The Annals of Statistics
1975, Vol. 3, No. 1, 241–245

LOCALLY MOST POWERFUL RANK TESTS FOR INDEPENDENCE WITH CENSORED DATA

BY SHINGO SHIRAHATA
Osaka University

In this paper locally most powerful rank tests for independence with censored data for a one-parameter family are derived. The statistic derived has discrete score functions and its asymptotic normality follows from a theorem essentially given by Ruymgaart [6].

1. Introduction. Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be a random sample of size n from a population with bivariate continuous distribution function (df) $H(x, y)$ having continuous marginal df's $F(x)$ and $G(y)$. We want to test whether the first coordinate variable X and the second coordinate variable Y are independent or not when H, F and G are unknown and yet, for some reason, we cannot wait until observations are performed for all sample units. Such situations may occur, for example, when we treat efficacies of two different drugs or lifetimes of two physical systems.

There are many censoring types in multivariate analysis; for example Watterson [8] subdivided censoring into three distinct types. In this paper we consider the censoring scheme of type A in [8] in which only the first n_1 ordered observations in the first coordinate and only the first n_2 ordered observations in the second coordinate are available for some fixed integers n_1 and n_2. In the special case $n_1(n_2) = n$, the scheme reduces to type C(B) in [8].

Let us denote the ith order statistics among $(X_1, \cdots, X_n)$ and $(Y_1, \cdots, Y_n)$ by X_{in} and Y_{in} respectively. The censoring scheme amounts to using only the ranks R_i of the $X_i \leqq X_{n_1n}$ and the ranks Q_i of the $Y_i \leqq Y_{n_2n}$. In this way we obtain a pair $(R_*, Q_*) = (R_{*1}, \cdots, R_{*n}, Q_{*1}, \cdots, Q_{*n})$ of lacunary rank vectors, where we put

$$(1.1)\qquad \begin{aligned} R_{*i} &= \#\{j \mid X_j \leqq X_i\} \quad \text{for} \quad X_i \leqq X_{n_1n}\,, \qquad R_{*i} = * \quad \text{for} \quad X_i > X_{n_1n}\,, \\ Q_{*i} &= \#\{j \mid Y_j \leqq Y_i\} \quad \text{for} \quad Y_i \leqq Y_{n_2n}\,, \qquad Q_{*i} = * \quad \text{for} \quad Y_i > Y_{n_2n}\,, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \cdots, n\,. \end{aligned}$$

These lacunary rank vectors may be completed by replacing each $*$ by an appropriately chosen natural number. Given (R_*, Q_*) we define an arbitrary completion $(\bar{R}, \bar{Q}) = (\bar{R}_1, \cdots, \bar{R}_n, \bar{Q}_1, \cdots, \bar{Q}_n)$ to be a pair of permutations of the numbers $(1, \cdots, n)$ such that moreover $\bar{R}_i = R_{*i}$ for i with $X_i \leqq X_{n_1n}$ and $\bar{Q}_i = Q_{*i}$ for i with $Y_i \leqq Y_{n_2n}$. The set of all possible completions of the pair (R_*, Q_*) will be denoted by

$$(1.2)\qquad C(R_*, Q_*) = \{(\bar{R}, \bar{Q}) \mid (\bar{R}, \bar{Q}) \text{ is a completion of } (R_*, Q_*)\}\,.$$

Received July 1973; revised February 1974.
AMS 1970 *subject classifications.* Primary 62G10; Secondary 62E20.
Key words and phrases. Censored data, rank test, independence, asymptotic normality.

