National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Canada

The Use of Spearman's Footrule in Testing for Trend When The Data is Incomplete

By Martin Charbonneau

A M.Sc. Thesis

submitted to the School of Graduate Studies and Research

in partial fulfillment of the requirements for

the Master's degree in Mathematics*

University of Ottawa

Ottawa, Ontario

Canada

March 31 1994

*The M.Sc. Program is a joint program with

Carleton University, administered by the Ottawa-Carleton

Institute of Mathematics and Statistics

ISBN 0-612-00451-1

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

The Use of Spearman's Footrule in Testing for Trend When The Data is Incomplete

By Martin Charbonneau

Work supported in part by Natural Sciences and Engineering Council of Canada, Grant A-9068

AMS 1991 Subject Classifications. Primary 62G10; secondary 62E20.

Key Words and phrases. Rank correlation, Spearman footrule metric, incomplete rankings, missing observations, test for trend and independence.

The Use of Spearman's Footrule in Testing for Trend When The Data is Incomplete

By Martin Charbonneau

Rank correlation has been used in numerous applications in tests for trend and for independence. It is possible to extend the definition of rank correlation to situations when the data is incomplete. In this thesis, it is shown that in such situations, a test of trend can be constructed through the use of Spearman's footrule.

1.Introduction

The Spearman and Kendall rank correlation measures have been widely used in the literature to test for trend or for independence in situations when the data has been collected at regular intervals of time. Such tests are popular in environmental sciences because few assumptions are required for their application and they provide efficient ways of testing for trend. It has been the common practice in situations when the data is incomplete, to ignore the time gaps and to act as if the available data were collected at regular time intervals. Alvo and Cabilio (1993) proposed a new approach in such situations which takes into account the length of the time between successive observations and in preliminary results showed that this always leads to an increase in efficiency for Spearman's rho statistic when testing for trend. Recently, Diaconis and Graham (1977) proposed the Spearman footrule as a test for trend when the data is complete and studied its asymptotic properties, thereby adding to the test bank of possible procedures. In this article, we define a modification of the Spearman footrule when some data is missing and develop its asymptotic properties.

Let $\mathcal{P}$ represent the collection of all possible rankings of t objects which, for convenience, are labelled 1,...,t. Denote the t! possible permutations of the integers 1,...,t by the column vectors

$$\nu_j=(\nu_j(1), \ldots, \nu_j(t))' \ , \ j=1,\ldots,t!$$

The correlation between permutations μ and ν can be defined in terms of the distance $d(\mu,\nu)$ between them as :

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

where M is the maximum of the value of $d(\mu,\nu)$ taken over all possible pairs μ and ν in $\mathcal{P}$. Examples of metrics over permutations may be found in Critchlow (1985). These include the metrics associated with Spearman, Kendall and the Spearman footrule:

(1.2) $$d_S(\mu,\nu)=\tfrac{1}{2}\sum_{i=1}^{t}[\mu(i)-\nu(i)]^2$$

(1.3) $$d_K(\mu,\nu)=\sum_{i<j}\Big\{1-\mathrm{sgn}[\mu(i)-\mu(j)]\mathrm{sgn}[\nu(i)-\nu(j)]\Big\}.$$

(1.4) $$d_F(\mu,\nu)=\tfrac{1}{2}\sum_{i=1}^{t}\mid\mu(i)-\nu(i)\mid.$$

These metrics have the property of right invariance, which means that the distance between two rankings remains unchanged under any permutation relabeling of the objects. Denote by $\Delta=(d(\mu,\nu))$ the (t! x t!) matrix of values of the distance. For the metrics above, it has been shown in Feigin and Alvo (1986) that there is a matrix T and a constant c such that

(1.5) $$\Delta = cJ - T'T$$

where J is the (t! x t!) matrix of 1's. If π denotes a probability distribution over $\mathcal{P}$, that is, a t!-vector of probabilities, tests of hypothesis may then be formulated in terms of $T\pi$. In this thesis, we will concern ourselves with the study of the Spearman footrule.

The matrix T corresponding to d_F is given by:

(1.6) $$T_F = (t_F(\nu_1), \ldots, t_F(\nu_{t!}))$$

of dimension (t^2 x t!), where each $t_F(\nu)$ is a column vector of length t^2 defined as follows:

(1.7) $$(t_F(\nu))((i-1)t+j) = I_{[\nu(i)\le j]} - j/t$$

where I is the indicator function and $1\le i,j\le t$.

The characteristic $T_F\pi$ may be viewed as the set of centered distribution functions for the ranks of each item. If we write

$$\mathcal{F}_i(j) = P(\nu(i)\le j) - j/t,\quad 1\le j\le t,$$

then $T_F\pi$ is equivalent to the set $\{\mathcal{F}_1, \ldots, \mathcal{F}_t\}$.

