Rank Tests For Unbalanced Two-Way Analysis of Variance

by

Wenzi Qiu
Department of Statistical Science, Southern Methodist University

Technical Report No. SMU/DS/TR-272

RANK TESTS FOR UNBALANCED TWO-WAY ANALYSIS OF VARIANCE

By Wenzi Qiu

Department of Statistical Science, Southern Methodist University

SUMMARY

In Analysis of Variance designs with independent observations, the rank transform test consists of ranking all observations together and then substituting the ranks for the observations in the classical F-test. Previous research has shown that the asymptotic distribution of the rank transform test is not always χ^2. In the present research, two rank tests, of which one is the rank transform, are derived in the way similar to the derivation of the classical F-test in Linear Models. Then, necessary and sufficient conditions are given for these rank test to be asymptotically χ^2 by the theory of linear rank statistics, and their limiting properties are studied by using theorems in linear models and quadratic forms. It is found that even when the rank transform test does not converge in distribution to a chi-sqaured distribution, does the other one under certain conditions. Finally, based on the above results, the application of these rank tests to unbalanced two-way designs is investigated.

1. Introduction

The study of the rank transform test for two-way layouts has been an area of research for longer than two decades. Two of the main approaches for obtaining asymptotic properties have been to either apply theorems about linear

Key words: Analysis of variance; Asymptotic normality; Linear rank statistics; Rank tests; Rank transform; Unbalanced two-way layouts.

rank statistics, or apply the central limit theorem to the results from empirical processes.

The first approach has been applied to scored ranks generated by real functions that have bounded second derivatives. Examples of such scores include Wilcoxon scores. This approach has yielded both the rank transform tests, such as presented by Lemmer and Stoke (1967), Lemmer (1968), Lemmer (1980), Iman (1974), Conover and Iman (1976), Conover and Iman (1980), Hora and Conover (1984), and other alternative rank tests as by Thompson (1991a). All tests proposed in those papers are for balanced two-way designs and for score functions with bounded second derivatives.

In the second approach, followed by Akritas (1990), the Wilcoxon scores are expressed as an additive model and scaled by an estimate of the unknown standard deviation of the rank. This approach makes the limiting distribution of the rank statistics easier to obtain, and can be applied to unbalanced designs. However, this method has only been applied to Wilcoxon scores, two-way nested models, and two-way models without interaction but with proportional sampling.

In the above research, three very interesting problems remain open. First, necessary and sufficient conditions for the rank transform tests to be asymptotically chi-squared are unknown. Second, rank tests, including rank transform tests, for unbalanced two-way layouts have not been fully studied. Third, the nature of these asymptotic tests have not been fully revealed.

The present research takes the first approach to study the above questions. The results obtained are based on the theorems about linear rank statistics due to Hájek (1968) and theorems about quadratic forms and linear models due to Searle (1971) and Kshirsagar (1983).

2. Definition of the Model and Test Statistics

Consider the linear model $Y = X\beta + \epsilon$, where ϵ is an $N \times 1$ vector of independent errors with absolutely continuous distribution functions, and X is a design matrix of $N \times P$ with I patterns in the rows, this means that there are I cells. Suppose each of the I patterns is repeated n_i, $1 \le i \le I$, times and that $N = \sum n_i$. Without loss of generality, the matrix X will be assumed not to contain any column of all 1's. For any square matrix M, let M^- denote a g-inverse of M with $\text{rank}(M^-) = \text{rank}(M)$. Suppose that the null hypothesis of interest is H_o: $K'\beta = 0$ where K' is an $S \times P$ matrix of full row rank such that $K' = K'(X'X)^- X'X$. Let $u(x-y) = 1$ if $x \ge y$ and $u(x-y) = 0$ if $x < y$. Then, the rank of Y_{ik} is:

$$R_{ij} = \sum_{i=1}^{I} \sum_{k=1}^{n_i} u(Y_{ij} - Y_{ik}).$$

Let $a_N = [a_N(R_{11}), \cdots, a_N(R_{In_I})]'$, generated by a real-valued, nondecreasing function $\phi(t)$ with $0<t<1$, denote the vector of scored ranks of the elements of Y, and define the vector $\hat{\beta}_r = (X'X)^- X' a_N$.

Under the classical assumptions, as described in Searle (1971, p.190 – 191), the two methods based on $\hat{\beta} = (X'X)^- X'Y$ to obtain test statistics for testing for H_o: $K'\beta = 0$ are:

i) the distributional method, which yields statistic

$$T = (K'\hat{\beta})'\{K'(X'X)^- K\sigma^2\}^{-1}(K'\hat{\beta}),$$

ii)the constrained least squares method, which yields

$$Q = (K'\hat{\beta})'\{K'(X'X)^- K\}^{-1}(K'\hat{\beta}).$$

Then, the classical F-test is

$$F = Q/(S\hat{\sigma}^2) = \sigma^2 T/(S\hat{\sigma}^2),$$

where $\hat{\sigma}^2 = (Y - X\beta)'(Y - X\beta)/\{N - \text{rank}(X)\}$.

To parallel the development of these two methods, replace $\hat{\beta}$ by $\hat{\beta}_r$. Define the asymptotic covariance matrix of $(K'\hat{\beta}_r)$ by $\Lambda_N = NK'(X'X)^- \Sigma_N (X'X)^- K$, the counterpart of T, Q and the rank transform, respectively, is:

$$T_r = (K'\hat{\beta}_r)'\Lambda_N^-(K'\hat{\beta}_r),$$

$$Q_r = (K'\hat{\beta}_r)'\{K'(X'X)^- K\}^{-1}(K'\hat{\beta}_r),$$

$$\hat{F}_r = Q_r/(S\hat{\sigma}_r^2),$$

where $\hat{\sigma}_r^2 = (a_N - X\hat{\beta}_r)'(a_N - X\hat{\beta}_r)/\{N - \text{rank}(X)\}$. If Λ_N is replaced by a consistent estimate, $\hat{\Lambda}_N$, it yields the test statistic $\hat{T}_r = (K'\hat{\beta}_r)'\hat{\Lambda}_N^-(K'\hat{\beta}_r)$. The asymptotic properties of statistics $\hat{T}_r$ can be discussed in terms of those of T_r.

