On asymptotic power of some tests for equality of distributions

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]) and the Anderson-Darling test (see [3]) that can be applied but in many cases these tests can be not powerful.

Zech and Aslan [4] suggested the test based 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. Recently Melas and Salnikov [5] introduced a similar but different test and provide a few analytical results on its power. In particular, it was proved that the test statistic is asymptotically distributed as the square of a Normal distribution. Here we derive an explicit formula for parameters of that distribution and establish a minimax property of the test.

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 variables ξ, such that

$$E[g^2(\xi)] < \infty, \tag{3}$$

where $g(x)$ is a given function, in particular, $g(x) = \ln(1 + x^2)$.

Many distributions and, in particular, the Cauchy distribution have this property for $g(x) = \ln(1 + x^2)$.

Consider the following test

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

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

where $g(x)$ is a given function. We will assume that it is non negative, symmetric around the origin and twice continuously differentiable. In [5] it was considered the case

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

this function 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 would like to have a test that is appropriate for the case where the basic distribution belongs to a rather general class of distributions and the alternative distribution differs only by shift and scale transformations.

Consider the class of distributions given by the property (3). Note that if the parameters are known the test based 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 a special distribution. We will prove that it is very efficient for all distributions with property (3) and possesses a remarkable minimax property.

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 a shift. To simplify notations assume that $m = n$. The case $m \neq n$ is similar. Now the criterion (4) - (5) assumes the form

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

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

Let $f(x)$ denotes the density of F_1. We will interest in the two distribution that are differ by a normalized shift bThe natural normalizing coefficient that characterize the "width" under H_0 is square root of the entropy of the difference of two independent random variables with density $f(x)$. Denote

$$J_h = J_h(f, g) = \int_R g(x - y - |h|/\sqrt{2H(f)n}) f(x)f(y)dxdy. \tag{8}$$

where $g(u)$ is a function such that the integral exists, $H(f)$ is the entropy of f. Assume also that $g(0) = 0$.

Since $g(x)$ is twice differentiable we obtain that for arbitrary density function $f(x)$ there exists the finite limit

$$J^*(h) = lim_{n\to\infty} n(J_h - J_0) \tag{9}$$

and it is equal to

$$(1/2)\frac{h^2}{2H(f)}\int_R g''_\theta(x - y - \theta)f(x)f(y)dxdy|_{\theta=0} \tag{10}$$

under the supposition that $g(x)$ is such that the differentiation under the integral is possible. Denote

$$\bar{b} = \sqrt{J^*(h)/h^2}.$$

The first 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). Let $g(x)$ be an arbitrary non negative, symmetric around the origin and twice continuously differentiable function and $g(0) = 0$. Then*
(i) under the condition $n \to \infty$ the distribution function of nT_n converges under H_0 to that of the random variable

$$2a^2(L)^2, \tag{11}$$

where L has the normal distribution with zero expectation and variance equal to 2 , $a^2 = J_0/2$.

(ii) Let $F_1(x) = F(x), F_2 = F(x+\theta)$, where F is an arbitrary distribution function that is symmetric around a point and possess property (3), $\theta = h/\sqrt{2H(f)n}, h$ is an arbitrary given number. Assume that $g(x)$ is such that the integral (10) is existed and finite. Then the distribution function of nT_n converges under H_1 to that of the random variable

$$2a^2(L+b)^2,$$

where $b = \bar{b}h$.

In this case the power of the criterion T_n with significance α is asymptotically equal to that is given by the formula

$$Pr\{L \geq z_{1-\alpha/2} - \bar{b}h/a\} + Pr\{L \leq -z_{1-\alpha/2} - \bar{b}h/a\},$$

where $z_{1-\alpha/2}$ is such that

$$Pr\{L \geq z_{1-\alpha/2}\} = \alpha/2.$$

The proof of this theorem and the next results are given in the Appendix.