The notion of correlation between two rankings has been used previously in nonparametric tests of trend and of independence (Mann (1945), Daniels (1950), Diaconis and Graham (1977)). In that context, it can be shown that the null distribution of a_F (the correlation based on the Spearman Footrule) properly standardized, is asymptotically normal as $t \to \infty$.

In Alvo and Cabilio (1991) an extension of the notion of distance applied to sets of incomplete permutations led to a generalization of the problem of m rankings to the case of incomplete block designs and to a re-interpretation of the Durbin statistic. This extension of the notion of distance was then used by Alvo and Cabilio (1992) to develop tests of trend based on the Kendall and Spearman metrics when the data is incomplete. In the present paper, we are concerned with developing a test of trend in such situations based on the Spearman footrule. We recall the notion of compatibility and the definition of distance between two incomplete rankings.

Definition 1.1. A complete ranking ν of t objects is said to be compatible with an incomplete ranking ν^* of k of these objects, if the relative ranking of every pair of objects ranked in ν^*, coincides with their relative ranking in ν.

The complete rankings $\{\nu_i\}$ may be ordered in some way and we may associate with every incomplete ranking ν^*, a (t! x 1) compatibility vector, whose i^{th} component is 1 or 0 according to whether ν_i is compatible to ν^* or not. We denote by $C(\nu^*)$ the compatibility vector of ν^*.

Definition 1.2. The distance between the incomplete rankings μ^* and ν^*, denoted by $d^*(\mu^*, \nu^*)$, is defined to be the average of all values $d(\mu_i,\nu_j)$ taken over all complete rankings μ_i, ν_j compatible with μ^* and ν^* respectively.

Note that in general d^* is not a metric since the distance thus defined between an incomplete ranking and itself is greater than 0. With this definition, the distance between two incomplete rankings $\mu^*$and $\nu^*$of k_1 and k_2 objects respectively, is given by

$$d^*(\mu^*, \nu^*) = \frac{1}{a_1 a_2} [C(\mu^*)]' \Delta [C(\nu^*)]$$

$$= \frac{1}{a_1 a_2} [C(\mu^*)]' [cJ - T'T][C(\nu^*)]$$

$$= c - \frac{1}{a_1 a_2} [C(\mu^*)]' T'T[C(\nu^*)] \qquad (1.8)$$

where the constants $a_i = t!/k_i!$, $i=1,2$, represent the number of complete rankings compatible with $\mu^*$and ν^* respectively. We can now define correlation between two incomplete rankings.

Definition 1.3. Let M and m be the maximum and minimum values of d^* respectively. The correlation between μ^* and ν^* is defined as

$$\alpha^* = 1 - \frac{2[d^* - m]}{M - m}. \tag{1.9}$$

The use of (1.8) shows that

$$\alpha^* = \frac{2\,[C(\mu^*)]'T'T[C(\nu^*)]}{a_1 a_2 (M-m)}\,. \tag{1.10}$$

The quantity $(M-m)/2$ in (1.10) is a standardizing constant. For the asymptotic results which follow, we shall be interested only in

$$A = \frac{[C(\mu^*)]'T'T[C(\nu^*)]}{a_1 a_2}\,. \tag{1.11}$$

From this point forward, we shall be interested in testing for an increasing trend in an incomplete sequence of data points; that is, the null hypothesis is that there is no trend and the alternative is that there is an increasing trend. Therefore, we shall set ν^* equal to the (complete) identity permutation (1,2,...,t). Thus k_1 becomes k, and k_2 is set to t. (To test for a decreasing trend, simply set ν^* equal to the permutation (t,...,2,1).

An example is provided at the end of Section 2, after an explicit expression for A_F has been computed.

We shall be interested in two cases, denoted hypotheses H_1 and H_2. Both are null hypotheses: there is no increasing trend in the data. Under H_1, we assume in addition that the pattern of the missing observations is fixed. Under H_2, we assume in addition that the pattern of missing observations is randomly selected. Under both hypotheses, the number of observations is known and equal to k, whereas the number of missing observations is equal to t−k.

In the next sections, we shall be concerned with the asymptotic normality of A_F under each of the two hypotheses H_1 and H_2. For both hypotheses, we assume that the rankings for which we have (possibly) incomplete data are in fact uniformly distributed over the t! permutations of (1,2,...,t).

