Evaluating the Performance of Some Statistical Location Difference Tests

Artem D. Cheremukhin

Department "Mathematics and Computer Science"
Nizhny Novgorod State Engineering and Economic University
Knyaginino, Russia
ngieu.cheremuhin@yandex.ru
[0000-0003-4076-5916]

Abstract—This article shows the results of a simulation experiment evaluating the comparative effectiveness of the application of classical and new shift tests in the context of errors of the second kind using the example of the normal, lognormal, exponential, gamma and Weibull distributions. The experiment procedure is described in detail, based on the results of 100 simulations for each type of distribution; a "binary tree" algorithm is used to solve the classification problem. The result of its application made it possible to detect "areas of effectiveness" of tests depending on the parameters of the scenario. At the same time, the Wilcoxon test showed the best comparative effectiveness, the second - the modified Yuen-Dixon t-test. It is noted that in many cases, with insignificant values of the shift parameter, all tests gave an erroneous conclusion. The conclusion presents further prospects and directions for the development of this topic.

Keywords - statistical hypotheses, distribution, shift tests, simulation modeling, errors of the second kind

I. INTRODUCTION

The development of mathematical statistics, on the one hand, and exploratory data analysis, on the other hand, led to the active emergence of new tools for testing statistical hypotheses.

Today, statistical hypotheses are an important tool for data analytics. Improper work with them can lead to erroneous management decisions and company losses. That is why it is now extremely important to use accurate and effective statistical tests to test the relevant hypotheses.

This work covers only a small part of this area and focuses on shift tests.

II. MATHERIALS AND METHODS

Currently, the field of data science related to the application of statistical tests includes:

- direct application of statistical tests in the process of data analysis. Traditionally, they are actively used in medical and psychological research, but recently there has been a tendency to use statistical tests in solving machine learning problems - for example, in [1] an example of their use in the process of solving a pattern recognition problem is shown;

- a classic application, which is to check the significance of the obtained statistical indicators. With the passage of time, new ways of checking indicators in increasingly complex and difficult tasks appear - an example is work [2] on assessing the significance of the cross-correlation coefficient of time series or work [3] on comparing the effectiveness of optimization algorithms based on statistical tests;

- the use of statistical tests not as research tools, but as objects. There are more and more studies that compare the performance of various tests [4, 5, 6] to identify the “minimum optimal set” that should be used when testing a particular statistical hypothesis.

The presented work is a brief report on this topic.

Suppose that there are two independent samples of different sizes from one distribution, but the values of one sample are "shifted" relative to the other by some fixed value. Statistical tests testing a given hypothesis are called shift tests.

The null hypothesis in tests of this type is formulated as follows: the shift parameter is 0. An alternative hypothesis is most often formulated as follows: the shift parameter is not equal to 0.

The purpose of this work is to develop and test a methodology for evaluating the effectiveness of different shift tests for data subject to different distribution laws.

The following statistical tests were selected for analysis:

- Two-sample Fried-Dehling (1) test based on Hodges-Lehmann estimator.
- Modified [1] median test.

Identify applicable funding agency here. If none, delete this text box.

- Test based on M-scores [2,3,4].
- Modified Yuen-Dixon t-test [5].
- Classic Mood's two-sample test.
- The classic Wilcoxon test.

This paper considers the comparative effectiveness of tests in the context of errors of only the second kind - when an incorrect null hypothesis about the equality of the shift parameter with zero is mistakenly accepted.

For this, the following methodology of the simulation experiment was applied:

- Values are randomly selected from the uniform distribution, which are defined as the center of the initial distribution, the shift parameter and the spread parameter.
- The number of observations in two samples (from 10 to 100) is determined randomly.
- Identify applicable funding agency here. If none, delete this text box.
- Two samples are generated and all the tests described above are performed.
- The obtained p-values of the tests are evaluated. The test with the smallest p-value is chosen as the best (the smaller the p-value, the more reliably the erroneous hypothesis of zero shift equality is rejected). If all tests have shown a p-value greater than 0.05, then it is noted that all tests were mistaken at this value of the incoming indicators.

Thus, for each distribution, a table with the following variables is obtained:

- Variables of the distribution parameters and the ratio of the shift to the central value.
- Sample size variables and their relationships.
- Best test number (0 if all tests are wrong).

