Stochastic process computational modeling for learning research

Oleksandr H. Kolgatin[1][0000−0001−8423−2359],
Larisa S. Kolgatina[2][0000−0003−2650−8921], and
Nadiia S. Ponomareva[3,4][0000−0001−9840−7287]

[1] Simon Kuznets Kharkiv National University of Economics,
9A Science Ave., Kharkiv, 61166, Ukraine
kolgatin@ukr.net
http://www.is.hneu.edu.ua/?q=node/294
[2] H. S. Skovoroda Kharkiv National Pedagogical University,
29 Alchevskyh Str., Kharkiv, 61002, Ukraine
LaraKL@ukr.net
http://hnpu.edu.ua/uk/kolgatina-larysa-sergiyivna
[3] Kharkiv University of Technology "STEP",
9/11 Malomyasnytska Str, Kharkiv, 61000, Ukraine
ponomareva.itstep@gmail.com
https://tinyurl.com/5xc89ntp
[4] Kryvyi Rih State Pedagogical University,
54 Gagarin Ave., Kryvyi Rih, 50086, Ukraine

Abstract. The goal of our research was to compare and systematize several approaches to non-parametric null hypothesis significance testing using computer-based statistical modeling. For teaching purposes, a statistical model for simulation of null hypothesis significance testing was created. The results were analyzed using Fisher's angular transformation, Chi-square, Mann-Whitney, and Fisher's exact tests. Appropriate software was created, allowing us to recommend new illustrative materials for expressing the limitations of the tests that were examined. Learning investigations as a technique of comprehending inductive statistics has been proposed, based on the fact that modern personal computers can run simulations in a reasonable amount of time with great precision. The collected results revealed that the most often used non-parametric tests for small samples have low power. Traditional null hypothesis significance testing does not allow students to analyze test power because the true differences between samples are unknown. As a result, in Ukrainian statistical education, including PhD studies, the emphasis must shift away from null hypothesis significance testing and toward statistical modeling as a modern and practical approach of establishing scientific hypotheses. This finding is supported by scientific papers and the American Statistical Association's recommendation.

Keywords: computational modeling · computer-based simulation · statistical hypothesis significance testing · education · learning research

1 Introduction

1.1 Statement of the problem

Computational modeling and using the computer-based models for simulation is an essential part of educational content and methodology. From one hand, computer-based simulation becomes one of the main method of pedagogical research that brings new facilities in forecast and proving the efficiency of new pedagogical technologies. From the other hand computational modeling and computer-based simulations provided by students give them new experience in difficult sided of educational content, facilitate competences in independent work and researcher's competences [17]. Thus, students' learning research with computational simulations has been developed as a method for improving students' self management through creative learning activity [3, 4]. Semerikov et al. [27, 28] have suggested to use computer simulation of neural networks using spreadsheets. These results give us possibility to introduce this modern technology in educational process of wide kinds of educational programs that are not directly connected with computer science. The elements of technique of using CoCalc at studying topic "Neural network and pattern recognition" of the special course "Foundations of Mathematic Informatics" are shown by Markova et al. [21]. The method of computational simulation and modeling is supported by work of Modlo and Semerikov [24], where new tools for modeling of electromechanical technical objects in cloud-based learning environment have been suggested. Khazina et al. [14] also considered computer modeling as a scientific means of training. So we can conclude that computational modeling and simulation is the popular and actual learning method. This field of research is very interesting for our present study, because it promotes development of computational modeling as a method of learning researches.

One of such fields of pedagogical investigations, where new information technologies can provide new level of understanding the modelled processes, is comparative pedagogical experiment and statistical hypothesis testing as a part of one. Computer-based statistical analysis becomes a major part of the monitoring of the learning resources quality [18]. 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 compared random variables in the studied parameter, and the alternative hypothesis, which argues that the observed differences are caused by the studied impact. 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 accept alternative hypothesis, if the probability of the type I error (the probability that the observed differences are due to random factors) does not exceed 5 %. Understanding the essence of statistical hypothesis testing is a hard problem. Thus, Castro Sotos et al. [6] pointed the

common misconceptions about statistical inference. They noted that in response to the persistence of the misconceptions, educational researchers and practitioners have initiated and promoted a thorough reform for teaching statistics. And one direction of this reform was the importance of integrating technology in the statistics classroom, using simulations to help students understand the ideas behind statistical processes [6]. Problems of using statistical hypothesis tests are so deep that discussions continue even now, after more than a hundred years after implementation this approach into the science. For example, scientists discuss the problem of dichotomization of p-values, because it makes matters worse (Wasserstein et al. [33]) and suggest to describe the data using other approaches (McShane et al. [23]). The use of computer-based modeling 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 [15].

Otherwise, the practice of statistical data analysis in Ukrainian pedagogical researches is grounded on traditional approach. So we need to show educational community modern computer-based techniques for data analysis that are based on simulation of stochastic processes. We also need to show comparison of these techniques with traditional criteria for null hypothesis significance testing. This work is devoted to simulation of using popular classical criteria of statistical hypothesis testing: Pearson's criterion Chi-square, Fisher's angular transformation and Mann-Whitney U criterion. 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 statistical analysis for small samples. Learning researches with appropriate models will be useful not only for students, but also for researchers to improve understanding the essence of statistical data processing in pedagogic research.