For the null 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. It has been shown in Alvo and Cabilio (1993) that for the Spearman and Kendall distances, under H_1,

$$E[t(\mu) \mid \mu^*] = T[C(\mu^*)]/a, \tag{1.12}$$

where $t(\mu)$ is a column of the matrix T. This motivates us to use a similar calculation to compute the statistic corresponding to the Spearman footrule.

Under H_2, however, we assume that the pattern of missing observations is randomly selected from the set of all possible patterns. This 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. Note that the situations considered here are distinct from those described in Dabrowska (1986) wherein the fact that the data are missing depends on the values of the data.

In the following lemma, we compute a precise expression for the statistic corresponding to the Spearman footrule under either H_1 or H_2. Define:

$$a(i,j) = \frac{\binom{j-1}{\mu^*(o_i)-1}\binom{t-j}{k-\mu^*(o_i)}}{\binom{t}{k}} - 1/t.$$

and set $$t_{F^*} = E[t_F(\mu) \mid \mu^*].$$

Lemma 1. For the Spearman footrule, under either H_1 or H_2, the $[(i-1)t+j]^{th}$ component of t_{F^*} is given by:

$$\sum_{l=1}^{j} a(i, l) \tag{1.13}$$

if the i^{th} item is ranked, and zero if the i^{th} item is not ranked.

Proof: Using the definition of t_F in (1.7), it is clear that

$$t_{F^*}[(i-1)t+j] = P[\mu(i) \leq j \mid \mu^*] - j/t.$$

If the i^{th} item is unranked in the incomplete permutation μ^*, then $\mu(i)$ can take any one of the t

possible values with probability $1/t$ and therefore the conditional expectation is 0. However, if the i^{th} item is ranked in the permutation μ^*, then we can write

$$P[\,\mu(i) \le j \mid \mu^*] = \sum_{l=1}^{j} P[\,\mu(i) = l \mid \mu^*] \ .$$

Now, the permutation μ must be compatible with μ^*. Given that $\mu(i) = l$, the remainder of μ obeys the same order relationships defined by the incomplete permutation μ^*. There are $l-1$ available numbers smaller than l from which $\mu^*(o_i)-1$ must be chosen, and similarly, $t-l$ available numbers greater than l from which $k-\mu^*(o_i)$ must be chosen (recall k is the number of known observations). The remaining numbers are used to fill the positions in μ corresponding to the blanks in μ^*. By the hypergeometric distribution, we get the result.

It follows that

$$A_F = t_F'(\nu)\ E[t_F(\mu) \mid \mu^*] \tag{1.14}$$

where ν is the natural order (1,2,...,t).

2.Computation of the test statistic A_F.

We will now compute the test statistic A_F defined in (1.14). By definition,

$$A_F = \sum_{i=1}^{t}\sum_{j=1}^{t} \Big\{\sum_{l=1}^{j} a(i,j)\Big\}\Big\{ I_{[c_i \le j]} - j/t\Big\}.$$

Lemma 2.1. The statistic A_F is equivalent to

$$\sum_{i=1}^{k} [\mu^*(o_i)\ \frac{t+1}{k+1} - c_i]\ I_{[c_i \ge \mu^*(o_i)]} \ .$$

The tests for either H_1 or H_2 reject whenever A_F is large.

Proof. We may write

$$A_F = \sum_{i=1}^{k} \sum_{l=\mu^*(o_i)}^{t-k+\mu^*(o_i)} \sum_{j=l}^{t-k+\mu^*(o_i)} a(i,j)\, I_{[c_i \le j]}.$$

Set

$$S_1 = \sum_{i=1}^{k} \sum_{l=\mu^*(o_i)}^{t-k+\mu^*(o_i)} \sum_{j=l}^{t-k+\mu^*(o_i)} a(i,l) I_{[c_i \le j]}\, I_{[c_i \ge \mu^*(o_i)]}\, I_{[c_i \ge l]}$$

$$S_2 = \sum_{i=1}^{k} \sum_{l=\mu^*(o_i)}^{t-k+\mu^*(o_i)} \sum_{j=l}^{t-k+\mu^*(o_i)} a(i,l)\, I_{[c_i \le j]}\, I_{[c_i \ge \mu^*(o_i)]}\, I_{[c_i < l]}$$

$$S_3 = \sum_{i=1}^{k} \sum_{l=\mu^*(o_i)}^{t-k+\mu^*(o_i)} \sum_{j=l}^{t-k+\mu^*(o_i)} a(i,l) I_{[c_i \le j]}\, I_{[c_i < \mu^*(o_i)]}\, I_{[c_i \ge l]}$$

$$S_4 = \sum_{i=1}^{k} \sum_{l=\mu^*(o_i)}^{t-k+\mu^*(o_i)} \sum_{j=l}^{t-k+\mu^*(o_i)} a(i,l) I_{[c_i \le j]}\, I_{[c_i < \mu^*(o_i)]}\, I_{[c_i < l]}.$$

