

ASYMPTOTIC NORMALITY OF RANK SUMS UNDER DEPENDENCY AND ITS APPLICATIONS TO THE TESTING PROBLEM

Tamura, Ryoji
Department of Mathematics, Faculty of Science, Kumamoto University

<https://doi.org/10.5109/13143>

出版情報：統計数理研究. 19 (3/4), pp.1-8, 1981-03. Research Association of Statistical
Sciences
バージョン：
権利関係：



ASYMPTOTIC NORMALITY OF RANK SUMS UNDER DEPENDENCY AND ITS APPLICATIONS TO THE TESTING PROBLEM

By

Ryoji TAMURA*

(Received March 7, 1980)

1. Introduction.

Let the distribution function of the random vector $X=(X_1, \dots, X_c)$ be $H(x_1, \dots, x_c)$ and the marginal distribution function of X_i be $F_i(x)$, $i=1, \dots, c$ where we assume H to be absolutely continuous. Further, let $X_\alpha=(X_{1\alpha}, \dots, X_{c\alpha})$, $\alpha=1, \dots, n$ be a random sample from the population with the distribution function H . Now we denote the rank of $X_{i\alpha}$ among nc random variables $\{X_{i\alpha}, i=1, \dots, c, \alpha=1, \dots, n\}$ by $R_{i\alpha}$ and define $R_i=\sum_{\alpha=1}^n R_{i\alpha}/n(n-1)$ for $i=1, \dots, c$.

Then this paper is concerned with the asymptotic distribution of the linear rank statistic $R_d=\sqrt{n}\sum_{i=1}^c d_i(R_i-p_i)$ and its applications to some testing problem where d_i 's are any constants which are not all equal and p_i 's are defined in section 2. The asymptotic distribution of rank statistics, which is one of the most essential parts in nonparametric theory, has been studied by many workers and in particular, it is well-known that Chernoff-Savage [1] and Hájek [2] have established most fruitful results in this field. They have discussed under the assumption that the components X_i 's are independent, but not considered for the case that the independence of X_i 's is violated. It is, therefore, of interest in studying the asymptotic distribution of R_d and its applications when X_i 's are not independent, though R_d is the simplest rank statistic.

In what follows, the summation \sum extends over all integers from 1 to c when the index is i, j, k or l and from 1 to n when the index is α, β, γ or δ .

2. Asymptotic normality of R_d .

The rank $R_{i\alpha}$ of $X_{i\alpha}$ may be written as follows,

$$(2.1) \quad R_{i\alpha} = \sum_j \sum_{\beta} u(X_{i\alpha} - X_{j\beta}) + 1,$$

where

* Department of Mathematics, Faculty of Science, Kumamoto University, Kumamoto.

$$u(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then we easily obtain

$$(2.2) \quad E(R_{i\alpha}) = (n-1)p_i + p'_i + 1$$

$$(2.3) \quad E(R_i) = p_i + (p'_i + 1)/(n-1),$$

where

$$p_i = \sum_j p_{ij}, \quad p_{ij} = E(F_j(X_i)), \quad p'_i = \sum_j p'_{ij}, \quad p'_{ij} = P(X_i > X_j)$$

and

$$p_{ij} = 1 - p_{ji}, \quad p_{ij} = 1/2 \quad \text{if } F_i = F_j, \quad p'_{ii} = 0.$$

Now let \mathfrak{L} be the class of statistics L of the form $L = \sum_{\alpha} k_{\alpha}(\mathbf{X}_{\alpha})$ where the functions k_{α} may be chosen arbitrarily except $E(k_{\alpha}^2(\mathbf{X}_{\alpha})) < \infty$. We shall first consider the projection of R_i on the space \mathfrak{L} .

