Original article
https://doi.org/10.21683/1729-2646-2020-20-2-28-34

Accounting for the effect of correlations by modulo averaging as part of neural network integration of statistical tests for small samples

Alexander I. Ivanov[1*], **Andrey G. Bannykh**[2], **Yulia I. Serikova**[2]
[1]Penza Research and Design Electrical Engineering Institute, Russian Federation, Penza, [2]Penza State University, Russian Federation, Penza
**ivan@pniei.penza.ru*

Alexander I. Ivanov

Andrey G. Bannykh

Yulia I. Serikova

Abstract. *The* **Aim** *of the paper is to demonstrate the advantages of taking into consideration real correlations by means of their symmetrization, which is significantly better than completely ignoring real correlations in cases of statistical estimation using small samples.* **Methods.** *Instead of real correlation numbers different in sign and modulo, identical values of correlation numbers moduli are used. It is shown that the equivalence of transformation to symmetrization is subject to the condition of identical probabilities of errors of the first and second kind for asymmetrical and equivalent symmetrical correlation matrices. The authors examine the procedure of accurate calculation of equal data correlation coefficients by trial and error and procedure of approximate calculation of symmetrical coefficients by averaging the moduli of real correlation numbers of an asymmetrical matrix.* **Results.** *The paper notes a practically linear dependence of equal probabilities of errors of the first and second kind from the dimension of the symmetrized problem being solved under logarithmic scale of the variables taken into consideration. That ultimately allows performing the examined calculations in table form using low-bit, low-power, inexpensive microcontrollers. The examined transformations have a quadratic computational complexity and come down to using pre-constructed 8-bit binary tables that associate the expected probability of errors of the first and second kind with the parameter of equal correlation of data. All the table calculations are correct and do not accumulate input data round-off errors.* **Conclusions.** *The now widely practiced complete disregard of the correlations when performing statistical analysis is very detrimental. It would be more correct to replace the matrices of real correlation numbers with symmetrical equivalents. The approximation error caused by simple averaging of the moduli of coefficient of asymmetrical matrices decreases as the square of their dimension or the square of the number of neurons that generalize classical statistical tests. When 16 and more neurons are used, the approximation error becomes negligible and can be disregarded.*

Keywords: *replacement of statistical test with equivalent neurons; multicriteria statistical analysis of small samples; accounting for the effect of correlations; symmetrization of correlation matrices.*

For citation: *Ivanov A.I., Bannykh A.G., Serikova Yu.I. Accounting for the effect of correlations by modulo averaging as part of neural network integration of statistical tests for small samples. Dependability. 2020;2: 28-34. https://doi.org/10.21683/1729-2646-2020-20-2-28-34*

Received on: *14.06.2019* / **Revised on:** *07.05.2020* / **For printing:** *17.06.2020.*

The problem of application of classical statistical tests with small samples

Pearson's statistical hypothesis test was created in 1900 and proved to be very effective. Naturally, in 1900 computer technology did not exist, so only relatively computationally simple tests could be created, studied and used. Person's test set the trend in statistical study for decades. As the result, hundreds of mathematicians in the XX century created about 200 statistical tests applicable under various limiting conditions.

Unfortunately, all known statistical tests provide poor results with small samples. In such areas as biometrics, medicine, biology, economy, the samples of actual data are small. This circumstance impedes reliable statistical estimation. Thus, Pearson's chi square test over a 21-experiment sample yields poorly shared states for normal data and uniformly distributed data (Fig. 1).

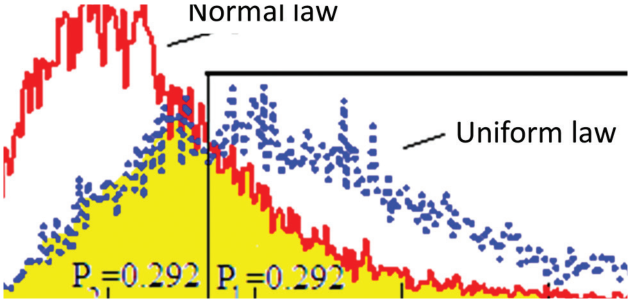

Fig. 1. The threshold of the chi square neuron $k = 7.5$ was defined based on the matching values of the probabilities of errors of the first and second kind $P_1 = P_2 = P_{EE} = 0.292$

The confidence probability of detection of normal data under the adopted conditions is not high: 0.708 (1 – 0.292 = 0.708), which makes practical application impossible. Practically acceptable confidence probabilities can be obtained only with large samples [1] of 200 and more examples.

