Aplikace matematiky

Dana Vorlíčková

Exact slopes of the rank statistics for the two-sample case under discrete distributions

Aplikace matematiky, Vol. 26 (1981), No. 6, 426–431

Persistent URL: http://dml.cz/dmlcz/103932

EXACT SLOPES OF THE RANK STATISTICS FOR THE TWO-SAMPLE CASE UNDER DISCRETE DISTRIBUTIONS

Dana Vorlíčková

(Received November 5, 1979)

1. INTRODUCTION

Let us suppose that samples of a size N are drawn from a purely discrete distribution, and that observations are integervalued. The set of integers C could be replaced by an arbitrary countable set. Moreover, the conclusions will continue to hold even for an arbitrary noncontinuous distribution.

For $u(x)$ defined as

$$u(x) = 1\,, \quad x \geqq 0\,, \quad u(x) = 0\,, \quad x < 0\,,$$

we put

$$R_i = \sum_{j=1}^{N} u(X_i - X_j)\,. \tag{1}$$

$R = (R_1, \ldots, R_N)$ will stand for the vector of ranks and $t = (t_1, \ldots, t_g)$ is the vector, the jth component of which is the number of equal observations in the jth tie.

We shall study the linear rank statistics for testing the hypothesis of randomness against the alternative of two samples when ties are present from the point of view of large deviations and exact slopes.

Let a_i, $1 \leqq i \leqq N$, be arbitrary scores and c_i, $1 \leqq i \leqq N$, regression constants. We consider two ways of treatment of ties: randomization by the help of a supplementary random sample $U_1, \ldots, U_N$ from the uniform distribution over the interval $(0, 1)$, and the method of averaged scores.

Randomization leads to the vector of ranks R^*, where

$$R_i^* = \sum_{j=1}^{N} u(X_i + U_i - X_j - U_j)\,, \quad 1 \leqq i \leqq N\,, \tag{2}$$

and the linear rank statistics

$$S_N^* = \sum_{i=1}^{N} c_i\, a(R_i^*)\,. \tag{3}$$

When the method of averaged scores is used we have

$$(4)\quad a(i, t) = \frac{1}{t_k} \sum_{j=T_{k-1}+1}^{T_k} a_j\,, \quad T_{k-1} = t_1 + \ldots + t_{k-1} < i \leqq t_1 + \ldots + t_k = T_k\,,$$

and the rank statistics

$$(5)\qquad \bar{S} = \sum_{i=1}^{N} c_i a(R_i, t)\,.$$

2. THE LAW OF LARGE NUMBERS

Let us fix two distribution functions of the discrete types F, G, and a number λ, $0 < \lambda < 1$. Throughout the paper let $X_1, \ldots, X_m$ be a random sample with the common distribution function F, $Y_1, \ldots, Y_n$ a random sample with the common distribution function G. Then, $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ forms a random sample from the distribution with the d.f. H, where

$$H = \frac{m}{N} F + \frac{n}{N} G\,, \quad m + n = N\,.$$

Put

$$\bar{\varphi} = \int_0^1 \varphi(u)\,\mathrm{d}u\,,$$

$$(6)\qquad a_N(i, \varphi) = E\,\varphi(U^{(i)})\,,$$

where $U^{(1)} < U^{(2)} < \ldots < U^{(N)}$ is an ordered sample from the uniform distribution over $(0, 1)$, and $\varphi(u)$, $0 < u < 1$, is an arbitrary nonconstant square-integrable function. The following theorem is a form of the law of large numbers for the two-sample statistics of the type (3).

Theorem 1. *Let us assume that the functions $a_N(1 + [uN], \varphi)$ have uniformly bounded variation on any closed subinterval of $(0, 1)$. Let $R_1^*, \ldots, R_N^*$ be the ranks corresponding to $(X_1 + U_1, \ldots, X_m + U_m, Y_1 + U_{m+1}, \ldots, Y_n + U_N)$, where $n + m = N$, $m/N \to \lambda$, $0 < \lambda < 1$, and $S_N^* = \sum_{i=1}^{m} a_N(R_i^*, \varphi)$. Then,*

$$(7)\qquad \frac{1}{N} S_N^* \to \lambda \int_0^1 \varphi(u)\,\bar{f}(u)\,\mathrm{d}u$$

with probability 1, *where*

$$(8)\qquad \bar{f}(u) = \frac{\mathrm{d}}{\mathrm{d}u} F^*(H^{*-1}(u))\,,$$

$$(9)\qquad H^* = \lambda F^* + (1 - \lambda)\,G^*\,,$$

$$(10)\qquad F^* = F \otimes \mathcal{U}\,, \quad G^* = G \otimes \mathcal{U}\,,$$

$\mathcal{U}$ is the distribution function of the uniform distribution over $(0, 1)$, *and the symbol* $\otimes$ *denotes the convolution.*

Proof. The distribution functions F^*, G^* have densities f^*, g^*. The ranks R_i^*, $1 \leqq i \leqq N$, correspond to the observations satisfying the alternative of two samples with densities f^*, g^*. The scores given by (6) satisfy the condition (3) from [1]. Therefore, the assertion follows from Theorem 1, [1]. Q.E.D.