It follows that $A_F = \sum_{i=1}^{4} S_i$. Note that using Feller (1968 , p.65, identity 12.16) we have

$$S_1 + S_2 = \sum_{i=1}^{k} \sum_{l=\mu^*(o_i)}^{t-k+\mu^*(o_i)} \sum_{j=c_i}^{t-k+\mu^*(o_i)} a(i,l)\, I_{[c_i \ge \mu^*(o_i)]}$$

$$= \sum_{i=1}^{k} \sum_{l=\mu^*(o_i)}^{t-k+\mu^*(o_i)} [t+1-k+\mu^*(o_i) - c_i]\, a(i,l)\, I_{[c_i \ge \mu^*(o_i)]}$$

$$= \sum_{i=1}^{k} [t+1-k+\mu^*(o_i) - c_i]\, I_{[c_i \ge \mu^*(o_i)]}.$$

On the other hand,

$$S_4 = \sum_{i=1}^{k} \sum_{l=\mu^*(o_i)}^{t-k+\mu^*(o_i)} \sum_{j=l}^{t-k+\mu^*(o_i)} a(i,l)\, I_{[c_i < \mu^*(o_i)]}\, I_{[c_i < l]}.$$

$$= \sum_{i=1}^{k} \sum_{l=\mu^*(o_i)}^{t-k+\mu^*(o_i)} [t+1-k+\mu^*(o_i)-l]\, a(i,l)\, I_{[c_i<\mu^*(o_i)]}.$$

$$= \sum_{i=1}^{k} [t+1-k+\mu^*(o_i)-\mu^*(o_i)\tfrac{t+1}{k+1}]\, I_{[c_i<\mu^*(o_i)]}.$$

The last equality is a consequence of the following identity due to Riordan (1968, p.10):

$$(t-k)! \sum_{s=i}^{t-k+i} s \binom{s-1}{i-1}\binom{t-s}{k-i} = i\binom{t+1}{t-k}(t-k)! = i\,\frac{(t+1)!}{(k+1)!}$$

Finally, in view of the indicator functions in its definition, $S_3 = 0$.

Hence,
$$A_F = \sum_{i=1}^{k} [t+1-k+\mu^*(o_i) - c_i\, I_{[c_i\geq\mu^*(o_i)]} - \mu^*(o_i)\,\tfrac{t+1}{k+1}\, I_{[c_i<\mu^*(o_i)]}\ ].$$

$$= \sum_{i=1}^{k} [t+1-k+\left\{\mu^*(o_i)\ \tfrac{t+1}{k+1} - c_i\right\} I_{[c_i\geq\mu^*(o_i)]} - \mu^*(o_i)\,\tfrac{t+1}{k+1}\ ]$$

$$= \sum_{i=1}^{k} [\mu^*(o_i)\ \tfrac{t+1}{k+1} - c_i]\, I_{[c_i\geq\mu^*(o_i)]} + k(t+1-k) + \frac{k(k+1)}{2} - \frac{k(t+1)}{2}$$

$$= \sum_{i=1}^{k} [\mu^*(o_i)\ \tfrac{t+1}{k+1} - c_i]\, I_{[c_i\geq\mu^*(o_i)]} + \frac{k(t-k)}{2} + k.$$

In the next section, we will study the asymptotic distribution of the statistic corresponding to the Spearman footrule when the locations of the missing data are fixed.

Example:

We obtain t=9 observations, of which are known k=5 . The data observed are:

(20 – 30 – 25 10 – 40 –).

This yields the incomplete ranking μ^*=(2 – 4 – 3 1 – 5 –).

We set ν^*=(1 2 3 4 5 6 7 8 9), the complete identity permutation, and determine the needed values as follows:

i	$\mu^*(o_i)$	c_i	$\mu^*(o_i)(10/6)-c_i$	$I_{[c_i \geq \mu^*(o_i)]}$
1	2	1	20/6−1	0
2	4	3	40/6−3	0
3	3	5	30/6−5	1
4	1	6	10/6−6	1
5	5	8	50/6−8	1

Therefore, we add only the last three elements of the fourth column, which yields A_F=−4.

3. Asymptotic results when the pattern of missing observations is fixed.

In this section, we first prove the asymptotic normality for the Spearman footrule under H_1. The theorem can also be used whenever one conditions on the observed pattern in much the same way as when ties are observed. We first quote a result due to Hoeffding (1951).

