ISSN 1063-4541, Vestnik St. Petersburg University. Mathematics, 2016, Vol. 49, No. 3, pp. 248–255. © Allerton Press, Inc., 2016.
Original Russian Text © V.B. Melas, D.I. Sal'nikov, A.O. Gudulina, 2016, published in Vestnik Sankt-Peterburgskogo Universiteta. Seriya 1. Matematika, Mekhanika, Astronomiya, 2016, No. 3, pp. 415–423.

MATHEMATICS

Numerical Comparison of Classical and Permutation Statistical Hypothesis Testing Methods

V. B. Melas, D. I. Sal'nikov, and A. O. Gudulina

St. Petersburg State University, Universitetskaya nab. 7-9, St. Petersburg, 199034 Russia

e-mail: vbmelas@post.ru, mejibkop.ru@mail.ru, anastasia.gudulina@gmail.com

Received February 11, 2016

Abstract—The article is devoted to the classical problem of statistical hypothesis testing for the equality of two distributions. For normal distributions, Student's test is optimal in many senses. However, in practice, distributions to be compared are often not normal and, generally speaking, unknown. When nothing is known about the distributions to be compared, one usually applies the nonparametric Kolmogorov–Smirnov test to solve this problem. In the present paper, methods are considered that are based on permutations and, in recent years, have attracted interest for their simplicity, universality, and relatively high efficiency. Methods of stochastic simulation are applied to the comparative analysis of the power of a few permutation tests and classical methods (such as the Kolmogorov–Smirnov test, Student's test, and the Mann–Whitney test) for a wide class of distribution functions. Normal distributions, Cauchy distributions, and their mixtures, as well as exponential, Weibull, Fisher's, and Student's distributions are considered. It is established that, for many typical distributions, the permutation method based on the sum of the absolute values of differences is the most powerful one. The advantage of this method over other ones is especially large when one compares symmetric distributions with the same centers. Thus, this permutation method can be recommended for application in cases when the distributions to be compared are different from normal ones.

Keywords: statistical hypothesis, permutation methods.

DOI: 10.3103/S1063454116030092

1. INTRODUCTION

The problem of testing the hypotheses of equality of two distributions is a classical problem of mathematical statistics and is of great theoretical and practical interest. It is well known (see, for example, [1]) that, in the case when both distributions are normal and have identical variances, the classical Student's test has a number of optimal properties. However, in practice, the distributions to be compared are often not normal and, generally speaking, unknown. In this case, Student's test is strongly competed by nonparametric tests, an important class of which are tests based on permutations.

In this paper, we present the results of analysis of the power of several permutation tests, as well as Student's test, the Kolmogorov–Smirnov test, and the Mann–Whitney test.

2. STATEMENT OF THE PROBLEM AND THE DESCRIPTION OF PERMUTATION TESTS

Consider the classical problem of testing the null hypothesis

$$H_0 : F_1 = F_2 \tag{1}$$

against the alternative hypothesis

$$H_1 : F_1 \neq F_2, \tag{2}$$

where F_1 and F_2 are distribution functions of general form, by the results of observations

$$Y_1 = (y_{1,1}, y_{1,2}, \ldots, y_{1,n_1}), \quad Y_2 = (y_{2,1}, y_{2,2}, \ldots, y_{2,n_2}). \tag{3}$$

For simplicity of notation, we assume without loss of generality that samples are balanced, i.e., that the equalities $n_1 = n_2 = n$ hold (in the case of unbalanced samples, the arguments are very similar). Define vectors

$$Z(\pi_0) = (y_{11}, \ldots, y_{1n}, y_{21}, \ldots, y_{2n}\}, \tag{4}$$

$$Z(\pi_k) = (\tilde{y}_{11}, \ldots, \tilde{y}_{1n}, \tilde{y}_{21}, \ldots, \tilde{y}_{2n}\}, \tag{5}$$

$$\begin{aligned} \tilde{y}_{1i_l} &= y_{2j_l}, \quad \tilde{y}_{2i_l} = y_{1j_l}, \quad l = 1,\ldots,k, \\ \tilde{y}_{1j} &= y_{1j}, \quad \tilde{y}_{2j} = y_{2j}, \quad j \neq j_1,\ldots,j_k, \end{aligned} \tag{6}$$

