1990

Developments In Rank Correlation Procedures For Trend Detection In The Analysis Of Water Quality Parameters

Paul Dexter Valz

DEVELOPMENTS IN RANK CORRELATION PROCEDURES FOR TREND DETECTION IN THE ANALYSIS OF WATER QUALITY PARAMETERS

by

PAUL DEXTER VALZ, M.Sc.

Department of Statistical and Actuarial Sciences

Submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario, Canada
January, 1990

National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

ISBN 0-315-55298-0

ABSTRACT

Paul Dexter Valz: Developments in rank correlation procedures for trend detection in the analysis of water quality parameters. Ph.D. thesis, The University of Western Ontario, December, 1989.

A primary objective of this thesis is the development of a partial rank correlation test which can be used to test the null hypothesis that two variables are, conditional upon the effect of a third variable, independent of each other and which does not require that the third variable be categorical. A parallel objective is the development of results for the distributions of Spearman's D and Kendall's S statistics; these being two of the most widely used rank correlation statistics when testing for trend or for a monotonic relationship between two variables.

Algorithms for enumerating the exact null distributions of Kendall's S and Spearman's D statistics, when there are ties in one or both of the rankings, are presented. An expression, which is used to provide a simple proof of the asymptotic normality of the score S when both rankings are tied, is obtained for the cumulant generating function of S. The usefulness of an Edgeworth approximation to the null distribution of S in the general case of tied rankings is investigated and compared with the standard normal approximation.

An algorithm for enumerating the exact distribution of Kendall's partial rank correlation statistic $t_{12.3}$, under the complete null hypothesis, is developed. Upper and lower bounds for $Var(t_{12.3})$ are established and a proof of the asymptotic normality of $t_{12.3}$ is given. A probability model, with the property that for the associated permutations $\mathcal{E}(t) = \tau$, is developed for the elements of an inversion vector. The variance of t under this probability model is derived, an application of the results to hypothesis testing is presented, and an algorithm for simulating rankings of size n, so that $\mathcal{E}(t) = \tau$, is given.

It is shown that the variance of $t_{12.3}$, under $H_o : \tau_{12.3} = 0$, varies with the underlying values of τ_{13} and τ_{23}. Straightforward derivations of the non-null variance of t and the covariance of t_{12} and t_{13} are presented. An asymptotic variance estimator for $t_{12.3}$ is derived and the asymptotic normality of $t_{12.3}$, under H_o and for the general case of variates with underlying parental correlation, is established. Monte Carlo simulation is used to show that when the magnitudes of t_{13} and t_{23} are both moderately large, $t_{12.3}$ is not a suitable statistic for testing the hypothesis $H_o' : X_1$ and X_2 are conditionally, given X_3, independent of each other. Consequently, a simulation study of partial Spearman's ρ is implemented. This study shows that $r_{s,12.3}$, when corrected for bias in $r_{s,12}$ etc., provides a satisfactory test statistic whose asymptotic distribution under $H_o : \rho_{s,12} = \rho_{s,13}\rho_{s,23}$ may be adequately approximated by its asymptotic distribution under the complete null hypothesis.

ACKNOWLEDGEMENTS

It is with pleasure that I express my deep appreciation and gratitude to my research supervisor, Dr. A. Ian McLeod, who suggested the research topic, reviewed and corrected each chapter of my thesis, and helped me to prepare five papers for publication. Additionally, he made the use of his wife's micro computer and his office facilities available to me, provided a great deal of support with the technical aspects of preparing this thesis, and helped me to combine the diverse topics covered into a presentable whole.

I would also like to express my appreciation to others who helped in various ways:

To Dr. Keith Hipel, Dr. Reginald Kulperger and Dr. David Bellhouse for the valuable and informative discussions which I had with them;

To Mrs. Maria Lavdas who patiently answered my many questions regarding the use of TeX and to Mr. Leslie Kwarciak who was always helpful whenever I encountered problems with the computer, terminals or printers;

To the Department of Statistics and Actuarial Sciences for the use of their computing facilities;

To the Faculty of Graduate Studies for financial support;

To Dr. David Bacon, without whose influence it would never have been undertaken, in appreciation of his help to me over many years; and

To my wife, Jo-Ann, and my children, Timothy and Marissa, who have patiently endured my preoccupation with the theoretical complexities of the thesis matter and who provided an unflagging source of encouragement during the difficult stages of my research.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 6

CHAPTER 7

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Monotonic trend detection in water quality data

Hirsch et al. (1982) and Berryman et al. (1988) have discussed several characteristics of water quality time series. Such data are usually found to follow non-normal distributions and feature cyclic variations and flow relatedness. Additionally, they may be mutually dependent and frequently contain missing or censored values. These characteristics led Hipel et al. (1988) to describe such data as "messy" environmental data and Berryman et al. to conclude that, when compared to other methods, nonparametric tests have less assumptions that limit their application to the analysis of water quality time series. Consequently, Hirsch et al. and Berryman et al. suggested that nonparametric tests be used to detect trends in water quality data.

