Research Journal of Applied Sciences 10 (12): 956-967, 2015
ISSN: 1815-932X
© Medwell Journals, 2015

Prospects for Multiple Reductions in Test Samples with a Multivariate, Multicriteria, The Neural Network Statistical Analysis of Biometric Data

[1]Berik Akhmetov, [2]Alexander Ivanov, [2]Alexander Malygin, [3]Zhibek Alibiyeva,
[3]Kaiyrkhan Mukapil, [3]Gulzhanat Beketova and [3]Nazym Zhumangalieva,
[1]Hoja Ahmet Yasaui International Kazakh-Turkish University, Turkistan, Kazakhstan
[2]Penza State University, Penza, Russia,
[3]K.I. Satpayev Kazakh National Research Technical University, Almaty, Kazakhstan,

Abstract: It is shown that the classical Chi-Square test has insufficient capacity for efficient processing of biometric data. It is shown that there is a possibility to increase the power of statistical processing through the use of several well-known statistical tests, through the neural network combining their private decisions. Contains tables of formulas promising statistical criteria that complement already used statistical tests. Considered the influence of quantization errors caused by the small amount of experience in the test sample. Proposed to raise the reliability of the estimates due to the digital smoothing of histograms with uniform quantization step. Shows the tables and nomograms to assess the reduction in the probability of errors of the first and second order transition to multivariate statistical analysis of biometric data.

Key words: A set of statistical criteria, the capacity of statistical criteria, multivariate statistical processing, artificial neural networks, reduction

INTRODUCTION

One of the most popular in the statistical analysis of the data is Pearson. In particular, only the Chi-Squared Pearson devoted entirely to the first part of the recommendations of the State Standard while all other criteria are described in the second part of the recommendations (Anonymous, 2002). Detailed description of Pearson in the first part of the recommendations of the State Standard reflects the high demand for this particular industry criteria. Most of the methods of statistical analysis of experimental data using built on the Chi-Square test:

$$\chi^2 = n\sum_{i=1}^{k}\frac{\left(\frac{b_i}{n}-\tilde{p}_i\right)^2}{\tilde{p}_i} \quad (1)$$

Where:

b_i = Number of experiments have got the ith interval histogram the expected theoretical probability of hitting the ith interval histogram
n = The number of tests in the test sample
k = Number of columns of the histogram

The popularity of using the Chi-Square of Pearson in the industry is largely due to the fact that when $n\to\infty$ its distribution is described by the gamma function with m = k-1 degrees of freedom:

$$p_{\chi^2}(n=\infty, m=k-1, x) = \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}x^{\frac{m}{2}-1}e^{\frac{-x}{2}} \quad (2)$$

Analytical description of Eq. 2 was obtained by Pearson in 1904 and played a crucial role in the first half of the 20th century when the computing power used in the statistical processing of data were very, very limited.

It should be emphasized that the presence of an analytical description of the Chi-Squared Pearson made this criterion is the most popular among practitioners (Kobzar, 2006; Mirvaliev and Nikulin, 1992; Lemeshko and Postovalov, 1998) and among researchers belonging to different school of mathematics (Aguirre and Nikulin, 1994; Mann and Wald, 1942; Gjlberg and Leuine, 1945; Aroian, 1973). Significant problems arise when trying to use of the Chi-Square test for the statistical analysis of biometric data. These problems are the subject of numerous articles in professional journals Oxford

Corresponding Author: Zhibek Alibiyeva, Institute of Information and Telecommunication Technologies, Kazakh National Technical University Named after K.I.Satpayev, Satpayev Street, 22, 050013 Almaty, Russia

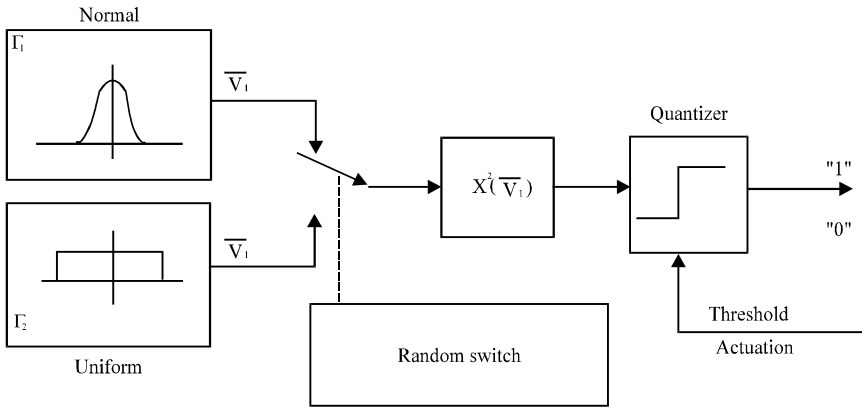

Fig. 1: Block diagram of the organization of the numerical experiment on capacity of one-dimensional Chi-Square test

"Biometrika" (Cochran, 1954; Gilbert, 1977; Pearson, 1959; Cadwell, 1952) which comes out since 1901 and in the last century was the most authoritative edition which reflected the many features of biometric statistics. Unfortunately, the traditional use of the Chi-Square test for multivariate dependent biometric data has been poorly studied and one-dimensional test circuits require large amounts of data. In particular, for decision-making on the one-dimensional criteria of the Pearson with 0.99 confidence level is necessary to use a sample of 400 test results. The use of such a large test samples is unacceptable for biometrics, it is necessary to achieve their reduction by about an order.

As a rule, biometric data samples are low. The number of analyzes of biomaterials in medicine is limited and each of analyzes has a sufficiently high cost. At the training of biometric authentication (Bolle *et al*., 2007; Akhmetov *et al*., 2013, 2014c) people feel comfortable when it is neccessary to show from 9-20 examples of their biometric image. If the requiring people to presenting their 30 or more examples of biometric images, people perceive it negatively. According to users, significantly decreases ergonomics biometric authentication because of the need to test and train them on large samples.

Assessment of capacity one-dimensional Chi-Square criterion of statistical verification of likelihood hypothesis test of the normal distribution: When organizing numerical experiment, starting from the fact that should be checked two statistical hypotheses. The first hypothesis is that these test samples have normal distribution of values. The second hypothesis is that the data of the same sample may have a normal distribution of values. As a consequence, the organization of a numerical experiment requires to use two pseudo random generator program data as shown in the block diagram of Fig. 1.

