Chapter 4
Testing for Randomness, Agreement, and Interaction

Suppose that n judges are asked to rank t contestants in accordance with some predetermined criterion. One immediate question that comes to mind is: are the judges ranking the contestants by selecting a ranking at random or is there some specific pattern for their choices? Placing this problem in a geometric setting, we may represent each ranking as a point in a t-dimensional space. If indeed the judges act in accordance with some specific nonrandom manner, the points would tend to cluster close together in one or more groups. Intuitively then, a test of randomness could be based on the average pairwise distance between points with large values of that statistic displaying evidence of the random pattern of the points.

In the literature, the Kendall W has been a widely used statistic whose asymptotic distribution was derived by Friedman (1937). Treating each judge as a block, it consists of calculating for each object the average of the ranks assigned by the judges and computing the variance of the averages. Small values of the test statistic are considered consistent with the null hypothesis of randomness. This test statistic is not always sensitive to patterns that may exist in the data. For example, if half the judges assign rankings in the natural order, $1, 2, \ldots, t$ and the other half assign rankings in the reverse order, $t, t-1, \ldots, 1$, then the value of the Kendall W statistic will be small and the null hypothesis will not be rejected. Such considerations lead one to inquire as to whether or not there are other test statistics with better performance.

4.1 Tests for Randomness

We begin with some notation. Let $\mathcal{P} = \{v_j\}$ be the set of $t!$ possible rankings of t objects and denote the rankings by

$$v_j = \left(v_j(1), \ldots, v_j(t)\right)', \; j = 1, \ldots, t!$$

M. Alvo, P.L.H. Yu, *Statistical Methods for Ranking Data*, Frontiers in Probability and the Statistical Sciences, DOI 10.1007/978-1-4939-1471-5_4

Suppose that we have a random sample of n rankings denoted by $R, \ldots, R_n$ observed from some population of rankers and suppose that each judge chooses a ranking in accordance with some distribution $\boldsymbol{p}$,

$$\boldsymbol{p} = (p_1, \ldots, p_{t!})'$$

where

$$p_j = P\left(R = v_j\right).$$

Our interest is in developing a test of the null hypothesis of randomness, namely

$$H_0 : \boldsymbol{p} = \boldsymbol{p}_0 = \mathbf{1}/t!$$

against the alternative

$$H_1 : \boldsymbol{p} \neq \boldsymbol{p}_0.$$

The null hypothesis indicates that each judge chooses a ranking at random from the population of possible rankings. Select a distance function, $d\left(R_k, R_l\right)$, between two rankings, R_k, R_l. A possible test statistic for testing the null hypothesis consists of computing the average pairwise distance between all the observed rankings

$$\bar{d}_n = \frac{1}{n\left(n-1\right)} \Sigma \Sigma_{k,l} d\left(R_k, R_l\right). \tag{4.1}$$

Under the null hypothesis, one would expect the average pairwise distance to be large or, equivalently from (4.1), the average pairwise correlation

$$\bar{\alpha}_n = 1 - \frac{2\bar{d}_n}{M} \tag{4.2}$$

to be small. Equivalently, one should reject the null hypothesis whenever $\bar{\alpha}_n$ is large. Note that we may write

$$d\left(R_k, R_l\right) = \Sigma_i \Sigma_j d\left(v_i, v_j\right) I\left[R_k = v_i\right] I\left[R_l = v_i\right]$$

where $I[B]$ is the indicator function taking value 1 if the event B is true and 0 otherwise. It follows that

$$\begin{aligned}
n\left(n-1\right)\bar{d}_n &= \Sigma_k \Sigma_l d\left(R_k, R_l\right) && (4.3)\\
&= \Sigma_k \Sigma_l \Sigma_i \Sigma_j d\left(v_i, v_j\right) I\left[R_k = v_i\right] I\left[R_l = v_j\right] && (4.4)\\
&= \Sigma_i \Sigma_j \left(\Sigma_k I\left[R_k = v_i\right]\right)\left(\Sigma_l I\left[R_l = v_j\right]\right) d\left(v_i, v_j\right) && (4.5)\\
&= \Sigma_i \Sigma_j N_i N_j d\left(v_i, v_j\right) && (4.6)\\
&= N' \Delta N && (4.7)
\end{aligned}$$

where $\Delta = \left(d\left(v_i, v_j\right)\right)$ is the matrix of pairwise distances and $N' = (N_1, \ldots, N_{t!})$ is the vector of frequencies with

$$N_i = \Sigma_k I\left[R_k = v_i\right].$$

We recognize that (4.7) is a quadratic form and that N is a multinomial random variable with mean and covariance respectively given by

$$EN = n\boldsymbol{p}, Cov(N) = n\sum,$$

where $\sum = (diag(\boldsymbol{p}) - \boldsymbol{pp'})$ and $diag(\boldsymbol{p})$ is a $t! \times t!$ diagonal matrix having entries p_i along the diagonal. Let $\hat{\boldsymbol{p}}_n = N/n$ and recall from (3.13)

$$Q = J - \left(\frac{2}{M}\right)\Delta.$$

Theorem 4.1. *(a) Under H_0, for $n \to \infty$, and $Q\boldsymbol{p}_0 = c^*\mathbf{1}$, we have that*

$$(n-1)\left(\bar{\alpha}_n - c^*\right) \Rightarrow_{\mathcal{L}} Z_0' Q Z_0 - 1 + c^*$$

where Z has a $t!$-variate normal distribution with mean 0 and covariance matrix $\Sigma_0 = (t!)^{-2}\left((t!I - J\right)$. Here I is the identity matrix and J is a $t! \times t!$ matrix of ones.

(b) Under H_1, for $n \to \infty$,

$$\sqrt{n}\left(\overline{\alpha}_n - \boldsymbol{p}'Q\boldsymbol{p}\right) \Rightarrow_{\mathcal{L}} 2Z'Q\boldsymbol{p}$$

where Z has a $t!$-variate normal distribution with mean 0 and covariance matrix

$$\sum = \left(diag(\boldsymbol{p}) - \boldsymbol{pp'}\right).$$

Proof. (a) Define

$$Z_n = n^{-1/2}\left(N - n\boldsymbol{p}\right).$$

A Taylor series expansion around $\hat{\boldsymbol{p}}_n = \boldsymbol{p}$ reveals the identity

$$\begin{aligned}\hat{\boldsymbol{p}}_n' Q \hat{\boldsymbol{p}}_n - \boldsymbol{p}'Q\boldsymbol{p} &= 2\left(\hat{\boldsymbol{p}}_n - \boldsymbol{p}\right)' Q\boldsymbol{p} + \left(\hat{\boldsymbol{p}}_n - \boldsymbol{p}\right)' Q\left(\hat{\boldsymbol{p}}_n - \boldsymbol{p}\right) \\ &= \frac{2}{\sqrt{n}} Z_n' Q\boldsymbol{p} + \frac{1}{n} Z_n' Q Z_n.\end{aligned} \tag{4.8}$$

Under the null hypothesis $\boldsymbol{p} = \boldsymbol{p}_0$ and $Q\boldsymbol{p}_0 = c^*\mathbf{1}$, so that $\boldsymbol{p}_0' Q \boldsymbol{p}_0 = c^*$. Now the relationships

$$Q = J - \frac{2}{M}\Delta$$
$$\overline{\alpha}_n = 1 - \binom{n}{2}^{-1} \left(N'\Delta N\right)/M$$

imply

$$(n-1)\,\bar{\alpha}_n + 1 = \left(n\,\hat{\boldsymbol{p}}_n' Q\,\hat{\boldsymbol{p}}_n\right)$$

