УДК 519: 24; 53; 57.017 DOI 10.21685/2307-4205-2019-2-4

A. I. Ivanov, S. E. Vjatchanin, E. A. Malygina, V. S. Lukin

PRECISION STATISTICS: NEUROET NETWORKING OF CHI-SQUARE TEST AND SHAPIRO–WILK TEST IN THE ANALYSIS OF SMALL SELECTIONS OF BIOMETRIC DATA

А. И. Иванов, С. Е. Вятчанин, Е. А. Малыгина, В. С. Лукин

ПРЕЦИЗИОННАЯ СТАТИСТИКА: НЕЙРОСЕТЕВОЕ ОБОБЩЕНИЕ ХИ-КВАДРАТ КРИТЕРИЯ И КРИТЕРИЯ ШАПИРО – УИЛКА ПРИ АНАЛИЗЕ МАЛЫХ ВЫБОРОК БИОМЕТРИЧЕСКИХ ДАННЫХ

Abstract. The aim of the paper is a neural network generalization of the Chi-square test and the Shapiro–Wilk test for the analysis of small samples of biometric data. It is shown that any of the statistical criteria can be represented in the form corresponding to a neuron having an input sorter, an adder and some functional converter. The generalization of two statistical criteria is accomplished by tuning the output quantizers of two neurons. The setting is always ambiguous for a predetermined value of the confidence probabilities of the generalized decisions. It is shown that the usual form of presentation of statistical criteria in the form of computational formulas and the tables of quantiles of confidence probability of the equivalent to their neural network description if the tables of the ratio of quantization thresholds providing a given level of confidence in a neural network generalization are given.

Keywords: the Chi-square test; the Shapiro–Wilk test; the neural network generalization of statistical criteria.

Аннотация. Целью работы является нейросетевое обобщение хи-квадрат критерия и критерия Шапиро – Уилка для анализа малых выборок биометрических данных. Показано, что любой из статистических критериев может быть представлен в форме соответствующего нейрона, имеющего сортировщик входных данных, сумматор и некоторый функциональный преобразователь. Обобщение двух статистических критериев осуществляется настройкой выходных квантователей двух нейронов. Настройка всегда неоднозначна для заранее заданного значения доверительных вероятностей принимаемых обобщенных решений. Показано, что обычная форма представления статистических критериев в виде вычислительных формул и таблиц квантилей доверительной вероятности эквивалентна их нейросетевому описанию, если приведена таблица соотношения порогов квантования, обеспечивающих заданный уровень доверительной вероятности нейросетевого обобщения.

Ключевые слова: хи-квадрат критерий; критерий Шапиро – Уилка; нейросетевое обобщение статистических критериев.

The problem of providing high reliability of neural network converters of biometrics-code during their training on small samples

Currently, the digital economy is being actively created. One of the most important elements of new technologies is the biometric-cryptographic authorization of users when providing them with "cloud" services. In Russia, this technology is built using large artificial neural networks which are automatically trained how to convert a unique biometric image of a person into his personal cryptographic key [1]. The domestic market of information security in terms of the use of artificial neural networks is regulated by FSTEC (Federal Service of Technical and Export Control) of Russia according to the legal documents of the Technical Committee for Standardization "Information Security" № 362. In terms of the cryptographic protection of neural network solutions, the domestic market is regulated by the FSS (Federal Security Service) of Russia, based on the documents of the technical Committee for Standardization № 026 "Cryptographic protection of information" [2].

The current requirements of domestic regulators for neural network converters biometry-code are currently much tougher than similar requirements of NIST USA and international Committees of standardization ISO / IEC JTC1 sc27 (Information Protection Technology) and ISO / IEC JTC1 sc37 (Biometry). The latter is due to the fact that the requirements of international standards are focused on international biometric passports, whose "biometric templates" are not designed to store them in the Internet clouds. This leads to the fact that the user is always forced to carry an RFID-identifier and always work in a trusted computing environment.

