On the asymptotic power of the new test for equality of two distributions

Viatcheslav Melas and Dmitrii Salnikov
St.Petersburg State University, Russia

1. Introduction

The presentation is devoted to a new test for equality of two distributions in a class of models recently introduced in (Melas and Salnikov,2020)

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

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

1. It is well known [see e.g. (Lehman,1986)] that in the case when both distributions differ only by shift and are normal the classical Student test has a few optimal properties.

2. If the distributions are not normal but still differs only by shift a widely popular Wilcoxon-Mann-Whitney (WMW) U-statistic is often used instead.

3. 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 Darling-Anderson that can be applied but in many cases these tests can be not powerful.

Zech and Aslan (2005) 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 consider a similar but different test and obtain a few analytical results on its power.

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 That are simmetric around a point and such that

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

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

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

Consider the following test

$$\Phi_A = \frac{1}{n^2} \sum_{1 \le i < j \le n} g(|X_i - X_j|), \Phi_B = \frac{1}{m^2} \sum_{1 \le i < j \le m} g(|Y_i - Y_j|), \quad (4)$$

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

where

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

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

We will show that the test is very appropriate for two distributions that differ either by shift or scale parameters and belong to a rather general class of distributions.

In particular, we consider the class of distributions satisfying (3) and are symmetric around a point, but the approach can be generalized for other classes of distributions.

Note that would be parameters are know the test basing on likelihood ratio is the most powerful among tests with a given parameters.

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

Basic analytical results

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

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

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

Denote by $C(u, v)$ the Cauchy distribution with shift u and scale parameter v.

The basic analytical results of Melas and Salnikov (2020) can be formulated in the following way.

Consider the problem of testing hypothesis on the equality of two distributions (1)-(2) where

$$F_1(x) = F(x), F_2(x) = F(x + h/\sqrt{n}),$$

$F(x)$ has the property (3) and is symmetric around a point.
Let us call these conditions as **basic conditons.**

Theorem

For the problem of testing equality of two distributions under **basic conditons** *with* $n \to \infty$ *the distribution function of* nT_n *converges under* H_0 *and under* H_1 *to that of the random value*

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

where Z *has the normal distribution with zero expectation and variance equal to 1,* a *and* b *are some numbers,* $b=0$ *for* H_0.

The following result can be easily derived from the above theorem.

Theorem

For the problem of testing equality of two distributions under **basic conditons** *with* $n \to \infty$ *the power of the criterion* T_n *with significance* α *is asymptotically equal to that is given by the formula*

$$Pr\{Z > z_{1-\alpha/2} - kh\} + Pr\{Z < -z_{1-\alpha/2} - kh\},$$

where k *is some coefficient,* $Pr\{Z > z_{1-\alpha/2}\} = 1 - \alpha/2$.

Note that the coefficient can be evaluated by simulation experiments and with given coefficient the power can be calculated by tables for the Normal distribution.

The value $k \approx 0.4125$ for the Cauchy distribution was estimated by the Least Squares technique by simulating 1000 samples $X \sim C(0,1)$, $Y \sim C(h/\sqrt{n},1)$ of size $n = 1000$, h takes values from 0 to 10.5 with a step of 0.5.

Table 1: Power of T_n, sim and its estimate for the Cauchy distribution, $X \sim C(0,1)$, $Y \sim C(h/\sqrt{n},1)$, samples size 1000, 1000 iterations

h	0	2	4	6	8	10
T_n, sim	4.6	11.9	34.6	69.9	92.5	98.9
Formula	5.0	13.1	37.8	69.7	91.0	98.5

The value $k \approx 0.6702$ for the Normal distribution was estimated similarly, h took values from 0 to 6.5 with a step of 0.5.

Table 2: Power of T_n, sim and its estimate for the Normal distribution, $X \sim N(0,1)$, $Y \sim N(h/\sqrt{n}, 1)$, samples size 1000, 1000 iterations

h	0	1	2	3	4	5	6
T_n, sim	5.2	11.3	27.4	51.6	77.0	91.2	98.5
Formula	5.0	10.3	26.8	52.0	76.4	91.8	98.0

Thus the simulation show that the formula present an approximation of the power of the test T_n with a good accuracy.

At the next tables results for Cauchy and Normal distributions of size $n = 100$ with $\alpha = 0.05$ are given.