The result of Theorem 1 allows to establish some interesting properties properties of the criterion (6)-(7). Denote

$$p(u) = \exp(-g(u))/\int_R \exp(-g(x))dx. \tag{12}$$

Let f(x) be a density function that is symmetric around zero and such that

$$J_0(f,g)/(2H(f)) \leq K, \tag{13}$$

where K is a given constant. Denote

$$K^* = \int_R g(x)p(x)dx/(H(p)),$$

where p is determined by formula (12). Set $v = 1$ if $K = K^*$, otherwise

$$v = arg \min_v \max_f \int_R g(v(x-y))/\sqrt{2H(f)n}f(x)f(y)dxdy, \tag{14}$$

where maximum is taken over all density functions that are symmetric around the origin and satisfy (13). Set $\tilde{K} = K^*$ if $K = K^*$ and

$$\tilde{K} = \min_v \max_f \int_R g(v(x-y))/\sqrt{2H(f)n} f(x)f(y)dxdy,$$

otherwise.

Note that the asymptotic power of the criterion T_n is increasing function of the magnitude

$$\bar{b}/a,, \tag{15}$$

where $a = \sqrt{J_0(f,g)}, \bar{b} = \sqrt{J''_h(f,g)|_h = 0}$.

Definition 1. *Let us call the magnitude (15) the asymptotic efficiency (for the case when the distribution can differ only by a shift) and denote it as*

$$Eff(T_n(f,g)).$$

Theorem 2. *.Assume that F_1 has a one dimensional distribution with density $f_1 = f$ that is symmetric in respect to zero and such that inequality (3)in power. Let we have*

$$J_0 \leq \tilde{K}. \tag{16}$$

Assume also that under H_1 the distributions are differ only by a shift and

$$F_1(x) = F(x), F_2 = F(x+\theta), \theta = h/\sqrt{2H(f)n},$$

where h is an arbitrary given number. Then the lower bound of the asymptotic power of the test Φ_{nn} among all distributions satisfying the properties above is achieved for f such that

$$f^*f(x) = p(vx),$$

where p is given by formula (12), v is determined by (14).

One more result concerns with the optimal choice of $g(x)$ for a fixed density function f. Denote by $G(f)$ the class of all twice differentiable functions $\tilde{g}$ such that $g(x) = 0$ and for a fixed symmetric around origin density functions f(x) the inequality

$$E\tilde{g}^2(x) < \infty$$

is in power. Note that $g(x) \in G(f)$ for any function f satisfying (3). Thus $G(f)$ is not empty for such functions.

Theorem 3. *For any function f satisfying (3)*

$$\max_{\tilde{g} \in G(f)} Eff(T_n(f,\tilde{g})) = Eff(T_n(f, g^*(f))), \tag{17}$$

where

$$g^*(f)) = \ln p(x)/p(0), p = f * f.$$

Let now $\Phi(K,g)$ be the class of all densities that are symmetric around the origin, satisfy inequality (3) and the inequality

$$Eg(\xi) \leq K,$$

where ξ is the sum of two independent random variables with density f(x). Denote by G the class of all twice differentiable functions $\tilde{g}$ such that $\tilde{g}(0) = 0$ and for any $f \in \Phi$ we have

$$J_0(f, \tilde{g}) \leq K.$$

.

Theorem 4. *The asymptotic efficiency possess the following property*

$$\max_{\tilde{g} \in G} \min_{f \in \Phi} Eff(f, \tilde{g}) = \min_{f \in \Phi} \max_{\tilde{g} \in G} Eff(f, \tilde{g}) \tag{18}$$

and the unique saddle point is (f_*, g_*), where $g_*(x) = g(vx),$, v is defined in Thorem 2, and f_* is such that $f_*^* f_* = p_*, p_*$ is determined by formula (12)with $g = g_*$.