Berryman et al. (1988) and Hipel et al. (1986) have used empirical studies to show that, for many practical problems, Spearman's and Kendall's tests are the most powerful tests available for detecting monotonic trends when the data are not subject to cyclic variation. These tests were both developed for the purpose of measuring the correlation between two sets of observations and have been shown, Daniels (1944), to be special cases of a generalized correlation coefficient. Despite Burr's (1960) work on the distribution of Kendall's score S, Panneton and Robillard (1972) stated that if both rankings of the observations contain ties, little is known about the exact distribution of S. The exact distribution of Spearman's score D, in an absence of ties, is much more difficult to ascertain and, so far as we know, no research has been carried out on the distribution of D when either one or both of the rankings contain ties. Chapters 2 and 3 of this thesis fully explore the null distributions of both S and D. Algorithms which are capable of enumerating the exact null distributions of these statistics, in the presence of ties, are developed in Chapter 2. Some theoretical contributions to the asymptotic null distribution of S, culminating in a comparison of the Edgeworth series approximation and the Normal approximation to the null distribution of S, when both rankings contain ties, are then presented in Chapter 3. The chapter concludes with a brief discussion of the relative merits of several approximations to the null distribution of D in an absence of ties.

1.2 Seasonality and the extraneous variable effect

Consider a multivariate suite of water quality data. There are, then, two practical considerations which lead to the research embodied in the remainder of this thesis. Firstly, the data are often subject to cyclic variation due to the effect of seasonality which Hirsch et al. (1982) define as the existence of different distributions for different times of the year. Secondly, the researcher may wish to determine whether an apparent trend is due to the effect of a third, or extraneous, variable. The seasons may be regarded as a variable in much the same way that time is regarded as a variable. The problem of seasonality then reduces to the problem of removing the effect of an extraneous variable, season, on the assessment of trends in water quality variables. This problem of an extraneous variable leads to a consideration of partial association statistics or partial rank correlation statistics.

Van Belle and Hughes (1984) have categorized nonparametric tests for detecting trends, in the presence of seasonality, into two main classes. One class is the set of aligned rank methods. Another class is the set of intrablock methods which compute a statistic such as Kendall's S for each block or season and then sum these to produce a single overall statistic. Such tests have been developed by Hirsch et al. (1982), Hirsch and Slack (1984), Van Belle and Hughes (1984) and Lettenmaier (1988). As is evident from Agresti's (1977) definition of partial association, stated in Section 6.4 of this thesis, the class of intrablock methods may be considered as a set of partial association statistics. However, they discard potentially valuable information, because they make no data comparisons across blocks. The primary objective of our research focuses upon finding a test which facilitates utilization of this discarded information. Also, it is desirable to have a test for trend, in the presence of an extraneous variable, which does not require that the extraneous variable be categorical.

Analogously to Daniel's generalized correlation coefficient, Somers (1959) showed that it is possible to define a generalized partial correlation coefficient. Thus there are two statistics which present themselves for our consideration; partial τ, based on Kendall's τ, and partial Spearman's ρ_s, based on Spearman's ρ_s. Kendall (1942) developed partial τ by using the concept of a fourfold table. Somers formally showed that partial τ is a special case of the generalized partial correlation coefficient. However, Kendall's work allows for both an intuitive interpretation and a more fundamental definition of partial τ. Perhaps, because of this, Hettmansperger (1984) stated that, "... τ (unlike ρ_s) can be extended to the case of partial correlation". Somer's work shows this to be incorrect. However, it is true that no intuitive interpretation, based on a more fundamental derivation than that of a generalized partial correlation coefficient, is available for partial Spearman's ρ_s. This may explain the absence of any research on the distribution of partial Spearman's ρ_s.

Our research begins with a study of Kendall's partial τ. The distribution of partial τ, under the complete null hypothesis of pairwise independence between variables, is developed in Chapter 4. Chapter 5 digresses to develop a probability model for the distribution of Kendall's t statistic when parental rank correlation exists. Distributional results pertaining to both τ and partial τ, in the presence of parental rank correlation, are then presented in Chapter 6. These results suggest that partial τ is not a suitable statistic for measuring the monotonic correlation between two variables, independently of the effect of an extraneous variable, when the underlying parental correlations are moderately high. Consequently, Chapter 7 explores the use of partial Spearman's ρ_s. Simulation studies suggest that partial Spearman's ρ_s, when adjusted for the bias in r_s, is an appropriate statistic for our purpose and indicate that this test will perform better than that of Hirsch et al. (1982) which is used for data which are uncorrelated across seasons. This thesis thus concludes with a recommendation that partial Spearman's ρ_s, adjusted for bias, be used as a tool for assessing the relationship between two variables independently of the effect of an extraneous variable. Further investigation into the asymptotic distribution of the statistic is also recommended.