For any (R_*, Q_*) the set $C(R_*, Q_*)$ contains $(n - n_1)!\,(n - n_2)!$ elements. For later use we shall introduce in an unambiguous way one special completion of (R_*, Q_*). This special completion will be denoted by $(R', Q') = (R_1', \cdots, R_n', Q_1', \cdots, Q_n') \in C(R_*, Q_*)$, where

$$(1.3)\qquad \begin{aligned} &R_i' = R_{*i} \quad \text{for} \quad X_i \leqq X_{n_1 n}\,, \\ &R_i' = n_1 + \#\{j \mid X_j > X_{n_1 n}, j \leqq i\} \quad \text{for} \quad X_i > X_{n_1 n}\,, \\ &Q_i' = Q_{*i} \quad \text{for} \quad Y_i \leqq Y_{n_2 n}\,, \\ &Q_i' = n_2 + \#\{j \mid Y_j > Y_{n_2 n}, j \leqq i\} \quad \text{for} \quad Y_i > Y_{n_2 n}\,. \end{aligned}$$

In Section 2 we shall derive the locally most powerful rank test (lmprt) which is based on (R_*, Q_*) against the alternative $H(x, y; \theta)$, $\theta > 0$, having a density function $h(x, y; \theta)$. The score functions (sf's) prove to be discontinuous when observations are censored. In Section 3 asymptotic distributions of linear rank statistics including the important special case in Section 2 will be given by the Theorem 2.1 in [6].

2. Locally most powerful rank tests. Suppose the df's $H(x, y; \theta)$, $\theta \geqq 0$, have density functions $h(x, y; \theta)$. Assume that $H(x, y; 0) = F(x)G(y)$ for some df's F and G. Furthermore, assume that there exists a nonconstant function

$$(2.1)\qquad \lim_{\theta\to 0} \phi(x, y; \theta) \equiv \lim_{\theta \to 0} (\partial/\partial\theta) \log h(x, y; \theta) = \phi(x, y)$$

satisfying

$$(2.2)\qquad 0 < \lim_{\theta\to 0} \textstyle\iint |\phi(x, y; \theta)|\, dH(x, y; \theta) = \iint |\phi(x, y)|\, dF(x)\, dG(y) < \infty\,.$$

Introduce the score constants

$$(2.3)\qquad \begin{aligned} a_n(i, j) = n^2\tbinom{n-1}{i-1}\tbinom{n-1}{j-1} \textstyle\iint \phi(x, y)[F(x)]^{i-1}[1 - F(x)]^{n-i}[G(y)]^{j-1} \\ \times [1 - G(y)]^{n-j}\, dF(x)\, dG(y)\,, \qquad i, j = 1, \cdots, n, \text{ and} \end{aligned}$$

$$(2.4)\qquad \begin{aligned} a_n'(i, j) &= a_n(i, j) && \text{for } i \leqq n_1, j \leqq n_2\,, \\ &= (n - n_1)^{-1} \textstyle\sum_{i=n_1+1}^{n} a_n(i, j) && \text{for } i > n_1, j \leqq n_2\,, \\ &= (n - n_2)^{-1} \textstyle\sum_{j=n_2+1}^{n} a_n(i, j) && \text{for } i \leqq n_1, j > n_2\,, \\ &= (n - n_1)^{-1}(n - n_2)^{-1} \textstyle\sum_{i=n_1+1}^{n} \sum_{j=n_2+1}^{n} a_n(i, j) \\ & && \text{for } i > n_1, j > n_2\,. \end{aligned}$$

Under the above conditions we can obtain the following theorem.

THEOREM 1. *If* (2.2) *holds, the test with critical region*

$$(2.5)\qquad S_n = \textstyle\sum_{i=1}^{n} a_n'(R_i', Q_i') \geqq k$$

is the lmprt *at the respective level to test the hypothesis of independence against the alternative $H(x, y; \theta)$, $\theta > 0$, on the basis of the n_1 smallest observations on X and of the n_2 smallest observations on Y where (R', Q') is given in Section* 1. *This completion (R', Q') may be replaced by any other completion of (R_*, Q_*).*

