Statistical data representation in the context of increasing construction efficiency

A Gazin\textsuperscript{1*}, A Ivanov\textsuperscript{1}, A Malygin\textsuperscript{1} and A Isaeva\textsuperscript{1}

\textsuperscript{1}Lipetsk State Pedagogical P. Semenov-Tyan-Shansky University (National Research University), 42, Lenina street, 398020, Lipetsk, Russia

E-mail: yearn@bk.ru

Abstract. The purpose of the paper is the development of new statistical tests and new forms of their representation which add classic tests and classic form of their description. The development of new building materials requires several years and is quite expensive. Builders cannot enter the market and recoup their investments because of the need to test a new matter for sufficiently large samples of buildings. The tests in terms of time and cost resources can be comparable to the ones for development of a new substance. Thus, the experiments can be very expensive, which makes it difficult to quickly get enough statistical data to increase construction efficiency. That is why it is important to develop new statistical tests and new forms of their representation. It is shown that the spectrum of a geometric mean molecule has a smaller number of spectral lines for small samples of biometric data as compared with the molecule of chi-square criterion. In addition, the probability amplitudes of the spectral components of a geometric mean molecule have a close to normal distribution. This advisable distinguishes the geometric mean molecule from the chi-square molecule. The criterion for the geometric mean of small samples has greater power as compared with the chi-square criterion. In order to increase the accuracy of statistical analysis of small samples, it is necessary to extend a variety of statistical criteria both in the usual continuous representation of their states, and in the analysis of their discrete spectra of the probability amplitudes. To increase the efficiency of the construction sector, it is necessary to transfer to the use of deep neural networks that analyze dozens of new and old statistical criteria, their equivalent neurons and equivalent artificial molecules. Technological capabilities for creating these structures already exist today.

1. Introduction
When analyzing the construction data, their sample is often small. A doctor can observe the biometric data of about 20 patients with the same diagnosis throughout a year. A similar pattern is observed among biologists, botanists, chemists, physicists. Biologists and botanists are primarily interested in rare species of animals and plants, which cannot be many by definition. The same problem arises when training the neural network based converters of the individuality into a cryptographic key code [1-4]. Users consider the biometric system to be friendly if it learns using 20 examples of the “Self” image (for example, it requires attaching a finger to a live-scan reader 20 times). If an identity-based biometric system will require 200 examples of a fingerprint pattern for its learning, then users interpret this system as not friendly. Fully the same arises when applying the statistical criteria [5-13]. To reliably check the normal-theory test, the chi-square criteria requires a sample of 200 examples or
more [1]. For small samples of 20 examples, the chi-square criteria gives unacceptably large probabilities of errors of the first and second kinds \( P_1 \approx P_2 \approx P_{EE} \approx 0.32 \) [5]

2. Materials and Methods

2.1. A neuron equivalent to the geometric mean square criteria

It must be emphasized that one of the methods to reduce the errors of statistical analysis of small samples is the use of several statistical tests. It is convenient to combine the results of different statistical tests if they are presented as artificial neurons. In particular, an artificial neuron acting as the square of the geometric mean criteria for a sample of 21 experiments and a histogram consisting of 6 intervals is described by the following system of functional relations:

\[
\begin{align*}
    x & \leftarrow \text{sort}(x) \\
    \Delta & \leftarrow \frac{x_{20} - x_0}{6} \\
    \Delta_i & \leftarrow x_0 + \Delta \cdot i \\
    g_{m^2} & \leftarrow \sum_{i=0}^{6} \left( P(\Delta_i) \cdot \overline{P}(\Delta_i) \right)^{1/2} \\
    z & \leftarrow "0" \text{ if } g_{m^2} \leq 3.5 \\
    z & \leftarrow "1" \text{ if } g_{m^2} > 3.5
\end{align*}
\]

(1)

where \( P(\Delta_i) \) is the experimental probability that the sample data fall within the histogram interval \( -\Delta_i \), \( \overline{P}(\Delta_i) \) is the theoretical probability that the sample data fall within the same histogram interval. Figure 1 shows the densities for distribution of values of the considered test (artificial neuron) for samples with a normal distribution and ones with a uniform distribution of values. Figure 1 shows that the distributions of the \( g_{m^2} \) values after their smoothing are well described by normal laws. The equally probable error values of the first and second kinds are \( P_1 \approx P_2 \approx P_{EE} \approx 0.19 \), which is significantly less than those for the chi-square criteria (neuron chi-square).