3. Limiting Properties of the Test Statistics

Once the asymptotic normality of $X'a_N$ is obtained, theorems in quadratic forms and linear models can be applied to find the limiting properties of all the test statistics contained $X'a_N$.

3.1 Asymptotic Normality of the Vector of Linear Rank Statistics

Let x_l be the lth row in X', $1 \le l \le P$, $\mathrm{x}_{ik,l}$ be the elements in x_l. Let

$$X'a_N = [L_{N,1}, \cdots, L_{N,P}]',$$

where $L_{N,l} = \sum_{i=1}^{I} \sum_{k=1}^{n_i} \mathrm{x}_{ik,l}\, a_N(R_{ik})$, $1 \le l \le P$. Then, $L_{N,l}$ is a simple linear rank statistic as defined in Hajek (1968) and Puri & Sen (1971). Denote the cumulative distribution function at y by $\Phi(y)$. Also define

$$H_N(\mathrm{y}) = \frac{1}{N}\sum_{i=1}^{I} n_i F_i(\mathrm{y}),\ \bar{\mathrm{x}}_l = \frac{1}{N_i}\sum_{i=1}^{I}\sum_{k=1}^{n_i} \mathrm{x}_{ik,l},$$

$$Z_{ik,l} = \frac{1}{N+1} \sum_{i'=1}^{I} \sum_{k'=1}^{n_{i'}} (x_{i'k',l} - x_{ik,l}) \int \{u(y - Y_{ik}) - F_i(y)\} \phi'\{H_N(y)\} dF_{i'}(y),$$

$$\sigma_{ll',N} = \mathrm{cov}(\sum_{i=1}^{I} \sum_{k=1}^{n_i} Z_{ik,l}, \sum_{i=1}^{I} \sum_{k=1}^{n_i} Z_{ik,l'}), \ \mu_{N,l} = \sum_{i=1}^{I} \sum_{k=1}^{n_i} x_{ik,l} \int \phi\{H_N(y)\} dF_i(y),$$

$$Z_{N,P} = [\sum_{i=1}^{I} \sum_{k=1}^{n_i} Z_{ik,1}, \cdots, \sum_{i=1}^{I} \sum_{k=1}^{n_i} Z_{ik,P}]', \ \mu_N = [\mu_{N,1}, \cdots, \mu_{N,P}]'.$$

Conditions for the linear rank statistic $L_{N,l}$ to be asymptotically normal given by Hájek (1968) in this setting is:

Theorem 3.1.1. Assume that ϕ has a bounded second derivative. Then, for every $\epsilon > 0$ there exists K_ϵ such that

$$\mathrm{var}(L_{N,l}) > K_\epsilon \max_{\substack{1 \le i \le I \\ 1 \le k \le n_i}} (x_{ik,l} - \bar{x}_l)^2 \tag{3.1.1}$$

entails

$$\sup_y |P(L_{N,l} - EL_{N,l} < y\{\mathrm{Var}(L_{N,l})\}^{\frac{1}{2}} - \Phi(y)| < \epsilon . \tag{3.1.2}$$

The assertion remains true if $\mathrm{var}L_{N,l}$ is replaced by $\sigma_{ll,N}$ in (3.1.1) and (3.1.2). The assertion is also true when $E(L_{N,l})$ is replaced by $\mu_{N,l}$ provided that

$$\frac{\sum_{i=1}^{I} \sum_{k=1}^{n_i} x_{ik,l}^2}{\sum_{i=1}^{I} \sum_{k=1}^{n_i} (x_{ik,l} - \bar{x}_l)^2} < \infty. \tag{3.1.3}$$

Let

$$m_l = \max_{\substack{1 \le i \le I \\ 1 \le k \le n_i}} (x_{ik,l} - \bar{x}_l)^2.$$

By Theorem 3.1.1, Corollary 3.1.2 follows straightforwards:

Corollary 3.1.2. Assume that ϕ has a bounded second derivative. If there exists positive constant B and $b > 0$ such that

$$\lim_{N \to \infty} \frac{\sigma_{ll,N}}{N^b m_l} \ge B, \tag{3.1.4}$$

then condition (3.1.1) is satisfied and the assertions of Theorem 3.1.1 follow.

Theorem 3.1.3. Assume that ϕ has a bounded second derivative. If

i) there exists a positive constant B_1 such that $m_l \le B_1 < \infty$ for all $1 \le i \le I$, $1 \le l \le P$;

ii) there exists a positive constant B_2 such that

$$\frac{\sum_{i=1}^{I}\sum_{k=1}^{n_i} x_{ik,l}^2}{\sum_{i=1}^{I}\sum_{k=1}^{n_i}(x_{ik,l}-\bar{x}_l)^2} \le B_2 < \infty, \text{ for all } 1 \le i \le I,\ 1 \le l \le P;$$

iii) there exist constants c_{lk} such that $0 < c_{ll} < \infty$, $|c_{lk}| < \infty$, and $\lim_{N\to\infty} \frac{1}{N}\sigma_{lk,N} = c_{lk}$ for $1 \le l,\ k \le P$;

then, $X'a_N$ is asymptotically multivariate normal with covariance matrix $N(c_{lk})$ and mean vector μ_N.