2.1. The projection of R_i on \mathfrak{L} . We denote the projection of R_i on \mathfrak{L} by T_i . Then we get the following by the multivariate form of the projection lemma due to Hájek [2]

$$(2.4) \quad T_i = \sum_{\alpha} E(R_i | \mathbf{X}_{\alpha}) - (n-1)E(R_i)$$

$$E(T_i) = E(R_i), \quad E(R_i - T_i)^2 = \text{Var}(R_i) - \text{Var}(T_i).$$

We may first obtain after some calculations,

$$E(R_{i\beta} | \mathbf{X}_{\alpha}) = \begin{cases} \sum_j [(n-1)F_j(X_{i\alpha}) + u(X_{i\alpha} - X_{j\beta})] + 1 & \beta = \alpha \\ \sum_j [(1 - F_i(X_{j\alpha})) + (n-2)p_{ij} + p'_{ij}] + 1 & \beta \neq \alpha \end{cases}$$

and hence

$$(2.5) \quad \begin{aligned} E(R_i | \mathbf{X}_{\alpha}) &= \frac{1}{n} \sum_j [F_j(X_{i\alpha}) + 1 - F_i(X_{j\alpha})] + \frac{n-2}{n} p_i \\ &\quad + \frac{1}{n} p'_i + \frac{1}{n(n-1)} \sum_j u(X_{i\alpha} - X_{j\alpha}) + \frac{1}{n-1} \end{aligned}$$

From (2.4) and (2.5), we also obtain

$$(2.6) \quad \begin{aligned} T_i - E(T_i) &= \frac{1}{n} \sum_{\alpha} \sum_j [F_j(X_{i\alpha}) + 1 - F_i(X_{j\alpha}) - 2p_{ij}] \\ &\quad + \frac{1}{n(n-1)} \sum_{\alpha} \sum_j u(X_{i\alpha} - X_{j\alpha}) - \frac{1}{n-1} p'_i. \end{aligned}$$

We shall show in Lemma 2.1 that R_i is asymptotically equivalent to its projection T_i .

LEMMA 2.1.

$$\sqrt{n}(R_i - T_i) \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty$$

PROOF. Since it follows from (2.4) that

$$nE(R_i - T_i)^2 = n \text{Var}(R_i) - n \text{Var}(T_i), \quad E(R_i) = E(T_i),$$

it suffices to show that

$$(2.7) \quad n(\text{Var}(R_i) - \text{Var}(T_i)) \longrightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Obviously,

$$n \text{Var}(R_i) = \frac{1}{n(n-1)^2} \left[\sum_{\alpha} \text{Var}(R_{i\alpha}) + \sum_{\alpha \neq \beta} \sum \text{Cov}(R_{i\alpha}, R_{i\beta}) \right].$$

After somewhat complicated calculations, we can get

$$\text{Var}(R_{i\alpha}) = (n-1)(n-2) \sum_k \sum_l [E(F_k(X_i)F_l(X_i)) - p_{ik}p_{il}] + O(n)$$

and for $\alpha \neq \beta$,

$$\begin{aligned} \text{Cov}(R_{i\alpha}, R_{i\beta}) &= (n-2) \sum_k \sum_l [E\{F_l(X_i)(1-F_i(X_k))\} + E\{F_k(X_i)(1-F_i(X_l))\} \\ &\quad + E\{(1-F_i(X_k))(1-F_i(X_l))\} - 3p_{ik}p_{il}] + O(1), \end{aligned}$$

and consequently,

$$(2.8) \quad \begin{aligned} n \text{Var}(R_i) &= \sum_k \sum_l E[(F_k(X_i) + 1 - F_i(X_k) - 2p_{ik}) \\ &\quad \cdot (F_l(X_i) + 1 - F_i(X_l) - 2p_{il})] + O\left(\frac{1}{n}\right). \end{aligned}$$

As for $\text{Var}(T_i)$, a direct calculation shows that

$$(2.9) \quad \begin{aligned} n \text{Var}(T_i) &= \frac{1}{n} \text{Var} \left(\sum_{\alpha} \sum_j [F_j(X_{i\alpha}) + 1 - F_i(X_{j\alpha}) - 2p_{ij}] + O\left(\frac{1}{n}\right) \right) \\ &= \sum_k \sum_l E[(F_k(X_i) + 1 - F_i(X_k) - 2p_{ik})(F_l(X_i) + 1 - F_i(X_l) - 2p_{il})] + O\left(\frac{1}{n}\right). \end{aligned}$$

Thus (2.7) easily follows from (2.8) and (2.9).