![Figure 1. Densities for distribution of the geometric mean square criteria for a sample of 21 experiments of normal and uniform data.](image)

2.2. Molecule of the geometric mean square

It is known that the analysis of small samples in chemistry, physics, and criminalistics by continual methods (weighing, titration, transmission) cannot give the results with very high accuracy. It is possible to assess the presence of impurities with a concentration of 0.1% by volume. The impurity concentrations less than approximately a billion can be detected if we proceed to the analysis of the position of the impurities discrete spectrum lines. While observing the flame color with the eyes, we cannot tell which gas burns out and which chemical elements are present in the flame. The situation changes if the light is dispersed by a prism according to Newton, directed onto the screen and the spectral lines of burning chemical elements are observed. Since the potential accuracy of discrete methods is much higher than the accuracy of continual procedures, we create a molecule of the
geometric mean square; the spectrum of output states, as well as a hydrogen molecule, is always linear (discrete). The mathematical molecule of the criteria under consideration is described by the following system of functional relations:

\[
\begin{align*}
\Delta & \leftarrow 1.0, \\
\Delta_i & \leftarrow -3.0 + i \cdot \Delta - E(x), \\
gm^2 & \leftarrow \sum_{i=0}^{5} \sqrt{P(\Delta_i) \cdot \hat{P}(\Delta_i)}.
\end{align*}
\]

(2)

where \(\sigma(x)\) is the standard deviation of the small sample, \(E(x)\) is the mathematical expectation of the normalized sample.

It must be emphasized that a quantifier is present in the \(gm^2\) neuron (1); it makes the output spectrum of neuron states discrete. Two states “0” and “1” may appear at its output. The \(gm^2\) molecule has a much more complex spectrum of its output states. The amplitudes of the probability for occurrence of various spectral lines of the \(gm^2\) molecule are illustrated in Figure 2.

Figure 2. Spectra of the output states of the molecule of the geometric mean square (17 most significant probability amplitudes) for a sample of 21 experiments.

As it is clear from Figure 2, the position of the molecule spectral lines coincides at normal and uniform data, but the amplitudes of the probability for their appearance differ significantly. The data on the 19 most significant spectral lines are given in Table 1.

Table 1. The values of 19 most significant amplitudes of the probabilities for appearance of spectral lines of the geometric mean square.

| Number | 1   | 2   | 3    | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|--------|-----|-----|------|-----|-----|-----|-----|-----|-----|-----|
| gm^2   | 4.7 | 4.9 | 4.98 | 5.12| 5.19| 5.26| 5.31| 5.4 | 5.47| 5.52|
| Normal | 0.003| 0.003| 0.07 | 0.236| 0.386| 0.07 | 0.66 | 1.07 | 0.29 | 0.89 |
| 10 \cdot \Psi(gm^2) | 0.166 | 0.733 | 0.05 | 1.782 | 0.202 | 0.03 | 2.55 | 0.38 | 0.05 | 2.08 |
| Uniform | 55.33| 244.3| 1.40 | 7.55 | 1.911| 2.333| 3.86 | 2.816| 5.80 | 2.34 |
| \frac{\max(\Psi)}{\min(\Psi)} | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| gm^2   | 5.6 | 5.67 | 5.73 | 5.8 | 5.88 | 5.93 | 6.01 | 6.08 | 6.21 |
| Normal | 1.79 | 0.56 | 0.72 | 1.40 | 0.44 | 0.28 | 0.49 | 0.165| 0.160|
| 10 \cdot \Psi(gm^2) | 0.39 | 0.05 | 0.94 | 0.24 | 0.11 | 0.20 | 0.048| 0.016| 0.003|
| Uniform | 4.59 | 11.2 | 1.31 | 5.833| 4.0 | 1.4 | 10.28 | 10.313| 53.33 |
As it is clear from table 1, the probability amplitudes of normal and uniform data can differ dozens of times. It is this circumstance that allows hoping for a significant increase in the accuracy of statistical estimates at small samples. If the ratio of the probability amplitudes with the same numbers is singular, then this spectral line is not fit for decisions making. The spectral lines with very different probability amplitudes are of the main interest. Three most informative spectral lines with numbers 1, 2, 19 as per Table 1 are marked with a fill. Note that these most informative spectral lines are at the beginning and at the end of the spectral sequence. In the central part of the spectral sequence, the lines’ informativity randomly varies over a fairly wide range. The second spectral line is the most informative for two spectra under consideration. If this is taken into account only, then it is possible to reach equally probable errors at the level of $P_1 \approx P_2 \approx P_{EE} \approx 0.004$. This indicator is more than 45 times better than the one obtained in classical continual calculations.

