Dana Vorlíčková

Remark on the rank tests in the case of censored samples

*Aplikace matematiky*, Vol. 20 (1975), No. 5, 372–377

Persistent URL: `http://dml.cz/dmlcz/103602`

# REMARK ON THE RANK TESTS IN THE CASE OF CENSORED SAMPLES

DANA VORLÍČKOVÁ

(Received June 10, 1974)

## 1. INTRODUCTION

The problem is to test the hypothesis of randomness of the observations $X_1, \ldots, X_N$ against a regression alternative by means of a rank test procedure. We know exact values of some observations but about the other ones we can only say that they lie in a known open interval.

Let us consider $k$ open intervals $(y_0, y_1), \ldots, (y_{k-1}, y_k)$, $y_0 < y_1 < \ldots < y_k$, and let $N_j$ observations lie in the $j$-th interval, $1 \leqq j \leqq N$, where $N_1, \ldots, N_k$ are random variables, $0 \leqq N_j \leqq N$, $\sum_{j=1}^{k} N_j \leqq N$. The exact values of the observations lying in the intervals $(y_{j-1}, y_j)$. $1 \leqq j \leqq k$, are unknown, the exact values of the other ones are known. It is possible to have $y_0 = -\infty$ or $y_k = \infty$. If $k = 1$ we have the situation discussed in [3] and [2]. In the former paper some asymptotically efficient rank test procedures are derived for the hypothesis of randomness against two- and $k$-sample alternatives. In the latter paper the locally most powerful rank test of the hypothesis of randomness against the alternative of two samples is derived and the asymptotic distribution of the two sample statistic is established.

We try to find a rank test based on the censored sample described above in another way than it is done in [2] and [3]. We see that it is irrelevant that we cannot distinguish the values of observations lying in the same interval $(y_{j-1}, y_j)$, $1 \leqq j \leqq k$, but it is essential that we cannot assign the exact ranks to them. We only know the set of possible values of ranks of these observations. Thus, we are in a similar situation as in the case of ties when a sample is taken from a noncontinuous distribution.

Without loss of generality we may suppose $k = 1$. Further, let $X_1, \ldots, X_N$ have a continuous distribution function $F$. We observe the following order statistics:

$$(1.1) \qquad X^{(1)} < \ldots < X^{(R_j)}, \quad X^{(R_j+N_1+1)} < \ldots < X^{(N)},$$

where $R_i$ stands for the rank of $X_i$, with probability 1. $N_1$ observations lie in the interval $(y_0, y_1) = I$ and we can say about their ranks $R_i$ only that $R_j < R_i \leqq R_j + N_1$. We shall treat the observations from $I$ as if all of them take on the same value $y \in I$ with the probability $F(y_1) - F(y_0)$ and form a tie. Then the censored sample under consideration behaves as a sample of random variables with the following common distribution function $H$:

$$(1.2) \qquad \begin{aligned} H(x) &= F(x), \quad x < y_0, x > y_1, \\ &= F(y_0), \quad y_0 \leqq x \leqq y, \\ &= F(y_1), \quad y < x \leqq y_1. \end{aligned}$$

The number $y \in I$ is arbitrary but fixed.

For this situation Conover's work [1] about rank tests under noncontinuous distributions is applicable.

Let $c_1, \ldots, c_N, a_1, \ldots, a_N$ be arbitrary real constants. Applying the method of randomization we get for the observations from $I$ the ranks $R_i^*$, $R_i^* = R_j + 1, R_j + 2, \ldots, R_j + N_1$, respectively, and for the test of the hypothesis of randomness $H_0$:

$$P(X_1 \leqq x_1, \ldots, X_N \leqq x_N) = \prod_{i=1}^{N} P(X_1 \leqq x_i) = \prod_{i=1}^{N} F(x_i)$$

we have the linear rank statistic

$$(1.3) \qquad S^* = \sum_{i=1}^{N} c_i\, a(R_i^*) = \sum_{X_i \notin I} c_i\, a(R_i) + \sum_{X_i \in I} c_i\, a(R_i^*).$$

Applying the method of averaged scores we obtain for testing $H_0$ the statistic

$$(1.4) \qquad \bar{S} = \sum_{i=1}^{N} c_i\, a(R_i, R),$$

where

$$(1.5) \qquad \begin{aligned} R_i &= \text{number of } X_j < X_i, \quad 1 \leqq j \leqq N, \quad X_i \notin I, \\ &= R_j + N_1, \quad X_i \in I, \end{aligned}$$

and

$$(1.6) \qquad \begin{aligned} a(i, R) &= \frac{1}{N_1} \sum_{k=R_j+1}^{R_j+N_1} a(k), \quad i = R_j + N_1, \\ &= a(i), \qquad\qquad i \neq R_j + N_1, \end{aligned}$$

so that

$$(1.7) \qquad \bar{S} = \sum_{X_i \notin I} c_i\, a(R_i) + a(R_j + N_1, R) \sum_{X_i \in I} c_i.$$

The statistic (1.7) coincides with the statistic (3.2) of [2] in the case of two samples with the respective scores and with $c_i = 1$ whenever $X_i$ is from the first sample, $c_i = 0$ otherwise, $1 \leqq i \leqq N$.

