AUSTRIAN JOURNAL OF STATISTICS
Volume 31 (2002), Number 4, 279–290

Nonparametric Rank Tests for Independence in Opinion Surveys

Philip L.H. Yu[+], K.F. Lam[+], and Mayer Alvo[*]
[+] The University of Hong Kong
[*] University of Ottawa, Canada

Abstract: Nonparametric rank tests for independence between two characteristics are commonly used in many social opinion surveys. When both characteristics are ordinal in nature, tests based on rank correlations such as those due to Spearman and Kendall are often used. The case where some ties exist has already been considered whereas Alvo and Cabilio (1995) have studied the case when there are missing values but no ties in the record. However, it frequently happens that the survey data may contain simultaneously many tied observations and/or many missing values. A naive approach is to simply discard the missing observations and then to make use of the rank correlations adjusted for ties. This approach would be less powerful as it does not fully utilize the information associated with the incomplete data set. In this article, we generalize Alvo and Cabilio's notion of distance between two rankings to incorporate tied and missing observations, and define new test statistics based on the Spearman and Kendall rank correlation coefficients. We determine the asymptotic distribution of the Spearman test statistic and compare its efficiency with the corresponding statistic based on the naive approach. The proposed test is then applied to a real data set collected from an opinion survey conducted in Hong Kong.

Keywords: Opinion Surveys; Asymptotic Relative Efficiency; Incomplete Rankings; Ties; Rank Correlation; Spearman and Kendall Distances.

1 Introduction

Discrete ordinal variables are very popularly seen in many social opinion surveys. To test for the independence between any two such variables, a nonparametric test based on Spearman rank correlation is commonly used. However, it frequently happens that the survey data may contain some missing values. For example in a public opinion survey carried out in early 1999 in Hong Kong by the Social Science Research Centre of the University of Hong Kong, it was of interest to determine whether the age of the respondents is related to the level of satisfaction of the Policy Address of the Chief Executive of the Hong Kong Special Administrative Region. The response is an ordinal variable with the seven options being: 1 - very satisfied, 2 - satisfied, 3 - neutral, 4 - unsatisfied, 5 - very unsatisfied, 6 - not sure, and 7 - refuse to answer. Options 6 and 7 will be classified as missing (non-response). The age at last birthday of the respondents was recorded when available. The following table shows the frequencies of female respondents classified by their responses and age groups.

It can be seen from Table 1 that the problem of missing values is quite severe, about 29.3% of respondents did not respond on either one or both questions. A naive approach

Table 1: Data from the public opinion survey

	Age Group						
Response	$18-24$	$25-34$	$35-44$	$45-54$	>54	$Missing$	$Sub-total$
1	0	0	0	0	0	0	0
2	10	12	17	16	10	5	70
3	16	29	37	21	3	7	113
4	8	9	10	7	1	3	38
5	1	1	3	1	0	0	6
$Missing$	10	14	19	10	14	6	73
$Sub-Total$	45	65	86	55	28	21	300

is to simply discard the missing observations and make use of the classical Spearman rank test for testing the independence between the two variables (see Lehmann, 1975, p. 301). Clearly, this approach would be less powerful as it does not utilize the information associated with the incomplete data set. Since the responses and the ages of the respondents are classified into a few categories, the problem of ties is also very severe. Another approach consists of analyzing the data as a contingency table. However, in that case, the natural ordering which exists among the age groups and similarly among the responses would not be used. The objective of this paper is to develop rank tests for testing independence between two ordinal variables which can incorporate the presence of both missing values and ties. Only tests based on Spearman and Kendall rank correlations will be considered here.

Rank-based correlations due to Spearman and Kendall play an important role as measures of association between two factors and as tests of independence between two random variables. When the data contain missing observations but no ties, Alvo and Cabilio (1995) proposed a new class of rank correlations based on the concept of distance between rankings and derived the corresponding asymptotic distributions of the test statistics. However, their method could not be directly applied to the above-mentioned two-way ordinal classification problem as the data contain many ties.

Suppose in a group of t respondents, each respondent is asked to assign scores on two characteristics, say A and B. Ranking the respondents according to their scores (called objects hereafter) on characteristics A and B results in two complete rankings of t objects $\boldsymbol{\mu} = (\mu(1), \mu(2), \cdots, \mu(t))$ and $\boldsymbol{\nu} = (\nu(1), \nu(2), \cdots, \nu(t))$, which can be viewed as permutations of the integers $(1, 2, \cdots, t)$, if there are no ties and missing values. The Spearman distance between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ given by

$$d_S(\boldsymbol{\mu}, \boldsymbol{\nu}) \;=\; \frac{1}{2}\sum_{i=1}^{t}(\mu(i) - \nu(i))^2$$

can also be expressed in terms of a similarity measure A_S

$$d_S(\boldsymbol{\mu}, \boldsymbol{\nu}) = c_S - A_S(\boldsymbol{\mu}, \boldsymbol{\nu}),$$

where

$$c_S = \frac{t(t^2-1)}{12}, \quad A_S(\boldsymbol{\mu},\boldsymbol{\nu}) = \sum_{i=1}^{t}\left(\mu(i)-\frac{t+1}{2}\right)\left(\nu(i)-\frac{t+1}{2}\right). \tag{1}$$