On using the multivariate central limit theorem for multinomial random variables (Timm 1975), it follows from (4.8)

$$\begin{aligned}(n-1)\left(\overline{\alpha}_n - c^*\right) + 1 - c^* &= n\left(\hat{\boldsymbol{p}}_n' Q\,\hat{\boldsymbol{p}}_n - \boldsymbol{p}_0' Q\,\boldsymbol{p}_0\right)\\ &= \quad Z_n' Q Z_n\\ &\Rightarrow \quad {}_{\mathcal{L}}Z' Q Z.\end{aligned}$$

(b) On the other hand if $\boldsymbol{p} \neq \boldsymbol{p}_0$ we have from (4.8)

$$\sqrt{n}\left(\hat{\boldsymbol{p}}_n' Q\,\hat{\boldsymbol{p}}_n - \boldsymbol{p}' Q\,\boldsymbol{p}\right) \Rightarrow_{\mathcal{L}} 2Z' Q\,\boldsymbol{p}$$
$$(n-1)\,\overline{\alpha}_n = -1 + n\,\hat{\boldsymbol{p}}_n' Q\,\hat{\boldsymbol{p}}_n$$

and it follows that

$$\sqrt{n}\left(\overline{\alpha}_n - \boldsymbol{p}' Q\,\boldsymbol{p}\right) \Rightarrow_{\mathcal{L}} 2Z' Q\,\boldsymbol{p}.$$

□

The distribution of $Z'QZ$ under the null hypothesis is that of a weighted chi square where the weights are given by the eigenvalues of the matrix

$$Q\Sigma_0 = (t!)^{-1}\left(Q - c^* J\right).$$

In what follows, we shall obtain properties of that matrix for both the Spearman and Kendall cases. For these cases, the constant $c^* = 0$ and hence

$$Q\Sigma_0 = (t!)^{-1} Q.$$

Before dealing with the specific distributions of the Spearman and Kendall statistics, we will need the following lemmas which are useful in their own right.

Lemma 4.1. *Let* $A(s,s',t,t') = \Sigma_\nu sgn(\nu(s)-\nu(t))\, sgn(\nu(s')-\nu(t'))$. *Then, under* H_0

$$A\left(s,s',t,t'\right) = \begin{cases} 0 & s \neq s', t \neq t', \\ t! & s = s', t = t', \\ \frac{t!}{3} & s = s', t \neq t', \\ -\frac{t!}{3} & s = t', s' \neq t. \end{cases} \tag{4.9}$$

Proof. Let R be a random ranking of t objects. Note that in distribution

$$sgn[R(s)-R(t)] =_d sgn[U-V]$$

where U, V are independent uniform random variables on $(0,1)$.

Let $Z = sgn[U_1-V_1]\, sgn[U_2-V_2]$ where U_1, U_2, V_1, V_2 are independent uniform random variables on $(0,1)$. It follows that

$$\begin{aligned} P(Z>0) &= P(U_1-V_1>0, U_2-V_2>0) + P(U_1-V_1<0, U_2-V_2<0) \\ &= P(U_1-V_1>0)\, P(U_2-V_2>0) + P(U_1-V_1<0)\, P(U_2-V_2<0) \\ &= \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{2}. \end{aligned}$$

Similarly, $P(Z<0) = \frac{1}{2}$. Also, if $Z_1 = sgn[U_1-V_1]\, sgn[U_1-V_2]$, then

$$\begin{aligned} P(Z_1>0) &= P(U_1-V_1>0, U_1-V_2>0) + P(U_1-V_1<0, U_1-V_2<0) \\ &= \int_0^1 P(x-V_1>0, x-V_2>0)\, dx + \int_0^1 P(x-V_1<0, x-V_2<0)\, dx \\ &= \int_0^1 P(x-V_1>0)\, P(x-V_2>0)\, dx + \int_0^1 P(x-V_1<0)\, P(x-V_2<0)\, dx \\ &= \int_0^1 x^2 dx + \int_0^1 (1-x)^2\, dx = \frac{2}{3}. \end{aligned}$$

It now follows that

$$\begin{aligned} \frac{1}{t!}\Sigma_\nu sgn(\nu(s)-\nu(t))\, sgn(\nu(s)-\nu(t')) &= E[sgn(R(s)-R(t))\, sgn(R(s)-R(t'))] \\ &= P(Z_1>0) - P(Z_1<0) = \frac{1}{3}. \end{aligned}$$

The other cases follow in a similar way. □

Lemma 4.2. *The matrices Q_S, Q_K satisfy*

(i) $Q_S^2 = \frac{t!}{t-1} Q_S$ *and hence,* $\frac{t-1}{t!} Q_S$ *is idempotent.*
(ii) $Q_K Q_S = \frac{2t!(t+1)}{3t(t-1)} Q_S$.
(iii) $Q_K = \frac{2(t+1)}{3t} Q_S + A, Q_S A = 0$.
(iv) $Q_K^2 = \frac{4t!(t+1)^2}{9t^2(t-1)} Q_S + \frac{2t!}{3t(t-1)} A$.

Proof. Setting $c_S = \frac{t\left(t^2-1\right)}{12}$ the matrix

$$c_S Q_S = \mathbf{T}_S' \mathbf{T}_S = (t-2)! c_S \left[tI - J\right].$$

In fact, the diagonal elements are equal to

$$(t-1)! \sum \left(i - \frac{t+1}{2}\right)^2 = (t-1)! c_S,$$

whereas the off-diagonal elements are equal to

$$(t-2)! \sum_{i \neq j} \left(i - \frac{t+1}{2}\right)\left(j - \frac{t+1}{2}\right) = -(t-2)! c_S.$$

The matrix Q_S is singular since the rows sum to 0. A generalized inverse of $c_S Q_S$ is given by $\frac{1}{(t-2)! t c_S} [I + J]$.

Now to show idempotency in (i), we see that

$$\begin{aligned}
(Q_S)^2 = \tfrac{1}{c_S^2} \mathbf{T}_S' \mathbf{T}_S \mathbf{T}_S' \mathbf{T}_S & \\
&= \frac{1}{c_S} (t-2)! \mathbf{T}_S' \left[t I_t - J\right] \mathbf{T}_S \\
&= \frac{1}{c_S} (t-2)! \left[t \mathbf{T}_S' \mathbf{T}_S\right] \\
&= \frac{t!}{t-1} (Q_S).
\end{aligned}$$

Next we prove (ii). We first note that the ith rank can be represented in terms of the remaining $(t-1)$ ranks as

$$\left[\nu(i) - \frac{t+1}{2}\right] = \frac{1}{2} \Sigma_l sgn\left[\nu(i) - \nu(l)\right]. \tag{4.10}$$

Part (ii) is equivalent to showing

$$\left(\mathbf{T}_K' \mathbf{T}_K\right)\left(\mathbf{T}_S' \mathbf{T}_S\right) = \frac{1}{c_K} \frac{t!(t+1)}{3} \left(\mathbf{T}_S' \mathbf{T}_S\right)$$

where $c_K = \frac{t(t-1)}{2}$.

