The Annals of Statistics
1976, Vol. 4, No. 2, 400-405

# ON THE ASYMPTOTIC NORMALITY OF RANK STATISTICS FOR THE TWO-SAMPLE PROBLEM

BY SHINGO SHIRAHATA

*Kyushu University*

Conditions to ensure the asymptotic normality of rank statistics having a scores generating function with finitely many jumps are obtained. These conditions are derived by studying rank statistics with a scores generating function $J$ such that $J(u) = 1$ or $0$ as $u \geqq s$ or $u < s$ for a fixed $s$, $0 < s < 1$. No differentiability conditions are imposed on the underlying distribution functions at the jump points of the scores generating function.

**1. Introduction.** Let $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$ be two independent random samples from populations with continuous distribution functions $F(x)$ and $G(x)$ respectively. Assume that there exists a real number $\lambda_0$ such that $0 < \lambda_0 \leqq \lambda = m/N \leqq 1 - \lambda_0$ where $N = m + n$. Denote by $H_N$ and $F_m$ the empirical distribution functions of the combined sample and of $X$'s respectively and put $H = \lambda F + (1 - \lambda)G$.

The purpose of this paper is to show the asymptotic normality of a rank statistic

$$S_N = N^{\frac{1}{2}}[\int K_N(N(N+1)^{-1}H_N)\, dF_m - \int K(H)\, dF] \tag{1.1}$$

where $K(u)$ is a piecewise continuous function on $(0, 1)$, while $K_N$ is a function which is constant in $I_i = ((i-1)/(N+1), i/(N+1)]$ and satisfies $\lim_{N\to\infty} K_N(u) = K(u)$. In order to study $S_N$, the statistic

$$T_N = N^{\frac{1}{2}}[\int J_N(N(N+1)^{-1}H_N)\, dF_m - \int J(H)\, dF] \tag{1.2}$$

plays an essential role where $J(u) = 1$ or $0$ as $u \geqq s$ or $u < s$ for a fixed $s$, $0 < s < 1$ with $J_N$ constant in each $I_i$ such that $\lim_{N\to\infty} J_N(u) = J(u)$. This is because $S_N$ can be represented as a sum of a rank statistic with a continuous scores generating function and a linear combination of a finite number of statistics of the form (1.2).

Rank statistics with discontinuous scores generating function will appear, for instance, for censored samples. In the censored sample problem, the recommended scores generating function has the form $L(u) + cJ(u)$, where $L$ is the optimal scores generating function for the uncensored case for $u \leqq s$ and $L(u) = L(s)$ for $u > s$ (see Gastwirth (1965) or Johnson and Mehrotra (1972)).

If the scores generating function is continuous and satisfies suitable regularity

Received January 1974; revised July 1975.
*AMS* 1970 *subject classifications*. Primary 62G10; Secondary 62E20.
*Key words and phrases*. Rank statistics, asymptotic normality, scores generating function with finitely many jumps, two-sample problem.

conditions, the asymptotic normality of rank statistics has been investigated by Chernoff and Savage (1958) and Hájek (1968) among others under fixed alternatives. Under the hypothesis of randomness and its contiguous alternatives, Hájek and Šidák (1967) proved the asymptotic normality for quite a wide class of scores generating functions. But they did not deal with fixed alternatives.

Dupač and Hájek (1969) showed the asymptotic normality of rank statistics with not necessarily continuous scores generating function, and its simple proof was given by Koul and Staudte (1972). Although their results are general, they imposed almost everywhere differentiability on the distribution functions (or condition (2.39) in [2]). However, the condition does not hold when $F$ and $G$ are uniform distributions over $(0, \frac{1}{4}) \cup (\frac{3}{4}, 1)$ and $(\frac{1}{4}, \frac{3}{4})$ respectively and $s = \frac{1}{2}$. Neither does it hold when $F$ has a density function $f(x) = \frac{1}{4}|x|^{\frac{1}{2}}$ for $|x| < 1$ and $f(x) = 0$ for $|x| \geqq 1$ and $G$ is a uniform distribution over $(-1, 1)$. By the result in Section 2, the asymptotic normality of the rank statistic holds when it has a scores generating function with a jump at $u = \frac{1}{2}$. This fact can not be derived from known results as far as the author is aware.

**2. Main theorem.** Let us begin with the following assumptions.

ASSUMPTION 2.1.

$$B_{N1} = N^{\frac{1}{2}} \int [J_N(N(N+1)^{-1}H_N) - J(N(N+1)^{-1}H_N)]\, dF_m \to_P 0 \quad \text{as } N \to \infty . \tag{2.1}$$

The assumption is automatically satisfied when $J_N(u) = J(i/(N+1))$ for $u \in I_i$.