Domestic neural network technologies are potentially deprived of this disadvantage, if neural network containers with biometric data are protected cryptographically [2]. In this case, it becomes possible to have electronic biometric passports safely stored in the Internet by the Federal Migration Service of Russia. Unfortunately, new technologies are poorly suited to well-tested traditional statistical methods of data processing [3, 4]. Thus, for the correct application of the Chi-square test, the standard recommendations [3] assume a sample of 200 experiments, while the real training samples contain about 20 examples. A similar situation arises when applying the Shapiro-Wilk test [4]. In order to overcome such significant limitations, it is necessary to radically change the algorithms of statistical data processing. For example, the testing of the quality of training of artificial neural networks must be performed in the Hamming convolution space [5, 6]. The transition from the space of ordinary codes with 256 possible states to the Hamming convolution space allows us to logarithmically reduce the number of states to 257.

The situation is approximately the same when we use 256 variations, long-known statistical criteria, in place of one test of statistical hypotheses. The task of this article is to describe the neural network generalization of the classical Chi-square test and the Shapiro-Wilk test. Using the generalization of these two classical criteria, we will try to estimate the resulting gain from their joint use.

Application of the Chi-square test for small biometric data samples

For estimates on the Chi-square test, it is necessary to find the maximum and minimum values of the data in the sample. Next, you need to set the number of intervals -k of the histogram and find the width of these intervals:

$$\Delta x = \frac{\max(x) - \min(x)}{k}. \tag{1}$$

Then you should count the number of examples in the analyzed sample, which fell into each of the intervals –k of the histograms. With this histogram forming, the minimum sample value is the left border of the first column of the histogram, and the maximum value of the sample coincides with the right border of the rightmost column of the histogram.

The Chi-square value of the criterion is calculated using the following formula:

$$\chi^2 = N \cdot \sum_{i=1}^{k} \frac{\left(\frac{n_i}{N} - P_i\right)}{P_i}, \tag{2}$$

where n_i – the number of experiments in the i-th column of the histogram, P_i – the expected probability of testing in the i-th column of the histogram.

The popularity of the Pearson's Chi-square test is due to the fact that an analytical description of the density distribution of values is known for it:

$$p(\chi^2, m) = \frac{1}{2^{\frac{m}{2}} \cdot G\left(\frac{m}{2}\right)} \left\{ x^{\frac{m}{2}-1} \cdot \exp\left(\frac{-x}{2}\right) \right\}, \tag{3}$$

where *m* – the number of degrees of freedom, *G* (.) – the Euler gamma function.

It is recommended [3] to choose the number of degrees of freedom by the formula:

$$m = k - 3 = k - 2 - 1, \tag{4}$$

when the problem of checking the normality of empirical statistics is solved. The formula (4) is usually justified by the fact that the normal distribution law is described by two statistical moments (mathematical expectation and standard deviation). In order to use the formula (2), we need to calculate two statistical moments, which should lead to a decrease in the number of degrees of freedom by two units. The greater the number of statistical moments described by the theoretical distribution law, the lower should be index of the number of degrees of freedom – m. If the theoretical law is described by d statistical moments, then the number of degrees of freedom should be [3]:

$$m = k - d - 1. \tag{5}$$

It should be noted that the above approach for large samples works well, but for small samples, the calculation of only integer degrees of freedom (5) does not work well. This is easily seen in 100,000 sample implementations consisting of 21 experiences of normal data distribution. In this numerical experiment, we observe the distribution of the Chi-square values of the criterion, shown in Fig. 1.