Proof. It is necessary to prove that for every vector λ such that $\lambda' N(c_{lk})\lambda > 0$ for N sufficiently large, the sequence $\lambda' X_N a_N$ is asymptotically normal with mean $\lambda'\mu_N$ and covariance matrix $\lambda' N(c_{lk})\lambda$. Assume $||\lambda|| = 1$. Rewrite $\lambda' X' a_N$ into

$$\lambda' X' a_N = (X\lambda)' a_N = \sum_{i=1}^{I}\sum_{k=1}^{n_i}\Big(\sum_{l=1}^{P}\lambda_l x_{ik,l}\Big) a_N(R_{ik})$$

with regression constants $\sum_{l=1}^{P}\lambda_l\, x_{ik,l}$, $1 \le i \le I$, $1 \le k \le n_i$. Define

$$Z_{ik,\lambda} = \frac{1}{N+1}\sum_{i'=1}^{I}\sum_{j=1}^{n_{i'}}\Big(\sum_{l=1}^{P}\lambda_l x_{i'j,l} - \sum_{l=1}^{P}\lambda_l x_{ik,l}\Big)\int\{u(y - Y_{ik}) - F_i(y)\}\phi'\{H_N(y)\}\,dF_{i'j}(y).$$

It follows that

$$Z_{ik,\lambda} = \sum_{l=1}^{P}\lambda_l Z_{ik,l},\quad \sum_{i=1}^{I}\sum_{k=1}^{n_i} = Z_{ik,\lambda} = \lambda' Z_N.$$

Denote $\mathrm{var}(\lambda' Z_N)$ by $\sigma_{N,\lambda}^2$, $\lim_{N\to\infty}\sigma_{N,\lambda}^2 \ge 0$. Therefore, (c_{lk}) is semi-positive. To verify (3.1.4) for λ such that $\lambda' N(c_{lk})\lambda > 0$, with condition iii), if $0 < \max_{i,k}\{\sum_{l=1}^{P}\lambda_l(x_{ik,l} - \bar{x}_l)\}^2 < \infty$ holds, $B = \lambda'(c_{lk})\lambda$. The inequality on the left following from: if $\max_{i,k}\{\sum_{l=1}^{P}\lambda_l(x_{ik,l} - \bar{x}_l)\}^2 = 0$, then, $\mathrm{var}(\lambda' Z_N) = 0$; and the one on the right from conditions i) and ii). Also, it is easy to show that condition (3.1.3) is satisfied as well. □

3.2 Limiting Properties of the Test Statistics

Assume that p_i is a rational number satisfying $0 < p_i < 1$ for all i, and

$$\lim_{N\to\infty} \frac{n_i}{N} = p_i \,. \tag{3.2.1}$$

Let $H(y) = \lim_{N\to\infty} H_N(y)$. Define

$$g_{i'}(Y_{ik}) = \int_{Y_{ik}}^{\infty} \phi'\{H(y)\}\, dF_{i'}(y),$$

and $P \times P$ matrix Σ with the $(ll')^{\text{th}}$ element

$$\sigma_{ll'} = \sum_{i=1}^{I} \sum_{i'=1}^{I'} \sum_{u=1}^{I} p_i p_{i'} p_u (x_{i'1,l} - x_{i1,l})(x_{u1,l'} - x_{i1,l'}) \operatorname{cov}\{g_{i'}(Y_{i1}),\, g_u(Y_{i1})\},$$

for $1 \le l,\, l' \le P$.

To obtain the asymptotic normality of $X'a_N$ under (3.2.1), three conditions in Theorem 3.1.3 need to be verified. Under the above assumptions, condition i) and ii) are satisfied, condition iii) will be verified in Lemma 3.2.1. Then, it can be seen that $\lim_{N\to\infty} \{K'N(X'X)^{-}K\}^{-1}$ and $\lim_{N\to\infty} N\Lambda_N$, denoted respectively by Γ and Λ, exist and are not null. The rank of Γ and Λ, a factor determining what Q_r and $\hat{T}_r$ test for, is discussed in Lemma 3.2.3. The limit of the asymptotic mean Q_r and T_r is given in Lemma 3.2.4. Finally, the limiting distribution of Q_r, T_r and $\hat{F}_r$ is given in Theorem 3.2.5, 3.2.6 and Corollary 3.2.7, respectively.

Lemma 3.2.1. Assume that the score function $\phi(t)$ is nondecreasing, and has a bounded second derivative. Under the assumptions of the above model, it follows that $0 < \sigma_{ll} < \infty$, $|\sigma_{ll'}| < \infty$, and $\lim_{N\to\infty} \sigma_{ll',N}/N = \sigma_{lk}$ for all $1 \le l,\, l' \le P$.

Proof. Simplifying the integral on the R.H.S of $Z_{ik,l}$ gives

$$\int \{u(y - Y_{ik}) - F_i(y)\}\phi'\{H_N(y)\}\, dF_u(y) \;= g_{N,u}(Y_{ik}) \;- C_i,$$

where

$$g_{N,u}(Y_{ik}) = \int_{Y_{ik}}^{\infty} \phi'\{H_N(y)\}\, dF_u(y),\; C_i = \int F_i(y)\; \phi'\{H_N(y)\}\, dF_u(y).$$

Then, $Z_{ik,l}$ can be written as

$$Z_{ik,l} = \frac{1}{N+1} \sum_{u=1}^{I} \sum_{w=1}^{n_u} (x_{uw,l} - x_{ik,l})\{g_{N,u}(Y_{ik}) - C_i\} .$$

Computations show that

$$\lim_{N\to\infty} \frac{1}{N} \sigma^2_{ll',N} = \sum_{i=1}^{I} \sum_{i'=1}^{I'} \sum_{u=1}^{I} p_i p_{i'} p_u (x_{i'1,l} - x_{i1,l})(x_{u1,l'} - x_{i1,l'})$$

$$\times \lim_{N\to\infty} \mathrm{cov}\{g_{N,i'}(Y_{i1}),\ g_{N,u}(Y_{i1})\}.$$

for every y, and since ϕ' is bounded and integrable, it follows from the Lebesgue's dominated convergence theorem that both

$$g_{i'}(Y_{ik}) = \lim_{N\to\infty} g_{N,i'}(Y_{ik}) = \int_{Y_{ik}}^{\infty} \phi'\{H(y)\}\, dF_{i'}(y)$$

and

$$\lim_{N\to\infty} \mathrm{cov}\{g_{N,i'}(Y_{ik}),\ g_{N,u}(Y_{ik})\} = \mathrm{cov}\{g_{i'}(Y_{ik}),\ g_u(Y_{ik})\}$$

exist. It is also true that $\mathrm{cov}\{g_{i'}(Y_{ik}),\ g_u(Y_{ik})\} \geq 0$, $(x_{i'k',l} - x_{ik,l})(x_{uw,l} - x_{ik,l}) \geq 0$, since $x_{uw,l}$ is either 0 or 1. Also, none of the p_i's are zero. Therefore, result follows. □

Theorem 3.2.2. Assume that ϕ is nondecreasing and has a bounded second derivative. Then, under the above model, $(X'a_N - \mu_N)/N^{\frac{1}{2}}$ converges to a multivariate normal vector with mean vector 0 and covariance matrix (σ_{lk}) .