1.2 Analysis of previous research

Last time researchers pay great attention for statistical modeling as an alternative approach to prove research hypothesises. Computer-based simulation has provided possibility to show the boundaries of using Pearson's criterion Chi-square at null hypothesis significance testing (Kolgatin [15]). Computational model for investigating the efficiency of statistical hypothesis testing was proposed. This model did not use any assumptions about probability distribution and test features. So it could be used for comparison of methods built on different principles. There was shown that Chi-square test and Fisher's angular transformation test in the studied range of sample sizes (from 9 to 200) do not provide good accuracy for frequency tables with 2 categories. The idea of these tests is to guarantee that we should obtain the error of the first type (type I error) in 5 % cases (5 % significance level were used), if the null hypothesis is true. Real value of the type I error was essentially differ within the interval from 0.04 to 0.08 instead of 0.05. Accuracy 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. Accuracy of the type I error estimation by Chi-square test for frequency tables with 3 categories is better even for very small samples sizes. This accuracy essentially depends on the number of measures in samples and is worse, when one of the sample is small and the other one is large. Therefore, some recommendations for combining categories to use the chi-square test for small sample sizes are debatable. Another result that was obtained in [15] is devoted to Chi-square test power, its ability to show differences between distributions. The type II error was quite high, it decreased with increasing the sample sizes, when 3 categories in frequencies tables were used instead of 2 categories. This question is discussed in this paper later in detail, but we can conclude here that this results correlate with the statement of the American Statistical Association (ASA) about the limitations of p values (Wasserstein and Lazar [32]).

Statistical modeling as a powerful alternative to null hypothesis significance testing was described by Lang et al. [19]. They noted that statistical modeling is a more complicated approach than null hypothesis significance testing, but this added complexity affords researchers the opportunity to quantify evidence in support of specific substantive hypotheses relative to competing hypotheses — not simply against a null hypothesis [19]. The authors underlined that the purpose of statistical modeling is to represent, as accurately and completely as possible, a data generation process, with the goal of understanding and gathering evidence about its structure [19]. These authors suggested and compared Bayesian and "frequentist" models for exploring how child temperament mediates the relationship between age and developmental progress in communication and motor skills [19].

Statistical modeling as an educational tool is analysed in many scientific works. A good review on the corresponded literature was suggested by Jamie [12]. The main idea of the authors is to use computer simulation methods (CSMs) for the purpose of clarifying abstract and difficult concepts and theorems of statistics. Some systems of computer mathematics and spreadsheets are considered: SAS PROC IML, Excel, MINITAB, SAS, SPSS. There was analysed the approaches to teach and illustrate such parts of statistical education: central limit theorem, Student's t-distribution, confidence intervals, binomial distribution, regression analysis, sampling distribution, survey sampling [12].

Many of the models for modeling the statistical hypothesis testing with educational purpose was suggested at the end of last century. Flusser and Hanna [9] have used BASIC computer programs to simulate a binomial experiment and test a simple statistical hypothesis. Taylor and Bosch [30] suggested interactive clinical trial simulation program that provides a few thousand simulations in about 5 minutes. Bradley et al. [5] have developed a comprehensive simulation laboratory for statistics that could work with real experimental data from database and generate samples according to given parameters. This software calculated p-value according F-test. Students could see that the decisions about the null hypothesis differ for various series and analyse the Type I and Type II errors. Ricketts and Berry [26] used statistical modeling in a package Resampling

Stats to demonstrate a histogram of Differences between means. This results, obtained for very small samples, helped students to understand the essence of null hypothesis significance testing without any formulas.

This software gave possibility to demonstrate the Type I and Type II errors for students, but did not produce enough performance for analysing the qualities of used criteria. Therefore, it is actual to develop some computer-based model for comparison of various approach for null hypothesis significance testing and analysing boundaries of its using. Such model will be useful not only for understanding the essence of null hypothesis significance testing, but it will be also useful for understanding the limitation of the traditional null hypothesis significance testing approach and will motivate pedagogical scientists to computational modeling as a perspective method of statistical data analysis.

1.3 Objectives

We have started this work in 2014 with the objectives to use computer-based statistical modeling for comparison and systematisation of various approaches to non-parametric null hypothesis significance testing. The accessible for Ukrainian students information in textbooks and handbooks was contradictory and not enough for confident and reasonable choice of the statistical method for data analysis in pedagogical researches. We have tried to develop a statistical model for providing learning researches with null hypothesis significance testing by university and postgraduate students.

But now we are finishing this work with the objectives to prove advantage of statistical modeling over null hypothesis significance testing. We are grounding on our own simulations, stormy development of information and communication technologies and newest publications in statistical scientific literature. The aim of this research is to show the limitations of classical null hypothesis significance testing and motivate students and researchers to computational modeling as an effective method of research hypothesises proving.

Such changing the objectives of our study leads to some inconsistency of this paper and deprives us of opportunities to introduce this analysis directly into educational process, because the program of statistical education should be revised taking into account obtained result. So we can suggest our results as a matter for critical thinking and developing educational programs to statistical educators.