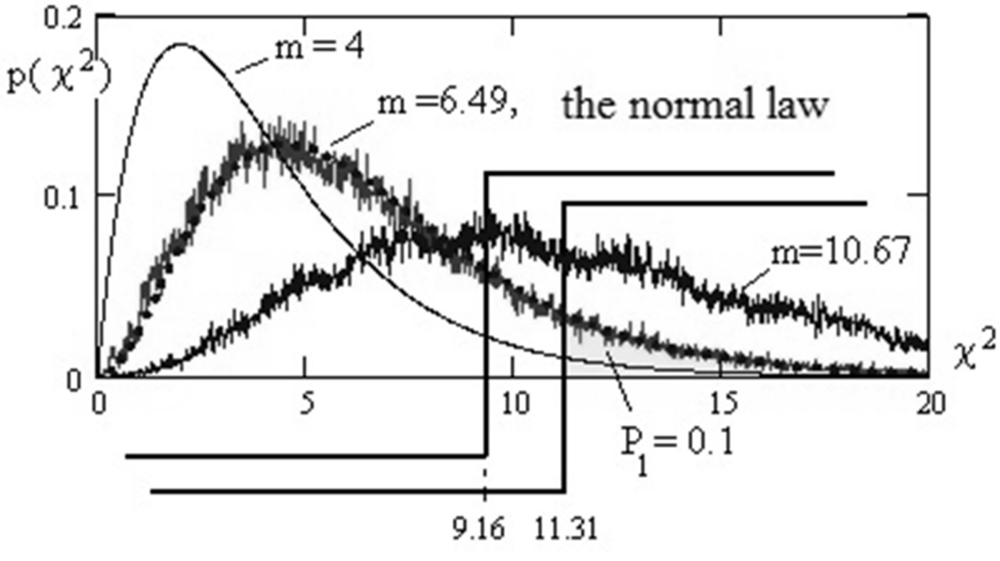


Fig. 1. Fractional number of degrees of freedom m = 6.49 for values of Chi-square values of sample distribution of 21 examples with normal data distribution and a histogram of 7 columns

Figure 1 shows that the recommended value in (5) of the member of degrees of freedom m = 4 is significantly different from the real value *m* = 6.49. As a result, the decision threshold with a confidence level of 0.9 should be $\chi^2 \geq 11.31$. For the confidence probability of 0.8, the decision threshold is reduced to $\chi^2 \geq 9.16$.

It should be emphasized that setting of thresholds according to formula (5) will lead to the effect of approximately twofold lowering of the probability of errors of the first kind – P_1 of the decisions made on small samples, which is unacceptable in the analysis of biometric data.

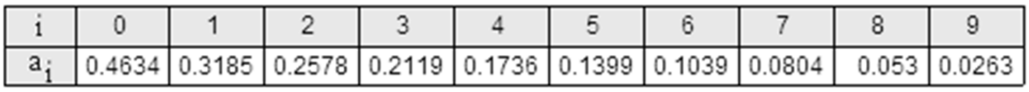
Application of the Shapiro-Wilk test for small samples of biometric data

By analogy with the use of the Chi-square test in the verification of statistical hypotheses, we have the right to use the Shapiro–Wilk test [4] for sampling in 21 examples:

$$\omega_{21}{}^2 = \frac{1}{\sigma^2(x)}\left[\sum_{i=0}^{9} a_i \cdot (x_{20-i} - x_i)\right]^2 , \qquad (6)$$

when a_i – the weight coefficients of the criterion, which are given in Table 1 for a given sample of 21 examples.

Table 1



i	0	1	2	3	4	5	6	7	8	9
a_i	0.4634	0.3185	0.2578	0.2119	0.1736	0.1399	0.1039	0.0804	0.053	0.0263

The distribution of the values of the Shapiro–Wilk test for the normal law and the uniform distribution law are shown in Fig. 2.

