APPROXIMATION OF THE DISTRIBUTION LAW OF THE SUM OF RANDOM BETA VARIABLES

Svetlana A. Oleinikova, Oleg Ja. Kravets

Voronezh Technical State University, Voronezh, Russian Federation
e-mails: s.a.oleynikova@gmail.com, csit@bk.ru
Russian Federation

Abstract: The subject of the research in this paper is the approximation of the probability density function (pdf) of random variable, which is the sum of beta values. Since the required sum cannot be expressed by any of the known laws, the aim is to find such approximation for the pdf of this sum that would have the least error. As a result the approximation of the desired law with the use of the beta distribution, were obtained.

Key words: beta-distribution, sum of random variables, approximation, numerical experiment, distribution law.

1. INTRODUCTION

The problem of approximation of the probability density function (pdf) of the sum of a finite number of beta variables is considered. This is a universal law that can be used to describe the majority of random phenomena with continuous distribution law. In particular, in the majority of cases investigating random events that can be described using unimodal continuous random variables which lie in a certain range of values can be approximated by beta distribution. Unfortunately, unlike the majority of distribution laws, the beta distribution does not have unique properties that allow analytically describe the required sum. Moreover, the specificity of this law is that the extracting of the multiple definite integral, which is necessary for determining the pdf of the sum of random variables is extremely difficult. In this regard, the problem of the approximating the pdf of the sum of beta values with minimum error is occurs.

In this paper an approach to determination of the approximation of the required law on the basis of the beta distribution is presented. The effectiveness of this approximation is justified by a series of numerical experiments that allow to compare the error of the obtained estimate with an accuracy of existing approaches, based on the normal law. Finally an example illustrating one of the areas of application of the results is given.

2. STATEMENT OF THE PROBLEM AND ITS PECULIARITIES

In general, the beta law determined by the density, defined on the interval [0,1] as follows [1]:

$$f_\eta(t) = \begin{cases} \dfrac{1}{B(p,q)(b-a)^{p+q-1}} t^{p-1}(1-t)^{q-1}, 0 \le t \le 1_i, \\ 0, t < 0, t > 1. \end{cases} \tag{1}$$

However, practical interest as a rule represents beta values determined at an arbitrary interval [a, b]. These is primarily due to the fact that the range of practical problems in this case is much broader, and, secondly, if a solution for the more general case was founded, then get the result for a particular case, which will be determined by a random variable (1), will not present any difficulties. Therefore, further we will consider the random variables defined on an arbitrary interval. In this case, the problem can be stated as follows.

The estimation of the distribution law of a random variable is the sum of random variables ξ_i, i=1,…,n, each of which is distributed by the law of beta in the range [a_i, b_i] with parameters p_i and q_i is searching. The pdf of each variable can be described by formula:

$$f_{\xi_i}(t) = \begin{cases} \dfrac{1}{B(p_i,q_i)(b_i-a_i)^{p_i+q_i-1}} (t-a_i)^{p_i-1}(b_i-t_i)^{q_i-1}, a_i \le t \le b_i, \\ 0, t < a_i, t > b_i. \end{cases} . \tag{2}$$

Partially the problem of searching the law for of the sum of these random variables has been solved earlier. In [2] a formula for determining the amount of beta-values, each of which is defined by the formula (1), was obtained. In [3], an approach to finding the sum of two random variables, each of which is defined by the formula (2), was given.

However, in general, the original problem is not solved. In particular, in [2] written, that in practice the mathematics can be intractable. This is primarily due to specifics of formula (2), which does not provide a compact and easy-to-use formula in finding the density of the sum of random variables.

Indeed, for two values ξ_1 and ξ_2 desired density is defined as follows [1]:

$$f_\eta(z) = \int_{-\infty}^{\infty} f_{\xi_1}(x) \cdot f_{\xi_2}(z-x)dx \cdot \tag{3}$$

In the case of addition of n random variables the multiple integral will be obtained. For this problem there are difficulties associated with the specifics of the beta distribution. In particular, already for n=2, the use of formula (3) leads to very cumbersome result, which is defined in terms of hypergeometric functions. Re-taking the integral of the resulting density, which should be done already for n=3 and above, is extremely difficult. It is not ruled out errors that inevitably arise when rounding and calculating of such a complicated expression. Therefore it is necessary to find approximation of the pdf of the sum of a finite number of variables that described by formula (2), allowing to use the known formulas with minimum error.