Unlike in 1900, we possess the capability to multiply the complexity of calculations as part of statistical analysis. For instance, we can use several different statistical tests. We can associate artificial neurons with each statistical test [2, 3] and use them simultaneously. Fig. 1 shows one of the chi square neuron settings with 5 inputs. Each of such inputs of the neuron analyzes one of the bins of the histogram of the tested sample.

The output comparator of the artificial neuron is set in such a way as to allow the probability P_1 of errors of the first kind to be close to the probability P_2 of errors of the second kind. This technique allows reducing the dimension of the problem at hand by replacing two variables with one $P_1 = P_2 = P_{EE}$. Formally, the variables P_1 and P_2 may differ from each other, however if they are artificially made identical (symmetrical), we will be able – by means of symmetrization – simplify the calculations.

Table 1 shows the values of matching probabilities of errors of the first and second kind for 8 different statistical tests (neurons), where:

χ^2 is the chi square test [2, 3, 4, 5];

ad^2 is the Anderson-Darling test [4, 5];

adL is the logarithmical form of the Anderson-Darling test [4, 5];

sg is the geometric mean test [6, 7, 8];

sg$_d$ is the differential-integral variant of the geometric mean test [5, 7];

ω^2 is the Cramér-von Mises test [5, 7];

ω^2_c is the Smirnov-Cramér-von Mises test [4, 5];

su^2 is the Shapiro-Wilk test [5, 9].

It is obvious that, using eight statistical tests instead of one is made easily possible through modern computer technology with low-bit microcontrollers (4-bit processors of RFID identification cards, 8-bit processors of modern controllers, low-power processors of SIM cards and microSD cards). The neural network implementation of such engineering solution will result in code condition 00000000, when all tests (all neurons) make a decision in favour of the normal distribution law of small sample values. If all neurons make a decision in favour of even distribution of values, the output code will be 11111111.

In practice, the bits of a neural network's output code do not always have identical states. In such cases the decision is made based on the majority of observed states. In other words, all codes with a majority of states 0 are taken as the decision of detection of normal value distribution in a 21-experiment input sample.

All transformations that can be performed using low-bit microcontrollers can also be performed on personal computers using appropriate software. Such approach is acceptable as part of scientific research; however, it cannot be used in large-scale biometrical calculations. In order to ensure compliance with cyber security requirements, biometric neural network calculations and cryptographic transformations must be performed only in a trusted computational environment, normally implemented on a low-bit, low-energy, low-cost microcontroller.

Rough statistical estimation under the hypothesis of complete absence of correlations between the responses of generalized statistical tests

Table 1 shows data of only 8 statistical tests (statistical neurons). For that reason, we can conduct a numerical experiment and identify the probabilities of each of the

Table 1. Values of error probability for criteria of statistical hypothesis testing for 21-test samples

No., i	1	2	3	4	5	6	7	8
Test	χ	ad^2	adL	sg	sg$_d$	ω	ω	su^2
P_{EEi}	0.292	0.349	0.320	0.320	0.278	0.351	0.311	0.322

256 code conditions. If we increase the number of neurons from 8 to 256, calculating the probabilities of all code conditions will be technically very complicated. As the number of simultaneously working neurons grows, the complexity of such computational task increases exponentially.

As we do not know how to precisely take into consideration the effect of correlations between the bits of the output code, we will opt for a simplification and accept the hypothesis of independence of the analyzed data. In this case the mutual reinforcement of the eight tests may be estimated as the product of equally possible errors from Table 1:

$$P_{EE(8)} = \prod_{i=1}^{8} P_{EEi} \approx 0,0001. \qquad (1)$$

The geometric mean of the probabilities P_{EE} of eight tests is 0.316. Assuming that 256 simultaneously used statistical tests are independent, and their harmonic mean is 0.316, we obtain a very optimistic estimate of the probability of errors:

$$P_{EE(256)} \approx \prod_{i=1}^{256} P_{EEi} \approx \left\{ \sqrt[8]{\prod_{i=1}^{8} P_{EEi}} \right\}^{256} \approx 0,316^{256} \approx 10^{-128}. \qquad (2)$$

The data of the actual numerical experiment for 8 statistical neurons from Table 1 are about 80 times worse than the optimistic estimate (1). That means that the hypothesis of independence of the conditions of statistical neurons is not applicable to our case. We cannot neglect the existing correlations while performing neural network integration of the many classical statistical tests.