4. Concluding remarks

. Note that the optimal choice of the function g up to a scale parameter coincides with the function used in constructing the class of density functions . Numerical stochastic simulations show that optimality does not great influenced by misspecification of this parameter. Also note that the parameter can be easily evaluated by sampling data. We should simply take

$$v_{opt} = \arg\min_v \frac{1}{n^2} \sum_{1 \leq i < j \leq n} g(v(X_i - X_j))/v^2$$

and use $g(v_{opt}x)$ instead of $g(x)$ in the definition of the criterion. Thus the approach has a good perspectives for practical applications.

5. Appendix

Proof of Theorem 1. Let us consider the test $(4) - (5)$ with the function $g(u) = u^2$.

Lemma 1. *For $g(x) = x^2$ the following identity holds*

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

where

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

Denote

$$Z = (X,Y) = (X_1, \ldots, X_n, Y_1, \ldots, Y_n), V(Z) = \frac{1}{2} \sum_{i=1}^{2n} \sum_{j=1}^{2n} (Z_i - Z_j)^2.$$

The proof follows from the known formula [see e.g. [6], p.296]

$$\frac{1}{n(n-1)}\sum_{1\le i<j\le n}(X_i-X_j)^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i-\bar{x})^2 \tag{19}$$

and the obvious identity

$$\sum_{i=1}^{2n}\sum_{j=1}^{2n}(Z_i-Z_j)^2=\sum_{i,j=1}^{n}(X_i-X_j)^2+\sum_{i,j=1}^{n}(Y_i-Y_j)^2+2\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-Y_j)^2, \tag{20}$$

by direct but non trivial calculations.

Really, let us use the standard notation

$$S_x^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i-\bar{x})^2$$

And S_y^2 and S_z^2 will be understood in the similar way. Denote

$$S_{xy}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-Y_j)^2.$$

Note that due to formula (19) for X replaced by Z

$$V(Z)=2n[\sum_{i=1}^{n}(X_i-(\bar{x}+\bar{y})/2)^2+\sum_{j=1}^{n}(Y_i-(\bar{x}+\bar{y})/2)^2]=2n(n-1)(S_x^2+S_y^2)+n^2(\bar{x}-\bar{y})^2. \tag{21}$$

From (19) and (20) we obtain

$$n^2S_{xy}=V(Z)-n(n-1)(S_x^2+S_y^2). \tag{22}$$

Therefore

$$S_{xy}=\frac{1}{n}(n-1)(S_x^2+S_y^2)+(\bar{x}-\bar{y})^2,$$

and we obtain

$$\Phi_{nn}=S_{xy}-\frac{1}{n}(n-1)(S_x^2+S_y^2)=(\bar{x}-\bar{y})^2.$$

Thus Lemma 1 is proved. It follows from this lemma, that the criterion Φ_{nn} in this case is equivalent to the criterion $(\bar{x}-\bar{y})^2$.

Let us turn to the proof of the theorem.

Assume that either H_0 or H_1 holds. Then due to the law of large numbers for $U-$statistics ([6]) each of the sums

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

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

tends to J_0.
Note that

$$nT_n = n[\Phi_{AB} - J_0] - n[\Phi_A - \frac{1}{2}J_0] - n[\Phi_B - \frac{1}{2}J_0].$$

Let us apply the limit theorem for U-statistics (see Theorem 7.1 [6]) to each of the three terms in brackets. We obtain that nT_n tends to a random variable with a finite variance. Note that the conditions of the limit theorem are fulfilled for distributions F_1 with the property (3).

Denote by

$$T_{n,g}(X,Y), X = (X_1, \ldots, X_n), Y = (Y_1, \ldots, Y_n)$$

the left hand side of (6). 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. Let

$$g^*(x) = x^2, \phi_n = T_{n,g}(X,Y),$$

$$\tilde{\phi}_n = T_{n,g}(\tilde{X},\tilde{Y}), \tilde{\psi}_n = T_{n,g^*}(\frac{\tilde{X}}{\sqrt{var\tilde{X}_1}}, \frac{\tilde{Y}}{\sqrt{var\tilde{Y}_1}}),$$

$$\tilde{r}_n = \frac{\tilde{\phi}_n/\tilde{\psi}_n}{var\tilde{X}_1}$$

and note that under the assumption of the theorem $var\tilde{X}_1 = var\tilde{Y}_1$.