3. NUMERICAL EXPERIMENT TO GENERATE HYPOTESES OF THE APPROXIMATION OF THE PDF OF SUM OF THE BETA VALUES

To analyze the specificity of the desired pdf the following computer experiment was conducted. Sets the number of random variables with beta distribution, and parameters of each variable as a result we obtain the statistical data on the random variable representing the sum of the given values. The input data for the experiment will be the following values:

- the number of random variables n, having a beta distribution;
- parameters p_i and q_i of the random variable ξ_i, i=1,…,n, which has a beta distribution;
- the lower and the upper bounds [a_i, b_i] of the random variable ξ_i, i=1,…,n.

We will explore the random variable:

$$\eta_n = \xi_1 + ... + \xi_n . \tag{4}$$

At the output it is necessary to get:

- numerical characteristics of the random variable η_n;
- histogram of the random variable η_n.

In details the experiment was described in [4]. By varying parameters of the individual beta values, as well as their amount, and summarizing the results of the experiments, the following conclusions were obtained.

1. If individual random quantities entering into the sum has symmetrical density, the histogram of the final distribution has a form close to the normal. Also close to the normal distribution the estimates of numerical characteristics of the final value (mathematical expectation, variance, skewness and kurtosis).

2. If individual random variables are asymmetric (both with positive and negative asymmetry), but the total asymmetry is close to 0, then from the point of view of the graphical representation and numerical characteristics the resulting distribution law is also close to normal.

3. In the other cases, the required law is visually close to the beta distribution. In particular, the histogram of the sum of five asymmetric random variables is shown in Fig. 1.

Thus, on the basis of the experiment we can advance a hypothesis about a possible approximation of the sum of beta-density variables using the normal or the beta distribution.

To confirm this hypothesis and selecting a single law for approximating the following experiment was conducted. Setting the number of random variables with beta distribution, as well as their parameters, it is necessary to find the numerical value of the desired density and contrast it with the density of the corresponding normal or beta distribution. This will require:

- developing an algorithm which allows to obtain a numerical estimate of the pdf of the sum of random variables with beta-density;

- determining the parameters of the final pdf under the assumption of a normal or beta distribution (with the given parameters and the number of the initial values);

- calculating the error of the approximation using the normal and the beta distribution.

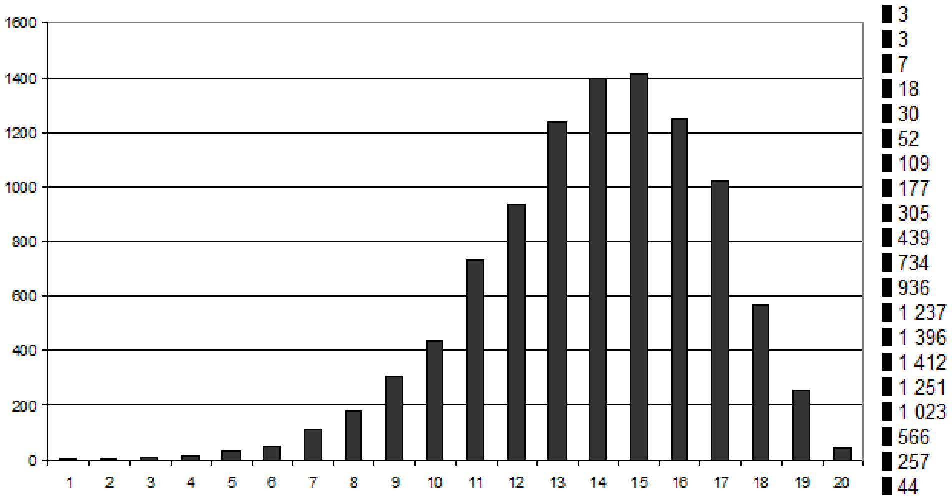

Figure 1. Histogram of the sum of five asymmetric random variables

Consider these tasks in details.