Let us now consider the statistic $\bar{S}_N = \sum_{i=1}^{m} a_N(R_i, t)$ with scores $a_N(i, t)$ defined by (4), where $a_i = a_N(i)$ satisfy

$$\int_0^1 (a_N(1 + [uN]) - \varphi(u))^2 \, du \to 0 \quad \text{as} \quad N \to \infty\,, \tag{11}$$

for a nonconstant and square integrable function φ, defined on $(0, 1)$. Let us denote $I_h = [H(h - 0), H(h))$, $h \in C$, $H = \lambda F + (1 - \lambda) G$, $\lambda = \lim_{N \to \infty} m/N$, $0 < \lambda < 1$. If there exists $h_0 \in C$ such that $H(x) = 0$ for all $x < h_0$, put $I_{h_0} = (H(h_0 - 0), H(h_0))$. Denoting by μ the Lebesgue measure let us define

$$\begin{aligned} \varphi_H(u) &= \frac{1}{\mu I_h} \int_{I_h} \varphi(v) \, dv\,, \quad u \in I_h\,, \quad \mu I_h > 0\,, \\ &= 0\,, \qquad\qquad\qquad\quad u \in I_h\,, \quad \mu I_h = 0\,. \end{aligned} \tag{12}$$

Theorem 2. *Let us assume that the functions* $a_N(1 + [uN])$ *have uniformly bounded variation on any closed subinterval of* $(0, 1)$, *and that*

$$0 < \int_0^1 (\varphi_H(u) - \bar{\varphi})^2 \, du < \infty\,, \tag{13}$$

for φ_H *defined by* (12). *Let* $R = (R_1, \ldots, R_N)$ *be the vector of ranks corresponding to* $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$, *where* $m + n = N$, $m/N \to \lambda$, $N \to \infty$, $0 < \lambda < 1$. *Then,*

$$\frac{\bar{S}_N}{N} \to \lambda \int_0^1 \varphi_H(u) \bar{f}(u) \, du + \lambda \int_0^1 (\varphi(u) - \varphi_H(u)) \, du \tag{14}$$

in probability corresponding to F, G, *where* $\bar{f}$ *is defined by* (8).

Proof. The vector t gives for every N a disjoint decomposition of the interval $(0, 1)$ corresponding to the empirical distribution function H_N based on the vector $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$, $\{(0, t_1/N), [t_1/N, T_2/N), \ldots, [T_{g-1}/N, 1)\}$, and $H_N \to H = \lambda F + (1 - \lambda) G$ uniformly with probability 1, according to the Giivenko-Cantelli lemma. Then, following the pattern of the proof of Theorem 3, [4], we obtain

$$\frac{\bar{S}_N - S_{NH}^* - E\bar{S}_N + ES_{NH}^*}{\sigma_H} \to 0 \tag{15}$$

in probability (corresponding to the d.f. H or F, G, respectively), where

$$S^*_{NH} = \sum_{i=1}^{m} a_N(R^*_i, \varphi_H) ,$$

$$ES^*_{NH} = \frac{m}{N} \sum_{i=1}^{N} a_N(i, \varphi_H) ,$$

$$\sigma^2_H = \frac{1}{N-1} \frac{mn}{N} \int_0^1 (\varphi_H(u) - \bar{\varphi})^2 \, du ,$$

$$E\bar{S}_N = \frac{m}{N} \sum_{i=1}^{m} a_N(i) .$$

Utilizing the assumption (13) and $m/N \to \lambda$ we have that

$$\frac{\bar{S}_N - S^*_{NH} - E\bar{S}_N + ES^*_{NH}}{N} \to 0 \quad \text{in probability} .$$

Now,

$$\lim_{P} \frac{\bar{S}_N - S^*_{NH}}{N} = \lim_{N \to \infty} \frac{E\bar{S}_N - ES^*_{NH}}{N} = \lambda \int_0^1 (\varphi(u) - \varphi_H(u)) \, du , \tag{16}$$

which together with Theorem 1 concludes the proof. Q.E.D.

