Canad. Math. Bull. Vol. **16** (3), 1973

ON A CLASS OF NONPARAMETRIC TESTS FOR INDEPENDENCE—BIVARIATE CASE(1)

BY
M. S. SRIVASTAVA

1. **Introduction.** Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be n mutually independent pairs of random variables with absolutely continuous (hereafter, a.c.) *pdf* given by

$$h(x, y; \rho) = f^{(\rho)}(x \mid y)g(y) \equiv f(e(\rho)x - b(\rho)u(y))g(y), \tag{1}$$

where $f^{(\rho)}$ denotes the conditional *pdf* of X given Y, $g(y)$ the marginal *pdf* of Y, $e(\rho)\to 1$ and $b(\rho)\to 0$ as $\rho\to 0$ and,

$$u(y) = -[g'(y)/g(y)]; \qquad g'(y) = (d/dy)(g(y)). \tag{2}$$

We wish to test the hypothesis

$$H_0 : \rho = 0 \tag{3}$$

against the alternative

$$K_n : \rho = n^{-1/2}b, \qquad 0 < b < \infty, \tag{4}$$

For the two-sided alternative we take $-\infty < b < \infty$. A feature of the model (1) is that it covers both-sided alternatives which have not been considered in the literature so far. One-sided alternatives have been considered by Konjin [6], Farlie [4] and Bhuchongkul [1], Hájek and Šidká [5]. However, these models seem to be far from satisfactory as pointed out by Hájek and Šidák [5]. We hope that the present approach may fill at least partially one of serious gaps mentioned by Hájek and Šidák [5] in the *preface* of their book.

The hypothesis H_0 is equivalent to testing the independence of X and Y, i.e.,

$$h(x, y; 0) = f^{(0)}(x \mid y)g(y) \equiv f(x)g(y). \tag{5}$$

The form of h is not known but we shall assume that $h \in \mathscr{h}$, where h denotes the class of all absolutely continuous two-dimensional *pdf*'s satisfying (1) and such that their marginals f and g satisfy

$$\int [f'(x)/f(x)]^2 f(x)\, dx < \infty, \qquad \int [g'(y)/g(y)]^2 g(y)\, dy < \infty. \tag{6}$$

we will refer to the above conditions as *Condition* (Cl) in the sequel.

Received by the editors July 22, 1970 and, in revised form, October 26, 1971.
(1) Research supported partially by Canada Council and National Research Council of Canada.

https://doi.org/10.4153/CMB-1973-054-9 Published online by Cambridge University Press

In this paper a class of rank-score tests for H_0 is proposed in §2 and is shown to be locally most powerful rank tests. In §3, the asymptotic non-null distribution of the test statistics is given and, in §4 the Pitman efficiency with respect to the parametric correlation coefficient is derived.

2. **Rank score tests.** Let

$$\psi(t,f) = -[f'(F^{-1}(t))/f(F^{-1}(t))], \qquad 0 \le t \le 1$$

where $f \in \mathscr{k}$ with distribution function F and F^{-1} is the inverse of F. We will consider only those a.c. f for which the score function $\psi(t,f)$ is a nondecreasing function of t. We refer to this condition in the sequel as Condition (C2). Let R_i and Q_i denote the ranks of X_i and Y_i among $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ respectively. Let $U_n^{(1)} < U_n^{(2)} < \cdots < U_n^{(n)}$ be an ordered sample from a uniform distribution over [0, 1]. Let

$$a_n(i,f) = E\psi(U_n^{(i)}, f). \tag{7}$$

We will show at the end of this section that the test with critical region

$$\sum_{i=1}^{n} a_n(R_i, f_0)a_n(Q_i, g_0) \ge k \tag{8}$$

is the locally most powerful rank test for H_0 against K_n at the respective level, where f_0 and g_0 are known densities belonging to the class $\mathscr{k}$. Under condition (C2) an asymptotically equivalent class of statistics is given by

$$T_n(f_0, g_0) = \frac{1}{n}\sum_{i=1}^{n} \psi_n\left(\frac{R_i}{n+1}, f_0\right)\psi_n\left(\frac{Q_i}{n+1}, g_0\right), \tag{9}$$

where

$$\psi_n(t,f) = \psi\left(\frac{j}{n+1}, f\right), \qquad \frac{(j-1)}{n} < t \le j/n. \tag{10}$$

We now turn to show that the critical region given by (8) is locally most powerful.