ASSUMPTION 2.2. There exist $j(s)$ and $k(s)$, $0 \leqq j(s), k(s) \leqq \infty$, such that for any $M > 0$

$$\max_{0<t<MN^{-\frac{1}{2}}} N^{\frac{1}{2}}|HF^{-1}FH^{-1}(s) - HF^{-1}(FH^{-1}(s) - t) - tj(s)| = o(1) \tag{2.2}$$

and

$$\max_{-MN^{-\frac{1}{2}}<t<0} N^{\frac{1}{2}}|HF^{-1}(FH^{-1}(s)+) - HF^{-1}(FH^{-1}(s) - t) - tk(s)| = o(1) . \tag{2.3}$$

The case $j(s) = \infty$ is to mean that the left-hand side of (2.2) with $tj(s)$ deleted diverges to infinity; the case $k(s) = \infty$ has a similar meaning. The inverse functions in (2.2) and (2.3) are defined by

$$F^{-1}(u) = \inf\{x : F(x) \geqq u\} \tag{2.4}$$

and

$$F^{-1}(u+) = \inf\{x : F(x) > u\} . \tag{2.5}$$

From (2.4) and (2.5), obviously

$$HF^{-1}FH^{-1}(s) \leqq s \leqq HF^{-1}(FH^{-1}(s)+) . \tag{2.6}$$

Therefore the pair $(F, G)$ can be classified into one of the four classes defined by

$$\mathscr{H}_1 = \{(F, G): HF^{-1}FH^{-1}(s) = s = HF^{-1}(FH^{-1}(s)+)\},$$
$$\mathscr{H}_2 = \{(F, G): HF^{-1}FH^{-1}(s) < s = HF^{-1}(FH^{-1}(s)+)\},$$
$$\mathscr{H}_3 = \{(F, G): HF^{-1}FH^{-1}(s) = s < HF^{-1}(FH^{-1}(s)+)\}$$

and

$$\mathscr{H}_4 = \{(F, G): HF^{-1}FH^{-1}(s) < s < HF^{-1}(FH^{-1}(s)+)\}.$$

The main result in this paper is the following theorem.

THEOREM 2.1. *Let $S_N$ have a scores generating function $K$ which satisfies $K = L + cJ$ for $c \neq 0$ such that $L$ is continuous in $(0, 1)$. Suppose that $S_N - cT_N$ is asymptotically equivalent to a sum*

$$D_1 m^{-\frac{1}{2}} \sum_{i=1}^{m} [B(X_i) - EB(X_i)] + D_2 n^{-\frac{1}{2}} \sum_{i=1}^{n} [C(Y_i) - EC(Y_i)] \tag{2.7}$$

*for some constants $D_1$ and $D_2$ where $EB^2(X_i) < \infty$ and $EC^2(Y_i) < \infty$; and suppose that $(F, G)$ satisfies Assumptions 2.1 and 2.2 and one of the following set of conditions:*

(1) *$(F, G)$ belongs to $\mathscr{H}_1$ and $0 < j(s) = k(s) < \infty$.*
(2) *$(F, G)$ belongs to $\mathscr{H}_2$ and $0 \leqq j(s) < \infty$, $k(s) = \infty$.*
(3) *$(F, G)$ belongs to $\mathscr{H}_3$ and $j(s) = \infty$, $0 \leqq k(s) < \infty$.*
(4) *$(F, G)$ belongs to $\mathscr{H}_4$.*
(5) *$j(s) = k(s) = \infty$.*

*Then $S_N$ is asymptotically normal.*

REMARK. If $S_N - cT_N$ and $L$ satisfy the assumptions of Theorem 1 in [1] or the weaker forms in Govindarajulu, Le Cam and Raghavachari (1965) or the assumptions of theorems in [6], $S_N - cT_N$ can be asymptotically written in the form (2.7). When $K$ is discontinuous at finitely many points, Theorem 2.1 also holds if $(F, G)$ satisfies the above-mentioned assumptions at each discontinuity point.

**3. Proof of the theorem.** The statistic $T_N$ can be decomposed into $T_N = T_{N1} + T_{N2} + B_{N1} + B_{N2}$ where

$$T_{N1} = N^{\frac{1}{2}} \int J(H)\, d(F_m - F), \tag{3.1}$$

$$T_{N2} = N^{\frac{1}{2}} \int [J(N(N+1)^{-1}H_N) - J(H)]\, dF, \tag{3.2}$$

$$B_{N2} = N^{\frac{1}{2}} \int [J(N(N+1)^{-1}H_N) - J(H)]\, d(F_m - F) \tag{3.3}$$

and where $B_{N1}$ is given by (2.1). It is easy to show that $B_{N2} = o_P(1)$. In view of Assumption 2.1, it is sufficient to consider only $T_{N1} + T_{N2}$. The term $T_{N1}$ can be rewritten as

$$T_{N1} = (\lambda m)^{-\frac{1}{2}} \sum_{i=1}^{m} [J(H(X_i)) - \int J(H)\, dF] \tag{3.4}$$

which is a sum of i.i.d. random variables with finite variance.

