Chapter 3
Correlation Analysis of Paired Ranking Data

3.1 Notion of Distance Between Two Rankings

A ranking represents the order of preference one has with respect to a set of t objects. If we label the objects by the integers 1 to t, a ranking can then be thought of as a permutation of the integers $(1, 2, \ldots, t)$. We may denote such a permutation by $\mu = (\mu(1), \mu(2), \ldots, \mu(t))'$ which may also be conceptualized as a point in t-dimensional space. It is natural to measure the spread between two individual permutations μ, ν by means of a distance function. There are several examples of distance functions that have been proposed in the literature. Here are a few:

Spearman

$$d_S(\mu, \nu) = \frac{1}{2} \sum_{i=1}^{t} (\mu(i) - \nu(i))^2 . \tag{3.1}$$

Kendall

$$d_K(\mu, \nu) = \sum_{i<j} \{1 - sgn\,(\mu(j) - \mu(i))\, sgn\,(\nu(j) - \nu(i))\}, \tag{3.2}$$

where $sgn(x)$ is either 1 or -1 depending on whether $x > 0$ or $x < 0$.

Hamming

$$d_H(\mu, \nu) = t - \sum_{i=1}^{t} \sum_{j=1}^{t} I\,(\mu(i) = j)\, I\,(\nu(i) = j) \tag{3.3}$$

where $I(.)$ is the indicator function taking values 1 or 0 depending on whether the statement in brackets holds or not.

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_3

Spearman Footrule

$$d_F(\mu, \nu) = \sum_{i=1}^{t} |\mu(i) - \nu(i)|. \tag{3.4}$$

The Spearman measure is not a proper "distance" in that it does not obey the triangular inequality property. We shall nonetheless refer to it as a distance function in this book. It is based upon squared Euclidean distance whereas the Footrule is based on the absolute deviations. The Kendall distance counts the number of "discordant" pairs whereas the Hamming distance counts the number of "mismatches." The Hamming distance has found uses in coding theory. These distances have the property of being invariant under any permutation relabeling of the objects. That is, for any permutations σ, μ, ν,

$$d(\mu, \nu) = d(\mu \circ \sigma, \nu \circ \sigma)$$

where $\mu \circ \sigma(i) = \mu(\sigma(i))$. This property is known as right invariance. Let $\Delta = \left(d\left(\mu_i, \mu_j\right)\right)$ denote the matrix of all pairwise distances. If d is right invariant, then it follows that there exists a constant $c > 0$ for which

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

where $\mathbf{1} = (1, 1, \ldots, 1)'$ is of dimension $t!$. Hence, c is equal to the average distance. It is straightforward to show that for the Spearman and Kendall distances

$$c_S = \frac{t(t^2 - 1)}{12}, c_K = \frac{t(t-1)}{2}.$$

Turning attention to the Hamming distance, we note that if $e = (1, 2, \ldots, t)'$, then

$$\begin{aligned} \Sigma_\mu d_H(\mu, e) &= \Sigma_\mu t - \Sigma_\mu \Sigma_i \Sigma_j I(\mu(i) = j) I(e(i) = j.) \\ &= t(t!) - t! \end{aligned}$$

and hence $c_H = (t - 1)$.

Example 3.1. Suppose that $t = 3$ and that the complete rankings are denoted by

$$\mu_1 = (1, 2, 3)', \mu_2 = (1, 3, 2)', \mu_3 = (2, 1, 3)', \mu_4 = (2, 3, 1)', \mu_5 = (3, 1, 2)',$$
$$\mu_6 = (3, 2, 1)'.$$

Using the above order of the permutations, we may write the matrix Δ of pairwise Spearman, Kendall, Hamming, and Footrule distances respectively as

$$\Delta_S = \begin{pmatrix} 0&1&1&3&3&4 \\ 1&0&3&1&4&3 \\ 1&3&0&4&1&3 \\ 3&1&4&0&3&1 \\ 3&4&1&3&0&1 \\ 4&3&3&1&1&0 \end{pmatrix}$$

$$\Delta_K = \begin{pmatrix} 0&2&2&4&4&6 \\ 2&0&4&2&6&4 \\ 2&4&0&6&2&4 \\ 4&2&6&0&4&2 \\ 4&6&2&4&0&2 \\ 6&4&4&2&2&0 \end{pmatrix}$$

$$\Delta_H = \begin{pmatrix} 0&2&2&3&3&2 \\ 2&0&3&2&2&3 \\ 2&3&0&2&2&3 \\ 3&2&2&0&3&2 \\ 3&2&2&3&0&2 \\ 2&3&3&2&2&0 \end{pmatrix}$$

$$\Delta_F = \begin{pmatrix} 0&2&2&4&4&4 \\ 2&0&4&2&4&4 \\ 2&4&0&4&2&4 \\ 4&2&4&0&4&2 \\ 4&4&2&4&0&2 \\ 4&4&4&2&2&0 \end{pmatrix}$$

These distances may alternatively be written in terms of a similarity function in the form

$$d(\mu, \nu) = c - \mathcal{A}(\mu, \nu), \tag{3.5}$$

Spearman:

$$\mathcal{A}_S = \mathcal{A}_S(\mu, \nu) = \sum_{i=1}^{t} \left(\mu(i) - \frac{t+1}{2} \right) \left(\nu(i) - \frac{t+1}{2} \right). \tag{3.6}$$

Kendall:

$$\mathcal{A}_K = \mathcal{A}_K(\mu, \nu) = \sum_{i<j} sgn\left(\mu(j) - \mu(i)\right) sgn\left(\nu(j) - \nu(i)\right). \tag{3.7}$$

Hamming:

$$\mathcal{A}_H(\mu,\nu)=\sum_{i=1}^{t}\sum_{j=1}^{t} I\left([\mu(i)=j]-\frac{1}{t}\right) I\left([\nu(i)=j]-\frac{1}{t}\right). \tag{3.8}$$

Footrule:

$$\mathcal{A}_F(\mu,\nu)=\sum_{i=1}^{t}\sum_{j=1}^{t} I\left([\mu(i)\le j]-\frac{j}{t}\right) I\left([\nu(i)\le j]-\frac{j}{t}\right).$$

The similarity measures may be also interpreted geometrically as inner products which sets the groundwork for defining correlation in the next section.

3.2 Correlation Between Two Rankings

The notion of correlation occurs frequently in statistics. For example, in regression analysis, one is interested in the correlation between two variables such as height and weight. Similarly, in nonparametric statistics, we shall be interested in the correlation between two rankings. Let $\mathcal{P}$ be the space of all possible permutations of the integers $1,2,\ldots,t$. We may define the correlation between two rankings μ,ν as

$$\alpha(\mu,\nu)=1-\frac{2d(\mu,\nu)}{M} \tag{3.9}$$

where M is the maximum value of the distance $d(\mu,\nu)$ taken over all possible pairs μ,ν in $\mathcal{P}$ (Diaconis and Graham 1977). In the case of the Spearman and Kendall distance, the maximum values occur when

$$\left(\mu(i)-\frac{t+1}{2}\right)=-\left(\nu(i)-\frac{t+1}{2}\right) \textit{ for all } i,$$

whereas the minimum occurs when

$$\left(\mu(i)-\frac{t+1}{2}\right)=\left(\nu(i)-\frac{t+1}{2}\right)$$

This is a consequence of the rearrangement inequality given as a lemma below.

Lemma 3.1. *Let $a_1,\ldots,a_t$ and $b_1,\ldots,b_t$ be real numbers, not necessarily positive with*

$$a_1\le\ldots\le a_t, b_1\le\ldots\le b_t$$

and let σ *be a permutation of the integers* $1, \ldots, t$. *Then*

$$a_1 b_t + \ldots + a_t b_1 \leq a_1 b_{\sigma(1)} + \ldots + a_t b_{\sigma(t)} \leq a_1 b_1 + \ldots + a_t b_t.$$

Proof. The proof follows by induction on t. □

It can be shown that for the Spearman and Kendall distances, the maximum is equal to twice the mean,

$$M_S = 2c_S, M_K = 2c_K. \tag{3.10}$$

In view of (3.10) we have

$$\alpha_S(\mu, \nu) = \frac{\mathcal{A}_S}{c_S}, \alpha_K(\mu, \nu) = \frac{\mathcal{A}_K}{c_K}. \tag{3.11}$$

Example 3.2 (Lehmann 1975, p. 298). Consider the test scores in Language and Arithmetic for a group of 9 students as shown in Table 3.1. The right-invariance property shared by the Spearman and Kendall distances enables us to rewrite the table in a more convenient fashion with one of the rankings in natural order as in Table 3.2. The Spearman and Kendall correlations are respectively 0.683 and 0.500. Here $c_S = 60, c_K = 36$.

The correlation coefficients based on these distances are of the multiplicative type (Kendall and Gibbons 1990); that is, there exists a function g such that

$$\alpha(\mu, \nu) = k_\mu k_\nu \sum_i \sum_j g(\mu(i), \mu(j)) g(\nu(i), \nu(j)) \tag{3.12}$$

Table 3.1 Language and Arithmetic scores

Student	1	2	3	4	5	6	7	8	9
Language	50	23	28	34	14	54	46	52	53
Arithmetic	38	28	14	26	18	40	23	30	27
Language ranks	6	2	3	4	1	9	5	7	8
Arithmetic ranks	8	6	1	4	2	9	3	7	5

Table 3.2 Language and Arithmetic scores rearranged

Student	5	2	3	4	7	1	8	9	6
Language	14	23	28	34	46	50	52	53	54
Arithmetic	18	28	14	26	23	38	30	27	40
Language ranks	1	2	3	4	5	6	7	8	9
Arithmetic ranks	2	6	1	4	3	8	7	5	9

where k_μ, k_ν are normalizing constants. The constants may be different depending on whether the coefficient is of type a or b. A type a correlation is used above. For Spearman and Kendall, the functions are, respectively,

$$g_S(\mu(i), \mu(j)) = (\mu(i) - \mu(j))$$

$$g_K(\mu(i), \mu(j)) = sgn[\mu(i) - \mu(j)].$$

For a type b correlation, the constants are given by

$$k_\mu = \sqrt{\Sigma_i \Sigma_j [g(\mu(i), \mu(j))]^2}.$$

We shall make use of a type b correlation when defining angular correlations in Sect. 3.6.