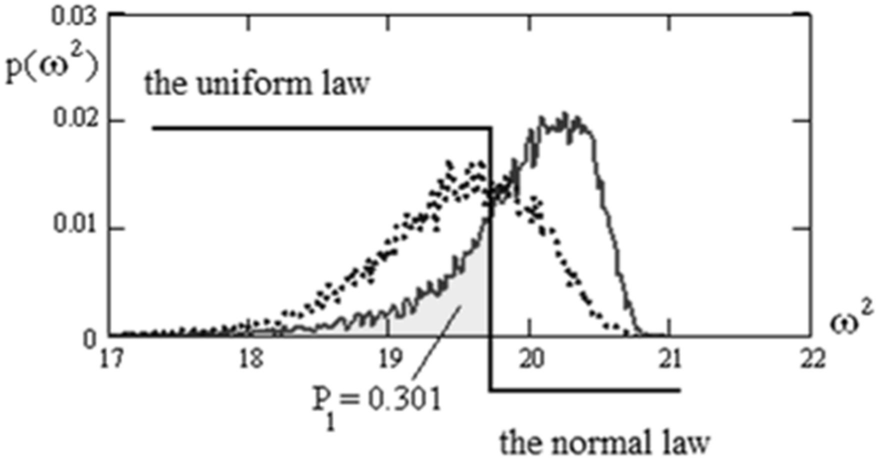


Fig. 2. Distributions of values of the Shapiro–Wilk test for the normal and uniform distributions of values

It can be seen from Fig. 2 that the distribution of the normal law is located in its right-hand side. The distribution of the values of the states of the criterion on the left-hand side of the figure corresponds to the substitution in the formula (6) of data with a uniform law. We divide these distribution laws by a quantizer with a threshold $\omega^2 = 19.8$ and a state "0" for normal data. State "1" will be most likely for uniform data. In this case, the probability of erroneous acceptance of the uniform law data for the normal law is $P_1 = 0.301$.

Representation of the chi-square criterion as an artificial neuron

It should be noted that the neural network emulator of quadratic forms is known in biometrics for a long time [7] and it is a special case of radially basic neurons with the transformation of data of the following form:

$$\begin{cases} y = \sum_{i=1}^{n}\left(\frac{(\nu_i - E(\nu))}{\sigma(\nu)}\right)^2 , \\ z(y) = "0" \; if \;\; y \le k, \\ z(y) = "1" \; if \;\; y > k, \end{cases} \qquad (7)$$

where k – the value of the threshold of the output quantizer of the neuron, ν_i – the value of the i-th biometric parameter in the sample, $E(\nu)$ – the mathematical expectation of the biometric parameter, $\sigma(\nu)$ – the standard deviation of the biometric parameter, n – the number of inputs in the neuron or the size of the training sample.

However, the power of the neural network transformation (7) is much inferior to the classical chi-quararate criterion. Align the power, if in addition to the transformation (7), sort the input data of the neuron [8]:

$$\begin{cases} \tilde{\nu} = sort(\nu), \\ \tilde{y} = \sum_{i=1}^{n} \left(\frac{(\tilde{\nu}_i - E(\tilde{\nu}))}{\sigma(\tilde{\nu})} \right)^2, \\ z(\tilde{y}) = "0" \; if \; \tilde{y} \le k, \\ z(\tilde{y}) = "1" \; if \; \tilde{y} > k. \end{cases} \tag{8}$$

All classical statistical criteria include a sorting operation, while the conventional neurons of this data processing do not use. Combining the sorting operation with conventional neural network transformation operations is technically easy to implement, however, the addition of this operation with other operations of artificial neurons gives them new opportunities.

In the end, we get a quadratic neuron that reproduces Pearson's statistics well. In this case, the Chi-square neuron will be described by a criterion similar to the classical Chi-square table of quantiles of confidence probabilities. Below it is a complete table describing the Chi-square of the neuron for the four values of the quantization thresholds.

Table 2

Name of conversion, sample size	Output comparator threshold	P_1 Normal law	P_2 Uniform law	$\frac{P_1+P_2}{2}$
Chi-square neuron to sample of 21 examples	$k = 9.16$	0.20	0.382	0.291
	$k = 10.16$	0.14	0.457	0.298
	$k = 11.31$	0.10	0.544	0.322
	$k = 12.31$	0.08	0.612	0.341

From Table 2 it is seen that an increase in the threshold of Chi-square neuron leads to a decrease in the errors of the first kind – P_1 (an erroneous refusal to recognize the presented distribution as normal). At the same time, the probability of errors of the second kind – P_2 9an erroneous recognition of the uniform distribution as normal) increases. Obviously, errors of the first and second kind can be reduced by increasing the size of the test sample. However, this path is difficult to implement in practice. A medical student or biologist who already has a sample of 21 examples, often can not increase it. In order to make a sample of 22 examples, a physician needs another patient with the necessary diagnosis. A biologist needs to get another copy of a rare plant or animal.