For simplicity, note that the first row and column entry in the matrix $\left(\mathbf{T}_K'\mathbf{T}_K\right)\mathbf{T}_S'$ is given by

$$\Sigma_{h=1}^{t!}\Sigma_{i<j}sgn\left[\nu_1(j)-\nu_1(i)\right]sgn\left[\nu_h(j)-\nu_h(i)\right]\left[\nu_h(1)-\frac{t+1}{2}\right]=$$

$$\frac{1}{2}\Sigma_{l=2}^{t}\Sigma_{i<j}sgn\left[\nu_1(j)-\nu_1(i)\right]\Sigma_{h=1}^{t!}sgn\left[\nu_h(j)-\nu_h(i)\right]sgn\left[\nu_h(1)-\nu_h(l)\right].$$

There are two cases to consider, namely $(i=1, j=l)$ and $(i\neq 1, j=l)$. It follows that

$$\begin{aligned}
-\frac{1}{2}\left\{t!\Sigma_{l=2}^{t}sgn\left[\nu_1(l)-\nu_1(1)\right]+\frac{t!}{3}\Sigma_{i\neq 1,j}\Sigma_{l\neq j}sgn\left[\nu_h(j)-\nu_h(l)\right]\right\} &= \\
\left\{\left[t!\left(\nu_1(1)-\frac{t+1}{2}\right)\right]-\frac{t!}{3}(t-2)\,\Sigma_{j\neq 1}\left(\nu_1(l)-\frac{t+1}{2}\right)\right\} &= \\
\left\{t!\left(\nu_1(1)-\frac{t+1}{2}\right)+\frac{t!}{3}(t-2)\left(\nu_1(1)-\frac{t+1}{2}\right)\right\} &= \\
\frac{(t+1)}{3}t!\left(\nu_1(1)-\frac{t+1}{2}\right).&
\end{aligned}$$

Other entries are treated similarly. Part (iii) follows directly from (ii).

To show part (iv) it suffices to show

$$\left(\mathbf{T}_K'\mathbf{T}_K\right)^2=\frac{t!}{3}\left(\mathbf{T}_K'\mathbf{T}_K\right)+\frac{4t!}{3}\left(\mathbf{T}_S'\mathbf{T}_S\right). \tag{4.11}$$

In fact, this follows since the rs term of the left-hand side of (4.11) is equal to

$$\begin{aligned}
&\sum_{k<l}\sum_{k'<l'}sgn\left[\nu_r(k)-\nu_r(l)\right]sgn\left[\nu_s\left(k'\right)-\nu_s\left(l'\right)\right] \\
&\quad\times\sum_{h=1}^{t!}sgn\left[\nu_h(k)-\nu_h(l)\right]sgn\left[\nu_h\left(k'\right)-\nu_h\left(l'\right)\right]= \\
&\quad\frac{t!}{3}\sum_{k<l}\sum_{k'<l'}sgn\left[\nu_r(k)-\nu_r(l)\right]sgn\left[\nu_s\left(k'\right)-\nu_s\left(l'\right)\right] \\
&\quad+\frac{t!}{3}\sum_{k<l}\sum_{k<l'}sgn\left[\nu_r(k)-\nu_r(l)\right]sgn\left[\nu_s(k)-\nu_s\left(l'\right)\right] \\
&=\frac{t!}{3}\sum_{k<l}\sum_{k'<l'}sgn\left[\nu_r(k)-\nu_r(l)\right]sgn\left[\nu_s\left(k'\right)-\nu_s\left(l'\right)\right] \\
&\quad+4\frac{t!}{3}\sum_{k}\left(\nu_r(k)-\frac{t+1}{2}\right)\left(\nu_s(k)-\frac{t+1}{2}\right).
\end{aligned}$$

□

Theorem 4.2. *The asymptotic distribution of the Spearman statistic under the null hypothesis of randomness is given by*

$$(t-1)\left\{(n-1)\,\overline{\rho}_n + 1\right\} \Rightarrow_L \chi^2_{t-1}. \tag{4.12}$$

The left-hand side of (4.12) can also be expressed as

$$\frac{12n}{t\,(t+1)} \sum_{i=1}^{t} \left(\bar{R}_i - \frac{t+1}{2}\right)^2$$

which is the usual Friedman statistic.

Proof. The asymptotic distribution of $(t-1)\left\{(n-1)\,\overline{\rho}_n + 1\right\}$ is that of a weighted χ^2 where the weights are determined by the eigenvalues of the idempotent matrix $\frac{(t-1)}{t!}\,Q_S$. Hence its eigenvalues are 0 or 1. Moreover, the rank of the matrix is $(t-1)$.

The left-hand side of (4.12) is equal to

$$\begin{aligned}(t-1)\,Z'QZ &= \frac{t-1}{c_S} n \parallel \mathbf{T}\hat{\mathbf{p}}_n \parallel^2 \\ &= \frac{12n\,(t-1)}{t\,(t^2-1)} \sum_{i=1}^{t} \left(\bar{R}_i - \frac{t+1}{2}\right)^2.\end{aligned}$$

□

Theorem 4.3. *The asymptotic distribution of the Kendall statistic under the null hypothesis of randomness is given by*

$$(n-1)\,\bar{\tau}_n + 1 \Rightarrow_L \frac{2}{3t\,(t-1)} \left\{(t+1)\,\chi^2_{t-1} + \chi^2_{\binom{t-1}{2}}\right\} - 1. \tag{4.13}$$

The left-hand side of (4.13) can also be expressed as

$$\frac{\sum (2x_i - n)^2}{n\binom{t}{2}}$$

where the summation is taken over all $\binom{t}{2}$ pairs of objects and x_i is the number of judges whose ranking of the pair i of objects agrees with the ordering of the same pair in a criterion ranking such as the natural ordering.

Proof. From Lemma 4.2,

$$A = Q_K - \frac{2\,(t+1)}{3t} Q_S$$

and it follows that

$$A^2 = \frac{2t!}{3t\,(t-1)}A.$$

This implies $\frac{3t(t-1)}{2t!}A$ is an idempotent matrix. Noting that

$$Trace\,(A) = \frac{t!\,(t-2)}{3t} = rank\,(A)\,,$$

we see that Q_K has two distinct nonzero eigenvalues,

$$\lambda_1 = \frac{2t!\,(t+1)}{3t\,(t-1)}, \lambda_2 = \frac{2t!}{3t\,(t-1)}$$

and (4.13) follows.

Now, let $a_{ij} = 1$, if judge j agrees with the ranking in pair i and $= -1$ if he disagrees. Then, setting $a_i = \sum_j a_{ij}$ and noting that if x_i =number of judges who agree with the ranking in pair i and y_i =number who disagree, we have

$$x_i + y_i = n, x_i - y_i = a_i\,,$$

then $a_i = 2x_i - n$. The left-hand side of (4.13) is equal to

$$\frac{\sum (a_i)^2}{n\binom{t}{2}}$$

and the result follows. □

The preceding theorems did not consider the situation where ties are possible in the rankings. This situation was considered in the literature for the case of the Spearman statistic (Lehmann 1975) wherein the asymptotic distribution is obtained by conditioning on the observed ties. Consider the following example where it may not be desirable to condition on the observed ties only. Suppose that tasters are asked to rank in order of preference each of three varieties of tea. If ties are permitted, the sample space would consist of all possible permutations, including those where either two or all three varieties are tied. Alvo and Cabilio (1985) derived the correction for ties under precisely such situations. This correction for ties is made once and for all. This approach allows for comparisons to be made when the same experiment is repeated. We recall for completeness the definition of a tied ordering, previously given in Chapter 3.

