MSC 62E15 **DOI: 10.14529/jcem190105**

PRECISION STATISTICS: FRACTIONAL NUMBER OF DEGREES OF FREEDOM CHI-SQUARE CRITERION FOR SMALL SAMPLES OF BIOMETRIC DATA

V. I. Volchikhin[1], president@pnzgu.ru,
A. I. Ivanov[2], ivan@pniei.penza.ru,
E. A. Malygina[1], mal890@yandex.ru,
E. N. Kupriyanov[1], ibst@pnzgu.ru,
Yu. I. Serikova[1], julia-ska@yandex.ru
[1]Penza State University, Penza, Russian Federation,
[2]Penza Scientific Research Electrotechnical Institute, Penza, Russian Federation

The article presents the results of numerical modeling of the distribution of chi-square criterion for small samples with a volume of 8 to 80 examples for normal distribution of values. It is shown that, for small samples, the recommendations of Gosstandart R 50.1.037-2002 give too optimistic estimates of confidence probability when testing the hypothesis of normality of the empirical law. Errors in the estimation of confidence probability can be eliminated if we turn to the use of fractional indices of number of degrees of freedom. The connection curves of a fractional number of degrees of freedom of a chi-square distribution with sample size are given. The decrease in estimation errors with increasing sample size of experimental data is shown. So with 21 experiences, it is necessary to increase by 62% the typical value of the number of degrees of freedom. With a sample size of 81 experience, increasing the number of degrees of freedom falls by 31%. The need to adjust the number of degrees of freedom is completely eliminated only with samples of more than 2000 experiments. The logarithmic approximation of the fractional number of degrees of freedom for 7 and 9 equal intervals of the histogram is given.

Keywords: chi-square criterion; small samples; fractional value of the number of degrees of freedom.

Introduction

Informatization of modern society leads to the need to store personal data on the Internet "clouds", which requires using of their cryptographic protection. For ordinary people it's difficult to remember a personal cryptographic key consisting of 256 bits (32 random characters in 8-bit encoding) and applied in accordance with the national standards for cryptographic transformations. To simplify the applying and mass using of cryptography in Russia and abroad, technologies of transforming unique biometric images of a person into his personal cryptographic key are actively being developed. The so-called "fuzzy extractors" are used abroad [1, 2], and in Russia, neural network converters of biometrics-code are being created [3, 4]. Using of these technologies makes applying of cryptography convenient for ordinary users. Market of products combining biometrics and cryptography in Russia falls under competence of the two regulators "FSTEC (Federal Service for Technical and Export Control) of Russia" and "FSB (Federal Security Service) of Russia". Both regulators have specific requirements for information security of software and hardware, formulated in the form of national standards or specifications [4–6]. One of the problems of neural network biometrics is its testing [7,8]. Neural networks start to work

well only after they have been trained on a sample of sufficient volume. Before learning a neural network, you need an input quality control of the user-submitted 20 examples of the biometric image "Own". If the sample consisted of 400 examples, then it would be easy to calculate the mathematical expectations and standard deviations [9] of controlled biometric parameters. It is difficult to calculate exactly the mathematical expectation, standard deviation, correlation coefficients on a sample of 20 examples. The situation is saved by the fact that most of the controlled biometric parameters have a normal law of distribution of values. Knowledge of the distribution law substantially increases the accuracy of the calculation of the lowest statistical moments [10–12].

In connection with the above, the validation of the hypothesis of normality of small samples of biometric data is relevant, since the recommendations of Gosstandard (State Standard) R 50.1.037-2002 [13] on the use of chi-square criterion are focused on the sample size of 200 or more examples. Moreover, the verification of classical recommendations [13] showed that they contain methodological errors of an extremely simplified calculation of the number of degrees of freedom. The purpose of this article is to show that the rejection of the traditional application of integer values of number of degrees of freedom leads to a significant increase in the accuracy of the analytical statistical description of the classical chi-square Pearson criterion. It makes the chi-square test applicable to small samples of biometric data.