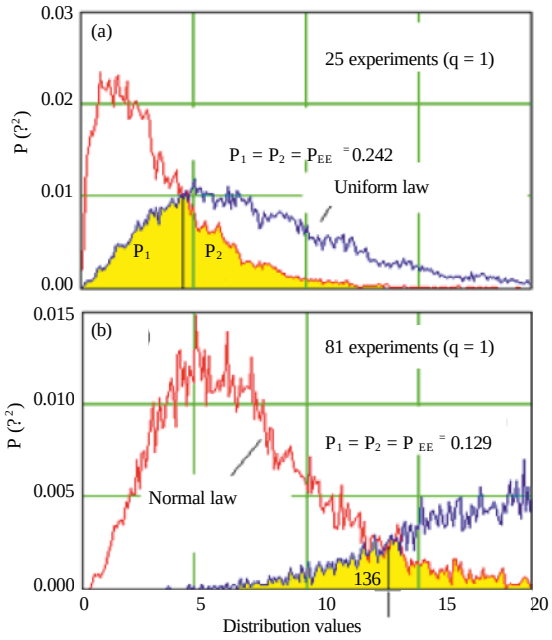

Fig. 2: Histograms of the distribution of values of one-dimensional Chi-Square criterion for testing the hypothesis of normality and uniformity hypothesis input

Each of the generators of random data G1 (normal data) and T2 (data uniformly distributed) randomly inputted into the calculator values of the Chi-Square test (1). Then the values Chi-Square test should be compared with a certain threshold quantizer. If the Chi-Square value less than the threshold, the decision about the normality of input investigated. If the value of the Chi-Square test (1) is higher than the threshold, then a decision is made about the most probable values of a uniform distribution law. Figure 2 shows plots of the histogram distribution of values of the Chi-Square test data obtained from the two program generators.

Table 1: The P_{EE} error probabilities for different thresholds and different values of the input dimension-q independent data

No. of experiments n		9	16	25	36	49	64	81	100	121
No. of columns of histogram (k)		3	4	5	6	7	8	9	10	11
Value of probabilities $P_{EE} = P_1 = P_2$; dimension of the task										
	q =1	0.420	0.320	0.280	0.220	0.160	0.140	0.130	0.120	0.090
	q =2	0.389	0.262	0.169	0.109	0.080	0.040	0.028	0.023	0.021
	q =3	0.355	0.216	0.119	0.068	0.032	0.024	0.013	0.009	0.006
	q =4	0.332	0.187	0.089	0.054	0.019	0.010	0.006	0.004	0.003
	q=5	0.304	0.154	0.061	0.027	0.012	0.006	0.004	0.002	0.001
Thresholds to ensure probabilities $P_{EE} = P_1 = P_2$; quantization thresholds										
	q =1	2.1	3.1	4.4	5.7	8.3	11.100	13.600	17.100	19.200
	q =2	2.2	3.2	4.8	6.8	9.1	11.500	14.400	17.900	20.100
	q =3	2.2	3.2	5.0	6.9	9.2	11.600	14.300	17.800	20.200
	q =4	2.1	3.2	4.9	6.9	9.2	11.500	14.500	17.800	20.100
	q =5	2.1	3.2	4.9	6.9	9.2	11.600	14.400	17.900	20.100

Numerical modeling results need further analysis which is complicated by the presence of two kinds of errors. The first kind error occurs (false rejection of true hypothesis) with probability P1 and occurs an error of the second kind (false acceptance incorrect hypothesis) with probability P2. Analyzing the probability of errors of the first and second kinds is difficult. In this regard, simplify the problem through its symmetrization and continue to consider only equal probability of errors of the first and second kinds P_{EE} = P1 = P2. Figure 2a, b marked pouring equal probabilities of errors of the first and second kind. This figure shows that by increasing the number of experiments, increases the number of columns and the histogram Type I and II errors falls equiprobable.

So, at 25 experiments uses k = 5 columns histogram that provides an equal probability of error 0.242 cm in Fig. 2. However, even at 81 experiences can be used 9 columns of the histogram that provides equal probability of occurrence of errors of the first and second kinds at the level of 0.129. Increasing the size of the test sample results in 3 times leads to reducing the likelihood of errors in two-fold. Observed nonlinear dependence of the number of tests in the test sample is growing much faster in comparison with the corresponding drop in the probability of error P_{EE}. The relationship between these two quantities is reflected in Table 1.

MATERIALS AND METHODS

Multivariate statistical analysis by adding private Chi-Square criteria: It should be noted that biometric data are multidimensional. Particularly, neural transmitter biometrics code freely available simulation environment "BioNeyroAvtograf", converts 416 of biometric parameters in the code of private key length of 256 bits. That is, there is an opportunity to analyze not just one but 416 biometric parameters. If we have a sample of 16 examples, then there is an opportunity to analyze 16×416 = 6656 samples. There is a real opportunity to

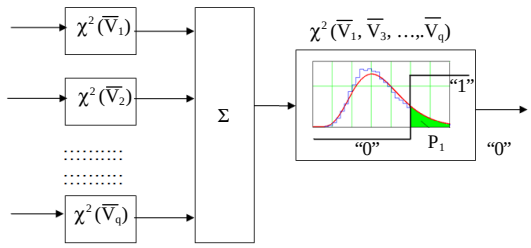

Fig. 3: Multivariate statistical processing of data by network of Pearson

increase the amount of data to be processed and thus raise the reliability of decisions (Serikova *et al*., 2015). For the multi-dimensional processing use adding the private criteria of Pearson:

$$\chi^2(\nu_1, \nu_2, \ldots, \nu_q) = \frac{\chi^2(\nu_1) + \chi^2(\nu_2) + \ldots + \chi^2(\nu_q))}{q} \quad (3)$$

Transformation Eq. 3 is equivalent to the use of the private criteria of Pearson, Pearson network structure is shown in Fig. 3. The network of private criteria of Chi-Square Pearson has input and output nonlinear transformations and linear summation of data between them. In fact, the conversion is carried out in accordance with the multi-dimensional model of Hammerstein-Wiener (Billings, 1980; Ivanov, 2002) which is formally a neuron with input nonlinear transformations of Pearson.

In the transition to the modeling of a network of private Chi-Squared criteria of Pearson it is enough instead of two software random number generators use q steam generators. When low-dimensional input data q≤16 special difficulties in programming numerical experiment does not arise. Table 1 shows the values of equal probabilities of errors of the first and second kind for

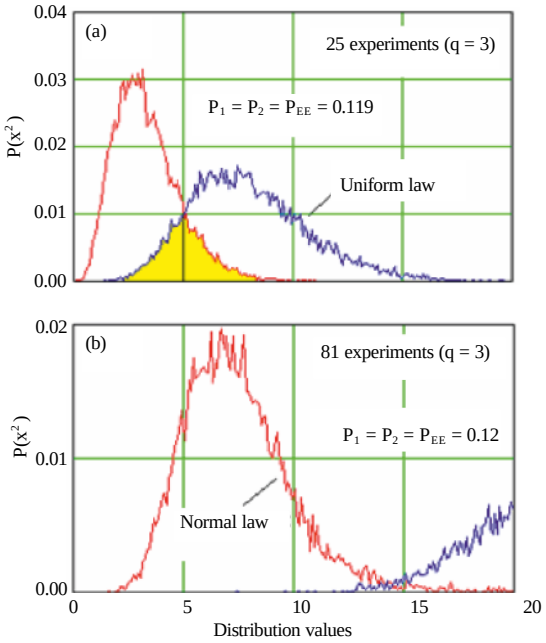

Fig. 4: Distribution of the output data of three dimensional network of Pearson for 25 and 81 runs

different values of the output of the quantizer thresholds Pearson networks as well as for different input dimension.

