UDC 004:37

Oleksandr Kolgatin
Kharkiv National Pedagogical University named after G.S. Skovoroda, Kharkiv, Ukraine

COMPUTER-BASED SIMULATION OF STOCHASTIC PROCESS FOR INVESTIGATION OF EFFICIENCY OF STATISTICAL HYPOTHESIS TESTING IN PEDAGOGICAL RESEARCH

DOI: 10.14308/ite000582

Efficiency of Chi-square criterion and Fisher's angular transformation for statistical hypothesis testing are investigated at small samples with computer-based model. The results show that both criteria give satisfactory accuracy even at small samples. Accuracy zones of determination of the false positive rate (type I error) are analysed on the base of stochastic computational experiments. Comparison of sensitivities of investigated tests is shown as a function of samples size for some cases.

Keywords. *Stochastic process, statistical hypothesis, small samples, Chi-square criterion, Fisher's angular transformation.*

1. Introduction

The use of information and communication technologies radically transforms all spheres of the education system, and pedagogical research is not an exception [1]. Computer-based statistical analysis becomes a major part of the monitoring of the learning resources quality [2]. Measurement of significance of each element in three-subject didactic model [3] becomes possible only with use of ICT. In our work, we consider statistical processing of results of pedagogical experiment as one aspect of pedagogical research. Traditionally this problem is solved by methods of mathematical statistics on the basis of statistical hypothesis testing. Two hypotheses are put forward: the null hypothesis, which states that there are no differences between the compared random variables in the studied parameter, and the alternative hypothesis, which argues that the observed differences are caused by the impact, the study of which is the purpose of the experiment. A researcher uses some criterion that integrates the observed differences in the numeric form and calculates the probability of obtaining the same or larger differences in random process to accept one of those hypothesises. The number of participants in pedagogical studies is usually small, so we usually accept alternative hypothesis if the probability of a type I error (the probability that the observed differences are due to random factors) does not exceed 5%. The use of computer-based modelling provides a new look at the system of inductive methods of statistics, gives possibility to highlight the most powerful methods and to determine the limits of their applicability, which is particularly important in psychological and pedagogical studies, where samples are small. This work is devoted to comparison of popular classical criteria of statistical hypothesis testing: Pearson's criterion Chi-square and Fisher's angular transformation. Information and communication technologies offer new perspectives for the analysis of the boundaries of these tests application, investigation of the criteria sensitivity, development of approaches to application of criteria for small samples that, in our view, is important to improve methods of statistical data processing in pedagogic research.

2. Objectives

The analysis of publications ([4], [5], [6], [7] etc.) regarding the use of the Pearson's criterion, which is built on the statistics of Chi-square, shows that the recommendations of different authors are somewhat different, but in general reflect the fact that the replacement of the real distribution of the criterion value on the distribution Chi-square is an approximation, the accuracy of which depends on the sample size. The authors recommend using this criterion at samples, which sizes ensure not less than 5-10 measurements in every category of the frequency tables. But analysis of the errors in significance level at deferent sizes of samples is not shown. Some authors recommend the Fisher's exact test ([7], [8]) and Fisher's angular transformation [5] as an alternative to Pearson's test for frequency tables of two categories. So the question of sensitivity of these tests is relevant.

3. Model and Algorithm

This analysis has been done with Chi-square criterion in the form:

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(E_{i,j}-T_{i,j}\right)^2}{T_{i,j}}, \qquad (1)$$

where $E_{i,j}, T_{i,j}$ – empirical and theoretical frequencies; *i, m* – index and number of categories; *j, k* – index and number of samples (*k* = 2 in this study).

The form of Chi-square criterion with Yates's correction for continuity was analysed in our paper [9] and here is not used. All studies in this work were carried out for significance level of 5%, the critical values of Chi-square criterion are assumed 3.841 for two categories and 5.991 for three categories.