The Spearman rank correlation ρ_S (Spearman, 1904) can be expressed in terms of c_S and A_S as $\rho_S(\boldsymbol{\mu},\boldsymbol{\nu}) = \frac{A_S(\boldsymbol{\mu},\boldsymbol{\nu})}{c_S}$. In an analogous manner, one may define the Kendall rank correlation with distance

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

where $sgn(x)$ is either 1 or -1 depending on whether $x>0$ or $x<0$. In fact, the Kendall rank correlation (Kendall, 1938) becomes $\rho_K(\boldsymbol{\mu},\boldsymbol{\nu}) = \frac{A_K(\boldsymbol{\mu},\boldsymbol{\nu})}{c_K}$ where

$$c_K = \frac{t(t-1)}{2}, \quad A_K(\boldsymbol{\mu},\boldsymbol{\nu}) = \sum_{i<j}\{sgn\,(\mu(i)-\mu(j))\,sgn\,(\nu(i)-\nu(j))\}.$$

A detailed and complete review of the distance based approach to the analysis of rank data is given in Alvo and Cabilio (1993).

In the next section, we recall the notion of compatibility of Alvo and Cabilio (1995) and use it in Section 3, to introduce new test statistics based on the Spearman and Kendall distances when both ties and missing observations are present. The proposed tests specialize to the known test statistics of Alvo and Cabilio (1995) when there are only missing observations and to the classical Spearman and Kendall tests when only ties are present. The asymptotic distribution of the new Spearman test statistic is derived. It is also found that the two proposed tests are asymptotically equivalent. Some remarks on the asymptotic efficiency of the new Spearman test are made. In Section 4, we apply the new Spearman test to the above opinion survey data. We conclude with some remarks in Section 5.

2 Extensions to Incomplete Rankings with Ties

In this section, we propose two new test statistics based on the Spearman and Kendall distances which make use of all the data. First of all, we introduce two separate definitions of compatibility for missing and tied data.

Definition 1 *A complete ranking of t objects is said to be* **compatible** *with an incomplete ranking of a subset of k of these objects, $2 \le k \le t$ if every pair of the specified k objects is given the same relative ranking in both rankings.*

A tied ordering of t objects is partitioned into e sets, $1 \le e \le t$, each containing g_i objects, $g_1 + g_2 + \cdots + g_e = t$, so that the g_i objects in each set share the rank $\sum_{j=1}^{i-1} g_j + g_i(g_i+1)/2, 1 \le i \le e$. Such a tie pattern is denoted by $\delta = (g_1, g_2, ..., g_e)$.

Definition 2 *A complete ranking of t objects is said to be* **compatible** *with a tied ranking of these objects with tie pattern $\delta = (g_1, g_2, ..., g_e)$ if every pair of objects which receive distinct ranks is given the same relative ranking in both rankings.*

The definition of compatibility when there are both ties and missing values is then a blend of the two previous definitions. We shall denote an incomplete ranking of k out of t objects with tie pattern δ by $\boldsymbol{\mu}^* = (\mu^*(1), \mu^*(2), \cdots, \mu^*(k))$. At times, this k-vector is written as a t-vector in which the missing ranks are denoted by the symbol '_'. In the presence of ties, the concept of compatibility suggests that $\mu^*(i)$ is defined as the midrank of all the items tied with item i. We shall denote by $C_\delta(\boldsymbol{\mu}^*)$ the class of complete rankings compatible with ranking $\boldsymbol{\mu}^*$.

Example. The observations $X_1 = X_2 < X_3$ and X_4 is missing yield $t = 4$, $k = 3$, $e = 2$, $g_1 = 2$, $g_2 = 1$ and the incomplete tied ranking written as a 4-vector becomes $\boldsymbol{\mu}^* = (1.5, 1.5, 3, \_)$. The tied ranking $(1.5, 1.5, 3, \_)$ can be viewed as the average of 2!1!=2 possible incomplete rankings namely, (1,2,3,_) and (2,1,3,_). The 4!/3!=4 complete rankings compatible to (1,2,3,_) are (1,2,3,4), (1,2,4,3), (1,3,4,2) and (2,3,4,1) while the 4 complete rankings compatible to (2,1,3,_) are (2,1,3,4), (2,1,4,3), (3,1,4,2) and (3,2,4,1). The total 4!2!1!/3!=8 complete rankings are then compatible with $(1.5, 1.5, 3, \_)$.

