Исследование мощности нового критерия проверки равенства распределений двух выборок

В.Б. Мелас[a], Д.И. Сальников[a]

[a] *Санкт-Петербургский государственный университет,*
Санкт-Петербург, Россия

Abstract

В статье представлен новый перестановочный критерий проверки равенства распределений двух выборок. С помощью стохастического моделирования показана высокая эффективность критерия. Для выборок из нормальных распределений и распределений Коши, отличающихся только сдвигом, достигаемая мощность критерия оказывается примерно равной мощности критериев Уилкоксона-Манна-Уитни и Колмогорова-Смирнова. Однако если распределения различаются параметром масштаба, мощность предлагаемого критерия значительно выше.

Keywords: Перестановочные критерии, критерии однородности, распределение Коши, нормальное распределение

1. Постановка задачи

Рассмотрим классическую задачу проверки гипотезы равенства распределений двух выборок

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

против альтернативной гипотезы

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

в случае двух независимых выборок $X = (X_1, \ldots, X_n)$ и $Y = (Y_1, \ldots, Y_m)$ с функциями распределения F_1 и F_2 соответственно.

Хорошо известно [see e.g. (Lehman,1986)], что в случае, когда оба распределения являются нормальными и различаются только параметром сдвига, классический критерий Стьюдента обладает некоторыми оптимальными свойствами. Если распределения не является нормальным, но все же отличаются только параметром сдвига, то вместо критерия Стьюдента часто используется широко известная U-статистика Вилкоксона-Манна-Уитни (WMW). Однако можно показать, что для выборок из нормальных распределений, отличающихся только параметром масштаба, мощность критерия WMW крайне мала. Если законы распределения выборок неизвестны, возможно применение некоторых универсальных методов, таких как критерии Колмогорова-Смирнова или Крамера-фон Мизеса (см. [**?**]), но во многих случаях эти критерии могут обладать невысокой мощностью.

В работе [**?**] был предложен критерий, основанный на U-статистике с логарифмическим ядром, и представлено численное исследование его мощности для одномерных и многомерных случаев в сравнении с несколькими альтернативными методами. Насколько известно авторам, нет аналитических результатов асимптотической мощности критерия. В данной работе представлен похожий, но отличный критерий, и приведены численные результаты его мощности в сравнении с широко известными критериями.

suggested the test basing on U-statistics with the logarithmic kernel and provided its numerical justification for one and many dimensional cases in comparison with a few alternative techniques. However,to the best authors knowledge there are no analytical results about its asymptotic power. Here we introduce a similar but different test and provide a few analytical results on its power.

2. The new test and its statistical motivation

Assume that the distribution functions F_1 and F_2 belongs to the class of distribution functions of random values ξ, such that

$$E[\ln(1+\xi^2)] < \infty. \tag{3}$$

Many distributions and, in particular, the Caushy distribution have this property.

Among all distributions with given parameters of shift and scale having this property the Caushy's one have the maximum entropy.

Consider the following test

$$\Phi_A = \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} g(X_i - X_j), \Phi_B = \frac{1}{m(m-1)} \sum_{1 \le i < j \le m} g(Y_i - Y_j), \tag{4}$$

$$\Phi_{AB} = -\frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} g(X_i - Y_j), \Phi_{nm} = \Phi_A + \Phi_B + \Phi_{AB}, \tag{5}$$

where

$$g(u) = -\ln(1+|u|^2)$$

is under a constant term precision the logarithm of the density of the standard Caushy distribution. (Note that Zech and Aslan (2005) took $g(u) = \ln(|u|)$).

We would like to have a test that is appropriate for two distributions that differ only by shift and scale parameters and belong to a rather general class of distributions.

In particular, we consider the class of distributions satisfying (3), but the approach can be generalized for other classes of distributions.

Consider the class of distributions given by the property (3). Note that if the parameters are known the test basing on likelihood ratio is the most powerful among tests with given parameters.

The test suggested above can be considered as an approximation of logarithm of this ratio for the Caushy distribution.