The criterion of Fisher's angular transformation are used in the form:

$$\varphi^* = 2\cdot\left|\arcsin\left(\frac{E_{1,1}}{n_1}\right)-\arcsin\left(\frac{E_{1,2}}{n_2}\right)\right|\sqrt{\frac{n_1 n_2}{n_1+n_2}}, \qquad (2)$$

where $E_{1,1}$ and $E_{1,2}$ – frequencies in one of the categories for samples 1 and 2; n_1 and n_2 – size of samples 1 and 2. The critical value of this criterion is assumed 1.64 at significance level of 5% [5, pp. 160–162].

The method of computational stochastic testing is used for investigation of sensitivity and specificity of χ^2 and φ^* criteria. A simple model [9] is prepared for calculation experiments: two series of numbers are created on the base of the same random number generator; the values obtained are distributed into m categories, we can control distribution, to ensure uniform distribution or the predominance of frequencies in certain categories; an empirical value of criterion is calculated for obtained frequencies tables and compared with the critical value of this criterion at the specified level of significance; decision about the possibility of rejection of the null hypothesis is made. We know that actually the null hypothesis is true, because both samples (series of numbers) are generated with one random number generator. But the alternative hypothesis will be accepted in some of the tests as the result of random factors. The relative frequency of such false decisions is estimated as the probability of a type I error and should correspond to the significance level that was used to choose critical value of a criterion.

We need a large number of trials to obtain a satisfactory precision of the analysis. 1000000 trials were conducted in computational experiments for each case. The precision of the obtained values of the probability of a type I error was estimated on the base of the standard deviation in consecutive identical trials. The estimated absolute error is about 0.0005 for 95% confidence interval. In some of the trials with very small samples were obtained zero values of the frequencies in some categories, and it was not possible to calculate the values of a criterion. These results were removed from the analysis, and, if their part in the total number of trials exceeded 1%, the study under appropriate conditions was not conducted.

Two unequal random number generators were used to analyse of criteria sensitivity. In such case we know that actually the alternative hypothesis is true, because samples (series of numbers) are generated with deferent random number generator. We can control the level of variation. The relative frequency of true positive decisions corresponds to the sensitivity of a criterion. This sensitivity is determined by the level of differences between the parameters of random number generators, which are used for samples.

4. Type I Error Estimation

A series of statistical tests was performed based on the proposed model for uniform distribution of values of the compared random variables in two and three categories. Complete enumeration of all possible combinations of sample sizes in the range from 9 to 120 was made for the three categories. The minimum sample size that was analysed is equal to 9 that is the average for 3 values in each category.

Analysis of the results of computational experiments for the case of the frequency distribution into three categories (fig. 1) gives rise to conclusions:

- decreasing the sample size, in general, reduces the precision of significance level, on which the statistic hypothesis testing is executed; the size of each sample is more crucial than the sum of the sizes of the two samples;
- error of estimating the probability of a type I error depends on the samples sizes not monotonically due to the discrete nature of the frequencies being compared;
- in the studied range of sample sizes (from 9 to 120) the maximum value of the probability of a type I error amounted α = 0,058 (for example, n_1 = 15 and n_2 = 12), the minimum α = 0,044 (for example, n_1 = 9 and n_2 = 120), thus, for very small sample sizes (but not less than 9) determination of significance level with application of the Chi-square test provides accuracy not worse than 16 %, with deviations either higher or lower.