where $\pi_k = \pi_k(s)$, $s = 1, 2, \ldots$, and $(C_n^k)^2$ are various methods of substitution of k elements from the second half for k elements from the first half. Denote by $Z = Z(\pi_0)$ the family of vectors (4), by $\bar{Y}$, the sample mean, and by Y_{med}, the median, and define criteria $K_i = K_i(Z)$, $i = 1, 2, \ldots, 6$, on the set Z:

$$K_1(Z) = (\bar{Y}_1 - \bar{Y}_2)^2, \quad K_2(Z) = \sum_{i,j=1}^{n} (X_{1i}(t) - X_{2j}(t))^2,$$

$$K_3(Z) = \frac{nK_1(Z)}{S^2(Z)}, \quad K_4(Z) = (Y_{1med} - Y_{2med})^2,$$

$$K_5(Z) = \left(\sum_{i=1}^{n} |Y_{1i} - Y_{1med}| - \sum_{i=1}^{n} |Y_{2i} - Y_{2med}| \right)^2, \quad K_6(Z) = \sum_{i,j=1}^{n} |Y_{1i} - Y_{2j}|,$$

where

$$S^2(Z) = S_1^2(Z) + S_2^2(Z),$$

$$S_1^2(Z) = \frac{1}{N} \sum_{t=1}^{N} \left(\sum_{i=1}^{n} \frac{(X_{1i}(t) - \bar{X}_1(t))^2}{n} \right),$$

$$S_2^2(Z) = \frac{1}{N} \sum_{t=1}^{N} \left(\sum_{i=1}^{n} \frac{(X_{2i}(t) - \bar{X}_2(t))^2}{n} \right).$$

For $Z = Z(\pi)$, $\pi = \pi_k(s)$, $s = 1, \ldots$, and $(C_n^k)^2$, $k = 1, 2, \ldots, n$, the functions $K_1, K_2, \ldots, K_6$ are defined by the same formulas in which $Z = Z(\pi_0)$ is replaced by $Z = Z(\pi)$. By the permutation K_i-test of the hypothesis H_0 we will mean the following algorithm.

Suppose given $r_2 = (C_n^k)^2$, where $k = n/2$, n is even, and let r_1 be the number of permutations π_k such that the inequality $K_6(Z(\pi_k)) > K_6(Z(\pi_0))$ is satisfied. Then, if $r_1/r_2 \geq \alpha$ for $K_1, \ldots, K_4, K_6$ and $r_1/r_2 \leq (1 - \alpha)$ for K_5, where α is a given significance level, the null hypothesis is not rejected. Otherwise the null hypothesis is rejected in favor of the alternative hypothesis.

In [2], for a special case of the hypothesis testing problem, criteria based on the norms of L_1 and L_2 were proposed, which provided a basis for the criteria considered. In the recent paper [3], it is shown that three permutation methods based on the norm of L_2 are equivalent to each other.

The power of criterion K_1 was studied by numerical methods in [4]. Criterion K_2 was introduced in [2]. Criterion K_3 is a natural generalization of the classical t-criterion and is analogous to the permutation criterion proposed in [5] and [6]. Criteria K_4 and K_6 were also considered in [2]. To the knowledge of the present authors, criterion K_5 is new.

As alternatives, we will consider Student's test (t.test), the Kolmogorov–Smirnov test (ks.test), and the Mann–Whitney test (wilcox.test). Student's test is considered as a test with optimal properties when comparing normal distributions with identical valiances. The Kolmogorov–Smirnov test is a nonparametric test based on a sample distribution function, whereby it is the most universal of all possible tests. The Mann–Whitney test is a nonparametric test based on ranks and, according to standard references, is the most powerful nonparametric test in the case of distributions differing only by a shift. The problem consists in the comparative analysis of the power of these tests for typical distributions F_1 and F_2. First, we

