Надежность № 2 2016
Dependability no.2 2016

Original article
УДК: 519.24; 57.017
DOI: 10.21683/1729-2640-2016-16-2-43-48

Estimation of quality of a small sampling biometric data using a more efficient form of the chi-square test

Berik B. Akhmetov, *International Informatization Academy (IIA), Turkestan, Kazakhstan, e-mail: berik.akhmetov@ayu.edu.kz*
Aleksander I. Ivanov, *Laboratory of biometric and neural network technologies, JSC Penza Research Electric and Technical Institute, Penza, Russia, e-mail: ivan@pniei.penza.ru*

Berik B. Akhmetov

Aleksander I. Ivanov

Abstract. Aim. *The purpose is to increase the power of the Pearson's chi-square test so that this test will become efficient on small test samplings. . It is necessary to reduce the scope of a test sample from 200 examples to 20 examples while maintaining the probability of errors of the first and the second kind. Selection of 20 examples of biometric images is considered by users to be a comfortable level of effort. The need to select more examples is perceived by users negatively.*
Methods. *The article offers one more (the second) form of the Pearson test that is much less sensitive to the scope of data in a test sampling. It is shown that a traditional form of the chi-square test is more sensitive to the scope of a test sampling than the Cramer-von Mises test. The offered (second) form of the chi-square test is less sensitive to the scope of a test sampling than a classical form of the chi-square test and less sensitive than the Cramer-von Mises test as well. This effect is achieved by the transition from the space of frequency of occurrence of events and probabilities of a group of similar events occurring in the space of more accurately evaluated junior statistical moments (mean and standard deviation). The fractal dimension of the new synthetic form of chi-square test coincides with the fractal dimension of the classical form of the chi-square test.* ***Results.*** *The offered second variant of the chi-square test is presumably one of the most powerful of all existing statistical tests. The analytical description of correlation of standard deviations of a classical form of the chi-square test and a new form of the chi-square test is given. The standard deviation of the second form of the chi-square test decreases by half on retention of a statistical expectation on samplings of the same scope. The latter is equivalent to a four-time reduction of the requirements to the scope of a test sampling within the interval from 16 to 20 examples. Power gain as the result of the application of a new test is growing with the growth of a test sampling scope.* ***Conclusions.*** *When creating a classical chi-square test in 1900, Pearson was guided by limited computing opportunities of the existing computer facilities, and he had to rely on the analytical relations that he found. Today the situation has changed and there are no more restrictions in relation to the engaged computing resources. However we continue to rely on those created with computing resources of 1900 by inertia. Probably, we should try to consider modern opportunities of computer facilities and to build more powerful options of statistical tests. Even if new tests will require a search of large number of possible states (they will have big tables calculated in advance instead of analytical relations), it is not a constraining factor today. When data is insufficient (in biometrics, in medicine, in economy) a computing complexity of statistical tests does not play a special role if the result of estimations is more accurate.*

Keywords: *multivariate statistical analysis, chi-square test, small samplings of test biometric data.*

For citation: *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, no. 2, pp. 43-48. (in Russian) DOI: 10.21683/1729-2640-2016-16-2-43-48*

Introduction

An information-oriented society requires active use of Internet resources. State and private organizations create personal user accounts on their web-sites. Unfortunately, the current practice of password security of the access to personal user accounts is quite vulnerable. Users are not able to remember long randomly chosen passwords. An owner of an information resource cannot be sure that it will be exactly his host who will get access to the personal account. A password may be intercepted by a backdoor. Besides it is quite easy to spoof an IP address of an Internet-user.