Note that the threshold value is equal to the probability of errors is almost the same for all indicators dimension Table 1. This is an extremely interesting fact that shows a significant simplification of the problem because of a correctly symmetrization. An illustration of this situation is Fig. 4 which shows the distribution of distances at the output of three-dimensional network of Pearson for the normal and uniform laws of the distribution of values.

If we compare Fig. 2 and 4, it is easy to identify the effect of linear separability of growth, considered the distribution of values as the number of experiments in the training set and the growth of a network of private dimension of Pearson. This means that by increasing the dimension of the statistical processing, we can substantially reduce the requirements to the dimensions of the training sample. Thus, when a one-dimensional treatment to receive PEE = 0.1 requires to use a test sample of 112 experiments. If using a two-dimensional statistical processing of data, for the same probability of error $P_{EE} = 0.1$ requires a sample of 41 experience. There is almost two-fold reduction in the requirements for the size of the training sample.

It should be noted that the data in Table 1 were obtained for the number of independent (uncorrelated state program generators). Actual biometric data is always dependent (Ivanov, 2002; Akhmetov *et al*., 2014b, 2012) which leads to a decrease in gain from the transition to multivariate statistical criteria. However, the gain of the multi-dimensional data is always present and is significant.

Analytical description of the Chi-Squared distributions for finite samples when testing the hypothesis of the normal distribution of one-dimensional values: An important property of the Chi-Squared distributions is that they have a precise analytical description, not only for an infinite amount of samples $n = \infty$. The results of numerical experiments have shown that for n = 9, 16, 25, 36, 49 density of Chi-Square distribution is described by Pearson gamma function with integer exponents number of degrees of freedom. In particular, the final sample of 16 experiments (four columns of the histogram), the density distribution is described by the following Eq. 4:

$$p_{\chi^2}(q=1, n=16, \quad m=3, x) = \frac{1}{2\times 2^{\frac{3}{2}}\times\Gamma\left(\frac{3}{2}\right)}\times(2x)^{\frac{3}{2}-1}\times e^{\frac{-2x}{2}} \tag{4}$$

For two-dimensional Chi-Square criteria Pearson density distribution is described by the same Eq. 4:

$$p_{\chi^2}(q=2, n=16, \quad m=5, x) = \frac{1}{3\times 2^{\frac{5}{2}}\times\Gamma\left(\frac{5}{2}\right)}\times(3x)^{\frac{5}{2}-1}\times e^{\frac{-3x}{2}} \tag{5}$$

Increasing the dimension of the input data, induction can obtain the following description of Chi-Square criteria distribution for an arbitrary value q:

$$p_{\chi^2}(q, n=16, x) = \frac{1}{(q+2)2^{\frac{(2q+2)}{2}}\times\Gamma\left(\frac{2q+2}{2}\right)}\times ((q+2)x)^{\frac{2q+2}{2}-1}\times e^{\frac{-(q+2)x}{2}} \tag{6}$$

It turns out that it can be used with analytical relations of the form Eq. 6 in order to build a table of quantile with confidence probability for multivariate Chi-Square criteria distribution of Pearson any dimension q. At least, this may be done for the final sample of 16 experiments. Presumably, that similar analytical expressions can be obtained for other test

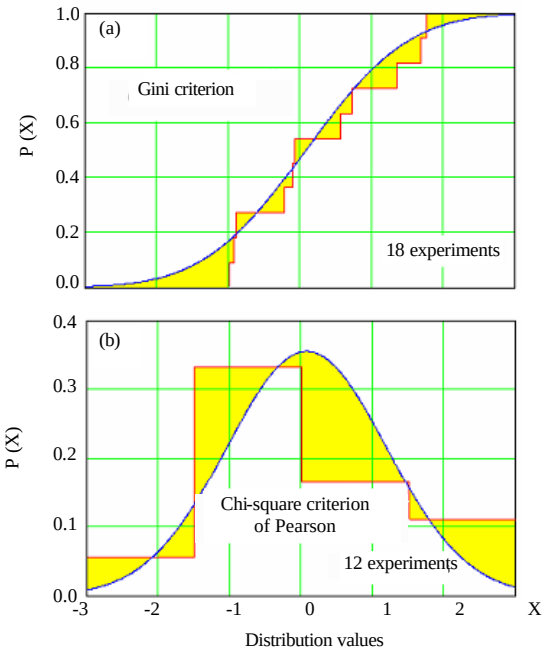

Fig. 5: The ratio of the quantization error at the approach of the probability distribution function and the probability density function for the same number of experiments

samples with the number of experiments coincides exactly with the square of the number of columns of the histogram. In all other cases, the Chi-Square distribution can not be accurately described by integer figures number of degrees of freedom. To describe them should be used fractional (fractal) figures the number of degrees of freedom.

Effect on Chi-Square power criterion quantization errors due to the small number of experiments in a test sample: One of the fundamental problems in statistics is that when assessing species distributions and their parameters, it is necessary to replace a seamless continuum of possible values of the data selecting from a finite number of examples. In particular, the hypothesis is verified if a normal distribution, the final volume of the test sample necessarily gives rise to sampling errors due to the finite number of experiments in a test sample. This situation is illustrated in Fig. 5 where the two approximations are constructed for a sample of 12 experiments. Figure 5a gives the approximation of the probability function used Gini criterion and Fig. 5b is given an approximation of the probability density used Chi-Square of Pearson.

Errors occur due to quantization in Fig. 5, marked with pouring. Comparing the Fig. 5a and b part, it is easy to verify that when approaching the density distribution of values p(x) quantization error is approximately three times more than the same error that occurs when approaching the probability function P(x). This happens due to the fact that the number of columns of the histogram approximating the density of the distribution of values is much smaller than the number of experiments in the test sample.

As a consequence, the Chi-Square criterion of Pearson and other statistical criteria based on a comparison of the theoretical density histogram with distribution of values turn out in the worst position with other criteria by comparing the probability functions. It is obvious that enhancing the power of statistical tests can be achieved by smoothing the sharp edges of step approximations of functions shown in Fig. 5. Particularly, can be built digital linear filter without phase distortion which will significantly increase the capacity of Chi-Square criterion by increasing the number of degrees of freedom (Akhmetov *et al*., 2014a).

The data smoothing by linear averaging filter: In the construction of the classical histogram the quantization is performed by discovered dynamic range of the experimental data on several intervals. It is usually assumed that in each interval histogram should get some experimental data, only in this case, the histogram will be similar to a controlled density distribution of the values of the investigated data.

For definiteness, use data received from generator of 64 normal random numbers. After that, calculate the standard deviation of the received sample and quantize its dynamic range of $E(x)\pm4\sigma(x)$ for 100 quanta. Such a choice of dynamic range ensures that the studied small samples will almost always fall within this dynamic range.

Next, use the first 8 samples alternately the next 12 samples and then other samples 16, 24, 32, 48, 64 created by a reference sample. Such a method of processing use allows you to monitor the impact of the volume of raw data on the reliability of the test sample to obtain statistical estimates in testing the hypothesis of the normal distribution of values.

Obviously, the introduction of substantial redundancy quantization intervals will give rise to a large number of pseudo histogram empty slots that shown in Fig. 6.