Table 1. Power of tests in the presence of a shift, $n = 30$

Distribution	F_1	F_2	K_1	K_4	K_5	K_6
Normal	(0, 1)	(0, 1)	0.045	0.049	0.045	0.046
		(0.25, 1)	0.16	0.132	0.129	0.148
		(0.5, 1)	0.475	0.385	0.372	0.446
		(0.75, 1)	0.815	0.707	0.689	0.787
		(1, 1)	0.969	0.915	0.904	0.96
Composite (95% of normal and 5% of Cauchy distributions)	(0, 1)	(0, 1)	0.056	0.055	0.054	0.055
		(0.25, 1)	0.134	0.123	0.122	0.137
		(0.5, 1)	0.376	0.352	0.345	0.409
		(0.75, 1)	0.642	0.659	0.64	0.723
		(1, 1)	0.823	0.887	0.872	0.929
Cauchy	(0, 1)	(0, 1)	0.049	0.048	0.048	0.049
		(0.5, 1)	0.074	0.218	0.223	0.122
		(1, 1)	0.129	0.611	0.643	0.36
		(1.5, 1)	0.217	0.888	0.913	0.668
		(2, 1)	0.299	0.979	0.986	0.874
Student's	(1, 0)	(1, 0)	0.049	0.05	0.05	0.049
		(1, 0.5)	0.2	0.35	0.353	0.275
		(1, 1)	0.502	0.856	0.853	0.719
		(1, 1.5)	0.698	0.984	0.987	0.922
		(1, 2)	0.794	0.999	0.999	0.975
Weibull	(1, 3)	(1, 3)	0.05	0.05	0.051	0.049
		(1, 2.5)	0.104	0.08	0.08	0.097
		(1, 2)	0.323	0.226	0.211	0.297
		(1, 1.5)	0.731	0.548	0.514	0.702
		(1, 1)	0.979	0.898	0.877	0.974

present results on the equivalence of some of the permutation tests and then compare the remaining tests by statistical modeling.

3. EQUIVALENCE OF SOME PERMUTATION TESTS

The following theorem establishes the equivalence of the three criteria, since each of them is characterized by the same power function.

Theorem 3.1. *For any distribution functions F_1 and F_2, the permutation criteria K_1, K_2, and K_3 for testing the null hypothesis H_0 defined by formula* (1) *against the alternative hypothesis H_1 defined by* (2) *are equivalent for any permutation and any arbitrarily defined significance level α.*

The proof of the theorem can be found in [3]. Now, consider two symmetric distributions with a common center. It follows from the form of the tests considered that the t-criterion and the criterion K_1 are completely useless in this case.

Theorem 3.2. *For any distribution functions F_1 and F_2 that are symmetric with respect to the same center, for testing the null hypothesis H_0 defined by formula* (1) *against the alternative hypothesis H_1 defined by* (2), *the power of the test K_1, as well as of Student's test, coincides with any arbitrarily defined significance level α.*

According to numerical experiments, the power of the Mann–Whitney test can be slightly higher than the significance level; however, it is also useless in this situation. Slightly better (but also insignificant) capabilities are exhibited in this situation by the tests K_4 and K_5. However, as is demonstrated below by

Table 2. Power of tests in the absence of a shift, $n = 30$

Distribution	F_1	F_2	K_1	K_4	K_5	K_6
Normal	(0, 1)	(0, 1)	0.048	0.047	0.045	0.046
		(0, 1.5)	0.05	0.065	0.059	0.139
		(0, 2)	0.052	0.103	0.082	0.464
		(0, 2.5)	0.053	0.155	0.111	0.795
		(0, 3)	0.053	0.202	0.136	0.944
Composite (95% of normal and 5% of Cauchy distributions)	(0, 1)	(0, 1)	0.05	0.047	0.048	0.049
		(0, 1.5)	0.05	0.063	0.057	0.125
		(0, 2)	0.05	0.103	0.083	0.406
		(0, 2.5)	0.053	0.142	0.107	0.709
		(0, 3)	0.054	0.184	0.128	0.893
Cauchy	(0, 1)	(0, 1)	0.047	0.048	0.048	0.049
		(0, 3)	0.058	0.175	0.127	0.419
		(0, 5)	0.052	0.324	0.217	0.743
		(0, 7)	0.052	0.446	0.294	0.877
		(0, 9)	0.056	0.532	0.36	0.935
Fisher's	(100, 2)	(100, 2)	0.048	0.05	0.049	0.047
		(100, 1.6)	0.084	0.064	0.06	0.085
		(100, 1.2)	0.24	0.13	0.113	0.244
		(100, 0.8)	0.643	0.357	0.302	0.654
		(100, 0.4)	0.98	0.869	0.813	0.98
Weibull	(5, 1)	(5, 1)	0.055	0.053	0.051	0.053
		(4, 1)	0.057	0.06	0.058	0.072
		(3, 1)	0.069	0.107	0.098	0.211
		(2, 1)	0.078	0.242	0.198	0.722
		(1, 1)	0.059	0.565	0.454	1