Further, the classification problem is solved for the obtained data through the use of the binary tree method – this allows us to formulate some statements about the areas of comparative effectiveness of tests.

Simulation modeling was conducted for 5 types of distribution: normal, lognormal, exponential, Weibull distribution, gamma distribution.

The simulation results are displayed in tables that show the corresponding constructed tree. Simulation modeling included 100 different scenarios; and the criterion for stopping less than 10 observations in its separate branch was accepted for the tree.

All calculations were performed using the robnptests package of the R language.

III.RESULTS

Consider the evaluation of the comparative effectiveness of tests for normal distribution.

The value of the average value for the samples ranged from 0.1-1000.0, the value of the shift did not exceed 1/10 of the average, the value of the standard deviation ranged from 0.1-10.0. The evaluation results are presented in the Table 1:

Table I. Comparative evaluation of tests effectiveness in the case of normal distribution

Number of split	Parameters of split		
	Condition	Best test	Percentage of scenarios with the best test selected
1	Root	Wilcox	89.0
2	Location difference < 5.14	No test	56.3
3	Location difference > 5.14	Wilcox	100.0

According to Table 1, the Wilcoxon test shows the greatest efficacy on average. In this case, if the shift value is greater than 5.14, then this test gives the best evaluation in 100% of cases, in the opposite case, all the considered tests give the wrong result.

Consider the evaluation of the comparative effectiveness of tests for the lognormal distribution.

The logarithm value of average value for the samples ranged from 0.01 to 10.0, the shift value did not exceed 1/10 of the average, the standard deviation value ranged from 0.01 to 5.0. The evaluation results are presented in Table 2:

Table II. Comparative evaluation of tests effectiveness in the case of normal distribution

Number of split	Parameters of split		
	Condition	Best test	Percentage of scenarios with the best test selected
1	Root	No test	60.0
2	SD < 2.34	No test	37.8
4	Mean < 3.49	No test	47.6
5	Mean > 3.49	Wilcox	66.7
6	Ratio Delta on Mean < 0.05	No test	63.3
7	Ratio Delta on Mean > 0.05	Wilcox	92.3
3	SD > 2.34	No test	78.1

According to Table 2, we note that on average all tests gave incorrect results. However, an area of effectiveness of the Wilcoxon test was identified - if the simultaneous logarithm of the average value is greater than 3.49 and the shift value is greater than 1/20 of the logarithm of the average value. In other cases, all the tests considered were erroneous.

Consider the evaluation of the comparative effectiveness of tests for exponential distribution.

The value of the average value for the samples ranged from 0.2-100.0, the shift value did not exceed 1/10 of the average. The results are presented in Table 3:

Table III. Comparative evaluation of tests effectiveness in the case of exponential distribution

Number of split	Parameters of split		
	Condition	**Best test**	**Percentage of scenarios with the best test selected**
1	Root	Yuen-Dixon trimmed t-test	40.0
2	Mean < 2.38	Yuen-Dixon trimmed t-test	80.0
3	Mean > 2.38	No test	65.4
4	Elements on sample < 23	No test	84.8
5	Elements on sample > 23	Yuen-Dixon trimmed t-test	40.9

According to Table 3, Yuen-Dixon trimmed t-test showed the highest efficiency on average. If the mathematical expectation is less than 2.38, then the test is recognized as the best in 80% of cases. Moreover, if the mathematical expectation is greater, then with small samples of up to 25 elements, all tests give the wrong result.

Consider the evaluation of the comparative effectiveness of tests for the Weibull distribution.

The shape parameter for the samples was changed in the range 0.01-500.0, the shift value did not exceed 1/10 of the shape value, the scale parameter value fluctuated in the range 0.1-50.0. The results are presented in Table 4.:

Table IV. Comparative evaluation of tests effectiveness in the case of weibull distribution

Number of split	Parameters of split		
	Condition	**Best test**	**Percentage of scenarios with the best test selected**
1	Root	Wilcox	57.0
2	Shape < 6.94	No test	68.7
3	Shape > 6.94	Wilcox	67.8
4	Ratio Delta on Scale < 0.015	No test	81.8
5	Ratio Delta on Scale > 0.015	Wilcox	75.4

According to Table 4, the Wilcoxon test showed the greatest effectiveness on average. In this case, if the value of the form parameter is low (less than 6.94), all tests are mistaken. Accordingly, the area of effectiveness of the Wilcoxon test is the value of the shape parameter greater than 6.94, and the ratio of the shift value to the shape parameter is greater than 0.015.