To deal with $T_{N2}$, we need the following lemma.

LEMMA 3.1. *It holds that*

$$T_{N2} = N^{\frac{1}{2}}[FH^{-1}(s) - F(W_\nu)] \tag{3.5}$$

*where* $W_\nu$ *is the* $\nu$th *order statistic in the combined sample and* $\nu$ *is the smallest integer not smaller than* $(N+1)s$.

PROOF. The integrand in (3.2) takes value one if $N(N+1)^{-1}H_N \geqq s > H$, minus one if $N(N+1)^{-1}H_N < s \leqq H$ and zero otherwise. The relation $W_\nu \leqq x$ holds if and only if $N(N+1)^{-1}H_N(x) \geqq s$. Combining these facts, we have (3.5).

Next, let us define

$$\begin{aligned} U_N = {} & (\lambda/m)^{\frac{1}{2}} \textstyle\sum_{i=1}^m [I(F(X_i) < FH^{-1}(s)) - FH^{-1}(s)] \\ & + [(1-\lambda)/n]^{\frac{1}{2}} \textstyle\sum_{i=1}^n [I(F(Y_i) < FH^{-1}(s)) - GF^{-1}FH^{-1}(s)] \end{aligned} \tag{3.6}$$

and

$$\begin{aligned} V_N = {} & (\lambda/m)^{\frac{1}{2}} \textstyle\sum_{i=1}^m [I(F(X_i) \leqq FH^{-1}(s)) - FH^{-1}(s)] \\ & + [(1-\lambda)/n]^{\frac{1}{2}} \textstyle\sum_{i=1}^n [I(F(Y_i) \leqq FH^{-1}(s)) - GF^{-1}(FH^{-1}(s)+)] \end{aligned} \tag{3.7}$$

where $I$ denotes the indicator function.

LEMMA 3.2. *If Assumption* 2.2 *holds for* $0 < j(s), k(s) < \infty$, *then* $T_{N2}$ *is asymptotically equivalent to* $U_N I(U_N \geqq 0)/j(s) + V_N I(V_N < 0)/k(s)$, $V_N I(V_N < 0)/k(s)$, $U_N I(U_N \geqq 0)/j(s)$ *and zero according as* $(F, G)$ *belongs to* $\mathscr{H}_1$, $\mathscr{H}_2$, $\mathscr{H}_3$ *and* $\mathscr{H}_4$.

PROOF. Let us denote by $H_{FN}$ the empirical distribution function of the combined sample of $F(X_i)$'s and $F(Y_i)$'s and put $H_F(u) = \lambda u + (1-\lambda)GF^{-1}(u+)$. From Lemma 3.1, the event $T_{N2} \leqq u$ is equivalent to $H_{FN}((FH^{-1}(s) - N^{-\frac{1}{2}}u)-) \leqq N^{-1}(\nu - 1)$ which is also equivalent to

$$\begin{aligned} & N^{\frac{1}{2}}[H_{FN}((FH^{-1}(s) - N^{-\frac{1}{2}}u)-) - H_F((FH^{-1}(s) - N^{-\frac{1}{2}}u)-) \\ & \qquad \leqq N^{\frac{1}{2}}[N^{-1}(\nu-1) - H_F((FH^{-1}(s) - N^{-\frac{1}{2}}u)-)] \,. \end{aligned} \tag{3.8}$$

When $u \geqq 0$, the left-hand side of (3.8) is asymptotically equivalent in probability to $N^{\frac{1}{2}}[H_{FN}(FH^{-1}(s)-) - H_F(FH^{-1}(s)-)]$ which is identical with $U_N$. This fact can be proved by the same arguments as in Ghosh (1971, pages 1958–1959). Similarly, it is asymptotically equivalent in probability to $V_N$ when $u < 0$. On the other hand, the right-hand side of (3.8) is

$$N^{\frac{1}{2}}[s + O(N^{-1}) - HF^{-1}((FH^{-1}(s) - N^{-\frac{1}{2}}u)-)] \,. \tag{3.9}$$

Since $H_F(s) = HF^{-1}(s+)$, Assumption 2.2 implies that, when $u \geqq 0$, (3.9) converges to $uj(s)$ for $\mathscr{H}_1$ and $\mathscr{H}_3$ and diverges to infinity for $\mathscr{H}_2$ and $\mathscr{H}_4$. When $u < 0$, the same assumption entails that (3.9) converges to $uk(s)$ for $\mathscr{H}_1$ and $\mathscr{H}_2$ and diverges to minus infinity for $\mathscr{H}_3$ and $\mathscr{H}_4$. From these facts follows the conclusion of the lemma.

