Kybernetika

Dana Vorlíčková

Rank statistics for two-sample location and scale problem for rounded-off data

Kybernetika, Vol. 27 (1991), No. 2, 120--124

Persistent URL: http://dml.cz/dmlcz/124520

Terms of use:

© Institute of Information Theory and Automation AS CR, 1991

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.

This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* http://project.dml.cz

RANK STATISTICS FOR TWO-SAMPLE LOCATION AND SCALE PROBLEM FOR ROUNDED-OFF DATA

DANA VORLÍČKOVÁ

The contribution deals with the rank statistics for testing randomness in the case of two samples which may differ in location and scale simultaneously, and when observations are rounded-off. The asymptotic properties of the vector of adapted linear rank statistics and of their quadratic forms are studied under the hypothesis and under the sequence of contiguous alternatives.

1. INTRODUCTION

To use a properly chosen rank statistic for testing the hypothesis of randomness against the alternative of two samples differing in location and scale simultaneously was proposed by several authors (see Goria, Vorlíčková [1] for a review). In just mentioned paper the authors studied asymptotic properties of vectors and quadratic forms of linear rank statistics for two-sample problem when the underlying distribution is continuous.

Discrete distributions have not location and scale parameters. We shall consider here the case of observations which are rounded-off, without loss of generality to integers.

Let $X_1, \ldots, X_m$ and $X_{m+1}, \ldots, X_N$, $N = m + n$, be two independent random samples. Let I denote the set of integers. Suppose that

$$\mathsf{P}(X_i = k) = d(k, \theta_1, \theta_2) = \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} h(x, \theta_1, \theta_2)\,\mathrm{d}x\,, \quad k \in I\,, \tag{1}$$

where

$$h(x, \theta_1, \theta_2) = \mathrm{e}^{-\theta_2} f(\mathrm{e}^{-\theta_2}(x - \theta_1))\,, \quad -\infty < x < \infty\,,$$

f is a density with a finite Fisher information. Then, the hypothesis of randomness and the alternative of two samples differing in location and scale simultaneously, which in fact concerns the original observations before rounding off, can be expressed with the help of probability functions in the following way:

$$\mathrm{H}_0\colon p(x_1, \ldots, x_N) = \prod_{i=1}^{N} d(x_i, 0, 0)\,, \quad x_i \in I\,. \quad i = 1, \ldots, N\,. \tag{2}$$

$$\text{A: } q(x_1, \ldots, x_N) = \prod_{i=1}^{m} d(x_i, \theta_1, \theta_2) \prod_{i=m+1}^{N} d(x_i, 0, 0), \quad x_i \in I, \tag{3}$$
$$\theta_1 \neq 0 \neq \theta_2, \quad i = 1, \ldots, N.$$

With respect to the fact that observations are integervalued ties may occur. Let jth tie contain t_j observations, $\sum_{j=1}^{i} t_j = T_i$, $i = 1, \ldots, g$, $\sum_{i=1}^{g} t_i = N$. Then, the ordered sample consists of g groups of equal observations $X_{(1)} = \ldots = X_{(t_1)} < X_{(t_1+1)} = \ldots$ $\ldots = X_{(T_2)} < \ldots < X_{(T_{g-1}+1)} = \ldots = X_{(N)}$.
If we define ranks as usually by

$$R_i = \sum_{i=1}^{N} I_{[0,\infty)}(X_i - X_j), \quad i = 1, \ldots, N, \tag{4}$$

we can see that all observations in the jth tie have the same rank $R_i = T_j$, $j = 1, \ldots, g$.

We shall use the method of averaged scores applied to linear rank statistics constructed for testing randomness against the alternative of two samples differing in location or scale, respectively, in the continuous case.

If $a_1, \ldots, a_N$ are arbitrary constants used as scores in a linear rank statistic $\sum_{i=1}^{N} c_i a_{R_i}$ the method of averaged scores leads to a statistic $\bar{S} = \sum_{i=1}^{N} c_i a(R_i, t)$, where

$$a(j, t) = (1/t_k) \sum_{i=T_{k-1}+1}^{T_k} a_i, \quad j = T_{k-1} + 1, \ldots, T_k, \quad k = 1, \ldots, g, \tag{5}$$

which depends on the vector of ties $t = (t_1, \ldots, t_g)$.