Lemma 3.2.3. Matrices Γ and Λ are of full rank.

Proof. Since all p_i's are rational numbers, there exist N^* and n_i^* such that $\lim_{N\to\infty} n_i^*/N^* = p_i$ for all i. Denote the corresponding design matrix by X_o. We have $\Gamma = K'\{N^*(X_o'X_o)^-\}K$. By the results found in Searle (1971), $K'\{N^*(X_o'X_o)^-\}K$ is of full rank.

Now, it will be proved that Λ is non-singular. First, define a partitioned

matrix Æ, which consists of $I \times I$ matrices. The ith diagonal one is an $n_i \times n_i$ matrix:

$$\Big(\sum_{i=1}^{I} n_i \mathrm{var}(g_{N,i}(Y_{i1})) - 2N\mathrm{cov}(g_{N,i}(Y_{i1}), \phi(H_N(Y_{ik})))\Big) J + N^2 \mathrm{var}[\phi(H_N(Y_{ik}))]\, E,$$

where E is an identity matrix, the ijth off-diagonal one is an $n_i \times n_j$ matirx :

$$\Big(n_i n_j \mathrm{cov}\Big(\sum_{u=1}^{I} g_{N,i}(Y_{u1}), \sum_{u=1}^{I} g_{N,j}(Y_{u1})\Big) - N\mathrm{cov}\Big(g_{N,i}(Y_{i1}), \phi(H_N(Y_{i1}))\Big)$$
$$- N\mathrm{cov}\Big(g_{N,j}(Y_{j1}), \phi(H_N(Y_{j1}))\Big)\Big) 1_{n_i} \cdot 1'_{n_j}.$$

The covariance matrix of $Z_{N,P}$ can be written as $X'ÆX$. The rank of Æ is N. Hence, the rank of $(\sigma_{lk,N})$, the covariance matrix of $Z_{N,P}$, is equal to Rank(X). When $N = N^*$, the rank of (σ_{lk,N^*}) equals to the rank of X_o, and $\Sigma = \frac{1}{N^*}(\sigma_{lk,N^*})$. It follows that Γ is nonsingular. □

Rewrite Q_r and T_r into:

$$Q_r = (N^{1/2}K'\hat{\beta}_r)'\{K'N(X'X)^- K\}^{-1}(N^{1/2}K'\hat{\beta}_r),$$

$$T_r = (N^{1/2}K'\hat{\beta}_r)'(N\Lambda_N)^{-1}(N^{1/2}K'\hat{\beta}_r).$$

$\lim_{N\to\infty} \mathrm{E}(Q_r)$ and $\lim_{N\to\infty} \mathrm{E}(T_r)$ depends on $\lim_{N\to\infty}\{\{(N^{1/2}K'(X'X)^- \mu_N\}$. Computations show that for every $1 \le l \le P$

$$\lim_{N\to\infty} \tfrac{1}{N}\mu_{N,l} = \sum_{i=1}^{I} p_i\, x_{i1,l} \int \phi\{H(y)\}\, dF_i(y).$$

Furthermore, $N(X'X)^-$ is finite. Thus, $\lim_{N\to\infty}(X'X)^- \mu_N$ is finite, and the expression $\lim_{N\to\infty}\{(N^{1/2}K'(X'X)^- \mu_N\}$ is a zero vector if and only if $\lim_{N\to\infty}\{K'(X'X)^- \mu_N\} = 0$.

Lemma 3.2.4. Under the above model and assumptions on the score function, it follows that $\lim_{N\to\infty} \mathrm{E}(Q_r) < \infty$ and $\lim_{N\to\infty} \mathrm{E}(T_r) < \infty$ if and only if $\lim_{N\to\infty}[K'(X'X)^- \mu_N] = 0$.

The following theorems are straightforward by the theorems of Searle (1971, P.69).

Theorem 3.2.5. Under the above model and assumptions on the score function, the statistic T_r converges in distribution to central χ^2 with S degree of freedom if and only if $\lim_{N\to\infty} \mathrm{E}\{K'(X'X)^- \mu_N\} = 0$.

Theorem 3.2.6. Under the above model and assumptions on the score function, the statistic Q_r converges in distribution to $\frac{1}{c}\chi^2(S, 0)$ if and only if i) $\lim_{N\to\infty} \{K'(X'X)^- \mu_N\} = 0$, ii) $\Gamma = c\Lambda$, where c is a positive constant.

Corollary 3.2.7. Under the above model and assumptions on the score function, the statistic $\hat{F}_r$ converges in distribution to $\frac{1}{c}\chi^2(S, 0)$ if and only if i) $\lim_{N\to\infty} \{K'(X'X)^- \mu_N\} = 0$, ii) $\lim_{N\to\infty} \Gamma/\hat{\sigma}_r^2 = c\Lambda$, where c is a positive constant.

Because both Λ and Γ are positive definite, there exists a full rank matrix D such that $\Lambda^{-1} = D'\Gamma^{-1}D$. Hence,

$$\lim_{N\to\infty} T_r = \lim_{N\to\infty} (N^{1/2}DK'\hat{\beta}_r)'\Gamma^{-1}(N^{1/2}DK'\hat{\beta}_r)$$

$$= \lim_{N\to\infty} (DK'\hat{\beta}_r)'\{K'(X'X)^- K\}^{-1}(DK'\hat{\beta}_r),$$

which resembles the form of $\lim_{N\to\infty} Q_r$. If $D = cE$, for some constant c, then Q_r is asymptotically proportional to a χ^2 random variable.