Consider the evaluation of the comparative effectiveness of tests for gamma distribution.

The shape parameter for the samples changed in the range 0.01-500.0, the shift value did not exceed 1/10 of its value, the scale parameters fluctuated in the range 0.1-50.0. The results are presented in Table 5:

Table V. Comparative evaluation of tests effectiveness in the case of gamma distribution

Number of split	Parameters of split		
	Condition	**Best test**	**Percentage of scenarios with the best test selected**
1	Root	Wilcox	73.0
2	The shape parameter for the samples changed in the range 0.01-500.0, the shift value did not exceed 1/10 of its value, the scale parameters fluctuated in the range 0.1-50.0. The results are presented in Table 5	No test	53.1
4	SD < 30.6	No test	31.7
5	SD > 30.6	No test	84.1
3	Delta > 4.64	Wilcox	100.0

According to Table 5, the Wilcoxon test shows the greatest effectiveness on average. If the shift value is greater than 4.64, then it is effective in 100% case, otherwise all tests give an error.

IV. CONCLUSION

Evaluation of the results of the conducted experiments shows the comparative superiority of the classic Wilcoxon test for 4 of the 5 considered distributions.

The results obtained will be expanded in further researches - it is necessary to consider more diverse distributions and tests, complicate the experiment by introducing more variables characterizing the sample and more iteration.

ACKNOWLEDGMENT

I want to express my gratitude to the staff of the Department of Humanities of Nizhny Novgorod State University of Engineering and Economics for their help in expression our ideas in foreign language.

REFERENCES

[1] R. Zhu, F. Zhou, W. Yang, J.-H. Xue, “Statistical hypothesis testing as a novel perspective of pooling for image quality assessment”, Signal Processing: Image Communication, 2023, vol. 114, 116942

[2] A.M. da Silva Filho, G.F. Zebende, A.P.N. de Castro, E.F. Guedes, “Statistical test for Multiple Detrended Cross-Correlation Coefficient”, Physica A, 2021, vol. 562, 125285

[3] J. Carrasco, S.Garcíaa, M.M.Rueda, S.Das, F. Herrera, “Recent trends in the use of statistical tests for comparing swarm and evolutionary computing algorithms: Practical guidelines and a critical review”, 2020, Swarm and Evolutionary Computation, 2020, vol. 54, 100655

[4] Y. Zhou, Y. Zhu, W. K. Wong, “Statistical tests for homogeneity of variance for clinical trials and recommendations”, Contemporary Clinical Trials Communications, 2023, vol. 33, 101119

[5] R. Kochana, L. Kovalchuk, O. Korchenko, N. Kuchynska, “Statistical Tests Independence Verification Methods”, 25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, Procedia Computer Science, 2021, vol. 192, pp. 2678–2688

[6] E. A. Luengo, M. B. L. Cerna, L. J. G. Villalba, J. Hernandez-Castro, “A new approach to analyze the independence of statistical tests of randomness”, Applied Mathematics and Computation, 2022, vol. 426, 127116

[7] R Fried, H. Dehling, “Robust nonparametric tests for the two-sample location problem.”, Statistical Methods & Applications, 2011, vol. 4(20), pp. 409–422.

[8] R. Fried, “On the online estimation of piecewise constant volatilities.” , Computational Statistics & Data Analysis, 2012, vol. 56(11), pp. 3080–3090.

[9] R. A. Maronna, R. H. Zamar, “Robust estimates of location and dispersion of high-dimensional datasets.”, Technometrics, 2012, vol. 44(4), pp. 307–317.

[10] B. Phipson, G. K. Smyth, “Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn.” Statistical Applications in Genetics and Molecular Biology, 2010, vol. 9(1), Article 39.

[11] K. K. Yuen, W. T. Dixon, “The approximate behaviour and performance of the two-sample trimmed t.” Biometrika, 60(2), 369–374