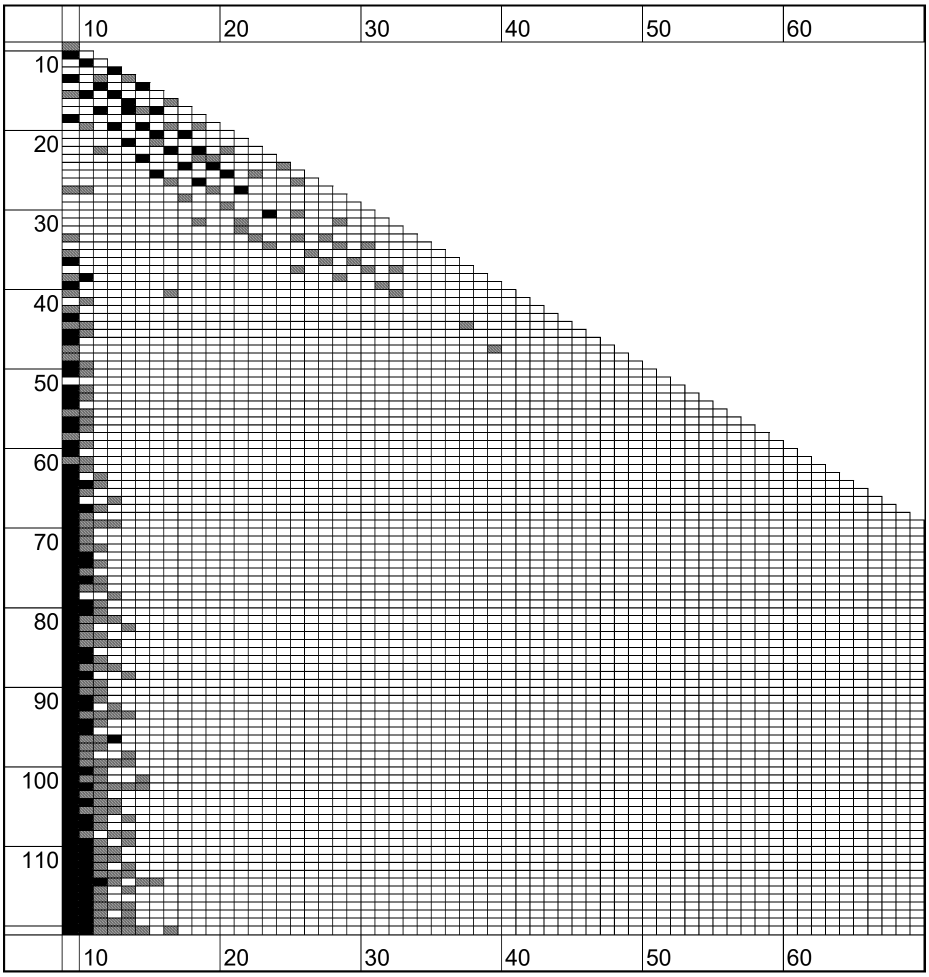

Fig. 1. Map of precision of type I error estimation in Chi-square test for 3 categories: full white cells – good precision, estimation in boundaries 4.75 ... 5.25; hatched cells – estimation in boundaries 4.65 ... 4.75 or 5.25 ... 5.35; full black cells – bad precision, estimation in boundaries 4.35 ... 4.65 or 5.35 ... 5.85.

Dependence of the precision of type I error estimation in Chi-square testes with samples of equal sizes n_1 =n_2 are shown (Fig 2) to demonstrate features of this criterion. Due to the complex behaviour of the dependence of the Type I of the Chi-square criterion on the size of samples, let the reader determines the acceptable limits of application of this criterion to compare two random variables, distributed into three categories, guided by the diagram (Fig. 1).