2.3. The use of symmetrisation in estimating the benefit from the transition to multivariate discrete spectral analysis of small samples

Supposing that the spectral lines with different numbers give independent estimates, then the final probability of errors can be estimated as the product of probabilities for errors appearance when analyzing each spectral line:

$$P_i = P_1 \approx P_2 \approx P_{EE} \approx \prod_{i=1}^{19} \frac{\min(\Psi_i)}{\max(\Psi_i)},$$

Estimate (3) is too optimistic, since technically it is extremely difficult to achieve independence of the analyzed data. Nevertheless, we can calculate this through the geometric mean of the averaged probability of errors of the first and second kinds for one (averaged) line of the spectra considered – $P_{EE} \approx 0.014$. As a result, the probability of errors for 19-dimensional generalization of independent data should decrease to a value of $P_{EE} \approx (0.014)^{19} \approx 10^{-15}$. This value is extremely small and technically unreachable due to the condition of independence of the generalized data, however, it is quite realistic to obtain the conditions under which the coefficient of equal correlation of the generalized data will be the final value. This pattern arises when a small sample of 42 experiments is analyzed. In this case, we can randomly select many smaller samples of 21 experiments from the initial relatively large sample. A subsamples size should be half as large as that of the original sample. This ratio results in the data are relatively weakly correlated $corr(.,.) \approx 0.5$. In this case, the maximum possible number of small incompletely repeating subsamples is determined through the following factorials:

$$\binom{42}{21} = \frac{42!}{21!(42 - 21)!} = 53\,825\,7874\,440.$$  \hspace{1cm} (4)

Such a scope of information is quite enough to build a spectrum with data accuracy that is much higher than the data given in Table 1. When calculating the data from Table 1, a significantly smaller sample was used in 999,999 experiments. It is managed to take into account the influence of correlation relations between data by means of simulation modelling. In order to obtain equally correlated data, it is necessary to multiply random, uncorrelated initial data by a symmetric linking matrix $[6, 10]$:

$$\begin{bmatrix} 1 & a & a & a \\ a & 1 & a & a \\ a & a & 1 & a \\ a & a & a & 1 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & r & r & r \\ r & 1 & r & r \\ r & r & 1 & r \\ r & r & r & 1 \end{bmatrix} \begin{bmatrix} \text{corr}(y_i, y_j) \end{bmatrix}.$$

For any matrices of any dimension, it is always possible to select the value of the linking matrix elements so that all elements of the correlation matrix outside the diagonal take the value $r = 0.5$. This
manoeuvre allows obtaining the values for reducing equally probable errors in the generalization of 2, 3, 4, 5, 6 spectral lines of the \( gm^2 \) molecule. The numerical experiment data are shown in Figure 3.

![Figure 3](image)

**Figure 3.** Numerical simulation results for equally correlated data.

Figure 3 shows that the maximum possible probability of errors will be \( 2.22 \times 10^{-6} \) for a sample of 42 experiments, provided that the \( \text{corr(.,.)} \approx 0.5 \) data of 19 spectral lines are equally correlated. In order to reduce an error to \( 10^{-7} \), it will be necessary to complicate the task by increasing a number of spectral lines to 57. This cannot be made for \( gm^2 \) molecules with the histogram having 6 intervals in 16 experiments of the sample. It is necessary to increase either a number of experiments or the sample size.