numerical examples, the criterion K_6 allows one to effectively test the hypothesis for the distributions described by this theorem.

4. COMPARISON OF THE POWERS OF PERMUTATION TESTS

Let us carry out a comparative analysis of the powers of the tests K_1, K_4, K_5, and K_6. We begin with the case when the distributions to be compared are normal. For normal distributions with identical variances, one should expect that a permutation analog of the t-criterion, i.e., the criterion K_1, turns out to be the best one. We should note at once that this conclusion is confirmed by the results of statistical modeling (see Table 1); however, the gain in power of the test K_1 over K_6 is only 3 to 5 percent. On the other hand, the criterion K_1 is absolutely useless when the distributions to be compared have identical means and differ only in variances, because in this case the power of the criterion is identically equal to the significance level α. Consider the following distributions:

(1) normal distributions with identical variances;

(2) normal distributions with identical means;

(3) composite distributions, 95% of which are normal distributions and 5% are Cauchy distributions;

(4) Cauchy distributions with identical centers;

(5) Cauchy distributions with identical width;

Table 3. Power of tests in the presence of a shift, $n = 10$

Distribution	F_1	F_2	K_6	*t.test*	*ks.test*	*wilcox.test*
Normal	(0, 1)	(0, 1)	0.052	0.05	0.013	0.045
		(0.5, 1)	0.17	0.176	0.056	0.156
		(1, 1)	0.533	0.556	0.24	0.511
		(1.5, 1)	0.866	0.886	0.579	0.857
		(2, 1)	0.982	0.988	0.862	0.98
Composite (95% of normal and 5% of Cauchy distributions)	(0, 1)	(0, 1)	0.051	0.044	0.013	0.044
		(0.5, 1)	0.154	0.146	0.045	0.143
		(1, 1)	0.481	0.447	0.211	0.452
		(1.5, 1)	0.801	0.739	0.507	0.772
		(2, 1)	0.956	0.871	0.792	0.941
Cauchy	(0, 1)	(0, 1)	0.05	0.018	0.012	0.042
		(1, 1)	0.187	0.068	0.106	0.191
		(2, 1)	0.481	0.183	0.383	0.468
		(3, 1)	0.739	0.317	0.652	0.684
		(4, 1)	0.872	0.419	0.806	0.808
Student's	(1, 0)	(1, 0)	0.05	0.018	0.012	0.044
		(1, 0.75)	0.238	0.092	0.105	0.254
		(1, 1.5)	0.602	0.266	0.454	0.689
		(1, 2.25)	0.792	0.376	0.75	0.897
		(1, 3)	0.876	0.457	0.894	0.963
Weibull	(1, 5)	(1, 5)	0.053	0.039	0.011	0.044
		(1, 4)	0.072	0.054	0.018	0.061
		(1, 3)	0.164	0.128	0.049	0.142
		(1, 2)	0.428	0.34	0.156	0.354
		(1, 1)	0.872	0.736	0.517	0.776

(6) Student's distributions with and without shift;

(7) Weibull distributions;

(8) Fisher's distributions with different parameters.

We carried out a simulation of these distributions for $n = 10$ and $n = 30$. Each experiment was repeated 10 000 times. For permutation tests, we chose 1600 random permutations (this amount was chosen on the basis of preliminary experiments). We present the results only for a sample size of $n = 30$, because, the results for $n = 10$ are quite analogous. Table 1 presents the results for the case when the distributions differ only by a shift.