Definition 4.1. A tied ordering of n objects is a partition into e sets, $1 \leq e \leq t$, each of which contains d_i objects, $d_1 + d_2 + \ldots + d_e = t$, so that the d_i objects in each set share the rank i, $1 \leq i \leq e$. Such a tie pattern is denoted by $\boldsymbol{\delta} = (d_1, d_2, \ldots, d_e)$. The ranking denoted by $\boldsymbol{v}_\delta = (v_\delta(1), v_\delta(2), \ldots, v_\delta(t))$, resulting from such an ordering, is a tied ranking and is one of $t!/(d_1!d_2!\ldots d_e!)$ possible permutations.

Let $k_i = \frac{t!}{d_1!\ldots d_e!}$. Then the total number of possible permutations is given by $k = \sum_{i=1}^{2^t-1} k_i$. Define $t_i = \frac{1}{12}\sum_{j=1}^{e}\left(d_{ij}^3 - d_{ij}\right)$, $\theta_i = 1 - \frac{12t_i}{t\left(t^2-1\right)}$, $\theta = \sum_{i=1}^{2^t-1} k_i\theta_i$.

Theorem 4.4. *(a) The asymptotic distribution of the Spearman statistic under the null hypothesis $H_0 : \mathbf{p}_i = \frac{1}{k}$ is as $n \to \infty$ given by*

$$(t-1)\,\frac{k}{\theta}\left\{(n-1)\,\overline{\rho}_n + \frac{\theta}{k}\right\} \to_L \chi^2_{t-1}.$$

(b) The asymptotic distribution of the Kendall statistic under the null hypothesis $H_0 : \mathbf{p}_i = \frac{1}{k}$ is given by

$$n\bar{\tau}_n \Rightarrow_L \frac{2}{3t\,(t-1)}\left\{\frac{\theta}{k}\,(t+1)\,\chi^2_{t-1} + \frac{3\,(\beta-2\gamma)}{k}\chi^2_{\binom{t-1}{2}}\right\} - \frac{\beta}{k}$$

where the two χ^2 variates are independent and

$$\beta = \sum_{i=1}^{2^t-1} k_i\beta_i, \beta_i = \left(t^2 - \sum_j d_{ij}^2\right) / \left(t\,(t-1)\right)$$

$$\gamma = \sum_{i=1}^{2^t-1} k_i\gamma_i, \gamma_i = \frac{1}{t-2}\left\{\theta_i\,\frac{t+1}{3} - \beta_i\right\}.$$

Proof. See Alvo and Cabilio (1985) for the proof. □

When ties are not allowed, $\theta_i = \beta_i = 1$, $\theta = \beta = k = t!$, $\gamma_i = \frac{1}{3}$, $\gamma = \frac{k}{3}$.

4.2 Tests for Agreement Among Groups

We may wish to compare two groups of patients with respect to how they perceive their hospitalization, those who require bed rest and those who are mobile in their recovery. Each patient is presented with a set of situations and asked to rank them in order of severity of stress. The result is that two sets of rankings are obtained and it is necessary to determine if the groups are responding in a similar manner. In another example Hollander and Sethuraman (1978) considered data of C. Sutton in his/her 1976 thesis on leisure preferences and attitudes on retirement of the elderly for 14 white and 13 black females in the age group 70–79 years. Each individual was asked: with which sex do you wish to spend your leisure? Each female was asked to rank the three responses: male(s), female(s) or both, assigning rank 1 for the most desired and 3 for the least desired. The first object in the ranking corresponds to "male," the second to "female," and the third to "both." It was desired to compare these two groups. The data is reproduced in Table 4.1.

Table 4.1 Sutton data on leisure preferences

Rankings	(123)	(132)	(213)	(231)	(312)	(321)
Frequencies for white females	0	0	1	0	7	6
Frequencies for black females	1	1	0	5	0	6

We begin with a general introduction to the concepts of diversity and dissimilarity. These concepts provide a generalization of the classical analysis of variance and are particularly applicable to data in the form of rankings. Consider a set of g populations where the individuals are characterized by a set of rankings chosen from the set of all possible rankings $\mathcal{P}$ in accordance with some distribution.

Definition 4.2. The diversity coefficient of the population whose distribution on the set of possible rankings is $\mathbf{p}_i$ is defined to be

$$H_i = \mathbf{p}_i' \Delta \mathbf{p}_i$$

where Δ is the matrix of pairwise distances between rankings. The diversity coefficient is the average difference between two randomly chosen individuals from the ith population.

Similarly, we may define the similarity coefficient when one individual is drawn from the ith and another from the jth population

$$H_{ij} = \mathbf{p}_i' \Delta \mathbf{p}_j . \tag{4.14}$$

The dissimilarity coefficient or between population diversity is then defined to be the difference

$$H_{ij} - \frac{1}{2}\left(H_i + H_j\right) = -\frac{1}{2}\left(\mathbf{p}_i - \mathbf{p}_j\right)' \Delta \left(\mathbf{p}_i - \mathbf{p}_j\right). \tag{4.15}$$

Suppose now that the individuals are mixed together in accordance with the proportions $\lambda_1, \ldots, \lambda_g$ such that $\sum_{i=1}^{g} \lambda_i = 1$. The convex set generated by the mixture leads to a new population with probability vector $\mathbf{p} = \sum_{i=1}^{g} \lambda_i \mathbf{p}_i$. The notions of diversity and between population diversity can now be formally defined.

Definition 4.3. The total diversity, the within population diversity, and the between population diversity are defined respectively to be

(i) $H(\mathbf{p}) = \mathbf{p}' \Delta \mathbf{p}$,
(ii) $H_W = \sum_i \lambda_i \mathbf{p}_i' \Delta \mathbf{p}_i$,
(iii) $H_B = -\sum_{i<j} \lambda_i \lambda_j \left(\mathbf{p}_i - \mathbf{p}_j\right)' \Delta \left(\mathbf{p}_i - \mathbf{p}_j\right)$.

It can be seen that

$$H(\mathbf{p}) = H_B + H_W .$$

The requirement that the between population diversity H_B be positive demands that

$$a'\Delta a \leq 0 \text{ whenever } \sum_{i=1}^{t!} a(i) = 0.$$

It can be seen from (4.15) that this condition is equivalent to requiring that H be a concave function. The concavity requirement imposes certain conditions on the distance function which must be verified for each potential distance measure. It does not follow from the right-invariance property. The following lemma makes this requirement more precise.

Lemma 4.3. *If the distance measure $d(\mu, \nu)$ is right invariant on the set of permutations, then there exists a constant $c > 0$ such that*

$$\Delta \mathbf{1} = (ct!)\,\mathbf{1}$$

and $H(p)$ is concave if and only if

$$Q^* = cJ - \Delta$$

is positive semidefinite. Moreover, in this case $H(p)$ has the maximum value c at $u = \frac{1}{t!}\mathbf{1}$.

Proof. The existence of the eigenvalue $ct!$ follows from the right-invariance property of the distance measure. We note that whenever $a'\mathbf{1} = 0$, for any $x = a + b\mathbf{1}$, which includes all points in $R^{t!}$

$$x'Q^*x = cb'Jb - x'\Delta x \geq 0.$$

Writing $\mathbf{p} = u + (\mathbf{p} - u)$ we note that since u is an eigenvector of Δ orthogonal to $(\mathbf{p} - u)$ we have

$$H(\mathbf{p}) = u'\Delta u + (\mathbf{p} - u)'\,\Delta\,(\mathbf{p} - u) \leq c$$

showing that for right-invariant measures, the uniform distribution over the set of all permutations is most diverse among diversity measures. □

Specializing to the Spearman and Kendall distances, we saw earlier in Chap. 3 that the matrix cQ can be expressed as

$$c_K Q_K = \mathbf{T}_K'\mathbf{T}_K,\ c_S Q_S = \mathbf{T}_S'\mathbf{T}_S.$$

In the next result we establish the link between the characteristic $\mathbf{T}\pi$ and the between population diversity, thus showing that it is this characteristic which forms the basis for inference when comparing populations.