A numerical algorithm for finding the sum of beta-density values based on recursion. The sum of the n arbitrary random variables can be defined as follows:

$$\eta_n = \xi_1 + ... + \xi_n = \eta_{n-1} + \xi_n , \tag{5}$$

$$\eta_{n-1} = \xi_1 + ... + \xi_{n-1} . \tag{6}$$

Similarly, the pdf of the random variable η_{n-1} can be described as follows:

$$\eta_{n-1} = \xi_1 + ... + \xi_{n-2} + \xi_{n-1} = \eta_{n-2} + \xi_{n-1} . \tag{7}$$

Continuing the same reasoning, and using the formula (3), we obtain:

$$f_{\eta_n}(x) = \int_{-\infty}^{\infty} f_{\xi_{n-1}}(x - x_{n-1}) \cdot \left(\int_{-\infty}^{\infty} f_{\xi_{n-2}}(x_{n-1} - x_{n-2}) \cdot \left(... \int_{-\infty}^{\infty} f_{\xi_2}(x_2) \cdot f_{\xi_1}(x_2 - x_1) dx_1 ... \right) dx_{n-2} \right) dx_{n-1} \tag{8}$$

A more detailed discussion about it, as well as the specifics of determining the pdf for variables with the beta distribution is given in [5]. The main difficulty in developing an algorithm that is designed for the numerical evaluation of the pdf of (4) is that the number of random variables (and, consequently, the number of integrals in (8)) is unknown beforehand. In this regard, the following algorithm based on recursion has been proposed. If the number of terms in the formula (4) is equals to 2 than the integral is calculated directly using the formula (3). Otherwise, to calculate the second factor of the integrand (8) a recursive call of the algorithm with the number of terms n-1 is used.

The parameters of the final distribution are determined under the assumption of independence of the random variables. In this case, the expectation and the variance of their sum will be determined by the formulas:

$$E(\eta) = E(\xi_1) + E(\xi_2) + ... + E(\xi_n); \tag{9}$$

and

$$Var(\eta) = Var(\xi_1) + Var(\xi_2) + ... + Var(\xi_n). \tag{10}$$

For normal law parameters a and σ will be directly determined by the formulas (9) and (10). For beta distributions first it is necessary to calculate the lower and upper bounds. They can be defined as follows:

$$a = \sum_{i=1}^{n} a_i \; ; \tag{11}$$

and

$$b = \sum_{i=1}^{n} b_i \cdot \tag{12}$$

Here a_i and b_i - the boundaries of the intervals of the individual terms. Next, set up a system of equations, including the formulas for the expectation and the variance of a beta value:

$$\begin{cases} E(\xi) = a + (b-a)\dfrac{p}{p+q} \\ Var(\xi) = (b-a)^2 \dfrac{pq}{(p+q)^2(p+q+1)} \end{cases} \tag{13}$$

Solving the system (13) with respect to the parameters p and q, we will have:

$$p = \frac{(b - E(\xi)) \cdot (E(\xi) - a)^2 - Var(\xi) \cdot (E(\xi) - a)}{Var(\xi) \cdot (b-a)}. \tag{14}$$

and

$$q = \frac{(b - E(\xi))^2 \cdot (E(\xi) - a) - Var(\xi) \cdot (b - E(\xi))}{Var(\xi) \cdot (b-a)}. \tag{15}$$

Next, it is need to evaluate the numerical approximation error of the pdf of the sum of random variables with beta distribution by the normal law and the law of beta. To do this, use the formula:

$$Er = \int_a^b \left| \tilde{f}(x) - f_\eta(x) \right| dx \;, \tag{16}$$

where $\tilde{f}(x)$ - approximation of the pdf of the sum of random variables; $f_\eta(x)$- pdf of the sum of random beta variables.

We will consistently change the parameters of the individual random values to estimate the error. In particular, it is interesting the following questions:

- how quickly the sum of the beta values converges to a normal distribution, and whether it is possible to estimate the pdf of the sum by another law, which will have a minimum error;
- how much the error increases with increasing asymmetry of the individual beta values;
- how will change the error with the change of distribution scope of the individual beta values.