2 Theoretical framework

Statistical modeling of various criteria for null hypothesis significance testing needs in preparing procedures of these criteria implementation. More over, some criteria, such as Pearson's Chi-square, Fisher's angular transformation, needs data in a form of frequency table. Our model generates the samples in metric scale, so the data should be collapsed into some intervals to obtain the frequency table.

Pearson's Chi-square criterion was used in the form:

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

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 by D'Agostino et al. [7], Kolgatin [16] and was not used here. All studies in this work were carried out for significance level of 5 %, the critical values of Chi-square criterion were assumed according to Verma [31].

The criterion of Fisher's angular transformation was used for 2-tails in such form:

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

where $E_{1,1}$ and $E_{1,2}$ – frequencies in one of the categories for samples 1 and 2; n_1 and n_2 – sizes of samples 1 and 2 [11]. The critical value of this criterion was assumed 1.96 at significance level of 5 % (2-tail). Mostly, Fisher's angular transformation is used for 1-sided test with critical value 1.64 [11]. The test power is higher in such case [15]. We used 2-sided test in this work to have correct comparison with Pearson's Chi-square test, which has no 1-sided form.

The words "exact test" are magical for some students and even researchers. Fisher's exact test for consistency in a 2×2 table was analysed by D'Agostino [7], Berkson [1], Liddell [20] etc. Their results were pessimistic. All these researchers believed that this test is exact only because it do not use any approximations. Theoretical basis of this test is not exact, so let try our simulations to understand and illustrate the problem. We have used 2-sided form of the Fisher's exact test criterion, which give us the p-value (probability of the Type I error) [13,25]. The probability of given observed frequencies combination can be calculated with the formula:

$$p^* = \frac{(n_1)!(n_2)!(n_a)!(n_b)!}{a_1!b_1!a_2!b_2!n!}, \tag{3}$$

where a_1, b_1, a_2, b_2 – observed frequencies in the samples A and B in the categories 1 and 2 accordantly; $n = a_1 + a_2 + b_1 + b_2$ – total number of measures; $n_1 = a_1 + b_1$ – number of measures in the category 1; $n_2 = a_2 + b_2$ – number of measures in the category 2; $n_a = a_1 + a_2$ – the size of the sample A; $n_b = b_1 + b_2$ – the size of the sample B.

The probability of random realisation of given combination and all other less probable combinations is

$$p = p^* + \sum_{\forall p_i < p^*, i \in [0; n_a]} p_i, \tag{4}$$

where

$$p_i = \frac{(n_1)!(n_2)!(n_a)!(n_b)!}{(i)!(n_1 - i)!(n_a - i)!(n_b - n_a + i)!n!}. \tag{5}$$

Mann-Whitney test and its modifications are the field of researchers' attention now and statistical modeling is the main method of comparison the efficiency of various modifications [10, 22]. The assumptions of this group of test were analysed by Fay and Proschan [8].

Mann-Whitney test was used in our work based on research by Sidorenko [29], Gubler and Genkin [11], Billiet [2] in the form

$$U = \min(U_a, U_b), \tag{6}$$

where

$$U_a = (n_a n_b) + \frac{n_a\,(n_a + 1)}{2} - T_a, \tag{7}$$

$$U_b = (n_a n_b) - U_a, \tag{8}$$

where n_a and n_b – the sizes of A and B samples accordantly; T_a – the sum of ranks in the sample A. The calculated values of Mann-Whitney criterion were compared with its critical values according the table, when both n_a and n_b were not grater than 30 [2]. Z-test for U criterion was used in cases, where at least one of the sample size was grater than 30 [2]:

$$Z = \frac{U - \frac{1}{2}n_a n_b}{\sqrt{\frac{n_a n_b (n_a + n_b + 1)}{12}}}. \tag{9}$$

3 Statistical model

The method of statistical modeling was used for investigation. The model allows to form 2 samples from one population or from different populations that have some differences in its probability distributions.

The first regime was used for Type I error investigation. Two series of numbers was created on the base of the same random number generator. The values obtained were distributed into m categories, we could control the distribution to ensure uniform distribution or the predominance of frequencies in certain categories; an empirical value of criterion was 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 was made. We knew that actually the null hypothesis was true, because both samples (series of numbers) were generated with one random number generator. But the alternative hypothesis was accepted in some of the tests as a result of random factors. The relative frequency of such false decisions was 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 needed 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 was about 0.0005 for 95 % confidence interval. So

we used all numbers with 2-3 significant digits and the last digit in all shown results is spare guard digit. We need in a guard digit for further data processing. The number of trials can be less in students' investigations for saving computational time when the power of tests are analysed. 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.

The second regime was used for investigation the power of the tests. Two unequal random number generators were used to analyse of criteria sensitivity. In such case we knew that actually the alternative hypothesis is true, because samples (series of numbers) were generated with different random number generator. We could control the level of variation. The relative frequency of true positive decisions corresponds to the criterion power that determined by the level of differences between the parameters of random number generators, which are used for samples.

