Software support for simulating and investigating the distribution laws of functions of random variables

Boris Yu. Lemeshko
Faculty of Applied Mathematics and Informatics (FAMI)
Novosibirsk State Technical University
20 Karl Marx Prospekt, Novosibirsk 630092
Russia
lemeshko@fpm.ami.nstu.ru

Dmitri V. Ogurtsov
FAMI
dogurtsov@gmail.com

Abstract

The probability distribution laws of various functions of random variables which obey different distribution laws are investigated using developed software applying statistical modeling methods. The effectiveness of this approach for investigating probability laws is demonstrated.

1 Introduction

In practice of statistical analysis of the number of problem definition is much more than proposed solutions in the classical mathematical statistics. A variety of distribution laws and a various complexity of functions of random variables and systems of random variables makes the use of the classical instrument s for determining the distribution function of the law is very time-taking task, which often don't have analytical solutions.

As a result various approximations are proposed, which allows simply enough to find the numerical characteristics of interest law distribution function. Unfortunately these approaches apply very rigorous restrictions on the tasks.

Generally the lack of necessary theoretical results is explained by the complexity and laboriousness of the decision obtained by analytic methods.

The development of computer methods for the study of statistical regularities, the properties of estimates and statistics of the various criteria of statistical hypotheses, building probabilistic models for the investigated regularities makes it less costly to obtain basic intellectual knowledge of mathematical statistics and, therefore, to get statistical conclusions in the analysis of data in various applications areas.

2 Problem definition

Required to determine the probabilistic characteristics of the variable Y inaccessible to direct measurement based on the available for multiple measurements of variables $X_1, X_2, ..., X_k$. It is assumed that

$$Y = \phi(X_1, X_2, ..., X_k) \tag{1}$$

where $\phi(\cdot)$-some known function. It is assumed that the distribution law of vector $\bar{X}$, or in the case of the independence of its components, the laws of the distribution of $X_1, X_2, ..., X_k$ (it may be the laws of distribution of errors of measurements) are known or can be obtained on the basis of the results of statistical analysis.

The classical approach for determining the law of probability distribution functions of random variables requires knowledge of the joint density $f(x_1, x_2, ..., x_k)$ of the random variables $X_1, X_2, ..., X_k$.

Let $X : \Omega \to R^n$ - random variable and $g : R^n \to R^n$ - continuously differentiable function such that $J_g(x) \neq 0, \forall x \in R^n$, where $J_g(x)$ - Jacobians of the function g at point x. Then the random variable is also absolutely continuous and its density has the form:

$$f_Y(y) = f_x(g^{-1}(y)) \mid J_{g^{-1}}(y) \mid \tag{2}$$

However, analytical solution obtained with the classical approach can be found only at some particular cases of the functions $Y = \phi(\bar{X})$ and density $f(x_1, x_2, ..., x_k)$.

As a result in determining the probability characteristics of the results of indirect measurements, described by the model $Y = \phi(\bar{X})$, in the case of noncorrelatedness of components $X_1, X_2, ..., X_k$ of vector $\bar{X}$ measured variables model linearization are recommended [(1)]

$$Y \approx \phi(\bar{M}) + (\bar{X} - \bar{M})^T \nabla\phi(\bar{M}) \tag{3}$$

where $\bar{M}$ - means vector, $\nabla\phi(\cdot)$ - the gradient function. This approach is simply to determine the characteristics of the random variable Y. Unfortunately, this approach is effective also in the comparatively rare cases when close to a linear function of $\phi(\bar{X})$.

3 Software description

Methods of computer modeling and analysis of statistical regularity supposes the development of software for investigation. Software is developed for the simulation of random functions of random variables with different distribution laws.

The software allows to solve the following tasks:

- Simulation samples of random variables with a given distribution law;
- Simulation samples of random vectors;
- Simulation samples of functions of random variables;
- Simulation samples of functions of random vectors.

The software consists of two programs. The first program (≪Operations with one-dimensional random variables≫) makes it possible to simulate the functions of one-dimensional random variables (Figure 1), second (≪Operations with multidimensional random variables≫) - the function of multidimentional random variables (Figure 2).