3. The analytical study of asymptotic power

Let us consider the case of two distributions having the property (3) and, in particular, the two that differ only by the shift parameters. To simplify notations assume that $m = n$. The case $m \ne n$ is similar. Now the criterion (4) - (5) assumes the form

$$T_n = \Phi_{nn} = \frac{1}{n^2} \sum_{i,j=1}^{n} \ln(1+(X_i - Y_j)^2) - \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} \ln(1+(X_i - X_j)^2) \tag{6}$$

$$-\frac{1}{n(n-1)}\sum_{1\le i<j\le n}\ln(1+(Y_i-Y_j)^2). \tag{7}$$

Denote by $C(u,v)$ the Caushy distribution with the density function

$$v/(\pi(v^2+(x-u)^2)).$$

Let $f_1(x)$ denotes the density of F_1 and $f_2(x)$ denotes the density of F_2. Denote

$$J_h = \int_R g(x-y-|h|/\sqrt{n})f_1(x)f_2(y)dxdy,$$

If there exists the limit

$$lim_{n\to\infty}n(J_h - J_0) \tag{8}$$

denote it by $J^*(h)$.

The basic analytical result of the present paper is the following

Theorem 1. *Consider the problem of testing hypothesis on the equality of two distributions (1)-(2) where both functions have the property (3). Then*
(i) under the condition $n\to\infty$ the distribution function of nT_n converges under H_0 to that of the random value

$$(aZ+b)^2, \tag{9}$$

where Z has the normal distribution with zero expectation and variance equal to 1, $a^2 = J_0/3, b = 0$.

(ii) Let $F_1 = F(\nu,\mu), F_2 = F(\nu+\theta,\mu)$, where ν and μ be the shift and scale parameters of the distribution F, F is arbitrary distribution with property (3) and $\theta = h/\sqrt{n}, h$ is an arbitrary given number. Then the distribution function of nT_n converges under H_1 to that of the random value

$$(aZ+b)^2,$$

where $a^2 = J_0/3, b = 0$ for the case of H_0 and $a^2 = J_0/3, b^2 = J^*(h)$ for H_1. In this case the power of the criterion T_n with significance α is asymptotically equal to that is given by the formula

$$Pr\{Z \ge z_{1-\alpha/2} - \sqrt{\frac{3J^*(h)}{J_0}}\} + Pr\{Z \le -z_{1-\alpha/2} - \sqrt{\frac{3J^*(h)}{J_0}}\}.$$

(iii) If $F_1 = C(\nu,1), F_2 = C(\nu+\theta)$ then in the part (ii) $a^2 = (2/3)\ln 3, b = h/3$ and

$$Pr\{Z \ge z_{1-\alpha/2} - (1/\sqrt{6}\ln 3)h\} + Pr\{Z \le -z_{1-\alpha/2} - (1/\sqrt{6}\ ln3)h\}.$$

The proof of the theorem is given in the Appendix.

4. Simulation results

We found by a stochastic simulation that the formula presents an approximation of the power of the test T_n with a good accuracy (see tables 1-3 below).

At the tables 1-12 results for cases n =100, 500, 1000 and different values of h with $\alpha = 0.05$ are given for normal and Caushy distributions that differ either by shift or by scale parameters.

The critical values were calculated in two way: by simulation of the initial distribution and by random permutations (we used 800 random permutation in all cases). It worth to be noted that the results are very similar. Since the permutation technics is more universal, it can be recommended for practical applications.

Note that in all these cases when the distributions differ only in the scale parameters the power of T_n and that of the Wilcoxon-Mann-Whitney and the Kolmogorov - Smirnov tests were approximately equal to each other. It can pointed out also that if the variances are not standard but are known we should simply make a the corresponding normalisation. But for the cases where the distributions differ in scale parameters the Wilcoxon-Mann-Whitney is not appropriate at all and the power of the Kolmogorov - Smirnov test is considerably lower.

Таблица 1: Power of tests for the Cauchy distribution,
$X \sim C(0,1)$, $Y \sim C(h/\sqrt{n},1)$,
samples size 500, 1000 iterations, 800 permutations in $T_n, perm$

h	$T_n, perm$	T_n, sim	$formula$	$wilcox.test$	$ks.test$
1	5.8	6.1	6.8	6.4	6.4
2	11.6	11.6	12.2	12.6	13.9
3	21	21.8	21.5	22.2	24.3
5	50.9	51	49.5	48	57.9
7	82.2	82.4	77.8	75.6	85.9
9	96.2	96.5	93.9	93.2	97.2

Таблица 2: Power of tests for the Cauchy distribution,
$X \sim C(0,1)$, $Y \sim C(h/\sqrt{n},1)$,
samples size 1000, 1000 iterations, 800 permutations in $T_n, perm$

h	$T_n, perm$	T_n, sim	$formula$	$wilcox.test$	$ks.test$
1	6.3	6	6.8	6.8	8.1
2	11.4	11.9	12.2	12.9	13.4
3	21	20.9	21.5	22.8	26.2
4	34.9	34.6	34.4	36.1	43
7	84	84.5	77.8	79.5	87.6
10	99	98.9	97.4	96.8	99.2