1. General Provisions for Practical Application of Chi-square Criterion

For estimates on the chi-square criterion, it is necessary to find maximum and minimum value of data in the sample. Next, you must specify the number of intervals -k histogram and find the width of these intervals:

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

Then, it is necessary to count number of examples of analyzed sample in each of k intervals of histogram. With this histogram forming, minimum sample value is the left border of the first column of histogram, and maximum value of the sample coincides with the right border of the rightmost column of histogram. The calculation of value of chi-square criterion is carried out according to the following formula:

$$\chi^2 = N\sum_{i=1}^{k} \frac{\frac{n_i}{N} - P_i}{P_i}, \tag{2}$$

where n_i is the number of experiments in the i-th column of the histogram, P_i is the expected probability that experiments get in the i-th column of the histogram. The popularity of chi-square Pearson criterion is due to the fact that it is known the analytical description of the density distribution of its values:

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

where m is the number of degrees of freedom, $\Gamma(.)$ is the Euler gamma function. It is recommended to choose the number of degrees of freedom according to the formula:

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

when the problem of checking normality of empirical statistics is solved. Formula (4) is usually justified by the fact that normal distribution law is described by two statistical moments (mathematical expectation and standard deviation). In order to use formula (2), we need to calculate two statistical moments, which should lead to a decrease in the number of degrees of freedom by two units. The larger number of statistical moments describes the theoretical law of distribution, the lower should be number of degrees of freedom - m. If the theoretical law is described in terms of d statistical moments, then the number of degrees of freedom should be [13]:

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

2. Numerical Experiment on the Estimation of Real Indicators of the Number of Degrees of Freedom

Pearson could not perform a numerical experiment in 1900, because there was no digital computers. Today, the situation has changed dramatically, any person who is able to use pseudo-random software data generators with normal law of distribution of values in such modeling environments as MathCAD, MathLAB, Mapl and others can perform the corresponding numerical experiment. The results of numerical simulation for a histogram of 7 columns for 21, 41, 81 experiments in the training set are shown in Fig. (1).

It can be seen from Fig. (1) that for small samples number of degrees of freedom of chi-square distribution is always fractional, while it decreases monotonically, tending to its limiting value $m = 4$, correctly indicated in 1900 by Pearson. It can also be seen from Fig. (1) that the real distributions of the values are noisy, but for them it is quite simple to construct the approximation in the form of the distribution (3) by choosing the value of the fractional number of degree of freedom.

3. Nomogram of the Connection of the Fractional Number of Degree of Freedom of Chi-square Distributions with the Sample Size

Obviously, for histograms with a given number of columns – k and a given value of the sample size – N, we can always carry out a numerical experiment and establish the dependence $m(N)$. Fig. (2) shows two such functional dependencies.

Presented dependencies are similar (almost parallel), their approximations are marked in the figure with a dotted line and are described by the following relations:

$$m_7(N) = 4.0 + 0.27|\lg(N) - 4|^{2.22}, \tag{6}$$

$$m_9(N) = 6.3 + 0.31|\lg(N) - 4|^{2.22}, \tag{7}$$

It turns out that the sample size of the order of 10 thousand examples is the point where the value of dimension coincides with the Pearson limit. Further increase in sample sizes

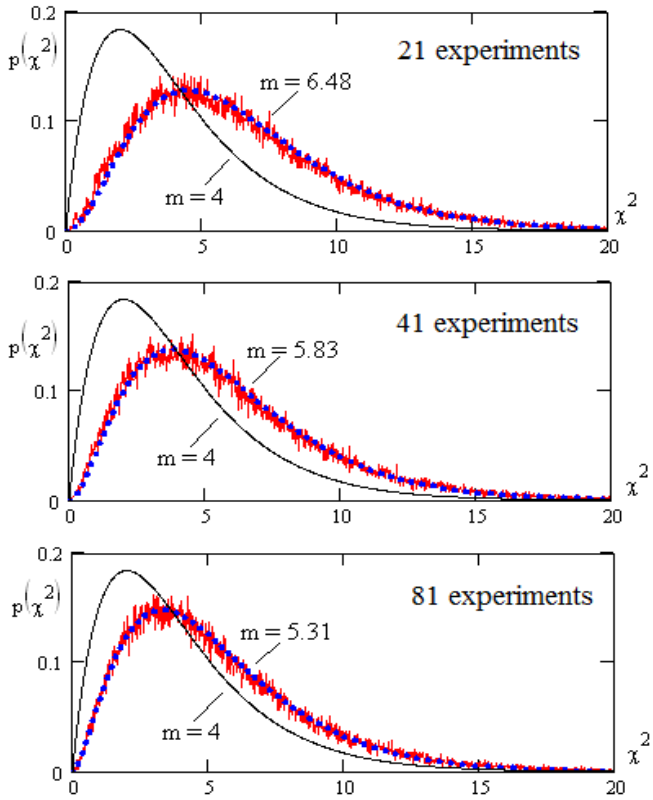

Fig. 1. Distributions of values of chi-square criterion for histograms with 7 columns(1)