PROOF. First observe that

$$P_\theta(R_*, Q_*) = \sum_{(\bar{R},\bar{Q}) \in C(R_*,Q_*)} \int_{(\bar{R},\bar{Q})} \prod_{i=1}^n h(x_i, y_i; \theta)\, dx_i\, dy_i\,, \tag{2.6}$$

so that in particular

$$P_0(R_*, Q_*) = (n-n_1)!\,(n-n_2)!\,(n!)^{-2}\,. \tag{2.7}$$

Since the differentiation of (2.6) with respect to θ at $\theta = 0$ can be performed under the integral sign by (2.2), we obtain

$$\begin{aligned}(\partial/\partial\theta)P_\theta(R_*, Q_*)|_{\theta=0} &= \sum_{(\bar{R},\bar{Q}) \in C(R_*,Q_*)} \sum_{i=1}^n \int_{(\bar{R},\bar{Q})} \phi(x_i, y_i) \prod_{j=1}^n dF(x_j)\, dG(y_j) \\ &= (n!)^{-2} \sum_{i=1}^n \sum_{(\bar{R},\bar{Q}) \in C(R_*,Q_*)} a_n(\bar{R}_i, \bar{Q}_i) \\ &= (n-n_1)!\,(n-n_2)!\,(n!)^{-2} \sum_{i=1}^n a_n'(R_i', Q_i')\,.\end{aligned}$$

The conclusion follows from the fact that the lmprt has critical region $[(\partial/\partial\theta)P_\theta(R_*, Q_*)/P_0(R_*, Q_*)]|_{\theta=0} \geqq k$. □

In most of the models proposed so far, for example bivariate normal, the general model due to Farlie [2], or Konijn [5] if restricted to one-parameter families, the function $\phi(x, y)$ is of the product form

$$\phi(x, y) = \phi_1(x)\phi_2(y)\,. \tag{2.8}$$

This is also the case with Hájek–Šidák's model [3], although the sf does not have the form (2.1) since $-\infty < \theta < \infty$ there. When (2.8) holds, let us put

$$\begin{aligned} a_n(i) &= n\tbinom{n-1}{i-1} \int \phi_1(x)[F(x)]^{i-1}[1-F(x)]^{n-i}\, dF(x) \quad \text{and} \\ b_n(i) &= n\tbinom{n-1}{i-1} \int \phi_2(y)[G(y)]^{i-1}[1-G(y)]^{n-i}\, dG(y)\,.\end{aligned} \tag{2.9}$$

Then the statistic S_n reduces to

$$S_n = \sum_{i=1}^n a_n'(R_i')b_n'(Q_i') \tag{2.10}$$

where

$$\begin{aligned} a_n'(i) &= a_n(i) && \text{for } i \leqq n_1\,, \\ &= (n-n_1)^{-1} \textstyle\sum_{j=n_1+1}^n a_n(j) && \text{for } i > n_1\,, \\ b_n'(i) &= b_n(i) && \text{for } i \leqq n_2\,, \\ &= (n-n_2)^{-1} \textstyle\sum_{j=n_2+1}^n b_n(j) && \text{for } i > n_2\,. \end{aligned} \tag{2.11}$$

This result is similar to the two-sample case obtained by Johnson–Mehrotra [4].

3. Asymptotic distribution. In this section our arguments concern the asymptotic normality of

$$S_n = \sum_{i=1}^n a_n(R_i)b_n(Q_i)\,, \tag{3.1}$$

which has just been suggested to be important in many applications. The asymptotic distribution of

$$T_n = n^{\frac{1}{2}}\left[\iint J_n(F_n)K_n(G_n)\, dH_n - \iint J(F)K(G)\, dH\right], \tag{3.2}$$

a standardized version of S_n, has been given by Bhuchongkul [1] and Ruymgaart, Shorack and van Zwet [7]. Here F_n, G_n and H_n are the empirical df's of $(X_1, \cdots, X_n)$, $(Y_1, \cdots, Y_n)$ and $((X_1, Y_1), \cdots, (X_n, Y_n))$ respectively and $J_n(u) = a_n(i)$,