It can be seen from the above that:

1) Only those null hypotheses which imply $\lim_{N\to\infty} \mathrm{E}K'\hat{\beta}_r = 0$ can be possibly tested by the tests based on Q_r and T_r. 2) The tests based on Q_r and T_r are consistent for all the alternative of $\lim_{N\to\infty} \mathrm{E}K'\hat{\beta}_r \neq 0$. 3) If these tests reject the null, a statistical conclusion of the existence of the effects being tested can be made. If the tests fail to reject the null, there still exists the possibility that the effects being tested exist, because $\lim_{N\to\infty} \mathrm{E}K'\hat{\beta}_r = 0$ does not necessarily imply $K'\beta = 0$.

4. Application

The application of T_r and Q_r on various two-way layouts is investigated based on the above results. In each case, the model is first defined. Then, the investigation under H_o includes three steps: i) check whether $\lim_{N\to\infty} N^{1/2}K'(X'X)^{-}\mu_N = 0$; ii) check whether $\Gamma = c\Lambda$; and iii) simplify T_r. In the process, the simplification of the matrices involved in Q_r and T_r are greatly simplified by using the results in Searle (1971) and Graybill (1970).

4.1 Two-Way Nested

Assume $Y_{ijk} = \alpha_i + \beta_{ij} + \epsilon_{ijk}$, $1 \le i \le I$, $1 \le j \le b_i$, $1 \le k \le n_{ij}$. Define $p_{ij} = \lim_{N\to\infty} \frac{n_{ij}}{N}$, and N, $a_{ij\cdot}$, $\bar{a}_{i\cdot\cdot}$, and $\bar{a}$ are as defined as in usual ANOVA. Let E denote an identity matrix, $K_l' = (E_{(b_l-1)\times(b_l-1)} \quad -1_{(b_l-1)})$, and $K' = (0_{m\times I} \ \mathrm{Diag}(K_l'))$, where $m = \sum(b_i - 1)$. Further, define $[(\beta_{ij} - \beta_{ib_i})]'$, with $1 \le i \le I$, $1 \le j \le b_i - 1$, a $(\sum(b_i - 1)) \times 1$ vector. We have

$$(X'X)^{-} = \begin{bmatrix} 0 & 0 \\ 0 & \mathrm{diag}(\frac{1}{n_{ij}}) \end{bmatrix}, \text{ with } 1 \le i \le I,\ 1 \le j \le b_i.$$

Then, the null hypothesis is: $K'\beta = (\beta_{ij} - \beta_{ib_i}) = 0$. It can be interpreted into: $F_{ij} = F_i$ for all j and i. By substituting the rank score for the observations, we have $(K'\hat{\beta}_r) = [(\hat{\beta}_{ij} - \hat{\beta}_{ib_i})]'$, where $\hat{\beta}_{ij} - \hat{\beta}_{ib_i} = \bar{a}_{ij\cdot} - \bar{a}_{ib_i\cdot}$, for $1 \le i \le I$, $1 \le j \le b_i - 1$.

It is easy to find that $K'(X'X)^{-}\mu_N = 0$ under H_o even for every N. Algebraic computation shows that $\Lambda = \mathrm{Diag}[\mathrm{var}\phi\{H(Y_{ij1})\}E_{(b_i-1)\times(b_i-1)}] \times \Gamma$, where $\Gamma = \mathrm{Diag}(D_i)$, $1 \le i \le I$, $D_i = (d_{j,j'})$, $d_{j,j} = \frac{1}{n_{ij}} + \frac{1}{n_{ib_i}}$, $d_{j,j'} = \frac{1}{n_{ib_i}}$ for $j' \ne j$, $1 \le j$, $j' \le b_i - 1$. Hence, T_r converges to a central χ^2 variable with $(\sum b_i - I)$ degree of freedom under H_o, but neither Q_r nor $\hat{F}_r$ does.

Under H_o, algebraic simplification gives

$$T_r = \sum_{i=1}^{I} \frac{1}{\text{var } \phi\{H(Y_{ij})\}} \sum_{j=1}^{b_j} (a_{ij\cdot} - \bar{a}_{i\cdot\cdot})^2, \tag{4.1.1}$$

for each i, the var$\phi\{H(Y_{ij})\}$'s are equal for all $1 \le j \le b_j$. It is proved in Appendix that the following are two consistent estimates of var$\phi\{H(Y_{ij})\}$:

$$\text{i) } \frac{1}{n_{i\cdot}} \sum_{j=1}^{b_i} \sum_{k=1}^{n_{ij}} (a_{ijk} - \bar{a}_{ij\cdot})^2, \quad \text{ii) } \frac{1}{n_{i\cdot}} \sum_{j=1}^{b_i} \sum_{k=1}^{n_{ij}} (a_{ijk} - \bar{a}_{i\cdot\cdot})^2.$$

When Wilcoxon score and ii) are chosen, $\hat{T}_r$ is exactly $(\sum b_i - I)$ times the one suggested by Akritas (1990).

4.2 Two-Way Without Interaction

Assume $Y_{ijk} = \alpha_i + \gamma_j + \epsilon_{ijk}$, $1 \le i \le I$, $1 \le j \le b$, $1 \le k \le n_{ij}$, ϵ_{ijk}'s are independently and identically distributed with the cumulative distribution function F_{ij}. The null hypothesis is $\gamma_1 = \cdots = \gamma_I$. This is equivalent to $F_{ij} = F_i$ for all i and j.