Таблица 3: Power of tests for the Cauchy distribution,
$X \sim C(0,1)$, $Y \sim C(0,1+h/\sqrt{n})$,
samples size 100, 1000 iterations, 800 permutations in $T_n, perm$

h	$T_n, perm$	T_n, sim	$wilcox.test$	$ks.test$
2	0.106	0.119	0.054	0.054
4	0.276	0.298	0.055	0.087
6	0.494	0.536	0.055	0.159
8	0.688	0.735	0.055	0.25
10	0.842	0.871	0.052	0.364

Таблица 4: Power of tests for the Cauchy distribution,
$X \sim C(0,1)$, $Y \sim C(0, 1 + h/\sqrt{n})$,
samples size 500, 1000 iterations, 800 permutations in $T_n, perm$

h	$T_n, perm$	T_n, sim	$wilcox.test$	$ks.test$
2	0.094	0.1	0.045	0.063
4	0.285	0.306	0.048	0.14
6	0.545	0.565	0.05	0.261
8	0.795	0.805	0.052	0.433
10	0.93	0.94	0.052	0.622

Таблица 5: Power of tests for the Cauchy distribution,
$X \sim C(0,1)$, $Y \sim C(0, 1 + h/\sqrt{n})$,
samples size 1000, 1000 iterations, 800 permutations in $T_n, perm$

h	$T_n, perm$	T_n, sim	$wilcox.test$	$ks.test$
2	0.102	0.105	0.05	0.076
4	0.324	0.338	0.052	0.138
6	0.611	0.628	0.052	0.279
8	0.848	0.856	0.052	0.474
10	0.961	0.971	0.054	0.679

5. Conclusion

In this paper we suggested a new test for equality of two distributions. Its asymptotic power was analytically established for the case of distributions that differ only by shift. For the case of Caushy distribution this formula was presented in a closed form since it proved to be possible to calculate the corresponding integrals analytically.

By stochastic simulation we found that for the Normal and Caushy distributions that differ only by shift the power of the new test is approximately equal to that of the Wilcoxon-Mann-Whitney and the Kolmogorov - Smirnov tests. However if the distributions differ by the scale parameter our simulations show that the new test is considerably better than the Kolmogorov - Smirnov test. And in this case the the Wilcoxon-Mann-Whitney test is not appropriate at all.

Acknowledgments

The authors are indebted to professor Yakov Nikitin for the help in calculating the integrals. Work of Viatcheslav Melas was supported by RFBR (grant N 20-01-00096).

6. Appendix

Proof of Theorem 1.

Let us begin with studying the asymptonic behaviour of the magnitude $E(nT_n)^2$.

Lemma 1. *(1) If hypothesis H_0 is satisfied and F_1 posesses property (3) then there exists a finite limit of $d = \lim E(nT_n)^2$ with $n \to \infty$.*

(2)If hypothesis H_1 is satisfied for $F_2(x) = F_1(x-\theta, \theta = h/\sqrt{n}$ and F_1 and F_2 possess property (3) then there exists a finite limit of $E(nT_n)^2$ for $n \to \infty$ and it is given by the formula

$$d + 2J^*(h)J_0 + J^*(h)^2.$$

Proof of the lemma.
Note that $(nT_n)^2$ is equal to

$$n^2[\frac{1}{n^2}\sum_{i,j=1}^{n}[g(X_i - Y_j) - J_0] - \frac{1}{n(n-1)}\sum_{i<j,i,j=1}^{n}[g(X_i - X_j) - J_0] - \frac{1}{n(n-1)}[\sum_{i<j,i,j=1}^{n}[g(Y_i - Y_j) - J_0]]^2,$$

where $g(z) = \ln(1+z^2)$.
The idea of the proof consists in the splitting the three squares of three sums/ including in this sum and three pairwise products into peculiar sums of the identical structure. Then to each peculiar sum either law of large numbers or central limit theorem is applied.