$K_n(u) = b_n(i)$ for $u \in [(i-1)/n, i/n)$ and yet $\lim_{n\to\infty} J_n(u) = J(u)$, $\lim_{n\to\infty} K_n(u) = K(u)$. They imposed continuity upon the sf's J and K. In our case, however, it is required to delete the restriction of continuity. In fact, assume that the uncensored scores have sf's ϕ_1 and ϕ_2 and assume that $n_1/n \to p$ and $n_2/n \to q$ for some $0 < p, q < 1$, then in view of (2.11) the sf's in the censored case are

$$\begin{aligned} J(u) &= \phi_1(u) && \text{for } 0 < u < p\,, \\ &= (1-p)^{-1} \textstyle\int_p^1 \phi_1(u)\,du && \text{for } p \leqq u < 1, \text{ and} \\ K(u) &= \phi_2(u) && \text{for } 0 < u < q\,, \\ &= (1-q)^{-1} \textstyle\int_q^1 \phi_2(u)\,du && \text{for } q \leqq u < 1\,, \end{aligned} \tag{3.3}$$

which are discontinuous. For discontinuous sf's, Hájek–Šidák [3] proved asymptotic normality of T_n under the null hypothesis, but they did not prove this under fixed alternatives.

Recently Ruymgaart [6] proved that T_n is asymptotically normal under fixed alternatives and local alternatives for sf's with a finite number of jumps. His conditions on the sf's seem to be quite general.

Here we give a theorem which is an important special case of [6].

ASSUMPTION 1. There are points $0 < s_1 < \cdots < s_\lambda < 1$ such that J is continuously differentiable in $(0, 1) - \{s_1, \cdots, s_\lambda\}$. A similar condition is imposed on K with respect to $\{t_1, \cdots, t_\nu\}$.

ASSUMPTION 2. The functions J_n, K_n, J and K satisfy $|J_n| \leqq Dr^a$, $|K_n| \leqq Dr^b$, $|J^{(i)}| \leqq Dr^{a+i}$, $|K^{(i)}| \leqq Dr^{b+i}$ for $i = 0, 1$ where defined on $(0, 1)$ for some positive constant D, a, b and for $r(u) = [u(1-u)]^{-1}$. The constants a and b satisfy either (i) $a = (\frac{1}{2} - \delta)/p_0$, $b = (\frac{1}{2} - \delta)/q_0$ for some $0 < \delta < \frac{1}{2}$ and some $p_0, q_0 > 1$ with $p_0^{-1} + q_0^{-1} = 1$ or (ii) $a = b = \frac{1}{2} - \delta$.

ASSUMPTION 3. For $F_n^* = [n/(n+1)]F_n$ and $G_n^* = [n/(n+1)]G_n$,

$$B_{0n}^* = n^{\frac{1}{2}} \iint [J_n(F_n)K_n(G_n) - J(F_n^*)K(G_n^*)]\,dH_n = o_p(1)\,.$$

ASSUMPTION 4. Denote a probability measure with df $F(x \mid G^{-1}(t))$ by μ_t. Then $\lim_{t\to t_i} \sup_A |\mu_t(A) - \mu_{t_i}(A)| = 0$ for $i = 1, \cdots, \lambda$. A similar condition holds for $G(y \mid F^{-1}(s))$.

ASSUMPTION 5. The condition $dG(y \mid x) \leqq D\,dG(y)$ holds in a neighborhood of $x = F^{-1}(s_i)$ and the condition $dF(x \mid y) \leqq dF(x)$ holds in a neighborhood of $y = G^{-1}(t_i)$.

Let us introduce the sets of bivariate df's $\mathscr{H} = \{H \mid H$ is continuous on the plane$\}$ and $\mathscr{H}_{C\delta} = \{H \in \mathscr{H} \mid dH \leqq C[r(F)r(G)]^{\delta/2}\,dF\,dG\}$, where δ is the same number as in Assumption 2, and $C \geqq 1$ is a fixed constant. For any real number v we define the function δ_v by

$$\delta_v(u) = 0 \quad \text{for} \quad u < v\,, \qquad \delta_v(u) = 1 \quad \text{for} \quad u \geqq v\,.$$

The conditional expectations in the theorem below are supposed to be obtained by integration with respect to the conditional probability measures considered in Assumption 4.