Lemma 3.1. (Hoeffding) Let $(R_1, \ldots, R_n)$ be a random vector which takes the n! permutations of (1, ..., n) with equal probabilities. Let c(i,j), i,j =1, ..., n be n^2 real numbers. Let $S_n = \sum_{i=1}^{n} c(i, R_i)$ and define

$$d(i,j) = c(i,j) - \frac{1}{n}\sum_{g=1}^{n} c(g,j) - \frac{1}{n}\sum_{h=1}^{n} c(i,h) + \frac{1}{n^2}\sum_{g=1}^{n}\sum_{h=1}^{n} c(g,h).$$

Then, the distribution of S_n is asymptotically normal with mean $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} c(i,j)$ and variance

$\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1}^{n} d^2(i,j)$ if

$$\text{Lim}_{n\to\infty} \frac{\max_{1\leq i\leq n} d^2(i,j)}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d^2(i,j)} = 0. \tag{3.1}$$

In the present context, define

$$c(i,j) = [j\,\tfrac{t+1}{k+1} - c_i]\; I_{[c_i \geq j]}$$

and set $c_{\cdot j} = \sum_{i=1}^{k} c(i,j)$, $c_{i\cdot} = \sum_{j=1}^{k} c(i,j)$, $c_{\cdot\cdot} = \sum_{i,j=1}^{k} c(i,j)$.

Theorem 1. Assume that $k \to \infty$, $t \to \infty$ with $k/t \to \lambda > 0$. Then, under H_1,

$$S_k = \sum_{i=1}^{k} [\mu^*(o_i)\;\tfrac{t+1}{k+1} - c_i]\; I_{[c_i \geq \mu^*(o_i)]}$$

is asymptotically normal with mean $\frac{c}{k} = \frac{1}{k^2} \sum_{i=1}^{k} [\, a_i(a_i+1)/2\, \frac{t+1}{k+1} - c_i a_i]$ where $a_i = \min(k, c_i)$ and variance $\sigma_F^2 = \frac{1}{k-1} \sum d^2(i,j)$, where $d(i,j)$ is given in Lemma 3.1.

Proof: We may apply Hoeffding's theorem with $n=k$, $R_i = \mu^*(o_i)$ and $c(i,j) = [j\,\frac{t+1}{k+1} - c_i]\; I_{[c_i \geq j]}$. It is necessary only to verify condition (3.1). First note that

$$E[S_k] = c/k \quad \text{and} \quad \mathrm{Var}[S_k] = \frac{1}{k-1} \sum_{i,j}^{k} c(i,j)^2 - \frac{1}{k^2} \sum_{i=1}^{k} c_{i\cdot}^{\,2} - \frac{1}{k^2} \sum_{j=1}^{k} c_{\cdot j}^{\,2} + \frac{1}{k} \sum_{i=1}^{k} c_{\cdot\cdot}^{\,2}.$$

It can be seen that $d(i,j) = O(t)$. Moreover,

$$\begin{aligned}
\sum_{i,j}^{k} c(i,j)^2 &= \sum_{i,j}^{k} [j^2 (\tfrac{t+1}{k+1})^2 + c_i^{\,2} - 2\,(\tfrac{t+1}{k+1})\, j\, c_i]\; I_{[c_i \geq j]} \\
&\approx \sum_{i,j}^{k} [j^2 \lambda^2 + c_i^{\,2} - 2\,\lambda j\, c_i]\; I_{[c_i \geq j]} \\
&= \sum_{i,j}^{k} [a_i(a_i+1)(2a_i+1)\lambda^2/6 + a_i c_i^{\,2} - \lambda j\, a_i(a_i+1) c_i] \\
&\approx \sum_{i,j}^{k} [a_i^{\,3}\lambda^2/3 + a_i c_i^{\,2} - \lambda j\, a_i^{\,2} c_i)]
\end{aligned}$$

Observe that, $\sum_{i=1}^{k} a_i^{\,3} \geq \sum_{i=1}^{k} i^3 \geq k^4/4 \approx (t/\lambda)^4/4$

$$\sum_{i=1}^{k} a_i^3 \le \sum_{i=1}^{k} k^3 = k^4 \approx (t/\lambda)^4$$

$$\sum_{i=1}^{k} a_i c_i \ge \sum_{i=1}^{k} i\, i^2 \ge k^4/4 \approx (t/\lambda)^4/4$$

$$\sum_{i=1}^{k} a_i c_i \le \sum_{i=1}^{k} k\,(t-i-1)^2 = \sum_{i=1}^{k} k\, t^2 \approx k^2 t^2 \approx t^4/\lambda^2$$

$$\sum_{i=1}^{k} a_i^2 c_i \ge \sum_{i=1}^{k} i^2\, i \approx k^4/4 \approx (t/\lambda)^4/4$$

and $$\sum_{i=1}^{k} a_i^2 c_i \le \sum_{i=1}^{k} k^2(t-i+1) = \sum_{i=1}^{k} k^2 t = k^3 t \approx t^4/\lambda^3.$$

Hence, $\sum_{i=1}^{k}\sum_{j=1}^{k} c(i,j)^2 \approx t^4 M_1$ where M_1 is a constant.