4. THE RESULTS OF THE EXPERIMENT

Let's analyze the results of the experiment. The dynamics of reduction of error with the increasing the number of terms is shown in Fig. 2.

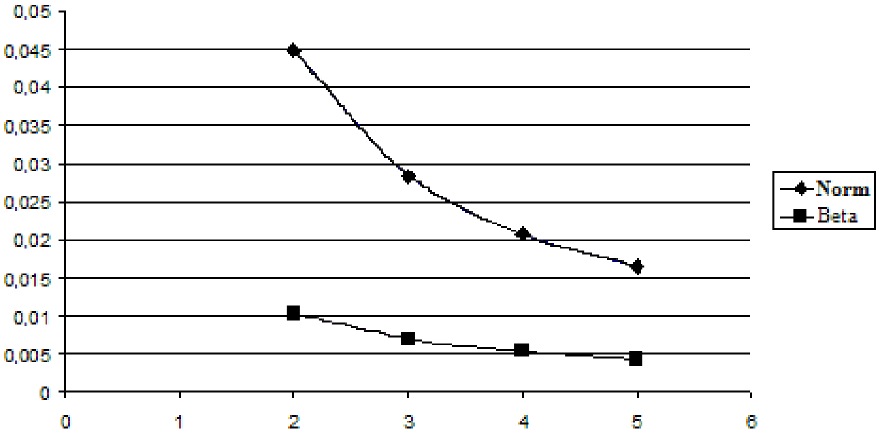

Figure 2. The reduction of errors by increasing the number of terms

The x-axis shows the number of terms, and the vertical axis - the value of the error. As can be seen from this figure, for two terms, the approximation error of beta distribution is about 4 times lower than the error of approximation of normal distribution. It is obvious that with increasing the terms the normal law approximation error is reduced much faster than the error of the beta approximation. It can also be prompted, that with a very large number of terms the approximation with the normal law will have a smaller error than the approximation with the beta distribution. However, taking into account the values of error in this case it can be concluded that in terms of the number of terms the approximation using the beta distribution is preferred.

Fig. 3 shows the dynamics of change of error with increasing asymmetry of the random variables. Without loss of generality, the parameter p of the initial beta values, was recorded with a value of 2, and the parameter q + 1 on the horizontal axis is presented. The ordinate axis on the figure shows the approximation error. Results of the experiment with other values of the parameters are broadly similar. In this case, it is also clear a preference for the approximation of the sum of beta values using the beta distribution. Then the changing of error with the changing of the initial scope of the beta values is analyzed.

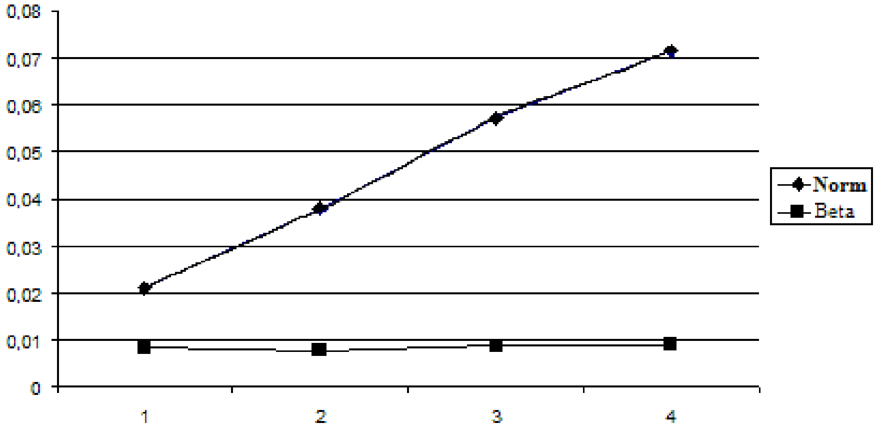

Figure 3. The dynamics of change the error with increasing asymmetry of the random variables

Fig. 4 shows the results of measurement of error for the sum of the four beta-values, three of which are distributed in the interval [0,1], and the fourth sweep sequentially increases (it is delayed by the abscissa).