3. Results

3.1. Algorithms complexity is the training of two types of artificial neural networks compressing intermediate information redundancy

It must be emphasized that in this paper we consider the comparative possibility of two options for the implementation of the statistical square criteria of the geometric mean. In this case, a continuous option of this statistical criterion in a sample of 21 experiments gives solutions with a low confidence probability of 0.81. The same statistical criterion, when it is executed in the form of a mathematical molecule and several of the most informative spectral lines are taken into account, has a much higher potential at the decision-making level with a confidence probability of 0.99999 and higher. The transition to spectral analysis should give very good results in the future, however, it will be necessary to eliminate the spectral information redundancy, for example, this can be done using "deep" neural networks. In this case, it will be necessary to apply the industrial technology of their training that is based on very large training samples (hundreds of millions of implementations of small samples of 21 experiments) and the use of powerful computers. Currently, this industrial technology is well developed to identify the faces [14-19], and it is also actively used to identify handwritten numbers and sounds of continuous speech. There is no doubt that this technology can also be used to identify the curves of small samples distribution laws. At the same time, there is no use to analyze the spectral lines of only one molecule of the geometric mean square with deep neural networks. Much more effective is the parallel use of many statistical tests and equivalent statistical molecules [20-21].

As a result, it turns out that it will be necessary to use dozens of parallel neurons for continuous statistical analysis of small samples or dozens of statistical molecules for the close neural analysis of their spectral lines. Figure 4 shows a decision circuit through analyzing continual data by several neurons. The intermediate code redundancy of this circuit in the simplest case is convolved by counting the majority of output states of neurons. The neurons’ training with this circuit of analysis does not cause problems.
**Figure 4.** Neural generalization of several statistical criteria (each neuron is the equivalent of one of the statistical tests).

**4. Discussion**

A much more complicated design is the modulation of the input data by extracting small subsamples of 21 experiments from a large sample of 42 experiments and the subsequent analysis of the most informative spectral lines by the deep neural network (Fig. 5). It is not difficult to write a software data modulator and artificial molecules. The learning process of “deep” neural networks is complex. It is assumed that by the end of 2020, a neural corrector with the structure of figure 4 will be created for small samples from 16 to 21 experiments, providing an confidence level of 0.99 acceptable for the practice. It is much more difficult to create more efficient structures for processing spectral lines with “deep” neural networks. Probably, it will only be possible to create and test structures from artificial molecules with close neural data compilation only after a few years.

**Figure 5.** Elimination of information redundancy of the majority of the most informative spectral lines of dozens of statistical molecules by using the pre-trained “deep” artificial neural network (ANN).

**5. Conclusions**

In the XX century, several dozens of statistical criteria were created for different eventualities, however, all of them have low capacity in small samples. It is obvious that by combining dozens of already known statistical criteria in the form of the artificial neurons networks, we are able to significantly increase the reliability of statistical estimates in small samples. In addition, already known statistical tests can be supplemented by new previously unknown ones. This paper covers one of the new statistical tests of the geometric mean square (currently, 9 statistical criteria of this family have been studied). Another fundamentally important technological point is the transition from
statistical criteria and their equivalent neurons to a discrete description in the form of statistical molecules. This transition seems strange, however, it should allow multiple increasing the confidence in statistical estimates. At least, the methods of discrete spectral analysis in chemistry, physics, and criminalistics have acquired a good reputation. Preliminary estimates of the capabilities of new methods have shown that the uncertainty interval for calculating the values of statistical estimates can be reduced by about 150 times. These are forecasts to be confirmed empirically. Real estimates of the correlation coefficients based on the use of a correlation molecule were performed, which reduced the calculation error in 9.1 times by means of 16 spectral lines.