Representation of the Shapiro-Wilk test as an artificial neuron

By analogy with the Chi-square neuron, we can create the Shapiro-Wilk neuron:

$$\begin{cases} \tilde{\nu} = sort(\nu), \\ \tilde{y} = \frac{1}{(\sigma(\tilde{\nu}))^2} \left[\sum_{i=0}^{9} a_i \cdot (\tilde{\nu}_{20-i} - \tilde{\nu}_i) \right]^2, \\ z(\tilde{y}) = "0" \; if \; \tilde{y} \ge k, \\ z(\tilde{y}) = "1" \; if \; \tilde{y} < k. \end{cases}$$

At different values of the quantization threshold, this artificial neuron will have different error ratios of the first and second kind (Table 3).

Table 3

Name of conversion, sample size	Output comparator threshold	P_1 Normal law	P_2 Uniform law	$\frac{P_1+P_2}{2}$
Shapiro–Wilk neuron to sample of in 21 examples	$k = 19.8$	0.31	0.30	0.305
	$k = 19.6$	0.19	0.44	0.315
	$k = 19.4$	0.12	0.57	0.345
	$k = 19.2$	0.08	0.71	0.395

Description of the network of two neurons, summarizing two classical cooperating statistical criteria

Domestic neural network biometry is based on the fact that one neuron is responsible for one bit of the output code [9–11]. Let's keep this rule, integrating the Chi-square neuron and the Shapiro–Wilk neuron into the network. In this situation, the output code corresponding to the state "00" will mean the confirmation of the hypothesis of normality of the analyzed data. The state of the output code "11" will be considered as a rejection of the hypothesis of normality by both criteria. The states "01" and "10" will be considered as indeterminate (they contradict each other). In order to reveal this uncertainty it is necessary to complicate the decisive rule, which is beyond the scope of this article.

Table 4

Name of conversion, sample size	Output comparators thresholds	P_1 Normal law	P_2 Uniform law	$\frac{P_1+P_2}{2}$
The first Chi-square neuron to sample in 21 examples The second Shapiro–Wilk neuron To sample in 21 examples	$k_1 = 9.16$ $k_2 = 19.0$	0.043	0.205	0.124
	$k = 9.16$ $k_2 = 19.2$	0.061	0.291	0.176
	$k_1 = 9.16$ $k_2 = 19.4$	0.081	0.395	0.238
	$k_1 = 9.16$ $k_2 = 19.6$	0.112	0.491	0.301
	$k_1 = 9.16$ $k_2 = 19.8$	0.145	0.556	0.351

If we compare the data of Table 1–3, we can observe a significant decrease in the probabilities of errors of the first and second kind for neural network generalization of two statistical criteria. Since the probabilities of errors of the first and second kind vary in different directions, when comparing data it is convenient to use their averaging (the last column of the compared tables).

In turn, comparing Table 1 and Table 2, we can make an unambiguous conclusion that the Chi-square neuron is more powerful than the Shapiro–Wilk neuron. According to the average error probabilities of the first and second kind, the Chi-square neuron gives a value of 0.291, while for the Shapiro–Wilk neuron this value is 0.305.

If we combine neurons according to the rule, the average value of the error probabilities can be reduced to 0.124 (more than twice).

Conclusion

Obviously, by analogy with the proposed neural network generalization of two statistical criteria, it is possible to construct a similar generalization for a larger number of statistical criteria. The addition of each new statistical criterion always improves the quality of the statistical decisions made by the neural network. The main condition for generalized criteria is that their values should not be strongly correlated. In particular, the positive result of the generalization described in this article is due to the fact that the data of the statistical criteria under consideration are not completely correlated $corr(\chi^2, \omega^2) = -0.726$ for a sample of 21 examples. The higher the modulus of the correlation coefficient of the combined statistical criteria, the worse the final result is obtained.