Let

$$a_{11} \leqq a_{12} \leqq \ldots \leqq a_{1N}, \tag{6}$$

$$a_{21} = a_{2N} \geqq a_{22} = a_{2,N-1} \geqq \ldots \tag{7}$$

be two sequences of scores, otherwise arbitrary. Put

$$\bar{S}_j = \bar{S}_{jN} = \sum_{i=1}^{m} a_j(R_i, t), \quad j = 1, 2. \tag{8}$$

Let us mention that $\bar{S}_1$, $\bar{S}_2$ are special types of statistic $\bar{S} = \sum_{i=1}^{N} c_i a(R_i, t)$ with $c_i = 1$, $i = 1, \ldots, m$, $c_i = 0$, $i = m + 1, \ldots, N$, and $a_i = a_{1i}$ or $a_i = a_{2i}$, respectively, $i = 1, \ldots, N$.

2. ASYMPTOTIC DISTRIBUTION OF $(\bar{S}_1, \bar{S}_2)$ UNDER H_0

Now, we shall investigate the asymptotic behaviour of the vector $(\bar{S}_1, \bar{S}_2)'$ which enables us to obtain a test statistic for testing H_0 against A given by (2), (3), respectively, with asymptotically χ^2-distribution.

We assume that scores $a_{ki} = a_{kN}(i)$, $i = 1, \ldots, N$, $k = 1, 2$, are generated by some nonconstant square integrable functions $\varphi_k(u)$, $0 < u < 1$, in such a way that

$$\lim_{N\to\infty} \int_0^1 (a_{kN}(1 + [uN]) - \varphi_k(u))^2 \, du = 0 \,, \quad k = 1, 2 \,, \tag{9}$$

and inequalities in (6), (7), respectively, are satisfied. Let D be the distribution function corresponding to the probability function $d(k, 0, 0)$ defined by (1). Put for $j = 1, 2$

$$\varphi_{jD}(u) = \frac{1}{d(k, 0, 0)} \int_{D(k-1)}^{D(k)} \varphi_j(u) \, du \,, u \in (D(k-1), D(k)] \,, \quad k \in I \,.$$

Denote

$$\begin{aligned} \bar{\varphi}_j &= \int_0^1 \varphi_j(u) \, du \,, \quad j = 1, 2 \,, \\ \Delta_{ii} &= \int_0^1 (\varphi_{iD}(u) - \bar{\varphi}_i)^2 \, du \,, \quad i = 1, 2 \,, \\ \Delta_{12} &= \Delta_{21} = \int_0^1 (\varphi_{1D}(u) - \bar{\varphi}_1)(\varphi_{2D}(u) - \bar{\varphi}_2) \, du \,, \\ \bar{a}_{iN} &= (1/N) \sum_{j=1}^{N} a_{iN}(j) \,, \quad i = 1, 2 \,. \end{aligned} \tag{10}$$

Theorem 1. Let $\min(m, n) \to \infty$ as $N \to \infty$ and let (9) hold. Then, the vector with elements $\bar{S}_{1N}, \bar{S}_{2N}$ defined by (8) is under $H_0 = H_{0N}$ asymptotically jointly normal with parameters $(m\bar{a}_{1N}, m\bar{a}_{2N})'$ and $mn\,\Delta/N$, where Δ is a matrix with elements (10).