For a multiplicative index, it can be shown that the correlation matrix is necessarily positive semidefinite (Quade 1972). Setting

$$Q = \left(J - \frac{2}{M}\Delta\right) \tag{3.13}$$

where $J = \mathbf{1}\mathbf{1}'$ and $\frac{M}{2} = c$, this implies that there exists a matrix $\mathbf{T}$ for which

$$Q = \frac{1}{c}(\mathbf{T}'\mathbf{T}). \tag{3.14}$$

It follows that the distance matrix for both Spearman and Kendall can be expressed as

$$\Delta = cJ - \mathbf{T}'\mathbf{T}. \tag{3.15}$$

From the form of the Spearman and Kendall similarity measures (3.12), it can be seen that the matrices $\mathbf{T}$ are respectively

$$\mathbf{T}_S = (t_S(\mu_1), \ldots, t_S(\mu_{t!}))' \tag{3.16}$$

where

$$t_S(\mu) = \left(\mu(1) - \frac{t+1}{2}, \ldots, \mu(t) - \frac{t+1}{2}\right)'$$

is the centered rank vector and

$$\mathbf{T}_K = (t_K(\mu_1), \ldots, t_K(\mu_{t!}))' \tag{3.17}$$

is of dimension $\binom{t}{2} \times t!$ where the qth element for $q = (i-1)\left(t - \frac{i}{2}\right) + (j-i), 1 \leq i < j \leq t$,

$$(t_K(\mu))_q = sgn[\mu(j) - \mu(i)].$$

For Hamming, we may write the t^2-dimensional vector where the (i, j)th element is

$$(t_H(\mu))_{ij} = \left(I[\mu(i) = j] - \frac{1}{t}\right)$$

for $1 \leq i, j \leq t$.

For the Footrule we have the t^2-dimensional vector where the qth element for $q = (i-1)t + j, 1 \leq i < j \leq t$

$$(t_F(\mu))_q = \left(I[\mu(i) \leq j] - \frac{j}{t}\right).$$

Example 3.3. Suppose that $t = 3$. Then, placing the rankings in the natural order of Example 3.1, we have that

$$\mathbf{T}_S = \begin{pmatrix} -1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & -1 & 1 & -1 & 0 \\ 1 & 0 & 1 & -1 & 0 & -1 \end{pmatrix}$$

and

$$\mathbf{T}_K = \begin{pmatrix} 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 \end{pmatrix}.$$

The notion of correlation is particularly useful in problems wherein one wishes to test for the independence of two variables as in Example 3.2 or for the existence of long-term monotone trend in the pH of a river. We will postpone a discussion of these important topics later in this chapter where it will be addressed in the general context of incomplete rankings.

3.3 Incomplete Rankings and the Notion of Compatibility

A judge may rank a complete set of candidates in accordance with some criterion. On occasion, however, data may be missing either at random or by design. For example, one or more candidates may not be ranked. In another example, the pH data on a lake may not be available for certain months in a year, thereby making it

impossible to test for a long-term trend using traditional nonparametric rank-based statistics. The option to ignore the missing data is unsatisfactory because it distorts the time scale. As we shall see later on, this option is always suboptimal when testing for trend. We address the topic in this section by first introducing the notion of compatibility.

Notation. Incomplete ranks will be denoted by "−" and corresponding incomplete rankings will be written with an upper script "*".

For example, the ranking $\mu^* = (2, -, 3, 4, 1)'$ indicates that object 2 is unranked among the five objects presented.

Definition 3.1. The complete ranking μ of t objects is said to be compatible with an incomplete ranking μ^* of a subset of k of these objects, $2 \leq k \leq t$, if the relative ranking of every pair of objects ranked in μ^* coincides with their relative ranking in μ.

An incomplete ranking gives rise to a class of order preserving complete rankings. Denoting by $C(\mu^*)$ the set of complete permutations compatible with $\mu^* = (2, -, 3, 4, 1)'$, we have that

$$C\left(\mu^*\right) = \left\{(2,5,3,4,1)', (2,4,3,5,1)', (2,3,4,5,1)', (3,2,4,5,1)', (3,1,4,5,2)'\right\}.$$

The total number of complete rankings of t objects compatible with an incomplete ranking of a subset of k objects is given by $t!/k!$. This follows from the fact that there are $\binom{t}{k}$ ways of choosing k integers for the ranked objects, one way in placing them to preserve the order and then $(t-k)!$ ways of rearranging the remaining integers. The product is thus

$$a = \binom{t}{k}(t-k)! = t!/k! \tag{3.18}$$

The notion of compatibility establishes a connection between an incomplete ranking and the class of complete rankings from which the incomplete ranking could have arisen. It seems natural as a result to extend the notion of distance to incomplete rankings by referring to the corresponding compatibility classes.

Definition 3.2. The distance $d^*(\mu^*, \nu^*)$ between two incomplete rankings μ^* and ν^* is defined to be the average of all values of the distances $d(\mu_{i,j})$ taken over all pairs of complete rankings $\mu_{i,j}$ compatible with μ^* and ν^*, respectively.

Example 3.4. Suppose that $t = 3, k = 2$. In that case, the possible incomplete rankings are denoted by

$$\nu_{11}^* = (1,2,-)', \nu_{12}^* = (2,1,-)', \nu_{21}^* = (1,-,2)', \nu_{22}^* = (2,-,1)',$$
$$\nu_{31}^* = (-,1,2)', \nu_{32}^* = (-,2,1)'$$

We may associate with every incomplete ranking a ($t!$ x 1) compatibility vector, also denoted by $C(\nu^*)$, whose ith component is 1 or 0 according to whether μ_i is compatible with ν^*. A summary can be provided by a compatibility matrix as follows.

$$C = \begin{array}{c} \\ \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{array} \begin{array}{c} \begin{array}{cccccc} \nu_{11}^* & \nu_{12}^* & \nu_{21}^* & \nu_{22}^* & \nu_{31}^* & \nu_{32}^* \end{array} \\ \begin{array}{cccccc} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{array} \end{array}$$

Consequently, the matrix of average pairwise Spearman distances for the incomplete rankings is given by the product $C_S' \Delta C_S / a^2$ where $a = t!/\, k! = 3$ and

$$C_S' \Delta C_S = \begin{array}{c} \\ \nu_{11}^* \\ \nu_{12}^* \\ \nu_{21}^* \\ \nu_{22}^* \\ \nu_{31}^* \\ \nu_{32}^* \end{array} \begin{array}{c} \begin{array}{cccccc} \nu_{11}^* & \nu_{12}^* & \nu_{21}^* & \nu_{22}^* & \nu_{31}^* & \nu_{32}^* \end{array} \\ \begin{array}{cccccc} 10 & 26 & 14 & 22 & 22 & 14 \\ 26 & 10 & 22 & 14 & 14 & 22 \\ 14 & 22 & 10 & 26 & 14 & 22 \\ 22 & 14 & 26 & 10 & 22 & 14 \\ 22 & 14 & 14 & 22 & 10 & 26 \\ 14 & 22 & 22 & 14 & 26 & 10 \end{array} \end{array}$$

We note from this example that the distance of an incomplete ranking to itself is 10 and not 0. In extending the notion of correlation to incomplete rankings, it will be necessary to take this into account.

For the Spearman and Kendall distances, we may re-express the distance $d^*(\mu^*, \nu^*)$ as

$$d^*(\mu^*, \nu^*) = \frac{1}{a^2} \left[C\left(\mu^*\right)\right]' \Delta \left[C\left(\nu^*\right)\right] \tag{3.19}$$

$$= \frac{1}{a^2} \left[C\left(\mu^*\right)\right]' \left(cJ - \mathbf{T}'\mathbf{T}\right) \left[C\left(\nu^*\right)\right] \tag{3.20}$$

$$= c - \mathcal{A}^*(\mu^*, \nu^*)$$

where

$$\mathcal{A}^*(\mu^*, \nu^*) = \frac{1}{a^2} \left[C\left(\mu^*\right)\right]' \mathbf{T}'\mathbf{T} \left[C\left(\nu^*\right)\right].$$

The latter may be viewed as the average of the $\mathcal{A}(\mu_i, \nu_j)$ taken over all complete rankings μ_i, ν_j compatible with μ^* and ν^*, respectively.

3.4 Correlation for Incomplete Rankings

At this point it is useful to derive an expression for an incomplete ranking μ^* given knowledge of its compatibility class $C(\mu^*)$. We shall assume that each complete ranking has the same probability of being selected, i.e., they are uniformly distributed over the $t!$ permutations of $(1, 2, \ldots, t)$.

Lemma 3.2. *The conditional distribution of the rank $\mu(i)$ given the compatibility class $C(\mu^*)$ generated by μ^* is given by*

$$P\left\{\mu(i) = j | C\left(\mu^*\right)\right\} = \binom{j-1}{\mu^*(i)-1}\binom{t-j}{k-\mu^*(i)}\binom{t}{k}^{-1}\delta(i) + \frac{1}{t}(1-\delta(i))$$

where $\delta(i)$ is either 1 or 0 depending on whether the object i is or is not ranked in the incomplete ranking. Here $\mu^*(i) \leq j \leq (t-k) + \mu^*(i)$, if object i is ranked whereas $1 \leq j \leq t$, if object i is not ranked.