Let $\beta = (\alpha_1, \cdots, \alpha_I, \gamma_1, \cdots, \gamma_b)'$, $K' = (0 \mid E \mid -1_{b-1})$, a $(b-1) \times (I+b)$ matrix. The null hypothesis can be written as: $K'\beta = [\gamma_1 - \gamma_b, \cdots, \gamma_{b-1} - \gamma_b]' = 0$. Next, let M be $I \times (b-1)$ matrix whose $(i,j)^{\text{th}}$ element is $m_{ij} = n_{ij}/n_{i\cdot}$, $1 \le i \le I$, $1 \le j \le b-1$, and let C be $(b-1) \times (b-1)$ matrix whose $(j,j')^{\text{th}}$ element is

$$c_{jj} = n_{\cdot j} - \sum_{i=1}^{I} \frac{n_{ij}^2}{n_{i\cdot}}, \quad c_{jj'} = -\sum_{i=1}^{I} \frac{n_{ij}\, n_{ij'}}{n_{i\cdot}} \quad \text{when } j' \ne j.$$

Also, let X be the design matrix. It follows from Searle (1971, p.268, p.280-281) that $\{K'(X'X)^- K\}^{-1} = C$ and $K'(X'X)^- = C^{-1}(-M' \mid E \mid 0)$. Thus, $K'\hat{\beta}_r = C^{-1}(-M' \mid E \mid 0)X'a_N$. Define the vector $\delta = (-M' \mid E \mid 0)X'a_N$. Then,

$$Q_r = (N^{-1/2}\delta')(NC)^{-1}(N^{-1/2}\delta),$$

$$T_r = (N^{-1/2}\delta')\{(-M' \mid E \mid 0)\Sigma(-M' \mid E \mid 0)\}^{-1}(N^{-1/2}\delta).$$

$(-M'|\ E\ |\ 0)\mu_N = 0$ implies $\lim_{N\to\infty}[N^{1/2}K'(X'X)^{-}\mu_N] = 0$. For checking ii), only need to check $\lim_{N\to\infty}(NC) \propto \lim_{N\to\infty}\{(-M'|\ E\ |\ 0)\Sigma(-M'|\ E\ |\ 0)\}$.

The lth element of δ is the linear rank statistic

$$\sum_{i=1}^{I}\sum_{k=1}^{n_{ij}} a_{ilk} - \sum_{i=1}^{I}\frac{n_{il}}{n_{i\cdot}}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} a_{ijk}.$$

After simplification, $(-M'|\ E\ |\ 0)\mu_N = 0$, and the diagonal element of

$$\lim_{N\to\infty}((-M'|\ E\ |\ 0)\Sigma(-M'|\ E\ |\ 0))$$

is

$$\sum_{i=1}^{I} p_{il}(1-\frac{p_{il}}{p_{i\cdot}})\mathrm{var}\phi\{H(Y_{ilk})\}\ , \tag{4.2.1}$$

the ldth off-diagonal element of $\lim_{N\to\infty}\{(-M'|\ E\ |\ 0)\Psi(-M'|\ E\ |\ 0)\}$ is

$$-\sum_{i=1}^{I} p_{il}\frac{p_{il}}{p_{i\cdot}}\ \mathrm{var}\phi\{H(Y_{idk})\}, \tag{4.2.2}$$

compared to

$$\lim_{N\to\infty} c_{jj} = p_{\cdot j} - \sum_{i=1}^{I}\frac{p_{ij}^2}{p_{i\cdot}}\ , \tag{4.2.3}$$

$$\lim_{N\to\infty} c_{jd} = -\sum_{i=1}^{I}\frac{p_{ij}p_{id}}{p_{i\cdot}}. \tag{4.2.4}$$

Hence, for a general sampling scheme, $\hat{T}_r = \delta'(\hat{\sigma}_{ll'})\delta \xrightarrow{\mathrm{P}} \chi^2_{(b-1)}$, where

$$\hat{\sigma}_{ll} = \sum_{i=1}^{I}\frac{n_{il}}{n_{i\cdot}}(1-\frac{n_{il}}{n_{i\cdot}})\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(a_{ijk}-\bar{a}_{i\cdot\cdot})^2,$$

$$\hat{\sigma}_{ld} = -\sum_{i=1}^{I}\frac{n_{il}n_{id}}{n_{i\cdot}^2}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(a_{ijk}-\bar{a}_{i\cdot\cdot})^2,\ d\neq l,$$

or

$$\hat{\sigma}_{ll} = \sum_{i=1}^{I}\frac{n_{il}}{n_{i\cdot}}(1-\frac{n_{il}}{n_{i\cdot}})\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(a_{ijk}-\bar{a}_{ij\cdot})^2,$$

$$\hat{\sigma}_{ld} = -\sum_{i=1}^{I}\frac{n_{il}n_{id}}{n_{i\cdot}^2}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(a_{ijk}-\bar{a}_{ij\cdot})^2,$$

but the rank transform is not.

Under the assumption of proportional sampling, namely $p_{ij} = p_{i\cdot}p_{\cdot j}$, (4.2.1) and (4.2.2) can be simplified to

$$p_{\cdot l}(1-p_{\cdot l})\sum_{i=1}^{I} p_{i\cdot}\mathrm{var}\phi(H(Y_{ilk})),$$

$$p_{\cdot l}\, p_{\cdot d}\sum_{i=1}^{I} p_i\mathrm{var}\phi\{H(Y_{ilk})\},$$

(4.2.3) and (4.2.4) to $p_{\cdot l}(1-p_{\cdot l})$ and $-p_{\cdot l}p_{\cdot d}$. Under the null hypothesis, by the result in Appendix, two consistent estimates of $\mathrm{var}\{\phi(H(Y_{ijk})\}$ are:

$$\text{i) } \frac{1}{n_{i\cdot}}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(a_{ijk}-\bar{a}_{i\cdot\cdot})^2, \text{ ii) } \frac{1}{n_{i\cdot}}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(a_{ijk}-\bar{a}_{ij\cdot})^2.$$

Hence

$$\hat{T}_{r_1}=\frac{N\sum_{j=1}^{b} n_{\cdot j}(\bar{a}_{\cdot j\cdot}-\bar{a}_{\cdots})^2}{\sum_{i=1}^{I}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(a_{ijk}-\bar{a}_{i\cdot\cdot})^2}, \text{ or } \hat{T}_{r_2}=\frac{N\sum_{j=1}^{b} n_{\cdot j}(\bar{a}_{\cdot j\cdot}-\bar{a}_{\cdots})^2}{\sum_{i=1}^{I}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(a_{ijk}-\bar{a}_{ij\cdot})^2},$$

and the rank transform test

$$\hat{F}_r=\frac{(N-I-b+1)\sum_{j=1}^{b} n_{\cdot j}(\bar{a}_{\cdot j\cdot}-\bar{a}_{\cdots})^2}{(b-1)\sum_{i=1}^{I}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(a_{ijk}-\bar{a}_{i\cdot\cdot}-\bar{a}_{\cdot j\cdot}+\bar{a}_{\cdots})^2}$$

converges to a $\chi^2_{(b-1)}/(b-1)$ variable, because

$$\lim_{N\to\infty}\frac{\sum_{i=1}^{I}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(a_{ijk}-\bar{a}_{i\cdot\cdot}-\bar{a}_{\cdot j\cdot}+\bar{a}_{\cdots})^2}{N-I-b+1}$$

converges in probability to $\sum_{i=1}^{I} p_i\mathrm{var}\phi\{H(Y_{ilk})\}$.