Then we have

$$n\phi_n \ge n\tilde{\phi}_n = var\tilde{X}_1 n\tilde{\psi}_n\tilde{r}_n$$

Let the hypothesis H_0 is true. Passing to the limit with $n \to \infty$ we obtain due to the large number law for U-statistics that $\tilde{r}_n$ tends to a constant with probability 1.
Denote

$$\hat{X}_i = \frac{\tilde{X}_i}{\sqrt{var\tilde{X}_i}}, \hat{Y}_i = \frac{\tilde{Y}_i}{\sqrt{var\tilde{Y}_i}}, i = 1, \ldots, n.$$

Note that due to Lemma 1

$$n\tilde{\psi}_n = (\sum_{i=1}^{n} \hat{X}_i/\sqrt{n} - \sum_{i=1}^{n} \hat{Y}_i/\sqrt{n})^2 \qquad (23)$$

and $n\tilde{\psi}_n$ tends to a random variable with Normal distribution with zero mean and variance 2 due to the central limit theorem for U-statistics. And we obtain that $\tilde{\phi}_n$ tends to a random variable with Normal distribution with zero mean and some variance $\tilde{a}^2$.Since C is arbitrary we obtain that the limiting distribution of $n\phi_n$ has the required form.

To calculate a, note that due to the above arguments we have

$$\lim_{n\to\infty} nT_n = a^2(V_1 - V_2)^2, \qquad (24)$$

where V_1 and V_2 are independent random variables with the standard normal distribution.

Taking expectation in the left hand side we obtain J_0. Really, since T_n is determined by equations (6)-(7) we have $nT_n = I_1 + I_2$, where

$$I_1 = \frac{1}{n} \sum_{1 \le i < j \le n} [2g(X_i - Y_j) - g(X_i - X_j) - g(Y_i - Y_j)],$$

$$I_2 = \frac{1}{n} \sum_{i=1}^{n} g(X_i - Y_i).$$

Note that $EI_1 = 0$ and $EI_2 = J_0$.

And the expectation of the right hand side is obviously $2a^2$. Thus $a^2 = J_0/2$.

Assume now that H_1 holds.

For determining b in the part (ii) of the theorem we now can use the equality

$$(aV + b)^2 = \lim_{n \to \infty} nT_n, \tag{25}$$

where $V = V_1 - V_2$ that follows from (23). If H_0 take place we obviously have $b = 0$.

In the case when H_1 take place EnT_n is asymptotically equivalent to

$$(n(J_h - J_0))^2 + En\hat{T_n}$$

where $\hat{T_n}$ received from T_n by replacing Y_i by $Y_i - b/\sqrt{n}$, $i = 1, \ldots, n$ and we obtain by passing to the limit with $n \to \infty$ that the right hand side 0f (25) is equal to

$$J_0 + (\bar{b}h)^2, \bar{b} = \sqrt{J^*(h)/h^2}.$$

And the asymptotic behaviour of the power announced in (ii) follows from the asymptotic normality of $\sqrt{nT_n}$ that completes the proof of the theorem.

Proof of Theorem 2. Let us begin with the case $K = K^*$. In this case as it was mentioned above v=1. Let us study the asymptotic efficiency

$$Eff(T_n(f,g)) = \bar{b}/a, a = \sqrt{J_0(f,g)}, \bar{b} = \sqrt{J''{}_h(f,g)|_h = 0}.$$

Note that

$$J_0(f,g) = Eg(\xi),$$

where ξ is the sum of two independent random variables with density f(x) (since f assumed to be symmetric around a point). Therefore the upper bound of a is achieved for arbitrary f such that

$$J_0(f,g) = K^*.$$

The proposition of the theorem in the case $K = K^*$ follows now from the next lemma.