Proof. If an object i is ranked in an incomplete ranking μ^* of k objects, then the number of complete rankings compatible with μ^* which assign rank j to object i is

$$\binom{j-1}{\mu^*(i)-1}\binom{t-j}{k-\mu^*(i)}(t-k)!$$

This consists of the number of ways of picking a set of $(\mu^*(i)-1)$ from the first $(j-1)$ integers and a set of $(k-\mu^*(i))$ from the last $(t-j)$ integers while allowing all possible permutations of the $(t-k)$ integers not picked. On the other hand, if object i is not ranked in μ^* then the number of such complete compatible rankings is given by

$$\binom{t-1}{k}(t-k-1)!$$

the number of ways of picking k from the $t-1$ integers not equal to j and allowing all possible permutations of the remaining $(t-k-1)$ integers. Dividing these by $\frac{t!}{k!}$ the number of complete rankings compatible with μ^* gives the result. □

In the next lemma, we show that it is possible to compute the value of a score function corresponding to an incomplete ranking from knowledge of the compatibility class. To this end, we make use of the conditional distribution of a complete ranking given its compatibility class and the fact that the conditional expectation of the score function corresponds to its projection onto that class. We apply this approach to compute the form of score functions for both the Spearman and Kendall distances.

Lemma 3.3. *Suppose that we select a complete ranking μ at random from the class of compatible rankings $C(\mu^*)$. Suppose that object s is ranked. Then (a)*

$$E\left[\left(\mu(s)-\frac{t+1}{2}\right) \mid \mathcal{C}(\mu^*)\right]=\frac{t+1}{k+1}\left(\mu^*(s)-\frac{k+1}{2}\right), \tag{3.21}$$

and (b) for any pair of objects $i<j$,

$$E\left[sgn\left(\mu(j)-\mu(i)\right) \mid \mathcal{C}(\mu^*)\right]=a(i,j), \tag{3.22}$$

where

$$a(i,j)=\begin{cases} sgn(\mu^*(j)-\mu^*(i)) & \text{if both objects } i \text{ and } j \text{ are ranked} \\ 1-\frac{2\mu^*(i)}{(k+1)} & \text{if only object } i \text{ is ranked} \\ \frac{2\mu^*(j)}{(k+1)}-1 & \text{if only object } j \text{ is ranked} \\ 0 & \text{otherwise} \end{cases} \tag{3.23}$$

Proof. To prove (a), recall the identity

$$\sum_{j=l}^{t-k+l}\binom{j-1}{l-1}\binom{t-j}{k-l}=\binom{t}{l}. \tag{3.24}$$

Consequently, we have that

$$\begin{aligned} E\left[\left(\mu(s)-\frac{t+1}{2}\right) \mid \mathcal{C}(\mu^*)\right] &= \sum_{j=\mu^*(s)}^{t-k+\mu^*(s)}\left(j-\frac{t+1}{2}\right)\binom{j-1}{\mu^*(s)-1}\binom{t-j}{k-\mu^*(s)}\Big/\binom{t}{l} \\ &= \frac{t+1}{k+1}\left(\mu^*(s)-\frac{k+1}{2}\right). \end{aligned}$$

For the proof of (b), let

$$\delta(s,j)=\begin{cases} 1 & \text{if judge } j \text{ ranks object } s \\ 0 & \text{otherwise} \end{cases}$$

and define

$$\varpi_j(s)=\mu_j^*(s)\,\delta(s,j)+\left(\frac{k+1}{2}\right)(1-\delta(s,j)) \tag{3.25}$$

so that the incomplete ranking takes value $\frac{k+1}{2}$ when an object is unranked. Note that for any complete ranking,

$$\mu(j)=\frac{t+1}{2}+\frac{1}{2}\sum_{i=1}^{t}sgn\left(\mu(j)-\mu(i)\right). \tag{3.26}$$

It is clear that if objects i and j are both ranked, then $a(i, j)$ is as stated. Suppose now only object j is ranked. The adjusted score becomes on using (3.21)

$$E\left[\mu(j) \mid \mathcal{C}(\mu^*)\right] = \frac{t+1}{2} + \frac{1}{2}E\left[\sum_{i=1}^{t} sgn\left(\mu(j) - \mu(i)\right) \mid \mathcal{C}(\mu^*)\right]$$

$$\begin{aligned}\frac{t+1}{k+1}\mu^*(j) &= \frac{t+1}{2} + \frac{1}{2}\sum_{i=1}^{k} sgn\left(\mu^*(j) - \mu^*(i)\right) + \frac{(t-k)}{2}a(i,j) \\ &= \frac{t+1}{2} + \left(\mu^*(j) - \frac{k+1}{2}\right) + \frac{(t-k)}{2}a(i,j).\end{aligned}$$

Hence, $a(i, j) = \left(\frac{2\mu^*(j)}{k+1} - 1\right)$. The case where only object i is ranked is dealt with similarly. □

In describing visualization techniques for incomplete ranking data, Kidwell et al. (2008) have noted the efficiency for computing the Kendall scores in (3.23). Next, we proceed to find the maximum and minimum distances when only k objects are ranked among the incomplete rankings.

Lemma 3.4. *(a) For the Spearman distance,*

$$m_S^* = c_S - \frac{(t+1)^2}{12}\frac{k(k-1)}{(k+1)}, M_S^* = c_S + \frac{(t+1)^2}{12}\frac{k(k-1)}{(k+1)}$$

where $c_S = \frac{t\left(t^2-1\right)}{12}$.

(b) For the Kendall distance,

$$m_K^* = c_K - \frac{(2t+k+3)}{6}\frac{k(k-1)}{(k+1)}, M_K^* = c_K + \frac{(2t+k+3)}{6}\frac{k(k-1)}{(k+1)}$$

where $c_K = \frac{t(t-1)}{2}$. *It follows that the correlation between the incomplete rankings* μ_1^*, μ_2^* *can be defined to be*

$$\alpha\left(\mu_1^*, \mu_2^*\right) = 1 - \frac{2\left[d_K^*\left(\mu_i^*, \mu_j^*\right) - m^*\right]}{M^* - m^*}. \tag{3.27}$$

Proof. The right-hand side of (3.21) provides a general expression for an incomplete ranking. It follows that the Spearman distance between two incomplete rankings with the same number of ranked objects is

$$d_S^*\left(\mu_i^*, \mu_j^*\right) = \frac{t(t+1)(2t+1)}{6} - \left(\frac{t+1}{k+1}\right)^2 \sum_{s=1}^{t} \varpi_i(s)\,\varpi_j(s)$$

and in the Kendall case, the distance may be written as

$$d_K^*\left(\mu_i^*, \mu_j^*\right) = \frac{t(t-1)}{2} - \sum_{q_1<q_2} a_i\left(q_1, q_2\right) a_j\left(q_1, q_2\right)$$

where $a_i\left(q_1, q_2\right)$is defined as in (3.23) and $\varpi_i(s)$ is given in 3.25. An application of the Cauchy–Schwarz inequality indicates that the upper bound of the Spearman distance occurs when $\mathbf{T}_S C\left(\mu_i^*\right) = -\mathbf{T}_S C\left(\mu_j^*\right)$ whereas the lower bound is achieved when $\mathbf{T}_S C\left(\mu_i^*\right) = \mathbf{T}_S C\left(\mu_j^*\right)$. If we let μ_j^* be the inverted ranking, that is, $\mu_j^*(s) = k+1-\mu_i^*(s)$ when object s is ranked by i, then $\varpi_j(s) = k+1-\varpi_i(s)$ and $\mathbf{T}_S C\left(\mu_i^*\right) = -\mathbf{T}_S C\left(\mu_j^*\right)$. Furthermore, for the Kendall scores, $a_j\left(q_1, q_2\right) = -a_i\left(q_1, q_2\right)$ and thus $\mathbf{T}_K C\left(\mu_i^*\right) = -\mathbf{T}_K C\left(\mu_j^*\right)$. A straightforward calculation of these distances using the incomplete ranking $(1, 2, \ldots, k, -, -, \ldots, -)'$ and its inversion yields the minimum and maximum for each distance. □

We quote without proof a result in Alvo and Cabilio (1995a) which allows for different numbers of observations missing at random.