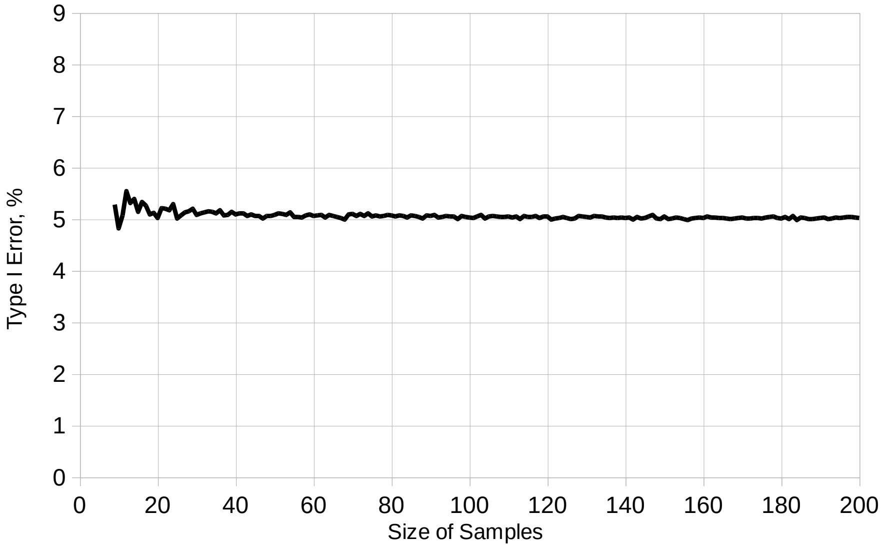

Fig. 2. Precision of type I error estimation in Chi-square tests for 3 categories at samples of equal sizes n_1=n_2=9 ... 200.

Both Chi-square and Fisher's angular transformation criteria can be used in case of two categories in frequencies tables. But the type I error (Fig. 3) is not so stable as it was observed at 3 categories in frequencies tables. Such instability reflects on accuracy of providing of significance level.

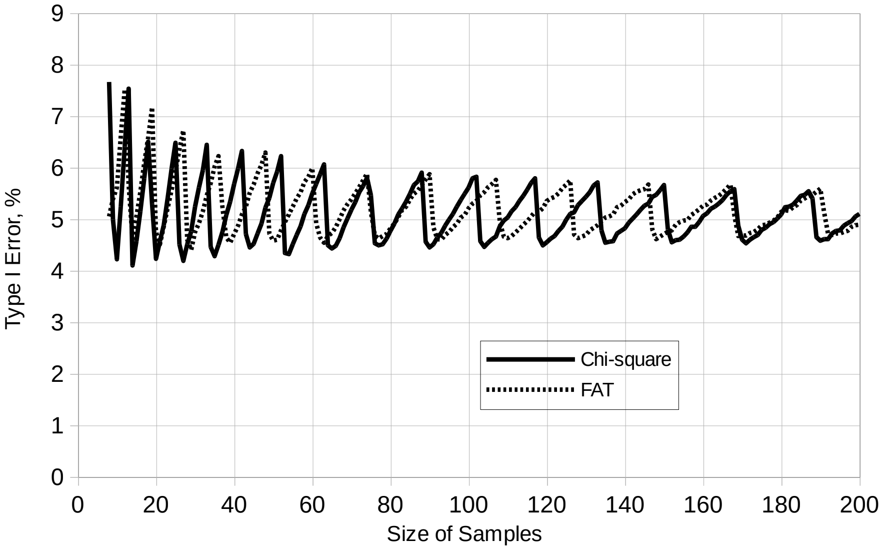

Fig. 3. Precision of type I error estimation in Chi-square and Fisher's Angular Transformation tests for 2 categories at samples of equal sizes n_1=n_2=8 ... 200.

5. Criteria Sensitivity

Sensitivity of statistical hypothesis test is determined by three factors: features of the criterion, significance level of the test as well as real differences between the compared distributions. The significance level is adopted 0.05 in this study as it was underlined above. So we should vary the distributions of probabilities in compared populations, from which two random samples are generated to carry out the statistical test. We use two independence random number generators with uniform distribution of values in such intervals: the first population from (0-d) to (1-d) and the second – from (0+d) to (1+d). Series of “measurements” distributed into 3 categories form the samples for executing the statistical hypothesis tests (Fig. 4).

