Method of Probability Distribution Fitting for Statistical Data with Small Sample Size

Valeriyi Kuzmin
Department of Telecommunication and Radioelectronic Systems
National Aviation University
Kyiv, Ukraine
kuzmin_vn@i.ua

Maksym Zaliskyi
Department of Telecommunication and Radioelectronic Systems
National Aviation University
Kyiv, Ukraine
maximus2812@ukr.net

Roman Odarchenko
Department of Telecommunication and Radioelectronic Systems
National Aviation University
Kyiv, Ukraine
odarchenko.r.s@ukr.net

Oksana Polishchuk
Educational and Scientific Institute of Innovative Educational Technologies
National Aviation University
Kyiv, Ukraine
opolishuk@ukr.net

Olga Ivanets
Department of Biocybernetics and AerospaceMedicine
National Aviation University
Kyiv, Ukraine
olchik2104@ ukr.net

Olga Shcherbyna
Department of electronics, robotics and technology of monitoring and internet of things
National Aviation University
Kyiv, Ukraine
shcherbyna_ol@nau.edu.ua

***Abstract*—The paper deals with a new approach for probability distribution fitting for empirical data with small sample size. The proposed method includes three steps: 1) outliers detection and correction; 2) transformation basis calculation; 3) basis function optimization. For the possibility of asymmetric distributions approximation, a piecewise linear basis function is used. During basis function optimization, the dependence of squared deviations sum on switching point abscissa is calculated. The mathematical formula for this dependence can be obtained by quadratic approximation according to the least squares method. The optimum of switching point abscissa coincides with minimum of obtained parabola. Method of probability distribution fitting for statistical data with small sample size is illustrated on the real empirical data example. For this example the best probability distribution fitting corresponds to the case of optimized piecewise linear basis function.**

Keywords—approximation; outliers detection; probability distribution fitting; basis function; optimization.

I. Introduction

At the present stage of science and technologies development, the researchers' attention is paid to the issues of real statistical data processing using various types of probability distributions, such as normal, lognormal, exponential, Weibull, Rayleigh, and others, as well as Pearson curves systems, Johnson curves, etc. [1]. These distributions can depend on one, two, three, and even four parameters.

At the same time, the choice of the probability distribution law in the case of small sample size is important and urgent problem. This is due to the fact that the lack of the data does not provide reliable information for the selection of the suitable probability distribution law [2]. The main purpose of choosing an appropriate probability distribution is it usage to calculate random variables of any probability [3]. Such tasks often come into existence in engineering practice during reliability calculations [4], econometric calculations [1], decision-making [5], etc. The researchers' tendency to describe random variables sets by single probability distribution laws does not always lead to correct results in case of solving specific engineering practical problems [6]. There are so many schemes and methods for generating random variables that it is impossible to describe these variables by single law. In authors opinion, when choosing a suitable distribution, special attention should be paid to distribution tails fitting (approximation). As the distribution tail, authors understand the specified size set of lower and upper values of order statistics.

The statistical tests may allow different distribution laws, but the final choice is subjective and not always correct [7].

II. Literature Analysis and Problem Statement

The small sample problem is widely studied in the scientific literature. The main aim of this problem is to obtain the maximum information in case of data lack. In [8] authors present the technique to determine the empirical probability density function and the empirical cumulative distribution function in case of small sample size. The estimates of probability density function is calculated for each values of order statistic in case of priori information about random variable range.

In paper [9] Weibull distribution was selected as the failure distribution for small samples. The parameters of this distribution are estimated based on regression analysis. Moreover, authors calculated the probability distribution of parameters, which combines prior information, results of Monte-Carlo simulation and is corrected by Bayes' theorem. Another way to analyze data with small sample size is data expansion [10]. Authors considers the segmented random sampling method that designed for the evaluation of failure time distribution of extremely small sample accelerated degradation test.

The probability-possibility transformation method for small sample size data based on Sison-Glaz's simultaneous confidence intervals is considered in [11]. The paper [12] deals with maximum entropy principle to fit probability distribution for small sample, and authors said that this approach avoids disadvantages which are inherent for the least square approximation method. In paper [13] the method of statistical processing of wind turbine operational data for small sample size based on classical approach, q-q plot construction and piecewise linear approximation is considered. In addition, literature analysis concerning statistical data processing showed that there is a need for efficient selection of mathematical models for approximating the distribution tails [14].

