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

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

バージョン：published
権利関係：

http://hdl.handle.net/2324/13143
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 \( X=(X_1, \ldots, X_c) \) be \( H(x_1, \ldots, x_c) \) and the marginal distribution function of \( X_i \) be \( F_i(x) \), \( i=1, \ldots, c \) where we assume \( H \) to be absolutely continuous. Further, let \( X^n_a=(X_{1a}, \ldots, X_{na}), \alpha=1, \ldots, n \) be a random sample from the population with the distribution function \( H \). Now we denote the rank of \( X_{ia} \) among \( nc \) random variables \( \{X_{ia}, i=1, \ldots, c, \alpha=1, \ldots, n\} \) by \( R_{ia} \) and define \( R_i=\sum_{\alpha=1}^{n} R_{ia}/n(n-1) \) for \( i=1, \ldots, c \).

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 \( \alpha, \beta, \gamma \) or \( \delta \).

2. Asymptotic normality of \( R_d \).

The rank \( R_{ia} \) of \( X_{ia} \) may be written as follows,

\[
R_{ia}=\sum_{j} \sum_{\beta} u(X_{ia}-X_{j\beta})+1,
\]

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

\begin{align*}
(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),
\end{align*}

where

\[ p_i = \sum_j p_{ij}, \quad p_{ij} = E(F_j(X_i)), \quad p_i' = \sum_j p_i' j, \quad p_i' = P(X_i > X_j) \]

and

\[ p_{ij} = 1 - p_{ji}, \quad p_{ij} = 1/2 \text{ if } F_i = F_j, \quad p_i' = 0. \]

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

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

\begin{align*}
(2.4) & \quad T_i = \sum_{\alpha} E(R_i|X_{\alpha}) - (n-1)E(R_i) \\
& \quad E(T_i) = E(R_i), \quad E(R_i - T_i)^2 = \text{Var}(R_i) - \text{Var}(T_i).
\end{align*}

We may first obtain after some calculations,

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

and hence

\begin{align*}
(2.5) & \quad E(R_i|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{align*}

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

\begin{align*}
(2.6) & \quad 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{align*}

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

**Lemma 2.1.**

\[ \sqrt{n} (R_i - T_i) \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty \]
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
\[ n(\text{Var}(R_i) - \text{Var}(T_i)) \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty. \] 
Obviously,
\[ n \text{Var}(R_i) = \frac{1}{n(n-1)^2} \left[ \sum \text{Var}(R_{ia}) + \sum \sum \text{Cov}(R_{ia}, R_{ij}) \right]. \]
After somewhat complicated calculations, we can get
\[ \text{Var}(R_{ia}) = (n-1)(n-2) \sum \sum \left[ E(F_k(X_i)F_j(X_i) - p_{ik} p_{ij}) \right] + O(n) \]
and for \( \alpha \neq \beta, \)
\[ \text{Cov}(R_{ia}, R_{ij}) = (n-2) \sum \sum \left[ E(F_i(X_j)(1-F_i(X_k))) + E(F_i(X_i)(1-F_i(X_k))) \right] \]
and consequently,
\[ n \text{Var}(R_i) = \sum \sum E[(F_k(X_i)+1-F_i(X_k)-2p_{ik}) \cdot (F_i(X_i)+1-F_i(X_k)-2p_{ij})] + O\left(\frac{1}{n}\right). \]
As for \( \text{Var}(T_i), \) a direct calculation shows that
\[ n \text{Var}(T_i) = \frac{1}{n} \text{Var} \left( \sum \sum [F_j(X_{ia})+1-F_j(X_{ja})-2p_{ja}] + O\left(\frac{1}{n}\right) \right) \]
\[ = \sum \sum E[(F_k(X_i)+1-F_i(X_k)-2p_{ik})(F_i(X_i)+1-F_i(X_k)-2p_{ij})] + O\left(\frac{1}{n}\right). \]
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.

\[ \sqrt{n} \sum_i d_i(R_i - p_i) - \sqrt{n} \sum_i T_i \rightarrow P \quad \text{as} \quad n \rightarrow \infty, \]
where
\[ T_i = \frac{1}{n} \sum \sum [F_j(X_{ia})+1-F_j(X_{ja})-2p_{ja}]. \]

PROOF. It is obvious from (2.6) and (2.3) that
\[ \sqrt{n} (T_i - E(T_i) - T_i) = \sum \sum (u(X_{ia} - X_{ja}) - p_{ij}) \rightarrow 0, \]
From these results and Lemma 2.1, we have

\[ \sqrt{n} \sum d_i (R_i - p_i - \hat{T}_i) = \sqrt{n} \sum d_i [(R_i - p_i - (R_i - E(R_i) + (R_i - T_i))] \]

\[ + \langle T_i - E(T_i) - \hat{T}_i \rangle \to 0. \]

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

\[ (2.11) \sigma^2(R_d) = \text{Var} \left( \sum_j d_i (F_j(X_i) - F_i(X_i)) \right). \]