Lemma 4.4. *For a right-invariant metric on the set of permutations, the between population diversity is given by*

$$\sum_{i<j} \lambda_i \lambda_j \left\| \mathbf{T}\mathbf{p}_i - \mathbf{T}\mathbf{p}_j \right\|_s^2 = tr\left\{ var\left(\mathbf{T}\mathbf{p}_I \right) \right\}$$

where $\|.\|_s$ *is the Euclidean norm in* R^s *and* I *has the distribution* $P(I = i) = \lambda_i$.

Proof. We note that since $\Delta = cJ - \mathbf{T}'\mathbf{T}$,

$$-\sum_{i<j} \lambda_i \lambda_j \left[\left(\mathbf{p}_i - \mathbf{p}_j \right)' \Delta \left(\mathbf{p}_i - \mathbf{p}_j \right) \right] =$$

$$\sum_{i<j} \lambda_i \lambda_j \left[\left(\mathbf{p}_i - \mathbf{p}_j \right)' \mathbf{T}'\mathbf{T} \left(\mathbf{p}_i - \mathbf{p}_j \right) \right] = \sum_{i<j} \lambda_i \lambda_j \left\| \mathbf{T}\mathbf{p}_i - \mathbf{T}\mathbf{p}_j \right\|_s^2 .$$

□

Suppose that we have a random sample of n_i judges from population i each of whom chooses a ranking in accordance with some distribution $\mathbf{p}_i$. Set $N = \sum n_i$. Given that the basis for inference for comparing two or more groups are the characteristics $\mathbf{Tp}$, consider therefore a test of the null hypothesis

$$H_0 : \mathbf{T}\mathbf{p}_1 = \mathbf{T}\mathbf{p}_2 = \ldots = \mathbf{T}\mathbf{p}_g \tag{4.16}$$

against the alternative that at least two among the $T\mathbf{p}_i$ are not equal. We observe for each group i, the relative frequency of occurrence of each ranking $\nu_l, l = 1, \ldots, t!$ denoted by $\hat{\mathbf{p}}_i(l), i = 1, \ldots, g$. Set $\hat{\mathbf{p}}_i = \left(\hat{\mathbf{p}}_i(1), \ldots, \hat{\mathbf{p}}_i(t!) \right)'$. A central limit theorem exists for each of the statistics $T\hat{\mathbf{p}}_i$.

Theorem 4.5. *Suppose that* $n_i / N \to \lambda_i > 0$ *as* $N \to \infty$. *Then*

(a)

$$\sqrt{n_i}\mathbf{T}\left(\hat{\mathbf{p}}_i - \mathbf{p}_i \right) \Rightarrow Z_i$$

where

$$Z_i \sim N_s\left(0, \mathbf{T}\Sigma_i\mathbf{T}' \right)$$

and the Z_i *are independent with*

$$\Sigma_i = \Pi_i - \mathbf{p}_i\mathbf{p}_i', \Pi_i = diag\left(\mathbf{p}_i(1), \ldots, \mathbf{p}_i(t!) \right).$$

(b) Under H_0,

$$\sqrt{N}\mathbf{T}\left(\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j \right) \Rightarrow N_s\left(0, \mathbf{T}\Sigma\mathbf{T}' \right)$$

where

$$\Sigma = \frac{\Sigma_i}{\lambda_i} + \frac{\Sigma_j}{\lambda_j}.$$

Moreover, for a consistent estimator $\hat{\Sigma}$ *of* Σ *and if* $\hat{D}$ *is the Moore–Penrose inverse of* $\mathbf{T}\hat{\Sigma}\mathbf{T}'$*, then*

$$N\left(\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j\right)' \mathbf{T}' \hat{D} \mathbf{T}\left(\hat{\mathbf{p}}_i - \hat{\mathbf{p}}_j\right) \Rightarrow \chi_r^2$$

where $r = rank(\mathbf{T}\Sigma\mathbf{T}')$.

Proof. (a) The multivariate central limit theorem applies to multinomial vectors

$$\sqrt{n_i}\left(\hat{\mathbf{p}}_i - \mathbf{p}_i\right) \Rightarrow N_{t!}\left(0, \Sigma_i\right).$$

The result follows.

(b) Using standard multivariate normal theory (Timm 1975), this part follows from the independence of the $Z_i's$ and the null hypothesis.

□

We note that the use of the Moore–Penrose inverse may be circumvented by choosing the matrix $\mathbf{T}$ so that $\mathbf{T}\Sigma\mathbf{T}'$ is of full rank. Hence, in the case of the Spearman distance, we may reduce the matrix $\mathbf{T}_S$ by using only the ranks of the first (t-1) objects. This problem does not immediately arise for the Kendall distance since there is no singularity in $\mathbf{T}_\mathbf{K}$.

An unbiased estimate of the covariance matrix Σ_i is given by

$$\hat{\Sigma}_i = \frac{n_i}{n_i - 1}\left(\hat{\Pi}_i - \hat{\mathbf{p}}_i \hat{\mathbf{p}}_i'\right)$$

where

$$\hat{\Pi}_i = diag\left(\hat{\mathbf{p}}_i(1), \ldots, \hat{\mathbf{p}}_i(t!)\right).$$

Suppose now that we are interested in the two-sample problem and that we wish to test the null hypothesis that

$$H_o(1) : \mathbf{T}\mathbf{p}_1 = \mathbf{T}\mathbf{p}_2.$$

Under $H_o(1)$, it follows that an estimate of the covariance matrix Σ in Theorem 4.5 is given by

$$\hat{\Sigma}_{Separate} = N\left(\frac{\hat{\Sigma}_1}{n_1} + \frac{\hat{\Sigma}_2}{n_2}\right).$$

The separate estimation of the covariances is appropriate in this case since the covariances are not assumed to be equal. In the situation when the null hypothesis is given by

$$H_o\,(2) : \mathbf{p}_1 = \mathbf{p}_2,$$

we may pool the separate estimates as

$$\hat{\Sigma}_{Pooled} = N\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\frac{(n_1 - 1)\,\hat{\Sigma}_1 + (n_2 - 1)\,\hat{\Sigma}_2}{N - 2}\right).$$

Hollander and Sethuraman (1978) actually used the combined estimate

$$\hat{\Sigma}_{combined} = \left(\frac{N-2}{N-1}\right)\hat{\Sigma}_{pooled} + \left(\frac{N}{N-1}\right)(f_1 - f_2)\,(f_1 - f_2)'$$

where f_1, f_2 are the frequency vectors. It should be noted that the estimates of the $(s \times s)$ covariance matrices are based on the observed score vectors $\{t\,(X_{ij})\}$; that is,

$$\mathbf{T}\hat{\Sigma}_i\mathbf{T}' = \frac{\sum_{j=1}^{n_i}\left(t\left(R_{ij}\right) - \bar{t}_{i.}\right)\left(t\left(R_{ij}\right) - \bar{t}_{i.}\right)'}{n_i - 1}$$

where R_{ij} is the observed ranking of judge j in group i and

$$\bar{t}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} t\left(R_{ij}\right).$$

Consequently, the calculations do not require computation of the individual covariance matrices $\hat{\Sigma}_i$.We may apply the methodology to the following example on leisure time preferences.