As seen in Fig. 6 pseudo histogram with multiple times reduced quantization step has blank spaces, even when using a sample of 64 experiments, as in this case, the number of micro intervals exceeds the number of experiments. In all cases, the pseudo histogram in shape is very different from the density of the normal distribution of values.

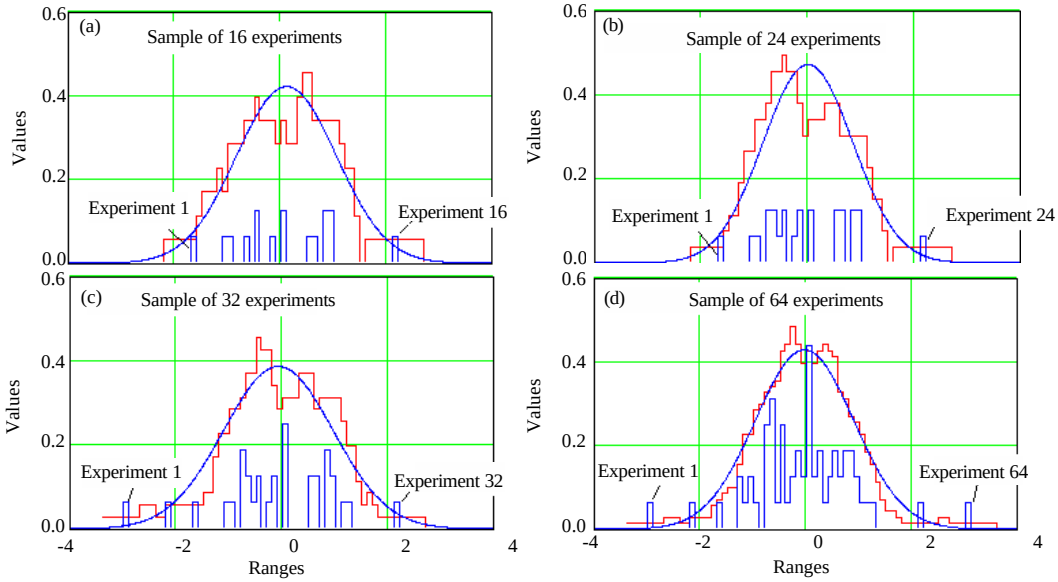

Fig. 6: Smoothing data by digital filter with bandwidth 11 samples by increasing monotonically with a test set of 16, 24, 32, 64 experiments

To recover the lost form of density distribution of values examined data by smoothing data of pseudo histogram by averaging filter, for example, a sliding window 11 samples wide. Used digital filter with zero phase shift, provided that the average for 11 samples of output is placed in the center of the viewing window. Figure 6 shows that after smoothing with this kind of filter the shape of the normal distribution of one-dimensional values is well restored. It should be noted that the number of columns of pseudo histogram with small intervals for n = 9, k = 3 is increased to a value:

$$\tilde{k}_1 = 10(k+2)+2 \qquad (7)$$

However, the first 10 microintervals will always be zero. Zero as well turn out the last 10 microintervals. This situation is illustrated Table 2. During smoothing the data the first 6 microintervals and the last 6 microintervals always stay zero. This means that the total number of intervals in histogram after smoothing will be described by the following relation:

$$\tilde{k}_2 = 10(k+2)-2\times 6 \qquad (8)$$

In particular for n = 9, k = 3 (Eq. 3 and 6) gives a value of 37 that is supported by data on the right side of Table 2. That is increasing the number of columns of histogram will be described by the following relation:

Table 2: An example of data arising from the digital smoothing with a ten-fold decrease in the quantization interval and the use of averaging filter with a window of 11 samples

Hist (intn, xn)		Hist (intn, xn)		C gn		C gn	
1	2	1	2	1	2	1	2
-	0	-	0	-	0	-	0.000
0	0	39	1	0	0	34	0.091
1	0	40	0	1	0	35	0.364
2	0	41	0	2	0	36	0.182
3	0	42	0	3	0	37	0.182
4	0	43	0	4	0	38	0.182
5	0	44	0	5	0	39	0.364
6	0	45	0	6	0.091	40	0.091
7	0	46	0	7	0.091	41	0.091
8	0	47	0	8	0.091	42	0.091
9	0	48	0	9	0.091	43	0.091
10	1	49	0	10	0.273	44	0.000
11	0	50	0	11	0.091	45	0.000

$$\frac{\tilde{k}_2}{n} = \frac{10(k+2)-2\times 6}{n} = \frac{10(\sqrt{n}+2)-2\times 6}{n} \qquad (9)$$

Precisely because of the effect of increasing the number of columns in the smoothed histogram is more accurate approximation of the original continuous density distribution of p(x) by its discrete (step) analog of.

RESULTS AND DISCUSSION

The variety of established and studied statistical criteria: Above in the text basically, it was a Chi-Square of Pearson

Table 3: The known statistical criteria

Criterion name and the year of creation	Formula
Chi-Square criterion or Pearson criterion 1900	$\int_{-\infty}^{+\infty} \frac{\{p(x)-\tilde{p}(x)\}^2}{\tilde{p}(x)}dx$
Criterion Cramer-Von Mises 1928	$\int_{-\infty}^{+\infty} \{P(x)-\tilde{P}(x)\}^2 dx$
Criterion of Kolmogorov-Smirnov 1933	$\sup_{-\infty<x<+\infty} \lvert P(x)-\tilde{P}(x)\rvert$
Criterion of Smirnov-Kramers Von Mises 1936	$\int_{-\infty}^{+\infty} \{P(x)-\tilde{P}(x)\}^2 d\tilde{P}(x)$
Criterion of Gini 1941	$\int_{-\infty}^{+\infty} \lvert P(x)-\tilde{P}(x)\rvert dx$
Criterion of Anderson-Darling 1952	$\int_{-\infty}^{+\infty} \frac{\{P(x)-\tilde{P}(x)\}^2}{\tilde{P}(x)\{1-\tilde{P}(x)\}} d\tilde{P}(x)$
Criterion of Cooper 1960	$\sup_{-\infty<x<+\infty} \{P(x)-\tilde{P}(x)\} + \sup_{-\infty<x<+\infty} \{\tilde{P}(x)-P(x)\}$
Criterion of Watson 1961	$\int_{-\infty}^{+\infty} \left\{\tilde{P}(x)-P(x)-\int_{-\infty}^{x}[\tilde{P}(x)-P(x)]d\tilde{P}(x)\right\} d\tilde{P}(x)$
Criterion of Frotsini 1978	$\int_{-\infty}^{+\infty} \lvert P(x)-\tilde{P}(x)\rvert d\tilde{P}(x)$
Differential version of Gini criterion 2006 (Akmetov *et al.*, 2012)	$\int_{-\infty}^{+\infty} \lvert p(x)-\tilde{p}(x)\rvert dx$

Table 4: The values equally probable errors for different volumes of test samples, using several statistical tests