It is known that one of the efficiency index to check the approximation quality is the sum of squared deviations (SSD). Assume that different functions for probability

978-1-7281-6760-2/20/$31.00 ©2020 IEEE

distribution fitting are described by $y_j(x_i)$, where x_i is a statistical data with sample size n, j is a number of function (the total quantity of such function is M). The selection of the best approximation is performed to provide SSD minimum. From mathematical point of view the aim of this paper is following equation solution

$$m = \inf\left(r \in \mathbf{N},\ 1 \le r \le M, 1 \le j \le M : S(y_r(x_i)) \le S(y_j(x_i))\right),$$

where S is SSD, $\mathbf{N}$ is a natural number.

III. Method of Probability Distribution Fitting

The method of probability distribution fitting consists of three steps: 1) outliers detection and correction; 2) calculation of transformation basis; 3) basis function optimization.

The analysis showed that there are several methods for outliers detection and correction. In one approach the outlier is detected and isn't taken into account in further calculations. According to winsorization approach the outlier is detected and replaced by the nearest value of order statistic. In this paper for outliers detection, Chauvenet's criterion will be used with transformation of the following type

$$Q_i = \mathrm{Me} \cdot F^{U_i V}, \tag{1}$$

where Q_i is an approximated (calculated) variable, F is a basis function, U_i is a quantile of normal distribution with zero expectation and standard deviation $\sigma = 1$, Me is a median of the sample, V is a variation coefficient.

To calculate the transformation basis it is necessary to obtain the transformed sample

$$z_i = \ln \frac{x_i^{(\text{order})}}{\mathrm{Me}}, \tag{2}$$

where $x_i^{(\text{order})}$ is an order statistic for initial data x_i.

The empirical probabilities for each observation of order statistic are calculated by the formula

$$p_i = \frac{i}{n}.$$

The quantiles of the standard normal distribution are calculated according to Kazakyavicius equation [15]

$$U_i = 2.0637\left(\ln\left(\frac{1}{1-p_i}\right) - 0.16\right)^{0.4274} - 1.5774. \tag{3}$$

Further the sums of first and last random variables in transformed order statistic are calculated

$$\Sigma_1 = \sum_{i=1}^{k} z_i\,;\quad \Sigma_2 = \sum_{i=n-k}^{n} z_i\,, \tag{4}$$

where k depends on sample size.

The products of the corresponding quantiles sum by the variation coefficient

$$\Sigma_{U\min} = V\sum_{i=1}^{k} U_i\,;\quad \Sigma_{U\max} = V\sum_{i=n-k}^{n} U_i\,. \tag{5}$$

Then transformation basis for minimum a_1 and maximum a_2 can be represented as

$$a_1 = e^{\Sigma_1 / \Sigma_{U\min}}\,;\quad a_2 = e^{\Sigma_2 / \Sigma_{U\max}}\,. \tag{6}$$

In this paper two options of basis function are considered

$$F_1(U_i) = \frac{a_1 e^{-U_i} + a_2 e^{U_i}}{e^{-U_i} + e^{U_i}}, \tag{7}$$

$$F_2(U_i) = a_1 + b(U_i + U_{\text{sw}})_+ - b(U_i - U_{\text{sw}})_+, \tag{8}$$

$$(U_i - U_{\text{sw}})_+ = \begin{cases} 0,\ \mathit{if}\ U_i < U_{\text{sw}}, \\ U_i - U_{\text{sw}},\ \mathit{if}\ U_i \ge U_{\text{sw}}, \end{cases}$$

$$(U_i - U_{\text{sw}})_+ = \frac{|U_i - U_{\text{sw}}| + (U_i - U_{\text{sw}})}{2},$$

where U_{sw} is a quantile value that corresponds to the switching point; b is a coefficient that depends on a_1, a_2 and U_{sw}. The coefficient b can be computed by formula

$$b = \frac{a_2 - a_1}{2U_{\text{sw}}}.$$

The second basis function corresponds to piecewise linear function with three segments. In case of basis function $F_2(U_i)$, it is possible to find optimum value of switching point $U_{\text{sw opt}}$. The optimization can be made according to the following rule:

$$U_{\text{sw opt}} = \arg\min\left(S(U_{\text{sw}})\right),$$

where $S(U_{\text{sw}})$ is SSD that calculated for different values of switching point U_{sw}. The best option of probability distribution fitting will correspond to the minimum SSD case.

To clarify the method, let us consider an example of statistical data processing.