Proof. It is necessary to prove that $b_1\bar{S}_{1N} + b_2\bar{S}_{2N}$ is asymptotically univariate normal with parameters $b_1 m\bar{a}_{1N} + b_2 m\bar{a}_{2N}$, $(b_1^2\Delta_{11} + b_2^2\Delta_{22} + 2b_1b_2\Delta_{12})\, mn/N$ for all real b_1, b_2. However, $b_1\bar{S}_{1N} + b_2\bar{S}_{2N}$ is a special case of the statistic $S_c = \sum c_i' a(R_i, t)$, where $c_i' = (c_{1i}, c_{2i})$, $a(i, t) = (a_1(i, t), \; a_2(i, t))'$. The assertion follows from Theorem 3.1 of [2], the assumptions of which are satisfied. □

Remark. The matrix Δ with elements (10) depends on the underlying distribution function. For testing purposes it may be replaced by a matrix $\hat{\Delta}$ with elements

$$\hat{\Delta}_{11} = \sum_{i=1}^{N} (a_1(i, t) - \bar{a}_1)^2/(N-1) \,, \quad \hat{\Delta}_{22} = \sum_{i=1}^{N} (a_2(i, t) - \bar{a}_2)^2/(N-1) \,,$$

$$\hat{\Delta}_{12} = \hat{\Delta}_{21} = \sum_{i=1}^{N} (a_1(i, t) - \bar{a}_1)(a_2(i, t) - \bar{a}_2)/(N-1) \,.$$

We can see it following the pattern of Part 3 in [3].

Corollary. Under the assumption of Theorem 1 the quadratic form

$$Q = (\bar{S}_{1N} - m\bar{a}_{1N}, \bar{S}_{2N} - m\bar{a}_{2N})(mn\hat{\Delta}/N)^{-1}(\bar{S}_{1N} - m\bar{a}_{1N}, \bar{S}_{2N} - m\bar{a}_{2N})' \tag{11}$$

has under H_0 asymptotically χ^2-distribution with 2 degrees of freedom.

3. ASYMPTOTIC DISTRIBUTION OF $(\bar{S}_{1N}, \bar{S}_{2N})$ UNDER ALTERNATIVES

We shall study now the asymptotic behaviour of the vector $(\bar{S}_{1N}, \bar{S}_{2N})'$ under a sequence of contiguous alternatives for which an appropriate theory is available. For contiguity of a sequence of alternatives (3) we need some additional assumptions:

(C) $\quad \boldsymbol{\theta} = (\theta_1, \theta_2)' \in \Theta$, Θ is an open set, $\Theta \ni (0, 0)' = \mathbf{0}$,

$$\dot{d}(x, \boldsymbol{\theta}) = (\partial/\partial\theta_1\, d(x, \theta_1, \theta_2), \partial/\partial\theta_2\, d(x, \theta_1, \theta_2))' = (\dot{d}_1(x, \boldsymbol{\theta}), \dot{d}_2(x, \boldsymbol{\theta}))'$$

exists, is continuous in θ_1, θ_2 for every $x \in I$,

information matrix

$$I(\boldsymbol{\theta}) = \mathsf{E}\left[\frac{\dot{d}(X, \boldsymbol{\theta})\, \dot{d}(X, \boldsymbol{\theta})'}{d^2(X, \boldsymbol{\theta})}\right] \quad \text{with diagonal elements } I_{jj}(\boldsymbol{\theta}) \text{ exists,}$$

$$\lim_{\|\boldsymbol{\theta}\| \to 0} I_{jj}(\boldsymbol{\theta}) = I_{jj}(\mathbf{0}) < \infty, \quad \text{where } \|\cdot\| \text{ is Euclidian norm,}$$

$$m = m_N \to \infty\,, \quad n = n_N \to \infty \quad \text{as} \quad N \to \infty\,,$$

$$\boldsymbol{\theta} = \boldsymbol{\theta}_N \in \Theta\,, \quad \|\boldsymbol{\theta}_N\| \to 0\,, \quad \text{as} \quad N \to \infty\,,$$

$$m\|\boldsymbol{\theta}\|^2 \leqq \delta < \infty\,,$$

$$m\boldsymbol{\theta}'\, I(\mathbf{0})\, \boldsymbol{\theta} \to \beta^2 < \infty\,.$$

The sequence of alternatives A_N given by q_N according to (3) is under (C) contiguous to the sequence of hypotheses H_{0N} given by p_N according to (2). It follows from [2], Theorem 4.1.