2.2. Asymptotic normality of $\sqrt{n} \sum_i d_i(R_i - p_i)$.

LEMMA 2.2.

$$\sqrt{n} \sum_i d_i(R_i - p_i) - \sqrt{n} \sum_i d_i \hat{T}_i \xrightarrow{P} 0 \quad \text{as } n \rightarrow \infty,$$

where

$$(2.10) \quad \hat{T}_i = \frac{1}{n} \sum_{\alpha} \sum_j [F_j(X_{i\alpha}) + 1 - F_i(X_{j\alpha}) - 2p_{ij}].$$

PROOF. It is obvious from (2.6) and (2.3) that

$$\sqrt{n} (T_i - E(T_i) - \hat{T}_i) = \sum_{\alpha} \sum_j (u(X_{i\alpha} - X_{j\alpha}) - p'_{ij}) \xrightarrow{P} 0,$$

$$\sqrt{n} (R_i - p_i - (R_i - E(R_i))) = \frac{\sqrt{n}}{n-1} (p'_i + 1) \longrightarrow 0.$$

From these results and Lemma 2.1, we have

$$\begin{aligned} \sqrt{n} \sum_i d_i (R_i - p_i - \hat{T}_i) &= \sqrt{n} \sum_i d_i [(R_i - p_i - (R_i - E(R_i)) + (R_i - T_i) \\ &\quad + (T_i - E(T_i) - \hat{T}_i)] \xrightarrow{P} 0. \end{aligned}$$

THEOREM 2.1. $\sqrt{n} \sum_i d_i (R_i - p_i)$ has the limiting normal distribution $N(0, \sigma^2(R_d))$ provided that $\sigma^2(R_d) \neq 0$, where

$$(2.11) \quad \sigma^2(R_d) = \text{Var} \left(\sum_i \sum_j d_i (F_j(X_i) - F_i(X_j)) \right).$$

PROOF. From Lemma 2.2, it suffices to show the asymptotic normality of $\sqrt{n} \sum_i d_i \hat{T}_i$. Now we may express $\sum_i d_i \hat{T}_i$ by

$$\sum_i d_i \hat{T}_i = \frac{1}{n} \sum_\alpha Y_\alpha, \quad \text{where } Y_\alpha = \sum_i \sum_j d_i [F_j(X_{i\alpha}) + 1 - F_i(X_{j\alpha}) - 2p_{ij}].$$

Then $\{Y_\alpha\}$, $\alpha=1, \dots, n$ are independent and identically distributed random variables with $\text{Var}(Y_\alpha) < \infty$. Thus we may apply the central limit theorem to show the asymptotic normality $N(0, 1)$ of $\sqrt{n} \sum_i d_i \hat{T}_i / \sigma(R_d)$ if $\sigma(R_d) \neq 0$.

2.3. A consistent estimator of $\sigma^2(R_d)$. In order to use R_d for the testing problem in section 3, we here derive a consistent estimator of $\sigma^2(R_d)$. Define S_α , $\alpha=1, \dots, n$ as follows,

$$(2.12) \quad S_\alpha = \frac{1}{n} \sum_i \sum_j \sum_\beta d_i [u(X_{i\alpha} - X_{j\beta}) + u(X_{i\beta} - X_{j\alpha})]$$

$$(2.13) \quad \hat{\sigma}_n^2 = \frac{1}{n} \sum_\alpha (S_\alpha - \bar{S})^2, \quad \text{where } \bar{S} = \frac{1}{n} \sum_\alpha S_\alpha.$$

LEMMA 2.3. $\hat{\sigma}_n^2$ is a consistent estimator of $\sigma^2(R_d)$.

PROOF. Notice first that $\{S_\alpha\}$, $\alpha=1, \dots, n$ are independent and identically distributed random variables. S_α^2 may be expressed as

$$S_\alpha^2 = \frac{1}{n^2} \sum_i \sum_j \sum_k \sum_l \sum_\beta \sum_\gamma d_i d_j v_{ik\alpha\beta} v_{jl\alpha\gamma},$$

where

$$v_{ik\alpha\beta} = u(X_{i\alpha} - X_{k\beta}) + u(X_{i\beta} - X_{k\alpha}).$$

Then we have

$$(2.14) \quad E(S_\alpha^2) = \sum_i \sum_j \sum_k \sum_l d_i d_j E[(F_k(X_i) + 1 - F_i(X_k))(F_l(X_j) + 1 - F_j(X_l))] + O\left(\frac{1}{n}\right).$$