4 Learning researches

4.1 Motivation with leading questions

One of the method to motivate students for learning research is to suggest them leading questions after a brief theoretical review [4]. Such questions attract students' attention to the most problem and debatable issues of the educational content. The limitation of using some theory, accuracy estimation, possible practical problems in specific cases are always the problem not only for students, but also for professional researchers. Concerning to the statistical education in using null hypothesis significance testing we would suggest students such leading questions:

- Is Type I error fixed, when null hypothesis significance testing?
- What factors affect the power of the null hypothesis significance testing?
- Should we collapse metric scale data into some intervals?
- Which tests should we use for small samples?
- What can we know about the test power when implement null hypothesis significance testing in practice?
- Can we prove that null hypothesis is true?

Students find the answers on these leading questions during independent work according to the plan of learning research. The answers can be not determined. So the process of making conclusions is creative for students.

Students should be equipped with the clear instruction for research steps and data collection. Also, some templates for conclusions should be prepared. We'll show the examples of tables and diagrams as possible results of investigations. The details of instructional materials is determined by the level of readiness of students for independent work.

Table 1. Type I error, when null hypothesis is true (example for equivalent samples sizes).

Size of the sample *A*	Size of the sample *B*	Frequency of null hypothesis rejection using the test, %					
		Pearson Chi-Square test for the number of categories			Fisher's angular transformation	Fisher's exact test	Mann-Whitney test
		2	3	5			
4	4	7.07				0.79	
5	5	6.05				2.17	3.18
6	6	5.79				0.63	4.11
7	7	5.71				1.29	3.81
8	8	7.64				2.09	4.96
9	9	4.99			9.02	2.10	4.01
10	10	4.21	4.83		4.84	1.26	4.03
data storing continue with the step given by a teacher							
198	198	5.01	5.02	5.02	5.01	3.95	5.00
199	199	5.07	5.04	5.01	5.07	4.00	4.98
200	200	5.10	5.00	5.04	5.01	4.02	4.96

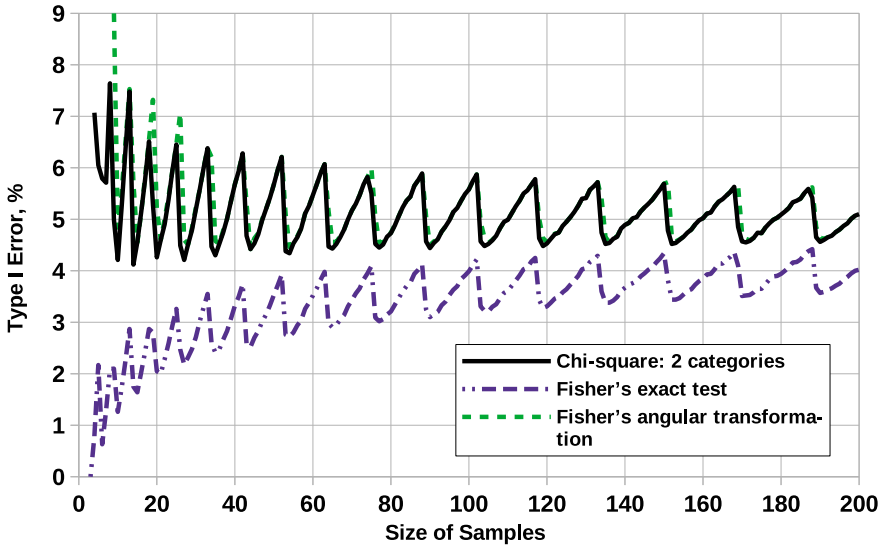

Fig. 1. Accuracy of Type I error estimation in Fisher's angular transformation, Chi-square and Fisher's exact tests for 2 categories for samples of equal sizes $n_a = n_b = 4 \ldots 200$.

4.2 Learning research of the non-parametric criteria performance with true null hypothesis

Now we will refer to students. You know some recommendations about limitations in using some null hypothesis significance tests. But how important are

Table 2. Conclusions according the accuracy of Type I error estimation with analysed tests.

Templates
Accuracy of Type I error estimation with analysed tests was ? $^{better}_{worse}$ when processing the data organised in 2 categories
Accuracy of Type I error estimation with analysed tests ? $^{disimproved}_{improved}$ with increasing the sizes of samples
The matter of the observed periodical behaviour of Type I error estimation at using Fisher's exact test for 2x2 frequency tables (2 categories) is ? $^{discrete\ matter\ of\ the\ base\ model\ of\ the\ criterion}_{low\ accuracy\ of\ statistical\ simulations}$
The matter of the observed periodical behaviour of Type I error estimation at using Fisher's angular transformation and Pearson's Chi-square test for 2x2 frequency tables (2 categories) is ? $^{aproximation\ error\ and\ discrete\ matter\ of\ the\ criteria}_{low\ accuracy\ of\ statistical\ simulations}$
Fisher's exact test was ? $^{less\ conservative}_{more\ conservative}$ than Fisher's angular transformation and Pearson's Chi-square test for 2x2 frequency tables (2 categories)
Collapsing data of small size samples into less number of categories ? $^{didn't\ lead\ to}_{led\ to}$ improving Type I error estimation
The observed behaviour ? $^{can\ differ}_{will\ be\ the\ same}$ in simulations with another probabilities distribution in population and another sizes of samples