LEMMA 3.3. *If Assumption* 2.2 *holds, then a statistic to which* $T_{N2}$ *is asymptotically equivalent is given in the table where the expression* $\mathscr{H}_i \to E$ *means that* *if* $(F, G)$ *belongs to* $\mathscr{H}_i$, $T_{N2}$ *diverges with positive probability, whereas* $\mathscr{H}_i \to \mathscr{H}_j$ *means that if* $(F, G)$ *belongs to* $\mathscr{H}_i$, $T_{N2}$ *is asymptotically equivalent to the same statistic as for* $\mathscr{H}_j$ *in Lemma* 3.2.

| | $k(s) = \infty$ | $0 < k(s) < \infty$ | $k(s) = 0$ |
|---|---|---|---|
| $j(s) = \infty$ | $\mathscr{H}_1, \mathscr{H}_2, \mathscr{H}_3,$ $\mathscr{H}_4 \to \mathscr{H}_4$ | $\mathscr{H}_1, \mathscr{H}_2 \to \mathscr{H}_2$ $\mathscr{H}_3, \mathscr{H}_4 \to \mathscr{H}_4$ | $\mathscr{H}_1, \mathscr{H}_2 \to E$ $\mathscr{H}_3, \mathscr{H}_4 \to \mathscr{H}_4$ |
| $0 < j(s) < \infty$ | $\mathscr{H}_1, \mathscr{H}_3 \to \mathscr{H}_3$ $\mathscr{H}_2, \mathscr{H}_4 \to \mathscr{H}_4$ | Lemma 3.2 | $\mathscr{H}_1, \mathscr{H}_2 \to E$ $\mathscr{H}_3 \to \mathscr{H}_3,$ $\mathscr{H}_4 \to \mathscr{H}_4$ |
| $j(s) = 0$ | $\mathscr{H}_1, \mathscr{H}_3 \to E$ $\mathscr{H}_2, \mathscr{H}_4 \to \mathscr{H}_4$ | $\mathscr{H}_1, \mathscr{H}_3 \to E$ $\mathscr{H}_2 \to \mathscr{H}_2,$ $\mathscr{H}_4 \to \mathscr{H}_4$ | $\mathscr{H}_1, \mathscr{H}_2, \mathscr{H}_3$ $\to E$ $\mathscr{H}_4 \to \mathscr{H}_4$ |

PROOF. Now we need no further arguments. For example if $j(s) = 0$ and $0 < k(s) < \infty$, then from the proof of Lemma 3.2 it holds that for $\mathscr{H}_1$ $\lim_{N\to\infty} P(T_{N2} \leqq u) = \frac{1}{2}$ for each $0 < u < \infty$. Thus for $\mathscr{H}_1$, $T_{N2}$ diverges to infinity with probability $\frac{1}{2}$. All other cases can be shown similarly.

PROOF OF THEOREM 2.1. If one of the conditions (2)—(5) holds, then by Lemma 3.3, $T_{N2} \to_P 0$ as $N \to \infty$. On the other hand, if (1) holds, then $U_N = V_N$ with probability one and Lemma 3.2 implies that $T_{N2}$ is asymptotically a sum of i.i.d. random variables. Combining these facts with (2.7) and (3.4), we obtain the asymptotic normality of $S_N$.

**Acknowledgment.** The author thanks Professor M. Okamoto and the referee for their encouragement and comments.

## REFERENCES

[1] CHERNOFF, H. and SAVAGE, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Statist.* **29** 972–994.

[2] DUPAČ, V. and HÁJEK, J. (1969). Asymptotic normality of simple linear rank statistics under alternatives II. *Ann. Math. Statist.* **40** 1992–2017.

[3] GASTWIRTH, J. L. (1965). Asymptotically most powerful rank tests for the two-sample problem with censored data. *Ann. Math. Statist.* **36** 1243–1247.

[4] GHOSH, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Statist.* **42** 1957–1961.

[5] GOVINDARAJULU, Z., LE CAM, L. and RAGHAVACHARI, M. (1966). Generalizations of theorems of Chernoff and Savage on the asymptotic normality of test statistics. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 609–638. Univ. of California Press.

[6] HÁJEK, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *Ann. Math. Statist.* **39** 325–346.

[7] HÁJEK, J. and ŠIDÁK, Z. (1967). *Theory of Rank Tests.* Academic Press, New York.

[8] 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.

[9] KOUL, H. L. and STAUDTE, R. G., JR. (1972). Weak convergence of weighted empirical cumulatives based on ranks. *Ann. Math. Statist.* **43** 832–841.

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