Let P denote that the probability is being computed under the alternative. Let $\mathbf{R}=(R_1, \ldots, R_n)$ and $\mathbf{Q}=(Q_1, \ldots, Q_n)$. Let $S=\{(x_i, y_i),\ i=1, 2, \ldots, n: \mathbf{R}=\mathbf{r}, \mathbf{Q}=\mathbf{q}\}$. We assume without loss of generality that $e(\rho)\equiv 1$, and $b(\rho)\equiv\rho$. Then

$$\begin{aligned} P\{\mathbf{R}=\mathbf{r}, \mathbf{Q}=\mathbf{q}\} &= \int_S \cdots \int \prod_{i=1}^{n} [f_0(x_i-\rho u(y_i))]g_0(y_i)\,dx_i\,dy_i \\ &= \int_S \cdots \int \prod_{i=1}^{n} f_0(x_i)g_0(y_i)\,dx_i\,dy_i \\ &\quad + \rho \int_S \cdots \int \rho^{-1}\left[\prod_{i=1}^{n} f_0(x_i-\rho u(y_i)) - \prod_{i=1}^{n} f(x_i)\right] \prod_{i=1}^{n} g_0(y_i)\,dx_i\,dy_i \end{aligned}$$

$$= (1/n!)^2 + \rho \sum_{k=1}^{n} \int \cdots \int_S [\rho^{-1}\{f_0(x_k - \rho u(y_k)) - f_0(x_k)\}]$$

$$\times \left[\prod_{l=k+1}^{n} f(x_l) \prod_{j=1}^{k-1} f_0(x_j - \rho u(y_j)) \right] \prod_{i=1}^{n} g_0(y_i)\, dx_i\, dy_i.$$

Following as in Hájek and Šidák [5, p. 71] it can be shown that (note that $u'(y) = -g_0'(y)/g_0(y)$)

$$\lim_{\rho \to 0} P\{\mathbf{R} = \mathbf{r}, \mathbf{Q} = \mathbf{q}\}$$

$$= (1/n!)^2 + \rho \sum_{k=1}^{n} \int \cdots \int_S [-f_0'(x_k)/f_0(x_k)] u(y_k) \prod_{i=1}^{n} f_0(x_i) g_0(y_i)\, dx_i\, dy_i$$

$$= (1/n!)^2 + \rho \sum_{k=1}^{n} E_0[-f_0'(x_k)/f_0(x_k) \mid \mathbf{R} = \mathbf{r}] E_0[-g_0'(y_k)/g_0(y_k) \mid Q = q]$$

$$= (1/n!)^2 + \rho \sum_{k=1}^{n} a_n(R_i, f_0) a_n(Q_i, q_0),$$

where E_0 denotes that the expectation is taken under the hypothesis of independence. ||

3. **Distribution of** T_n. The following equivalent form of the statistic T_n seems more convenient to work with. Let us rearrange all n pairs of observations according to the magnitude of their second coordinate into the sequence (X_{d_1}, Y_{d_1}), $(X_{d_2}, Y_{d_2}), \ldots, (X_{d_n}, Y_{d_n})$ in such a way that $Y_{d_1} < Y_{d_2} < \cdots < Y_{d_n}$. Let R_i^0 be the rank of X_{d_i} among $X_1, \ldots, X_n$. Then

$$T_n(f_0, g_0) = \frac{1}{n} \sum_{i=1}^{n} \psi_n\left(\frac{R_i^0}{n+1}, f_0\right) \psi_n\left(\frac{i}{n+1}, g_0\right). \tag{11}$$

Hájek and Šidák [5, p. 168] have shown that the limiting null-distribution of the test statistic $T_n(f_0, g_0)$ is normal with mean 0 and variance $\gamma^2\delta^2/n$, where

$$\gamma^2 = \int_0^1 \psi^2(t, f_0)\, dt \quad \text{and} \quad \delta^2 = \int_0^1 \psi^2(t, g_0)\, dt. \tag{12}$$

It thus remains to obtain the limiting non-null distribution of T_n (under near alternatives); we obtain this under the following additional conditions

(i) $\psi(\cdot, f)$ satisfies conditions A^* and E of Chernoff, Gastwirth and John [2, p. 61].

(C3) (ii) $\displaystyle\int_0^1 \psi(t, g_0)\psi'(t, g)[t(1-t)]^{1/2}\, dt < \infty; \qquad \psi' = (d\psi/dt)$

(iii) $\displaystyle\tau^2 \equiv \underset{s<t}{\int_0^1 \int_0^1} \psi(s, g_0)\psi'(s, g)\psi(t, g_0)\psi'(t, g) s(1-t)\, ds\, dt < \infty.$

https://doi.org/10.4153/CMB-1973-054-9 Published online by Cambridge University Press

First, we note that under condition (C1) (see Appendix)

$$\sum_{i=1}^{n}[u(Y_i)]^2 = O_p(n) \quad \text{and} \quad \max_{1\le i\le n}|u(Y_i)| = o_p(n^{1/2}) \tag{13}$$

since $u(y)=-g'(y)/g(y)$. Hence from Hájek and Šidák [5], we have conditionally given $y_1, \ldots, y_n$,

$$L(T_n \mid y_1, \ldots, y_n) \to N(\mu_n, \gamma^2\delta^2/n), \tag{14}$$

where L denotes the distribution "of" and

$$\mu_n = n^{-1}\rho \mathbf{c}'\mathbf{u}_n \int_0^1 \psi(t, f_0)\psi(t, f)\, dt, \tag{15}$$

$$\mathbf{c}' = \left(\psi_n\left(\frac{1}{n+1}, f_0\right), \ldots, \psi_n\left(\frac{n}{n+1}, f_0\right)\right). \tag{16}$$

It follows from Chernoff et al. [2] that under conditions (C1)–(C3) (see Moore(2), [7] also).

