}{\theta_5} - \exp\left(\frac{y - \theta_4}{\theta_5} \right) \right\} .$$

TABLE 9. Results of a Check of the Goodness of Fit with the Distribution (3) of a Sample of Values of the Function $Y = \sin X_1 \cos X_2 X_3 - \sin X_4 \cos X_5 X_6$

Criterion	Value of the statistics	Level of signficance achieved
Pearson χ^2 for $k = 15$	15.0640	0.1796
Kolmogorov	0.4796	0.8847
Mises ω^2	0.0423	0.7028
Anderson–Darling Ω^2	0.3647	0.6035

In Table 7, we show, for this case, the actual confidence intervals for Y, which emphasize the asymmetric form of the laws.

In Fig. 3, we show empirical distributions of similar products of random quantities, but which belong to a normal law with shift and scale parameters equal to unity. The distributions of Y for this case for $k = \overline{2, 4}$ are described quite well by mixtures of two parametric models, and for $k = 5$ by mixtures of three parametric models. The differences between the actual confidence intervals and those obtained as a result of linearization are shown in Table 8.

When the accuracy of measurements of X_i is increased (with a reduction in D_i) the distribution $Y = \prod_{i=1}^{k} X_i$ approaches a normal law. For example, when $M_i = 1$ and $D_i = 10^{-4}$ the distribution of Y agrees quite well with a normal law $N(0.99973, 0.02232)$, constructed from a modeled sample. In this case, linearization gives identical results $N(1, 0.02236)$. The same is observed when M_i increases for constant D_i.

In conclusion, we will illustrate to what extent good models can be constructed for arbitrary functions for systems of random quantities. For example, for the function $Y = \sin X_1 \cos X_2 X_3 - \sin X_4 \cos X_5 X_6$, where $X_1 \in N(0, 1)$, $X_2 \in \mathrm{rav}(0, 1)$, $X_3, X_6 \in \exp(0, 1)$, $X_4 \in N(0, 4)$, and $X_5 \in \mathrm{rav}(0, 2)$, i.e., they belong to normal, uniform, and exponential laws with the indicated shift and scale parameters, a very good model turns out to be a distribution with density

$$f(y) = \frac{\theta_3}{2\theta_2 \Gamma(1/\theta_3)} \exp\left\{ -\left(\frac{|y - \theta_1|}{\theta_2} \right)^{\theta_3} \right\} \tag{3}$$

and estimates of the parameters $\theta_1 = 0.0014$, $\theta_2 = 0.4461$, and $\theta_3 = 0.7922$. The degree of closeness of the empirical distribution obtained to the theoretical one (3) is indicated by the high levels of significance reached with respect to the goodness of fit criteria employed (when checking complex hypotheses [6, 7]), represented in Table 9.

Hence, methods of statistical modeling, in conjunction with appropriate software, which enable approximate mathematical models to be constructed for empirical distributions (including in the form of mixtures of different parametric laws), are an effective instrument for investigating the distribution laws of functions of random quantities and for investigating probability laws, which arise in problems of metrology.

The distributions of functions of random quantities X_i do not depend solely on the form of the distribution laws of X_i and may change over wide ranges depending on the parameters of these laws. Using methods of statistical modeling to investigate the distribution law of Y, one can either construct an approximate model, which approximates this law in a specific case, or investigate the conditions which justify the use of linearization.

An increase in the accuracy of measurements of X_i under certain conditions, although by no means always, helps to make the distribution of the value of Y, representing the function X_i, become closer to a normal law.

The use of statistical modeling and specialized software, an example of which is the Interval Statistics ISW system [11], enables good approximate mathematical models of the distribution laws of functions of random quantities to be constructed (including in the form of mixtures of parametric models of the laws), when this law cannot be found analytically.

This research was supported by the Russian Foundation for Basic Research (Project No. 06-01-00059-a) and the Ministry of Education and Science of the Russian Federation (Project No. 2006-RI-19.0/001/119).

REFERENCES

1. V. P. Chistyakov, *A Course in Probability Theory* [in Russian], Nauka, Moscow (1982).
2. E. I. Gurskii, *Theory of Probability with Elements of Mathematical Statistics* [in Russian], Vysshaya Shkola, Moscow (1971).
3. MI 2083-90, GSI, *Indirect Measurements: Determination of the Results of Measurements and the Estimation of Their Errors.*
4. S. F. Levin, *Izmer. Tekh.*, No. 3, 5 (2004); *Measurement Techniques*, **47**, No. 3, 216 (2004).
5. V. P. Shevchuk and D. N. Lyasin, *Izmer. Tekh.,* No. 10, 16 (2004); *Measurement Techniques*, **47**, No. 10, 969 (2004).
6. R50.1.033-2001, *Recommendations on Standardization: Applied Statistics: Rules for Checking the Goodness of Fit of an Experimental Distribution with a Theoretical One: Part I: Chi-Squared Type Criteria.*
7. R50.1.037-2002, *Recommendations on Standardization: Applied Statistics: Rules for Checking the Goodness of Fit of an Experimental Distribution with a Theoretical One: Part II: Nonparametric Criteria.*
8. G. Marsaglia, *J. Amer. Statist. Ass.*, **60**, 193 (1965).
9. G. Marsaglia, *J. Statist. Software*, **16**, No. 4 (2006), http://www.jstatsoft.org/vll/i04/.
10. B. Yu. Lemeshko, *Zavod. Lab.*, **64,** No. 1, 56 (1998).
11. B. Yu. Lemeshko and S. N. Postovalov, *Computer Technologies for Data Analysis and for Investigating Statistical Laws* [in Russian], Izd. NGTU, Novosibirsk (2004).