The notion of distance of Alvo and Cabilio (1995) can be generalized to include both missing and tied rankings as follows.

Definition 3 The Generalized Distance *Let* $\boldsymbol{\mu}^*$ $(\boldsymbol{\nu}^*)$ be an incomplete ranking of k_1 (k_2) out of t objects with tie pattern $\delta_1 = (g_{11}, g_{12}, \cdots, g_{1e_1})$ $(\delta_2 = (g_{21}, g_{22}, \cdots, g_{2e_2}))$. The distance between two incomplete rankings $\boldsymbol{\mu}^*$ and $\boldsymbol{\nu}^*$ is defined to be the average of all distances $d(\boldsymbol{\mu}, \boldsymbol{\nu})$ taken over all pairs of complete rankings $\boldsymbol{\mu}_i$ and $\boldsymbol{\nu}_j$, compatible with $\boldsymbol{\mu}^*$ and $\boldsymbol{\nu}^*$, respectively. More formally, let $\kappa_1 = t!(g_{11}!g_{12}!\cdots g_{1e_1}!)/k_1!$ and $\kappa_2 = t!(g_{21}!g_{22}!\cdots g_{2e_2}!)/k_2!$ be the total number of complete t-rankings compatible to $\boldsymbol{\mu}^*$ and $\boldsymbol{\nu}^*$ respectively, and set $\kappa = \kappa_1\kappa_2$. Then we have

$$d^*(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*) = \frac{1}{\kappa} \sum_{\boldsymbol{\mu}_i \in \mathcal{C}_{\delta_1}(\mu^*)} \sum_{\boldsymbol{\nu}_j \in \mathcal{C}_{\delta_2}(\nu^*)} d(\boldsymbol{\mu}_i, \boldsymbol{\nu}_j).$$

It follows from Alvo and Cabilio (1995) that the generalized Spearman distance $d_S^*(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*)$ can be expressed in terms of a similarity measure A_S as

$$d_S^*(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*) = c_S - A_S^*(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*)$$

with

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

where $\delta(j) = 1$ if both rankings of item j are not missing, or 0 otherwise. Similarly, it is readily seen that the generalized Kendall similarity measure is given by

$$A_K^*(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*) = \sum_{i<j} a_1(i,j) a_2(i,j)$$

where the $a_1(i,j)$'s are the scores for ranking 1 given by

$$a_1(i,j) = \begin{cases} 0 & \text{in case of a tie} \\ 1 - \frac{2\mu^*(i)}{k_1+1} & \delta(i) = 1, \delta(j) = 0 \\ \frac{2\mu^*(j)}{k_1+1} - 1 & \delta(i) = 0, \delta(j) = 1 \\ sgn\left(\mu^*(i) - \mu^*(j)\right) & \delta(i) = 1, \delta(j) = 1. \end{cases}$$

The scores for ranking 2 are defined similarly.

In the next section, we shall study a null hypotheses H_0 for testing independence between two random variables. We assume that the number of ranked observations in $\boldsymbol{\mu}^*$ and $\boldsymbol{\nu}^*$ are fixed as are the tie patterns and the pattern of missing observations. Moreover, under H_0, the elements in $C_{\delta_1}(\boldsymbol{\mu}^*)$ are equally likely and are independent of those in $C_{\delta_2}(\boldsymbol{\nu}^*)$. Hence, it is easily shown that under H_0, the measures $A^*(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*)$ for both Spearman and Kendall are conditional expectations given the classes of complete compatible rankings:

$$A^*(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*) = E\left[A(\boldsymbol{\mu}, \boldsymbol{\nu}) | C_{\delta_1}(\boldsymbol{\mu}^*), C_{\delta_2}(\boldsymbol{\nu}^*)\right].$$

As noted in Alvo and Cabilio (1995), this remark along with the fact that

$$E\left(A_K - \frac{4}{t}A_S\right)^2 = O\left(t^2\right)$$

implies that the generalized test statistics are asymptotically equivalent as $t \to \infty$. Consequently, in what follows we shall be concerned only with the Spearman case.

2.1 Asymptotic Distribution of A_S^*

The generalized Spearman distance remains unchanged under any permutation relabeling of the items. This is a property known as right invariance (see Alvo and Cabilio, 1993). Assuming $k_1 \leq k_2$, we may consequently relabel the items in such a way that in ranking 2, the first k_2 objects are the one ranked and similarly tied items can be arranged arbitrarily among themselves in any sequence accordingly. Hence, the missing items can be placed at the end and the new rankings $\nu^*(j)$ appear in natural order in ranking 2. We let o_j be the label of the j^{th} item ranked in ranking 1, and k^* be the number of items ranked in ranking 1 among the k_2 ranked in ranking 2. Define

$$o_j^* = \begin{cases} \nu^*(o_j) & \text{if } 1 \leq j \leq k^* \\ \dfrac{k_2+1}{2} & \text{if } k^*+1 \leq j \leq k_1, \end{cases}$$

and $\bar{o} = \sum_{j=1}^{k_1} \frac{o_j^*}{k_1}$.