The square of the first sum,

$$n^2\{\frac{1}{n^2}\sum_{i,j=1}^{n}[g(X_i - Y_j) - J_0]\}^2,$$

can be represented by sum of the following peculiar sums.

$$1)\ n^2(\frac{1}{n^2})^2\sum_{i,j=1}^{n}[(g(X_i - Y_j) - J_1]^2,$$

$$2a)\ n^2(\frac{1}{n^2})^2\sum_{i,j=1,i\neq j}^{n}\sum_{k=1}^{n}[(g(X_i - Y_j) - J_0][(g(X_k - Y_j) - J_0],$$

$$2b)\ n^2(\frac{1}{n^2})^2\sum_{i,j=1,i\neq j}^{n}\sum_{k=1}^{n}[g(Y_k - X_i) - J_0][g(Y_k - X_j) - J_0],$$

$$3)\ n^2(\frac{1}{n^2})^2\sum_{i,j=1}^{n}\sum_{l,k=1,(l\neq k)or(i\neq j)}^{n}[g(X_i - Y_j) - J_0][g(X_l - Y_k)^2) - J_0].$$

Similar expression can be obtained for each of the two other squares and three pairwise products. Note the limit with $n \to \infty$ for the peculiar sums of the tipe 1) is finite due to the law of large numbers.

The peculiar sums of type 3) consist of multiplications of indepent terms with zero expectation. Therefore for any n the expectation of these peculiar sums is zero and the limit is also equal to 0.

Consider the peculiar sum $ES^2_{xy,2a}$. It can be written as $I_1 - I_2$,

$$I_1 = \frac{1}{n}\sum_{k=1}^{n}\{[\sum_{i=1}^{n}(g(x_k - y_i) - J_0)]/\sqrt{n}][\sum_{j=1}^{n} g(x_k - y_j) - J_0)]/\sqrt{n}]\},$$

$$I_2 = \frac{1}{n^2}\sum_{i,j=1}^{n}(g(x_i - y_j)^2.$$

Note that I_2 tends with $n \to \infty$ to a finite limit due to the law of large numers. And due to the central limit theorem under fixed X_k the random value

$$\sum_{i=1}^{n}[g(x_k - y_i) - J_1]/\sqrt{n}$$

tends to normal random value with zero expectation and a finite variance.

Others sum of type 2) have a similar behaviour.Thus under H_0 there exists a final limit

$$\lim_{n\to\infty} E(nT_n)^2.$$

Denote this limit by d.

Thus the first part of the lemma is proved.

Let now H_1 holds with $|h| > 0$. In this case only the behavior of the following sums

$$n^2\{\frac{1}{n^2}\sum_{i,j=1}^{n}[g(X_i - Y_j) - J_0]\}^2,$$

$$S_{xy,xx,2} = n^2(\frac{1}{2n(n-1)n^2})\sum_{k=1}^{n}\sum_{i,j=1,i\neq j,k,j\neq k}^{n}[(\ln(1+(X_k - Y_i)^2) - J_1][(\ln(1+(X_k - X_j)^2) - J_1],$$

$$S_{xy,xx,3} = n^2(\frac{1}{2n(n-1)n^2})\sum_{i,j=1}^{n}\sum_{k=1,l=1,l\neq k}^{n}[(\ln(1+(X_k - Y_i)^2) - J_1][(\ln(1+(X_l - X_j)^2) - J_1].$$

and $S_{xy,yy,2}, S_{xy,xy,3}$ that are determined in a similar way that $S_{xy,xx,2}, S_{xy,xx,3}$. By a direct calculation we will obtain that

$$\lim_{n\to\infty} E(nT_n)^2 = d + 2J^*(h)J_0 + J^*(h)^2.$$

Lemma is proved.

Lemma 2. *For $g(x) = x^2$ the following identity holds*

$$\Phi_{nm} = (\bar{x} - \bar{y})^2,$$

where

$$\bar{x} = (\sum_{i=1}^{n} X_i)/n, \bar{y} = (\sum_{i=1}^{m} Y_i)/m.$$

The proof follows from the known formula [see f.e.[**?**], p.296]

$$\frac{1}{n(n-1)}\sum_{1\le i<j\le n}(X_i-X_j)^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i-\bar{x})^2.$$

by direct calculations.

Assume that H_0 holds. Let C be an arbitrary positive number,

$$\tilde{X}=(\tilde{X_1},\ldots,\tilde{X_n}),\ \ \tilde{Y}=(\tilde{Y_1},\ldots,\tilde{Y_n}),$$

where $\tilde{X_i}=X_i$, if $|X_i|\le C$ and $\tilde{X_i}=C$ if $X_i>0$, $\tilde{X_i}=-C$ if $X_i<0$ otherwise. And $\tilde{Y_i}$ are determined similarly. Note that $0\le\ln(1+x^2)\le x^2$. Therefore there exists a value t that depends from $\tilde{X}$ and $\tilde{Y}$ such that

$$n\{\frac{1}{n^2}\sum_{i,j=1}^{n}\ln(1+(\tilde{X_i}-\tilde{Y_j})^2-\frac{1}{n(n-1)}\sum_{i<j}\ln(1+(\tilde{X_i}-\tilde{X_j})^2)- \tag{10}$$

$$\frac{1}{n(n-1)}\sum_{i<j}\ln(1+(\tilde{Y_i}-\tilde{Y_j})^2)\}=t(\sum_{i=1}^{n}\tilde{X_i}/\sqrt{n}-\sum_{i=1}^{n}\tilde{Y_i}/\sqrt{n})^2. \tag{11}$$