3. PROBABILITY OF LARGE DEVIATIONS AND EXACT SLOPES

Proceeding in the same way as in the proof of Theorem 1 we have, according to [3],

$$\lim_{N \to \infty} \frac{1}{N} \log P\left(\frac{1}{N} S^*_N > \varrho_N\right) = -b(\lambda, \varrho) , \tag{17}$$

for any sequence $\varrho_N \to \varrho$, $N \to \infty$, where for

$$\lambda \int_0^1 \varphi(u) \, du < \varrho < \sup_A \left\{\int_A \varphi(u) \, du : \int_A du = \lambda\right\}, \tag{18}$$

$b(\lambda, \varrho)$ equals

$$b(\lambda, \varrho) = \varrho a + (1 - \lambda) \log b - \int_0^1 \log(\exp\{a \, \varphi(u)\} + b) \, du - \tag{19}$$
$$- \lambda \log \lambda - (1 - \lambda) \log(1 - \lambda) ,$$

a, b being the unique solution of the system of equations

$$\int_0^1 (1 + b \exp\{-a \, \varphi(u)\})^{-1} \, du = \lambda , \tag{20}$$

$$\int_0^1 \varphi(u) (1 + b \exp\{-a \, \varphi(u)\})^{-1} \, du = \varrho .$$

The function $b(\lambda, \varrho)$ is continuous in ϱ . φ is the score-generating function from (6). Hence, the (strong) exact slope of S_N^* is

$$2\,b(\lambda, \varrho)|_{\varrho=\lim S_N^*/N} = 2\,b\left(\lambda, \lambda \int_0^1 \varphi(u)\,\bar{f}(u)\,\mathrm{d}u\right),$$

where lim denotes the limit with probability 1 and $\bar{f}$ is given by (8).

Further, in virtue of (16), (17) and the continuity of $b(\lambda, \varrho)$ and provided $\bar{S}_N$ satisfies the assumptions of Theorem 2, we have

$$\lim_{N\to\infty} \frac{1}{N} \log P\left(\frac{\bar{S}_N}{N} - \lambda \int_0^1 (\varphi(u) - \varphi_H(u))\,\mathrm{d}u > \varrho_N\right) = -b_H(\lambda, \varrho)\,,$$

or, equivalently,

$$(21)\quad \lim_{N\to\infty} \frac{1}{N} \log P\left(\frac{\bar{S}_N}{N} > \varrho_N + \frac{m}{N}\int_0^1 (a_N(1 + [uN]) - a_N(1 + [uN], \varphi_H))\,\mathrm{d}u\right) = \\ = -b_H(\lambda, \varrho)\,,$$

where $b_H(\lambda, \varrho)$ for $\varrho = \lim \varrho_N$ satisfying (18) is given by (19) and (20) with the function φ replaced by φ_H defined by (12). Then, the (weak) exact slope of $\bar{S}_N$ equals

$$2\,b_H\left(\lambda, \lambda \int_0^1 \varphi_H(u)\,\bar{f}(u)\,\mathrm{d}u\right).$$

The validity of (18) can be achieved by an appropriate choice of the score-generating function.

Remark. If we introduce $L_N(x) = P(S_N/N > x)$ and consider such alternatives that large values of S_N lead to rejection of the hypothesis, then, according to [2], the strong exact slope of S_N is the limit with probability 1 of $-2(1/N)\log L_N$, the weak exact slope is the limit of $-2(1/N)\log L_N$ in probability. The definition used by Woodworth corresponds to the latter kind of convergence.

Comparing the exact slopes of two test statistics we can evaluate the Bahadur efficiency of tests.

The relations (17) and (21), respectively, give also probabilities of large deviations, more precisely $\lim (1/N) \log P(S_N^*/N > \varrho^*)$, for $\varrho^* = \lim S_N^*/N$, and $\lim (1/N)$ . . $\log P(\bar{S}_N/N > \bar{\varrho})$ for $\bar{\varrho} = \lim\limits_{P} \bar{S}_N/N$.

References

[1] *J. Hájek:* Asymptotic sufficiency of the vector of ranks in the Bahadur sense. Ann. Statist. 2 (1974), 1105—1125.

[2] *M. Raghavachari:* On the theorem of Bahadur on the rate of convergence of test statistics. Ann. Math. Statist. 41 (1970), 1695—1699.

[3] *G. G. Woodworth:* Large deviations and Bahadur efficiency of linear rank statistics. Ann. Math. Statist. 41 (1970), 251—283.

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

Souhrn

PŘESNÝ SPÁD POŘADOVÝCH STATISTIK PRO PŘÍPAD DVOU VÝBĚRŮ Z DISKRETNÍCH ROZDĚLENÍ

Dana Vorlíčková

Uvažujme lineární pořadové statistiky pro test hypotézy náhodnosti proti obecné alternativě dvou výběrů za předpokladu, že oba výběry pocházejí z diskrétních (celočíselných) rozdělení, a to buď se znáhodněnými pořadími nebo se zprůměrovanými skóry. S využitím toho, že statistiky se znáhodněnými pořadími se chovají tak, jakoby výběry pocházely ze spojitých rozdělení, odvodíme pro ně pomocí [1] a [3] zákon velkých čísel a přesný spád. Pomocí těchto výsledků pak získáme, když aplikujeme postup z [4], slabý zákon velkých čísel a přesný spád pro statistiky se zprůměrovanými skóry.

Author's address: RNDr. *Dana Vorlíčková,* CSc., katedra pravděpodobnosti a matematické statistiky MFF UK, Sokolovská 83, 186 00 Praha 8.