Milan Krišťák

Some rank tests of independence and the question of their power-function

Aplikace matematiky, Vol. 16 (1971), No. 6, 412–420

Persistent URL: `http://dml.cz/dmlcz/103376`

SOME RANK TESTS OF INDEPENDENCE AND THE QUESTION OF THEIR POWER-FUNCTION

MILAN KRIŠŤÁK

(Received July 10, 1970)

The paper deals with the problem of testing independence of a pair of random variables X, Y by locally most powerful rank tests. Theorem 1 gives a solution to this problem. A similar theorem is proved in [2] (II.4.11) under the assumptions that f' and g' are continuous almost everywhere, whereas we suppose only integrability of the derivatives f' and g'. Theorem 2 gives the derivative of the powerfunction of the S-test at the point $\Delta = 0$.

Two locally most powerful rank tests of independence for double-exponentially and normally distributed random variables W and W^*, which are based on general results of the first section and [2], are introduced. The power-functions of the U-test in a neighborhood of the point $\Delta = 0$ for both cases are given numerically.

1. LOCALLY MOST POWERFUL RANK TEST OF INDEPENDENCE

Let $(X_1, Y_1), \ldots, (X_N, Y_N)$ denote a random sample from a bivariate population. We shall test a composite hypothesis

$$H_0: \quad P(X_i \leqq x_i,\ Y_i \leqq y_i,\ i = 1, \ldots, N) = \prod_{i=1}^{N} F^*(x_i)\, G^*(y_i)$$

where F^*, G^* are arbitrary continuous distribution functions of the random variables X_i, Y_i, $i = 1, \ldots, N$. This hypothesis will be tested against a simple alternative
H_Δ: The density of the simultaneous distribution of the $2N$-dimensional random variable $(X, Y) = (X_1, Y_1, \ldots, X_N, Y_N)$ equals

$$p_\Delta(x, y) = \prod_{i=1}^{N} h_\Delta(x_i, y_i)$$

where

$$(1) \qquad h_\Delta(x_i, y_i) = \int_{-\infty}^{\infty} f(x_i - \Delta z_i)\, g(y_i - \Delta z_i)\, \mathrm{d}M(z_i)\,, \quad i = 1, \ldots, N\,,$$

$\Delta > 0$ denotes a real parameter and $M(z)$ is an arbitrary distribution function of the random variables Z_i, $i = 1, \ldots, N$, with a positive and finite variance σ^2, i.e.

$$0 < \int_{-\infty}^{\infty} z^2 \, \mathrm{d}M(z) - \left(\int_{-\infty}^{\infty} z \, \mathrm{d}M(z) \right)^2 = \sigma^2 < \infty .$$

We shall assume that both f and g are on finite intervals absolutely continuous densities of known types of the random variables

$$(2) \qquad W_i = X_i - \Delta Z_i \quad \text{and} \quad W_i^* = Y_i - \Delta Z_i ,$$

i.e. that for arbitrary $-\infty < a < b < \infty$ there exist functions f' and g' such that

$$\int_a^b f'(t) \, \mathrm{d}t = f(b) - f(a) \quad \text{and} \quad \int_a^b g'(t) \, \mathrm{d}t = g(b) - g(a) ,$$

and let furthermore

$$(3) \qquad \int_{-\infty}^{\infty} |f'(t)| \, \mathrm{d}t < \infty \quad \text{and} \quad \int_{-\infty}^{\infty} |g'(t)| \, \mathrm{d}t < \infty .$$

Remark 1. Under the alternative we suppose that

$$X_i = W_i + \Delta Z_i \quad \text{and} \quad Y_i = W_i^* + \Delta Z_i , \quad i = 1, \ldots, N ,$$

where W_i, W_i^* and Z_i are mutually independent random variables. Thus we have

$$\operatorname{cov}(X_i, Y_i) = \Delta^2 \operatorname{var}(Z_i) ,$$

hence we shall test the null hypothesis $\Delta = 0$ against the alternative hypothesis $\Delta > 0$.