IV. Analysis of Initial Statistical Data

Consider the empirical data presented in [16]. The data are shown in the Table I. This sample is described by following characteristics: mathematical expectation $\bar{x} = 79.804$; standard deviation $\sigma = 13.23$; median $\mathrm{Me} = 80.65$; variation coefficient $V = 0.166$. The histogram for six grouping intervals is shown in Fig. 1.

TABLE I. Initial Statistical Data

Initial statistical data, x_i							
68	74.4	81,7	87	77	72.4	81.3	82
94.7	79.8	74.4	82	80.1	79.7	91	72.3
86.4	81	84	87.2	73	80	102.4	88.2

Initial statistical data, x_i							
80.3	70.1	59.2	101.7	90.1	50.2	52	93.5
84.4	83	70.4	72.2	99.4	69.4	98	91.5
81.6	66.1	84.3	67.3	89.1	77	79.6	80
84.1		54.7		45.4		115.2	

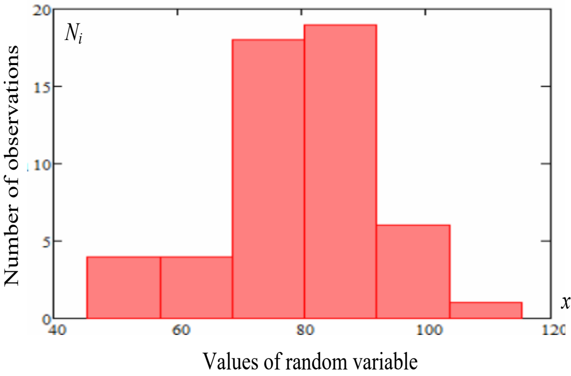

Fig. 1. The histogram for initial data.

Even a visual analysis of histogram shows that the data can be described by a normal distribution. The tests carried out according to chi-squared test and Kolmogorov-Smirnov test also confirmed this hypothesis. In the classical procedure of statistical data analysis, an approximation is performed using the normal distribution [17]. Results of this approximation will be compared with new approach of probability distribution fitting.

V. Detection and Correction of Outliers

Let us perform tails approximation according to the [18].

After formation and analysis of order statistic for initial data, minimum and maximum values of order statistic are determined as possible outliers. These values are 45.4 and 115.2. Determine the required number of observations for tail approximation $k \approx 1,5\sqrt[3]{n}$. In this case $k = 5.599$. Therefore, six observations will be used.

The expression (1) will be used for approximation. In this equation the basic function can be equal to one of two constant values: for the lower tail a_1 and for the upper tail a_2. Since the minimum and maximum values are candidates for the outlier, they won't be taken into account when calculating the basis of the transformation. So the decision will be made taking into consideration only five values. After transform (2) calculation, the values for the lower and upper tail are obtained. The corresponding quantiles are computed according to the equation (3). The results of calculation are listed in the Table II. According to the equations (4) and (5), the sums will be equal to following values: $\Sigma_1 = -1.809$; $\Sigma_2 = 1.035$; $\Sigma_{U\,\min} = -1.213$; $\Sigma_{U\,\max} = 1.213$. Taking into account (6), the transformation bases $a_1 = 4.442$; $a_2 = 2.347$.

TABLE II. Transformed Values and Corresponding Quantiles

$z_{i\text{ lower}}$	–0.474	–0.439	–0.388	–0.309	–0.199
$z_{i\text{ upper}}$	0.161	0.195	0.209	0.232	0.239
$U_{i\text{ lower}}$	–1.777	–1.584	–1.436	–1.314	–1.209
$U_{i\text{ upper}}$	1.209	1.314	1.436	1.584	1.777

The Chauvenet's criterion involves finding the critical values for empirical probabilities of two times larger sample sizes than the initial. So

$$p_{\min} = \frac{1}{2n+1} \text{ and } p_{\max} = \frac{2n}{2n+1} .$$

Then the critical values will be as follows

$$z_{\min}{}^{*} = \mathrm{Me} \cdot a_1{}^{U_{\min} V} = 45.173 ;$$

$$z_{\max}{}^{*} = \mathrm{Me} \cdot a_1{}^{U_{\max} V} = 112.361 .$$

Based on these calculations, it can be concluded that the maximum value 115.2 of the initial sample is an outlier. Therefore, this value is replaced by $z_{\max}{}^{*} = 112.361$. The minimum value is not an outlier. As a result, the corrected sample was obtained.

