Synthesis of four new neuro-statistical tests for testing the hypothesis of independence of small samples of biometric data

V I Volchikhin, A I Ivanov, T A Zolotareva and D M Skudnev
Lipetsk State Pedagogical P.Semenov-Tyan-Shansky University, Lipetsk, 398020, Russia

Abstract. The paper considers the analysis of small samples according to several statistical criteria to test the hypothesis of independence, since the direct calculation of the correlation coefficients using the Pearson formula gives an unacceptably high error on small biometric samples. Each of the classical statistical criteria for testing the hypothesis of independence can be replaced with an equivalent artificial neuron. Neuron training is performed based on the condition of obtaining equal probabilities of errors of the first and second kind. To improve the quality of decisions made, it is necessary to use a variety of statistical criteria, both known and new. It is necessary to form networks of artificial neurons, generalizing the number of artificial neurons that is necessary for practical use. It is shown that the classical formula for calculating the correlation coefficients can be modified with four options. This allows you to create a network of 5 artificial neurons, which is not yet able to reduce the probability of errors in comparison with the classical formula. A gain in the confidence level in the future can only be obtained when using a network of more than 23 artificial neurons, if we apply the simplest code to detect and correct errors.

1. Introduction
When solving a number of practical problems (medicine, biology, biometrics, economics, physics, chemistry) it is difficult to obtain large data samples. As a result, it is necessary to statistically analyze small samples of 16 to 21 experiments. Usually, in statistical analysis, the first statistical moments are assessed, such as: mathematical expectation, standard deviation and correlation coefficient.

Significant attention has been paid to statistical analysis of small samples, both in the last century and in the 21st century. The results achieved in this direction are reflected in the handbook on mathematical statistics. The reference book contains a description of about 200 statistical tests, including this source contains a description of more than 30 statistical tests created to test the hypothesis of data independence.

The problem of testing the hypothesis of data independence is illustrated in figure 1, which shows the distributions of the values of the correlation coefficients calculated using the classical Pearson-Edgeworth-Weldon formula (1890).
Due to the small sample size of 16 experiments, the independent data \( r = 0.0 \) give values that fall in the range from \( r = -0.75 \) to \( r = +0.75 \). Unfortunately, the classical formula for calculating the correlation coefficients is based on the use of two mathematical expectations \( E(.) \) and two standard deviations \( \sigma(.) \):

\[
r(x, y) = \frac{1}{16} \sum_{i=1}^{16} \frac{(E(x_i) - x_i)(E(y_i) - y_i)}{\sigma(x_i) \sigma(y)}
\]

The calculation of the lowest statistical moments on small samples leads to the appearance of significant errors \( \Delta E(.) \) and \( \Delta \sigma(.) \). In calculations of the form (1), the errors of the initial data \( \Delta E(.) \) and \( \Delta \sigma(.) \) are accumulated. This is one of the main reasons why the estimation of the correlation coefficients according to the classical formula (1) has a significant error - \( \Delta r(x, y) \). This error turns out to be acceptable only for sufficiently large samples of 160 to 200 samples.

The purpose of this article is to show that there are ways to solve the problem that, in the future, will make it possible to perform statistical analysis of small samples of 16 to 20 experiments with an acceptable confidence level for practice.

2. Neural network generalization of the set of statistical criteria for testing the hypothesis of independence of small samples of biometric data

The problem of statistical evaluations on small samples in neural network biometrics is especially urgent. Thus, the national standard GOST R 52633.5-2011 is focused on the automatic training of a network of artificial neurons using 20 examples of the “Own” image. Moreover, for most of the biometric parameters of a “good” training sample, the distribution laws should be close to normal. That is, before training, we must test the hypothesis of normality using one of the 21 known statistical tests for testing the hypothesis of normality.