Lemma 3.5. *For fixed $k_1 \leq k_2$ suppose the pattern of missing observations is randomly selected from the set of all possible patterns. Then, for the Spearman and Kendall cases, the minimum and maximum values of the distance are of the form*

$$m^* = c - \gamma(i), \; M^* = c + \gamma(i)$$

where the $\gamma(i)$ are given as

$$\gamma_S(1) = \frac{(t+1)^2\left(k_1-1\right)\left(3k_2-k_1\right)}{24\left(k_2+1\right)}, \; k_1 \, odd$$

$$\gamma_S(2) = \frac{(t+1)^2 k_1\left(k_1\left(3k_2-k_1\right)-2\right)}{24\left(k_1+1\right)\left(k_2+1\right)}, \; k_1 \, even$$

$$\gamma_K(1) = \frac{\left(k_1-1\right)\left(t\left(3k_2-k_1\right)+k_2\left(k_1+3\right)\right)}{6\left(k_2+1\right)}, \; k_1 \, odd$$

$$\gamma_K(2) = \frac{k_1\left(3k_1k_2(t+1)-\left(k_1^2+2\right)(t-k_2)-3\left(k_2+1\right)\right)}{6\left(k_1+1\right)\left(k_2+1\right)}, \; k_1 \, even$$

Consider now two independent rankings of length k_1, k_2, respectively, with $2 \leq k_1 \leq k_2 \leq t$. It follows from (3.6) and Lemma 3.3 that

$$\mathcal{A}_S^*(\mu^*, \nu^*) = E\left[\mathcal{A}_S(\mu, \nu) \mid \mathcal{C}(\mu^*), \mathcal{C}(\nu^*)\right] \tag{3.28}$$

$$= \frac{(t+1)^2}{\left(k_1+1\right)\left(k_2+1\right)} \sum_{s=1}^{t}\left(\mu^*(s) - \frac{k_2+1}{2}\right)\left(\nu^*(s) - \frac{k_1+1}{2}\right)\delta\left(s, \mu^*\right)\delta\left(s, \nu^*\right)$$

$$= \frac{(t+1)^2}{\left(k_1+1\right)\left(k_2+1\right)} \sum_{i=1}^{k^*}\left(o_i - \frac{k_2+1}{2}\right)\left(\mu^*\left(o_i\right) - \frac{k_1+1}{2}\right) \tag{3.29}$$

Table 3.3 Language and arithmetic scores revisited

Student	1	2	3	4	5	6	7	8	9
Arithmetic (2)	14	18	23	26	27	30	40	–	–
Language (1)	28	14	46	–	53	–	54	50	–
Ranking (2)	1	2	3	4	5	6	7	–	–
Ranking (1)	2	1	3	–	5	–	6	4	–

where k^* is the number of objects ranked in ranking 1 among the k_2 objects ranked in ranking 2 and o_i is the label of the ith object ranked in ranking 1. Here, $\delta(s, \mu^*)$ takes value 1 if object s is ranked by μ^* and value 0 otherwise. Note that

$$o_i = i + l_i,$$

where l_i = number of objects unranked in ranking 1 which are to the left of the object being ranked. Similarly from (3.7) we have that

$$\mathcal{A}_K^*(\mu^*, \nu^*) = E\left[\mathcal{A}_K(\mu, \nu) \mid \mathcal{C}(\mu^*), \mathcal{C}(\nu^*)\right] \tag{3.30}$$

$$= \sum_{i<j} a_1(i, j)\, a_2(i, j)\,. \tag{3.31}$$

Example 3.5. Consider the test scores in Language (ranking 1) and Arithmetic (ranking 2) of a group of nine students in Table 3.3. The original data was altered by removing certain values, with the remaining observations reordered and ranked as follows.

Here $t = 9, k_1 = 6, k_2 = 7, k^* = 5, o_1 = 1, o_2 = 2, o_3 = 3, o_4 = 5, o_5 = 7, o_6 = 8$, and $l_1 = l_2 = l_3 = 0, l_4 = 1, l_5 = l_6 = 2$. Further,

$$\mu^*(o_1) = 2, \mu^*(o_2) = 1, \mu^*(o_3) = 3, \mu^*(o_4) = 5, \mu^*(o_5) = 6, \mu^*(o_6) = 4.$$

Hence $A_S^* = 33.9286$ and $A_K^* = 4$.

3.4.1 Asymptotic Normality of the Spearman and Kendall Test Statistics

The main objective of this section is to demonstrate the asymptotic normality of the similarity measures due to Spearman and Kendall in the case of incomplete rankings. Specifically, we shall be concerned with the asymptotic distributions of both $\mathcal{A}_S^*$,$\mathcal{A}_K^*$ under each of two possible null hypotheses H_1 and H_2. For both hypotheses we assume that k_1, k_2, the number of ranked observations, are fixed and the rankings for which we have (possibly) incomplete data are uniformly distributed over the $t!$ permutations of $(1, 2, \ldots, t)$.

- Under hypothesis H_1, we assume that the pattern of missing observations is fixed, so that all inference in this case is conditional on such a pattern.
- Under H_2, we assume that the patterns of missing observations are randomly selected from the set of all possible patterns. The latter situation would arise in practice if unranked objects occur by chance. An example would be testing for trend in water quality data when the historical data is incomplete.

We begin with the definition of a linear rank statistic.

Definition 3.3. Let $\{a(i)\}$ and $\{c(i)\}$ be two sets of constants. A statistic of the form

$$S = \Sigma_{i=1}^{N} c(i)\, a(R_i)$$

where $R = (R_1, \ldots, R_N)$ is a vector of ranks is called a linear rank statistic. The constants $a(i)$ are called scores whereas the $c(i)$ are called regression coefficients.

Many test statistics are of this form. For example, suppose that we have a random sample of n observations from a population and N-n from another. We are interested in testing the null hypothesis that the two populations are the same against the alternative that they differ only in location. Rank all N observations together. The Wilcoxon statistic then considers only the ranks of one of the populations by choosing

$$c(i) = \begin{cases} 0 & i = 1, \ldots, n \\ 1 & i = n+1, \ldots, N. \end{cases}$$

Lemma 3.6. *Suppose that R is uniformly distributed over the set of permutations in $\mathcal{P}$. Then*

(i) for $i = 1, \ldots, N$, $E(R_i) = \frac{N+1}{2}$, $Var(R_i) = \frac{(N^2-1)}{12}$ *and for* $i \neq j$, $Cov(R_i, R_j) = -\frac{N+1}{12}$ *and*

(ii)

$$ES = N\bar{c}\bar{a}$$

and

$$Var\, S = \frac{1}{N-1} \Sigma (c(i) - \bar{c})^2 \Sigma (a(i) - \bar{a})^2$$

where $\bar{a}$ and $\bar{c}$ represent the corresponding means.

Proof. The proof of this lemma is given in (Hájek and Sidak 1967). □

The following theorem states that under certain conditions, linear rank statistics are asymptotically normally distributed. We shall consider square integrable

functions ϕ defined on $(0, 1)$ which have the property that they can be written as the difference of two nondecreasing functions and satisfy

$$0 < \int_0^1 \left[\phi(u) - \bar{\phi}\right]^2 du < \infty$$

where $\bar{\phi} = \int_0^1 \phi(u)\, du$.

Theorem 3.1. *Suppose that R is uniformly distributed over the set of permutations in $\mathcal{P}$. Let the score function be given by $a(i) = \phi\left(\frac{i}{N}\right)$ where $\phi()$ is a square integrable score function. Then S is asymptotically normally distributed as $N \to \infty$ with mean $N\bar{c}\bar{a}$ and variance*

$$Var\, S = \frac{1}{N-1} \Sigma_{i=1}^N (c(i) - \bar{c})^2 \, \Sigma_{i=1}^N (a(i) - \bar{a})^2$$

provided

$$\frac{\sum_{i=1}^N (c(i) - \bar{c})^2}{\max_{1 \le i \le N} (c(i) - \bar{c})^2} \to \infty.$$

Proof. The proof of this important result is given in (Hájek and Sidak 1967). □

We may now apply Theorem 3.1 to obtain the asymptotic normality of the Spearman test statistic in the case of incomplete rankings under Hypothesis 1 wherein the pattern of missing data is fixed. Set

$$\sigma_S^2 = \frac{1}{12}\left[\frac{(t+1)^2}{(k_2+1)}\right]^2 \sum_{i=1}^{k_1} \left(o_i^* - \bar{o}_1\right)^2, \tag{3.32}$$

where

$$o_i^* = \begin{cases} o_i & \text{if } 1 \le i \le k^* \\ \frac{k_2+1}{2} & \text{if } k^*+1 \le i \le k_1 \end{cases} \tag{3.33}$$

and $\bar{o}_1 = \left(\sum_{i=1}^{k_1} o_i^*\right)/k_1$. Also set $\bar{o}^* = \left(\sum_{i=1}^{k^*} o_i\right)/k^*$.

Theorem 3.2. *Assume that $k^* \to \infty$ (and hence $k_1 \to \infty, k_2 \to \infty, t \to \infty$) with $k^*/t \to \lambda > 0$, where λ is a finite constant. Then, under H_1, whereby the pattern of missing data is fixed, $\mathcal{A}_S^*$ given in (3.28) is asymptotically normal with mean 0 and variance σ_S^2.*

Proof. The proof hinges on the fact that $\mathcal{A}_S^*$ is a linear rank statistic. In fact

$$\mathcal{A}_S^* = \frac{(t+1)^2}{(k_1+1)(k_2+1)} \sum_{i=1}^{k_1} \left(o_i^* - \frac{k_2+1}{2}\right)\left(\mu^*(o_i) - \frac{k_1+1}{2}\right)$$
$$= \frac{(t+1)^2}{(k_1+1)(k_2+1)} \sum_{i=1}^{k_1} \left(o_i^* - \bar{o}_1\right)\left(\mu^*(o_i)\right).$$

The normality follows provided

$$\frac{\sum_{i=1}^{k_1}\left(o_i^* - \bar{o}_1\right)^2}{\max\left(o_i^* - \bar{o}_1\right)^2} \to \infty.$$

Now

$$\Sigma_{i=1}^{k_1}\left(o_i^* - \bar{o_1}\right)^2 = \Sigma_{i=1}^{k^*}\left(o_i^* - \bar{o_1}\right)^2 + k^*\left(\bar{o}^* - \bar{o}_1\right)^2 + \left(k_1 - k^*\right)\left(\frac{k_2+1}{2} - \bar{o}_1\right)^2$$
$$\geq k^*\left(k^{*2} - 1\right)/12.$$

Further, $\left(o_i^* - \bar{o}_1\right)^2 \leq (t-1)^2$, so that the result follows on letting $k^* \to \infty$ with $k^*/t \to \lambda$. □

The exact variance of $\mathcal{A}_S^*$ under H_1, which is recommended in applications of Theorem 3.2, is related to σ_S^2 by

$$Var(\mathcal{A}_S^*) = \frac{k_1}{k_1+1}\sigma_S^2$$

(Lehmann 1975 (A. 49) p. 334). That is, the asymptotic variance given in the theorem is essentially the actual variance of $\mathcal{A}_S^*$. In any application, the calculation of the variance of $\mathcal{A}_S^*$ is a straightforward computation. Next, we consider the asymptotic distribution of $\mathcal{A}_S^*$ and $\mathcal{A}_K^*$ when the pattern of missing observations is random.