Accounting for correlations through their symmetrization: estimation of the correctness of the hypothesis of equal correlation of the responses of the generalized statistical test

As the real correlations cannot be disregarded, since about 1999, neural network biometrics [10, 11, 12, 13] have been using the practical technique of symmetrization of correlations. The essence of the technique consists in the fact that the actual correlation number matrix is replaced with some equivalent with identical elements out of range:

$$\begin{bmatrix} 1 & r_1 & r_2 & ::: & r_n \\ r_1 & 1 & r_{n+1} & ::: & r_{2n-2} \\ r_2 & r_{n+1} & 1 & ::: & r_{3n-3} \\ ::: & ::: & ::: & ::: & ::: \\ r_n & r_{2n-2} & r_{3n-3} & ::: & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & \tilde r & \tilde r & ::: & \tilde r \\ \tilde r & 1 & \tilde r & ::: & \tilde r \\ \tilde r & \tilde r & 1 & ::: & \tilde r \\ ::: & ::: & ::: & ::: & ::: \\ \tilde r & \tilde r & \tilde r & ::: & 1 \end{bmatrix}. \ (3)$$

The condition of correct symmetrization (3) comes down to matching probabilities of errors of the first and second kind for the initial asymmetrical model and the final symmetrical model:

$$P_{EE} \approx P_2\left\{\left[r_{i,j}\right]\right\} \approx P_1\left\{\left[r_{i,j}\right]\right\} \approx \tilde P_2\left\{\left[\tilde r\right]\right\} \approx \tilde P_1\left\{\left[\tilde r\right]\right\}. \qquad (4)$$

For any actual correlation matrix, a symmetrical equivalent correlation matrix can be chosen that would have identical values data correlation coefficient. In order perform an accurate symmetrization, an iterative fitting of parameter $\tilde r$ is required. Such approach to the solution of the problem is similar to training artificial neurons through an iterative algorithm by criterion of systems movement towards the fulfillment of condition (4). The computational complexity of such iterative processes strongly depends on the dimension of the problems at hand. It is generally believed that iterative fitting as part of neural network has a polynomial computational complexity (for our case, the polynomial order is always lower than the dimension of the symmetrized matrix).

It is interesting to note that the first approximation of the equal correlation coefficients can be obtained through a simple averaging of the correlation coefficient modules of the initial asymmetrical matrix (this procedure has a quadratic computational complexity):

$$\tilde r \approx \frac{2}{n^2 - n} \cdot \sum_{i=1}^{\frac{n^2-n}{2}} \left| r_i \right|, \qquad (5)$$

where i is the numbers of the correlation coefficients outside the diagonal of the initial asymmetrical correlation matrix.

Table 2. Correlation numbers between pairs of examined statistical tests

	χ	ad^2	adL	sg	sg$_d$	ω	ω	su^2
χ	**1**	0.423	0.672	0.037	-0.042	0.559	0.401	-0.726
ad^2	0.423	**1**	0.644	0.018	-0.145	0.226	0.393	-0.113
adL	0.672	0.644	**1**	0.056	0.209	0.827	0.832	-0.917
sg	0.037	0.018	0.056	**1**	0.132	0.414	0.402	-0.212
sg$_d$	-0.042	-0.145	0.209	0.132	**1**	-0.242	-0.142	-0.041
ω	0.559	0.226	0.827	0.414	-0.242	**1**	0.885	-0.667
ω	0.401	0.393	0.832	0.402	-0.142	0.885	**1**	-0.764
su^2	-0.726	-0.113	-0.917	-0.212	-0.041	-0.667	-0.764	**1**

It is obvious that formula (5) is an approximation, therefore it is required to evaluate the approximation error $\Delta\tilde{r}$ as the dimension function n of the matrix. In order to evaluate the rate of error reduction, let us use the correlations of the 8 neural network implementations of statistical tests, whose data is given in Table 2.