To protect the access to user accounts, the technologies of personal biometric authentication are currently being developed by means of transformation of personal biometric data into a cryptographic key of a person, or into a long randomly chosen password. The following biometric images are used: an image of finger mark [1], an image of eye iris [2], voice password [3], hand-written password [4], an image of blood vessels of an eye ground or a palm [5]. Naturally, biometrics-code transformers cannot be ideal, they have

the probabilities of errors of the first and the second kind. It becomes necessary to test the errors of the first and the second kind on real biometric data. Moreover, when setting the "indistinct extractors" [1, 2, 3] and training the neural network transformers [4, 5] it is necessary to control the absence of gross errors in biometric data. Basically, on a small number of examples of a biometric image it is necessary to control the indicator of relationship of the biometric data distribution to the multivariate normal law [6]. Formally, for this purpose we can use a simple univariate Pearson chi-square test [7, 8], but such approach is far from the best one. In this article we will try to show that a classic form of the Pearson chi-square test is by far not the only one, i.e. it is possible to set the task on searching of the most effective Pearson's functionalities, considering different peculiarities of their practical application.

Occurrence of quantization noise at the statistical processing of small samplings

Let us consider the simplest situation, when a test or a learning sampling is represented by 9 examples of the "Self" image. Since a continuous function of probability $P(x)$ of the first biometric parameter v_1 is a small sampling function, we have to describe it by a step monotonously increasing function $\tilde{P}(x)$, as it is shown on the left part of Figure 1.

To construct a step monotonously increasing approximation $\tilde{P}(x)$, it is necessary to sort biometric data in its ascending order:

$$x_i = sort(v_{1,i}) \text{ for } i = 0, 1, 2, \ldots, n, \qquad (1)$$

where n is a dimension of a test sampling, or a number of quanta of approximation of a monotone function of probability.

In this case a monotonously increasing step function will be described by the following piecewise constant approximation:

$$\tilde{P}(x_i) = \frac{i}{n}. \qquad (2)$$

An approximation error or a quantization noise is found as a difference of a continuous probability function and its step approximation:

$$\Delta P(x) = P(x) - \tilde{P}(x). \qquad (3)$$

The lower part of Figure 1 shows the functions of the quantization error or quantization noise caused by small test samplings.

In the context of mentioned above, the Kolmogorov-Smirnov test [7] should be considered as a search of the maximum value of the module of approximation error:

$$\sup_{-\infty<v<+\infty} \left|P(x) - \tilde{P}(x)\right| = \max \left|\Delta P(x_i)\right| \qquad (4)$$

or a choice of the biggest from local maximums of the quantization noise.

From the same perspective, the Cramer-von Mises test [7] is the estimation of standard deviation of the quantization noise of the continuous probability function:

$$\int_{-\infty}^{\infty} \left\{P(x) - \tilde{P}(x)\right\}^2 \cdot dx = \int_{-\infty}^{\infty} \left\{E(\Delta P(x)) - \Delta P(x)\right\}^2 \cdot dx =$$

$$= \int_{-\infty}^{\infty} \left\{\Delta P(x)\right\}^2 \cdot dx = \sigma^2(\Delta P(x)), \qquad (5)$$

if the condition of a zero statistical expectation of a quantization noise is fulfilled $E(\Delta P(x))=0$.

It should be emphasized that the Kolmogorov-Smirnov test (4) always has a lower power in comparison to the Cramer-von Mises test (5). The Kolmogorov-Smirnov test (4) is a point test, and the Cramer-von Mises test (5) is integral.

It is evident that with the increase of n of a test sampling, both these statistical tests are getting power of estimations, however, the estimation by an integral test is always more reliable than the point estimation.