Theorem 3.3. *Let $k_1 \to \infty$ (and hence $k_2 \to \infty, t \to \infty$) with $k_1/t \to \lambda > 0$, where λ is a finite constant. Then, under H_2, whereby the pattern of missing data is random, $\mathcal{A}_S^*$ is asymptotically normal with mean 0 and variance*

$$Var\left(\mathcal{A}_S^*\right) = \frac{(t+1)^4}{144(t-1)}\kappa_1\kappa_2, \tag{3.34}$$

with

$$\kappa_i = \frac{k_i(k_i-1)}{(k_i+1)}, i = 1, 2.$$

Proof. Define $\mathbf{U} = (U_1, U_2, \ldots, U_t)$ as the random vector uniformly distributed over the permutations of $(1, 2, \ldots, k_1, \frac{k_1+1}{2}, \ldots, \frac{k_1+1}{2})$. In this case, the extended Spearman distance may be written as

$$d_S^* = \frac{t(t+1)(2t+1)}{6} - \mathcal{A}_S^* \tag{3.35}$$

$$= \frac{(t+1)^2}{(k_1+1)(k_2+1)} \left[\sum_{i=1}^{k_2} i U_i + \frac{k_2+1}{2} \sum_{i=k_2+1}^{t} U_i \right]. \tag{3.36}$$

The result follows from the combinatorial central limit theorem of Hoeffding (see Appendix B.1) applied to the quantity within square brackets above. □

Theorem 3.4. *$\mathcal{A}_K^*$ is asymptotically equivalent to $\mathcal{A}_S^*$ under both hypotheses H_1 and H_2. Hence, $\mathcal{A}_K^*$ is asymptotically normal with mean 0 and variance $\left(\frac{16}{t^2}\right) Var\left(\mathcal{A}_S^*\right)$.*

Proof. We know from (Hájek and Sidak 1967) that for the complete case

$$E\left(\mathcal{A}_K - \frac{4}{t}\mathcal{A}_S\right)^2 = \frac{(t-1)(t-2)}{18}$$

and that, moreover,

$$\frac{12\mathcal{A}_S}{t(t+1)\sqrt{t-1}} \Rightarrow N(0,1) \text{ as } t \to \infty.$$

Consequently, we have

$$\frac{6\mathcal{A}_K}{\sqrt{2t(t-1)(2t+5)}} \Rightarrow N(0,1).$$

From Jensen's inequality

$$E\left(\mathcal{A}_K^* - \frac{4}{t}\mathcal{A}_S^*\right)^2 = E\left(E^2\left(\left(\mathcal{A}_K - \frac{4}{t}\mathcal{A}_S\right)|\mathcal{C}(\mu^*), \mathcal{C}(\nu^*)\right)\right)$$

$$\leq E\left(E\left(\mathcal{A}_K - \frac{4}{t}\mathcal{A}_S\right)^2 |\mathcal{C}(\mu^*), \mathcal{C}(\nu^*)\right) = O\left(t^2\right)$$

and consequently the asymptotic normality of $\mathcal{A}_S^*$ will imply the asymptotic normality of $\mathcal{A}_K^*$. □

Example 3.6. We return to Example 3.2 wherein we wish to test the hypothesis of independence against the alternative of a positive correlation. For the complete data, the value of $\mathcal{A}_S$ is 41, and from the tables, under the randomness hypothesis,

$P(\mathcal{A}_S \geq 41) = 0.0252$, whereas the use of the asymptotic result gives a p-value of $1 - \Phi(1.9328) = 0.0266$, where Φ is the cumulative distribution function of a standard normal. For the data in Example 3.5, the value of $\mathcal{A}_S^*$ for the reduced data is calculated to be 33.9286. An application of the theorem yields that under H_1, the p-value is $P\left(\mathcal{A}_S^* \geq 33.9286\right) = 0.0178$. On the other hand, if all observations with missing values are deleted, we obtain a reduced value of $\mathcal{A}_S = 9$ with $t = 5$, and from the tables $P(\mathcal{A}_S \geq 9) = 0.0417$.

3.4.2 Asymptotic Efficiency

We now turn to the question of the efficiency which is further discussed in Appendix B.4. Let $X_1, X_2, \ldots, X_t$ be independent random variables whose joint density under the alternative is described by

$$q_d = \prod_{i=1}^{t} f_0\left(x_i - d_i\right)$$

where f_0 is a known density having finite Fisher information $I\left(f_0\right)$ and $\mathbf{d} = (d_1, d_2, \ldots, d_t)$ is an arbitrary vector. In the notation of our tests, $k_2 = t$, and write $k_1 = k$, the actual number of X_i's observed. Recalling that o_i is the label of the ith object ranked, the Spearman test which deletes all missing observations is based on the Spearman correlation of the reduced sample of k pairs, and the test statistic may be written as

$$A_{RS} = (t+1) \sum_{i=1}^{k} \left(i - \frac{k+1}{2}\right) \left(\frac{\mu^*\left(o_i\right)}{t+1}\right).$$

Since $k = k_1 = k^*$ and consequently $o_i = o_i^*$, the statistic $\mathcal{A}_S^*$ may be written as

$$\mathcal{A}_S^* = \frac{(t+1)}{(k+1)} \sum_{i=1}^{k} \left(o_i - \frac{t+1}{2}\right) \left(\mu^*\left(o_i\right) - \frac{k+1}{2}\right).$$

Hence,

$$\mathcal{A}_S^* = \frac{(t+1)}{(k+1)} \left\{ A_{RS} + \sum_{i=1}^{k} \left(\mu^*\left(o_i\right) - \frac{k+1}{2}\right)\left(o_i - i\right) \right\}.$$

The weight $(o_i - i)$ represents the number of time points to the left of o_i for which there are no observations. Similarly,

$$\mathcal{A}_K^* = A_{RK} + \frac{4}{k+1} \sum_{i=1}^{k} \left(\mu^*\left(o_i\right) - \frac{k+1}{2}\right)\left(o_i - i\right)$$

where

$$A_{RK} = \sum_{i<j}^{k} sgn\left(\mu^*\left(o_j\right) - \mu^*\left(o_i\right)\right).$$

Set $d_i^* = d_{o_i}$ and $\overline{d} = \sum_{i=1}^{t} d_i/t$. Under the alternative q_d, provided

$$\max_{1\le i\le t}\left(d_i - \overline{d}\right)^2 \to 0 \text{ and } I\left(f_0\right)\sum_{i=1}^{t}\left(d_i - \overline{d}\right)^2 \to b^2, 0 < b^2 < \infty,$$

both A_{RS} and A_S^* are asymptotically normal with means and variances given respectively by $\left(\mu_R, \sigma_{RS}^2\right)$ and $\left(\mu_S, \sigma_S^2\right)$, where

$$\mu_{RS} = (t+1)\sum_{i=1}^{k}\left(i - \frac{k+1}{2}\right)\left(d_i^* - \overline{d}\right)\int_0^1 u\,\phi\left(u, f_0\right)du$$

$$\mu_S = \frac{(t+1)^2}{(k+1)}\sum_{i=1}^{k}\left(o_i - \overline{o}\right)\left(d_i^* - \overline{d}\right)\int_0^1 u\,\phi\left(u, f_0\right)du.$$

$$\sigma_{RS}^2 = \frac{(t+1)^2}{12}\sum_{i=1}^{k}\left(i - \frac{k+1}{2}\right)^2,\ \sigma_S^2 = \frac{(t+1)^4}{12\left(k+1\right)^2}\sum_{i=1}^{k}\left(o_i - \overline{o}\right)^2.$$

Here $\phi\left(u, f\right) = \left[f'\left(F^{-1}\left(u\right)\right)\right]/\left[f\left(F^{-1}\left(u\right)\right)\right], 0 < u < 1$, and F is the cumulative distribution of f.

Shifting now to the efficiencies, it is seen that the asymptotic efficiencies as $k \to \infty$, for A_{RS} and $\mathcal{A}_S^*$ are respectively given by

$$e_{RS} = \lim \frac{\left[\sum_{i=1}^{k}\left(i - \frac{k+1}{2}\right)\left(d_i^* - \overline{d}\right)\right]^2}{\sum_{i=1}^{k}\left(i - \frac{k+1}{2}\right)^2\sum_{i=1}^{t}\left(d_i - \overline{d}\right)^2}Q_1$$

$$e_S = \lim \frac{\left[\sum_{i=1}^{k}\left(o_i - \overline{o}\right)\left(d_i^* - \overline{d}\right)\right]^2}{\sum_{i=1}^{k}\left(o_i - \overline{o}\right)^2\sum_{i=1}^{t}\left(d_i - \overline{d}\right)^2}Q_1,$$

where Q_1 is a positive function of f_0 and the limit is taken as $t \to \infty, k \to \infty$, with $k/t \to \lambda > 0$. The asymptotic relative efficiency of $\mathcal{A}_S^*$ relative to A_{RS} is then given by the ratio e_S/e_{RS} (Appendix B.4).