4.3 Two-Way With Interaction

Assume $Y_{ijk}=\alpha_i+\beta_j+\gamma_{ij}+\epsilon_{ijk}$, $1\le i\le I$, $1\le j\le b$, $1\le k\le n_{ij}$. Let $\theta_{ij,i'j'}=\gamma_{ij}-\gamma_{ij'}-\gamma_{i'j}+\gamma_{i'j'}$. By Searle (1987, p.97), there are $Ib(I-1)(b-1)/4$ different θ's. The maximum number of LIN θ's is $(I-1)(b-1)$. Let θ be a vector, of which the elements are a set of $(I-1)(b-1)$ LIN θ's. The null hypothesis of no

interaction can be expressed as $\theta = 0$, that is $K'\beta = \theta = 0$. Denote $\theta^r_{ij,i'j'} = \bar{a}_{ij\cdot} - \bar{a}_{ij'\cdot} - \bar{a}_{i'j\cdot} + \bar{a}_{i'j'\cdot}$, we use $K'\hat{\beta}_r = \theta^r$ in the rank tests discussed here.

First, verify whether $E\theta^r$ converges in probability to 0 under H_o. The asymptotic mean of $\theta^r_{ij,i'j'}$ converges in probability to

$$\int \phi\{H(y)\}\,dF_{ij}(y) - \int \phi\{H(y)\}\,dF_{ij'}(y) - \int \phi\{H(y)\}\,dF_{i'j}(y) + \int \phi\{H(y)\}\,dF_{i'j'}(y), \tag{4.3.1}$$

which is not necessarily equal to zero.

When the design is balanced and when $\phi\{H(y)\} = H(y) = \sum\sum p_{ij}F_{ij}(y)$, which is the Wilcoxon score, then, by Thompson (1991b), (4.3.1) is zero when either there is only one main effect, or when $I = b = 2$ with the presence of both main effects. Furthermore, in the latter case, only when the underlying distributions have common support, which implies that

$$\int F(y+\alpha-\beta)dF(y+\alpha+\beta) = \int F(y-2\beta)dF(y), \tag{4.3.2}$$

$$\int F(y+\alpha+\beta)dF(y+\alpha-\beta) = \int F(y+2\beta)dF(y), \tag{4.3.3}$$

$$\int \{F(y+\delta)dF + F(y-\delta)\}\,dF = c. \tag{4.3.4}$$

Here only the case with the presence of both main effects is discussed.

Similar to Thompson (1991b), expression (4.3.1) becomes

$$\begin{aligned}
&\tfrac{1}{2}(p_{11}-p_{12}-p_{21}+p_{22}) + (p_{12}-p_{22})\int F(y-2\beta)dF(y) + (p_{21}-p_{22})\int F(y-2\alpha)dF \\
&+ p_{22}\int F(y-2\alpha-2\beta)dF + p_{11}\int F(y+2\alpha+2\beta)dF(y) - (p_{11}-p_{12})\int F(y+2\alpha)dF \\
&-(p_{11}-p_{21})\int F(y+2\beta)dF - p_{21}\int F(y-2\alpha+2\beta)dF - p_{12}\int F(y+2\alpha-2\beta)dF.
\end{aligned} \tag{4.3.6}$$

When $\int \{F(y+\delta)+F(y-\delta)\}\,dF$ is a constant c, using the relationship of

$\int F(y-\delta)dF = c - \int F(y+\delta)dF$, equation (4.3.6) becomes

$$\tfrac{1}{2}(p_{11}-p_{12}-p_{21}+p_{22}) + c(p_{12}-p_{22}) - (p_{21}-p_{11}-p_{12}+p_{22})\int F(y+2\beta)dF$$

$$-(p_{12}-p_{11}-p_{21}+p_{22})\int F(y+2\alpha)dF - (p_{22}-p_{11})\int F(y+2\alpha+2\beta)dF(y)$$

$$-(p_{21}-p_{12})\int F(y+2\alpha-2\beta)dF. \tag{4.3.7}$$

Equation (4.3.7) is a function of α, β and c, and it is constantly zero, regardless to α, β, c and the F that satisfies (4.3.2), (4.3.3) and (4.3.4), if and only if all p_{ij}'s are equal, where $i=1,\ 2$ and $j=1,\ 2$. That is, if and only if the design is balanced. Counter example: when $c=\frac{1}{2}$, $p_{11}=p_{22}$, $p_{12}=p_{21}$, but $p_{11}\neq p_{21}$, (4.3.7) equals $\frac{1}{2}(p_{11}-p_{12}-p_{21}+p_{22})+\frac{1}{2}(p_{12}-p_{22})\neq 0$.