The correlation data from Table 2 can be used in estimating the degree of convergence of the examined computational operation. For that purpose, it will suffice to randomly select sets of three out of the eight statistical tests and apply approximate relationship (5) to their data. The histogram of the results of such calculations is shown in Fig. 2 (red line).

This procedure must also be done with sets of five randomly selected from the data of Table 2. As the result, we obtain a histogram of data also shown in Fig. 2 (blue line).

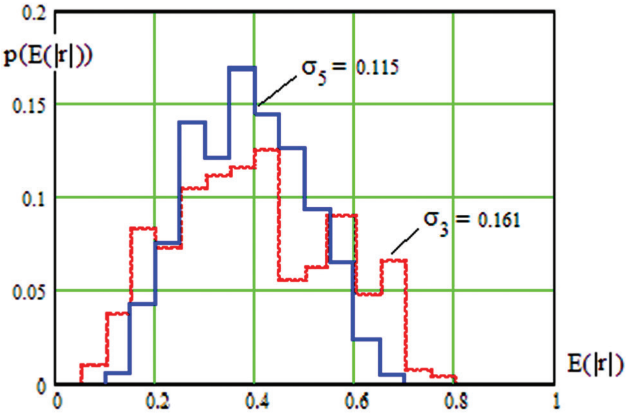

Fig. 2. Histogram of the value distribution of effective moduli of correlation numbers of non-repeating sets of three and five statistical tests from Table 2

Fig. 2 shows that as matrix dimension grows, the standard deviate of data decreases from value σ_3=0,161 to value σ_5=0,115. As the matrix dimension further grows, the distributions of possible values of effective moduli contract. Additionally, normalization of the distributions of possible values of calculation errors $\Delta\tilde{r}$ of symmetrization can be observed.

Numerical estimation of convergence of the symmetrization procedure of correlations of actual biometric data

It should be noted that activities aimed at neural network integration of several statistical tests started only recently [2, 4, 5] and, as consequence, actual statistical data is not yet sufficient. In neural network biometrics the situation is completely different [10, 11, 12, 13]. The biometric neural network authentication technology has been in active development in Russia and other countries since the beginning of the XXI century. As consequence, large anonymized biometric databases have been created using various technical methods, however, they cannot be used due to ethical limitations. Access to such reliable information if restricted both in Russia and abroad.

Ethical restrictions are removed if the problem of access to large volumes of reliable biometric information is solved using the BioNeiroAvtograf simulation environment [14, 15]. That is a free software product that is designed in such a way as to allow Russian-speaking universities to organize their educational process. The product analyzes the dynamics of handwritten reproduction of letters and/or words using a mouse or any graphic tablet. Using two-dimensional Fourier transform, BioNeiroAvtograf extracts 416 biometric parameters from handwriting dynamics and based on the GOST R 52633.5-2011 standard trains a single-layer 256-neuron network.

All data on the biometric parameters, weighting parameters and neuron connections are observable [15] (stored in viewable *.txt files). Using that data, let us generate a training database out of 30 examples of the handwritten