Example 4.1. Sutton data was analyzed in Feigin and Alvo (1986) using both the Spearman and Kendall test statistics. The total diversity was apportioned as indicated in Table 4.2. It can be seen that there is strong evidence that the two groups of females differ significantly.

The hypothesis expressed in (4.16) can alternatively be tested by using general multivariate analysis of variance methods. We do not pursue this further but instead refer the reader to Timm (1975).

Table 4.2 Analysis of the Sutton data

	Spearman	Kendall
Within	0.88	1.51
Between	0.41	0.54
Total	1.29	2.05
	χ^2_2	χ^2_3
Separate	28.0	28.1
Pooled	28.5	28.5

4.3 Test for Interaction in a Two-Way Layout

In this section, we consider the general two-factor design with equal numbers of replications in each cell. Such designs are utilized in statistics to test for main effects and for interactions in a variety of experiments. In more recent times, they have been applied in a genetics environment in order to understand the underlying biological mechanisms. See Gao and Alvo (2005b) for an application in a more general situation. In the gene expression data of *Drosoplila melanogaster* (Jin et al. 2001) for example, there are 24 cDNA microarrays, 6 for each combination of two genotypes (Oregon R and Samarkand) and two sexes. As each array used two different dyes, there were in total 48 separate labeling reactions. Focusing on the individual expression level of a gene and its relationship with genotypes and sexes, the objective of the study was to identify genes whose expression levels are affected by the interaction between the two factors. For such data, the assumption of normality for the error terms is not warranted and consequently, nonparametric procedures are needed. We shall consider a nonparametric test for interaction based on the row ranks and column ranks of the data.

We consider the following general two-way layout with interaction

$$X_{ijn} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijn}, i = 1, \ldots, I, j = 1, \ldots, J, n = 1, \ldots, N$$

where X_{ijin} is the response, $\{\alpha_i\}$ and $\{\beta_j\}$ are main effects, $\{\gamma_{ij}\}$ are interaction effects, and $\{\epsilon_{ijn}\}$ are independent and identically distributed according to a continuous cumulative distribution F_{ij}. We wish to test the null hypothesis of no interaction effects

$$H_0 : \gamma_{ij} = 0 \text{ for all } i, j$$

against the alternative

$$H_1 : \gamma_{ij} \neq 0 \text{ for some } i, j.$$

We propose a test statistic based on both row and column ranks. This statistic is invariant under monotone transformations and therefore can be applied directly on the original data. In order to motivate the test, let R_{ijn} be the rank of X_{ijn} with respect to the entries in the ith row. Similarly, let C_{ijn}be the rank of X_{ijn} with respect to the entries in the jth column. Define the score

$$a_{ijn} = \frac{R_{ijn}}{NJ+1} + \frac{C_{ijn}}{NI+1}. \tag{4.17}$$

Set the indicator function

$$u(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0. \end{cases}$$

It then follows that

$$\begin{aligned} E\left(a_{ijn}\right) &= \frac{1}{NJ+1} \sum_{b=1}^{J} \sum_{n'=1}^{N} Eu\left(X_{ijn} - X_{ibn'}\right) + \frac{1}{NI+1} \sum_{a=1}^{I} \sum_{n'=1}^{N} Eu\left(X_{ijn} - X_{ajn'}\right) \\ &= \frac{1}{NJ+1}\left(N \sum_{b=1}^{J} \int F_{ib} dF_{ij} + \frac{1}{2}\right) + \frac{1}{NI+1}\left(N \sum_{a=1}^{I} \int F_{aj} dF_{ij} + \frac{1}{2}\right). \end{aligned}$$

Under the null hypothesis of no interaction effects

$$\sum_{b} \int F_{ib} dF_{ij} = \int F\left(x - \alpha_i - \beta_0\right) dF\left(x - \alpha_i - \beta_j\right) = \int F\left(x + \beta_j - \beta_0\right)$$

which does not depend on i. Similarly, $\sum_a \int F_{aj} dF_{ij}$ does not depend on j. Setting

$$\overline{a}_{ij.} = \frac{1}{N} \sum_{n} a_{ijn}, \overline{a}_{i.} = \frac{1}{NJ} \sum_{n} a_{ijn}, \overline{a}_{.j.} = \frac{1}{NI} \sum_{n} a_{ijn}, \overline{a}_{...} = \frac{1}{NIJ} \sum_{n} a_{ijn}$$

it follows that

$$E\left(\overline{a}_{ij.} - \overline{a}_{i..} - \overline{a}_{.j.} + \overline{a}_{...}\right) = 0.$$

The quantity $\overline{a}_{ij.} - \overline{a}_{i..} - \overline{a}_{.j.} + \overline{a}_{...}$ serves as the nonparametric analogue of $\overline{X}_{ij.} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{X}_{...}$ which is the measure of the interaction effect appearing in the F statistic in the usual normal theory test for interaction.

4.3.1 Proposed Row–Column Test Statistic

In light of the motivation for the test statistic, define the sum of the row ranks in the (i, j) cell

$$\mathbf{S}_N(i,j) = \frac{1}{NJ+1}\sum_{n=1}^{N} R_{ijn}$$

and set

$$\mathbf{S}_N = \left(\mathbf{S}_N(1,1), \ldots, \mathbf{S}_N(I,J)\right)'.$$

Similarly, define the sum of the column ranks in the (i, j) cell

$$\mathbf{T}_N(i,j) = \frac{1}{NI+1}\sum_{n=1}^{N} C_{ijn}$$

and set

$$\mathbf{T}_N = \left(\mathbf{T}_N(1,1), \ldots, \mathbf{T}_N(I,J)\right)'.$$

Let $\mathbf{I}_I$ and $\mathbf{I}_J$ represent the I × I and J × J identity matrices, respectively, and let $\mathbf{J}_I$ and $\mathbf{J}_J$ represent the I × I and J × J matrices with all elements equal to one, respectively. Set

$$\mathbf{A} = \mathbf{J}_I \otimes \left(-\frac{1}{I}\mathbf{I}_J\right) + \mathbf{I}_I \otimes \mathbf{I}_J,$$

$$\mathbf{B} = \mathbf{I}_J \otimes \left(\mathbf{I}_J - \frac{1}{J}\mathbf{J}_J\right).$$

We note that the (i, j) term of $\frac{1}{N}(\mathbf{AS}_N + \mathbf{BT}_N)$ is $\left(\overline{a}_{ij.} - \overline{a}_{i..} - \overline{a}_{.j.} + \overline{a}_{...}\right)$. The proposed test statistic is then given by

$$W = \frac{1}{N}(\mathbf{AS}_N + \mathbf{BT}_N)' \left(\hat{\sum}\right)^{-} (\mathbf{AS}_N + \mathbf{BT}_N) \tag{4.18}$$

where $\left(\hat{\sum}\right)^{-}$ is the generalized inverse of the estimate of the variance-covariance matrix of $\mathbf{AS}_N + \mathbf{BT}_N$. The covariance matrix is not of full rank since there exist I+J-1 linear combinations of $\mathbf{AS}_N + \mathbf{BT}_N$ which are constants. We may obtain a formal expression for the estimate of the covariance matrix. Let

$$\sum_1 = \lim \frac{1}{N} Var\left(\mathbf{S_N}\right), \sum_2 = \lim \frac{1}{N} Var\left(\mathbf{T_N}\right), \sum_{12} = \lim \frac{1}{N} cov\left(\mathbf{S_N}, \boldsymbol{T_N}\right).$$

Then

$$\hat{\sum} = \mathbf{A}\hat{\sum}_1\mathbf{A}' + \mathbf{B}\hat{\sum}_2\mathbf{B}' + 2\mathbf{A}\hat{\sum}_{12}\mathbf{B}'$$

where estimates of the covariances denoted by hats will be given in the next section.