	The number of tests in the test sample								
Differnt criterion	9	16	25	36	49	64	81	100	121
Values of equally probable errors $P_1 = P_2 = P_{EE}$									
Criterion of Cooper 1960	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.500	0.500
Criterion of Gini 1941	0.5	0.497	0.482	0.417	0.348	0.269	0.225	0.205	0.186
Criterion of Kolmogorov-Smirnov 1933	0.46	0.44	0.345	0.315	0.239	0.232	0.215	0.201	0.177
Criterion of Frotsini 1978	0.439	0.38	0.325	0.268	0.212	0.172	0.154	0.107	0.089
Chi-Squared Pearson criterion 1900	0.42	0.32	0.28	0.22	0.16	0.14	0.13	0.12	0.090
Criterion of Cramer-Von Mises 1928	0.356	0.306	0.24	0.215	0.155	0.121	0.102	0.082	0.061
Differential version of Gini criterion 2006	0.281	0.202	0.162	0.101	0.07	0.05	0.03	0.02	0.010

as the most sought after and most widespread in practice. This criterion has appeared before the others and by far the most studied. However, in parallel with the study of Chi-Square of Pearson for more than a century mathematical thought has also created other criteria (Kobzar, 2006). The names of most of these criteria, the time of their creation and calculating formula shown in Table 3.

From Table 3 seen that the statistical criteria set up gradually. The most recent was a differential criterion Gini (Malygin *et al*., 2006) specifically for the processing of biometric data. This criterion was the most powerful and constructed by replacing the original criteria Gini (1941) (Kobzar, 2006), the probability function on their derivatives (density distribution of probabilities). Each of the criteria of Table 3 has different power when testing the hypothesis of normality of the distribution of values for the alternative hypothesis of the second law of uniform distribution of values.

From Table 4 seen that the Chi-Square criterion is not the best. There are criteria with significantly more power, offer much less important error probabilities of the first and second kinds in recognition of two statistical hypotheses.

Statistical criteria supplementing already known criteria: According to the literature cited in (Kobzar, 2006), the creation and study of statistical criteria continues, starting with the creation of the first Pearson Chi-Square test in 1900. In this case, all the currently known criteria established as heuristics. That is a system of criteria is not exhaustive. Therefore, try to organize the known criteria and assess the completeness of previous studies.

We note that the criteria Gini Frotsini and Watson (lines 5, 8, 9, Table 3) constructed using modules difference observed function of probability and

Table 5: Statistical integral criteria, using squares of the differences of the probability function

Integration over	Integration over
$\int_{-\infty}^{+\infty} \lvert P(x) - \tilde{P}(x) \rvert dx$	$\int_{-\infty}^{+\infty} \lvert P(x) - \tilde{P}(x) \rvert d\tilde{P}(x)$
Criterion of Gini in 1941y	Criterion of Frotsini in 1941y
$\int_{-\infty}^{+\infty} \frac{\lvert P(x) - \tilde{P}(x) \rvert}{\tilde{P}(x)} dx$	$\int_{-\infty}^{+\infty} \frac{\lvert P(x) - \tilde{P}(x) \rvert}{\tilde{P}(x)} d\tilde{P}(x)$
$\int_{-\infty}^{+\infty} \frac{\lvert P(x) - \tilde{P}(x) \rvert}{(1 - \tilde{P}(x))} dx$	$\int_{-\infty}^{+\infty} \frac{\lvert P(x) - \tilde{P}(x) \rvert}{(1 - \tilde{P}(x))} d\tilde{P}(x)$
$\int_{-\infty}^{+\infty} \frac{\lvert P(x) - \tilde{P}(x) \rvert}{\sqrt{\tilde{P}(x)(1 - \tilde{P}(x))}} dx$	$\int_{-\infty}^{+\infty} \frac{\lvert P(x) - \tilde{P}(x) \rvert}{\sqrt{\tilde{P}(x) \cdot (1 - \tilde{P}(x))}} d\tilde{P}(x)$
$\int_{-\infty}^{+\infty} \left\{ \tilde{P}(x) - P(x) - \int_{-\infty}^{x} \left[\tilde{P}(x) - P(x) \right] dx \right\} dx$	$\int_{-\infty}^{+\infty} \left\{ \tilde{P}(x) - P(x) - \int_{-\infty}^{x} \begin{bmatrix} \tilde{P}(x) - \\ P(x) \end{bmatrix} d\tilde{P}(x) \right\} d\tilde{P}(x)$
	Criterion of Watson in 1961y

Table 6: Statistical integral criteria, using squares of the differences of the probability function according to new criteria

Integration over	Integration over
$\int_{-\infty}^{+\infty} \left\{ P(x) - \tilde{P}(x) \right\}^2 dx$	$\int_{-\infty}^{+\infty} \left\{ P(x) - \tilde{P}(x) \right\}^2 d\tilde{P}(x)$
Criterion of Cramer-von Mises in 1928y	Criterion of Smirnov-Cramer-von Mises in 1936y
$\int_{-\infty}^{+\infty} \frac{\left\{ P(x) - \tilde{P}(x) \right\}^2}{\tilde{P}(x)} dx$	$\int_{-\infty}^{+\infty} \frac{\left\{ P(x) - \tilde{P}(x) \right\}^2}{\tilde{P}(x)} d\tilde{P}(x)$
$\int_{-\infty}^{+\infty} \frac{\left\{ P(x) - \tilde{P}(x) \right\}^2}{(1 - \tilde{P}(x))} dx$	$\int_{-\infty}^{+\infty} \frac{\left\{ P(x) - \tilde{P}(x) \right\}^2}{(1 - \tilde{P}(x))} d\tilde{P}(x)$
$\int_{-\infty}^{+\infty} \frac{\left\{ P(x) - \tilde{P}(x) \right\}^2}{\tilde{P}(x) \cdot (1 - \tilde{P}(x))} dx$	$\int_{-\infty}^{+\infty} \frac{\left\{ P(x) - \tilde{P}(x) \right\}^2}{\tilde{P}(x) \cdot \left\{ 1 - \tilde{P}(x) \right\}} d\tilde{P}(x)$
$\int_{-\infty}^{+\infty} \left\{ \left\{ \tilde{P}(x) - P(x) \right\}^2 - \int_{-\infty}^{x} \left\{ \tilde{P}(x) - P(x) \right\}^2 dx \right\} dx$	$\int_{-\infty}^{+\infty} \left\{ \left\{ \tilde{P}(x) - P(x) \right\}^2 - \int_{-\infty}^{x} \begin{Bmatrix} \tilde{P}(x) - \\ P(x) \end{Bmatrix}^2 d\tilde{P}(x) \right\} d\tilde{P}(x)$
	Criterion of Anderson-Darling in 1952y

Table 7: Statistical integral criteria, using the difference in the probability density function