It is also obvious that any of the known statistical criteria can be represented as a certain neuron with a certain quantization threshold. In Table 4, each row connects the different quantization thresholds of two different neurons. If the number of generalized statistical criteria increases, then the number of related quantization thresholds should increase proportionally. Finding the optimal ratio of thresholds for neurons of different statistical criteria is an independent task.

References

1. GOST R 52633.5-2011 Information protection. Information protection technology. Automatic learning of neural network converters biometry-access code.
2. Technical specification (project, public discussion started on 01.02.2017 by members of TK 26 "Cryptographic protection of information") PROTECTION OF NEUROET-NET BIOMETRIC CONTAINERS WITH THE USE OF CRYPTOGRAPHIC ALGORITHMS.
3. Р 50.1.037-2002 Recommendations of standardization. Applied statistics. Rules for verifying the agreement between the experimental distribution and the theoretical distribution. Part I. Criteria of the type χ2. State standard of Russia. – Moscow, 2001. – 140 p.
4. *Kobzar, A. I.* Applied mathematical statistics. For engineers and scientists / A. I. Kobzar. – Moscow : FIZMATLIT, 2006. – 816 p.
5. GOST R 52633.3-2011 Information protection. Information protection technology. Testing the robustness of highly reliable biometric protection to assault attacks.
6. *Akhmetov, B. S.* Algorithms for testing of biometric-neural network mechanisms of information protection / B. S. Akhmetov, V. I. Volchikhin, A. I. Ivanov, A.Y. Malygin. – Kazakhstan, Almaty, KazNTU Satpayev, 2013. – 152 p. – URL: http://portal.kazntu.kz/files/publicate/2014-01-04-11940.pdf
7. *Akhmetov, B. B.* Multidimensional statistics of essentially dependent biometric data, which are generated by neural network emulators of quadratic forms : monograph / B. B. Akhmetov, A. I. Ivanov. – Kazakhstan, Almaty : From LEM, 2016. – 86 p.
8. *Volchikhin, V. I.* Absolutely stable algorithm of automatic learning of networks of probabilistic neurons "Kramer-von Mises" on small samples of biometric data / V. I. Volchikhin, A. I. Ivanov, S. E. Vjatchanin, E. A. Malygina // Izvestiya Vysshikh Uchebnykh Zavedenii. The Volga region. Technical science. – 2017. – № 2 (42). – P. 55–65.
9. Justification and selection of statistical criteria for correct evaluation of small samples data of biometric images / A. I. Ivanov, E. A. Malygina, Yu. I. Serikova, S. E. Vjatchanin, E. N. Kupriyanov // Proceedings of the International Symposium Reliability and quality. – 2018. – Vol. 1. – P. 176–178.
10. Neural networks improve the quality of decisions by switching from a double quantization functions to functions of quantization with multiple levels / E. A. Malygina, A. I. Solopov, Yu. I. Serikova, A. I. Gazin, E. N. Kupriyanov // Proceedings of the International Symposium Reliability and quality. – 2018. – Vol. 1. – P. 182–183.
11. Features testing neural network converters biometrics code on small test samples of alien images / V. I. Volchihin, A. Yu. Malygin, I. V. Urnev, A. V. Serikov, E. N. Kupriyanov // Proceedings of the International Symposium Reliability and quality. – 2018. – Vol. 1. – P. 52–53.