For constructing the right hand side we applied Lemma 2. Note that for distributions F_1 and F_2 satisfying (3) it follows from Lemma 1 that the variance of the left hand side is finite. Therefore the variance of the right hand side is also finite for arbitrary C. Passing to the limit with $n\to\infty$ we obtain due to the central limit theorem that the right hand side has the limit distribution of the form (9) where Z has the normal distribution with zero expectation and variance equal to 1. And its variance is equal to the variance of the left hand side of (11). Since C is arbitrary we obtain that the limiting distribution has the required form for H_0.
For determining a and b in the part (ii) of the theorem we now can use the equality

$$E((aZ+b)^2)^2=\lim_{n\to\infty}E(nT_n)^2, \tag{12}$$

that follows from (11).

Since $EZ^2=1, EZ^4=3$, we have for the left hand side (12)

$$3a^4+6a^2b^2+b^4. \tag{13}$$

The formula for the left hand side follows from Lemma 1. And the asymptotic behaviour of the power follows from the asymptotic normality of $\sqrt{n}T_n$. In order to calculate the right hand side of (12)in (iii) the following result is crucial.

Lemma 3. *If X and Y are independent random values with the distribution $C(0,1)$, then*

$$E\ln(1+(X-Y)^2)=\ln 9,\ \ E\ln(1+(X-Y-\theta)^2)-\ln 9=ln(1+\theta^2/9).$$

In order to prove this Lemma we need the following integrals

$$\int_R\frac{\ln(1+(x-y)^2)}{\pi(1+y^2)}dy=\ln(4+x^2),$$

$$\int_R \frac{\ln(4+x^2)}{\pi(1+x^2)}dx = \ln 9,$$

([**?**] 4.296.2 and 4.295.7.)

$$\int_R \frac{\ln(4+(x+\theta)^2)}{\pi(x^2+1)}dx = \ln(9+\theta^2),$$

[see [**?**], formula (2.6.14.19)]. Using these integrals we obtain

$$E\ln(1+(X-Y-\theta)^2) - \ln 9 = 2\int_R\int_R \frac{\ln(1+(x-y-\theta)^2)}{\pi^2(1+x^2)(1+y^2)}dxdy - \ln 9$$

$$= \int_R \frac{\ln(4+(y+\theta)^2)}{\pi(1+y^2)}dy - \ln 9 = \ln(9+\theta^2) - \ln 9 = \ln(1+\theta^2/9).$$

Submitting here $\theta = 0$ we obtain both formulas of the Lemma. Note that $\theta^2 = nh^2$ and

$$\lim_{n\to\infty} n\ln(1+\theta^2/9) = (1/9)h^2.$$

Therefore we obtain for the right hand side (8) with some algebra

$$3a^4 + \frac{(2\ln 9)h^2}{9} + \frac{h^4}{81}. \tag{14}$$

From Lemma 3 and (14) we obtain

$$b = \frac{1}{3}h,\ a^2 = \frac{2}{3}\ln 3.$$

The formula for the power follows from the form of the limiting distribution (9).

7. References

Lehmann E. (1986). Testing Statistical Hypotheses, Probability and Statistics Series, Wiley.

Zech, G. and Aslan, B.(2005). New test for the multivariate two-sample problem based on the concept of minimum energy.Journal of Statistical Computation and Simulation 75(2), 109–119.

Wassily Hoeffding, A class of statistics with asymptotically normal distribution. Ann. Math. Statistics 19 (1948), 293–325.

Buening, H. (2001). Kolmogorov-Smirnov- and Cram‘er-von Mises type two-sample tests with various weight functions. Communications in Statistics- Simulation and Computation, 30, 847-865.

I.S. Gradshteyn and I.M. Ryzhik. Table of Integrals, series and products.Seventh edition AMSTERDAM,BOST LONDON

A. P. Prudnikov, Yu. A. Brychkov, and O. I. Marichev, Integrals and Series. Elementary Functions (Nauka, Moscow, 1981) [in Russian].