Modules of difference	Squares of differences
$\int_{-\infty}^{+\infty} \frac{\lvert p(x) - \tilde{p}(x) \rvert}{\tilde{p}(x)} dx$	$\int_{-\infty}^{+\infty} \frac{\left\{ p(x) - \tilde{p}(x) \right\}^2}{p(x)} dx$
	Chi-Squared Pearson criterion in 1900y
$\int_{-\infty}^{+\infty} \lvert p(x) - \tilde{p}(x) \rvert dx$	$\int_{-\infty}^{+\infty} \left\{ p(x) - \tilde{p}(x) \right\}^2 dx$
Differential criterion of Gini in 2006y	
$\int_{-\infty}^{+\infty} \lvert p(x) - \tilde{p}(x) \rvert \tilde{p}(x) dx$	$\int_{-\infty}^{+\infty} \left\{ p(x) - \tilde{p}(x) \right\}^2 \times \tilde{p}(x) dx$
$\int_{-\infty}^{+\infty} \lvert p(x) - \tilde{p}(x) \rvert \tilde{p}^2(x) dx$	$\int_{-\infty}^{+\infty} \left\{ p(x) - \tilde{p}(x) \right\}^2 \tilde{p}^2(x) dx$
$\int_{-\infty}^{+\infty} \lvert p(x) - \tilde{p}(x) \rvert \tilde{p}^k(x) dx$	$\int_{-\infty}^{+\infty} \left\{ p(x) - \tilde{p}(x) \right\}^2 \tilde{p}^k(x) dx$

Table 8: Statistical criteria constructed on registered amplitudes approximation error

Probability function	The probability density functions
$\sup_{-\infty<x<+\infty} \lvert P(x) - \tilde{P}(x) \rvert$	$\sup_{-\infty<x<+\infty} \lvert p(x) - \tilde{p}(x) \rvert$
Criterion of Kolmogorov-Smirnov in 1933y	
$\sup_{-\infty<x<+\infty} \left\{ P(x) - \tilde{P}(x) \right\} +$	$\sup_{-\infty<x<+\infty} \left\{ p(x) - \tilde{p}(x) \right\} +$
$\sup_{-\infty<x<+\infty} \left\{ \tilde{P}(x) - P(x) \right\}$	$\sup_{-\infty<x<+\infty} \left\{ \tilde{p}(x) - p(x) \right\}$
Criterion of Cooper in 1960y	
$\sup_{-\infty<x<+\infty} \left\{ P(x) - \tilde{P}(x) \right\}^2$	$\sup_{-\infty<x<+\infty} \left\{ p(x) - \tilde{p}(x) \right\}^2$
$\sup_{-\infty<x<+\infty} \left\{ P(x) - \tilde{P}(x) \right\}^2 +$	$\sup_{-\infty<x<+\infty} \left\{ p(x) - \tilde{p}(x) \right\}^2 +$
$\sup_{-\infty<x<+\infty} \left\{ \tilde{P}(x) - P(x) \right\}^2$	$\sup_{-\infty<x<+\infty} \left\{ \tilde{p}(x) - p(x) \right\}^2$
$\sup_{-\infty<x<+\infty} \left\{ P(x) - \tilde{P}(x) \right\}^4$	$\sup_{-\infty<x<+\infty} \left\{ p(x) - \tilde{p}(x) \right\}^4$

theoretical probability function corresponding to test hypotheses. All these criteria are summarized in Table 4 and divided by type of the variable of integration in different variants of weighing modules difference. As a result, we get 7 possible options but previously unknown, criteria for statistical hypothesis testing (Table 5). A similar procedure of generalization was performed to criteria:

- Cramer-von Mises 1928
- Smirnov, Cramer-Von Mises 1936
- Anderson-Darling 1952
- The result of this synthesis is still unknown seven statistical criteria previously arranged in Table 6

Another option of creation the criteria is to use the difference observed density distribution value as it approaches the theoretical density and histogram-corresponding to test hypotheses. In this case, we get a number of new criteria which are given in Table 7.

A special place is occupied by statistical criteria based on the analysis of the upper limit of divergence probability functions and the corresponding density distribution of values. Known and new statistical criteria of this class are placed in the Table 8.

Thus, a variety of known statistical criteria can be significantly enhanced with new, as yet unexplored with mathematical constructs. At the same time, expecting from such studies a significant increase in power of the criteria is not necessary. Presumably that the above options for new, not yet investigated criteria that can significantly improve the situation (slightly reduce the volume requirements of the test sample).

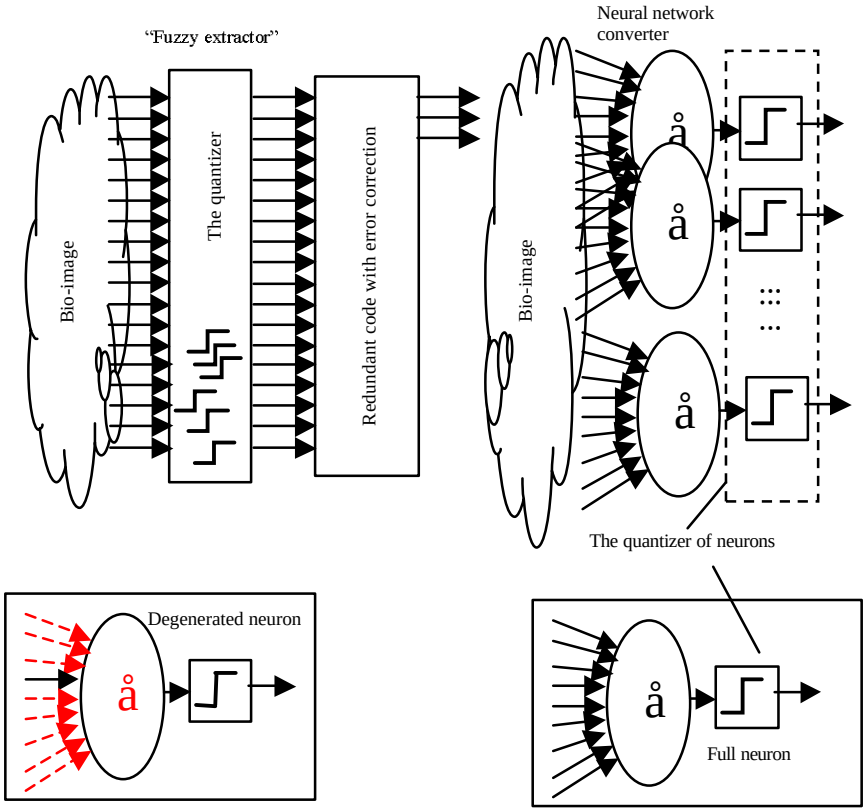

Fig. 7: Scheme of two types of converters of biometrics code "fuzzy extractors" (left panel) and a complete neural network (right panel)

Formation of the generalized one-dimensional statistical criterion that combines several particular criteria: The fact that many statistical criteria can be used to combine them into a generalized criterion. Thus, generalization can be built in two ways. The first way is to use so-called "fuzzy extractors" the second way is to use high-grade artificial neural networks. Both of these technological areas are well explored in the development of converters biometrics code. Block diagrams explaining operation of these technology, generalizations of partial results are shown in Fig. 7.