Let $R = (R_1, \ldots, R_N)$ be the random vector of ranks of the random variables $X_1, \ldots, X_N$ in their ordered sequence $X^{(1)} < \ldots < X^{(N)}$, i.e.

$$X_i = X^{(R_i)} , \quad i = 1, \ldots, N ,$$

and let $D = (D_1, \ldots, D_N)$ denote the inverse permutation to $(R_1, \ldots, R_N)$. Thus D is the vector of antiranks of the random variables $X_1, \ldots, X_N$, i.e.

$$X^{(i)} = X_{D_i} , \quad i = 1, \ldots, N .$$

Similarly let $Q = (Q_1, \ldots, Q_N)$ be the vector of ranks of the random variables $Y_1, \ldots, Y_N$ in their ordered sequence $Y^{(1)} < \ldots < Y^{(N)}$, i.e.

$$Y_i = Y^{(Q_i)} , \quad i = 1, \ldots, N .$$

Now denote F^{-1} and G^{-1} the inverse functions of the distribution functions of the random variables W and W^* respectively, and similarly as in [2] (I.2.4) define for

$\lambda \in (0, 1)$ the functions

$$(4) \qquad \varphi(\lambda) = -\frac{f'(F^{-1}(\lambda))}{f(F^{-1}(\lambda))} \quad \text{and} \quad \psi(\lambda) = -\frac{g'(G^{-1}(\lambda))}{g(G^{-1}(\lambda))}.$$

Introduce the following scores

$$(5) \qquad a_i = E\,\varphi(C^{(i)}) \quad \text{and} \quad b_i = E\,\psi(C^{(i)})$$

where $C^{(1)} < \ldots < C^{(N)}$ is an ordered sample from the uniform distribution on $(0, 1)$.

Definition 1. *Let $\{p_\Delta\}$, $\Delta \geqq 0$ is a set of densities, and suppose that $p_0 \in H_0$. Then a rank test will be called a locally most powerful rank test for H_0 against $\Delta > 0$ at some level α, iff it is uniformly most powerful among all rank tests at the level α for H_0 against p_Δ, $\Delta \in (0, \delta)$ for some $\delta > 0$.*

Considering this definition we shall construct for some right-hand neighborhood of the point $\Delta = 0$ a uniformly most powerful rank test of the hypothesis H_0 against H_Δ. We shall consider the least favourable particular null hypothesis, which is nearest to the alternative hypothesis H_Δ that the distribution of the random variable (X, Y) is determined by the density $f_\Delta(x)\, g_\Delta(y)$, where

$$f_\Delta(x) = \int_{-\infty}^{\infty} f(x - \Delta z)\, \mathrm{d}M(z) \quad \text{and} \quad g_\Delta(y) = \int_{-\infty}^{\infty} g(y - \Delta z)\, \mathrm{d}M(z).$$

Now we can formulate the following main theorem.

Theorem 1. *The locally most powerful rank test for H_0 against H_Δ at the level α_k is, under the above assumptions, the test with the critical region*

$$(6) \qquad S = S(R, Q) = \sum_{i=1}^{N} a_{R_i} b_{Q_i} \geqq k,$$

where α_k equals the probability of the event (6) under H_0.

In the proof of this theorem we can use the same procedure as in the proof of theorem II.4.11 from [2], only instead of the assumption that f' and g' are continuous almost everywhere, which is used for proving (10), p. 77 in [2], we directly use the property of their integrability. First, we introduce the following definition.

Definition 2. *A point x will be called Lebesgue's point of the function f iff $f(x) \neq \pm\infty$ and*

$$\lim_{h \to 0} \frac{1}{h} \int_x^{x+h} |f(t) - f(x)|\, \mathrm{d}t = 0.$$

For $z \neq z'$ we hawe

$$\frac{1}{\Delta(z - z')} [f(x - \Delta z) - f(x - \Delta z')] = \frac{1}{\Delta(z - z')} \int_{x - \Delta z'}^{x - \Delta z} f'(t) \, \mathrm{d}t \,,$$

furthermore for each Lebesgue's point x of the function f' is

$$\lim_{\substack{\delta_1 \to 0, \delta_2 \to 0 \\ \delta_1 \neq \delta_2}} \frac{1}{\delta_1 - \delta_2} \int_{x - \delta_2}^{x - \delta_1} f'(t) \, \mathrm{d}t = f'(x) \,,$$

and similarly for g'. Thus, in each Lebesgue's point of the funktion f', or g', formula (10) in [2] holds.

Since the theorem 5, IX, §4 in [3] holds clearly also for the whole real line, in view of (3) almost every point of the interval $(-\infty, \infty)$ is Lebesgue's point of the functions f' and g', consequently (10) in [2] holds almost everywhere.