Note that in all cases when the tested distributions differs only by shift the power of T_n and that of the Wilcoxon-Mann-Whitney and the Kolmogorov-Smirnov tests were approximately equal to each other in the case

But if the distributions differ by scale the new test is considerably better than alternative ones.

Table 3: Cauchy distribution, $X \sim C(0,1)$, $Y \sim C(h/\sqrt{n},1)$, $n = 100$

h	$T_n, perm$	T_n, sim	$Formula$	$wilcox.test$	$ks.test$	$ad.test$
1	6.4	6.3	7.0	6.6	6.3	7.1
2	10.1	10.6	13.1	11.9	11.1	11.6
3	19.6	20.3	23.6	20.5	20.2	20.7
5	50.9	50.5	54.1	48.5	53.1	52.2
7	82	82.3	82.3	77.2	83.6	80.7
9	96.7	96.8	96.0	91.5	96.5	95.2

Table 4: Cauchy distribution, $X \sim C(0,1)$, $Y \sim C(0, 1+h/\sqrt{n})$, $n = 100$

h	$T_n, perm$	T_n, sim	$wilcox.test$	$ks.test$	$ad.test$
2	10.6	11.9	5.4	5.4	6.9
4	27.6	29.8	5.5	8.7	11.3
6	49.4	53.6	5.5	15.9	22.2
8	68.8	73.5	5.5	25	37.7
10	84.2	87.1	5.2	36.4	55.4

Table 5: Normal distribution, $X \sim N(0,1)$, $Y \sim N(h/\sqrt{n},1)$, $n = 100$

h	$T_n, perm$	T_n, sim	$Formula$	$wilcox.test$	$ks.test$	$ad.test$
1	11.1	11.3	10.3	12.5	9.5	12.2
2	29.3	29	26.8	31.1	20.5	29.6
3	52.4	53.4	52.0	55.8	42	55
4	77.5	77.5	76.4	80.6	64.9	78.9
5	91.9	92.5	91.8	93.1	84.7	93.1

Table 6: Normal distribution, $X \sim N(0,1)$, $Y \sim N(0, 1+h/\sqrt{n})$, $n = 100$

h	$T_n, perm$	T_n, sim	$wilcox.test$	$ks.test$	$ad.test$
1	8.1	8.7	6.4	5.3	7.3
2	15	17.4	6.3	7.2	12.7
3	30.5	34.2	6.6	10.7	24.0
4	50.6	57.1	6.7	16.7	39.9
5	70.8	76.7	6.5	24.8	59.9

4.Proof of Theorems

Lemma

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

$$\Phi_{nm} = (\bar{x} - \bar{y})^2, \bar{x} = (\sum_{i=1}^{n} X_i)/n, \bar{y} = (\sum_{i=1}^{m} Y_i)/m.$$

The proof follows from the known formula (Hoeffding, 1946)

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

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

$$\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_n), \quad \tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_n),$$

where $\tilde{X}_i = X_i$, if $|X_i| \leq C$ 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 $0 \leq \ln(1 + x^2) \leq x^2$. Therefore there exists a value t that depends from $\tilde{X}$ and $\tilde{Y}$ such that (our **basic equation**)

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

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

Note that for distributions of random values ξ^2 with finite expectation of $\ln(1+\xi^2)$ **it can be shown by standard but tedious calculations that the variance of the left hand side of the basic equation is finite**.

Therefore the variance of the right hand side of the basic equation 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 $(aZ + b)^2$ 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 the basic equation. Since C is arbitrary we obtain that the limiting distribution has the required form for H_0.

References

Melas, V., Salnikov, D. (2020). New test for equality of two disributions. In: Kozyrev D. (Ed.), THE 5th INTERNATIONAL CONFERENCE ON STOCHASTIC METHODS (ICSM-5): Proceedings of the international scientific conference, pp. 125-129.

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. (1948) A class of statistics with asymptotically normal distribution. Ann. Math. Stat. 19, 293-325.

Conclusion

In this paper we consider a new test for equality of two distributions. Its asymptotic power was analytically established for the case of Normal and Cauchy 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

Work at the paper was supported by RFBR (grant N 20-01-00096).

Also we would like to thank organizers and participants who visited this presentation.