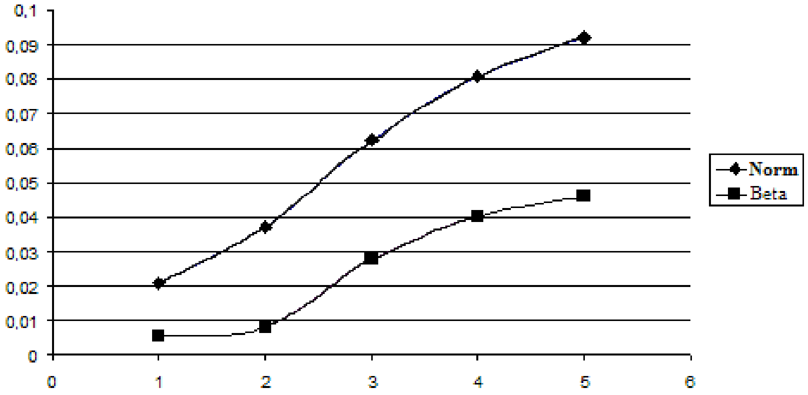

Figure 4. The dynamics of change the error with increasing the initial scope of the random variables

Based on the graphic illustrations of experiments, as well as the data obtained in the experiment it can be concluded the feasibility of using beta distribution for approximating the pdf of the sum of random variables which have beta distribution.

As shown by the results, in 98% of cases the error in approximation of the pdf of variable (4) using the beta distribution will be lower than in the approximation using the normal distribution. The average value of the error of approximation using the beta law will depend primarily on the intervals span in which each term is distributed. At the same time the dependence of the symmetry of the random variables, as well as the number of terms this rating (as opposed to the normal law) extremely insignificant.

5. APPLICATIONS

One of the applications of the results is the tasks of project management. The project is a set of mutually-dependent works with a random duration of service. The most appropriate law to estimate the unknown time is the beta distribution [6]. The duration of the project will be a random variable. It is obvious that the evaluation of the pdf of this value is interesting not only in the planning stages, but also in the analysis of possible situations related to the late completion of all works. Given the fact that the project delay may lead to a variety of adverse situations, including fines, the estimate of the law of the random variable that describes the duration of the project, is extremely important practical task. At the present time for such evaluation is used method PERT [6]. According to its assumptions, the project duration is normally distributed random variable $\eta \sim N(a, \sigma)$ with the following parameters:

$$a = E(\xi_1) + E(\xi_2) + \ldots + E(\xi_n); \tag{17}$$

and

$$\sigma = \sqrt{Var(\xi_1) + Var(\xi_2) + \ldots + Var(\xi_n)}. \tag{18}$$

In the formulas (17) and (18) $\xi_1,\ldots,\xi_n$ represent the random variables that describe the duration of works lying on the critical path. Consider the correction of method PERT taking into account the results that have been obtained. In this case, we will assume that the duration of the project is distributed according to the law of beta with parameters (14) and (15). Apply the obtained results in practice. Without loss of generality, consider a project, given the network graph shown in Fig. 5. Here the graph edges corresponds to works, edge weights are designated works rooms; tops in the squares - the events that marks the beginning or ending of works.

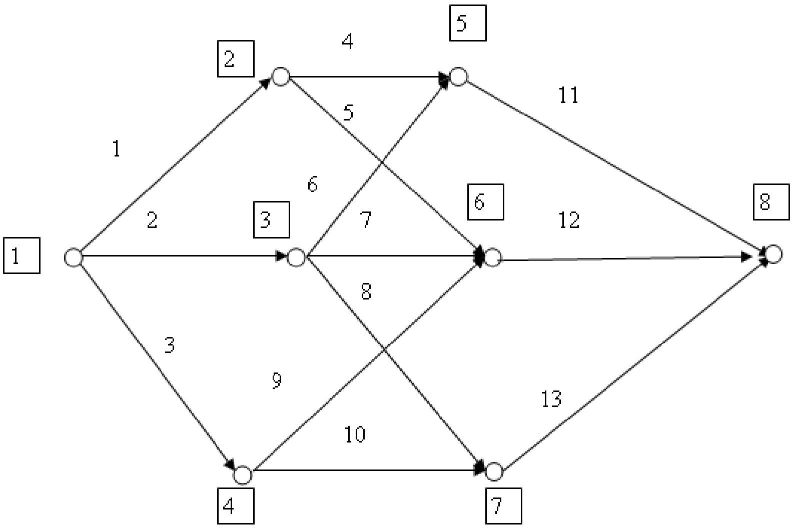

Figure 5. The example of the network graph

Let the works are given of its durations shown in Table 1.