Similarly,

$$\sum_{i=1}^{k} c_i^2 = \sum_{i=1}^{k}\left[\sum_{j=1}^{k} \left(j\,\frac{t+1}{k+1} - c_i\right) I_{[c_i \ge j]}\right]^2$$

$$= \sum_{i=1}^{k}\left[a_i(a_i+1)\,\frac{t+1}{2(k+1)} - a_i\, c_i\right]^2$$

$$= \sum_{i=1}^{k}\left[a_i^2\ \frac{\lambda}{2} - a_i\, c_i\right]^2$$

$$= \sum_{i=1}^{k}\left[a_i^4\ \frac{\lambda^2}{4} + a_i^2\, c_i^2 - a_i^3 \lambda\, c_i\right] \approx t^5 M_2$$

where M_2 is a constant.

$$\sum_{j=1}^{k} c_j^2 = \sum_{j=1}^{k}\left[\sum_{i=1}^{k} \left(j\,\frac{t+1}{k+1} - c_i\right) I_{[c_i \ge j]}\right]^2$$

$$= \sum_{j=1}^{k}\left[t_i\,\frac{t+1}{k+1} - c_i\right]^2$$

$$=\sum_{i=1}^{k}[a_i{}^2\ \frac{\lambda}{2}-a_i\ c_i{}^*]^2\ \approx\sum_{j=1}^{k}\ c_j{}^{*2}$$

where $t_j=\sum_{i=1}^{k} I_{[c_i\geq j]}$ and $c_i{}^*=\sum_{i=1}^{k} c_i\ I_{[c_i\geq j]}$.

Since, $k-j\leq t_j\leq k$ and $(k-j)(k-j+1)/2\leq c_j{}^*\leq k(2t-k+1)/2$. we have $t_j/c_j{}^*\to 0$ as $k\to\infty$.

Finally,

$$\sum_{j=1}^{k} c_j{}^2\qquad\approx\sum_{j=1}^{k} c_j{}^{*2}\approx t^5M_3$$

and

$$c=\sum_{i,j}^{k} c(i,j)\approx\sum_{i=1}^{k}\sum_{j=1}^{k} c_j{}^*=k\sum_{j=1}^{k} c_j{}^*\approx t^4M_4.$$

where M_3 and M_4 are constants.

It follows that

$$Var[S_k]\approx t^3[\lambda M_1-\lambda^2M_2-\lambda^2M_3+\lambda M_4].$$

Consequently, (3.1) is proved.

4. Asymptotic results when the pattern of missing observations is random.

In this section, we show that A_F is asymptotically normal when the pattern of missing observations is random. In this case, the asymptotics are as $k\to\infty$. We may view this situation as one where the pattern consisting of k items is first chosen at random with probability $1/(\binom{t}{k})$. This determines the items to be ranked which therefore determines the scores c_i. Once the k items are determined, they are then ranked. This way of viewing the random case will help us in computing the mean and variance of the test statistic.

Theorem 2. Let $k \to \infty$, $t \to \infty$ with $k/t \to \lambda$. Write $\theta = \lambda^{-1}$. Then, under H_2, A_F is asymptotically normal with mean

$\approx [(k^2/3)(1/2\theta - 1)] + o(k^2)$ and variance $\approx k^3(\frac{\theta^2}{12} - \frac{\theta}{6} + \frac{1}{6} - \frac{13}{180\theta} - \frac{5}{180\theta^2} + \frac{1}{36\theta^3}) + o(k^3)$.

Proof: Consider

$$E[A_F] = E\{E[A_F|c_i]\} = \frac{1}{t} \sum_{i=1}^{t} \sum_{j=1}^{k} [j\theta - i]\, I_{[i \geq j]}$$

$$= \frac{1}{t} \sum_{j=1}^{k} \sum_{i=1}^{t} [j\theta - i]\, I_{[i \geq j]}$$

$$\approx k^2 M_0$$

where $M_0 = (\frac{1}{6\theta} - \frac{1}{3})$.

Setting $U = \sum_{i=1}^{k} [\mu_i\theta - c_j]^2\, I_{[c_i \geq \mu_i]}$ and $V = \sum_{i \neq j}^{k} [\mu_i\theta - c_j]\, I_{[c_i \geq \mu_i]}\, [\mu_j\theta - c_j]\, I_{[c_i \geq \mu_j]}$.