**Proof.** From Lemma 2.2, it suffices to show the asymptotic normality of \( \sqrt{n} \sum d_i \hat{T}_i \). Now we may express \( \sum d_i \hat{T}_i \) by

\[ \sum d_i \hat{T}_i = \frac{1}{n} \sum_a Y_a, \text{ where } Y_a = \sum_i \sum_j [F_j(X_{ia}) + 1 - F_i(X_{ja}) - 2p_{ij}]. \]

Then \( \{Y_a\}, a = 1, \ldots, n \) are independent and identically distributed random variables with \( \text{Var}(Y_a) < \infty \). Thus we may apply the central limit theorem to show the asymptotic normality \( N(0, 1) \) of \( \sqrt{n} \sum d_i \hat{T}_i / \sigma(R_d) \) 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 \( S_a, a = 1, \ldots, n \) as follows,

\[ (2.12) S_a = \frac{1}{n} \sum_j \sum \sum d_i [u(X_{ia} - X_j) + u(X_{ia} - X_j)] \]

\[ (2.13) \hat{\sigma}_a^2 = \frac{1}{n} \sum (S_a - \bar{S})^2, \text{ where } \bar{S} = \frac{1}{n} S_a. \]

**Lemma 2.3.** \( \hat{\sigma}_a^2 \) is a consistent estimator of \( \sigma^2(R_d) \).

**Proof.** Notice first that \( \{S_a\}, a = 1, \ldots, n \) are independent and identically distributed random variables. \( S_a \) may be expressed as

\[ S_a = \frac{1}{n^2} \sum_i \sum_j \sum_k \sum_l \sum \sum d_i d_j v_{ik\alpha\beta} v_{jl\gamma}, \]

where

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

Then we have

\[ (2.14) E(S_a^2) = \sum_i \sum_j \sum k \sum l d_i d_j E[u(F_k(X_i) + 1 - F_i(X_k))(F_l(X_j) + 1 - F_i(X_l))] + \mathcal{O} \left( \frac{1}{n} \right). \]

Since \( \bar{S} \) may be written as
Asymptotic Normality of Rank Sums Under Dependency

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

we get from (2.3)

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

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

\[ E(\hat{\theta}_n) = \frac{\sum_i \sum_j \sum_{\alpha} \sum_{\beta} d_i d_j E[(F_\alpha(X_i) + 1 - F_i(X_i) - 2p_{ik})(F_{\beta}(X_j) + 1 - F_j(X_i) - 2p_{ij})] + O(1)}{n} \]

\[ = \text{Var}\left[ \sum_i \sum_j d_i (F_\alpha(X_i) - F_i(X_i)) \right] + O\left(\frac{1}{n}\right) \rightarrow \sigma^2(R_d) \quad \text{as} \quad n \rightarrow \infty. \]

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

\[ \hat{\theta}_n = \frac{1}{n^3} \sum_{\alpha} \left( \sum_{\beta} v_{\alpha\beta} \right)^2 - \frac{1}{n^4} \left( \sum_{\alpha} \sum_{\beta} v_{\alpha\beta} \right)^2, \]

we easily find the following

\[ \text{Var}(\hat{\theta}_n) = \frac{1}{n^4} \text{Var}\left( \sum_{\alpha} \sum_{\beta} \sum_{\gamma} v_{\alpha\beta\gamma} \right) + \frac{1}{n^5} \text{Var}\left( \sum_{\alpha} \sum_{\beta} \sum_{\gamma} v_{\alpha\beta\gamma} \right) \]

\[ - \frac{2}{n^7} \text{Cov}\left( \sum_{\alpha} \sum_{\beta} \sum_{\gamma} v_{\alpha\beta\gamma}, \sum_{\alpha} \sum_{\beta} \sum_{\gamma} v_{\alpha\beta\gamma} \right). \]

After some elementary but tedious calculations, it follows that

\[ \text{Var}\left( \sum_{\alpha} \sum_{\beta} \sum_{\gamma} v_{\alpha\beta\gamma} \right) = O(n^5) \]

\[ (2.18) \quad \text{Var}\left( \sum_{\alpha} \sum_{\beta} \sum_{\gamma} v_{\alpha\beta\gamma} \right) = O(n^7) \]

\[ \text{Cov}\left( \sum_{\alpha} \sum_{\beta} \sum_{\gamma} v_{\alpha\beta\gamma}, \sum_{\alpha} \sum_{\beta} \sum_{\gamma} v_{\alpha\beta\gamma} \right) = O(n^6). \]

Thus we obtain from (2.17) and (2.18)

\[ (2.19) \quad \text{Var}(\hat{\theta}_n) \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty. \]

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

**Corollary 2.1.** The statistic \( \sqrt{n} \sum d_i (R_i - p_i)/\hat{\theta}_n \) 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
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