Biometric-neural network authentication technologies are focused on the information security market, which is regulated by two state services, Russian Federal Security Service and the FSTEC of Russia. That is, mass products of biometric-neural network authentication will need to have certificates from these two organizations. In this regard, “FSTEC of Russia” for future certification has created 7 existing standards through the efforts of the technical committee for standardization No. 362 “Information Security”. Technical Committee for Standardization No. 26 “Cryptographic Information Security” has developed a technical specification for recurrent encryption and decryption of data tables of trained neurons. The introduction of this technical specification on the territory of the Russian Federation is expected to occur in the second half of 2021.

It should be noted that the developed domestic standards and technical specifications apply only to neural networks with data accumulation in linear space. We can talk about a high level of national
standardization of the technology for processing biometric data by networks of perceptrons (neurons that accumulate data by weighted summation).

Apparently, the next class relevant for standardization will be networks consisting of quadratic neurons. This type of artificial neurons accumulates relatively poor "raw" biometric data in quadratic spaces [1-5].

In general, all quadratic neurons can be described in terms of the quadratic form:

\[
\begin{align*}
\text{where } \hat{e}^2 & \leftarrow [E(x) - \bar{x}]^T \cdot [r]^{-1} \cdot [E(x) - \bar{x}] \\
\text{if } e^2 < k_1 & \rightarrow "00", \\
\text{if } k_1 < e^2 < k_2 & \rightarrow "01", \\
\text{if } k_2 < e^2 < k_3 & \rightarrow "10", \\
\text{if } e^2 > k_3 & \rightarrow "11", \\
\end{align*}
\]

(2)

where \(\bar{x}\) is the vector of "raw" normalized biometric parameters with unit standard deviation \(\sigma(x) = 1\); 
\(r^{-1}\) — inverse correlation matrix of "raw" biometric parameters; \(z(.)\) is a four-level output quantizer of an artificial neuron with three comparison thresholds \(\{k_1, k_2, k_3\}\).

The most difficult operation in setting up an artificial neuron (2) is the ill-conditioned operation of calculating the inverse correlation matrix. Obviously, this operation becomes robust if the input data is not correlated (not dependent). In this case, the correlation matrix turns out to be unit and its inversion becomes stable.

Nevertheless, the problem of accurately calculating the correlation coefficients on small samples remains. So, if we receive data of 16 examples, then we are not able to test the hypothesis of independence \(r = 0.0\) with acceptable accuracy. The classical formula (1) in place of the real value \(r = 0.0\) will give values lying in the range from \(r = -0.75\) to \(r = +0.75\) (see figure 1). Such significant deviations are unacceptable for practice; however, we can nevertheless create an artificial neuron based on the classical formula (1) capable of distinguishing between the state \(r = 0.0\) and the state \(r = 0.5\) with the probabilities of errors of the first and second kind \(P_1=P_2=0.146\).

It is obvious that for all 15 known classical statistical criteria for testing the hypothesis of independence [1], an artificial neuron can be constructed. That is, today a network of 15 artificial neurons can be built, generalizing the statistical criteria for testing the independence hypothesis created in the last century.

By analogy with the criteria for testing the hypothesis of checking the data normality, work was carried out to create new statistical tests for testing the hypothesis of independence. In particular, a criterion (artificial neuron) with two linear quantizers [6, 7] and an artificial neuron with two elliptical quantizers [2] were created. Also, a fractal-correlation functional was created, built on the ascending ordering of one of the analyzed variables [3]. As a result, at the moment, we are able to create a network of 17 neurons, generalizing 17 currently known statistical criteria for testing the hypothesis of independence.