As an example, consider the following measurements (X_1, X_2) from $t = 10$ individuals.

j-th Individual	1	2	3	4	5	6	7	8	9	10
X_2	12	17	17	17	24	29	33	35	_	_
X_1	15	37	_	31	18	42	_	39	_	37
ranking 2 $\nu^*(j)$	1	3	3	3	5	6	7	8	_	_
ranking 1 $\mu^*(j)$	1	4.5	_	3	2	7	_	6	_	4.5

Here $k_1 = 7$, $k_2 = 8$, $k^* = 6$, $o_1 = 1$, $o_2 = 2$, $o_3 = 4$, $o_4 = 5$, $o_5 = 6$, $o_6 = 8$, $o_7 = 10$, $o_1^* = 1$, $o_2^* = 3$, $o_3^* = 3$, $o_4^* = 5$, $o_5^* = 6$, $o_6^* = 8$, $o_7^* = 4.5$.

For the analyses which follow, we shall focus on the similarity measure A_S^*. Following Lehmann (1975, p. 360), let $U_1, \cdots, U_{k_1}$ be independent random variables uniformly

distributed on $(0,1)$, let R_j be the rank of U_j $(j = 1, \cdots, k_1)$ and define the function $a_{k_1}(u)$ by

$$k_1 a_{k_1}(u) = \begin{cases} \frac{1}{2}(g_{11}+1) & \text{if } 0 < u \le \frac{g_{11}}{k_1} \\ g_{11} + \frac{1}{2}(g_{12}+1) & \text{if } \frac{g_{11}}{k_1} < u \le \frac{g_{11}+g_{12}}{k_1} \\ \quad \vdots & \quad \vdots \\ (k_1 - g_{1,e_1}) + \frac{1}{2}(g_{1,e_1}+1) & \text{if } \frac{k_1 - g_{1,e_1}}{k_1} < u \le 1. \end{cases} \tag{2}$$

Then the k_1 tuple $\left(a_{k_1}\left(\frac{R_1}{k_1}\right), \cdots, a_{k_1}\left(\frac{R_{k_1}}{k_1}\right)\right)$ and $\left(\frac{\mu^*(o_1)}{k_1}, \cdots, \frac{\mu^*(o_{k_1})}{k_1}\right)$ have the same distribution. Re-write A_S^* as

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

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

$$= \frac{(t+1)^2\, k_1}{(k_1+1)(k_2+1)} \sum_{j=1}^{k_1} \left[o_j^* - \bar{o}\right] \frac{\mu^*(o_j)}{k_1}. \tag{3}$$

It follows that under H_0, A_S^* has the same distribution as S_{k_1} where S_{k_1} is a linear rank statistic of the form $S_{k_1} = \sum_{j=1}^{k_1} b_j\, a_{k_1}\left(\frac{R_j}{k_1}\right)$, with $b_j = \frac{(t+1)^2\, k_1}{(k_1+1)(k_2+1)}\left[o_j^* - \bar{o}\right]$.

Theorem 1 *Assume that*

(i) $k^* \to \infty$ *(hence* $k_1 \to \infty$, $k_2 \to \infty$ *and* $t \to \infty$*) with* $k^*/t \to \lambda > 0$ *(and hence* $\lambda \le \frac{k_1}{t} < 1$ *and* $\lambda \le \frac{k_2}{t} < 1$*);*

(ii) $\max\limits_{j=1,\cdots,e_1} \frac{g_{1j}}{k^*}$ *is bounded away from 1;*

(iii) $\max\limits_{j=1,\cdots,e_2} \frac{g_{2j}}{k^*}$ *is bounded away from 1.*

Then under H_0*, the distribution of* S_{k_1} *is asymptotically normal as* $k^* \to \infty$.

Proof: See the Appendix.

Using Theorem 1, A_S^* is also asymptotically normally distributed under H_0. Moreover, by applying Theorem 'a' on p. 160 of Hájek and Šidák (1967), it can be shown that under H_0, the expected value of A_S^* is zero and the exact variance of A_S^* is

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

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

2.2 Efficiency of the Test Statistic

An important consideration in rank tests is the efficiency of the test statistic. In particular, we would like to compare the proposed statistic A_S^* with the Spearman statistic obtained by discarding all the missing observations. It is shown below that under the location shift alternative to H_0, A_S^* is always more powerful than the corresponding Spearman statistic based on the reduced sample.