The asymptotic variance of $(\bar{R}_{11\cdot}-\bar{R}_{12\cdot}-\bar{R}_{21\cdot}+\bar{R}_{22\cdot})$ is $\mathrm{var}(\sum\sum\sum Z_{ijk})$. With the Wilcoxon score, it can be simplified to

$$\begin{aligned}
\mathrm{var}(Z_{11k}) &= 4n^{-2}\ \mathrm{var}\{F_{12}(Y_{111})+F_{21}(Y_{111})\},\\
\mathrm{var}(Z_{12k}) &= 4n^{-2}\ \mathrm{var}\{F_{11}(Y_{121})+F_{22}(Y_{121})\},\\
\mathrm{var}(Z_{21k}) &= 4n^{-2}\ \mathrm{var}\{F_{11}(Y_{211})+F_{22}(Y_{211})\},\\
\mathrm{var}(Z_{22k}) &= 4n^{-2}\ \mathrm{var}\{F_{12}(Y_{221})+F_{21}(Y_{221})\}.
\end{aligned}$$

Let $\hat{\sigma}^2$ be a consistent estimate of $\mathrm{var}(\sum\sum\sum Z_{ijk})$. Further, let $R_{uvk,(uv,ij,i'j')}$ be the rank of Y_{uvk} among Y_{ijw}'s, $Y_{i'j'w}$'s and Y_{uvw}'s, $1\le w\le n$, $R_{uvk,(uv)}$ be the rank of Y_{uvk} among Y_{uvw}'s, $1\le w\le n$, and

$$R^*_{uvk} = R_{uvk,(uv,ij,i'j')} - R_{uvk,(uv)},\ \bar{R}^*_{uv} = \frac{1}{n}\sum_{k=1}^{n} R^*_{uvk}.$$

It can be shown by the way similar to the proof in the appendix that

$$\frac{1}{n^3}\sum(R^*_{11k}-\bar{R}^*_{11})^2 \xrightarrow{P} \mathrm{var}\{F_{12}(Y_{111})+F_{21}(Y_{111})\},$$

$$\hat{\sigma}^2 = \frac{1}{4n^4}\sum\{(R^*_{11k} - \bar{R}^*_{11})^2 + (R^*_{12k} - \bar{R}^*_{12})^2 + (R^*_{21} - \bar{R}^*_{21})^2 + (R^*_{22k} - \bar{R}^*_{22})^2\},$$

$$\hat{\sigma}_r^2 \xrightarrow{P} \frac{1}{4}\sum\sum \mathrm{var}[\phi(H(Y_{ij1}))] .$$

Thus,

$$T_r = \frac{(\bar{R}_{11\cdot} - \bar{R}_{12\cdot} - \bar{R}_{21\cdot} + \bar{R}_{22\cdot})^2}{\sigma^2} \xrightarrow{P} \chi_1^2,$$

$$\frac{\hat{\sigma}_r^2}{4n^2\sigma^2}\hat{F}_r \xrightarrow{P} \chi_1^2.$$

It can be seen that neither Q_r nor $\hat{T}_r$ is distribution-free in the case with the presence of two main effects for testing for interaction.

APPENDIX

Consistent Estimates of the Unknown Parameter

First define $\rho_{ijk,N} = (N+1)^{-1}R_{ijk}$, $\bar{\phi}_{ij,N} = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\phi(\rho_{ijk,N})$, where $\phi(\rho_{ijk,N})$ is the score function, and $\bar{\phi}_H = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}\phi\{H(Y_{ijk})\}$.

Theorem 4.1 Under the conditions set forth in the previous section, it follows that for all $1 \le i \le P$, $\frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}}(\phi(\rho_{ijk,N}) - \bar{\phi}_{i,N})^2$ converges in probability to $\mathrm{var}\phi\{H(Y_{ij1})\}$.

Proof. Let n_{ij} be denoted by n. Theorem 4.1 will follow if

$$\lim_{\substack{N\to\infty \\ n\to\infty}} P\{|\tfrac{1}{n}\sum_{k=1}^{n_i}(\phi(\rho_{ijk,N}) - \bar{\phi}_{ij,N})^2 - \tfrac{1}{n}\sum_{k=1}^{n}[\phi\{H(Y_{ijk})\} - \bar{\phi}_H]^2| > \epsilon\} = 0 \qquad (1)$$

holds for every $\epsilon > 0$. By some algebraic simplification and Triangle-Inequality, (1) can be simplified to

$$\lim_{\substack{N\to\infty \\ n\to\infty}} P\left\{\tfrac{1}{n}\sum_{k=1}^{n}|\phi(\rho_{ijk,N}) - \phi\{H(Y_{ijk})\}| > \epsilon\right\} = 0,\ \epsilon > 0. \qquad (2)$$

Equation (2) will follow if

$$\lim_{\substack{N\to\infty \\ n\to\infty}} E\left(\tfrac{1}{n}\sum_{k=1}^{n}|\phi(\rho_{ijk,N}) - \phi\{H(Y_{ijk})\}|\right) = 0. \qquad (3)$$

LEMMER, H. H. (1964). Verdelingsvrye toetsingsmetodes vir die probleem van twee of eer steekproewe. *D.Sc. thesis*, University of Pretoria.

LEMMER, H. H. & STOKER, D. J. (1967). A distribution-free analysis of variance for the two-way classification. *South African Statistical Journal*, 1, 67-74.

LEMMER, H. H., STOKER, D. J. & REINACH, S. G. (1968). A distribution-free analysis of variance technique for block designs. *South African Statistical Journal*, 2, 9-32

LEMMER, H. H. (1980). Some empirical results on the two-way analysis of variance by ranks. Comm. Statist. A 9, 1427-1438.

PURI, M. L. & SEN, P. K. (1985). *Nonparametric methods in general linear models*. New York: Wiley.

SEARLE, S. R. (1971). *Linear Models*. New York: Wiley.

SEARLE, S. R. (1987). *Linear Models for Unbalanced Data*. New York: Wiley.

SEARLE, S. R. (1983). *Matrix Algebra Useful for Statistics*. New York: Wiley.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

STRANG, GILBERT (1976). *Linear Algebra and Its Applications*. New York: Academic Press.

THOMPSON, G. L. & AMMANN, L. P. (1989). Efficacies of rank-transform statistics in two-way models with no interaction. *J. Am. Statist. Assoc.* 84, 325-330.

THOMPSON, G. L. & AMMANN L. P. (1990). Efficiencies of interblock rank statistics for repeated measures designs. *J. Am. Statist. Assoc.* 85, 519-528.

THOMPSON, G. L. (1991a). A unified approach to rank tests for multivariate and repeated measures design. *J. Am. Statist. Assoc.* 86, 410-419.

THOMPSON, G. L. (1991b). A note on the rank transform for interactions. *Biometrika*, 78,697-701.