It follows similarly that

$$E[U] = E\{E[U|c_i]\} = \frac{1}{t} \sum_{i=1}^{t} \sum_{j=1}^{k} [j\theta - i]^2\, I_{[i \geq j]}$$

$$\approx \frac{1}{t} \{\theta^2 + \frac{k^3}{3} - \theta^2 \frac{k^4}{4} + t^3 \frac{k}{3} - \frac{k^4}{12} - \theta t^2 \frac{k^2}{2} + \theta \frac{k^3}{4}\}$$

$$\approx k^3 M_1$$

where $M_1 = \{\frac{\theta^2}{6} - \frac{\theta}{4} + \frac{1}{4} - \frac{1}{12\theta}\}$.

Also

$$E[V] = E\{E[V|c_i, c_j]\} = \frac{1}{t(t-1)} \sum_{p \neq q}^{k} \sum_{i \neq j}^{t} [p\theta - i]\, [q\theta - j]\, I_{[i \geq p]} I_{[j \geq q]}$$

$$= \frac{1}{t(t-1)} \sum_{p \neq q}^{k} \{ \sum_{i=j}^{t} [p\theta - i]\, [q\theta - j]\, I_{[i \geq p]}\, I_{[j \geq q]} - \sum_{i=1}^{t} [p\theta - i]\, [q\theta - i]\, I_{[i \geq p]} I_{[i \geq q]} \}$$

$$= \frac{1}{t(t-1)} \sum_{p \neq q}^{k} \{ a_p a_q - b(p, q)\}$$

where

$$a_p = \sum_{i=1}^{t} [p\theta - i] I_{[i \geq p]} \quad \text{and} \quad b(p, q) = \sum_{i=1}^{t} [p\theta - i]\,[q\theta - i]\, I_{[i \geq p]} I_{[i \geq q]}.$$

Hence,

$$E[V] = \frac{1}{t(t-1)} \{ \sum_{p, q}^{k} a_p a_q - 2 \sum_{p<q}^{k} b(p, q) - \sum_{p=1}^{k} a_p^2 \}.$$

Now,

$$\frac{1}{t(t-1)} \{ \sum_{p, q}^{k} a_p a_q \} = \frac{1}{t(t-1)} \{ \sum_{p}^{k} a_p \}^2 = \frac{t}{(t-1)} \{ \frac{1}{t} \sum_{p}^{k} a_p \}^2 = \frac{t}{(t-1)} \{E[A_F]\}^2.$$

Also, $\frac{1}{t(t-1)} \sum_{p<q}^{k} b(p, q) = \sum_{p<q}^{k} \sum_{i=1}^{t} [pq\theta^2 - \theta(p+q)i + i^2]\, I_{[i \geq \max(p, q)]}$

$$= \frac{1}{t(t-1)} \sum_{p<q}^{k} \sum_{i=1}^{t} [pq\theta^2 - \theta(p+q)i + i^2]\, I_{[i \geq q]}$$

$$\approx \frac{1}{t(t-1)} \sum_{p<q}^{k} [pq\theta^2(t-q) - \theta(p+q)(t^2 - q^2)/2 + (t^3 - q^3)/3]$$

$$\approx k^3 M_2$$

where $M_2 = \{\frac{\theta}{24} - \frac{1}{10} + \frac{3}{20\theta} - \frac{1}{15\theta^2}\}$.

Similarly,

$$\frac{1}{t(t-1)} \{ \sum_{p}^{k} a_p \}^2 \approx k^3 M_3$$

where $M_3 = \{\frac{\theta^2}{12} - \frac{\theta}{6} + \frac{17}{60} - \frac{1}{5\theta} + \frac{1}{20\theta^2}\}$.

It follows that,

$$\begin{aligned}
\mathrm{Var}[A_F] &= E[A_F^2] - \left\{E[A_F]\right\}^2, \\
&= E[U] + E[V] - \left\{E[A_F]\right\}^2 \\
&= k^3 (M_1 - 2M_2 - M_3) + \frac{1}{t-1}\left\{E[A_F]\right\}^2 \\
&= k^3 (M_1 - 2M_2 - M_3 + \frac{M_0^2}{\theta}) \\
&\approx k^3 (\frac{\theta^2}{12} - \frac{\theta}{6} + \frac{1}{6} - \frac{13}{180\theta} - \frac{5}{180\theta^2} + \frac{1}{36\theta^3}) + o(k^3).
\end{aligned}$$

The result follows immediately by Lemma 3.1.