Following the approach of Hájek and Sidák (1967), let $X_1, \cdots, X_t$ be independent random variables whose joint density under the location shift alternative to H_0 is given by $q_\beta = \prod_{j=1}^{t} f_0(x_j - \beta_j)$ where f_0 is a known density function having finite Fisher information $I(f_0)$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_t)$ is an arbitrary vector. Upon deletion of all pairs with missing values, we let $k_2 = t$, and $k_1 = k$ where k is simply the actual number of X's observed. Therefore, the Spearman type statistic based on the reduced sample can be written in the form

$$\bar{A}_S = \frac{(t+1)^2}{k+1} \sum_{j=1}^{k} \left[o_j^{\#} - \frac{k+1}{2} \right] \left[\frac{\mu^*(o_j)}{t+1} \right]$$

where $o_j^{\#}$ is defined as the midrank of the j^{th} item ranked in ranking 1. The statistic (3) can be expressed as

$$A_S^* = \frac{(t+1)^2}{k+1} \sum_{j=1}^{k} [\nu^*(o_j) - \bar{o}] \left[\frac{\mu^*(o_j)}{t+1} \right]$$

since $k = k_1 = k^*$ and $o_j^* = \nu^*(o_j)$. On setting $\bar{\beta} = \sum_{i=1}^{t} \beta_i / t$ and provided that

$$\max_{1 \le i \le t} (\beta_i - \bar{\beta})^2 \to 0 \quad \text{and} \quad I(f_0) \sum_{i=1}^{t} (\beta_i - \bar{\beta})^2 \to b^2 \ \text{ for } 0 < b^2 < \infty,$$

it follows immediately that, under the alternative q_β, both $\bar{A}_S$ and A_S^* are asymptotically normal with means and variances given respectively by

$$E[\bar{A}_S | q_\beta] = \frac{(t+1)^2}{k+1} \sum_{j=1}^{k} \left(o_j^{\#} - \frac{k+1}{2} \right) \left(\beta_{o_j} - \bar{\beta} \right) \int_0^1 u \, \phi(u, f_0) du,$$

$$E[A_S^* | q_\beta] = \frac{(t+1)^2}{k+1} \sum_{j=1}^{k} (\nu^*(o_j) - \bar{o}) \left(\beta_{o_j} - \bar{\beta} \right) \int_0^1 u \, \phi(u, f_0) du,$$

$$Var[\bar{A}_S | q_\beta] = \frac{(t+1)^4}{12(k+1)^2} \sum_{j=1}^{k} \left[o_j^{\#} - \frac{k+1}{2} \right]^2,$$

$$Var[A_S^* | q_\beta] = \frac{(t+1)^4}{12(k+1)^2} \sum_{j=1}^{k} [\nu^*(o_j) - \bar{o}]^2.$$

Here $\phi(u, f_0) = [f'(F^{-1}(u))]/[f(F^{-1}(u))]$ for $0 < u < 1$, and F is the cumulative distribution function of f.

Moreover, the asymptotic efficiencies for $\bar{A}_S$ and A_S^* can be obtained as

$$e_{\bar{A}_S} = \lim \frac{\left[\sum_{j=1}^{k}\left(o_j^{\#} - \frac{k+1}{2}\right)\left(\beta_{o_j} - \bar{\beta}\right)\right]^2}{\sum_{j=1}^{k}\left(o_j^{\#} - \frac{k+1}{2}\right)^2 \sum_{j=1}^{t}\left(\beta_j - \bar{\beta}\right)^2} Q_1,$$

$$e_{A_S^*} = \lim \frac{\left[\sum_{j=1}^{k}\left(\nu^*(o_j) - \bar{o}\right)\left(\beta_{o_j} - \bar{\beta}\right)\right]^2}{\sum_{j=1}^{k}\left(\nu^*(o_j) - \bar{o}\right)^2 \sum_{j=1}^{t}\left(\beta_j - \bar{\beta}\right)^2} Q_1$$

where Q_1 is a positive function of f_0 and the limit is taken as $t \to \infty$, $k \to \infty$ with $k/t \to \lambda > 0$. Therefore, the asymptotic relative efficiency of A_S^* relative to $\bar{A}_S$ is given by $e_{A_S^*}/e_{\bar{A}_S}$. Consider the case where $\beta_{o_j} = \nu^*(o_j)$ $(j = 1, \cdots, k)$ and the remaining β_j's are arbitrary. This situation includes alternatives of the form $E(X_i) = \psi_0 + \psi_1 R(X_i)$ for $\psi_1 > 0$ where $R(X_i)$ is just the midrank of item j. It can be seen that irrespective of the density f_0, the asymptotic relative efficiency of A_S^* relative to $\bar{A}_S$ is given by $ARE(A_S^*, \bar{A}_S) = \lim_{k\to\infty} R(k, \boldsymbol{\nu}^*)$ where

$$R(k, \boldsymbol{\nu}^*) = \frac{\sum_{j=1}^{k}\left(o_j^{\#} - \frac{k+1}{2}\right)^2 \sum_{j=1}^{k}\left(\nu^*(o_j) - \bar{o}\right)^2}{\left[\sum_{j=1}^{k}\left(o_j^{\#} - \frac{k+1}{2}\right)\left(\nu^*(o_j) - \bar{o}\right)\right]^2} \geq 1.$$