The remainder of the proof is the same as the proof of theorem II. 4.11 in [2].

Note that for arbitrary fixed ranks $R_i = r_i$, $Q_i = q_i$, $i = 1, \ldots, N$, according to the last relation in the proof of the quoted theorem from [2] we have, under the alternative H_Δ,

$$P(R = r,\ Q = q / H_\Delta) = [1 + \Delta^2 \sigma^2 \, S(r, q) + o(\Delta^2)] \, (N!)^{-2}$$

where $\lim_{\Delta \to 0} o(\Delta^2) = 0$.

We can consider the critical region of the S-test, say $\mathscr{D}$, which is given by (6), as a subset of the pairs of permutations (r, q). Consequently, for the power-function of the S-test in a sufficiently small right-hand neighborhood of the point $\Delta = 0$ it holds

$$(7) \quad P((R, Q) \in \mathscr{D} / H_\Delta) = \sum_{(r,q) \in \mathscr{D}} [1 + \Delta^2 \sigma^2 \, S(r, q) + o(\Delta^2)] \, (N!)^{-2} \,.$$

By (7) we immediately obtain the following theorem.

Theorem 2. *The derivative of the power-function of the S-test at the point* $\Delta = 0$ *equals*

$$(8) \qquad \frac{\partial}{\partial \Delta^2} P((R, Q) \in \mathscr{D} / H_\Delta) = (N!)^{-2} \, \sigma^2 \sum_{(r,q) \in \mathscr{D}} S(r, q) \,.$$

Remark 2. If the subset $\mathscr{D}$ is defined by the rank statistic

$$S(t) = \sum_{j=1}^{N} a_j b_{t_j}$$

where $t_j = q_{d_j}$, then we can consider $\mathscr{D}$ as a subset of the permutations $t = (t_1, \ldots$ $\ldots, t_N)$. The derivative of the power-function of this test is by (8) equal to

$$\frac{\partial}{\partial \Delta^2} P(T \in \mathscr{D} | H_\Delta) = (N!)^{-1} \sigma^2 \sum_{t \in \mathscr{D}} S(t) . \tag{9}$$

We shall use these results in subsequent sections.

2. TWO RANK TESTS OF INDEPENDENCE FOR DOUBLE-EXPONENTIAL AND NORMAL DISTRIBUTIONS

We first suppose that the random variables W and W^* have the double-exponential density, i.e.

$$f(x) = g(x) = \tfrac{1}{2} e^{-|x|} . \tag{10}$$

It is easily seen that all assumptions from the first section are satisfied, and the functions (4) are equal to

$$\varphi(\lambda) = \psi(\lambda) = \operatorname{sgn}(\lambda - \tfrac{1}{2}) . \tag{4a}$$

If we now introduce the scores

$$a_i = b_i = E \operatorname{sgn}(C^{(i)} - \tfrac{1}{2}) \tag{5a}$$

where $C^{(i)}$ have the same meaning as in (5), then, by theorem 1, the locally most powerful rank test of H_0 against H_Δ at the respective level can be based on the statistic

$$S_1 = \sum_{i=1}^{N} E[\operatorname{sgn}(C^{(R_i)} - \tfrac{1}{2})] \, E[\operatorname{sgn}(C^{(Q_i)} - \tfrac{1}{2})] .$$

If we introduce the function

$$u(x) = \tfrac{1}{2}(\operatorname{sgn} x + 1) ,$$

then for the scores (5a) holds

(5aa)

$$a_i = b_i = E\left[2u(C^{(i)} - \tfrac{1}{2}) - 1\right] = 2 \sum_{j=0}^{i-1} \binom{N}{j} (\tfrac{1}{2})^N - 1 = 1 - 2 \sum_{j=1}^{N} \binom{N}{j} (\tfrac{1}{2})^N ,$$
$$i = 1, \ldots, N .$$

We are able to calculate the scores (5aa) with the aid of the tables [4]. These scores are given in table 1 for the sample size $N = 6$.