Table 1 shows that, in the case of normal distributions that differ only by a shift, as well as in the case of the Weibull distribution, the best criterion is K_1; however, the criterion K_6 is inferior to it in power only by one or two percent. Other criteria are significantly inferior in power. For the Cauchy and Student's distributions, the best criterion is K_5. For the composite distribution, the criterion K_6 is significantly, by 10 percent in power, superior to other criteria in most cases.

Table 2 shows that, in the absence of a shift, all criteria, except for K_6, are absolutely useless for symmetric distributions, which can be considered as an illustration to Theorem 2. However, the criterion K_6 is effective. In the case of Fisher's distribution with different second parameter (the number of degrees of freedom of the χ-squared distribution in the denominator), the most effective criteria are K_1 and K_6, which have approximately equal powers.

Table 4. Power of tests in the presence of a shift, $n = 30$

Distribution	F_1	F_2	K_6	*t.test*	*ks.test*	*wilcox.test*
Normal	(0, 1)	(0, 1)	0.049	0.049	0.034	0.048
		(0.25, 1)	0.151	0.158	0.098	0.154
		(0.5, 1)	0.445	0.475	0.315	0.456
		(0.75, 1)	0.778	0.81	0.638	0.788
		(1, 1)	0.958	0.97	0.883	0.961
Composite (95% of normal and 5% of Cauchy distributions)	(0, 1)	(0, 1)	0.05	0.043	0.033	0.05
		(0.25, 1)	0.136	0.117	0.094	0.141
		(0.5, 1)	0.407	0.342	0.29	0.41
		(0.75, 1)	0.726	0.61	0.584	0.734
		(1, 1)	0.93	0.783	0.844	0.932
Cauchy	(0, 1)	(0, 1)	0.051	0.02	0.037	0.05
		(0.5, 1)	0.118	0.033	0.164	0.171
		(1, 1)	0.369	0.073	0.554	0.51
		(1.5, 1)	0.665	0.134	0.864	0.794
		(2, 1)	0.874	0.21	0.974	0.935
Student's	(1, 0)	(1, 0)	0.051	0.02	0.036	0.048
		(1, 0.5)	0.279	0.102	0.29	0.376
		(1, 1)	0.732	0.298	0.821	0.887
		(1, 1.5)	0.929	0.466	0.988	0.995
		(1, 2)	0.971	0.541	1	1
Weibull	(1, 3)	(1, 3)	0.057	0.054	0.038	0.056
		(1, 2.5)	0.101	0.102	0.061	0.092
		(1, 2)	0.305	0.314	0.184	0.268
		(1, 1.5)	0.7	0.716	0.483	0.619
		(1, 1)	0.973	0.975	0.878	0.937

5. COMPARISON OF THE BEST PERMUTATION AND NONPERMUTATION CRITERIA

On the basis of the analysis carried out in the previous section, we will compare the criterion K_6, which in most cases turns out to be the best among permutation criteria, with the t-criterion and the Kolmogorov–Smirnov and Mann–Whitney criteria. Again, we begin with distributions that differ by a shift.

Notice that, for small samples, which are considered in Table 3, the permutation test K_6 is superior in power to other tests. Student's criterion is especially ineffective for the Cauchy and Student's distributions. The Kolmogorov–Smirnov test is appreciably inferior to the test K_6; it is especially ineffective for the Weibull distribution. In the next table, we consider the same distributions, but now for a three-times larger sample.

In this case, the test K_6 is slightly inferior to the Mann–Whitney and Kolmogorov–Smirnov criteria in the case of the Cauchy and Student's distributions. The Kolmogorov–Smirnov test already demonstrates low efficiency for such samples—its power is close to or higher than that of K_6 except for the cases of normal or close-to-normal distributions. Student's criterion, conversely, is the most effective in the case of a normal and Weibull distributions.

For distributions without shift, we consider only samples of size $n = 30$. The results are presented in Table 5.