Note that $R(k, \boldsymbol{\nu}^*) > 1$ in most cases. One exception would be the case of no tied and no missing observations in which case both A_S^* and $\bar{A}_S$ reduce to A_S.

As in Alvo and Cabilio (1995) we may illustrate the results of the calculation of this efficiency. Suppose that $t = 19, k =$7, $o_1^* = 1.5$, $o_2^* = 8.5$, $o_3^* = 8.5$, $o_4^* = 10$, $o_5^* = 11$, $o_6^* = 18$,, $o_7^* = 19$; as well, $o_1^{\#} = 1, o_2^{\#} = o_3^{\#} = 2.5, o_4^{\#} = 4, o_5^{\#} = 5, o_6^{\#} = 6, o_7^{\#} = 7$. Then the ratio of the efficiencies is 1.088. On the other hand, if $o_1^* = 1.5$, $o_2^* = 8$,, $o_3^* = 9$, $o_4^* = 10$, $o_5^* = 11$, $o_6^* = 12$,, $o_7^* = 18.5$ and, $o_1^{\#} = 1, o_2^{\#} = 2, o_3^{\#} = 3, o_4^{\#} = 4, o_5^{\#} = 5, o_6^{\#} = 6, o_7^{\#} = 7$, then the ratio of the efficiencies is 1.163.

3 Opinion Survey Data-Revisited

In this section, we apply the proposed Spearman rank test to the opinion survey data to test whether the level of satisfaction of the Policy Address of the Chief Executive depends on the age of the respondents. For benchmark comparison, we consider a reduced sample whereby we discard all the observations with at least one missing variable and apply the classical Spearman rank test. The values of the test statistics and their p-values are tabulated in Table 2.

From the results shown in Table 2, it is seen that the p-values of the two tests do not give consensus results. At the 5% significance level, the test based on the reduced sample rejects H_0 but the test based on the complete sample does not reject H_0. This indicates that the proposed test provides a simple way of handling the missing values in a fair manner such that the information carried in the partially missing observations is also utilized. As the standardized statistic is negative, this implies that A_S^* is below its expected value and

Table 2: Results of the analyses

Test based on	A_S^*	Standardized statistic	p-value
reduced sample	-107165.00	-2.2039	0.0276
complete sample	-210504.34	-1.8812	0.0600

hence indicates a negative association between age and the level of satisfaction. That is, young people tend to be less satisfied with the Policy address. We also performed a test of independence using a contingency table analysis of the same data whereby the "missing" categories for age and response were dropped. The row corresponding to response "1" was also dropped since there are no occurrences. The chi-square statistic based on 12 degrees of freedom yielded a value of 15.806 and the corresponding p-value is 0.200.

4 Concluding Remarks

Rank tests are widely applicable in many contexts. However, one main disadvantage of the rank tests is that they may not be applicable when the data contain missing observations and/or tied values. The problem appears very often in two-way ordinal classifications used in analyzing survey data. In this paper, we proposed a rank test for independence which is a generalization of the Spearman rank correlation based on a natural extension of the concept of distance between two incomplete rankings in order to include data consisting of both missing observations as well as ties. The test is simple and easily applicable. The test statistic reduces to the classical Spearman/Kendall rank statistic when there are no missing values; it reduces to the test proposed by Alvo and Cabilio (1995) when there are missing values but no ties.

Sometimes, we might want to have a measure of association to indicate the direction of influence between the two characteristics if the test for independence is rejected. It is easy to do so by defining a correlation measure in terms of the generalized Spearman distance $d_S^*(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*) = c_S - A_S^*(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*)$ between $\boldsymbol{\mu}^*$ and $\boldsymbol{\nu}^*$ as

$$\alpha^*(\boldsymbol{\mu}^*, \boldsymbol{\nu}^*) \;=\; 1 \;-\; \frac{2(d^* - m)}{M - m}$$

where M and m be the maximum and minimum value of the generalized Spearman distance d^* taken over all possible patterns of the missing and tied observations when the number of tied groups in rankings 1 and 2, e_1 and e_2, respectively are fixed. Note that $-1 \leq \alpha^* \leq 1$. The calculations of M and m are not straightforward and this interesting problem is worthwhile for future research.