It should be emphasized that the "fuzzy extractors" for biometric applications are mainly developed by the English-speaking professionals. Work on the creation of this technology started in the late last century and is actively developed in the zero years of this century (Monrose *et al*., 2001; Juels and Sudan, 2002; Verbitskiy *et al*., 2003; Dodis *et al*., 2004; Yang and Verbauwhede, 2005; Ramirez-Ruiz *et al*., 2006; Cauchie *et al*., 2006; Arakala *et al*., 2007; Lee *et al*., 2007; Nandakumar *et al*., 2007; Balakirsky *et al*., 2009). In Russian language researches of "fuzzy extractors" a lot less (Chmorra, 2011; Ushmaev and Kuznetsov, 2012), it is due to the fact that in Russia, in parallel with the "fuzzy extractors" developed another direction, exploring the use of large and very large artificial neural networks (Ivanov, 2000, 2004; Volchikhin *et al*., 2005; Yazov, *et al*., 2012) for converting biometrics in personal cryptographic key code.

"Fuzzy extractors" are based on the fact that each biometric parameter is compared with a certain threshold (quantized), eventually yielding two possible states "0" or "1". If a biometric image is 416 biometrics. Ie a biometric image of a handwritten word among the "BioNeyroAvtograf" could be transformed into a 416-bit output code. Thus up to 30% discharges of BioKOD are incorrect because of the natural instability of the biometric image. All the "fuzzy extractors" are built on the fact that, they correct errors of BioKOD with any known classical redundant self-correcting codes (Morelos-Zaragoza, 2007). Building efficient code that can correct 30% of errors today technically impossible. In this regard, "fuzzy extractors" disguise approximately 20% of the most volatile biometric parameters, remaining 10% are detecting and correcting.

Relation to the subject of this article "fuzzy extractors" can be used to summarize several statistical

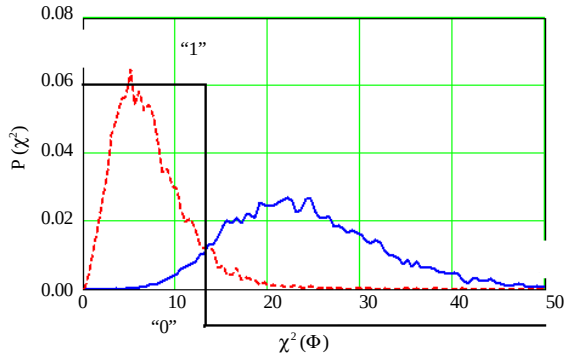

Fig. 8: Allocation of data from a normal distribution of values (dashed line) when checking the first hypothesis

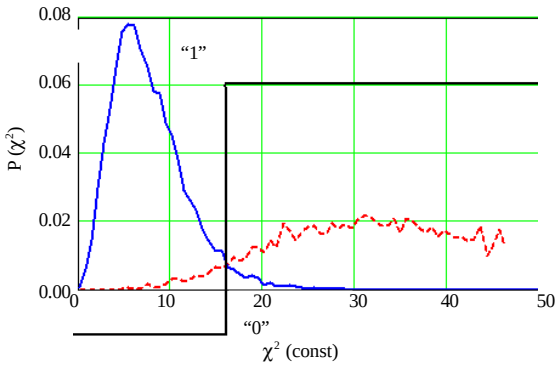

Fig. 9: Allocation of data with a normal distribution of values (dashed line) when checking the second hypothesis

criteria, each statistical criterion should be used twice. For example, the classic Chi-Square criterion of Pearson (1) can be used to estimate the closeness of the data to the hypothesis of the normal distribution:

$$\chi^2(\Phi) = n\sum_{i=1}^{k}\frac{\left(\frac{b_i}{n} - \frac{1}{\sigma(x)\sqrt{2\pi}}\int_{x_i}^{x_{i+1}}\exp\left\{\frac{-(E(x)-u)^2}{2(\sigma(x))^2}\right\}du\right)^2}{\frac{1}{\sigma(x)\sqrt{2\pi}}\int_{x_i}^{x_{i+1}}\exp\left\{\frac{-(E(x)-u)^2}{2(\sigma(x))^2}\right\}du} \quad (10)$$

Where the limits of integration x_1, x_2,...,x_n is boundaries of uniform intervals on which builds the histogram of frequencies of occurrence data in a test sample consisting of n experiments.

Figure 8 shows the curves of the histogram distribution of values of Chi-Square criterion for the data obtained from the two generator program (Fig. 1), giving 81 count sampling (k-9).

Figure 8 shows that the comparator makes a decision on an input of normal sequence should give state of "1" in the range of 0-14. The switching threshold of the comparator in state "0"-14. In this case, the probability of errors of the first and second kind are the same $P_1 = P_2 = P_{EE} = 0.054$. Applied to verification of the hypothesis of uniform distribution law values of Pearson criterion (1) will have Eq. 11:

$$\chi^2(\text{const}) = n\cdot\sum_{i=1}^{k}\frac{\left(\frac{b_i}{n} - \frac{1}{k}\right)^2}{\frac{1}{k}} \quad (11)$$

Figure 9 shows curves of distribution of histograms of the Chi-Square values for criterion data obtained from the two generator program (Fig. 1) giving the reference sample of 81 (k = 9).

Figure 9 shows that the comparator makes a decision on detecting an input of normal sequence should give state of "1" in the range of 17 and above. The switching threshold of the comparator in the state "0"-16. In this case, the probability of errors of the first and second kind are the same.

The result is that all the statistical criteria that are in Table 4 below criterion line Chi-Square will give the codes with 5.4% error. If you use a self-correcting code with 100% redundancy (6% correcting errors) then editing the error can be even for a single criterion for the code of 2 bits. That is the union of several statistical criteria "fuzzy extractors" giving 4, 6, 8, 10 or more bits is quite real.

Conducted in the Russia and Kazakhstan, studies have shown that, at least, converters of biometry code using "fuzzy extractors" are inferior to their main characteristics of neural network drives. In this regard, the transition to the neural network generalization of several statistical criteria should allow to obtain better results. Thus, when such generalizations need to carry out training of artificial neural networks. For this purpose can be used the first standardized learning algorithm or any other of the known learning algorithms (Wasserman, 2006).

CONCLUSION

People are able to learn and test the quality of their training on a small number of examples. How is this done and which mathematics is the basis of our abilities is still unknown. Nevertheless, we can confidently assert

that created in the last century one-criterial and one-dimensional mathematical statistics, has significant reserves of growth of its effectiveness. It is necessary to organize the research in the areas of research outlined above and their combinations.

It is interesting to note that mathematical thought evolves cyclically. Authoritative magazine of mathematical statistics, published in 1901 in Oxford, not by chance is called "Biometrika". Biometric data at the beginning of the last century needed statistical processing, at the same time took a special position and gave rise to many problems.

In the early 21st century, observed the first active development of biometric technologies and the emergence of several new biometric magazines. The main idea of this article is that the ideas generated at the development of biometric technologies in the 21st century are very, very conformable with the classical postulates of mathematical statistics. Moreover, they are likely to be useful for the next round of development of the mathematical statistics (multidimensional and multicriteria).

REFERENCES

Anonymous, 2002. Applied Statistics. Validation rules experienced distribution agreement with the theoretical P 50.1.037, Part II. Nonparametric tests. State Standard of Russia. Moscow, pp: 67.

