Commentationes Mathematicae Universitatis Carolinae

Dana Vorlíčková

Rank tests of independence when samples are drawn from noncontinuous distributions or censored

Commentationes Mathematicae Universitatis Carolinae, Vol. 17 (1976), No. 3, 557--565

Persistent URL: http://dml.cz/dmlcz/105717

COMMENTATIONES MATHEMATICAE UNIVERSITATIS CAROLINAE

17,3 (1976)

RANK TESTS OF INDEPENDENCE WHEN SAMPLES ARE DRAWN FROM NONCONTINUOUS DISTRIBUTIONS OR CENSORED

Dana VORLÍČKOVÁ, Praha

Abstract: In this paper there are considered rank tests of independence when underlying distributions may be noncontinuous or observations are not exactly measurable in some intervals. The asymptotic normality of the test statistics obtained by the method of averaged scores is derived under the hypothesis.

Key words and phrases: Rank test, hypothesis of independence, censored sample, linear rank statistic, averaged scores.

AMS: 62G10 Ref. Ž.: 9.742

---

1. Introduction. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be a random sample from a two-dimensional distribution which is arbitrary. Then, ties of equal X- and Y- observations may occur. Let us denote by $T = (T_1,\ldots,T_g)$ resp. $V = (V_1,\ldots,V_h)$ the vectors of lengths of ties X- resp. Y- observations. Obviously $T_1 +\ldots+ T_g = n = V_1 +\ldots+ V_h$. Let R_i stand for the rank of X_i in the sequence $X_1,\ldots,X_n$, let Q_i stand for the rank of Y_i in the sequence $Y_1,\ldots,Y_n$. Let us denote $(R_1,\ldots,R_n)$ by R and $(Q_1,\ldots,Q_n)$ by Q.

A similar situation occurs when observations may have a continuous distribution but we cannot distinguish exact values of X_i's lying in the same interval (x_{j-1},x_j),

$1 \leq j \leq k$ or of Y_i's from the interval (y_{j-1}, y_j), $1 \leq j \leq m$. Such a sort of a sample we shall call a censored sample. Without loss of generality we can suppose $k = m = 1$. For the construction of vectors of ranks observations from the same interval may be treated as if all of them take on the same value from the interval and form a tie. Then, we may consider the censored two-dimensional sample as a sample from a distribution with marginal distributions functions given by 1.2 from [7] adding the index 1 or 2.

It is obvious now, how censored samples correspond to samples from noncontinuous distributions (cf. [7] for a sample from a one-dimensional distribution) so we may restrict ourselves to the latter ones in the sequel.

2. Rank statistic and its distribution under hypothesis. We say that the two-dimensional observations $(X_1,Y_1),\ldots,(X_n,Y_n)$ satisfy the hypothesis of independence when samples $(X_1,\ldots,X_n)$ and $(Y_1,\ldots,Y_n)$ are independent, the observations from the same sample being independent and identically distributed.

Expressed by the common distribution function of two-dimensional observations (X_i,Y_i), $1 \leq i \leq n$:

$$H(x,y) = H_1(x)\, H_2(y), \quad -\infty < x,y < \infty ,$$

where H_1, H_2 are arbitrary distribution functions.

The linear rank statistic for testing hypothesis of independence has a form $S = \sum_{i=1}^{n} a(R_i)\, b(Q_i)$, where a_i, b_i, $1 \leq i \leq n$, are arbitrary constants, in the case of continuity.

Because the distribution of R and Q in the noncontinuity case depends on the vector T resp. V even under the hypothesis we cannot use the statistic S. Then, analogically to the case of the hypothesis of randomness, we can use the method of randomization for both vectors of ranks or the method of averaged scores for both sequences of scores.

Applying the method of randomization we get the statistic

$$S^* = \sum_{i=1}^{m} a(R_i^*)\, b(Q_i^*),$$

which behaves quite analogically as the statistic S under continuous distributions.

We shall consider the statistic $\bar{S}$ obtained by the method of averaged scores:

$$\bar{S} = \sum_{i=1}^{m} a(R_i,T)\, b\,(Q_i,V), \tag{1}$$

where

$$a(i,T) = \frac{1}{T_k} \sum_{j=T_1+\ldots+T_{k-1}+1}^{T_1+\ldots+T_k} a_j ,$$

$$T_1 +\ldots+ T_{k-1} < i \leq T_1 +\ldots+ T_k,\quad 1 \leq k \leq g,$$

and scores $b(j,V)$, $1 \leq j \leq n$, are defined analogically.

The distribution of $\bar{S}$ can be considered only conditionally given T, V so that its evaluation is difficult. Then, an investigation of asymptotic normality of $\bar{S}$ is very important.