Classic variant of the Pearson chi-square test

General practice of check of statistical hypotheses in most sectors of industry is reduced to the construction of

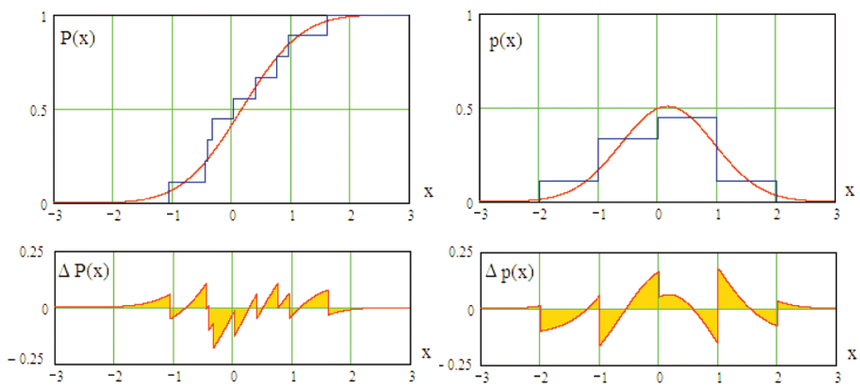

Fig. 1. Effects of quantization of a continuous probability of values distribution and a continuous density of values distribution by 9 examples that cause continuous noise of a quantization error

histograms of the available data (right part of Figure 1) and to the calculation of the classic chi-square test:

$$\chi^2 = \sum_{i=1}^{k} \left\{ \frac{\left(\frac{n_i}{n} - P_i \right)^2}{P_i} \right\} = \sum_{i=1}^{k} \left\{ \frac{\left(\tilde{P}_i - P_i \right)^2}{P_i} \right\}, \qquad (6)$$

where n_i is the number of samples, occurring in the i-th column of the histogram, P_i is the probability of occurrence in the i-th column of the histogram of the theoretical distribution, k is the number of columns of the histogram.

Wide application of chi-square test is determined by the fact that for this test we know the analytical description of distribution density:

$$p(\chi^2, m) = \frac{1}{m} \left\{ \frac{1}{2^{\frac{m}{2}} \cdot \Gamma\left(\frac{m}{2}\right)} \left\{ x^{\frac{m}{2}-1} \cdot \exp\left(\frac{-x}{2}\right) \right\} \right\}, \qquad (7)$$

where G(·) is a gamma-function, m is the number of degrees of freedom.

The number of degrees of freedom m can be set in different ways [8]. For instance, it can be defined through the scope n of a test sampling:

$$m = \sqrt{n} - 3 = k - 3, \qquad (8)$$

if the number k of the histogram columns is chosen by the rounding off up to the nearest integer of the value $\sqrt{n}$:

$$k = round\left(\sqrt{n}\right). \qquad (9)$$

Let us note that the value of the number k of the histogram columns and the value of the number m of the degrees of freedom for a classic chi-square test always turns out to be much smaller in comparison to the scope n of a test sampling. So, the error of step approximation of density of distribution of the values $\Delta p(x) = p(x) - \tilde{p}(x)$ (right part of Figure 1) is always more than the error of approximation of a probability function (3). So in the left part of Figure 1 the approximation of the probability function is constructed using 9 steps, whereas the function of approximation of probability distribution is constructed using just 4 steps in the right part of Figure 1. The quantization noises of the Cramer-von Mises test turn out to be less than the quantization noises of the classic chi-square test (6).

It means that the power of the Cramer-von Mises test is always higher than the power of the classic Pearson chi-square test (3).

Comparison by power of the Cramer-von Mises test with the classic chi-square test

We shall proceed from the fact that biometric data for each of the parameters under control is distributed normally. Then the quality of data of one parameter can be estimated by both, the Cramer-von Mises test, and the chi-square test [7, 8]. To compare the tests let us use the data distribution by the uniform law as an alternative. The results of the numerical simulation for the samplings of 9 examples are shown in Figure 2.

When making a decision, a match threshold plays an important role. Each match threshold gives its probability value P_1 for the errors of the first kind and probability value P_2 for the errors of the second kind. To exclude uncertainty of a match threshold, let us compare the results in the point with equal probability of errors $P_1 = P_2 = P_{EE}$.