Table 5. Power of tests in the absence of a shift, $n = 30$

Distribution	F_1	F_2	K_6	*t.test*	*ks.test*	*wilcox.test*
Normal	(0, 1)	(0, 1)	0.051	0.052	0.034	0.05
		(0, 1.5)	0.141	0.047	0.075	0.051
		(0, 2)	0.464	0.051	0.189	0.062
		(0, 2.5)	0.796	0.05	0.349	0.066
		(0, 3)	0.95	0.048	0.513	0.066
Composite (95% of normal and 5% of Cauchy distributions)	(0, 1)	(0, 1)	0.053	0.042	0.037	0.052
		(0, 1.5)	0.126	0.043	0.068	0.054
		(0, 2)	0.41	0.046	0.166	0.061
		(0, 2.5)	0.712	0.045	0.295	0.065
		(0, 3)	0.894	0.046	0.439	0.064
Cauchy	(0, 1)	(0, 1)	0.058	0.023	0.037	0.052
		(0, 3)	0.407	0.018	0.231	0.058
		(0, 5)	0.738	0.021	0.507	0.07
		(0, 7)	0.883	0.023	0.69	0.076
		(0, 9)	0.937	0.022	0.808	0.081
Fisher's	(100, 2)	(100, 2)	0.051	0.014	0.036	0.052
		(100, 1.6)	0.087	0.022	0.043	0.058
		(100, 1.2)	0.246	0.039	0.095	0.115
		(100, 0.8)	0.645	0.051	0.318	0.297
		(100, 0.4)	0.982	0.013	0.879	0.768
Weibull	(5, 1)	(5, 1)	0.048	0.049	0.034	0.048
		(4, 1)	0.07	0.055	0.045	0.053
		(3, 1)	0.216	0.069	0.12	0.079
		(2, 1)	0.718	0.07	0.363	0.133
		(1, 1)	0.999	0.048	0.889	0.248

Table 5 shows that Student's and Mann–Whitney tests are useless in this case. The test K_6 is significantly superior to the Kolmogorov–Smirnov test, especially in the case of a normal distribution and a mixture of the normal and Cauchy distributions.

6. CONCLUSIONS

Stochastic modeling is a universal method of research that allows one to estimate the efficiency of statistical procedures in cases when this cannot be done by analytical methods. A comparative estimate of the powers of permutation tests and the classical Student's, Kolmogorov–Smirnov, and Mann–Whitney tests for solving the hypothesis testing problem for the equality of two distributions has shown that a test based on the sum of the absolute values of differences of elements of two samples in most cases is superior in power to all the other tests considered. The advantage of this test is especially large if the centers of the distributions to be compared coincide.

ACKNOWLEDGMENTS

This work was supported by St. Petersburg State University, project no. 6.38.435.2015.

REFERENCES

1. E. L. Lehmann, *Testing Statistical Hypotheses* (Wiley, New York, 1959).
2. M. Sirski, “On the statistical analysis of functional data arising from designed experiments,” PhD Thesis (Univ. of Manitoba, Manitoba, 2012).
3. L. Corain, V. B. Melas, A. Pepelyshev, and L. Salmaso, “New insights on permutation approach for hypothesis testing on functional data,” Adv. Data Anal. Classif. **8** (3), 339–356 (2014).
4. J. Sturino, I. Zorych, B. Mallick, et al., “Statistical methods for comparative phenomics using high-throughput phenotype microarrays,” Int. J. Biostat. **6**, 3–4 (2010).
5. D. Cox and J. Lee, “Pointwise testing with functional data using the Westfall-Young randomization method,” Biometrika **95**, 621–634 (2008).
6. J. Ramsay, G. Hooker, and S. Graves, *Functional Data Analysis with R and MATLAB* (Springer-Verlag, New York, 2009).
7. S. Keller-McNulty and J. J. Higgins, “Effect of tail weight and outliers on power and type-i error of robust permutation tests for location,” Commun. Stat. – Simul. Comput. **16**, 17–35 (1987).
8. E. S. Edgington, “Approximate randomization tests,” J. Psychol. **72**, 143–149 (1969).
9. P. I. Good, *Resampling Methods: A Practical Guide to Data Analysis*, 3rd ed (Birkhäuser, Boston, 2005).

Translated by I. Nikitin