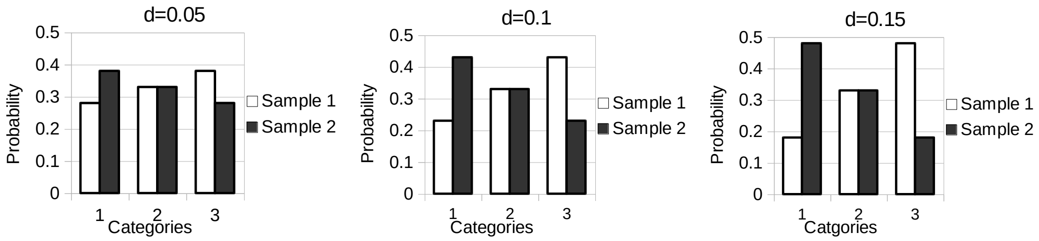

Fig. 1. Distribution of probabilities by 3 categories in random number generators for samples 1 and 2, being compared.

As a result of the computational experiments one can see that sensitivity of Chi-square test is low at small samples (Fig. 5). We need more than 50 “measurements” to provide the type II error less 5% even for samples that very differ (d = 0.15). Sensitivity to small differences (d=0.05) stays lower, than 70%, in all diapason of computational experiments. Distribution of “measurements” into 4, 5 or 6 categories leads to small increasing of Chi-square criterion sensitivity. We observed up to 18 % increasing in case of random number generators of our study. But this effect is determined by the form of probability distributions in compared populations. One can estimate how the number of categories influences on the Chi-square criterion sensitivity, using the formula (1).

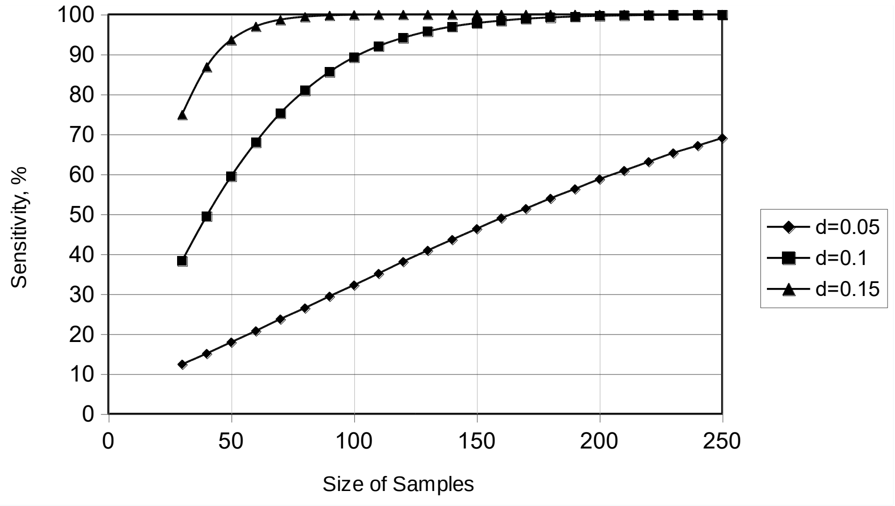

Fig. 2. Sensitivity of Chi-square test with frequencies tables of 3 categories.

We can compare the sensitivity of Chi-square and Fisher’s angular transformation (FAT) criteria, if the “measurements” are distributed into 2 categories (Fig. 6). The series of “measurements” are generated by the same random number generators as it was describe above, which are deferent for samples 1 and 2.

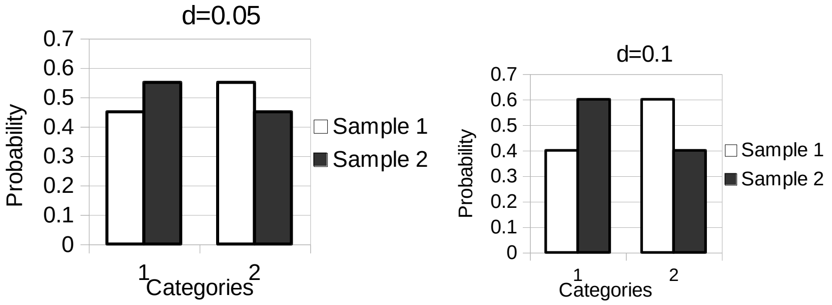

Fig. 6. Distribution of probabilities by 2 categories in random number generators for samples 1 and 2, being compared.