Figure 2 shows that the distribution of data received by the Cramer-von Mises test gives the value $P_1 = P_2 = P_{EE} = 0.306$. Under the same conditions the chi-square test gives the value of equally probable errors $P_1 = P_2 = P_{EE} = 0.327$. The results are approximately 9% worse. It means that the chi-square test requires the sampling of 10 examples, whereas for the Cramer-von Mises test only 9 examples are required. The relief in the requirements to the dimensions of a test sampling is explained by the fact that the quantization error of the probability function $P(x)$ turns out to be smaller than the quantization error of the distribution density $p(x)$ (see Figure 1).

Calculation procedure of the Cramer-von Mises test is approximately $\sqrt{n}$ times more effective for the suppression

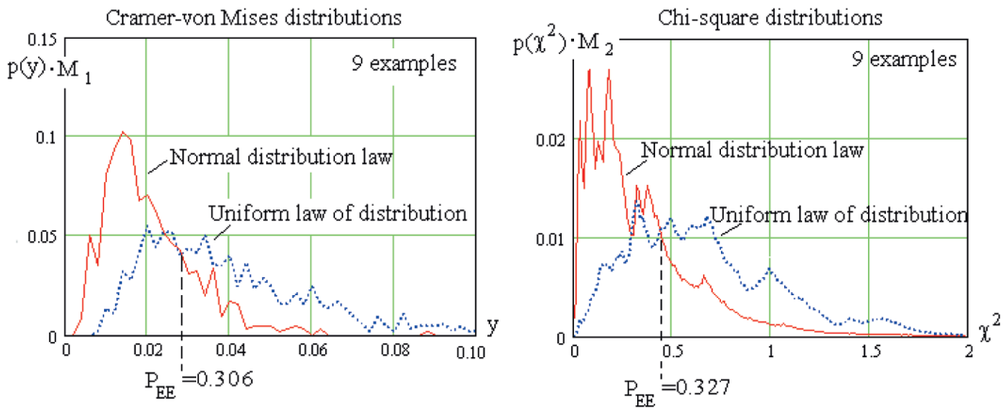

Fig. 2. Distributions of the values of the Cramer-von Mises test and chi-square test for the normal distribution law, and for its alternative in form of the uniform law of distribution for the samplings of 9 examples

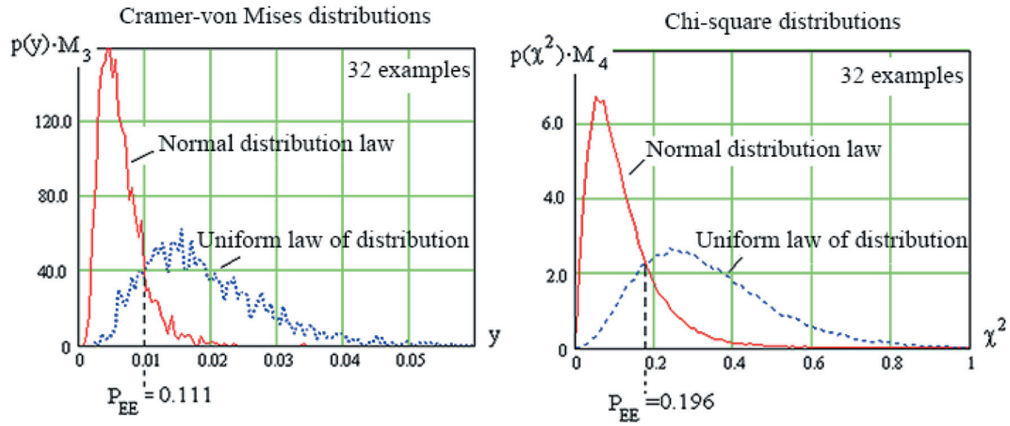

Fig. 3. Distributions of the values of the Cramer-von Mises test and chi-square test for the normal distribution law, and for its alternative in form of the uniform law of distribution for the samplings of 32 examples

of quantization noises in comparison to the data calculation by the chi-square test. The more a test sampling is, the stronger the effect of a more intense suppression of quantization noises is. Figure 3 shows the simulation data for the sampling of 32 examples.