APPENDIX: Proof of Theorem 1

From Lehmann (1975, Corollary 4 on p. 358), if

$$\frac{\max_i (o_i^* - \bar{o})^2}{\sum_{i=1}^{k_1} \left[o_i^* - \bar{o}\right]^2 / k_1} \quad \text{is bounded as } k_1 \to \infty$$

and the functions a_{k_1} satisfy the following three conditions:

(A.1) there exist constants $-\infty < m < M < \infty$ such that $m \le a_{k_1}(u) \le M$ for all u and all k_1;

(A.2) the variance $Var\left[a_{k_1}(U)\right]$ are bounded away from zero;

(A.3) the expectations $E\left[a_{k_1}(U_1) - a_{k_1}\left(\frac{R_1}{k_1}\right)\right]^2$ tend to zero;

then S_{k_1} is asymptotically normally distributed.
Let $\bar{o}^* = \sum_{j=1}^{k^*} \frac{o_j^*}{k^*}$. Note that

$$\begin{aligned}
\sum_{i=1}^{k_1} \left[o_i^* - \bar{o}\right]^2 &= \sum_{i=1}^{k^*} \left[\nu^*(o_i) - \bar{o}^*\right]^2 + k^*(\bar{o}^* - \bar{o})^2 + (k_1 - k^*)\left[\frac{k_2+1}{2} - \bar{o}\right]^2 \\
&\ge \sum_{i=1}^{k^*} \left[\nu^*(o_i) - \bar{o}^*\right]^2 \ge \frac{1}{12}\left[k^*(k^{*2}-1) - \sum_{i=1}^{e_2} g_{2i}^*\left(g_{2i}^* - 1\right)\right] \\
&= \frac{1}{12}\left(k^{*3} - \sum_{i=1}^{e_2} {g_{2i}^*}^3\right).
\end{aligned}$$

Moreover, $(o_i^* - \bar{o})^2 \le (t-1)^2$. Hence,

$$\begin{aligned}
\frac{\max_i (o_i^* - \bar{o})^2}{\sum_{i=1}^{k_1} \left[o_i^* - \bar{o}\right]^2 / k_1} &\le \frac{12\,(t-1)^2\,k_1}{k^{*3} - \sum_{i=1}^{e_2} {g_{2i}^*}^3} \\
&< \frac{12}{\left(\frac{k^*}{t}\right)^3 - \sum_{i=1}^{e_2}\left(\frac{g_{2i}^*}{t}\right)^3} = \frac{12}{\left(\frac{k^*}{t}\right)^3\left[1 - \sum_{i=1}^{e_2}\left(\frac{g_{2i}^*}{k^*}\right)^3\right]}.
\end{aligned}$$

Since $\max(g_{2i}/k^*)$ is bounded away from 1 as $k^* \to \infty$, there exists an ϵ_2 $(0 < \epsilon_2 < 1)$ such that $g_{2i} \le (1-\epsilon_2)k^*$ for all i. Therefore,

$$\frac{\max_i (o_i^* - \bar{o})^2}{\sum_{i=1}^{k_1} \left[o_i^* - \bar{o}\right]^2 / k_1} < \frac{12}{\lambda^3\left[1 - (1-\epsilon_2)^2 \sum_{i=1}^{e_2} \frac{g_{2i}^*}{k^*}\right]} = \frac{12}{\lambda^3\left[1 - (1-\epsilon_2)^2\right]}$$

and hence, the above condition is satisfied. Further, condition (A.1) is obviously satisfied with $m = 0$ and $M = 1$. As to condition (A.2), note that from (2), $k_1 a_{k_1}$ are just the

midranks of the observations in ranking 1, and it is well known (see Lehmann 1975, p. 294) that

$$Var\left[a_{k_1}(U)\right] \;=\; \left(\frac{1}{12k_1^2}\right)\left[(k_1^2-1)-\frac{\sum_{i=1}^{e_1} g_{1i}(g_{1i}^2-1)}{k_1}\right] \;=\; \frac{1}{12}\left[1-\sum_{i=1}^{e_1}\left(\frac{g_{1i}}{k_1}\right)^3\right].$$

By an argument similar to the above with $0<\epsilon_1<1$, such that $g_{1i}\le(1-\epsilon_1)k_1$ for all i, we have

$$Var\left[a_{k_1}(U)\right] \;\ge\; \frac{1}{12}\left[1-(1-\epsilon_1)^2\right] \;>\; 0$$

which shows that condition (A.2) is also satisfied.