THEOREM 2. *If $H \in \mathscr{H}$ and Assumptions* 1, 2 (i) *and* 3–5 *are satisfied or $H \in \mathscr{H}_{C\delta}$ and Assumptions,* 1, 2 (ii) *and* 3–5 *are satisfied, then the asymptotic normality*

$$T_n \to_d N(0, \sigma^2) \qquad as \quad n \to \infty \tag{3.4}$$

holds. Here

$$\sigma^2 = \mathrm{Var}\,[J(F(X))K(G(Y)) + \textstyle\int_0^1 (\delta_{F(X)}(s) - s)E(K(G(Y)) \,|\, F(X) = s)\, dJ(s) + \int_0^1 (\delta_{G(Y)}(t) - t)E(J(F(X)) \,|\, G(Y) = t)\, dK(t)]\,. \tag{3.5}$$

Moreover, the asymptotic normality in (3.4) *is uniform on each subclass $\mathscr{H}'$ of $\mathscr{H}$ or $\mathscr{H}_{C\delta}$ for which Assumptions* 3–5 *hold uniformly and on which $\sigma^2 = \sigma^2(H)$ is bounded away from zero.*

REMARK 1. The above theorem is essentially a special case of Theorem 2.1 [6] and the differences between our assumptions and those in [6] are due to the absence of a density in our Assumptions 4 and 5.

REMARK 2. If we introduce the sets $S_{\beta 1n} = [F^{-1}(s_1 - n^{-\frac{1}{2}}\beta),\ F^{-1}(s_1 + n^{-\frac{1}{2}}\beta)]$, $\Omega_{\beta 1n} = \{\omega \,|\, n^{\frac{1}{2}} \sup |F_n - F| < \beta\}$ and $S_{\beta 2n}$, $\Omega_{\beta 2n}$ similarly defined for G, we can simplify the proof of Theorem 2.1 in [6] by avoiding the technical Lemmas 4.2, 4.3 in [6]. This is because the introduced set $\Omega_{\beta 1n}$ has the property that for any $\varepsilon > 0$, $P(\Omega_{\beta 1n}) > 1 - \varepsilon$ uniformly in n and F for large β and that by integrating over $S_{\beta 1n}$ the factor $n^{\frac{1}{2}}$ cancels out.

Acknowledgment. I wish to thank Professor M. Okamoto for his helpful guidance. I also wish to thank the referee for his relevant comments and for drawing my attention to Ruymgaart's paper.

REFERENCES

[1] BHUCHONGKUL, S. (1964). A class of nonparametric tests for independence in bivariate populations. *Ann. Math. Statist.* **35** 138–149.
[2] FARLIE, D. J. G. (1960). The performance of some correlation coefficients for a general bivariate distribution. *Biometrika* **47** 307–323.
[3] HÁJEK, J. and ŠIDÁK, Z. (1967). *Theory of Rank Tests.* Academic Press, New York.
[4] JOHNSON, R. A. and MEHROTRA, K. G. (1972). Locally most powerful rank tests for the two-sample problem with censored data. *Ann. Math. Statist.* **43** 823–831.
[5] KONIJN, H. S. (1956). On the power of certain tests for independence in bivariate populations. *Ann. Math. Statist.* **27** 300–323.
[6] RUYMGAART, F. H. Asymptotic normality of nonparametric tests for independence. *Ann. Statist.* **2** 892–910.
[7] RUYMGAART, F. H., SHORACK, G. R. and VAN ZWET, W. R. (1972). Asymptotic normality of nonparametric tests for independence. *Ann. Math. Statist.* **43** 1122–1135.
[8] WATTERSON, G. A. (1959). Linear estimation in censored samples from multivariate normal populations. *Ann. Math. Statist.* **30** 814–824.

DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE
KYUSHU UNIVERSITY
HIGASHI-KU, HAKOZAKI
FUKUOKA, 812, JAPAN