Lemma 2. *Under the conditions described in Theorem 2 with $K = K^*$ the value $\bar{b}$ achieves its lower bound if and only if*

$$f(x) = p(x). \tag{26}$$

The proof is based on the property to be well known under the title of maximal entropy: among all random variables with density ψ satisfying the inequality

$$\int_R g(x)\phi(x)dx = K^* \tag{27}$$

the variable possess the maximum entropy if and only it is equal to p(x) determined by equality (12).

Denote

$$\tilde{J}(h) = \int_R \varphi(z)g(z+h)dz,$$

where $\varphi(z)$ is the density of the difference between two independent random variables with the density f.

Note that $\bar{b}$ can be written as

$$\frac{[\tilde{J}(h)]''|_{h=0}}{H(\varphi)}. \tag{28}$$

Thus in the new notations

$$\bar{b} = \lim_{n\to\infty} \frac{\tilde{J}(h) - \tilde{J}(0)}{\frac{1}{n}h^2 H(\varphi)}. \tag{29}$$

Note that $\tilde{J}(0) = J_0$ and it is assumed to be K. Denote by φ_n the function φ for which the lower bound of

$$\frac{\tilde{J}(h) - \tilde{J}(0)}{\frac{1}{n}h^2 H(\varphi)}$$

is achieved. Then obviously φ_n with $n \to \infty$ tends to a function (denote it φ^*) such that

$$H(\varphi^*) = supH(\varphi),$$

where the upper bound is taken over densities such that

$$\int_{-\infty}^{\infty} \varphi(x)g(x))dx = K^*.$$

Due to the property mentioned above we have

$$\varphi^*(z) = p(z).$$

Therefore the lower bound of $\bar{b}$ is achieved if and only if ξ has the density function p(x). Thus the lemma is proved. The case of arbitrary K can be considered in the same way. Theorem 2 is proved.

Proof of theorem 3. The proposition of the theorem follows from the known inequality

$$\int_{-\infty}^{\infty} f(x)\ln q(x)dx \geq \int_{-\infty}^{\infty} q(x)\ln q(x)dx$$

with equality if and only if q=f that is valid for arbitrary densities and the arguments similar to that was used in proof of Theorem 2.

Proof of Theorem 4. The proposition follows from that Theorems 2 and 3. Really we obtain due to Theorem 3

$$Eff(f_*, g_*) = \max_{\tilde{g}\in G} Eff(f_*, \tilde{g}) \geq \max_{\tilde{g}\in G} \min_{f\in\Phi} Eff(f, \tilde{g}) \geq \min_{f\in\Phi} | \max_{\tilde{g}\in G} Eff(f, \tilde{g}) \geq \min_{f\in\Phi} Eff(f, g_*). \tag{30}$$

* And by Theorem 2

$$\min_{f\in\Phi} Eff(f, g_*) = Eff(f_*, g_*).$$

The uniqueness of the saddle point follows from the corresponding statements in Theorems 2 and 3.

Acknowledgments

The work of V.Melas at the paper was partly supported by RFBR (grant N 20-01-00096).

References

[1] Lehmann E.: Testing Statistical Hypotheses, Probability and Statistics Series, Wiley (1986).

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

[3] Anderson T.W.: Anderson–Darling Tests of Goodness-of-Fit. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg.(2011) https://doi.org/10.1007/978-3-642-04898-2_118

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

[5] Melas V. and Salnikov D.On Asymptotic Power of the New Test for Equality of Two Distributions A. N. Shiryaev et al. (eds.), Recent Developments in Stochastic Methods and Applications, Springer Proceedings in Mathematics and Statistics 371 (2021)

[6] Wassily Hoeffding: A class of statistics with asymptotically normal distribution. Ann. Math. Statistics 19, pp. 293-325 (1948).