Now consider the case where $d_i^* = o_i, \bar{d} = \bar{o}, i = 1, \ldots, k$ and the remaining d_i are arbitrary, a situation which includes alternatives of the form $EX_i = \beta_0 + \beta i, \beta > 0$. It can be shown that irrespective of the density f_0, the asymptotic relative efficiency of $\mathcal{A}_S^*$ relative to A_{RS} is given by

$$ARE\left(\mathcal{A}_S^*, A_{RS}\right) = \lim_{k\to\infty} R\left(k, \mathbf{o_k}\right),$$

where $\mathbf{o_k} = (o_1, \ldots, o_k)$ and

$$R(k, \mathbf{o_k}) = \frac{\sum_{i=1}^{k} \left(i - \frac{k+1}{2}\right)^2 \sum_{i=1}^{k} (o_i - \overline{o})^2}{\left[\sum_{i=1}^{k} \left(i - \frac{k+1}{2}\right)(o_i - \overline{o})\right]^2} \geq 1.$$

Note that $R(k, \mathbf{o_k}) > 1$ unless the o_i's are equally spaced.

In order to illustrate the magnitude of this efficiency, suppose for example that $t = 19, k = 7, o_1 = 1, o_2 = 2, o_3 = 3, o_4 = 10, o_5 = 17, o_6 = 18, o_7 = 19$, then the ratio of the efficacies of A_S^* to A_{RS} is 1.086. On the other hand, if $o_1 = 1$, $o_2 = 8, o_3 = 9, o_4 = 10, o_5 = 11, o_6 = 12, o_7 = 19$, then that ratio is 1.176.

3.5 Tied Rankings and the Notion of Compatibility

The notion of compatibility may also be extended to deal with tied rankings. As an example, suppose that objects 1 and 2 are equally preferred whereas object 3 is least preferred. Such a ranking would be compatible with the rankings $(1, 2, 3)$ and $(2, 1, 3)$ in that both are plausible. The average of the rankings in the compatibility class, which as we shall see results from the use of the Spearman distance, will then be the ranking

$$\frac{1}{2}[(1, 2, 3) + (2, 1, 3)] = (1.5, 1.5, 3)$$

to be presented in this case. It is seen that the notion of compatibility serves to justify the use of the midrank when ties exist. Formally we can define tied orderings as follows.

Definition 3.4. A tied ordering of t objects is a partition into e sets, $1 \leq e \leq t$, each containing 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 $\delta = (d_1, d_2, \ldots, d_e)$. The ranking denoted by $\mu_\delta = (\mu_\delta(1), \ldots, \mu_\delta(t))$ resulting from such an ordering is a tied ranking and is one of $\frac{t!}{d_1! d_2! \ldots d_e!}$ possible permutations.

Associated with every tied ranking we may define a $t! \times (\frac{t!}{d_1! d_2! \ldots d_e!})$ matrix of compatibility D_δ. Yu et al. (2002) considered the problem of testing for independence between two random variables when the tie patterns and the pattern of missing observations are fixed. Specifically, let μ^* be an incomplete ranking of k_1 out of t objects with tie pattern $\delta_1 = (d_{11}, \ldots, d_{1e_1})$. Similarly, let ν^* be an incomplete ranking of k_2 out of t objects with tie pattern $\delta_2 = (d_{21}, \ldots, d_{2e_2})$. The Spearman similarity measure between two incomplete rankings μ^*, ν^* is defined to be

$$A_S^* = \frac{(t+1)^2}{(k_1+1)(k_2+1)} \sum_{j=1}^{t} \delta(j) \left[\mu^*(j) - \frac{k_1+1}{2}\right] \left[\nu^*(j) - \frac{k_2+1}{2}\right]$$

where $\delta(j) = 1$ if both rankings of object j are not missing and 0 otherwise.

Table 3.4 Data from the public opinion survey

	Response						
Education level	1	2	3	4	5	Missing	Subtotal
Primary or below	2	35	23	7	3	33	103
Secondary	2	72	129	37	6	53	299
Matriculated	0	9	9	7	0	3	28
Tertiary, nondegree	1	9	6	6	0	5	27
Tertiary, degree	0	22	28	7	6	6	69
Missing	0	2	3	0	0	1	6
Subtotal	5	149	198	64	15	101	532

Theorem 3.5. *Let k^* be the number of objects ranked in ranking 1 among the k_2 objects ranked in ranking 2. Let $2 \le k_1 \le k_2 \le t$. Assume that*

(i) $k^* \to \infty$, *(and hence* $k_1 \to \infty, k_2 \to \infty, t \to \infty$*) with* $k^*/t \to \lambda > 0$.
(ii) $\max_{j=1,\cdots,e_1} \frac{g_{1j}}{k^*}$ *is bounded away from 1.*
(iii) $\max_{j=1,\cdots,e_2} \frac{g_{2j}}{k^*}$ *is bounded away from 1.*

Then, under the null hypothesis of independence whereby the pattern of ties and missing data is fixed, A_S^* is asymptotically normal with mean 0 *and exact variance*

$$Var\left(A_S^*\right) = \left[\frac{(t+1)^2 k_1}{(k_1+1)(k_2+1)}\right]^2 \frac{\sum_{j=1}^{k_1}\left(o_j^* - \bar{o}\right)^2}{12}\left\{1 - \frac{\sum_{j=1}^{e_1}\left(g_{1j}^3 - g_{1j}\right)}{k_1^3 - k_1}\right\}.$$

Proof. See Yu et al. (2002). □

Example 3.7. In a public opinion survey held in 1999 in Hong Kong, it was of interest to determine whether the education level of the respondents is related to the level of dissatisfaction of the Policy Address of the Chief Executive of the Hong Kong Special Administrative Region. The response is an ordinal variable having seven options as follows: (1), very satisfied; (2), satisfied; (3), neutral; (4), unsatisfied; (5), very unsatisfied; (6), not sure; and (7), refuse to answer. Options (6) and (7) were combined and listed as "missing." Table 3.4 displays the frequencies of the respondents listed by option and by education level.

It is noted that about 19.9 % of the respondents did not respond either to one or to both questions. Moreover, since the education levels are grouped into a few categories, the problem of ties cannot be ignored. One alternative approach for analyzing this data is as a contingency table. In that case, however, the ordering among the education levels and separately among the responses would not be taken into account. The results of the analysis shown in Table 3.5 reveal that at the 5 % significance level, the test based on the reduced sample (which discards all observations with at least one missing variable) cannot reject the hypothesis of

Table 3.5 Results of the analyses

Test	Statistic	Standardized statistic	p-value
Reduced sample	494,132.0	1.9075	0.0564
Complete sample	786,633.2	1.9690	0.0490

Table 3.6 Wind direction in degrees

6 a.m.	356	97	211	262	343	292	157	302	324	85	324
Noon	119	162	221	259	270	29	97	292	40	313	94
6 a.m.	85	324	340	157	238	254	146	232	122	329	
Noon	45	47	108	221	248	270	45	23	270	119	
Data replaced by their ranks											
6 a.m.	21	3	8	12	20	13	6.5	14	16	1.5	16
Noon	10.5	12	13.5	16	18	2	8	20	3	21	7
6 a.m.	1.5	16	19	6.5	10	11	5	9	4	18	
Noon	4.5	6	9	13.5	15	18	4.5	1	18	10.5	

independence whereas the one based on the complete sample can. Since the test statistic is positive, this implies that there is a positive association between education level and level of dissatisfaction. More highly educated respondents tend to be less satisfied with the Policy Address. The analysis by means of a contingency table whereby the missing categories for education and response were dropped leads to a chi-square statistic with a value of 35.2161 on 16 degrees of freedom and a p-value of 0.0037.

3.6 Angular Correlations

There has been a great deal of interest in directional statistics in the literature. Consider the following example on wind directions whereby we are interested in testing for independence between the 6 a.m. and the noon readings. The data shown in Table 3.6 can be viewed as points on the unit circle and cannot be dealt with by simply computing the usual rank correlation. The reason is that the larger ranks are close to the smaller ranks. Hence, for example, for the noon readings, angle 23 is closer to angle 313 than to angle 248. Yet, the ranks imply an opposite interpretation. In the table, tied ranks were replaced by their midranks.

Example 3.8 (Johnson and Wehrly 1977). Wind directions were recorded at 6 a.m. and at 12 noon on each day at a weather station for 21 consecutive days. It is desired to test for independence. Tied rankings were replaced by their midranks (Table 3.6).

Excellent review articles along with additional references are given by Mardia (1975, 1976) and Jupp and Mardia (1989). Typically, data is provided in the form

of directions either in two- or three-dimensional space or as rotations in such a space. The data may take on a variety of forms. It may consist of a unit vector of directions, pairs of such vectors, or a vector of directions along with a corresponding random variable on the line. Examples of applications are to be found in the fields of astronomy, biology, geology, medicine, and meteorology (Downs 1973; Johnson and Wehrly 1977; Breckling 1989). A large number of the works presented deal with the study of inference from parametric models. In this section, we define a corresponding notion of angular correlation using the ranks of the data.

Let X and Y be random vectors with covariance matrix Σ partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

and suppose Σ_{11} and Σ_{22} are non-singular of ranks p and q, respectively.

Definition 3.5 (Jupp and Mardia 1989). The correlation coefficient γ_{XY} between X and Y is defined to be the trace γ of the matrix

$$\gamma_{XY} = Tr[\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}].$$

It follows that, $\gamma_{XY} = \sum_{i=1}^{s} \lambda_i^2$ where the λ_i are the canonical correlations and $s = min(p, q)$. This coefficient satisfies the property of invariance under rotation and reflection in addition to the usual properties of a correlation.