References
[1] Volchikhin V I, Ivanov A I, Bezyaev A V, Kupriyanov E N 2019 Inzhenern. Tekhn. Sis. Engin. Techn. Sys Vol 29(2), pp 205–217, https://doi.org/10.15507/2658-4123.029/2019.02.205-217.
[2] Ivanov A I, Kupriyanov E N, Tureev S V 2019 Dependab. Vol 19(2), pp 22–27 https://doi.org/10.21683/1729-2646-2019-19-2-22-27
[3] Volchikhin V I, Ivanov A I, Gazin A I, Bannykh A G 2017 J. Comp. Engin. Math. Vol 4(4), pp 53–63 https://doi.org/10.14529/jcem170405
[4] Volchikhin V I, Ivanov A I, Pushchenko D V, Akhmetov B B, Vyatchanin S E 2017 Izvest. Vyssh. Uchebn. Zaveden. Povolzh. Reg. Tekhn. Nauk Vol 1, pp 5–15 https://doi.org/10.21685/2072-3059-2017-1-5
[5] Volchikhin V I, Ivanov A I, Serikov A V, Serikova Yu I 2017 Izvest. Vyssh. Uchebn. Zaveden. Povolzh. Reg. Tekhn. Nauk. Bullet. High. Edu. Instit. Volg. Reg. Techn. Sci. Vol 3(43), pp 70–80 https://doi.org/10.21685/2072-3059-2017-3-6
[6] Volchikhin V I, Ivanov A I, Gazin A I, Bannykh A G 2017 J. Comp. Engin. Math. Vol 4(4), pp 53–63 https://doi.org/10.14529/jcem170405
[7] Akhmetov B B, Ivanov A I 2016 Dependab. Vol 16(2), pp 43–48 https://doi.org/10.21683/1729-2640-2016-16-2-43-48
[8] Kodra S 2015 Fuzzy extractors: How to generate strong keys from biometrics and other noisy data
[9] Saleem H, Albermany S, Hadi H 2018 Intern. J. Engin. Techn. Vol 7(3.27), pp 129–131 https://doi.org/10.14419/ijet.v7i3.27.17672
[10] Rice J A 2007 Mathematical Statistics and Data Analysis (3rd edn.) (Ser.: Duxbury advanced) (Belmont: Thomson Brooks/Cole)
[11] Zhang L, Nie Y 2014 Comp. Model. New Techn. Vol 18(12C), pp 46–55
[12] Rolke W, Gongora C G 2020 Comp. Stat. arXiv:2005.02793 https://doi.org/10.1007/s00180-020-00997-x
[13] Huang R, Cui H 2015 J. Syst. Sci. Complex. Vol 28, pp 439–450 https://doi.org/10.1007/s11424-015-0351-2
[14] Akhmetov B, Ivanov A, Gilmudinov A, Funtikova Yu 2015 The family of chi-square molecules Pearson: Software-continuum quantum accelerators of high-dimensional calculations. In Software-Continuum Quantum Accelerators of High-Dimensional Calculations 15th International Conference on Control, Automation and Systems (ICCAS 2015) pp 1337–1341 https://doi.org/10.1109/ICCAS.2015.7364846
[15] Ivanov A I, Bezyaev A V, Gazin A I 2017 J. Comp. Engin. Math. Vol 4(2), pp 3–13 https://doi.org/10.14529/jcem170201
[16] Lozhnikov P S, Sulavko A E 2018 J. Phys. Conf. Ser. Vol 1050(1), Article 012047 https://doi.org/10.1088/1742-6596/1050/1/012047
[17] Ivanov A I, Lozhnikov P S, Bannykh A G 2019 J. Phys. Conf. Ser. Vol 1260(2), Article 022003 https://doi.org/10.1088/1742-6596/1260/2/022003
[18] Akhmetov B, Ivanov A I, Lozhnikov P S 2016 Multivariate statistical analysis of handwritten images via higher order correlation coefficients. In 2016 International Siberian Conference on
Control and Communications (SIBCON) pp 1–3
https://doi.org/10.1109/SIBCON.2016.7491759

[19] Ivanov A I, Bannykh A G, Serikova Y I 2020 Dependab. Vol 20(2), pp 28–34
https://doi.org/10.21683/1729-2646-2020-20-2-28-34

[20] LeCun Y, Bengio Y, Hinton G 2015 Nat. Vol 521, pp 436–444
http://dx.doi.org/10.1038/nature14539

[21] Goodfellow I, Bengio Y, Courville A 2016 Deep Learning (Ser.: Adaptive Computation and Machine Learning) (Cambridge, Massachusetts; London: MIT Press)