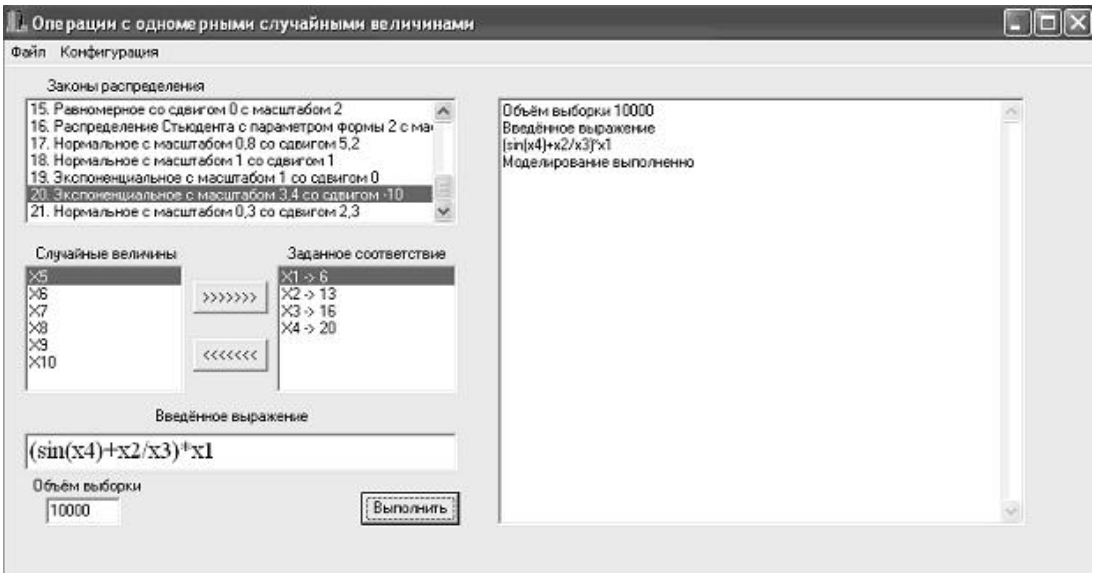

Figure 1: Dialog box first program

4 Results of experiments

In [(2)] on example the function $Y = X_1/X_2$ shows the difference in the solutions obtained using the classical approach and as a result of linearization, underscores the unallowable large errors resulting from the use of the linearization method. At the same time [(3)] demonstrated the effectiveness and collection of tasks, in which can be applied the method of statistical modeling.

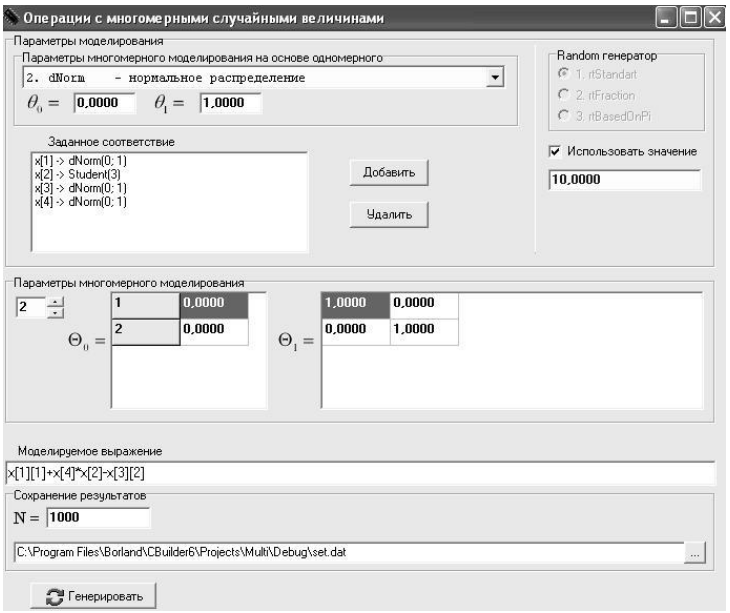

Figure 2: Dialog box second program

Consider some examples showing the accuracy of modeling.

To verify the goodness used following criteria: χ^2 Pirson with $k = 15$, Kolmogorov, ω Mises, Ω Anderson–Darling

Example 1. $y = 2x_1 - 5x_2 + 3x_3$, where $x_1 \in N(3, 7)$, $x_2 \in N(1, 1)$, $x_3 \in N(3, 5)$. Theoretically the distribution is the normal law . Significance levels for all applying criteria says about very good goodness between the empirical distribution obtained from simulations and theoretical distribution.

Example 2. $y = x_1 x_2$, where $x_1 \in N(0, 1)$, $x_2 \in Exp(0, 1)$. In this case, the theoretical distribution of function is unknown and the identification of law distribution was applied (verified by a composite hypothesis). The best model for the empirical distribution has a normal law with parameters shift $\theta_1 = 20.1906$ and scale $\theta_2 = 9.2237$. The results saying about very good goodness.

Example 3. $y = x[1][1] + x[1][2]$, i.e. summed up the first and second component of the vector x, where $x[1] \in N(0, 1)$ with the vector of means $\Theta_0 = [2, 1]^T$ and covariance matrix $\Theta_1 = \begin{bmatrix} 4 & 0 \\ 0 & 9 \end{bmatrix}$. Analytically is that $y \in N(3, \sqrt{13})$. The results showing good goodness between the modeled expression and expected results.