References

1. *GOST R 52633.5-2011 "Information protection. Information protection technology. Automatic learning of neural network converters biometry-access code."*
2. *Technical specification (project, public discussion started on 01.02.2017 by members of TK 26 "Cryptographic protection of information") PROTECTION OF NEUROET-NET BIOMETRIC CONTAINERS WITH THE USE OF CRYPTOGRAPHIC ALGORITHMS*
3. *R 50.1.037-2002 Recommendations of standardization. Applied statistics. Rules for verifying the agreement between the experimental distribution and the theoretical distribution. Part I. Criteria of the type χ2. State standard of Russia*. Moscow, 2001, 140 p.
4. Kobzar A. I. *Applied mathematical statistics. For engineers and scientists*. Moscow: FIZMATLIT, 2006, 816 p.
5. *GOST R 52633.3-2011 "Information protection. Information protection technology. Testing the robustness of highly reliable biometric protection to assault attacks."*
6. Akhmetov B. S., Volchikhin V. I., Ivanov A. I., Malygin A. Y. *Algorithms for testing of biometric-neural network mechanisms of information protection*. Kazakhstan, Almaty, KazNTU Satpayev, 2013, 152 p. Available at: http://portal.kazntu.kz/files/publicate/2014-01-04-11940.pdf
7. Akhmetov B. B., Ivanov A. I. *Multidimensional statistics of essentially dependent biometric data, which are generated by neu-ral network emulators of quadratic forms: monograph*. Kazakhstan, Almaty. From LEM, 2016, 86 p.

8. Volchikhin V. I., Ivanov A. I., Vjatchanin S. E., Malygina E. A. University proceedings. Volga region. Engineering sciences. 2017, no. 2 (42), pp. 55–65.
9. Ivanov A. I., Malygina E. A., Serikova Yu. I., Vjatchanin S. E., Kupriyanov E. N. *Proceedings of the International Symposium Reliability and quality*. 2018, vol. 1, pp. 176–178.
10. Malygina E. A., Solopov A. I., Serikova Yu. I., Gazin A. I., Kupriyanov E. N. *Proceedings of the International Symposium Reliability and quality*. 2018, vol. 1, pp. 182–183.
11. Volchihin V. I., Malygin A. Yu., Urnev I. V., Serikov A. V., Kupriyanov E. N. *Proceedings of the International Symposium Reliability and quality*. 2018, vol. 1, pp. 52–53.

Иванов Александр Иванович
доктор технических наук, доцент,
ведущий научный сотрудник,
Пензенский научно-исследовательский
электротехнический институт
(440000, Россия, Пенза, Советская площадь, 9)
E-mail: bio.ivan.penza@mail.ru

Вятчанин Сергей Евгеньевич
доцент, начальник кафедры радио- и
космической связи,
Пензенский государственный университет
(440026, Россия, г. Пенза, ул. Красная, 40)
E-mail: ivo@.pnzgu.ru

Малыгина Елена Александровна
кандидат технических наук, научный сотрудник,
межотраслевая лаборатория тестирования
биометрических устройств и технологий,
Пензенский государственный университет
(440026, Россия, г. Пенза, ул. Красная, 40)
E-mail: e-mail: ivo@.pnzgu.ru

Лукин Виталий Сергеевич
аспирант,
Пензенский государственный университет
(440026, Россия, г. Пенза, ул. Красная, 40)
E-mail: ivo@.pnzgu.ru

Ivanov Alexander Ivanovich
doctor of technical sciences, associate professor,
senior researcher,
Penza Scientific and Research
Electrotechnical Institute
(440000, 9 Soviet square, Penza, Russia)

Vjatchanin Sergej Evgenevich
associate professor,
head of sub-department
of radio-space communications,
Penza State University
(440026, 40 Krasnaya street, Penza, Russia)

Malygina Elena Aleksandrovna
candidate of technical sciences, researcher,
interdisciplinary laboratory testing
of biometric devices and technologies,
Penza State University
(440026, 40 Krasnaya street, Penza, Russia)

Lukin Vitaliy Sergeevich
postgraduate student,
Penza State University
(440026, 40 Krasnaya street, Penza, Russia)

Образец цитирования:

Ivanov, A. I. Precision statistics: neuroet networking of chi-square test and shapiro-wilk test in the analysis of small selections of biometric data / A. I. Ivanov, S. E. Vjatchanin, E. A. Malygina, V. S. Lukin // Надежность и качество сложных систем. – 2019. – № 2 (26). – С. 27–34. – DOI 10.21685/2307-4205-2019-2-4.