To show that condition (A.3) is satisfied, recall the definition of R_1 and let $I_h=1$ if $0<u<\sum_{i=1}^{h} g_{1i}/k_1$ or $I_h=0$ otherwise for $h=1,\cdots,e_1$; $J_h=1$ if $0<\frac{R_1}{k_1}\le\sum_{i=1}^{h} g_{1i}/k_1$ or $J_h=0$ otherwise for $h=1,\cdots,e_1$ and $a_{k_1}^{(h)}(u)=1$ if $\sum_{i=1}^{h-1} g_{1i}/k_1<u\le\sum_{i=1}^{h} g_{1i}/k_1$ for $h=1,\cdots,e_1$. It can be seen that

$$a_{k_1}(u)=\frac{1}{2k_1}(g_{11}+1)a_{k_1}^{(1)}(u)+\cdots+\frac{1}{k_1}\left[(k_1-g_{1,e_1})+\frac{1}{2}(g_{1,e_1}+1)\right]a_{k_1}^{(e_1)}(u).$$

From the inequality $(X_1+\cdots+X_{e_1})^2\le e_1(X_1^2+\cdots+X_{e_1}^2)$ and since the coefficients of $a_{k_1}^{(j)}(u)$ are all less than or equal to 1, it follows that

$$0\;\le\;\left[a_{k_1}(U_1)-a_{k_1}\left(\frac{R_1}{k_1}\right)\right]^2\;\le\;e_1\sum_{i=1}^{e_1}\left[a_{k_1}^{(i)}(U_1)-a_{k_1}^{(i)}\left(\frac{R_1}{k_1}\right)\right]^2.$$

Note that $\sum_j^{k_1}\left[a_{k_1}^{(1)}(U_j)-a_{k_1}^{(1)}\left(\frac{R_j}{k_1}\right)\right]^2=|W-g_{11}|$ where W is the number of U's less than or equal to g_{11}/k_1, the variables $\left[a_{k_1}^{(1)}(U_j)-a_{k_1}^{(1)}\left(\frac{R_j}{k_1}\right)\right]$ are independently and identically distributed with mean 0, and W has the binomial distribution with k_1 trials and probability of success g_{11}/k_1. Consequently, we have

$$E\left[a_{k_1}^{(1)}(U_1)-a_{k_1}^{(1)}\left(\frac{R_1}{k_1}\right)\right]^2=\frac{E|W-d_{11}|}{k_1}\le\sqrt{\frac{1}{k_1}\,\frac{g_{11}}{k_1}\left(1-\frac{g_{11}}{k_1}\right)}\;\to\;0$$

as $k_1\to\infty$ and g_{11}/k_1 is bounded away from 1. Similarly, for $h<e_1$,

$$\begin{aligned}0\le\;& E\left[a_{k_1}^{(h+1)}(U_1)-a_{k_1}^{(h+1)}\left(\frac{R_1}{k_1}\right)\right]^2\\ =\;& E\left[I_{h+1}-I_h-J_{h+1}+J_h\right]^2\\ \le\;& E\left[I_{h+1}-J_{h+1}\right]^2+E\left[I_h-J_h\right]^2\to 0\qquad\text{as } k_1\to\infty.\end{aligned}$$

Hence, we have $E\left[a_{k_1}(U_1)-a_{k_1}\left(\frac{R_1}{k_1}\right)\right]^2\to 0$ as $k_1\to\infty$ and this completes the proof.

Acknowledgments

The research of Philip L.H. Yu and K.F. Lam was supported by the CRCG grant 337/017-/0014 of the University of Hong Kong and partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 7169/98H), whereas the research of M. Alvo was supported by the Natural Sciences and Engineering Council of Canada Grant OGP0009068. The authors would like to thank the Social Sciences Research Centre of the University of Hong Kong for providing the data set.

References

M. Alvo and P. Cabilio. Rank correlations and the analysis of rank-based experimental design. In M.A. Flinger and S.J. Verducci, editors, *Probability Models and Statistical Analyses for Ranked Data. Lecture Notes in Statistics*, volume 80, pages 140–154. Springer, New York, 1993.

M. Alvo and P. Cabilio. Rank correlation methods for missing data. *The Canadian J. of Statistics*, 23(4):345–358, 1995.

J. Hájek and Z. Šidák. *Theory of Rank Tests*. Academic Press, New York, 1967.

M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

E.L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco, 1975.

C. Spearman. The proof and measurement of association between two things. *Am. J. of Psychol.*, 15:72–101, 1904.

Authors' addresses:

Philip L.H. Yu
K.F. Lam
Department of Statistics and Actuarial Science
The University of Hong Kong
Pokfulam Road
Hong Kong
Tel.: (852) 2857-8321
Fax: (852) 2858-9041
E-mail: plhyu@hku.hk

Mayer Alvo
Department of Mathematics and Statistics
University of Ottawa
585 King Edward Ave.
P.O. Box 450, Stn. A
Ottawa, Ontario K1N 6N5
Canada
Tel.: (613) 562-5864
Fax: (613) 562-5776
E-mail: MALVO@science.uottawa.ca