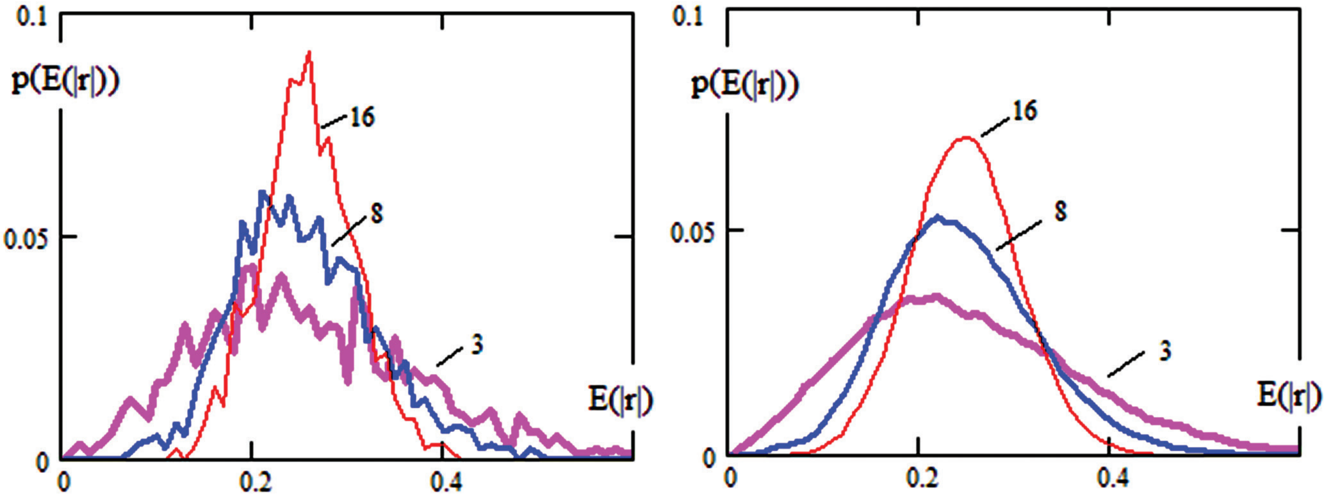

Fig. 3. Examples of distribution of symmetrization parameter values $\tilde{r}$ without smoothing (left part of the figure) and after smoothing (right part of the figure) for matrix dimensions from 3 to 16

word “Penza” in one person’s handwriting. Loading data on 30 examples of 416 biometric parameters in MathCAD enables us building a 416 × 416 matrix of correlation numbers. Ultimately, we obtain an amount of data that is much larger than in an 8 × 8 matrix of Table 2.

That allows randomly generating 1024 samples of 3 biometric parameters each and modulo-averaging their coefficient correlation numbers. The resulting distribution of the values of symmetrization results is given in Fig. 3. Similar distributions are shown in the figure for random samples of 8 and 16 biometric parameters.

Fig. 3 shows that the constructed distributions normalize sufficiently quickly. In case of symmetrization of correlation coefficients of a 16 × 16 or larger matrix, the distribution can be considered to be normal. In other words, the distributions are normalized sooner than for the chi square test. It is allowable to replace asymmetrical chi square distributions with normal ones only when 32 and more parameters are taken into consideration. The matters of approximation of chi square distributions by other laws are examined in more detail in [16]. The effect of data normalization for the considered symmetrization procedures ensues sooner as compared with normalization of data of a well-researched chi-square test.

Another important feature of symmetrization is that the uncertainty introduced by this simplification monotonously declines $\sigma_3 > \sigma_4 > ... > \sigma_{256}$. For that exact reason, accounting for mutual correlations for vectors of the length of 256 binary states of a long password or cryptographic key enables sufficiently accurate predictions if a simple symmetrization procedure is used [12, 13]. In the first approximation it can be considered that the uncertainty decreases proportionally to $\sqrt{n^2 - n}/\sqrt{2}$. That means that the standard deviation σ_3=0,161 (see Fig. 2) in case of neural network integration of 100 statistical tests must decrease about 50 times to $\sigma_{100} \approx$ 0,0032.

Simple nomogram for predicting the operational quality of neural network integrations of statistical tests of various dimension

A sufficiently accurate prediction of the attainable probabilities of error under various conditions is possible if simulation tools are used to reproduce the operation of 1, 2, …, 8 neurons under various values of equal correlation coefficients $\tilde{r} = \{0,3, 0,4, ..., 0,7\}$. The results of simulation are well described with a linear approximation in logarithmical coordinates [17] as shown in Fig. 4.

The nomogram in Fig. 4 calculated for the probabilities of error in each of the neuron shows the geometric mean value of the geometrical probabilities of errors in each neuron 0.316. This nomogram easily transforms for other geometric mean values of the probabilities of errors in each of the neurons. For that purpose, it suffices to offset data upwards, if the probability of errors increases and downwards if the probability of errors decreases.

Fig. 3 shows that driving the strength of statistical tests up is less profitable than driving their correlation down. Thus, under correlation value $\tilde{r}$ = 0,4, in a group of 8 examined tests, the probability of errors of 0.001 would require 70 neurons (70 statistical tests). If the level of mutual correlation is reduced to $\tilde{r}$ = 0,3, the same level of probability of failures would require only 17 neurons (17 statistical tests).