## 2. ASYMPTOTIC DISTRIBUTION OF TEST STATISTICS UNDER $H_0$

Let $\varphi(u)$, $0 < u < 1$, (throughout the paper) be an arbitrary nonconstant square-integrable function. Let the scores $a_i$, $1 \leqq i \leqq N$, satisfy

$$\int_0^1 (a(1 + [uN]) - \varphi(u))^2 \, du \to 0 \,, \quad \text{as} \quad N \to \infty \,, \tag{2.1}$$

and let for the regression constants $c_i$, $1 \leqq i \leqq N$,

$$\sum_{i=1}^{N} (c_i - \bar{c})^2 / \max_{1 \leqq i \leqq N} (c_i - \bar{c})^2 \to \infty \quad \text{as} \quad N \to \infty \tag{2.2}$$

hold.

**Theorem 2.1.** *Under* $H_0$, (2.1), *and* (2.2), *the statistic* (1.3) *is asymptotically normal* $((1/N) \sum_{i=1}^{N} c_i \sum_{i=1}^{N} a_i, (1/(N-1)) . \sum_{i=1}^{N} (c_i - \bar{c})^2 \sum_{i=1}^{N} (a_i - \bar{a})^2)$.

Proof. The assertion follows from [1], Theorem 4.4.

Denote by $G$ the common distribution function of $Y_i = H(X_i)$, $1 \leqq i \leqq N$, where $X_i$'s have the distribution function $H$ given by (1.2), under $H_0$. Let $G(\{y\}) = P(Y = y)$ in discontinuity points of $G$ and equal zero elsewhere. Let us define

$$\begin{aligned} \varphi_a(u) &= \varphi(u) \quad \text{if} \quad G(\{G^{-1}(u)\}) = 0 \,, \\ &= \frac{1}{b(u) - a(u)} \int_{a(u)}^{b(u)} \varphi(t) \, dt \,, \quad u \in (a(u), b(u)) \,, \end{aligned} \tag{2.3}$$

where $a(u) = G^{-1}(u) - G(\{G^{-1}(u)\})$ and $b(u) = G^{-1}(u)$.

Put $\bar{\varphi} = \int_0^1 \varphi(t) \, dt$.

**Theorem 2.2.** *Under* $H_0$, (2.1), (2.2), *and* $\int_0^1 (\varphi_a(u) - \bar{\varphi})^2 \, du > 0$, *the conditional distribution of the statistic* (1.7) *given* $N_1$ *observations lie in* $I$ *is asymptotically normal* $(N\bar{a}\bar{c}, \operatorname{var}(\bar{S} \mid R))$, *where*

$$\operatorname{var}(\bar{S} \mid R) = \frac{1}{N-1} \sum_{i=1}^{N} (c_i - \bar{c})^2 \sum_{j=1}^{N} (a(j, R) - \bar{a})^2 \,, \tag{2.4}$$

*with* $a(j, R)$ *given by* (1.5).

Proof. The assertion follows from [1], Theorem 4.2.

## 3. ASYMPTOTIC DISTRIBUTION OF TEST STATISTICS UNDER ALTERNATIVES

Let us consider the alternative

$$P(X_1 \leqq x_1, \ldots, X_N \leqq x_N) = \prod_{i=1}^{N} F(x_i, \theta_i), \tag{3.1}$$

where $\theta_i$, $1 \leqq i \leqq N$, are real parameters, satisfying

$$\max_{1 \leqq i \leqq N} \theta_i^2 \to 0 \quad \text{as} \quad N \to \infty, \tag{3.2}$$

and $F$ is a continuous distribution function. The censored sample, which is observed, will be considered as a sample of random variables with the distribution functions $H(., \theta_i)$ given by (1.2) with $F(x) = F(x, \theta_i)$, $1 \leqq i \leqq N$.