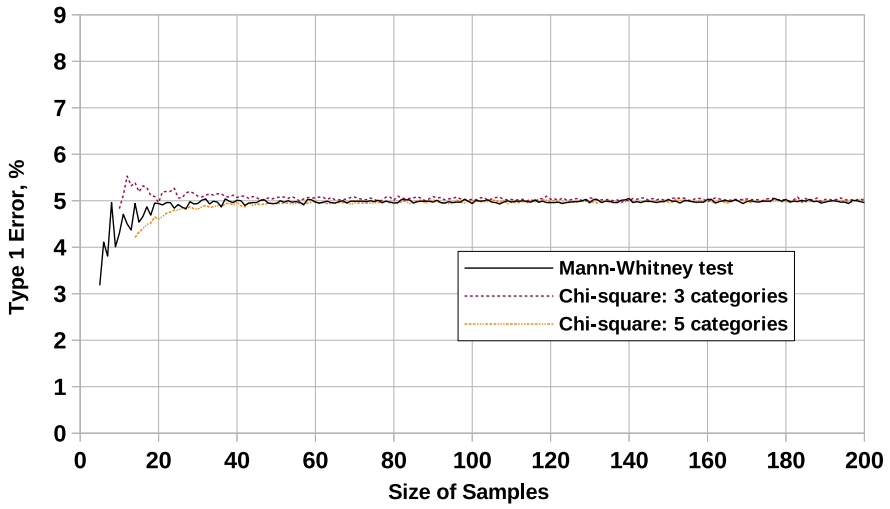

Fig. 2. Accuracy of Type I error estimation in Mann-Whitney and Chi-square tests for samples of equal sizes $n_a = n_b = 5...200$.

each of these limitations? What error can took place in each practical case of the test using. Textbooks do not give us detailed information. We can find the answers in professional research papers, but, may be, this source of information is not so easy to use. May be, some practical questions was not analysed in scientific works yet. So we need to master the method of statistical modeling

Table 3. Conclusions according the power of analysed tests with given data.

Templates
Collapsing data of small size samples into less number of categories ? $^{didn't\ lead\ to}_{led\ to}$ improving the power of null hypothesis significance testing
Power of the analysed tests ? $^{disimproved}_{improved}$ with increasing the sizes of samples asymptotically
To satisfy the Type II error less than 5 % (power grater than 95 %) at using Chi-square test for 2x2 frequency tables (2 categories) we needed samples sizes $n_a =$ _____ and $n_b =$ _____
To satisfy the Type II error less than 5 % (power grater than 95 %) at using Chi-square test for 3x2 frequency tables (3 categories) we needed samples sizes $n_a =$ _____ and $n_b =$ _____
To satisfy the Type II error less than 5 % (power grater than 95 %) at using Chi-square test for 5x2 frequency tables (5 categories) we needed samples sizes $n_a =$ _____ and $n_b =$ _____
To satisfy the Type II error less than 5 % (power grater than 95 %) at using Fisher's exact test for 2x2 frequency tables (2 categories) we needed samples sizes $n_a =$ _____ and $n_b =$ _____
To satisfy the Type II error less than 5 % (power grater than 95 %) at using Fisher's angular transformation test for 2x2 frequency tables (2 categories) we needed samples sizes $n_a =$ _____ and $n_b =$ _____
To satisfy the Type II error less than 5 % (power grater than 95 %) at using Mann-Whitney test we needed samples sizes $n_a =$ _____ and $n_b =$ _____
The analysed tests can be ranged according to their power in the such order: 1. (with the most power) _____; 2. _____; 3. _____; 4. _____; 5. _____; 6. _____
Power of some analysed tests can be improved in the case of one-sided hypothesis testing: 1. _____; 2. _____
The observed behaviour ? $^{can\ differ}_{will\ be\ the\ same}$ in simulations with another probabilities distribution in population and another sizes of samples

to explore some specific practical problems. Study the above statistical model and use it to test non-parametric criteria performance with true null hypothesis. This model tries to use Pearson's Chi-square, Fisher's angular transformation, Fisher's exact test for consistency in a 2×2 table, Mann-Whitney test for testing null hypothesis for 2 samples of some given probability distribution. Both samples are the random samples from the unique population. So, the ideal test

should decline the null hypothesis in 5 % cases (Type I error on significance level 5 %). Try your simulations for given probability distribution in population according to your individual variant and fill in the table 1. It will be useful to create this table in some spreadsheet by copying the data from the used software output. Draw diagrams according to the obtained data (see the examples on figure 1 and figure 2). Analyse your results and draw conclusions according to the templates in the table 2.