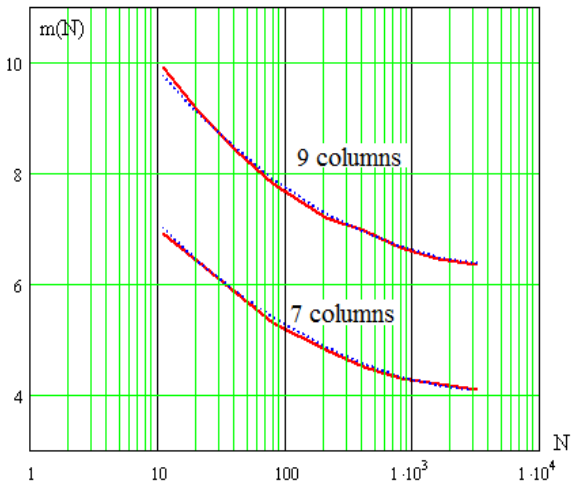

Fig. 2. Nomogram of the connection of the fractional number of degree of freedom of chi-square distributions for histograms with 7 and 9 columns(2)

practically does not affect the value of the fractional part of number of degrees of freedom. Number of degrees of freedom in this case becomes really an integer and, for a normal law, is actually described by the generally accepted relation (4).

Conclusion

Market regulators "FSTEC of Russia" and "FSB of Russia" make very high demands to domestic biometric technologies of information security on their reliability. If we rely on the established theoretical propositions of mathematical statistics, these rigid requirements can not be met. According to the recommendations formed in the last century [13], it is not possible to correctly describe the neural network converters of biometrics-code. The development of technology requires not only the creation of new standards for them [4–7], but also the adjustment of the usual statistical theories. In this article, we tried to show that gross estimates of number of degrees of freedom in the form of a series of integers are no longer sufficient. It is necessary to use more subtle procedures and consider number of degrees of freedom of chi-square criterion as fractional value. At the sample of 81 example , the calculated confidence probabilities turn out to be 2 times worse than for whole number of degrees of freedom. Of fundamental importance is also the fact that statistical estimates by chi-square criterion become more accurate and can be used for smaller samples.

References

1. Monrose F., Reiter M., Li Q., Wetzel S. Cryptographic Key Generation from Voice. *Proc. IEEE Symp. on Security and Privacy*, 2001, pp. 202–213.

2. Dodis Y., Reyzin L., Smith A. Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy. *Proceedings of International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004*. Berlin, Heidelberg, Springer, 2004, pp. 523–540. DOI: 10.1007/978-3-540-24676-3_31.

3. Volchikhin V. I., Ivanov A. I., Funtikov V. A. [*Fast Learning Algorithms for Neural Network Mechanisms of Biometric Cryptographic Data Protection: Monograph*]. PGU Publ., Penza, 2005. (in Russian)

4. *GOST R 52633.0-2006. [Information Protection. Information Protection Technology. Requirements for Means of Highly Reliable Biometric Authentication]*. Moscow, 2006. (in Russian)

5. *GOST R 52633.5-2011. [Information Protection. Information Protection Technology. Automatic Learning of Neural Network Converters Biometry-Access Code]*. Moscow, 2011. (in Russian)

6. *Technical Specification (Project, Public Discussion Started from 01.02.2017 by Members of TK 26 "Cryptographic protection of information"). Protection of Neural Network Biometric Containers using Cryptographic Algorithms*. (in Russian)

7. *GOST R 52633.3-2011. [Information Protection. Information Protection Technology. Testing of Stability of Highly Reliable Biometric Protection to Brute Force Attacks]*. Moscow, 2011. (in Russian)