Let us consider scores generated by one of the following ways:

$$a_n(i) = \varphi\left(\frac{i}{n+1}\right),\quad 1 \leq i \leq n, \tag{2}$$

(3) $$a_n(i) = E\varphi(U_n^{(i)}), \quad 1 \le i \le n,$$

where $\varphi(t)$, $0 < t < 1$, is a nonconstant function expressible as a finite sum of monotone square-integrable functions. $U_n^{(i)}$ denotes the i-th order statistic in a sample of a size n from the uniform distribution on (0,1).

Hájek [3] published the following assertion about the special statistic $\bar{S}_n$ as Theorem 6.3:

Theorem 1. Let the hypothesis of independence hold and let $a_n(i)$, $1 \le i \le n$, satisfy (2) or (3). Put

(4) $$\bar{S}_n = \sum_{i=1}^{m} a_n(R_i,T)\, a_n(Q_i,V).$$

Then, for every $\varepsilon > 0$, $\eta > 0$ such $n(\varepsilon,\eta)$ exists, that $n > n(\varepsilon,\eta)$ together with

(5) $$\min\left\{\sum_{i=1}^{m} (a_n(i,T) - \bar{a}_n)^2,\ \sum_{i=1}^{m} (a_n(i,V) - \bar{a}_n)^2\right\} > \eta \sum_{i=1}^{m} (a_n(i) - \bar{a}_n)^2$$

implies

(6) $$\sup_{-\infty<x<\infty} \left| P\left(\bar{S}_n \le E\bar{S}_n + x\sqrt{\operatorname{var}\,(\bar{S}_n/(T,V))/(T,V))}\right) - \Phi(x)\right| < \varepsilon .$$

Let G_1 be the common distribution of $F_1(X_i)$, $1 \le i \le n$, G_2 be the common distribution function of $F_2(Y_i)$, $1 \le i \le n$, under the hypothesis of independence, where F_1 resp. F_2 denotes the common distribution function of X_i's resp. Y_i's. Put $dG_j(\{y\}) = P(F_j(\cdot) = y)$, $j = 1,2$, in the points of

discontinuity of G_j, and equal zero elsewhere. For $j = 1,2$ let us define:

$$\varphi_j(u) = \varphi(u), \text{ when } dG_j(\{G_j^{-1}(u)\}) = 0,$$

$$= \frac{1}{b_j(u) - a_j(u)} \int_{a_j(u)}^{b_j(u)} \varphi(t)dt, \tag{7}$$

$u \in (a_j(u), b_j(u))$, where

$$a_j(u) = G_j^{-1}(u) - dG_j(\{G_j^{-1}(u)\}), \quad b_j(u) = G_j^{-1}(u).$$

Now, when we consider another form of the asymptotic normality, we can prove the following assertion.

Theorem 2. Let the hypothesis of independence hold and let the scores $a_n(i)$, $1 \leq i \leq n$, satisfy (2) or (3) with φ such that $0 < \int_0^1 (\varphi_j(u) - \bar{\varphi}_j)du < \infty$, $\bar{\varphi}_j = \int_0^1 \varphi_j(u)du$, $j = 1,2$, where φ_j is defined by (7). Then, for every $\varepsilon > 0$ there exists such $n(\varepsilon)$ that (6) holds for every $n > n(\varepsilon)$ with probability greater than $1 - \varepsilon$.

Proof. There exists a permutation R' resp. Q' for every vector of ranks R resp. Q such that $R(R')$ and $Q(Q')$ have components ordered according to their magnitude. Putting $Q_0 = Q(R')$ or $R_0 = R(Q')$ it follows from the definition of averaged scores for $\bar{S}_n$ given by (4) that

$$\bar{S}_n = \sum_{i=1}^{n} a_n(i,T)\, a_n(Q_{oi},V) = \sum_{j=1}^{n} a_n(j,V)\, a_n(R_{oj},T).$$

We choose the first form. The vector Q_0 has under the hypothesis the same distribution as the vector Q. The sequence $\{a_n(i,T)\}$ will play the role of regression constants. We prove at first that

(9) $$\frac{\max\limits_{1 \le i \le m} (a_m(i,T) - \bar{a}_m)^2}{\sum\limits_{i=1}^{m} (a_m(i,T) - \bar{a}_m)^2} \longrightarrow 0 \text{ in probability.}$$

Now,

$$1/n \sum_{i=1}^{m} (a_n(i,T) - \bar{a}_n)^2 \xrightarrow{P} \int_0^1 (\varphi_1(u) - \bar{\varphi}_1)^2 du > 0$$

as $n \to \infty$ (cf.[6], the proof of Theorem 3.4). Obviously, $\max (a_n(i,T) - \bar{a}_n)^2 \le \max (a_n(i) - \bar{a}_n)^2$ holds. Then, if scores satisfy (2) we can prove that $1/n \max (a_n(i) - \bar{a}_n)^2 \to 0$ following the pattern of the proof of Theorem 23.b of [9].