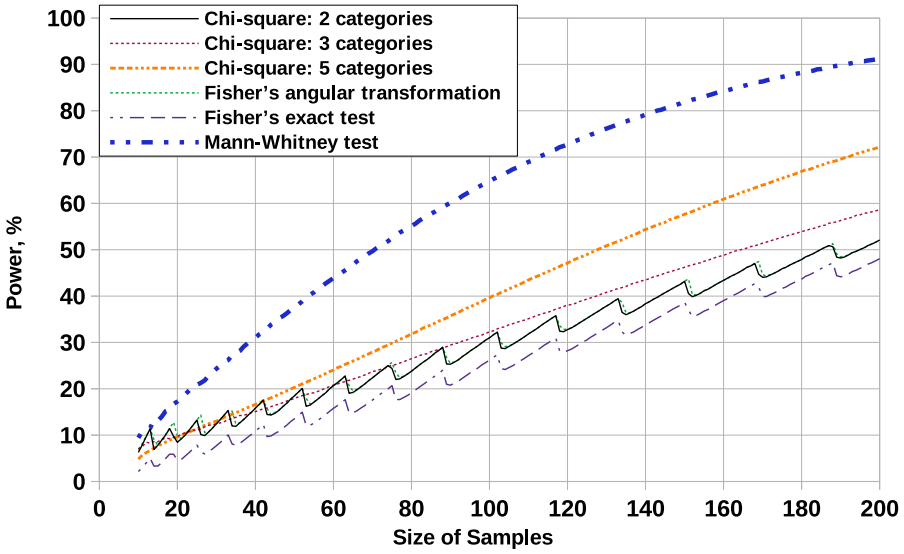

Fig. 3. Power of Fisher's angular transformation, Chi-square, Mann-Whitney and Fisher's exact tests for samples of equal sizes with uniform probability distributions in diapasons $[-0.05; 0.95]$ and $[0.05; 1.05]$.

4.3 Learning research of the non-parametric criteria power

We continue to refer to students. Let analyse the power of the tests. You remember that, according to the rules, we reject the null hypothesis, if the test allows us. But we never say that we accept the null hypothesis. We should say that we can not reject it. Now we should understand the matter of such rule. Use the null hypothesis significance test for two samples with different probability distributions obtained by different random generators with given different parameters (according to your individual variant). Store the results in the form of the table 1. The form is the same, but now we know that null hypothesis is false. So the data in the table will show the power of the tests. Organise your data using diagrams and show the theoretical probability distribution in your samples, using the information about your random generator. We have used the

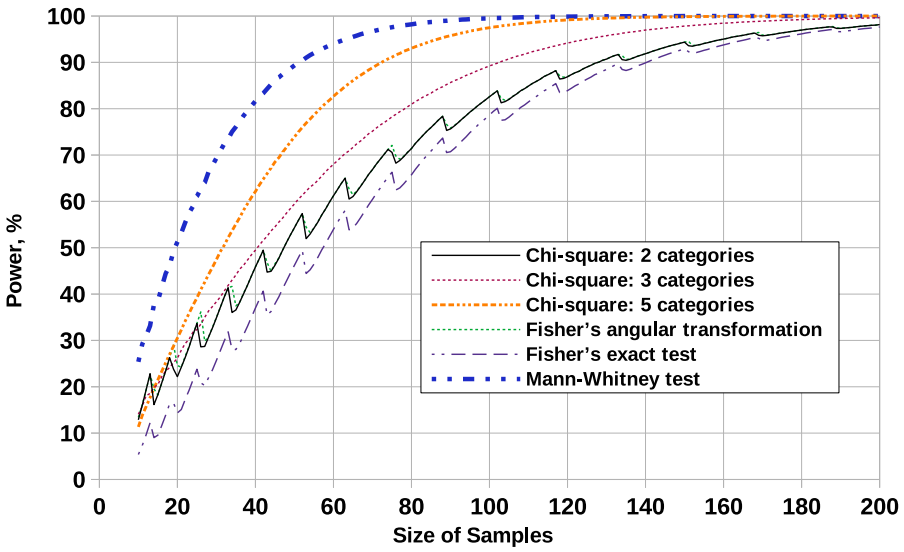

Fig. 4. Power of Fisher's angular transformation, Chi-square, Mann-Whitney and Fisher's exact tests for samples of equal sizes with uniform probability distributions in diapasons $[-0.1; 0.9]$ and $[0.1; 1.1]$.

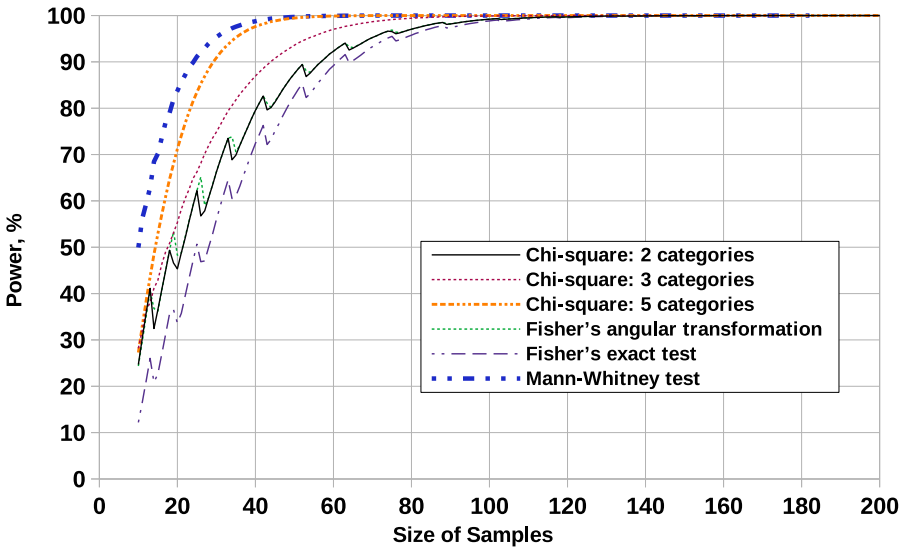

Fig. 5. Power of Fisher's angular transformation, Chi-square, Mann-Whitney and Fisher's exact tests for samples of equal sizes with uniform probability distributions in diapasons $[-0.15; 0.85]$ and $[0.15; 1.15]$.