3. Synthesis of four new statistical tests for testing the hypothesis of independence of small sample data

Since an increase in the number of statistically criteria (the number of artificial neurons) is beneficial, we will try to synthesize four more new statistical criteria. In this case, we will synthesize new statistical criteria (new neurons) based on the classical Pearson-Edgeworth-Weldon formula (1), which has been used in statistical evaluations for more than 130 years. One of the modifications of formula (1) can be constructed by rejecting data normalization through division by standard deviations (division by integral characteristics \(\sigma(x), \sigma(y)\)). Instead of integral normalizing statistics, we will use particular statistics calculated for each point. As private normalizing statistics, we will use the squares of the distance from each sample point to the center of the two-dimensional data distribution:

\[
\hat{r}(x, y) = \frac{1}{16} \sum_{i=1}^{16} \frac{(E(x) - x_i)(E(y) - y_j)}{(E(x) - x_0)^2 + (E(y) - y_0)^2}
\]

(3)
The results of numerical modeling of the new criterion (3) are shown in figure 2.

![Figure 2. Distribution of estimates of the correlation coefficients calculated by the formula, which is an analogue of the classical formula.](image)

If we compare the data in figure 1 and figure 2, it becomes obvious that the power of the new criterion decreases in comparison with the classical criterion (1). For the classical formula, the same probabilities of errors of the first and second kind of its neuron are $P_1 \approx P_2 \approx 0.146$. For the modified formula (2), the equiprobable errors of its neuron are higher than $P_1 \approx P_2 \approx 0.211$. We observe a decrease in the power of the new statistical criterion in comparison with the classic by 1.45 times. However, it is of fundamental importance that two neurons, equivalent to calculations by two formulas, have a significant independent component of their output states $corr(r, \bar{r}) = 0.75$. If the responses of two neurons were completely dependent, then it would not be advisable to use such neurons in one neural network.

Other variants of modification of the classical formula (1) are also possible. The studies carried out have shown that the following modifications give acceptable results:

- \[ \bar{r}_1(x, y) = \frac{1}{16} \sum_{i=1}^{16} \frac{(\min(x)-x_i)(\min(y)-y_i)}{((\min(x)-x_i)^2+(\min(y)-y_i)^2} \] (4)

- \[ \bar{r}_2(x, y) = \frac{1}{16} \sum_{i=1}^{16} \frac{(\min(x)-x_i)(\max(y)-y_i)}{((\min(x)-x_i)^2+(\max(y)-y_i)^2} \] (5)

- \[ \bar{r}_3(x, y) = \frac{1}{16} \sum_{i=1}^{16} \frac{(\max(x)-x_i)(\max(y)-y_i)}{((\max(x)-x_i)^2+(\max(y)-y_i)^2} \] (6)

A numerical experiment on modeling new statistical criteria (4), (5), (6) gives distributions similar to the distributions in figure 2. Equal probabilities of errors of the first and second kind $P_1 \approx P_2$ and the coefficients of the correlation entanglement of the output states of the five neurons considered above are given in table 1.

**Table 1. Statistical parameters of the neural network generalization of five criteria.**

| Formula | $P_1 \approx P_2$ | Correlation coefficient matrix |
|---------|-------------------|--------------------------------|
| $r$ (1) | 0.146             | $r$ (1) | 1   | 0.75 | 0.338 | 0.344 | 0.316 |
| $\bar{r}$ (2) | 0.211             | $\bar{r}$ (2) | 0.75 | 1   | 0.317 | 0.324 | 0.295 |
| $\bar{r}_1$ (3) | 0.334             | $\bar{r}_1$ (3) | 0.338 | 0.317 | 1   | -0.17 | 0.089 |
| $\bar{r}_2$ (4) | 0.395             | $\bar{r}_2$ (4) | 0.344 | 0.324 | -0.17 | 1   | -0.192 |
| $\bar{r}_3$ (5) | 0.328             | $\bar{r}_3$ (5) | 0.316 | 0.295 | 0.089 | -0.192 | 1   |
4. Prediction of the required number of neurons by symmetrizing the task of multivariate neural network data analysis

It is rather difficult to conduct a direct numerical experiment that takes into account different error probabilities for different neurons (left side of the table 1) and reproduces their asymmetric correlation matrix (right side of the table 1) for networks with a large number of artificial neurons. With an increase in the number of neurons taken into account, the requirements for the speed of the calculator and the requirements for its RAM in direct modeling grow exponentially.