Table 1. Temporal characteristics of works

№ of work	*min*	*max*	*expectation*
1	*5*	*10*	*9*
2	*3*	*6*	*4*
3	*6*	*8*	*7*
4	*4*	*7*	*6*
5	*4*	*7*	*5*
6	*2*	*5*	*3*
7	*4*	*8*	*6*
8	*4*	*6*	*5*
9	*6*	*8*	*7*
10	*2*	*6*	*4*
11	*9*	*13*	*12*
12	*2*	*6*	*3*
13	*5*	*7*	*6*

Min is the minimum time for which the given work can be done; max - the longest time; expectation - the expectation of the beta distribution, showing the expected duration of this work. An approach to find the expectation of the beta-distribution on the basis of the relationship between the mode and the variance was offered in [7]. The works standing on the critical path are highlighted in bold.

Imitate a process of implementation of the project with the help of a specially developed simulation system. As a result we need to get:

- histogram of a random variable described by the duration of the project;
- estimation of probability of completion of the project in a given interval on the basis of statistical data;
- estimation of this probability using the normal and the beta distributions.

During the simulation of the project 10,000 times, the service durations sample was obtained, which histogram is shown in Fig. 6. Obviously, the kind of the histogram is different from the normal curve. We use the formulas (9) and (10) to find the final expectation and variance. Then the necessary evaluations will be as follows. Expectation:

$$E(\eta) = 9 + 6 + 12 = 27 .$$

Variance:

$$Var(\eta) = \frac{(10-5)^2}{36} + \frac{(7-4)^2}{36} + \frac{(13-9)^2}{36} = 1{,}3889 .$$

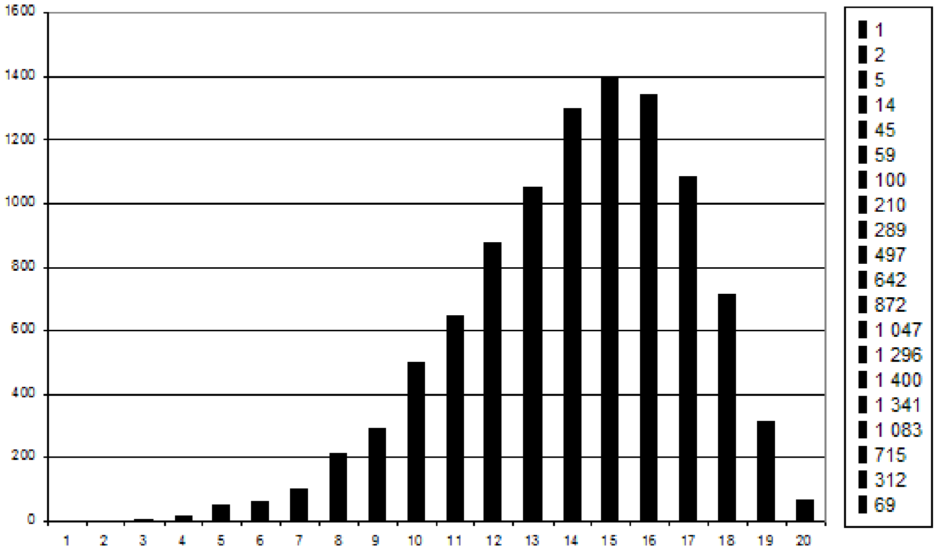

Figure 6. Histogram of the duration of the project

The probability of falling into the set interval will be calculated using the known formula:

$$P(a < \eta < b) = F_\eta(b) - F_\eta(a). \tag{19}$$

Calculate the parameters for the final beta distribution. For this we use formulas (14) and (15). We get p=13,83; q=4,61.