Since \bar{S} may be written as

$$\bar{S} = \frac{1}{n^2} \sum_i \sum_j \sum_\alpha \sum_\beta d_i v_{ij\alpha\beta} = \frac{2(n-1)}{n} \sum_i d_i R_i,$$

we get from (2.3)

$$(2.15) \quad E(\bar{S}) = 2 \sum_i d_i p_i + O\left(\frac{1}{n}\right).$$

Therefore we easily find from (2.14) and (2.15) that

$$(2.16) \quad \begin{aligned} E(\hat{\sigma}_n^2) &= \sum_i \sum_j \sum_k \sum_l d_i d_j E[(F_k(X_i) + 1 - F_i(X_k) - 2p_{ik})(F_l(X_j) + 1 - F_j(X_l) - 2p_{jl})] + O\left(\frac{1}{n}\right) \\ &= \text{Var} \left[\sum_i \sum_j d_i (F_j(X_i) - F_i(X_j)) \right] + O\left(\frac{1}{n}\right) \longrightarrow \sigma^2(R_d) \quad \text{as } n \rightarrow \infty. \end{aligned}$$

We shall turn to show that $\text{Var}(\hat{\sigma}_n^2) \rightarrow 0$ as $n \rightarrow \infty$. Since $\hat{\sigma}_n^2$ may be written as

$$\hat{\sigma}_n^2 = \frac{1}{n^3} \sum_\alpha \left(\sum_\beta v_{\alpha\beta} \right)^2 - \frac{1}{n^4} \left(\sum_\alpha \sum_\beta v_{\alpha\beta} \right)^2, \quad \text{where } v_{\alpha\beta} = \sum_i \sum_j d_i v_{ij\alpha\beta},$$

we easily find the following

$$(2.17) \quad \begin{aligned} \text{Var}(\hat{\sigma}_n^2) &= \frac{1}{n^6} \text{Var} \left(\sum_\alpha \sum_\beta \sum_\gamma v_{\alpha\beta} v_{\alpha\gamma} \right) + \frac{1}{n^8} \text{Var} \left(\sum_\alpha \sum_\beta \sum_\gamma \sum_\delta v_{\alpha\beta} v_{\gamma\delta} \right) \\ &\quad - \frac{2}{n^7} \text{Cov} \left(\sum_\alpha \sum_\beta \sum_\gamma v_{\alpha\beta} v_{\alpha\gamma}, \sum_\alpha \sum_\beta \sum_\gamma \sum_\delta v_{\alpha\beta} v_{\gamma\delta} \right). \end{aligned}$$

After some elementary but tedious calculations, it follows that

$$(2.18) \quad \begin{aligned} \text{Var} \left(\sum_\alpha \sum_\beta \sum_\gamma v_{\alpha\beta} v_{\alpha\gamma} \right) &= O(n^5) \\ \text{Var} \left(\sum_\alpha \sum_\beta \sum_\gamma \sum_\delta v_{\alpha\beta} v_{\gamma\delta} \right) &= O(n^7) \\ \text{Cov} \left(\sum_\alpha \sum_\beta \sum_\gamma v_{\alpha\beta} v_{\alpha\gamma}, \sum_\alpha \sum_\beta \sum_\gamma \sum_\delta v_{\alpha\beta} v_{\gamma\delta} \right) &= O(n^6). \end{aligned}$$

Thus we obtain from (2.17) and (2.18)

$$(2.19) \quad \text{Var}(\hat{\sigma}_n^2) \longrightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Lemma 2.3 follows from (2.16) and (2.19).

COROLLARY 2.1. *The statistic $\sqrt{n} \sum d_i (R_i - p_i) / \hat{\sigma}_n$ has the limiting standard normal distribution.*

The proof is easily shown from Theorem 2.1 and Lemma 2.3.

3. A class of tests for homogeneity of marginal distributions.

As an application of the results in section 2, we here consider the problem of testing hypothesis

$$H_0: F_1(x) = \cdots = F_c(x)$$

against the ordered alternative

$$H_1: F_1(x) \geq \cdots \geq F_c(x),$$

where at least one of the inequalities is strict.