Suppose now that θ and φ are circular variables with $0 \leq \theta, \varphi \leq 2\pi$. Define the directional vectors $t_1'(\theta) = (\cos\theta, \sin\theta)$, $t_2'(\varphi) = (\cos\varphi, \sin\varphi)$, and let Σ be the covariance matrix of t_1 and t_2. It is seen that

$$\begin{aligned} \gamma_{\theta\varphi} = [&\rho_{cc}^2 + \rho_{cs}^2 + \rho_{sc}^2 + \rho_{ss}^2 + 2(\rho_{cc}\rho_{ss} - \rho_{cs}\rho_{sc})\rho_1\rho_2 - 2(\rho_{cc}\rho_{cs} \\ &+\rho_{sc}\rho_{ss})\rho_1 - 2(\rho_{cc}\rho_{sc} + \rho_{cs}\rho_{ss})\rho_2]/[(1-\rho_1^2)(1-\rho_2^2)]. \end{aligned} \tag{3.37}$$

where $\rho_{cc} = corr(\cos\theta, \cos\varphi)$, $\rho_{cs} = corr(\cos\theta, \sin\varphi)$, etc., and $\rho_1 = corr(\cos\theta, \sin\theta)$, $\rho_2 = corr(\cos\varphi, \sin\varphi)$.

Let (θ_i, φ_i) for $i = 1, \ldots, n$ be a random sample of n pairs of angles which define points on the unit circle. Without loss in generality assume that the ranks of the θ's are the natural integers 1,...,n whereas the corresponding ranks of the φ's are denoted by $R_1, \ldots, R_n$. Let

$$\eta^{(1)} = (\cos\frac{2\pi}{n}, \cos\frac{4\pi}{n}, \ldots, \cos 2\pi)', \eta^{(2)} = (\sin\frac{2\pi}{n}, \sin\frac{4\pi}{n}, \ldots, \sin 2\pi)'$$

$$\nu^{(1)} = (\cos\frac{2\pi R_1}{n}, \cos\frac{2\pi R_2}{n}, \ldots, \cos\frac{2\pi R_n}{n})', \nu^{(2)} = (\sin\frac{2\pi R_1}{n}, \sin\frac{2\pi R_2}{n}, \ldots, \sin\frac{2\pi R_n}{n})'.$$

We may formally construct on the basis of the sample the matrix of pairwise correlations

$$\Upsilon_{12} = \begin{pmatrix} \rho(\eta^{(1)}, \nu^{(1)}) & \rho(\eta^{(1)}, \nu^{(2)}) \\ \rho(\eta^{(2)}, \nu^{(1)}) & \rho(\eta^{(2)}, \nu^{(2)}) \end{pmatrix}$$

where $\rho(\eta, \nu)$ is a measure of correlation between η and ν. We shall consider correlations based on the Spearman and Kendall distance functions in subsequent sections and we will determine the corresponding asymptotic distributions of the correlation coefficients as $n \to \infty$.

3.6.1 Spearman Distance

We shall consider the Kendall notion of a type b correlation (Kendall and Gibbons 1990) given by

$$\begin{aligned} \rho_S(\eta, \nu) &= \frac{\sum_{i \neq j} (\eta_i - \eta_j)(\nu_i - \nu_j)}{\sqrt{\sum_{i \neq j} (\eta_i - \eta_j)^2 \sum_{i \neq j} (\nu_i - \nu_j)^2}} \\ &= \frac{2}{n} \eta' \nu. \end{aligned}$$

It is straightforward to show

$$\sum_{i=1}^{n} \cos \frac{2\pi i}{n} = \sum_{i=1}^{n} \sin \frac{2\pi i}{n} = \sum_{i=1}^{n} \cos \frac{2\pi i}{n} \sin \frac{2\pi i}{n} = 0$$

and

$$\sum_{i=1}^{n} \cos^2 \frac{2\pi i}{n} = \sum_{i=1}^{n} \sin^2 \frac{2\pi i}{n} = \frac{n}{2}.$$

It follows that $\Sigma_{11} = \Sigma_{22} = \frac{n}{2} I$. The sample estimate of Σ_{12} is given by

$$\Upsilon_{12}^{S} = \frac{2}{n} \begin{pmatrix} T_{cc} & T_{cs} \\ T_{sc} & T_{ss} \end{pmatrix}$$

where $T_{cc} = \eta^{(1)'} \nu^{(1)}, T_{cs} = \eta^{(1)'} \nu^{(2)}, T_{sc} = \eta^{(2)'} \nu^{(1)}, T_{ss} = \eta^{(2)'} \nu^{(2)}$.

We recognize that the $\mathbf{T}'s$ are measures of correlation in the Spearman sense. Consequently, the sample correlation using Spearman distance becomes

$$\gamma_S = \frac{4}{n^2} \left(\mathbf{T}_{cc}^2 + \mathbf{T}_{ss}^2 + \mathbf{T}_{cs}^2 + \mathbf{T}_{sc}^2 \right).$$

3.6.2 Kendall Distance

Recalling the Kendall measure of distance defined by

$$d_K(\eta, \nu) = \sum_{i<j} \{1 - sgn(\eta_i - \eta_j) sgn(\nu_i - \nu_j)\}$$

where sgn indicates the sign function, we may define a corresponding type b correlation as

$$\begin{aligned}\rho_K(\eta, \nu) &= \frac{\sum_{i \neq j} sgn(\eta_i - \eta_j) sgn(\nu_i - \nu_j)}{\sqrt{\sum_{i \neq j} (sgn(\eta_i - \eta_j))^2} \sqrt{\sum_{i \neq j} (sgn(\nu_i - \nu_j))^2}} \\ &= \frac{\sum_{i \neq j} sgn(\eta_i - \eta_j) sgn(\nu_i - \nu_j)}{\sqrt{A(\eta) A(\nu)}},\end{aligned}$$

where $A(\eta) = \#\left(\text{pairs } (i, j), i \neq j \,|\eta_i \neq \eta_j\right)$. It is easy to see that Σ_{11} and Σ_{22} are diagonal matrices. In fact, the off-diagonal terms are equal to

$$\begin{aligned}&\sum_{i \neq j} sgn\left(\cos\frac{2\pi i}{n} - \cos\frac{2\pi j}{n}\right) sgn\left(\sin\frac{2\pi i}{n} - \sin\frac{2\pi j}{n}\right) \\ &= -4 \sum_{i \neq j} sgn\left(\sin\frac{\pi(i+j)}{n} \sin\frac{\pi(i-j)}{n}\right) sgn\left(\cos\frac{\pi(i+j)}{n} \sin\frac{\pi(i-j)}{n}\right) \\ &= -2 \sum_{i \neq j} sgn\left(\sin\frac{2\pi(i+j)}{n}\right) = 0.\end{aligned}$$

The normalization in the Kendall case is somewhat delicate and depends in part on the parity of n. For example, for $n = 10$, there are five pairs of equal values in the set $\left\{\sin\frac{2\pi i}{n}\right\}$ whereas for $n = 11$, all the values are distinct. In general, the number of equal pairs is at most $O(n)$. The sample estimate of Υ_{12} is given by

$$\Upsilon_{12}^K = \begin{pmatrix} K_{cc} & K_{cs} \\ K_{sc} & K_{ss} \end{pmatrix}$$

where $K_{cc} = \rho_K(\eta^{(1)}, \nu^{(1)})$, $K_{cs} = \rho_K(\eta^{(1)}, \nu^{(2)})$, $K_{sc} = \rho_K(\eta^{(2)}, \nu^{(1)})$, $K_{ss} = \rho_K(\eta^{(2)}, \nu^{(2)})$.

It follows that the sample correlation coefficient in the Kendall case is given by

$$\gamma_K = \left(K_{cc}^2 + K_{ss}^2 + K_{cs}^2 + K_{sc}^2\right).$$

In the following sections, we shall derive the asymptotic null distributions of the test statistics induced by the Spearman and Kendall distances.

3.6.3 Asymptotic Distributions

We are interested in testing the null hypothesis that the circular variables θ, φ are independent. In terms of the ranks, assuming no ties, this translates into the hypothesis H_0 that all permutations of the integers $1, \ldots, n$ are equally likely.

Theorem 3.6. *The asymptotic null distribution of $n\gamma_S$ as $n \to \infty$ is χ^2_4.*

Proof. The joint distribution of $T_{cc}, T_{ss}, T_{cs}, T_{sc}$ is asymptotically normal. In fact, for arbitrary $\{a_i\}$, consider the linear combination

$$a_1 T_{cc} + a_2 T_{ss} + a_3 T_{cs} + a_4 T_{sc} =$$

$$\sum_{i=1}^{n} [\cos \frac{2\pi R_i}{n} (a_1 \cos \frac{2\pi i}{n} + a_2 \sin \frac{2\pi i}{n}) + \sin \frac{2\pi R_i}{n} (a_3 \cos \frac{2\pi i}{n} + a_4 \sin \frac{2\pi i}{n})].$$

Let

$$d(i, j) = \cos \frac{2\pi i}{n} (a_1 \cos \frac{2\pi j}{n} + a_2 \sin \frac{2\pi j}{n}) + \sin \frac{2\pi i}{n} (a_3 \cos \frac{2\pi j}{n} + a_4 \sin \frac{2\pi j}{n}).$$

Since

$$max d_n^2(i, j) \leq 4(a_1^2 + a_2^2 + a_3^2 + a_4^2)$$

and the variance

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} d_n^2(i, j) = \frac{n}{4} (a_1^2 + a_2^2 + a_3^2 + a_4^2)$$

we have that

$$\frac{max\ d_n^2(i, j)}{\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} d_n^2(i, j)} \to 0\ as\ n \to \infty.$$

The result follows on using Hoeffding's combinatorial central limit theorem (see Appendix B.1). Hence Υ_{12}^S is multivariate normal and the theorem follows. □

A similar result holds for the Kendall tau statistic.