VI. Approximation of the Corrected Sample

Let us recalculate the characteristics of corrected sample. As the result, the numerical values are obtained: $\bar{x} = 79.749$, $\sigma = 13.089$, $V = 0.164$, $a_1 = 4.691$, $a_2 = 2.426$.

Perform an approximation of the corrected data using the basis function (7). This option of the basic function allows to determine it uniquely. The advantage of this basis function is its smoothness. For the researched statistical data, the basis function can be written as follows

$$F_1(U_i) = \frac{4.691e^{-U_i} + 2.426e^{U_i}}{e^{-U_i} + e^{U_i}} .$$

In case of basis function (8), it is necessary to solve optimization problem. Therefore, for the five options of switching points U_{sw} ={1; 1.1; 1.2; 1.3; 1.4}, the SSDs S = {716.498; 407.801; 245.770; 313.300; 513.572} were calculated for all 52 values of the corrected sample. The obtained dependence is approximated by second order parabola with the ordinary least squares method usage. This parabola can be written as follows

$$S(U_{\text{sw}}) = 13690 - 21890 U_{\text{sw}} + 8911 U_{\text{sw}}{}^2 .$$

The parabola minimum corresponds to the value

$$U_{\text{sw opt}} = -\frac{-21890}{2 \cdot 8911} = 1.228 .$$

So the optimal basis function (8) has the form:

$$F_2(U_i) = 4.691 - 0.987(U_i + 1.228)_+ + 0,987(U_i - 1.228)_+ .$$

The results of data approximation with normal distribution and basis functions (7), (8) usage are shown in Fig. 2.

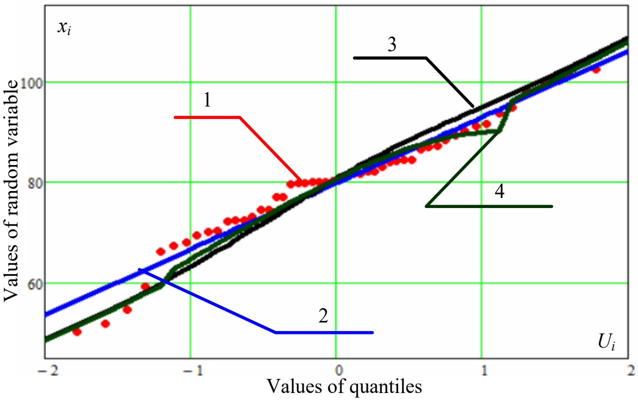

Fig. 2. Approximation of empirical data: 1 – initial data, 2 – data approximation using normal distribution, 3 – data approximation using basis function (7), 4 – data approximation using optimal basis function (8).

The SSD for the first and second half of order statistic in case of different approximation functions usage:

a) data approximation using normal distribution: $S_{\text{lower}} = 282.484$ and $S_{\text{upper}} = 60.624$;

b) data approximation using basis function (7): $S_{\text{lower}} = 322.238$ and $S_{\text{upper}} = 196.859$;

c) data approximation using optimal basis function (8): $S_{\text{lower}} = 126.415$ and $S_{\text{upper}} = 122.959$.

The approximation by a normal distribution has a large unevenness in the description of the lower and upper parts of the order statistic. Data approximation using the optimal basis function (8) is uniform for the lower and upper parts of the order statistic and has a minimum value of SSD for all sample. So, for given initial data, the final version of the individual probability distribution fitting with elements of optimization is optimal basis function (8).

VII. Conclusion

The paper considers the problem of analyzing a sample for outliers presence, correcting detected outliers, as well as several approximation options for corrected data: using the normal distribution, and using a new approach with two variants of basis functions.

The analysis of the researched approximation options showed that the best option for the minimum SSD criterion is approximation using piecewise linear basis function. This method of probability distribution fitting allows to approximate the real samples of random variables with the right or left asymmetry. However, its usage is limited only to unimodal distributions.

The research showed that as a result of individual fitting for specific empirical data, it is possible to obtain a more adequate mathematical model of sample description in comparison with classical probabilistic laws.

References

[1] D.C. Montgomery, and G.C. Runger. Applied Statistics and Probability for Engineers, Fours Edition, NJ: John Wiley & Sons, 2007, 768 p.

[2] D.V. Gaskarov, and V.I. Shapovalov. Small Sample, Moscow, Statistics, 1978, 248 p. (in Russian).