For this problem, Tamura [4] has proposed a test W under that X_i 's are not independent, that is

$$W = \sum_{i < j} W_{ij}, \quad W_{ij} = \sum_{\alpha \neq \beta} u(X_{j\beta} - X_{i\alpha}),$$

where we note that W_{ij} is the U -statistic proposed by Raviv [3] for the two-sample problem. We here propose a class of tests R_d which reject H_0 if

$$(3.1) \quad \sqrt{n} \sum_i d_i \left(R_i - \frac{c}{2} \right) \geq \hat{\sigma}_n z_\alpha,$$

where d_i 's are any constants satisfying $d_i \leq \cdots \leq d_c$ (at least one of inequalities is strict) and z_α is the $(1-\alpha)$ percentile point of the standard normal distribution $\Phi(x)$. Then it is easily shown from Corollary 2.1 that the test R_d has asymptotic level α of significance.

It is very important in applications how to choose d_i 's, but we can get no answer for this problem. We can only show later that a test R_d with $d_i = i$ for $i=1, \dots, c$ is asymptotically equivalent to the test W .

3.1. Asymptotic power of R_d . We shall now derive the asymptotic power of the test R_d for the translation alternative

$$H_1^*: F_i(x) = F(x - \theta_i / \sqrt{N}), \quad i=1, \dots, c,$$

where $\theta_1 \leq \cdots \leq \theta_c$ (at least one of inequalities is strict) and $n/N \rightarrow \lambda$, $0 < \lambda_0 \leq \lambda \leq \lambda_1$, as $N \rightarrow \infty$. Denoting the limiting power function of the test R_d be $\beta_R(\theta)$, then we have

$$(3.2) \quad \begin{aligned} \beta_R(\theta) &= \lim_{N \rightarrow \infty} P \left[\sqrt{n} \sum_i d_i \left(R_i - \frac{c}{2} \right) \geq \hat{\sigma}_n z_\alpha \mid H_1^* \right] \\ &= \lim_{N \rightarrow \infty} P \left[\sqrt{n} \sum_i d_i (R_i - p_i^*) \geq \hat{\sigma}_n z_\alpha + \sqrt{n} \sum_i d_i \left(\frac{c}{2} - p_i^* \right) \mid H_1^* \right], \end{aligned}$$

where

$$p_i^* = \sum_j p_{ij}^*, \quad p_{ij}^* = E[F(X_i - \theta_j / \sqrt{N}) \mid H_1^*].$$

We here notice that (i) under H_1^*

$$(3.3) \quad \begin{aligned} \hat{\sigma}_n^2 &\rightarrow \sigma_0^2(R_d) = \text{Var} \left(\sum_i \sum_j d_i (F(X_i) - F(X_j)) \mid H_0 \right) \\ &= \frac{c^2}{12} \sum_i \sum_j (d_i - \bar{d})(d_j - \bar{d}) \rho_0(F(X_i), F(X_j)), \end{aligned}$$

where $\rho_0(F(X_i), F(X_j))$ is the correlation coefficient of $F(X_i)$ and $F(X_j)$ under H_0 and

$\bar{d} = \sum_i d_i/c$ and (ii) under the assumption that the density f of F is bounded,

$$(3.4) \quad \begin{aligned} \sqrt{n} \left(\frac{c}{2} - p_i^* \right) &= \sqrt{n} \int_{-\infty}^{\infty} [F(x + (\theta_i - \theta_j)/\sqrt{N}) - F(x)] dF(x) \\ &\longrightarrow \sqrt{\lambda} (\theta_i - \theta_j) \int_{-\infty}^{\infty} f(x) dF(x) \quad \text{as } N \rightarrow \infty. \end{aligned}$$

From Corollary 2.1, (3.3) and (3.4), $\beta_R(\theta)$ is represented as follows,

$$(3.5) \quad \beta_R(\theta) = 1 - \Phi(z - \sqrt{\lambda} c \left(\int_{-\infty}^{\infty} f^2(x) dx \sum_i (d_i - \bar{d})(\theta_i - \bar{\theta}) / \sigma_0(R_d) \right)),$$

where $\bar{\theta} = \sum_i \theta_i/c$.