Theorem 3.7. *The asymptotic null distribution of $\frac{9}{4} n\gamma_K$ as $n \to \infty$ is χ^2_4.*

Proof. See Alvo (1998) for the proof. A different proof can make use of the asymptotic equivalence between the Kendall and Spearman coefficients in general. □

Example 3.9. We revisit the wind direction data. We calculate

$$\Upsilon_{12}^S = \begin{pmatrix} -0.246 & 0.306 \\ -0.376 & -0.452 \end{pmatrix}$$

and hence $n\gamma_S = 21(0.50047) = 10.51$ with a p-value of 0.0327. Consequently, we conclude that there is evidence that the 6 a.m. and noon wind directions are significantly correlated.

It is interesting to compare this result with the usual product moment correlation between the two angular measurements. The latter yields a value equal to -0.04, thereby implying that the variables are independent. On the other hand, restricting attention only to the pairs of measurements for which the 6 a.m. readings are below 180° the value of the product moment correlation is 0.512 while for pairs for which the 6 a.m. readings are above 180° it is -0.475. These results taken separately imply a fair degree of dependence. The test statistic γ_S takes into account the fact that very small and very large angles (mod 2π) are close to one another.

For the Kendall statistic, we may also calculate

$$\Upsilon_{12}^K = \begin{pmatrix} -0.1822 & 0.2097 \\ -0.3106 & -0.3637 \end{pmatrix}$$

and hence $\frac{9n}{4}\ \gamma_K = \frac{9(21)}{4}(0.3056) = 14.44$ with a p-value of 0.006. It is clear that with either the Spearman or the Kendall statistic, the hypothesis of independence is in doubt.

3.7 Angle-Linear Correlation

Suppose that we are now interested in defining the correlation between an angle θ and a real valued random variable X. It can be shown that the correlation coefficient in that case is given by

$$\gamma_L = [\rho_{xc}^2 + \rho_{xs}^2 - 2\rho_{xc}\rho_{xs}\rho_{cs}]/(1 - \rho_{cs}^2)$$

where

$$\rho_{xc} = corr(X, \cos\theta), \rho_{xs} = corr(X, \sin\theta), \rho_{cs} = corr(\cos\theta, \sin\theta).$$

In the nonparametric context, let (X_i, θ_i) for $i = 1, \ldots, n$ be a random sample of linear-angular measurements. Let $\{R_i\}$ be the ranks of the $\{X_i\}$ and let $\{S_i\}$ be the ranks of the $\{\theta_i\}$. We may assume without loss in generality that the S_i are in natural order $1, 2, \ldots, n$. Based on the Spearman measure of distance, the sample angular-linear correlation is defined by

$$\gamma_{LS} = \frac{[T_{xc}^2 + T_{xs}^2]}{\frac{n}{2}\left(\frac{n(n^2-1)}{12}\right)}$$

where $T_{xc} = \sum R_i \cos\left(\frac{2\pi i}{n}\right)$, $T_{xs} = \sum R_i \sin\left(\frac{2\pi i}{n}\right)$. Similarly, for the Kendall measure, the angular-linear correlation is then given by

$$\gamma_{LK} = [K_{xc}^2 + K_{xs}^2]$$

where

$$K_{xc} = \frac{\sum_{i \neq j}[sgn(R_i - R_j)sgn(\cos\left(\frac{2\pi i}{n}\right) - \cos\left(\frac{2\pi j}{n}\right))]}{\sqrt{[n(n-1)]}\sqrt{\sum_{i \neq j}(sgn(\eta_i^{(1)} - \eta_j^{(1)}))^2}}$$

$$K_{xs} = \frac{\sum_{i \neq j}[sgn(R_i - R_j)sgn(\sin\left(\frac{2\pi i}{n}\right) - \sin\left(\frac{2\pi j}{n}\right))]}{\sqrt{[n(n-1)]}\sqrt{\sum_{i \neq j}(sgn(\eta_i^{(2)} - \eta_j^{(2)}))^2}}.$$

We may now prove a theorem giving the asymptotic distributions of γ_{LS} and γ_{LK} under the null hypothesis that all vectors of ranks $(R_1, \ldots, R_n)$ are equally likely.

Theorem 3.8. *The asymptotic null distribution of $n\gamma_{LS}$ as $n \to \infty$ is χ_2^2.*

Proof. The joint distribution of T_{xc}, T_{xs} is asymptotically normal. In fact, for arbitrary constants a_1, a_2, consider the linear combination

$$a_1 T_{xc} + a_2 T_{xs} = \sum_{i=1}^{n}[R_i(a_1 \cos\frac{2\pi i}{n} + a_2 \sin\frac{2\pi i}{n}).$$

This is a linear rank statistic for which the conditions in Hoeffding (1951) are satisfied. In fact, let

$$d_n(i, j) = (i - \frac{n+1}{2})(a_1 \cos\frac{2\pi i}{n} + a_2 \sin\frac{2\pi i}{n}).$$

The variance is then equal to

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_n^2(i, j) = \frac{1}{4}(a_1^2 + a_2^2)\frac{n(n^2-1)}{12}$$

and we have that

$$\frac{max\ d_n^2(i, j)}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_n^2(i, j)} \to 0$$

as $n \to \infty$. The result follows. □

Theorem 3.9. *The asymptotic null distribution of $\frac{9n}{4}\gamma_{LK}$ as $n \to \infty$ is χ_2^2.*

Table 3.7 Wind direction and ozone concentration

Wind direction	327	91	88	305	344	270	67	21	281	
Ozone concentration	28.0	85.2	80.5	4.7	45.9	12.7	72.5	56.6	31.5	
Wind direction	8	204	86	333	18	57	6	11	27	84
Ozone concentration	112.0	20.0	72.5	16.0	45.9	32.6	56.6	52.6	91.8	55.2

Proof. For arbitrary constants a_1, a_2, consider the linear combination

$$\sum_{i \neq j}^{n} sgn(R_i - R_j)b_{ij}$$

where

$$b_{ij} = [(a_1 sgn(\cos\frac{2\pi i}{n} - \cos\frac{2\pi j}{n}) + a_2 sgn(\sin\frac{2\pi i}{n} - \sin\frac{2\pi j}{n})].$$

Using a result of Daniels (1950), the asymptotic normality of K_{xc} and K_{xs} follows. □

Example 3.10 (Johnson and Wehrly 1977). We consider data on wind direction and ozone concentration collected at a weather station for 19 days at 4-day intervals. The readings are given in Table 3.7.

The Spearman test statistic to be $n\gamma_{LS} = 19\,(0.3751) = 7.13$ which has a p-value equal to 0.0283. On the other hand the Kendall statistic is given by $\frac{9n}{4}\gamma_{LK} = \frac{9(19)}{4}\,(0.1595) = 6.82$ for a p-value of 0.033. Both statistics imply that there is a fair degree of dependence between wind direction and ozone concentration.

Chapter Notes

In this chapter, the traditional rank correlation has been extended to include incomplete rankings. This was made possible using the notion of compatibility which was developed by Alvo and Cabilio in a series of papers. Cabilio and Tilley (1999) report the results of a simulation study where they considered linear, quadratic, and square root trends. They observed that when there were no missing observations, the Spearman statistic was more powerful than Kendall's. In the incomplete case, however, the new Kendall statistic has superior power for more patterns.

The calculation of the exact variance of $\mathcal{A}_K^*$ under H_2, in Theorem 3.4, is more involved, and the reader is referred to Alvo and Cabilio (1992), where it is shown that

$$Var\left(\mathcal{A}_K^*\right) = \frac{\kappa_1\kappa_2}{9t\,(t-1)}\left[\frac{(2t+k_1+3)\,(2t+k_2+3)}{2} + \frac{\left(t^2-k_1-2\right)\left(t^2-k_2-2\right)}{(t-2)}\right].$$

An important application of the results presented above which do not discard missing data is in tests of trend where $k_2 = t$ and $k_1 < t$. It is seen that in this context, the superiority of the extended Spearman statistic is established through the calculation of its asymptotic relative efficiency relative to the "naive" statistic. (Alvo and Cabilio 1994) applied these methods to test for trend in precipitation data for St John and Fredericton (NB) and showed that the extended statistic based on Spearman distance is more sensitive in detecting trends than the statistic which ignores the missing observations. Tables of selected critical values of $\mathcal{A}_S^*$ and $\mathcal{A}_K^*$ for the trend case when $k \geq t/2$ have been developed for both hypotheses (Alvo and Cabilio 1993). The results of this section have been extended to the case of ties (Yu et al. 2002) and applied to deal with tests of independence in opinion surveys. A further extension to assess trend in proportions appears in Chap. 7.

Alvo and Smrz (2005) proposed an arc model which serves as a good approximation to Kendall distance.

Although not considered in this book, Alvo and Park (2002) were concerned with multivariate tests of trend when the data are partially incomplete. Such is the case in environmental studies when pH data for one or more lakes are often recorded over regular time intervals and examined for monotone increasing or decreasing trends in order to test for trend in acidification. In monitoring recovering patients, one looks for trends in their vital signs which are often multivariate data in nature. There may be as many as 20–30 blood constituents measured weekly over a period of several months or years. In those case, the use of separate tests on each constituent is inefficient.