uniform random generators with different means to obtain the examples (see figure 3, figure 4, figure 5). Analyse your results and draw conclusions according to the templates in the table 3.

As we can see, the main problem at using null hypothesis significance testing is unknown power of tests in practical tasks. The power of tests will be low, if null hypothesis is false with low differences between compared populations. The tests do not give us mechanism to estimate the power in such cases. So statistical modeling is more appropriate method of data analysing, because the model give us possibility to estimate data distribution and confidence intervals.

5 Conclusions

Statistical model for simulation of null hypothesis significance testing has been built. Fisher's angular transformation, Chi-square, Mann-Whitney and Fisher's exact tests were analysed. Appropriate software has been developed and gave us possibility to suggest new illustrative materials for describing the limitations of analysed tests.

Learning researches in inductive statistics have been suggested on the base of statistical modeling. This didactic materials can be useful for master and PhD students in pedagogics. Suggested methods contain new views on the use of null hypothesis significance testing. We stress that collapsing data into less number of categories decrease the efficiency of tests and does not give any advantage in accuracy of significance level providing.

We suggest to change the accents in Ukrainian statistical education, including PhD studies, from using null hypothesis significance testing to statistical modeling as a modern and effective method of proving the scientific hypothesises. We ground on results of our simulations suggested in this paper, possibilities of modern information and communication technologies, literature review and the opinion of American Statistical Association.

The field of further research is in developing the courseware for teaching the inductive statistics based on statistical modeling. Studying the null hypothesis significance tests should be considered as an auxiliary simplified methods.

References

1. Berkson, J.: In dispraise of the exact test: Do the marginal totals of the 2x2 table contain relevant information respecting the table proportions? Journal of Statistical Planning and Inference **2**(1), 27–42 (1978). https://doi.org/10.1016/0378-3758(78)90019-8
2. Billiet, P.: The Mann-Whitney U-test – analysis of 2-between-group data with a quantitative response variable (2003), https://psych.unl.edu/psycrs/handcomp/hcmann.PDF
3. Bilousova, L.I., Kolgatin, O.H., Kolgatina, L.S., Kuzminska, O.H.: Introspection as a condition of students' self-management in programming training. In: Proceedings of the 1st Symposium on Advances in Educational Technology - Volume 1: AET. pp. 142–153. INSTICC, SciTePress (2022). https://doi.org/10.5220/0010922000003364

4. Bilousova, L.I., Kolgatina, L.S., Kolgatin, O.H.: Computer simulation as a method of learning research in computational mathematics. CEUR Workshop Proceedings **2393**, 880–894 (2019)
5. Bradley, D.R., Hemstreet, R.L., Ziegenhagen, S.T.: A simulation laboratory for statistics. Behavior Research Methods, Instruments, and Computers **24**(2), 190–204 (1992). https://doi.org/10.3758/BF03203496, https://link.springer.com/content/pdf/10.3758/BF03203496.pdf
6. Castro Sotos, A.E., Vanhoof, S., Van den Noortgate, W., Onghena, P.: How confident are students in their misconceptions about hypothesis tests? Journal of Statistics Education **17**(2) (2009). https://doi.org/10.1080/10691898.2009.11889514
7. D'Agostino, R.B., Chase, W., Belanger, A.: The appropriateness of some common procedures for testing the equality of two independent binomial populations. The American Statistician **42**(3), 198–202 (1988), http://www.jstor.org/stable/2685002
8. Fay, M.P., Proschan, M.A.: Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Statistics Surveys **4**, 1–39 (2010). https://doi.org/10.1214/09-SS051
9. Flusser, P., Hanna, D.: Computer simulation of the testing of a statistical hypothesis. Mathematics and Computer Education **25**(2), 158 (1991), https://www.learntechlib.org/p/144840
10. Fong, Y., Huang, Y.: Modified Wilcoxon-Mann-Whitney test and power against strong null. The American Statistician **73**(1), 43–49 (2019). https://doi.org/10.1080/00031305.2017.1328375
11. Gubler, Y.V., Genkin, A.A.: Primeneniye Neparametricheskikh Metodov Statistiki v Mediko-Biologicheskikh Issledovaniyakh (Application of Nonparametric Methods of Statistics in Biomedical Research). Meditsina, Leningradskoye otdeleniye, Leningrad (1973)
12. Jamie, D.M.: Using computer simulation methods to teach statistics: A review of the literature. Journal of Statistics Education **10**(1) (2002). https://doi.org/10.1080/10691898.2002.11910548
13. Kanji, G.K.: 100 Statistical Tests. SAGE Publications, London - Thousand Oaks - New Delhi (2006)
14. Khazina, S.A., Ramskyi, Y.S., Eylon, B.S.: Computer modeling as a scientific means of training prospective physics teachers. In: EDULEARN16 Proceedings. pp. 7699–7709. 8th International Conference on Education and New Learning Technologies, IATED (4-6 July 2016). https://doi.org/10.21125/edulearn.2016.0694
15. Kolgatin, O.: Computer-based simulation of stochastic process for investigation of efficiency of statistical hypothesis testing in pedagogical research. Journal of Information Technologies in Education (ITE) (27), 007–014 (Oct 2016). https://doi.org/10.14308/ite000582, http://ite.kspu.edu/index.php/ite/article/view/101
16. Kolgatin, O.H.: Informatsionnyye tekhnologii v nauchno-pedagogicheskikh issledovaniyakh (Information technologies in educational researches). Upravlyayushchiye Sistemy i Mashiny (Control Systems and Machines) **255**(1), 66–72 (2015)
17. Kolgatin, O.H., Kolgatina, L.S., Ponomareva, N.S., Shmeltser, E.O., Uchitel, A.D.: Systematicity of students' independent work in cloud learning environment of the course "educational electronic resources for primary school" for the future teachers of primary schools. In: Proceedings of the 1st Symposium on Advances in Educational Technology - Volume 1: AET. pp. 538–549. INSTICC, SciTePress (2022). https://doi.org/10.5220/0010926000003364