Put

$$h(x, \theta) = \frac{\mathrm{d}H(x, \theta)}{\mathrm{d}H(x, 0)} \tag{3.3}$$

and define the generalized Fisher's information as

$$I(H, \theta) = \int_{-\infty}^{\infty} \left[\frac{(\partial/\partial\theta)\, h(x, \theta)}{h(x, \theta)}\right]^2 \mathrm{d}H(x, \theta). \tag{3.4}$$

Suppose

$$0 < \lim_{\theta \to 0} I(H, \theta) = I(H) < \infty, \tag{3.5}$$

and

$$\lim_{N \to \infty} I(H) \sum_{i=1}^{N} \theta_i^2 = b^2, \quad 0 < b^2 < \infty. \tag{3.6}$$

Let $J$ be an open interval containing zero. We shall consider a family of $h$ given by (3.3) satisfying the following conditions for a.a. $\theta \in J$ with respect to $H(x, 0)$:

(3.7) a) $h(x, \theta)$ exists,

b) $$\frac{\partial}{\partial\theta} h(x, \theta)\Big|_{\theta=0} = \lim_{\theta \to 0} \frac{h(x, \theta) - h(x, 0)}{\theta} \text{ exists},$$

c) $$h(x, 0) = \lim_{\theta \to 0} h(x, \theta) \text{ exists}.$$

A family of population distribution functions $F$ is determined by the conditions (3.7), too, through the relation (1.2) for $F$ and $H$.

**Theorem 3.1.** *Let* (3.2), (3.5), (3.6), (3.7) *hold. Then, under* (3.1), (2.1), *and* (2.2), *the statistic* (1.3) *is asymptotically normal with parameters*

$$N\bar{c}\bar{a} + \sum_{i=1}^{N} c_i \theta_i \int_0^1 \varphi(u)\, \varphi(u, H, 0)\, \mathrm{d}u, \tag{3.8}$$

*and*

$$\sum_{i=1}^{N} (c_i - \bar{c})^2 \int_0^1 (\varphi(u) - \bar{\varphi})^2 \, du \,, \tag{3.9}$$

*where*

$$\varphi(u, H, 0) = \frac{(\partial/\partial\theta)\, h(H^{-1}(u, \theta), \theta)|_{\theta=0}}{h(H^{-1}(u, 0), 0)} \,.$$

Proof. The assertion follows from [1], Theorem 8.3 and the proof of Theorem 4.4.

Let $\varphi_a$ be defined by (2.3) with $G$ corresponding to $H(x, 0)$ and let

$$\int_0^1 (\varphi_a(u) - \bar{\varphi})^2 \, du > 0. \tag{3.10}$$

**Theorem 3.2.** *Let* (3.2), (3.5), (3.6), (3.7) *hold. Then, under* (3.1), (2.1), (2.2), *and* (3.10), *the statistic* (1.7) *is asymptotically normal with parameters* (3.8) *and*

$$\sum_{i=1}^{N} (c_i - \bar{c})^2 \int_0^1 (\varphi_a(u) - \bar{\varphi})^2 \, du \,.$$

Proof. The assertion follows from [1], Theorem 8.3 and the proof of Theorem 4.2.

Remark. Conover's conclusions about the efficiency of the tests continue to hold even for rank tests in the case of censored samples.

The considerations introduced above are obviously applicable also in the case of noncontinuous distribution functions $F$ of $X_i$, $1 \leqq i \leqq N$. It is only necessary to take into account, in addition, that ties of exactly measurable observations may occur.

A method analogical to the described one may be also used for censored samples in the case of testing symmetry but we omit it because of its similarity to the case of the hypothesis of randomness.

*References*

[1] *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.

[2] *Johnson R. A., Mehrotra K. G.:* Locally most powerful rank tests for the two sample problem with censored data. Ann. Math. Statist. *43* (1972), 823—831.

[3] *Peto R., Peto J.:* Asymptotically efficient rank invariant test procedures. J. Roy. Statist. Soc. Ser. A, *135* (1972), 185—206.

Souhrn

# POZNÁMKA K POŘADOVÝM TESTŮM V PŘÍPADĚ CENSOROVANÝCH VÝBĚRŮ

DANA VORLÍČKOVÁ

Uvažujme náhodný výběr, o jehož některých pozorováních víme pouze, že leží v intervalu $(y_{j-1}, y_j)$, $1 \leqq j \leqq k$, $y_0 < \ldots < y_k$, přičemž může být $y_0 = -\infty$ nebo $y_k = \infty$. U ostatních pozorování známe přesné hodnoty. U pozorování, která padnou do téhož intervalu, nelze rozlišit pořadí. Proto při hledání pořadového testu postupujeme tak, jakoby pozorování padnuvší do téhož intervalu nabyla téže hodnoty s pravděpodobností rovnou rozdílu hodnot distribuční funkce v koncových bodech intervalu. S censorovaným výběrem pak zacházíme jako s výběrem z nespojitého rozdělení, v němž nastaly shody, a při konstrukci pořadových statistik užijeme buď metody znáhodnění nebo metody průměrných skórů. S využitím teorie z [1] pak dostáváme asymptotické rozdělení těchto statistik za hypotézy náhodnosti i při kontiguitních alternativách.

Analogického postupu by se dalo použít i pro testování hypotézy symetrie.

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