[3] I.G. Prokopenko, S.V. Migel, and K.I. Prokopenko, “Signal modeling for the efficient target detection tasks”, International Radar Symposium, June 19-21, 2013, (Dresden, Germany), Proceedings, Vol. II, pp. 976-982.

[4] M. Zaliskyi, and O. Solomentsev, “Method of Sequential Estimation of Statistical Distribution Parameters”, IEEE 3rd International Conference Methods and Systems of Navigation and Motion Control (MSNMC), October 14-17, 2014 (Kyiv, Ukraine), Proceedings, pp. 135-138.

[5] O. Solomentsev, M. Zaliskyi, T. Herasymenko, O. Kozhokhina, and Yu. Petrova “Data Processing in Case of Radio Equipment Reliability Parameters Monitoring”, Advances in Wireless and Optical Communications (RTUWO 2018), November 15-16, 2018 (Riga, Latvia), Proceedings, pp. 219-222.

[6] A. Goncharenko, “Aircraft operation depending upon the uncertainty of maintenance alternatives”, Aviation, 2017, Volume 21 No 4, pp. 126-131, https://doi.org/10.3846/16487788.2017.1415227.

[7] N.S. Kuzmenko, I.V. Ostroumov, and K. Marais, “An Accuracy and Availability Estimation of Aircraft Positioning by Navigational Aids”. Methods and Systems of Navigation and Motion Control (MSNMC), October 16-18, 2018 (Kyiv, Ukraine), Proceedings, pp. 36-40.

[8] R. Florescu, and N. Thirer, “Distribution Laws of Small Size Samples. Metrological Implementation”, IEEE 24th Convention of Electrical & Electronics Engineers in Israel, November 15-17, 2006, (Eilat, Israel), Proceedings, pp. 79-81, doi:10.1109/eeei.2006.321099.

[9] Z. Dai, Z. Wang, and Y. Jiao, “Bayes Monte-Carlo Assessment Method of Protection Systems Reliability Based on Small Failure Sample Data”, IEEE Transactions on Power Delivery, 2014, Vol. 29, No. 4, pp. 1841-1848, doi:10.1109/tpwrd.2014.2316915.

[10] H. Zhang, H. Yuan, and P. Li, “Estimation Method for Extremely Small Sample Accelerated Degradation Test Data”, First International Conference on Reliability Systems Engineering (ICRSE), October 21-23, 2015, (Beijing, China), Proceedings, pp. 1-5, doi:10.1109/icrse.2015.7366417.

[11] Y. Hou, and B. Yang, “Probability-possibility Transformation for Small Sample Size Data”, Seventh International Conference on Fuzzy Systems and Knowledge Discovery, August 10-12, 2010, (Yantai, China), Proceedings, pp. 1720-1724, doi:10.1109/fskd.2010.5569396.

[12] Yang Qingnian, and Sima Yuzhou, “The Fitting Method of Parameter Distributions in Geotechnical Engineering under Small Sample”, 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), August 8-10, 2011, (Dengleng, China), Proceedings, pp. 7366-7369, doi:10.1109/aimsec.2011.6010605.

[13] M. Zaliskyi, Yu. Petrova, M. Asanov, and E. Bekirov, “Statistical Data Processing during Wind Generators Operation”, International Journal of Electrical and Electronic Engineering & Telecommunications, 2019, Vol. 8, No. 1, pp. 33-38, doi:10.18178/ijeetc.8.1.33-38.

[14] Y. Hryshchenko, “Reliability Problem of Ergatic Control Systems in Aviation”. IEEE 4th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC), October 18-20, 2016 (Kyiv, Ukraine), Proceedings, pp. 126-129.

[15] K.A. Kazakyavicius, “Approximate Formulas for Mechanical Results Statistical Processing”, Zavodskaya laboratoriya, 1988, Vol. 54, No. 12, pp. 82-85.

[16] B.A. Dospehov, Methods of field experiments, Moscow, Kolos, 1973, 336 p. (in Russian).

[17] V.N. Kuzmin, “Extending the Possibilities of the Bootstrap Method using Analytical Approximation”, 13th International Scientific Conference named after Academician M. Kravchuk, May 13-15, 2010, (Kyiv, Ukraine), Proceedings, p. 12.

[18] V.N. Kuzmin, and P.I. Bidyuk, “A New Approach to Detection and Correction of Outliers”, 14th International Scientific Conference named after Academician M. Kravchuk, April 19-21, 2012, (Kyiv, Ukraine), Proceedings, pp. 13-15.