The results show that sensitivity of Fisher's angular transformation (FAT) criterion is higher than Chi-square criterion sensitivity at 2 categories in all diapason of computational experiments (Fig. 7). So it should be concluded that Chi-square criterion has no advantages in using for statistical hypothesis testing, if the measurements are grouped into 2 categories. But Chi-square criterion can be more effective at using distribution into several categories.

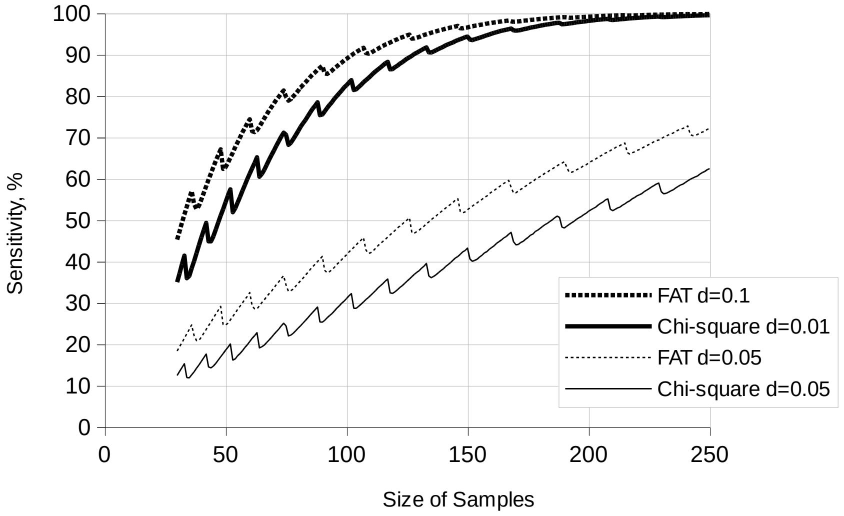

Fig. 7. Sensitivity of Chi-square and Fisher's Angular Transformation tests with frequencies tables of 2 categories.

6. Conclusions

Computational model for investigating the efficiency of statistical hypothesis testing is supposed. This model does not use any assumptions about probability distribution and test features. So it can be used for comparison of methods built on different principles.

Accuracy of the type I error estimation with Chi-square and Fisher's angular transformation criteria are investigated at small samples in computational experiments. Determination of significance level (α = 0,05) with application of the Chi-square test in the studied range of sample sizes (from 9 to 120) at frequencies tables with 3 categories provides accuracy not worse than 16 %, with deviations either higher or lower. Accuracy is not worse than 5% (from 0.0475 to 0.525), if sizes of both samples grater than 48.

Determination of significance level with application of the Chi-square test and Fisher's angular transformation test in the studied range of sample sizes (from 9 to 200) is in the interval from 0.04 to 0.08 instead of 0.05 that should be guaranteed by the test. Precision of the type I error estimation is better (in the interval from 0.04 to 0.06) if the sizes of samples are not less than 70.

Investigation of sensitivity of Chi-square test at some models of probability distributions in comprised populations shows that sensitivity is not high, it increases, if samples sizes are grater, and if we use more categories in frequencies tables.

Comparison of sensitivities of Chi-square and Fisher's angular transformation tests show that Chi-square criterion has no advantages in using for statistical hypothesis testing, if the measurements are grouped into 2 categories.

The perspectives of continuation of this study we see in investigation of precision and sensitivity of other criteria for statistical hypothesis testing, which are in use in pedagogical researches.