It is possible to simplify the problem of numerical simulation by symmetrizing it [4]. In symmetrization, it is assumed that all neurons have the same probabilities of errors of the first and second kind. The estimation of the equivalent error probabilities is performed by calculating the geometric mean:

$$ P_1 \approx P_2 \approx 0.271 \approx \sqrt[5]{0.146 \cdot 0.211 \cdot 0.334 \cdot 0.395 \cdot 0.328}. $$

In this case, the correlation matrix of artificial neurons must also be symmetrical. All its elements outside the diagonal must be the same and must coincide with the mean of the absolute values of the correlation coefficients of the real matrix:

$$ R_{\tilde{1}} \approx R_{\tilde{2}} \approx R_{\tilde{3}} \approx R_{\tilde{4}} \approx R_{\tilde{5}} \approx \frac{1}{5} \left( \left| R_{ij} \right| \right) $$

For the correlation matrix of the right side of table 1, the mean of the modules of the correlation coefficients is $\tilde{R} \approx 0.36$.

Numerical modeling of a symmetric neural network is greatly simplified and allows reproducing the probabilities of the code states of 1, 3, 5, 9 artificial neurons on an ordinary computer at a computer time of no more than 10 minutes. In our case, for five neurons, their ideal code will have five “0” states. That is, each of the neurons recognizes the tested sequences as independent or negatively correlated. For independent input data received from the software generator, the code states “00000” appear with a probability of 0.335. With a probability of 0.275, one in five neurons refuses to detect independent data. The worst case, when all neurons make a mistake, giving the state “11111” appears with a probability of 0.021. The amplitudes of the probability of occurrence in the bits of the code of erroneous states “1” are shown in figure 3.

![Figure 3](image3.png)

**Figure 3.** An example of the distribution of the probability amplitudes of 6 lines of the Hamming distance spectrum for the output codes of a network of 5 artificial neurons when exposed to independent data.
Note that the output code of a network of 5 neurons has 5-fold redundancy. This redundancy can be eliminated by applying some error detection and correction code. In the simplest case, a code based on counting the “0” states in its bits can be used. If the number of states “0” in the code bits is more than states “1”, then a decision is made to detect independent or negatively correlated data.

To predict the expected results, it is fundamentally important that, in logarithmic coordinates, the error probabilities are linearly related to the number of symmetric neurons, as shown in figure 4.

This makes it easy to extrapolate the results of numerical modeling of networks with a small number of artificial neurons to networks consisting of tens or even several hundred artificial neurons.

5. Transition from static neural network analysis of small samples to software support for their neurodynamic analysis

Unfortunately, increasing the number of statistical criteria used with subsequent analysis of redundant output codes in statics is ineffective. It is necessary to move from analysis in statics to analysis in neurodynamics. For example, this can be done by using the main sample of 21 experiments and then generating many smaller subsamples from it. Figure 5 shows a diagram of the organization of such data processing.

The organization is based on a random selection of 5 experiments in the main sample and their removal from it [4]. This is done by a special program “modulating” the input data of the neural network. In the mode of software support for neurodynamics, constantly changing subsamples of 16 experiments are sent to the inputs of the neural network. Accordingly, the neural network reacts to them with 5-bit output codes. The result is a stream of flickering states that can be used to make an unambiguous decision.

![Figure 4](image-url)

**Figure 4.** Solid linear - forecast of error probabilities when using the simplest error correction code; dashed line - forecast based on the use of a complex error correction algorithm.
Figure 5. Transition to neurodynamics by using a sample of 21 experiments, random selection of 5 experiments and discarding them.

As in the analysis of data in statics, the decision is made by counting the states “0” and states “1” in the analyzed stream. Upon detection of the fact that the number of states “0” exceeds the number of detected states “1”, a decision is made to confirm the hypothesis of data independence in a small sample of 21 experiments. In this case, the reliability of the decisions made turns out to be significantly higher than the static analysis of a single sample. The increase in the reliability of the final solution is due to the effect of averaging the set of possible close solutions.