Theorem 2. Let (9) and (C) hold. Then, the vector with elements $\bar{S}_{1N}$, $\bar{S}_{2N}$ is under A_N asymptotically jointly normal with expectations

$$m\bar{a}_{jN} + \frac{mn}{N} \sum_k \frac{1}{d(k, \mathbf{0})} \left[\dot{d}_1(k, \mathbf{0})\, \theta_1 + \dot{d}_2(k, \mathbf{0})\, \theta_2\right] \int_{D(k-1)}^{D(k)} \varphi_j(u)\, \mathrm{d}u\,, \quad j = 1, 2\,,$$

and variance matrix $mn\Delta/N$, where Δ has elements (10).

Proof. We proceed in the same way as in the proof of Theorem 1 using Theorem 4.1 from [2] instead of Theorem 3.1. □

Corollary. Under the assumptions of Theorem 2 the quadratic form (11) has asymptotically noncentral χ^2-distribution with 2 degrees of freedom. The parameter of noncentrality is given by the difference between asymptotical expectations under the alternative and the hypothesis.

4. REMARKS TO THE ASYMPTOTIC DISTRIBUTION OF RANDOMIZED AND MIDRANK STATISTICS

When ties are present also other methods of treatment of ties can be used. Randomization needs an additional sample from the uniform distribution over (0, 1). Ordering observations $X_i + U_i$ (or $R_i + U_i$) instead of X_i, $i = 1, \ldots, N$, we obtain a vector of ranks $R_1^*, \ldots, R_N^*$ which has the same properties as a vector of ranks in the continuous case so that for rank statistics depending on R^* we receive nothing new. Moreover, the R^*-tests are usually asymptotically less powerful then averaged-scores tests (see [3], Part 5).

We meet another situation when midranks are used. In this case we work with midranks $\tilde{R}_i = \frac{1}{2}(T_{j-1} + 1 + T_j)$, if $R_i = T_j$, $j = 1, \ldots g$, $i = 1, \ldots, N$, and with statistics

$$\tilde{S}_j = \sum_{i=1}^{m} a_j(\tilde{R}_i), \quad j = 1, 2.$$

It can be expected that $(\tilde{S}_1, \tilde{S}_2)'$ behaves similarly as the vector of the corresponding averaged-scores statistics, however, the assumptions have to be modified slightly. Scores $a_j(i)$ are generated by nonconstant functions φ_j, $j = 1, 2$, which are continuous in points $\frac{1}{2}(D(k-1) + D(k))$, $k \in I$, $a_j(\frac{1}{2}(1 + [2uN])) \to \varphi_j(u)$, $0 < u < 1$, as it can be deduced from [4]. The role of functions $\varphi_{jD}(u)$ would be played by functions $\varphi_{jm}(u)$, $j = 1, 2$, where

$$\varphi_{jm}(u) = \varphi_j(\tfrac{1}{2}(D(k-1) + D(k))), \quad u \in (D(k-1), D(k)], \quad k \in I.$$

(Received June 4, 1990.)

REFERENCES

[1] M. N. Goria and D. Vorlíčková: On the asymptotic properties of rank statistics for the two-sample location and scale problem. Apl. mat. *30* (1985), 425—434.

[2] A. R. Padmanabhan and M. L. Puri: Theory of nonparametric statistics for rounded-off data with applications. Statistics (Math. Operationsforsch. Statist. Ser. Statist.) *14* (1983), 301—349.

[3] D. Vorlíčková: Asymptotic properties of rank tests under discrete distributions. Z. Warsch. verw. Geb. *14* (1970), 275—289.

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

RNDr. Dana Vorlíčková, CSc., katedra pravděpodobnosti a matematické statistiky matematicko-fyzikální fakulty Univerzity Karlovy (Department of Probability and Statistics, Faculty of Mathematics and Physics — Charles University), Sokolovská 83, 186 00 Praha 8. Czechoslovakia.