REFERENCES

1. Monaco, A. F.; Bird, E. M.: Electronic Scientific-Educational Space and the Prospects of their Development in the Context of Supporting the Mass and Continuity. Control Systems and Machines, N 4, 83-92 (2012)
2. Kravtsov, H.: Methods and Technologies for the Quality Monitoring of Electronic Educational Resources [Electronic Resource]. Proc. 11-th Int. Conf. ICTERI 2015, Lviv, Ukraine, May 14-16, 2015, CEUR-WS.org/Vol-1356, ISSN 1613-0073, P.311-325. http://ceur-ws.org/Vol-1356/paper_109.pdf
3. Spivakovskiy A., Petukhova L., Spivakovska E., Kotkova V., Kravtsov H.: Three-Subjective Didactic Model. Information and Communication Technologies in Education, Research, and Industrial Applications. Communications in Computer and Information Science. Springer International Publishing, V 412, 252-273, (2013)
4. Grabar, M. I.; Krasnyanskaya, K. A.: Application of Mathematical Statistics in Educational Research. Nonparametric Methods. Pedagogika, Moscow, (1977)
5. Sidorenko, E. V.: Methods of Mathematical Processing in Psychology. "Speech", St. Petersburg, (2002)
6. Godfrua, J.: What is Psychology. T. 2. Mir, Moscow, (1992)
7. Gubler, E. V.; Genkin, A. A.: Use of Non-parametric Criteria of Statistics in Biomedical Research. Medicine, Leningrad, (1973)
8. Langsrud, Øyvind: Fisher's Exact Test [Electronic Resource]. http://www.langsrud.com /fisher.htm
9. Kolgatin, O. G.: Information Technology in Educational Research. Control Systems and Machines, 255, N1, 66-72, (2015)

Стаття надійшла до редакції 28.05.16

Колгатін О. Г.

Харківський національний педагогічний університет імені Г. С. Сковороди, Харків, Україна

КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ СТОХАСТИЧНИХ ПРОЦЕСІВ ДЛЯ ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ СТАТИСТИЧНОЇ ГІПОТЕЗИ ТЕСТУВАННЯ У ПЕДАГОГІЧНИХ ДОСЛІДЖЕННЯХ

Ефективність критерію Хі-квадрат і тригонометричне перетворення Фішера для статистичної перевірки гіпотез розглядаються на малих зразках на основі комп'ютерної моделі. Результати показують, що обидва критерії дають задовільну точність навіть при невеликих зразках. Точність зони визначення помилкового результату наукового дослідження (тип I помилки) аналізуються на основі стохастичних обчислювальних експериментів. Порівняння чутливості досліджуваних випробувань показана залежно від розміру зразків для деяких випадків.

Ключові слова: стохастичний процес, статистична гіпотеза, невеликі зразки, критерій хі-квадрат, тригонометричне перетворення Фішера.

Колгатин А. Г.

Харьковский национальный педагогический университет имени Г. С. Сковороды, Харьков, Украина

КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ СТОХАСТИЧЕСКОГО ПРОЦЕССА ДЛЯ ИССЛЕДОВАНИЯ ЭФФЕКТИВНОСТИ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ ТЕСТИРОВАНИЯ В ПЕДАГОГИЧЕСКИХ ИССЛЕДОВАНИЙ

Эффективность критерия Хи-квадрат и тригонометрическое преобразование Фишера для статистической проверки гипотез расматриваются на малых образцах на основе компьютерной модели. Результаты показывают, что оба критерия дают удовлетворительную точность даже при небольших образцах. Точность зоны определения ошибочного результата научного исследования (тип I ошибки) анализируются на основе стохастических вычислительных экспериментов. Сравнение чувствительности исследуемых испытаний показана в зависимости от размера образцов для некоторых случаев.

Ключевые слова: стохастический процесс, статистическая гипотеза, небольшие образцы, критерий хи-квадрат, тригонометрическое преобразование Фишера.