6. Results discussion
As a result, it turns out that the most powerful in the considered five statistical criteria is the classical criterion calculated by the formula (1). However, four new, less powerful statistical tests are useful. The network of five considered neurons gives a probability of errors of 0.198, which is worse than the probability of errors of 0.146, which is given by the classical formula. If we use the most primitive error detection and correction code, then the networks of artificial neurons will work more reliably than the classical formula (1) only if they have more than 23 artificial neurons in their composition. At the same time, the preprint [1] shows the possibility of neural network generalization of 17 currently known statistical criteria. That is, to obtain results better than the classical Pearson-Edgeworth-Weldon criterion (1), additional accounting for 4 new statistical criteria is not enough. Several newer statistical tests need to be synthesized. To obtain practical confidence probabilities, it will be necessary to additionally synthesize more than 100 new statistical criteria and their equivalent neurons in place of 4 new statistical criteria.

In comparison with the prospect of implementing a multicriteria neural network analysis of small samples in statics, it is advantageous to switch to the analysis of small samples in dynamics. In the case considered by us, using the main sample of 21 experiments and reducing its volume to 16 experiments by random thinning of the 5 artificial neurons considered in the article, it is already enough to obtain solutions with a confidence level higher than that given by one classical Pearson-Edgeworth-Weldon neuron (one). Nevertheless, in order to achieve a confidence level of 0.996, significant research work will have to be done, increasing the number of statistical tests known to date from 21 to 100. Approximately 80 more new statistical tests will have to be synthesized.
References

[1] Ivanov A I and Zolotareva T A 2020 Artificial Intelligence in Protected Execution: Synthesis of Statistical Neural Network Automata for Multicriteria Testing of the Hypothesis of Independence of Small Samples of Biometric Data Preprint (Penza: Publishing house of "Penza State University") 105 ISBN 978-5-907364-24-0 https://www.elibrary.ru/item.asp?id=44402980

[2] Serikov A V and Kachalin S V 2019 Correlation molecule with elliptical quantizers for computations on small training samples Proceedings of the 1st All-Russian Scientific and Technical Conference “Security of Information Technologies” (Penza) 123-9 https://www.elibrary.ru/item.asp?id=40845233&pf=1

[3] Ivanov A I, Lozhnikov P S and Serikova Yu I 2016 Reducing the Size of a Sample Sufficient for Learning Due to the Symmetrization of Correlation Relationships Between Biometric Data Cybernetics and Systems Analysis 3 49-56 https://link.springer.com/article/10.1007/s10559-016-9838-x

[4] Kulagin V, Ivanov A, Gazin A and Akhmetov B 2016 Cyclic continual quantum computing: power amplification of the chi-square test on small samples Analytics 5(30) 88-95

[5] Sulavko A E 2020 Biometric authentication based on a network of Bayesian hyperbolic neurons with three-level quantizers Proceedings of the XI All-Russian Scientific and Practical Conference of Students, Postgraduates, Education and Industry Workers (Omsk) 199-206 https://www.elibrary.ru/item.asp?id=43962653

[6] Duda R and Hart P 1976 Image recognition and scene analysis (Moscow, Russia: Publishing house “Mir”) 511

[7] Hunt E 1978 Artificial intelligence (Moscow, Russia: Mir) 558

[8] Tou J T and Gonzalez R C 1980 Pattern Recognition Principles (Moscow, Russia: Sov. Radio) 408

[9] Lawson C and Hanson R 1966 Numerical solution of problems by the least squares method (Moscow, Russia: Nauka) 230

[10] Forsythe J and Moler K 1969 Numerical solution of systems of linear algebraic equations (Moscow, Russia: Mir)

[11] Rice J 1984 Matrix calculations and mathematical support (Moscow, Russia: Mir) 412