According to II.4.3 and III.6.1 in [2] we can say that an approximate locally most powerful rank test of H_0 against H_Δ can be based on the statistic

$$S_1^* = \sum_{i=1}^{N} \operatorname{sgn}(R_i - \tfrac{1}{2}(N+1)) \operatorname{sgn}(Q_i - \tfrac{1}{2}(N+1)) .$$

If we now introduce the statistic

$$U = \sum_{i=1}^{N} u[(R_i - \tfrac{1}{2}(N+1))(Q_i - \tfrac{1}{2}(N+1))] ,$$

then according to the definition of the function u we can write

$$S_1^* = 2U - N .$$

Consequently, the statistic U represents the same test as the statistic S_1^*.

Further, if the random variables W and W^* have the standardized normal densities f and g, then also all assumptions from the first section are satisfied. The functions (4) are then equal to

$$\varphi(\lambda) = \psi(\lambda) = \Phi^{-1}(\lambda) \tag{4b}$$

where Φ^{-1} denotes the inverse function of the standardized normal distribution function. The locally most powerful rank test of H_0 against H_Δ can be based on the statistic

$$S_2 = \sum_{i=1}^{N} a_{R_i} b_{Q_i}$$

with

$$a_i = b_i = E(V^{(i)}) = E[\Phi^{-1}(C^{(i)})] , \tag{5b}$$

$V^{(i)}$ and $C^{(i)}$, $i = 1, \ldots, N$, being the ordered samples from the standardized normal and from the uniform on (0, 1) distributions, respectively. These values (5b) are also shown in table 1 for $N = 6$. The test S_2 is introduced in [2] as the Fisher-Yates (normal scores) test. According to (2), III.6.1. in [2], for the correlation coefficient ϱ of the random variables X, Y holds

$$\varrho = \frac{\Delta^2}{1 + \Delta^2} ,$$

hence for $\varrho \to 0$ and for arbitrary fixed ranks $R = r$, $Q = q$ the following relation holds:

$$\frac{\partial}{\partial \varrho} P(R = r, Q = q / H_\Delta) = \frac{\partial}{\partial \Delta^2} P(R = r, Q = q / H_\Delta) . \tag{11}$$

Now we shall study the test of H_0 against H_Δ based on the statistic

$$U = \sum_{i=1}^{N} u[(i - \tfrac{1}{2}(N + 1))(T_i - \tfrac{1}{2}(N + 1))]$$

where $T_i = Q_{D_i}$.

If we denote the critical region of this test by

$$\mathscr{D}_1 = \{T = t;\ U = U(t) \geqq 2k\}$$

where k is determined by the required level of significance α, i.e.

$$P(U \geqq 2k/H_0) \leqq \alpha\,, \tag{12}$$

then by (9), under the assumption $\sigma^2 = 1$,

$$\frac{\partial}{\partial \Delta^2} P(T \in \mathscr{D}_1/H_\Delta) = (N!)^{-1} \sum_{t \in \mathscr{D}_1} S(t) \tag{13}$$

where

$$S(t) = \sum_{i=1}^{N} a_i b_{t_i}\,. \tag{14}$$

The statistic U for even sample sizes $N = 2n$ equals the number of pairs (X_i, Y_i) in their correlation diagram, which have both coordinates simultaneously either above, or below, of their sample medians. According to (3) in [1], or problem 4, IV, in [2], we can write the left-hand side of (12)

$$P(U \geqq 2k/H_0) = \left[\binom{n}{k}^2 + \binom{n}{k+1}^2 + \ldots + \binom{n}{n}^2\right]\binom{N}{n}^{-1}.$$

Accordingly we can determine the number k for given α for the size $N = 2n$.

If the random variables W and W^* have the double-exponential distribution, then the scores in (14) are determined by (5aa). We can in this case calculate the sums (14), which are denoted by $S_1(t)$. We have $\sum_{j=1}^{36} S_1(t^j) = 102{,}141\,2$ for $N = 6$, where the vectors of the ranks t for which $U = 6$ are denoted by t^j, $j = 1, \ldots, 36$. We can approximately determine the power-function of the U-test (for $\Delta \to 0$) for the level $\alpha = 0{,}05$ and the size $N = 6$ as follows:

$$P(U = 6/H_\Delta) \cong P(U = 6) + \Delta^2 \left[\frac{\partial}{\partial \Delta^2} P(U = 6/H_\Delta)\right]_{\Delta^2 = 0} =$$

$$= 0{\cdot}05 + (N!)^{-1} \Delta^2 \sum_{j=1}^{36} S_1(t^j) = P_I\,.$$

The values P_I for $\Delta^2 = 0{\cdot}15$; $0{\cdot}10$; $0{\cdot}05$; $0{\cdot}03$; and $0{\cdot}01$ are shown in table 2.