Aguirre, N. and M. Nikulin, 1994. Chi-Squared goodness-of-fit test for the family of logistic distributions. Kybernetika, 30: 214-222.

Akhmetov, B.S., A.I. Ivanov, V.A. Funtikov, A.V. Bezyaev and E.A. Malygina, 2014c. The technology of using large neural networks for fuzzy transformation of biometric data in the key code access: Monograph. Publisher LEM. Almaty, pp: 144.

Akhmetov, B.S., A.A. Doszhanova, A.I. Ivanov, T.S. Kartbayev and A.Yu. Malygin, 2013. Biometric Technology in Securing the Internet Using Large Neural Network Technology. World Acad. Sci. Eng. Technol., Singapore, pp: 129-138.

Akhmetov, B.S., D.N. Nadeev, V.A. Funtikov, A.I. Ivanov, A.Yu. Malygin, 2014b. Risk assessment of highly reliable biometrics. Monograph. Publisher LEM. Almaty, pp: 108.

Akhmetov, B.S., A.I. Ivanov, V.A. Funtikov and I.V. Urnev, 2012. Evaluation of multidimensional entropy on short strings of biometric codes with dependent bits. PIERS Proceedings, Moscow, Russia, pp: 66-69.

Akhmetov, B.S, A.I. Ivanov, N.I. Serikova, Yu.V. Funtikova, 2014a. The algorithm is an artificial increase in the number of degrees of freedom in the analysis of biometric data by the criterion of consent Chi-Square test. Bulletin of National Academy of Sciences of the Republic of Kazakhstan, 5: 28-34.

Arakala, A. and J. Jeffers, K.J. Horadam, 2007. Fuzzy extractors for minutiae-based fingerprint authentication. Adv. Biomet., (LNCS 4642), Springer, pp: 760-769.

Aroian, L.A., 1973. A new approximation to the levels of significance of the Chi-Square distribution. Ann. Math. Stat., 14: 93-95.

Bolle, R.M., J.H. Connell, S. Pankanti, N.K. Ratha and A.W. Senior, 2007. Guide to Biometrics. Moscow, Technosphere, pp: 368.

Billings, S.A., 1980. Identification of nonlinear system (A survey). Proc. IEEE, part D, 127: 272-285.

Chmorra, A.L., 2011. Masking key using biometrics problems of information transmission, 2: 128-143.

Cochran, W.G., 1954. Some methods of strengthening the common 2 tests. Biometrika, 10: 417.

Cadwell, J.H., 1952. Approximating to the distributions of measures of dispersion by a power of 2. Biom., 40: 336-346.

Cauchie, S. and T. Brouard, H. Cardot, 2006. From features extraction to strong security in mobile environment: A new hybrid system. On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, Springer, pp: 89-498.

Dodis, Y., L. Reyzin and A. Smith, 2004. Fuzzy extractors: How to generate strong keys from biometrics and other noisy. In EuroCrypt, pp: 523-540

Gilbert, R.J., 1977. A sample formula for cuterpolating tables of 2. Biom., 33: 383-385.

Gjlberg, H. and H. Leuine, 1945. Approximate formulas for the percentage points and normalization of t and 2. Ann. Math. Stat., 17: 216-225.

Ivanov, A.I., 2002. Neural network technology of biometric authentication of users of open systems. Abstract for the degree of doctor of technical sciences, specialty. Penza State University. Penza, pp: 34.

Ivanov, A.I., 2000. Biometric identification on the dynamics of unconscious movements: Monograph. Penza State University, Penza, pp: 178.

Ivanov, A.I., 2004. Neural network algorithms for biometric identification. The15 book series "Neurocomputers and their application", Radiotechnika, pp: 144.

Juels, A. and M. Sudan, 2002. A Fuzzy Vault Scheme. IEEE International Symposium on Information Theory.

Kobzar, A.I., 2006. Applied Mathematical Statistics. For engineers and scientists'. FIZMATLIT, pp: 816.

Lemeshko, B.Y. and S.N. Postovalov, 1998. On the dependence of the limiting distributions of statistics 2 Pearson and likelihood ratio of the method of grouping data. Factory Laboratory, 64: 56-63.

Lee, Y.J., K. Bae, S.J. Lee, K.R. Park and J. Kim, 2007. Biometric key binding: Fuzzy vault based on Iris images. Proceedings of 2nd Inter. Conf. Biom., Seoul, South Korea, pp: 800-808.

Morelos-Zaragoza, R., 2007. Art error-correcting coding. M.: Technosphere, pp: 320.

Malygin, A.Yu., V.I. Volchikhin, A.I. Ivanov and V.A. Funtikov, 2006. Fast algorithms for testing neural network mechanisms biometrics, cryptographic protection of information. Penza State University, Penza, pp: 161.

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

Mirvaliev, M. and M.S. Nikulin, 1992. 'Fit of Chi-Square type'. Factory Laboratory, 58: 52-58.

Mann, N.V. and A. Wald, 1942. On the choice of the number of class intervals in the application of the chi square test. Ann. Math. Stat., 13: 306-317.

Nandakumar, K., A.K. Jain, S. Pankanti, 2007. Fingerprint-Based fuzzy vault: Implementation and performance. IEEE transactions on information forensics and security, 2: 744-757.

Pearson, E.S., 1959. Note on an approximation to the distribution of non-central 2. Biom., 46: 364-366.

Ramirez-Ruiz, J., C. Pfeiffer, J. Nolazco-Flores, 2006. Cryptographic keys generation using finger codes. Adv. Artif. Intell., IBERAMIA-SBIA, (LNCS 4140), pp: 178-187.

Serikova, N.I., V.I. Volchikhin, A.I. Ivanov, N.I. Serikova and Y. Funtikova, 2015. The effect of reducing the size of the test sample by switching to multivariate statistical analysis of biometric data. Proceedings of the higher educational institutions. Tech. Sci., Penza State University, Penza, 1: 86-91.

Ushmaev, O.V., V.V. Kuznetsov, 2012. Algorithms secure verification based on the binary representation of the topology of a fingerprint. Inf. Appl., 6: 132-140.

Volchikhin, V.I., A.I. Ivanov and V.A. Funtikov, 2005. Fast algorithms for training neural network mechanisms biometrics, cryptographic protection of information: monograph. Penza State University, Penza, pp: 273.

Verbitskiy, E., P. Tuyls, D. Denteneer and J.P. Linnartz, 2003. Reliable Biometric authentication with privacy protection. In Proc. 24th Benelux Symposium on Information theory.

Wasserman, F., 2006. Neurocomputing technique: Theory and practice. Mir, 1992, pp: 240. The 47 Simon Haykin. Neural networks: aA complete course. Williams, pp: 1104.

Yang, S., I. Verbauwhede, 2005. Automatic secure fingerprint verification system based on fuzzy vault scheme. Proc. IEEE ICASSP, pp: 609-612.

Yazov, Yu.K., V.I. Volchikhin, A.I. Ivanov, V.A. Funtikov and I.G. Nazarov, 2012. Neural protection of personal biometric data. Radiotechnika, pp: 157.