8. Malygin A. Yu., Volchikhin V. I., Ivanov A. I., Funtikov V. A. [*Fast Testing Algorithms for Neural Network Mechanisms of Biometric Cryptographic Data Protection*]. PGU Publ., Penza, 2006. (in Russian)

9. *Environment Modeling "BioNeuroAvtograf"*, available at: http://xn–h1aanh6e.xn–p1ai/activity/science/noc.htm (accessed on January 10, 2019). (in Russian)

10. Volchikhin V. I., Ivanov A. I., Serikova Yu. I. Compensation of Methodological Errors in Calculations of Standard Deviations and Correlation Coefficient Occuring due to Small Sample Sizes. *University Proceedings. Volga Region. Engineering Sciences*, 2016, no. 1 (37), pp. 103–110. (in Russian)

11. Volchikhin V. I., Ivanov A. I., Akhmetov B. B., Serikova Yu. I. The Fractal-Correlation Functional Used when Searching for Pairs of Weakly Dependent Biometric Data in Small Samples. *University Proceedings. Volga Region. Engineering Sciences*. 2016, no. 4 (40), pp. 27–36. DOI: 10.21685/2072-3059–2016-4-3.

12. Kulagin V. P., Ivanov A. I., Serikova Yu. I. Correction of Methodical and Casual Components of Errors of Calculation of the Coefficients of Correlation Arising on Small Selections of Biometric Data. *Information Technologies*. 2016, vol. 22, no. 9, pp. 705–710. (in Russian)

13. [*P 50.1.037-2002. Recommendations for Standardization. Applied Statistics. Rules for Verifying the Agreement Between the Experimental Distribution and the Theoretical Distribution. Part I. Criteria of the Type Chi-square. Gosstandart of Russia*]. Moscow, 2002.

Vladimir I. Volchikhin, DSc (Techn), Professor, President of Penza State University (Penza, Russian Federation), president@pnzgu.ru.

Alexander I. Ivanov, DSc (Techn), Associate Professor, Head of the Laboratory of Biometric and Neural Network Technologies, Penza Scientific Research Electrotechnical Institute (Penza, Russian Federation), ivan@pniei.penza.ru.

Elena A. Malygina, PhD (Techn), Researcher of the Interdisciplinary Laboratory Testing of Biometric Devices and Technologies, Penza State University (Penza, Russian Federation), mal890@yandex.ru.

Evgenii N. Kupriyanov, Postgraduate Student, Department of Information Security of Systems and Technologies, Penza State University (Penza, Russian Federation), ibst@pnzgu.ru.

Yuliya I. Serikova, Postgraduate Student, Department of Computer Engineering, Penza State University (Penza, Russian Federation), julia-ska@yandex.ru.

Received January 24, 2019.

УДК 519.24; 53; 57.017 **DOI: 10.14529/jcem190105**

ПРЕЦИЗИОННАЯ СТАТИСТИКА: ДРОБНЫЙ ПОКАЗАТЕЛЬ ЧИСЛА СТЕПЕНЕЙ СВОБОДЫ ХИ-КВАДРАТ КРИТЕРИЯ ДЛЯ МАЛЫХ ВЫБОРОК БИОМЕТРИЧЕСКИХ ДАННЫХ

В. И. Волчихин, А. И. Иванов, Е. А. Малыгина, Е. Н. Куприянов, Ю. И. Серикова

В работе даны результаты численного моделирования распределения данных хи-квадрат критерия для малых выборок объемом от 8 до 80 примеров для нормального закона распределения значений. Показано, что для малых выборок рекомендации Госстандарта Р 50.1.037-2002 дают слишком оптимистические оценки доверительной вероятности при проверке гипотезы нормальности эмпирического закона. Ошибки оценки доверительной вероятности могут быть устранены, если перейти к использованию дробных показателей числа степеней свободы. Приводятся кривые связи дробного числа степеней свободы хи-квадрат распределения с размером выборки. Показано снижение ошибок оценки с ростом объема выборки экспериментальных данных. Так при 21 опыте необходимо увеличение на 62% типового значения числа степеней свободы. При объеме выборки в 81 опыт увеличивать число степеней свободы приходится на 31%. Необходимость в корректировках числа степеней свободы полностью отпадает только при выборках более 2000 опытов. Дано логарифмическое приближение дробного числа степеней свободы для 7 и 9 равных интервалов гистограммы.