Denoting the scores defined by (3) as $a_{n3}(i)$ we have:

$1/n \max (a_{n3}(i) - \bar{a}_{n3})^2 \le 4/n \max (a_{n3}(i) - a_n(i))^2 +$

$+ 4/n \max (a_n(i) - \bar{a}_n)^2 + 2/n (\bar{a}_{n3} - \bar{a}_n)^2,$

where $a_n(i)$ satisfy (2).

$1/n \max (a_{n3}(i) - a_n(i))^2 \le 1/n \sum_{i=1}^{m} (a_{n3}(i) - a_n(i))^2 =$

$= \int_0^1 [a_{n3}(1 + [un]) - a_n(1 + [un])]^2 \to 0$

according to the proof of Theorem V.1.6 a and Lemma V.1.6 a of [4] . Similarly, $1/n (\bar{a}_{n3} - \bar{a}_n)^2 \to 0$ as $n \to \infty$.

It follows from [4], Theorem V.1.4 b and Lemma V.1.6 a that scores defined by (2) or (3) satisfy the condition

$$\int_0^1 (a(1 + [un]) - \varphi(u))^2 du \to 0.$$

Then, applying Theorem 4.2 from [2] we may conclude the proof. Q.E.D.

Remark. Theorems 1 and 2 obviously hold for statis-

tic (1) when both sequences of scores $\{a_n(i)\}$, $\{b_n(i)\}$ satisfy their assumptions.

The assertions of Theorems 1 and 2 give us the asymptotic normality of the conditional distribution with natural parameters so that the corresponding test procedure is nonparametric. The other authors obtained results in which the statistic $n^{-\frac{1}{2}}\bar{S}_n$ is asymptotically normal with the variance depending on a distribution of observations.

We can find such a result in Ruymgaart [5], under the hypothesis and under the fixed alternative for distributions with a finite number of points of discontinuity, or in Behnen [1]. Behnen considered the more general statistic $S = m^{-\frac{1}{2}} \sum_{i=1}^{m} b(R_i, Q_i)$.

Shirahata [8] derived the locally most powerful rank test of the hypothesis of independence for the case of censored samples from continuous distributions when the first n_1 X- and n_2 Y- order statistics were observed (n_1, n_2 are fixed nonrandom numbers). He obtained again - as a test statistic - the statistic of a type (1) and the assertion about the asymptotic normality analogical to Ruymgaart's Theorem 2.1.

3. Distribution of rank statistics under alternatives. Conditions of the asymptotic normality of statistics $n^{-\frac{1}{2}}\bar{S}_n$ under alternatives are derived for continuous observations in Ruymgaart [5]. Samples from general distributions are considered only by Behnen [1] who investigated the statistics

with random ranks and statistics with averaged scores under contiguous alternatives. The alternatives, however, are rather artificial and the author does not introduce an example of their use. It follows from Behnen´s investigations of the asymptotic power that the $\bar{S}$-test is asymptotically better than S^*-test.

References

[1] BEHNEN K.: Asymptotic properties of averaged scores rank tests of independence, Proc. Prague Symp. on asympt. stat. Sept. 1973. Ed. J. Hájek. Karlova universita, Praha 1974.

[2] CONOVER W.J.: Rank tests for one sample, two samples and k samples without the assumption of a continuous distribution function, Ann. Statist. 1 (1973), 1105-1125.

[3] HÁJEK J.: Miscelaneous problems of rank test theory, Nonpar. techniques in stat. inf., ed. M.L. Puri (1970).

[4] HÁJEK J., ŠIDÁK Zb.: Theory of rank tests, Academia, Praha 1967.

[5] RUYMGAART F.H.: Asymptotic normality of nonparametric tests for independence, Ann. Statist. 2(1974), 892-910.

[6] VORLÍČKOVÁ D.: Asymptotic properties of rank tests under discrete distributions, Z. Wahrscheinlichkeitstheorie verw. Geb. 14(1970), 275-289.

[7] VORLÍČKOVÁ D.: Remark on the rank tests in the case of censored samples, Apl. mat. 20(1975), 372-377.

[8] SHIRAHATA Shingo: Locally most powerful rank tests for independence with censored data, Ann. Statist. 3(1975), 241-245.

[9] HÁJEK J., VORLÍČKOVÁ D.: Neparametrické metody (skriptum), SPN 1967.

Matematicko-fyzikální fakulta
Karlova universita
Sokolovská 83, 18600 Praha 8
Československo

(Oblatum 26.2. 1976)