Figure 3 shows that for the sampling of 32 examples, the Cramer-von Mises test gives дает $P_{EE} = 0.111$, which is 43% less than the chi-square test gives: $P_{EE} = 0.196$. In the first approximation we may expect about 40% decline of the scope of the test sampling if to pass form the chi-square test to the Cramer-von Mises test.

One more variant of the chi-square test making the best use of the limited scope of the test sampling

Basically, both the Cramer-von Mises test and the Pearson chi-square test are the schemes of the suppression of quantization noises. I.e. we can try to amplify the property of these statistical functionalities to suppress the quantization noises. For example, we can use a supplementary digital-data filter configured to smooth the rises of piecewise constant approximation of the function of value distribution density [9, 10].

One more way is to check other possible variants of the calculation of the chi-square test. In particular, biometrics has been using the so called Pearson functionalities (networks of Pearson functionalities [11]) for the data preliminary normalized by a standard deviation:

$$\chi_2^{\,2} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\left(E(x)-x_i\right)^2}{\sigma(x)}\right\}, \qquad (12)$$

where $E(x)$ is a statistical expectation of data of the test sampling, $\sigma(x)\approx 1$ is a standard deviation of the preliminary normalized data of the test sampling.

It should be noted that at the processing of biometric data, a preliminary normalizing of data is usually made by its standard deviation, in attempt to fulfill the condition $\sigma(x)=1$. However, on small samplings, this condition cannot be fulfilled. For instance, a relative error of the calculation of standard deviation on small samplings of 20 examples is random and can comprise up to ±30%, at the smaller samplings an error may be even more. To compensate normalizing errors in formula (12) there occurs the term close to, but always different from entity.

Let us note that the equation (12) makes the summing up of squared deviations by all calculations of the test sampling, whereas the classic chi-square test (6) sums up the squared deviations only by the number of histogram columns. As $n > k$, then we can expect a higher power of the chi-square test (12) in comparison to the similar classic chi-square test (6).

Comparison by power of two variants of the chi-square test

The Cramer-von Mises test turns out to be more powerful than the classic Pearson chi-square test due to the fact that it presses upon the quantization noises of a smaller amplitude (let us compare the left and the right parts of Figure 1). However, for the Cramer-von Mises test there is no analytical description, and that is its huge disadvantage. That is why we shall further compare only the powers of two modifications of the chi-square tests (6) and (12).

The results of simulation modeling with 9 and 32 examples in the test sampling for the test (12) are provided in Figure 4.

If to compare the crossing of the distributions on the left graph of Figure 4 that gives $P_{EE} = 0.106$, and the analogous crossing on the right graph of Figure 2 that gives $P_{EE} = 0.327$, then we will get approximately a three-time profit in power between two variants of tests (confidence in the solutions based on them). With a growth of the scope of the test sampling, the profit of the second form of the chi-square test increases. If to compare the data of crossing of the distributions on the right graph of Figure 4, providing $P_{EE} = 0.007$, with the data on the

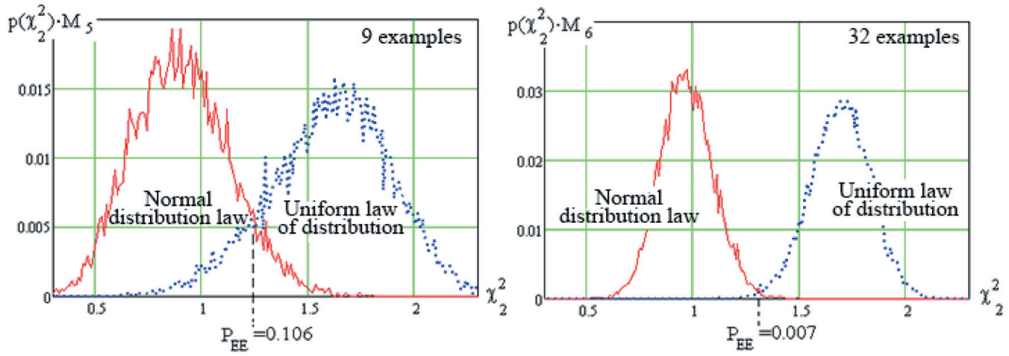

Fig. 4. Histograms of the values distribution of the second, more effective form of the chi-square test