Conclusion

In this paper we attempted to show that methods of symmetrization of multidimensional problems are sufficiently simple and efficient. Upon the symmetrization of the error probability estimation of several neurons accounting for their mutual correlations, a simple nomogram is constructed

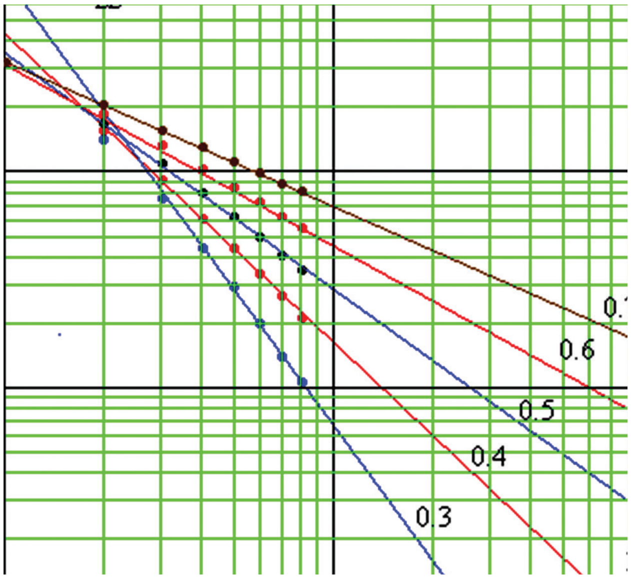

Fig. 4. Nomogram of association of identical probabilities P_{EE} of errors of the first and second kind of neural network integration for averaged values of correlation numbers 0.3, 0.4, 0.5, 0.6, 0.7

that predicts how many neurons will be required in order to achieve a certain probability of errors of the first and second kind.

Currently, the available computing facilities do not impose any restrictions on the number of statistical tests generalized by a neural network. It is only a matter of the degree of mutual correlation of hundreds of classical statistical tests. Unfortunately, most classical statistical tests provide strongly correlated results. The high level of their correlation is the next technical limitation. That indicates the growing relevance of the problem associated with the synthesis of new statistical tests, of which the data is weakly correlated in relation to the majority of known statistical tests.

Nevertheless, it is safe to say that in the years to come the confidence probability of statistical estimations based on small samples should significantly increase. Neural network integration of hundreds of already known statistical tests is not a complex scientific problem, but rather a sufficiently simple engineering task. Additionally, the approximations set forth in this paper allow taking into consideration the effect of correlations on the implementation of computations using low-bit, low-power microcontrollers of RFID cards, SIM cards and microSD cards, which should facilitate widespread application of the examined transformations as part of the solution of problems associated with biometric cryptographic authentication of persons.

References

1. R 50.1.037-2002. Recommendations for standardization. Applied statistics. Rules of check of experimental and theoretical distribution of the consent. Part I. Chi-square criteria. Moscow: Gosstandart Rossii; 2001. (in Russ.)

2. Ivanov A.I., Kupriyanov E.N., Tureev S.V. Neural network integration of classical statistical tests for processing small samples of biometrics data. Dependability 2019;2:22-27. DOI: 10.21683/1729-2646-2019-19-2-22-27.

3. Akhmetov B.B., Ivanov A.I. Estimation of quality of a small sampling biometric data using a more efficient form of the chi-square test. Dependability 2016;16(2):43-48.

4. Volchikhin V.I., Ivanov A.I., Bezyaev A.V., Kupriyanov E.N. The Neural Network Analysis of Normality of Small Samples of Biometric Data through Using the Chi-Square Test and Anderson–Darling Criteria. *Engineering Technologies and Systems*. 2019;29(2):205-217. DOI: 10.15507/2658-4123.029/2019.02.205-217.

5. Ivanov A.I., Bannych A.G., Kupriyanov E.N. et al. Collection of artificial neuron equivalent statistical criteria for their use when testing the hypothesis of normality of small samples of biometric data. Proceedings of the I All-Russian Science and Technology Conference Security of Information Technology. Penza. 2019. 156-164.

6. Perfilov K.A. [Criterion of geometric mean used for validity verification of the statistical hypotheses of biometric data distribution]. Proceedings of the Science and Technology Conference of the Penza Information Technology Security Cluster. Penza. 2014;9:92-93. [accessed 14.04.2020]. Available at: http://www.pniei.penza.ru/RV-conf/T9/S92. (in Russ.)