Let $Y = \prod_{i=1}^{k} X_i$, where X_i - nonintercorrelated variates with mean M_i and variance D_i. In accordance with (3) $Y \approx \sum_{i=1}^{k} X_i \prod_{j=1, j\neq i}^{k} M_j - (k-1)\prod_{j=1}^{k} M_j$, the mean $E[Y] \approx \prod_{j=1}^{k} M_i$ and variance $D[Y] \approx \sum_{i=1}^{k} D_i (\prod_{j=1, j\neq i}^{k} M_j)^2$. In the case where X_i belongs to the standard normal law for $k = 2,\bar{}5$ the using of the linearization is not possible for trivial reasons: the variance is zero. The results of modeling the distribution of Y in this case represent the asymmetrical laws with zero median. These distributions can not adequately describe one parametric model law, but they are well approximated by mixtures of species [(3)]:

$$\alpha \frac{\theta_3}{2\theta_2 \Gamma(1/\theta_3)} exp(-(\frac{\mid y - \theta_1 \mid}{\theta_2})^{\theta_3}) + (1-\alpha)\frac{1}{\theta_1} exp(\frac{y-\theta_4}{\theta_5} - exp(\frac{y-\theta_4}{\theta_5})) \tag{4}$$

In the case of normal laws with different parameters the situation has changed. Fig. 3 presents the empirical distribution of this kind of products of random variables, but by the normal law with parameters shift to 4 and scale equal to 3. The distributions of Y for the case when $k = 2,\bar{}4$ can be described a mixture of two and when $k = 5$ - three-parametric models.

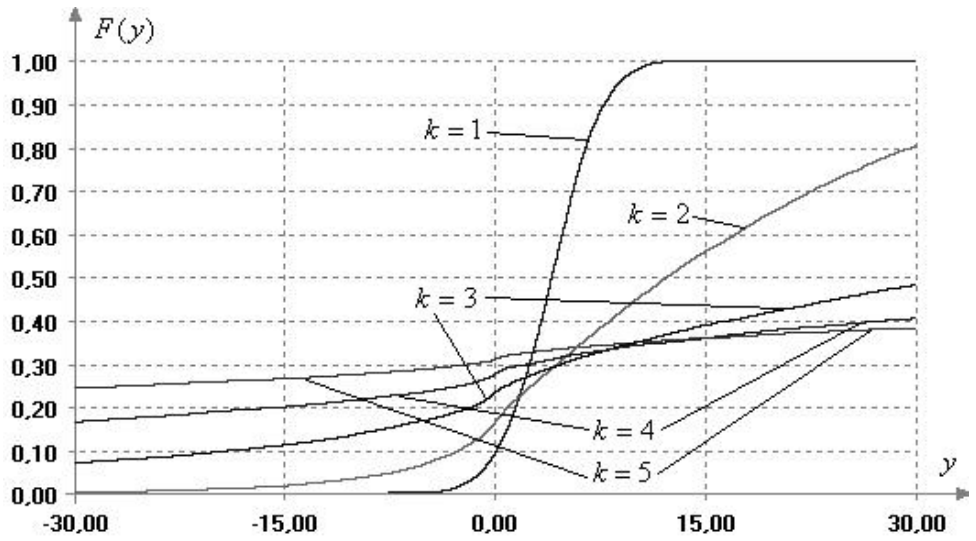

Figure 3: Empirical distribution of products of normal variables($k = \overline{1,5}$) with parameters shift equal 4 scale equal 3

5 Conclusion

Thus the methods of statistical modeling with the software that allows to build approximate mathematical model for the empirical distributions (including mixtures of different parametric laws), are an effective tool to study the laws of the distribution functions of random variables for the study of probabilistic laws, appeared in various technical applications.

Distributions of functions of random variables X_i depend not only on the type of distributions of X_i and can vary greatly changes with other parameters of those laws [(3)]. Using statistical modeling techniques for the study of law distribution of Y, you can either build a close model approximating the law, or to ascertain conditions for the validity of the application of linearization.

Using statistical modeling, and specialized software, such as developing system "Interval statistics" ISW [(4)], allows to construct a good approximate mathematical model of the distribution laws of functions of random variables (including mixtures of parametric laws), when the law was not able to find analytically.

References

[1] MI 2083-90. GSI. Indirect measurements. Definition results of measurements and estimation of their errors. (in Russian)

[2] Levin S.F. The cast scheme in the method of indirect measurement / / Measurement Technology, 2004. - N 3. - P.5-9. (in Russian)

[3] Lemeshko B.Yu., Ogurtsov D.V. Statistical modeling as an effective instrument for investigating the distribution laws of functions of random quantities / / Measurement Techniques, 2007. V.50, N 6. - P. 593-600

[4] Lemeshko B.Yu., Postovalov S.N. Computer technology of data analysis and research of statistical regularities. - Novosibirsk: Izd NSTU, 2004. - 119 pp. (in Russian)