right graph of Figure 3, providing $P_{EE} = 0.196$, then we will get a 28 times profit.

It turns out that both, the Cramer-von Mises test and the second form of the chi-square test are more powerful than the classic chi-square test due to the calculation of these two tests by the whole sampling. The classic chi-square test looses out to these two tests because it sums up a squared error by number of columns of the empirical histogram. It is quite easy to make sure that the Cramer-von Mises test is in its power within the interval between two forms of the chi-square tests.

Analytical description of the second form of the chi-square test

The essential property of the second form of the chi-square test is that for independent data its statistical properties are very well described by normal distribution laws. And the statistical expectation for the distribution of formula (12) with an absence of a normalizing factor $1/n$ is close to the scope of the test sampling:

$$E(\chi_2^2) \approx n - 0,778. \qquad (13)$$

This statement is illustrated by the positions of maximums of unbroken curved lines of Figure 5. I.e. the value of statistical expectation of the second form of the chi-square test is almost always taken (with accuracy to the correction 0.778) from the classic chi-square distribution with the number of degrees of freedom $m = n$ (dotted curves in Figure 5).

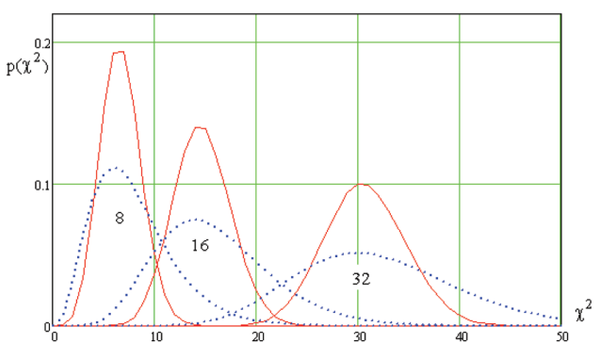

Fig. 5. Distributions of values of the second form of the chi-square test (unbroken curved lines) for $n = 8, 16, 32$ and classic chi-square distributions, when the number of degrees of freedom coincides with the scope of a sampling (dotted curves)

Figure 5 shows that standard deviations of normal distribution laws are always smaller than standard deviations of classic Pearson chi-square tests. At statistical calculations with an engineering accuracy standard deviations are described by the following equation:

$$\sigma(n, \chi_2^2) \approx \sqrt{\frac{E(\chi_2^2)}{2}} \approx \sqrt{\frac{n - 0,778}{2}} \qquad (14).$$

It should be underlined that the accuracy of approximation (14) increases with a normalizing of values distribution of the second form of the chi-square distributions. So for $n = 8, 16, 32$ a relative error of approximation of standard deviation shall be $\Delta\sigma = 3.10\%, 1.70\%, 0.75\%$, which is quite acceptable for an engineering practice of statistical processing of biometric data.

Conclusion

We are used to the fact that for reliable statistical estimations the samplings with hundreds of examples are required. Only in the case, when we have a large sampling and rely on standardized recommendations [8], there is a confidence in the quality of the performed statistical analysis. This is the current technical practice.

This article showed that the power of the chi-square test can be essentially increased, i.e. reliable estimations can be obtained on much smaller data samplings. It is very important for practice, especially if a destructing testing of costly products is performed. Already existing practice of statistical processing of biometric data proves true of this important stipulation.