7. Ivanov A.I., Malygina E.A., Perfilov P.A. et al. The comparison of performance between the criterion mean geometric and the criterion of Cramér-von Mises on a small sample of biometric data. *Models, Systems, Networks in Economics, Engineering, Nature and Society*. 2016;2:155-158.

8. Ivanov A.I., Perfilov K.A., Malygina E.A. Multivariate statistical analysis of the quality of biometric data on extremely small samples using the criteria of the geometric mean tests calculated for the analyzed probability functions. *Measuring. Monitoring. Management. Control.* 2016;2(16):58-66. (in Russ.)

9. Ivanov A.I., Vjatchanin S.E., Malygina E.A. et al. Precision statistics: neuron networking of chi-square test and Shapiro-Wilk test in the analysis of small selections of biometric data. *Proceedings of the International Symposium Dependability and Quality*. 2019;2:131-134. (in Russ.)

10. Ivanov A.I. [Biometric identification of a person based on the dynamics of unconscious movement. Monograph]. Penza: PSU Publishing; 2000. (in Russ.)

11. Ivanov A.I. [Neural network technology of biometric identification of open system users. Author’s summary of the Doctor of Engineering thesis per study program 05.13.01 System analysis, management and processing of information. Penza; 2002. (in Russ.)

12. Malygin A.Yu., Volchikhin V.I., Ivanov A.I. et al. [Fast algorithms of testing neural network mechanisms of biometric cryptographic protection of information]. Penza: Penza State University Publishing; 2006. (in Russ.)

13. Akhmetov B.S., Volchikhin V.I., Ivanov A.I., et al. [Algorithms for testing biometric neural network mechanisms of information protection]. Kazakhstan, Almaty: Satbayev University; 2013. [accessed 14.04.2020]. URL: http://portal.kazntu.kz/files/publicate/2014-01-04-11940.pdf. (in Russ.)

14. Ivanov A.I., Zakharov O.S. [The BioNeiroAvtograf simulation environment: a software product (created by the Laboratory of Biometric and Neural Network Technology, freely available since 2009 on the website of the Penza Electrical Engineering Research and Development Institute)]. [accessed 14.04.2020]. Available at: http://пниэи.рф/activity/science/noc/bioneuroautograph.zip. (in Russ.)

15. Ivanov A.I. [Automatic training of large artificial neural networks in biometric applications: a study guide to laboratory works performed in the BioNeiroAvtograf simulation environment]. Penza: Penza Electrical Engineering Research and Development Institute; 2013. [accessed 14.04.2020]. Available at: http://пниэи.рф/activity/science/noc.htm. (in Russ.)

16. Kobzar A.I. [Applied mathematical statistics. For engineers and researchers]. Moscow: FIZMATLIT; 2006. (in Russ.)

17. Ivanov A.I., Lozhnikov P.S., Bannykh A.G. A simple nomogram for fast computing the code entropy for 256-bit codes that artificial neural networks output. Journal of Physics: Conference Series. 2019;1260(2):022003.

About the authors

Alexander I. Ivanov, Doctor of Engineering, Associate Professor, Academic Advisor, Penza Research and Design Electrical Engineering Institute, Russian Federation, Penza, 9 Sovetskaya Str., phone: (841 2) 59 33 10, e-mail: ivan@pniei.penza.ru.

Andrey G. Bannykh, third year post-graduate student, Department of Information Security Technology, Penza State University, 440026, Russian Federation, Penza, 40 Krasnaya Str., 40, phone: (841 2) 36 82 23, e-mail: ibst@pnzgy.ru.

Yulia I. Serikova, third year post-graduate student, Department of Computer Technology, Penza State University, 440026, Russian Federation, Penza, 40 Krasnaya Str., e-mail: julia-ska@yandex.ru.

The authors' contribution

A.I. Ivanov proposed a method for estimating the correctness of procedures of approximate calculation of equal correlation coefficients through simple averaging of real coefficient modules of an asymmetrical correlation matrix.

A.G. Bannykh synthesized 8-bit tables that associate predictable probabilities of errors of the first and second kind with the equal correlation parameter for a predefined number of artificial neurons in a log grid.

Yu.I. Serikova developed the software to supervise the degree of convergence of the computational processes considered in the paper.