4.3.2 *Asymptotic Distribution of the Test Statistic Under the Null Hypothesis*

The asymptotic distribution of the test statistic W in (4.18) is a consequence of the general theory for linear rank statistics. We begin by recalling some theorems of Hajek.

Let $X_1, \ldots, X_N$ be independent random variables with continuous distribution functions $F_1, \ldots, F_N$, respectively. Let R_i be the rank of X_i among $X_1, \ldots, X_N$ and let $c_i, i = 1, \ldots, N$ be regression coefficients. Let $\alpha_N(x)$ be generated by a real values function $\phi(x)$ having a second derivative as

$$\alpha_N(i) = \phi\left(\frac{i}{N+1}\right).$$

A simple linear rank statistic takes the form

$$\mathbf{S} = \sum_{i=1}^{N} c_i \alpha_N(R_i).$$

Let

$$\begin{aligned}
\overline{c} &= \frac{1}{N}\sum c_i, \\
\overline{\phi} &= \int_0^1 \phi(x)\, dx, \\
H(x) &= \frac{1}{N}\sum F_i(x), \\
\mu &= \sum c_i \int \phi(H(x))\, dF_i(x).
\end{aligned}$$

We quote the following two theorems from Hajek (1968).

Theorem 4.6. *Let*

$$L_i(x) = \frac{1}{N}\sum_{j=1}^{N}(c_j - c_i)\int [u(y-x) - F_i(x)]\,\phi'(H(x))\, dF_j(x)$$

and

$$\sigma^2 = \sum var\left(L_i\left(X_i\right)\right).$$

If for every $\varepsilon > 0$, *there exists* K_ϵ *such that*

$$Var\left(\mathbf{S}\right) > K_\epsilon \max_{1 \le i \le N} \left(c_i - \overline{c}\right)^2,$$

then

$$\max_{\infty < x < \infty} \left| P\left(\mathbf{S} - E\mathbf{S} < x\left(var\mathbf{S}\right)^{1/2}\right) - \Phi\left(x\right) \right| < \epsilon$$

where Φ *denotes the standard normal distribution function. The conclusion still holds if* $var(\mathbf{S})$ *is replaced by* σ^2. *If* $\sum c_i^2$ *is bounded by a multiple of* $\sum \left(c_i - \overline{c}\right)^2$, $E\mathbf{S}$ *can be replaced by* μ *in the conclusion.*

We note that an integration by parts yields

$$\int \left[u\left(y - x\right) - F_i\left(x\right)\right] \phi'\left(H\left(x\right)\right) dF_i\left(x\right) = \int_x^\infty \phi'\left(H\left(y\right)\right) dF_j\left(y\right) + \text{constant}.$$

Moreover, $EL_i\left(X_i\right) = 0$.

The proof of Theorem 4.6 makes use of a projection argument. It is shown that the statistic $\mathbf{S} - E\mathbf{S}$ can be approximated best in the mean square sense by the statistic

$$\hat{\mathbf{S}} = \sum_{i=1}^{N} L_i\left(X_i\right)$$

which is the projection onto the Hilbert space generated by sums of independent square integrable linear functions of the X_i. The next result from Hajek makes this notion more precise.

Theorem 4.7. *Let* $Z_i = L_i\left(X_i\right)$. *There exists a constant M independent of N such that*

$$E\left(\mathbf{S} - E\mathbf{S} - \sum_{i=1}^{N} Z_i\right)^2 \le \frac{M}{N} \sum_{i=1}^{N} \left(c_i - \overline{c}\right)^2$$

and

$$E\left(\mathbf{S} - \mu\right)^2 \le \frac{M}{N} \sum_{i=1}^{N} c_i^2.$$

The proof of the asymptotic normality of our test statistic rests on extending Hajek's result to the study of composite linear rank statistics. We illustrate this result in the following simple situation whereby $X_1, \ldots, X_{n_1}, \ldots, X_{n_1+n_2}, \ldots, X_N$ are independent random variables. Consider the two simple linear rank statistics

$$\mathbf{S}_1 = \sum_{i=1}^{n_1+n_2} c_i^{(1)} a\left(R_i^{(1)}\right),$$

$$\mathbf{S}_2 = \sum_{i=n_1+1}^{N} c_i^{(2)} a\left(R_i^{(2)}\right),$$

where $R_i^{(1)}$ is the rank of X_i among $\{X_1, \ldots, X_{n_1+n_2}\}$ and $R_i^{(2)}$ is the rank of X_i among $\{X_{n_1+1}, \ldots, X_N\}$. We are interested in the asymptotic normality of the composite linear rank statistic formed by the sum

$$\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2.$$

This is done by adapting the projection argument. First, $\mathbf{S}_1$ is projected onto the space spanned by linear combinations of $\{X_1, \ldots, X_{n_1+n_2}\}$. Next, $\mathbf{S}_2$ is projected onto the space spanned by linear combinations of $\{X_{n_1+1}, \ldots, X_N\}$. Then the sum is projected onto the combined space $\{X_1, \ldots, X_N\}$. Let

$$W_i = \begin{cases} Z_i & i = 1, \ldots, n_1, \\ Z_i + Z_i^* & i = n_1 + 1, \ldots, n_1 + n_2, \\ Z_i^* & i = n_1 + n_2 + 1, \ldots, N, \end{cases}$$

where $Z_i = L_i(X_i)$ and $Z_i^* = L_i^*(X_i)$ are the respective projections. Here,

$$L_i(x) = \frac{1}{n_1 + n_2} \sum_{j=1}^{n_1+n_2} \left(c_j^{(1)} - c_i^{(1)}\right) \int \left[u(y - x) - F_i(x)\right] \phi'(H_1(x))\, dF_j(x),$$

$$L_i^*(x) = \frac{1}{n_2 + n_3} \sum_{j=1}^{n_2+n_3} \left(c_j^{(2)} - c_i^{(2)}\right) \int \left[u(y - x) - F_i(x)\right] \phi'(H_2(x))\, dF_j(x),$$

with

$$H_1(x) = \frac{1}{n_1 + n_2} \sum_{i=1}^{n_1+n_2} F_i(x), H_2(x) = \frac{1}{n_2 + n_3} \sum_{i=1}^{n_2+n_3} F_i(x).$$

We note that $EW_i = 0$.

Theorem 4.8. *Let $\boldsymbol{S_1}, \boldsymbol{S_2}$ be defined as above. Also let*

$$L_N = \max\left\{\sup\left(c_j^{(1)} - \overline{c}^{(1)}\right)^2, \sup\left(c_j^{(2)} - \overline{c}^{(2)}\right)^2\right\}$$

and

$$\sigma_N^2 = Var\left(\sum_{i=1}^{N} W_i\right).$$

If the following condition holds

$$\lim \frac{L_N}{\sigma_N^2} = 0, as \ \min\left(n_1 + n_2, n_2 + n_3\right) \to \infty,$$

then

$$\frac{\mathbf{S}_1 + \mathbf{S}_2 - E\left(\mathbf{S}_1 + \mathbf{S}_2\right)}{\sigma_N} \Rightarrow_L N\left(0, 1\right).$$

Proof. From (4.7), there exist constants M_1, M_2 such that

$$E\left(\mathbf{S}_1 - E\mathbf{S}_1 - \sum_{i=1}^{n_1+n_2} Z_i\right)^2 \leq \frac{M_1}{n_1 + n_2} \sum_{i=1}^{n_1+n_2} \left(c_i^{(1)} - \overline{c}^{(1)}\right)^2,$$

$$E\left(\mathbf{S}_2 - E\mathbf{S}_2 - \sum_{i=n_1+1}^{N} Z_i^*\right)^2 \leq \frac{M_2}{n_2 + n_3} \sum_{i=n_1+1}^{N} \left(c_i^{(2)} - \overline{c}^{(2)}\right)^2.$$

Hence,

$$E\left(\mathbf{S}_1 + \mathbf{S}_2 - E\left(\mathbf{S}_1 + \mathbf{S}_2\right) - \sum_{i=1}^{N} W_i\right)^2 \leq 2\left(M_1 + M_2\right) L_M.$$

It remains to show that $\sum_{i=1}^{N} W_i \sigma_N$ is asymptotically normally distributed with mean 0 and variance 1. This follows from the Lindeberg theorem. □

We state the general limiting distribution of the vector $(\mathbf{S}_N, \mathbf{T}_N)$.