18. Kravtsov, H.M.: Methods and technologies for the quality monitoring of electronic educational resources. CEUR Workshop Proceedings **1356**, 311–325 (2015)
19. Lang, K.M., Sweet, S.J., Grandfield, E.M.: Getting beyond the Null: Statistical Modeling as an Alternative Framework for Inference in Developmental Science. Research in Human Development **14**(4), 287–304 (2017). https://doi.org/10.1080/15427609.2017.1371567
20. Liddell, D.: Practical tests of 2 × 2 contingency tables. Journal of the Royal Statistical Society. Series D (The Statistician) **25**(4), 295–304 (1976). https://doi.org/10.2307/2988087
21. Markova, O., Semerikov, S., Popel, M.: CoCalc as a learning tool for neural network simulation in the special course "Foundations of mathematic informatics". CEUR Workshop Proceedings **2104**, 388–403 (2018)
22. Marx, A., Backes, C., Meese, E., Lenhof, H.P., Keller, A.: EDISON-WMW: Exact dynamic programing solution of the Wilcoxon-Mann-Whitney test. Genomics, Proteomics and Bioinformatics **14**(1), 55–61 (2016). https://doi.org/10.1016/j.gpb.2015.11.004
23. McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L.: Abandon Statistical Significance. The American Statistician **73**(sup1), 235–245 (2019). https://doi.org/10.1080/00031305.2018.1527253
24. Modlo, Y.O., Semerikov, S.O.: Xcos on Web as a promising learning tool for Bachelor's of Electromechanics modeling of technical objects. CEUR Workshop Proceedings **2168**, 34–41 (2018)
25. Preacher, K.J.: Calculation for Fisher's exact test (2021), http://quantpsy.org/fisher/fisher.html
26. Ricketts, C., Berry, J.: Teaching statistics through resampling. Teaching Statistics **16**(2), 41–44 (1994). https://doi.org/10.1111/j.1467-9639.1994.tb00685.x
27. Semerikov, S.O., Teplytskyi, I.O., Yechkalo, Y.V., Kiv, A.E.: Computer Simulation of Neural Networks Using Spreadsheets: The Dawn of the Age of Camelot. CEUR Workshop Proceedings **2257**, 122–147 (2018)
28. Semerikov, S.O., Teplytskyi, I.O., Yechkalo, Y.V., Markova, O.M., Soloviev, V.N.: Computer Simulation of Neural Networks Using Spreadsheets: Dr. Anderson, Welcome Back. CEUR Workshop Proceedings **2393**, 833–848 (2019)
29. Sidorenko, Y.V.: Metody Matematicheskoy Obrabotki v Psikhologii (Methods of Mathematical Processing in Psychology). Rech, St. Petersburg (2002), https://www.sgu.ru/sites/default/files/textdocsfiles/2014/02/19/sidorenko.pdf
30. Taylor, D.W., Bosch, E.G.: CTS: A clinical trials simulator. Statistics in Medicine **9**(7), 787–801 (1990). https://doi.org/10.1002/sim.4780090708
31. Verma, J.P.: Data Analysis in Management with SPSS Software. Springer, India (2013). https://doi.org/10.1007/978-81-322-0786-3
32. Wasserstein, R.L., Lazar, N.A.: The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician **70**(2), 129–133 (2016). https://doi.org/10.1080/00031305.2016.1154108
33. Wasserstein, R.L., Schirm, A.L., Lazar, N.A.: Moving to a World Beyond "p < 0.05". The American Statistician **73**(sup1), 1–19 (2019). https://doi.org/10.1080/00031305.2019.1583913