# New test for equality of two disributions

Viatcheslav Melas[a,*], Dmitrii Salnikov[a]

[a]*St. Petersburg State University*
*Department of Mathematics*
*St. Petersburg , Russia*

**Abstract**

The paper introduces a new test for equality of two distributions in a class of models. We proved analytically and by stochastic simulation that the test possesses high efficiency .

*Keywords:* Test for equality of two distributions, Asymptotic efficiency, Caushy distribution

## 1. Formulation of the problem

Let us consider the classical problem of testing hypothesis on the equality of two distributions

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

against the alternative

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

In the case of two independent samples $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$ with the distributions functions $F_1$ and $F_2$ respectively.

It is well known [see e.g. [1]] that in the case when both distributions differ only by the means and are normal the classical Student test has a few optimal properties. If the distributions are not normal but still differs only by means a widely popular Wilcoxon-Mann-Whitney (WMW) U-statistic is often used instead. However, it can be shown that if two normal populations differ only in variances, the power of WMW test is very low. If distributions are arbitrary there are some universal techniques such as tests by Kolmogorov - Smirnov and Cramer-von Mises (see [2]) that can be applied but in many cases these tests can be not powerful.

Recently [3] 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.

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

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

[*]Corresponding author
*Email addresses:* `vbmelas@yandex.ru` (Viatcheslav Melas), `mejibkop.ru@mail.ru` (Dmitrii Salnikov)

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.

## 2. The 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 $m \neq n$ is similar. Now the criterion (4) 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

$$1/(\pi(v^2 + (x-u)^2)).$$

The basic 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*

$$(aZ + b)^2, \tag{8}$$

*where $Z$ has the normal distribution with zero expectation and variance equal to 1, $a$ and $b$ are some numbers.*
*(ii) Let $F_1 = C(0, 1), F_2 = C(\theta, 1)$, where $\theta = h/\sqrt{n}, h$ is an arbitrary given number. Then $a^2 = (2/3)\ln 3, b = 0$ for the case of $H_0$ and $a^2 = (2/3)\ln 3, b = h/3$ for $H_1$. In this case the power of the criterion $T_n$ with significance $\alpha$ is asymptotically equal to that is given by the formula*

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

The proof of the theorem is given in the Appendix.

We found by a stochastic simulation that the formula present an approximation of the power of the test $T_n$ with accuracy 5%. Namely, cases $n$ = 250, 500, 1000 , $h$=1,2,3,5,7,9 were considered with $\alpha = 0.05$ and in all these cases the power of $T_n$ and that of the Wilcoxon-Mann-Whitney and the Kolmogorov - Smirnov tests were approximately equal to each other. However for the case of Caushy distributions with different scale parameters $T_n$ proved to be much more efficient than both other tests.

## 3. Conclusion

In this paper we suggested a new test for equality of two distributions. Its asymptotic power was analytically established for the case of Caushy distributions that differ only by shift. By stochastic simulation we found that in this case its power is approximately equal to that of the Wilcoxon-Mann-Whitney and the Kolmogorov - Smirnov tests. But if the distributions differ also by the scale parameter simulations show that the new test is considerably better than the alternative tests.


## 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).


## 4. Appendix

Proof of Theorem 1.

**Lemma 1.** *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.[4], 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,

$$\tilde{X} = (\tilde{X}_1, \dots, \tilde{X}_n), \;\; \tilde{Y} = (\tilde{Y}_1, \dots, \tilde{Y}_n),$$

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

$$n\{\frac{1}{n^2} \sum_{i,j=1}^{n} \ln(1 + (\tilde{X}_i - \tilde{Y}_j)^2) - \frac{1}{n(n-1)} \sum_{i<j} \ln(1 + (\tilde{X}_i - X_j)^2) - \tag{9}$$

$$\frac{1}{n(n-1)}\sum_{i<j}\ln(1+(\tilde{Y_i}-\tilde{Y_j})^2)]^2\}\}=t(\sum_{i=1}^{n}\tilde{X}_i/\sqrt{n}-\sum_{i=1}^{n}\tilde{Y}_i/\sqrt{n})^2. \tag{10}$$

For constructing the right hand side we applied Lemma 1. Note that for distributions $F_1$ and $F_2$ satisfying (3) it can be shown by standard but tedious calculations 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 (8) 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 (10). 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

$$E((aZ+b)^2)^2=\lim_{n\to\infty}E(nT_n)^2, \tag{11}$$

that follows from (10).

Since $EZ^2=1, EZ^4=3$, we have for the left hand side (11)

$$3a^4+6a^2b^2+b^4. \tag{12}$$

In order to calculate the right hand side of (11) the following result is crucial.

**Lemma 2.** *If $X$ and $Y$ are independent random values with the distribution $C(0,1)$, then*

$$E\ln(1+(X-Y)^2)=\ln 9, \quad E\ln(1+(X-Y-\theta)^2)-\ln 9=ln(1+\theta^2/9). \tag{13}$$

In order to prove this Lemma we need the following integrals

$$\int_R \frac{\ln(1+(x-y)^2)}{\pi(1+y^2)}dy=\ln(4+x^2), \tag{14}$$

$$\int_R \frac{\ln(4+x^2)}{\pi(1+x^2)}dx=\ln 9, \tag{15}$$

([5] 4.296.2 and 4.295.7.)

$$\int_R \frac{\ln(4+(x+\theta)^2)}{\pi(x^2+1)}dx=\ln(9+\theta^2), \tag{16}$$

[see [6], formula (2.6.14.19)]. Using these integrals we obtain

$$E\ln(1+(X-Y-\theta)^2)-\ln 9=2\int_R\int_R \frac{\ln(1+(x-y-\theta)^2)}{\pi^2(1+x^2)(1+y^2)}dxdy-\ln 9 \tag{17}$$

$$=\int_R \frac{\ln(4+(y+\theta)^2)}{\pi(1+y^2)}dy-\ln 9=\ln(9+\theta^2)-\ln 9=\ln(1+\theta^2/9). \tag{18}$$

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

$$3a^4 + \frac{(2\ln 9)h^2}{9} + \frac{h^4}{81}. \tag{19}$$

From (12) and (19) 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 (8).

## References


[1] E. Lehmann, Testing Statistical Hypotheses, Probability and Statistics Series, Wiley, New York, 1986.

[2] H. Buening, Kolmogorov-smirnov- and cram‘er-von mises type two-sample tests with various weight functions., Communications in Statistics-Simulation and Computation 30 (2001) 847–865.

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

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

[5] I. Gradshteyn, I. Ryzhik, Table of Integrals, series and products. Seventh edition, Elsevier Inc., AMSTERDAM, 2007.

[6] A. P. Prudnikov, Y. A. Brychkov, O. I. Marichev, Integrals and Series [in Russian], Nauka, Moscow, 1981.