Theorem 4.9. *Under the assumption that the errors $\{\epsilon_{ijn}\}$ are independent identically distributed in the two-way layout, we have that as $N \to \infty$*

(i) $\frac{1}{\sqrt{N}}\begin{pmatrix}\mathbf{S}_N - E(\mathbf{S}_N)\\ \mathbf{T}_N - E(\mathbf{T}_N)\end{pmatrix} \Longrightarrow N_{2IJ}(0, \sum)$

(ii) $W \Longrightarrow \chi^2_{(I-1)(J-1)}$ *under* H_0

(iii) $W \Longrightarrow \chi^2_{(I-1)(J-1)}(\delta)$ *under* H_1 *where* δ *is the noncentrality parameter under Pitman alternatives*

Proof. The proof makes use of projection arguments and Theorems 4.7 and 4.8 above. We refer the reader to Gao and Alvo (2005a) for details of the proof. □

To estimate the covariance of the test statistics define the following variables involving the empirical distribution functions:

$$C_{abn}^{(i,j)} = \begin{cases} -\frac{1}{NJ}\sum_{n'=1}^{N} u\left(X_{abn} - X_{ajn'}\right) & a = i, b \neq j, \\ \frac{1}{NJ}\sum_{j\neq j'}\sum_{n'=1}^{N} u\left(X_{abn} - X_{aj'n'}\right) & a = i, b = j, \\ 0 & a \neq i. \end{cases}$$

The fact that $C_{abn}^{(i,j)} - W_{abn}^{(i,j)} \to 0$ almost surely leads to the following consistent estimator

$$\hat{\sigma}_1^2(i,j,i',j') = \sum_{a,b}\frac{1}{N}\sum_{n=1}^{N}\left(C_{abn}^{(i,j)} - \overline{C_{ab.}^{(i,j)}}\right)\left(C_{abn}^{(i',j')} - \overline{C_{ab.}^{(i',j')}}\right).$$

Similarly, defining

$$G_{abn}^{(i,j)} = \begin{cases} -\frac{1}{NI}\sum_{n'=1}^{N} u\left(X_{abn} - X_{ajn'}\right) & a \neq i, b = j, \\ \frac{1}{NI}\sum_{i\neq i'}\sum_{n'=1}^{N} u\left(X_{abn} - X_{aj'n'}\right) & a = i, b = j, \\ 0 & b \neq j, \end{cases}$$

we may construct consistent estimators for Σ_2 and Σ_{12}, respectively,

$$\hat{\sigma}_2^2(i,j,i',j') = \sum_{a,b}\frac{1}{N}\sum_{n=1}^{N}\left(G_{abn}^{(i,j)} - \overline{G_{ab.}^{(i,j)}}\right)\left(G_{abn}^{(i',j')} - \overline{G_{ab.}^{(i',j')}}\right),$$

$$\hat{\sigma}_{12}^2(i,j,i',j') = \sum_{a,b}\frac{1}{N}\sum_{n=1}^{N}\left(C_{abn}^{(i,j)} - \overline{C_{ab.}^{(i,j)}}\right)\left(G_{abn}^{(i',j')} - \overline{G_{ab.}^{(i',j')}}\right).$$

Gao and Alvo (2005a) report the result of some simulation studies which compare the proposed row–column statistic with the aligned test as well as the rank transform test. It is shown that the row–column test performs very well under a variety of underlying distributions including the normal, contaminated normal, and Cauchy. The following example was also considered.

Example 4.2. Consider the gene expression data of *D. melanogaster* of Jin et al. (2001). The gene fs(1) $k10$ is known to be expressed in reproductive systems and its expression level was reportedly affected by the gender and genotype interaction. The row–column statistic was applied to this data to account for the genotype, the gender, and the genotype-gender interaction. It was found that the interaction effect was statistically significant with a p-value equal to 0.004. The parametric F statistic and the aligned rank transform using the residuals yielded similar results. In order to illustrate the robustness of the nonparametric procedures, the analyses were redone with the first observation changed to an arbitrarily large number. The performance of the F statistic was severely affected and yielded a nonsignificant result. On the other hand, the nonparametric procedures were unaffected.

Next, we recall that in an example of a 3×4 factorial design considered by Box and Cox (1964) it was claimed that only after application of a nonlinear transformation can the error term be stabilized and the data made suitable for standard statistical analysis. We applied the row–column procedure to the untransformed data and obtained a p-value of 0.44. Thus the hypothesis of no interaction was not rejected, a finding that concurs with Box and Cox. The aligned test on the other hand yielded a p-value of 0.02 which indicates the presence of interaction. However, for the transformed data, the aligned test with a p-value of 0.45 did not reject the null hypothesis.

Chapter Notes

Alvo et al. (1982) developed a new approach to test for randomness. This allowed the consideration of various distance functions including Kendall's distance. Theorems 4.2 and 4.3 provide the asymptotic distributions of the Spearman and Kendall test statistics in the complete randomized block design. Iman and Davenport (1980) describe the F distribution approximation to the Friedman statistic which is used later in Chap. 10.

One question of interest in connection with the asymptotic results is how accurate are the asymptotic distributions. Alvo and Cabilio (1984) considered the accuracy of the asymptotic distribution of Kendall's test statistic and compared it to other approximations for small values of t and n. In addition, tables of the exact distribution were computed for $t = 3, n = 3, \ldots, 19; t = 4, n = 3, \ldots, 9$; and $t = 5, n = 3, 4, 5$. Some exact calculations are made of the Bahadur efficiency where it is demonstrated that the Kendall tau is more efficient.

Feigin and Alvo (1986) considered the two-group problem by placing it in the context of diversity and described an extensive discussion of the literature on the subject. Bu et al. (2009) developed an extension of the two-sample situation to the case where there are missing data. Although not discussed in this book, it may be of interest to consider the problem of paired comparisons whereby a judge ranks a set of objects before and after a treatment.

Gao and Alvo (2005b) provide a brief historical look at the analysis of unbalanced two-way layout with interaction effects. Using the notion of a weighted rank, they present tests for both main effects as well as for interaction effects. In addition, there is a discussion of the asymptotic relative efficiency of the proposed tests relative to the parametric F test. Various simulations further exemplify the power of the proposed tests. In a specific application, it is shown that the test statistic is the most robust in the presence of extreme outliers compared to other procedures.

Gao et al. (2008) also consider nonparametric multiple comparison procedures for unbalanced one-way factorial designs whereas Gao and Alvo (2008) treat nonparametric multiple comparison procedures for unbalanced two-way layouts.