If the random variables W and W^* have the standardized normal distribution then the derivative of the power-function in a neighborhood of $\varrho = 0$ of the U-test of the hypothesis $\varrho = 0$ against the alternative $\varrho > 0$ has, according to (11) and (13), the following form:

$$\frac{\partial}{\partial \varrho} P(T \in \mathscr{D}_1 / \varrho > 0) = (N!)^{-1} \sum_{t \in \mathscr{D}_1} S_2(t)$$

where $S_2(t)$ are given by (14) with the scores (5b). In this case for $N = 6$ we have $\sum_{j=1}^{36} S_2(t^j) = 106{,}034\,8$. The approximation of the power-function of the U-test in this case is

$$P(U = 6/\varrho > 0) \cong P(U = 6/\varrho = 0) + \varrho \left[\frac{\partial}{\partial \varrho} P(U = 6/\varrho > 0)\right]_{\varrho = 0} =$$

$$= 0{\cdot}05 + (N!)^{-1} \varrho \sum_{j=1}^{36} S_2(t^j) = P_{II} .$$

The values P_{II} for $\varrho = 0{\cdot}15$; $0{\cdot}10$; $0{\cdot}05$; $0{\cdot}03$; and $0{\cdot}01$ are given in table 3.

Table 1

i	1	2	3	4	5	6
$a_i = b_i$ (5aa)	$-0{\cdot}969$	$-0{\cdot}781$	$-0{\cdot}313$	$0{\cdot}313$	$0{\cdot}781$	$0{\cdot}969$
$a_i = b_i$ (5b)	$-1{,}27$	$-0{\cdot}64$	$0{\cdot}20$	$0{\cdot}20$	$0{\cdot}64$	$1{\cdot}27$

Table 2

Δ^2	P_I
0·15	0·071 3
0·10	0·064 2
0·05	0·057 1
0·03	0·054 2
0·01	0·051 4

Table 3

ϱ	P_{II}
0·15	0·072 1
0·10	0·064 7
0·05	0·057 4
0·03	0·054 4
0·01	0·051 5

We see that the values P_I and P_{II} differ relatively little although the U-test was constructed for the double-exponential distribution.

References

[1] *Elandt, Regina:* Exact and Approximate Power of the Non-parametric Test of Tendecy. Ann. Math. Stat. 33, 471—481, 1962.
[2] *J. Hájek, Z. Šidák:* Theory of Rank Tests, Academia Praha 1967.
[3] *I. P. Natanson:* Teorija funkcij veščestvennoj peremenoj, Moskva 1957.
[4] Tables of the Binomial Probability Distribution, Nat. Bur. of Stand. Appl. Math. Ser. 6, 1950.

Súhrn

NIEKTORÉ PORADOVÉ TESTY NEZÁVISLOSTI A OTÁZKA ICH SILOFUNKCIE

MILAN KRIŠŤÁK

V článku sa rieši problém testovania nezávislosti dvojíc náhodných veličín $X = W + \Delta Z$, $Y = W^* + \Delta Z$ pomocou lokálne najsilnejších poradových testov v okolí bodu $\Delta = 0$. Veta 1 je uvedená za trocha slabších predpokladov než je v [2] veta II.4.11 (vynecháva sa predpoklad o spojitosti funkcií f' a g' skoro všade). Veta 2 dáva tvar derivácie silofunkcie takýchto testov v bode $\Delta = 0$. Pre dvojne-exponenciálne a normálne rozdelenie náhodných veličín Wa W^* sú uvedené takéto testy. Mediánový U-test je pre dvojne-exponenciálne rozdelenie pri párnych rozsahoch $N = 2n$ podobný s modifikovaným U-testom, ktorým sa zaoberá R. Elandtová v [1], ale pre nepárne rozsahy sú to rôzne testy. Numericky sú vypočítané hodnoty siolofunkcií oboch našich testov v okolí bodov $\Delta = \varrho = 0$.

Author's address: Milan Krišťák, Katedra matematiky a dg. na Stavebnej fakulte SVŠT v Bratislave, Gottwaldovo nám. 2.