Ключевые слова: хи-квадрат критерий; малые выборки; дробное значение числа степеней свободы.

Литература

1. Monrose, F. Cryptographic Key Generation from Voice / F. Monrose, M. Reiter, Q. Li, S. Wetzel // Proc. IEEE Symp. on Security and Privacy. – 2001. – P. 202–213.
2. Dodis, Y. Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy / Y. Dodis, L. Reyzin, A. Smith // Proceedings of International Conference on the Theory and Applications of Cryptographic Techniques, Interlaken, Switzerland, May 2-6, 2004. – Berlin, Heidelberg: Springer, 2004. – P. 523–540.
3. Волчихин, В. И. Быстрые алгоритмы обучения нейросетевых механизмов биометрико-криптографической защиты информации: монография / В. И. Волчихин, А. И. Иванов, В. А. Фунтиков. – Пенза: Изд-во ПГУ, 2005.
4. ГОСТ Р 52633.0-2006. Защита информации. Техника защиты информации. Требования к средствам высоконадежной биометрической аутентификации. – М., 2006.
5. ГОСТ Р 52633.5-2011. Защита информации. Техника защиты информации. Автоматическое обучение нейросетевых преобразователей биометрия-код доступа. – М., 2011.
6. Техническая спецификация (проект, публичное обсуждение начато с 01.02.2017 членами ТК 26 «Криптографическая защита информации»). Защита нейросетевых биометрических контейнеров с использованием криптографических алгоритмов.

7. ГОСТ Р 52633.3-2011. Защита информации. Техника защиты информации. Тестирование стойкости средств высоконадежной биометрической защиты к атакам подбора. – М., 2011.
8. Малыгин, А. Ю. Быстрые алгоритмы тестирования нейросетевых механизмов биометрико-криптографической защиты информации / А. Ю. Малыгин, В. И. Волчихин, А. И. Иванов, В. А. Фунтиков. – Пенза: Изд-во ПГУ, 2006.
9. Среда моделирования «БиоНейроАвтограф» [Электронный ресурс]. – URL: http://xn–h1aanh6e.xn–p1ai/activity/science/noc.htm (дата обращения: 10.01.2018).
10. Волчихин, В. И. Компенсация методических погрешностей вычисления стандартных отклонений и коэффициентов корреляции, возникающих из-за малого объема выборок / В. И. Волчихин, А. И. Иванов, Ю. И. Серикова // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2016. – № 1 (37). – С. 103–110.
11. Волчихин, В. И. Фрактально-корреляционный функционал, используемый при поиске пар слабо зависимых биометрических данных в малых выборках / В. И. Волчихин, А. И. Иванов, Б. Б. Ахметов, Ю. И. Серикова // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2016. – № 4 (40). – С. 27–36.
12. Кулагин, В. П. Корректировка методических и случайных составляющих погрешностей вычисления коэффициентов корреляции, возникающих на малых выборках биометрических данных / В. П. Кулагин, А. И. Иванов, Ю. И. Серикова // Информационные технологии. – 2016. – Т. 22, № 9. – С. 705–710.
13. Р 50.1.037-2002 Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть I. Критерии типа хи-квадрат. Госстандарт России. – М., 2002.

Волчихин Владимир Иванович, доктор технических наук, профессор, президент, Пензенский государственный университет (Пенза, Российская Федерация), president@pnzgu.ru.

Иванов Александр Иванович, доктор технических наук, доцент, начальник, лаборатория биометрических и нейросетевых технологий, Пензенский научно-исследовательский электротехнический институт (Пенза, Российская Федерация), ivan@pniei.penza.ru.

Малыгина Елена Александровна, кандидат технических наук, научный сотрудник, межотраслевая лаборатория тестирования биометрических устройств и технологий, Пензенский государственный университет (Пенза, Российская Федерация), mal890@yandex.ru.

Куприянов Евгений Николаевич, аспирант, кафедра информационная безопасность систем и технологий, Пензенский государственный университет (Пенза, Российская Федерация),ibst@pnzgu.ru.

Серикова Юлия Игоревна, аспирант, кафедра вычислительной техники, Пензенский государственный университет (Пенза, Российская Федерация), juliaska@yandex.ru.

Поступила в редакцию 24 января 2019 г.