The boundaries of the beta distribution are determined by the formulas (11) and (12): a=18; b=30.

The research results are given in Table 2. Without loss of generality, the number of runs of the model is chosen 10000.

In the column "Statistics" the probability obtained on the basis of statistical data is calculated. In the column "Normal" the probability calculated by the normal distribution law is shown. In the column "Beta" the probability value calculated on the basis of the beta distribution is given.

Table 2. The results of the probabilistic estimates

Interval	***Statistics***	***Normal***	***Beta***
(a-σ;a)	*0,2988*	*0,3413*	*0,3018*
(a-σ;a +σ)	*0,6862*	*0,6828*	*0,6761*
(a;a +σ)	*0,3874*	*0,3413*	*0,3743*
(a+σ;a +2σ)	*0,1542*	*0,1359*	*0,1578*
Less than a-3σ	*0,0043*	*0,0013*	*0,0043*

Based on the results presented in Table 2, as well as the other similar results obtained in the course of the simulation of performance of different projects, it can be concluded that the estimates obtained using the beta distribution as an approximation of the pdf of the execution time of the project allow to obtain the solution of the problem with greater accuracy compared with existing analogues.

6. CONCLUSION

The aim of this work was to find such approximation of the distribution law of the sum of random variables with beta distribution that would have the least error compared to other analogs. The following results were obtained:

1. Experimentally it has been hypothesized about the possibility of approximating the sum of beta-values using a beta distribution.
2. A software tool that allows to get the numerical value of the error of the desired density distribution using the normal law and the beta law was developed. The heart of this program is based on a recursive algorithm that allows to determine the pdf of the sum of random variables with the given beta density numerically.
3. The computational experiment, the purpose of which was to determine the best approximation by a comparative analysis of errors in a variety of conditions, was set. The experimental results show the feasibility of using the beta distribution as the best approximation of the pdf of the sum of beta variables.

The example, illustrating practical significance of the obtained results, is presented. This is the project management tasks with a random time of execution of certain works. The results allow to obtain more accurate estimates of the unknown probabilities and, as a consequence, to reduce the probability of errors in planning.

REFERENCES

[1] Kobzar A.I. Applied Mathematical Statistics. For engineers and scientists/ Prikladnaya matematicheskaya statistika. Dlya inzhenerov i nauchnykh rabotnikov (in Russian), Moscow, 2006, 816 p.

[2] Gupta A.K., Nadarajah S. Handbook of Beta Distribution and Its Applications, New York: Marsel Dekker, 2004, 579 p.

[3] Johnson N., Kotz S., Balakrishnan N. Continuous univariate distributions. Vol. 2, Second edition, New York: John Willey & Sons Press, 1995, 752 p.

[4] Oleynikova S.A. Computer experiment for the analysis of the distribution law of the random quantity that determines the duration of the project in problems of network planning and management. *Economics and Management of Control Systems*, **3(9)**, 2013, pp. 91-97.

[5] Oleynikova S.A. Recursive numerical method for the experimental evaluation of the distribution law of the duration of the project in network planning and management tasks. *Software Systems and Computational Methods*, 1, 2015, pp. 69-78.

[6] Golenko-Ginzburg D., Gonik A. Project planning and controlling by stochastic network models. *Managing and Modeling Complex Projects*, **17**, 1997, pp. 21-43.

[7] Oleinikova S.A., Kravets O.J., Silnov D.S. Analytical estimates for the expectation of the beta distribution on the basis of the known values of the variance and mode. *Information (Japan)*, **19** (vol. 2), 2016, pp. 343-351.

Information about the authors:

Svetlana Alexandrovna Oleinikova – Assistant professor at Voronezh State Technical University, Faculty of Information Technologies and Computer Security. Areas of Research are Network planning and management in stochastic systems;

Oleg Jakovlevich Kravets – Professor at Voronezh State Technical University, Faculty of Information Technologies and Computer Security. Areas of Research are Network planning, Network Routing and modeling.

Manuscript received on 17 April 2017