$$L\left[n^{1/2}\left(\frac{\mathbf{c}'\mathbf{u}_n}{n} - \theta\right)\right] \to N(0, \tau^2) \tag{17}$$

where

$$\theta = \int_0^1 \psi(t, g_0)\psi(t, g)\, dt, \qquad \text{and } \tau^2 \text{ is defined in (C3).} \tag{18}$$

Hence (see Appendix)

$$L(T_n) \to N(\xi_n, \eta_n^2) \tag{19}$$

where

$$\xi_n = \rho_n\theta \int_0^1 \psi(t, f)\psi(t, f_0)\, dt, \quad \eta_n^2 = (\gamma^2\delta^2/n) + (\rho_n^2\tau^2/n)\left[\int_0^1 \psi(t, f)\psi(t, f_0)\, dt\right]. \tag{20}$$

4. **Asymptotic efficiency.** The parametric test r_n is based on the sample correlation coefficient,

$$r_n = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \Big/ \left[\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2\right]^{1/2}. \tag{21}$$

Cramér [3, pp. 359–366] shows that

$$E(r_n) = \rho + O(n^{-1}) \quad \text{and} \quad \mathrm{Var}_0(r_n) = 1/n \tag{22}$$

where ρ denotes the correlation coefficient between X and Y, and Var_0 denotes the

(2) I am indebted to Y. S. Lee for pointing out this reference which led to Chernoff et al. [2].

variance under the hypothesis. Hence, the Pitman efficiency of the rank-score tests T_n relative to the correlation coefficient r_n-test is given by

$$(23)\qquad \begin{aligned} e(T_n, r_n) &= \lim_{n\to\infty} \frac{[\partial E(T_n)/\partial\rho|_{\rho=0}]^2/\mathrm{Var}_0(T_n)}{[\partial E(r_n)/\partial\rho|_{\rho=0}]^2/\mathrm{Var}_0(r_n)} \\ &= \left[\int_0^1 \psi(t, g_0)\psi(t, g)\, dt\right]^2 \left[\int_0^1 \psi(t, f_0)\psi(t, f)\, dt\right] \Big/ \gamma^2\delta^2. \end{aligned}$$

It follows from Chernoff-Savage [6] that $e(T_n, r_n) \geq 1$, the equality holds only if f and g are normal.

The expression (23) has been conjectured by Hájek and Šidák [5, p. 222].

ACKNOWLEDGEMENT. I wish to thank the referee for his helpful suggestions in condensing the proofs.

APPENDIX

Proof of (13). Since $u(Y_1), \ldots, u(Y_n)$ are *iid* random variables with finite expectation (also finite variance), the first part of (13) follows from the Kolmogorov's strong law of large numbers. The second part of (13) follows from the following more general result; the proof parallels to that of Gnedenko and Kolmogorov [p. 105]: Limit distributions for sums of independent random variables; translated by K. L. Chung, Addison Wesley, Reading, Mass.

LEMMA. *Let $X_1, X_2, \ldots$ be a sequence of random variables with distribution functions $F_1, F_2, \ldots$. Then $X_n \to 0$ in probability if and only if the following two conditions are satisfied:*

$$\text{(i)} \int_{|x|>1} dF_n(x) \to 0,$$

$$\text{(ii)} \int_{|x|\leq 1} x^2\, dF_n(x) \to 0.$$

Proof of (19). Since for any two-dimensional r.v. (X_n, Y_n)

$$\lim_{n\to\infty} P[X_n \leq x, Y_n \leq y] = \lim_{n\to\infty} \int_{-\infty}^{y} P[X_n \leq x \mid y]\, dG_n(y)$$

we get

$$\lim_{n\to\infty} P[X_n \leq x, Y_n \leq y] = \int_{-\infty}^{y} \lim_{n\to\infty} P[X_n \leq x \mid y]\, dG(y)$$

if $G_n(y) \to G(y)$ for every continuity point y of $G(y)$ and $\lim_{n\to\infty} P[X_n \leq x \mid y]$ exists for all y.

References

1. S. Bhuchongkul, *A class of non-parametric tests for independence in Bivariate populations*, Ann. Math. Statist. **35** (1964) 138–149.

2. H. Chernoff, J. L. Gastwirth, and M. V. Johns, *Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation*, Ann. Math. Statist. **38** (1967), 52–72.

3. H. Cramér, *Mathematical methods of statistics*, Princeton Univ. Press, Princeton, N.J., 1946.

4. D. J. G. Farlie, *The asymptotic efficiency of Daniel's generalized correlation coefficients*, J. Roy. Statist. Soc. Ser. B. **23** (1961) 128–142.

5. J. Hájek and Z. Šidák, *Theory of rank tests*, Academic Press, New York, 1967.

6. H. S. Konijn, *On the power of certain tests for independence in bivariate populations*, Ann. Math. Statist. **27** (1956), 300–323. Correction **27** (1958), p. 935.

7. D. S. Moore, *An elementary proof of asymptotic normality of linear functions of order statistics*, Ann. Math. Statist. **39** (1968), 263–265.

University of Toronto,
Toronto, Ontario