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

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

$$\sqrt{n} \sum_i d_i \left( R_i - \frac{c}{2} \right) \geq \sigma_{n} z_{\alpha},$$

where $d_i$'s are any constants satisfying $d_1 \leq \cdots \leq d_c$ (at least one of inequalities is strict) and $z_{\alpha}$ 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 $\alpha$ 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 $i=1, \cdots, c$ 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

$$H^\natural_i : F_i(x) = F(x - \theta_i / \sqrt{N}), \quad i = 1, \cdots, c,$$

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

$$\beta_R(\theta) = \lim_{N \rightarrow \infty} \left[ \sqrt{n} \sum_i d_i \left( R_i - \frac{c}{2} \right) \geq \sigma_{n} z_{\alpha} | H^\natural_i \right]$$

$$= \lim_{N \rightarrow \infty} \left[ \sqrt{n} \sum_i d_i \left( R_i - \mu_i \right) \geq \sigma_{n} z_{\alpha} + \sqrt{n} \sum_i d_i \left( \frac{c}{2} - \mu_i \right) | H^\natural_i \right],$$

where

$$\mu_i = \sum_j \mu_{ij}, \quad \mu_{ij} = E[F(X_i - \theta_j / \sqrt{N}) | H^\natural_i].$$

We here notice that (i) under $H^\natural_i$

$$\sigma_n^2 \rightarrow \sigma^2 (R_d) = \text{Var} \left( \sum_i \sum_j d_i(F(X_i) - F(X_j)) | H_0 \right)$$

$$= \frac{c^2}{12} \sum_i \sum_j (d_i - \bar{d}_i) (d_j - \bar{d}_j) \rho_{ij}(F(X_i), F(X_j)),$$

where $\rho_{ij}(F(X_i), F(X_j))$ is the correlation coefficient of $F(X_i)$ and $F(X_j)$ under $H_0$ and
Asymptotic normality of rank sums under dependency

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

\[
\sqrt{n} \left( \frac{c/2 - \theta^2}{\sqrt{n}} \right) = \sqrt{n} \int_{-\infty}^{\infty} \left[ F(x + (\theta_i - \theta_j)/\sqrt{N}) - F(x) \right] dF(x) \\
\rightarrow \sqrt{n} (\theta_i - \theta_j) \int_{-\infty}^{\infty} f(x) dF(x) \quad \text{as} \quad N \rightarrow \infty.
\]

From Corollary 2.1, (3.3) and (3.4), \( \beta_{\rho}(\theta) \) is represented as follows,

\[
\beta_{\rho}(\theta) = 1 - \Phi \left( z_a - \sqrt{\lambda} \left( \int_{-\infty}^{\infty} f^2(x) dx \sum_i (d_i - \bar{d})(\theta_i - \theta) / \sigma_{\rho}(R_d) \right) \right),
\]

where \( \theta = \sum \theta_i / c \).

3.2. Comparison with other tests.

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

\[
W \geq 1 - \Phi \left( z_a - \sqrt{\lambda} \left( \int_{-\infty}^{\infty} f^2(x) dx \sum_{i \neq j} (\theta_i - \theta_j) / \sigma_{\rho}(R_d) \right) \right) / A_\lambda,
\]

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

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

\[
\beta_{\rho}(\theta) = 1 - \Phi \left( z_a - \sqrt{\lambda} \left( \int_{-\infty}^{\infty} f^2(x) dx \sum_{i \neq j} (\theta_i - \theta_j) \right) / A_\lambda \right),
\]

where

\[
A_\lambda = \text{Var} \left( \sum_{i \neq j} [F(X_i) - F(X_j)] | H_0 \right).
\]

Comparing the test \( R_d \) when \( d_i = i \) with the test \( W \) in the limiting power, we can easily show \( \beta_{\rho}(\theta) = \beta_{\rho}(\theta) \) by noticing the identities

\[
\sum_i \sum_j i(F(X_i) - F(X_j)) = \frac{c}{2} \sum_{i \neq j} (F(X_i) - F(X_j)),
\]

\[
\sum_i (\theta_i - \theta) = \sum_{i \neq j} (\theta_i - \theta_j) / 2
\]

and consequently, \( \sigma_{\rho}^2(R_d) = c^2 A_\lambda^2 / 4 \).

(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,

\[
T_d = \sqrt{n} \sum_i d_i (X_i - \bar{X}), \quad \bar{X}_i = \sum_{a} X_{ia} / n, \quad \bar{X} = \sum_i \bar{X}_i / c.
\]

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

\[
T_d \approx \sigma_{\rho}(T_d) z_a, \quad \sigma_{\rho}(T_d) = \sigma_{\rho}(X) \sum_i \sum_j (d_i - \bar{d})(d_j - \bar{d}) \rho_{\rho}(X_i, X_j)
\]
where $\sigma^2(X)$ and $\rho(X_1, X_2)$ 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 $H_1^+$

\[(3.10) \quad \beta_T(\theta) = 1 - \Phi(z_a - \sqrt{\lambda} \sum_i (d_i - \bar{d}) (\theta_i - \bar{\theta}) / \sigma_T(T_d)).\]

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

\[(3.11) \quad e(R_d, T_d) = 12 \sigma^2(X) \left( \int_{-\infty}^{\infty} f^2(x) dx \right)^2 \sum_i \sum_j (d_i - \bar{d})(d_j - \bar{d}) \rho(X_i, X_j) \]
\[\times \left( \sum_i \sum_j (d_i - \bar{d})(d_j - \bar{d}) \rho(F(X_i), F(X_j)) \right).\]

References


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