Essential estimation resources are not a problem nowadays. We can make a statistical data processing more complicated, for instance, making it multivariate [11]. Today we have a technical capability of multiply complicating the applied methods of statistical analysis. In the last century we used to take one test and had to be satisfied with its results, but today we can use dozens of well-known statistical tests and, if necessary, create new tests especially for a certain practical task.

References

1. Ramírez-Ruiz J., Pfeiffer C., Nolazco-Flores J. Cryptographic Keys Generation Using FingerCodes. //Advances in Artificial Intelligence – IBERAMIA-SBIA 2006 (LNCS 4140), p. 178-187, 2006

2. Monrose F., Reiter M., Li Q., Wetzel S. Cryptographic key generation from voice. In Proc. IEEE Symp. on Security and Privacy, 2001

3. Feng Hao, Ross Anderson and John Daugman. Crypto with Biometrics Effectively, IEEE TRANSACTIONS ON COMPUTERS, VOL. 55, NO. 9, SEPTEMBER 2006.

4. Yazov Y.K. and others. Neural network protection of personal biometric data. // Y.K. Yazov (editor and author), co-authors V.I. Volchikhin, A.I. Ivanov, V.A. Funtikov, I.G. Nazarov // M.: Radiotechnics, 2012. 157 p. IBSN 978-5-88070-044-8.

5. Akhmetov B.S., Ivanov A.I., Funtikov V.A., Bezyaev A.V., Malygina E.A. Technology of use of large neural networks to transform indistinct biometric data into the access key code. Monograph, Kazakhstan, Almaty, LLP Publishing House LEM, 2014. -144 p.

6. Akhmetov B.S., Volchikhin V.I, Ivanov A.I., Malygin A.Y. Algorithms of testing of biometric and neural network mechanisms of information security, Kazakhstan, Almaty, KazNTU after K.I.Satpayev, 2013.- 152 p. ISBN 978-101-228-586-4.

7. Kobzar A.I. Practical mathematical statistics for engineers and researchers. M. FIZMATLIT, 2006, 816 p.

8. R 50.1.037-2002 Standartization recommendations. Practical statistics. Rules of check of fitting of practical distribution to the theoretical one. Part I. $\chi 2$ criteria. State Standard of Russia. Moscow-2001, 140 p.

9. Akhmetov B.S., Ivanov A.I., Serikova N.I., Funtikova Y.V. Algorithm of imitative increase of number of degrees of freedom at the analysis of biometric data by the goodness-of-fit of the chi-square test. Reporter of the National academy of sciences of the Republic of Kazakhstan. No.5, 2014. p. 28-:-34.

10. Serikova N.I., Ivanov A.I., Kachalin S.V. Biometrical statistics: smoothing of histograms constructed on a small learning sampling. /Reporter of SibSAU 2014 No. 3(55) p.146-150.

11. Akhmetov B.B., Ivanov A.I., Bezyaev A.V., Funtikova Y.V. Multivariate statistical analysis of biometric data by the network of partial Pearson test. // Reporter of the National academy of sciences of the Republic of Kazakhstan. – Almaty, 2015. No.1. P. 5-11.

About the authors

Berik B. Akhmetov – PhD., Member of the International Informatization Academy (IIA), Vice-president of the International Hoca Ahmet Yesevi Turkish-Kazakh University.

29 B.Sattarkhanov Avenue, Bld. of Rectorate, Turkestan, 161200, Kazakhstan, tel.: +7 (72533) 3-35-77, e-mail: berik.akhmetov@ayu.edu.kz

Alexander I. Ivanov – Dr. Sci., Associate Professor – Head of Laboratory of biometric and neural network technologies, JSC Penza Research Electric and Technical Institute.

9 Sovetskaya Str., Penza, 440000, Russia, tel.: +7 (841-2) 59-33-10, e-mail: ivan@pniei.penza.ru