3.2. Comparison with other tests.

(a) The test W . The test W reject H_0 if

$$(3.6) \quad W \geq \frac{1}{4} c(c-1) + \hat{A}_n z_\alpha / \sqrt{n},$$

where \hat{A}_n^2 is a consistent estimator of $A^2 = \text{Var} \left(\sum_{i < j} [F_j(X_i) - F_i(X_j)] \right)$ as defined in [4].

Then the limiting power of the test W under H_1^* has been expressed by

$$(3.7) \quad \beta_W(\theta) = 1 - \Phi(z_\alpha - \sqrt{\lambda} \left(\int_{-\infty}^{\infty} f^2(x) dx \sum_{i < j} (\theta_j - \theta_i) \right) / A_0),$$

where

$$A_0^2 = \text{Var} \left(\sum_{i < j} [F(X_i) - F(X_j)] | H_0 \right).$$

Comparing the test R_d when $d_i = i$ with the test W in the limiting power, we can easily show $\beta_R(\theta) = \beta_W(\theta)$ by noticing the identities

$$\sum_i \sum_j i(F(X_i) - F(X_j)) = \frac{c}{2} \sum_{i < j} (F(X_i) - F(X_j)),$$

$$\sum_i i(\theta_i - \bar{\theta}) = \sum_{i < j} (\theta_j - \theta_i) / 2$$

and consequently, $\sigma_0^2(R_d) = c^2 A_0^2 / 4$.

(b) The test based on the sample means. Finally, we shall make large sample comparison between the test R_d and the corresponding competitor T_d based on Student statistic,

$$(3.8) \quad T_d = \sqrt{n} \sum_i d_i (\bar{X}_i - \bar{X}), \quad \bar{X}_i = \sum_\alpha X_{i\alpha} / n, \quad \bar{X} = \sum_i \bar{X}_i / c,$$

This statistic will be used for our testing problem assuming that $\text{Cov}(X)$ is known. Then the test T_d of asymptotic level α reject H_0 if

$$(3.9) \quad T_d \geq \sigma_0(T_d) z_\alpha, \quad \sigma_0^2(T_d) = \sigma_0(X) \sum_i \sum_j (d_i - \bar{d})(d_j - \bar{d}) \rho_0(X_i, X_j)$$

where $\sigma_0^2(X)$ and $\rho_0(X_i, X_j)$ are respectively the variance of X_i and the correlation of X_i and X_j under H_0 .

Asymptotic normality of T_d is easily shown as in [4] and consequently, we can get the limiting power under H^*

$$(3.10) \quad \beta_T(\theta) = 1 - \Phi(z_\alpha - \sqrt{\lambda} \sum_i (d_i - \bar{d})(\theta_i - \bar{\theta}) / \sigma_0(T_d)).$$

From (3.4) and (3.10), the Pitman relative efficiency of R_d with respect to T_d may be expressed as follows

$$(3.11) \quad e(R_d, T_d) = 12\sigma_0^2(X) \left(\int_{-\infty}^{\infty} f^2(x) dx \right)^2 \sum_i \sum_j (d_i - \bar{d})(d_j - \bar{d}) \rho_0(X_i, X_j) \\ \div \sum_i \sum_j (d_i - \bar{d})(d_j - \bar{d}) \rho_0(F(X_i), F(X_j)).$$

References

- [1] CHERENOFF, H. and SAVAGE, I.R.: *Asymptotic normality and efficiency of certain non-parametric test statistics*. Ann. Math. Statist. **29** (1958) 972-994.
- [2] HÁJEK, J.: *Asymptotic normality of simple linear rank statistics under alternatives*. Ann. Math. Statist. **39** (1968) 325-346.
- [3] RAVIV, A.: *A non-parametric test for comparing two non-independent distributions*. J. Roy. Statist. Soc. B **40** (1978) 253-261.
- [4] TAMURA, R.: *A test of Wilcoxon type for homogeneity of marginal distributions against ordered alternatives*. Kumamoto J. Sci. (Math.) **14** (1980) 34-41.

Department of Mathematics
Faculty of Science
Kumamoto University

Editorial: This paper is the last publication by Professor Ryoji Tamura who died on November 13, 1980 in Kumamoto.