Introduction
The new test and its statistical motivation
The analytical study of asymptotic power
Simulation results
Proof of Theorems
References
Conclusion
Acknowledgments

New test for equality of two disributions

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

1. Introduction
2. The new test and its statistical motivation
3. The analytical study of asymptotic power
4. Simulation results
5. Proof of Theorems
6. References
7. Conclusion
8. Acknowledgments

1. Introduction

We introduce a new test for equality of two distributions in a class of models.

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

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

It is well known [see e.g. (Lehman,1986)] 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 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 introduce a similar but different test and provide 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, 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. (Note that Zech and Aslan (2005) took $g(x) = \ln(|x|)$).

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.

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 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 Caushy distribution.

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 $m \neq n$ 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 shift u and scale parameter v.

The basic analytical results of the present paper consist in the following two theorems

Theorem

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

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

where Z has the normal distribution with zero expectation and variance equal to 1, a and b are some numbers.

Theorem

Let under assumtions of the previuos theorem $F_1 = C(0,1), F_2 = C(\theta,1)$, *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* α *is asymptotically equal to that is given by the formula*

$$Pr\{Z \geq z_{1-\alpha/2} - (1/\sqrt{6\ln 3})h\} + Pr\{Z \leq -z_{1-\alpha/2} - (1/\sqrt{6\ \mathit{ln}3})h\}$$

We found by a stochastic simulation that the formula present an approximation of the power of the test T_n with a good accuracy.

At the next tables resuts for cases $n = 500, 1000$, h=1,2,3,5,7,9 with $\alpha = 0.05$ are given.

Note that 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.

Table: 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

Table: 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

Table: 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

Table: 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

Table: 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

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.

Lemma

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

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

In order to prove this Lemma we need the following integrals

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

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

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

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 of the basic equation with some algebra

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

And 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.

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), 109119.

Wassily Hoeffding, A class of statistics with asymptotically normal distribution. Ann. Math. Statistics 19 (1948), 293325.

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,BOSTON,HEIDELBERG, LONDON

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

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).

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