5. An Example:

To demonstrate the usefulness of the test for incomplete data, we consider 98 monthly January precipitation data for the city of Fredericton. The test based on the complete set of 98 data (k=98) yields a value of $A_F = -1702.5$. The mean and variance are -1600.67 and 10 457.7, respectively; the standard deviation is 102.26. Normalizing A_F yields a value of -0.996, so the test does not reject the null hypothesis, which states that there is no trend in the data.

Now if we choose k=60 observations at random from the 98 data points, we can construct the test for the incomplete case, as per section 4. The value of A_F was computed for 300 random samples of 60 observations; a histogram of these values is shown in Figure 5.1.

With t=98 and k=60, Theorem 4.1 allows us to compute the mean and variance of A_F, which are -832.65 and 24 436 respectively. If we normalize A_F, we see that the null hypothesis must be rejected when A_F is larger than -526 to insure 95% certainty. In Figure 5.1, it is clear that A_F never exceeds this value, and therefore, we never reject the null hypothesis.

Thus, we see that the same result was obtained using only 61% of the available data. If, in actuality, 39% of tha data had been missing, the test for trend with incomplete data would have proven as effective as the complete version.

Figure 5.1: Histogram of A_F values

Midpoint	Count	
-1150	2	*
-1100	1	*
-1050	7	****
-1000	16	********
-950	39	**********
-900	57	**********
-850	61	**********
-800	51	**********
-750	37	**********
-700	18	*********
-650	8	****
-600	3	**

Each * represents 2 observations.

6. Discussion.

In this thesis, a rank-based test statistics using the Spearman footrule was defined to handle the situation when the data is incomplete. It was shown that the statistic is asymptotically normal under two different scenarios whenever the number of ranked items increases subject to a rate condition. In the first, the pattern of missing observations is assumed fixed whereas in the second situation, the pattern of missing observations is assumed to occur randomly. This statistic can be used in tests of trend and of independence. It remains to produce tables to be used when the sample size is moderate and to evaluate this test against the statistics obtained by Alvo and Cabilio (1993) based on the Kendall and Spearman metrics.

REFERENCES

Alvo, M., and Cabilio, P. (1992). Correlation methods for incomplete rankings. Technical Report Series of the Laboratory for Research in Statistics and Probability. No.200. Carleton University and University of Ottawa.

Alvo, M., and Cabilio, P. (1993). Tables of critical values of rank tests for trend when the data is incomplete. Technical Report Series of the Laboratory for Research in Statistics and Probability. No.230 Carleton University and University of Ottawa.

Alvo, M., Cabilio, P., and Feigin, P.D. (1982). Asymptotic theory for measures of concordance with special reference to average Kendall tau. Ann. Statist. 10 1269-1276.

Alvo, M., and Cabilio, P. (1991). On the balanced incomplete block design for rankings. Ann. Statist. 19 1597-1613.

Dabrowska, D.M. (1986). Rank tests for independence for bivariate censored data. Ann.Statist. 14,250-264.

Critchlow, Douglas E. (1985). Metric Methods for Analyzing Partially Ranked Data. Lecture Notes in Statistics. Springer-Verlag. Berlin.

Daniels, H.E. (1944). The relation between measures of correlation in the universe of sample permutations. Biometrika 33 129-135.

Daniels, H.E. (1950). Rank correlation and population models. JRSS ser. B, 171-181.

Feigin, P., and Alvo, M. (1986). Intergroup diversity and concordance for ranking data: An approach via metrics for permutations. Ann. Statist. 14 691-707.

Feller, W. (1968). An Introduction to Probability Theory and its Applications. John Wiley & Sons. Inc.

Hájek, J., and Sidák, Z. (1967). Theory of Rank Tests. Academic Press, New York.

Hoeffding, W.(1951). A combinatorial central limit theorem. Ann. of Math. Statist. 22, 558-566.

Jirina, M. (1976). On the asymptotic normality of Kendall's rank correlation statistic. Ann. Statist. 4 214-215.

Kendall, M.G. (1975). Rank Correlation Methods. Griffin, London.

Kendall, M.G., and Stuart, A. (1979). The Advanced Theory of Statistics, Vol 2, Fourth Edition. Griffin, London.

Lehmann, E.L. (1975). Nonparametrics: Statistical Methods Based on Ranks. Holden-Day Inc., San Francisco, California.

Mann, H.B. (1945). Nonparametric tests against trend. Econometrica 13, 245-259.

Riordan, J. (1968). Combinatorial Identities. Wiley, New York.

Serfling, Robert, J. (1980). Approximate Theorems of Mathematical Statistics. John Wiley & Sons. New York.