REFERENCES

Agresti, A. (1977). Considerations in measuring partial association for ordinal categorical data. *Journal of the American Statistical Association*, **72**, 37–45.

Berryman, D., Bobée, B., Cluis, D. and Haemmerli, J. (1988). Nonparametric tests for trend detection in water quality time series. *Water Resources Bulletin*, **24**, 545–556.

Burr, E.J. (1960). The distribution of Kendall's score S for a pair of tied rankings. *Biometrika*, **47**, 151–171.

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

Hettmansperger, T.P. (1984). *Statistical inference based on ranks.* John Wiley & Sons.

Hipel, K.W., McLeod, A.I. and Fosu, P.K. (1986). Empirical power comparisons of some tests for trend. *Statistical Aspects of Water Quality Monitoring*, ed. El-Shaarawi A.H. and R.E. Kwiatkowski, Elsevier, 347–362.

Hipel, K.W., McLeod, A.I. and Weiler, R.R. (1988). Data analysis of water quality time series in lake Erie. *Water Resources Bulletin*, **24**, 533–544.

Hirsch, R.M., Slack, J.R. and Smith, R.A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, **18**, 107–121.

Hirsch, R.M. and Slack, J.R. (1984). A nonparametric trend test for seasonal data with serial dependence. *Water Resources Research*, **20**, 727–732.

Kendall, M.G. (1942). Partial rank correlation. *Biometrika*, **32**, 277–283.

Lettenmaier, D.P. (1988). Multivariate nonparametric tests for trend in water quality. *Water Resources Bulletin*, **24**, 505–512.

Panneton, M. and Robillard, P. (1972). The exact distribution of Kendall's S with ties in one ranking. *Applied Statistics*, **21**, 321–323.

Somers, R.H. (1959). The rank analogue of product-moment partial correlation and regression, with application to manifold, ordered contingency tables. *Biometrika*, **46**, 241–246.

Van Belle, G. and Hughes, J.P. (1984). Nonparametric tests for trend in water quality. *Water Resources Research*, **20**, 127–136.

Chapter 2

THE DISTRIBUTIONS OF KENDALL'S S AND SPEARMAN'S ρ_s WITH TIES IN ONE OR BOTH OF THE RANKINGS

Algorithms for enumerating the exact null distributions of Kendall's S statistic and of Spearman's ρ_s statistic, when there are ties in both of the rankings, are presented. For the case where there are ties in only one ranking a recursion formula is derived which yields the exact null distribution of Kendall's S. An application of this result is illustrated by a brief analysis of some meteorological data.

Let $\{r_{1i}\}$ and $\{r_{2i}\}$ be the rankings of n pairs of observations on two random variables X and Y. Then Kendall's rank correlation statistic S is defined as

$$S = \sum_{i>j} \operatorname{sign}((r_{1i} - r_{1j})(r_{2i} - r_{2j})), \tag{2.1}$$

where sign($\bullet$) denotes the signum function. The exact distribution of S, for the case of no ties in either ranking, has been derived by both Kendall (1938) and Mann (1945) under the null hypothesis that X and Y are independent.

For the case where there are ties in one ranking, Sillitto (1947) has tabulated the null distributions of S for any number of tied pairs or tied triplets up to and including $n = 10$. However, as Kendall (1975, Section 4.10) observes, "no complete tables are available owing to the large number of possibilities". Robillard and Panneton (1972) developed an algorithm, based on the frequency generating function, for enumerating the exact null distribution of S with ties in one ranking. However, they note that, "the maximum value of n to be used is essentially limited by the word length of the machine, which must be capable of representing $n!$ as an integer".

For the case where there are ties in both rankings, Burr (1960) has enumerated the null distributions for $n = 3, 4, 5,$ and 6. However, the approach which was used possesses the following limitations as stated by Burr: 1) Calculations of the second kind ($n > 6$) are sometimes tedious; 2) ... the computer should take great care to avoid such mistakes; 3) The whole calculation described above, ..., was completed in about one hour. It is thus evident that the approach used is time consuming, tedious and prone to error. To circumvent these problems, Burr suggests using an independent short method for finding the exact significance levels of two or three extreme values of S and of $-S$. However, this method can also lead to errors and, as shall be seen, it led to Burr stating an incorrect significance level in one of his examples.

Given the rankings specified in eqn. (2.1), Spearman's ρ_s is defined as

$$\rho_s = 1 - \frac{6D}{n^3 - n}$$

where

$$D = \sum_i (r_{1i} - r_{2i})^2 \ . \tag{2.2}$$

The equation giving ρ_s as a function of D needs to be modified to account for ties, Kendall (1975, eqn. (3.8)). However, ρ_s remains a linear function of D and, therefore, only the distribution of D is subsequently considered. Kendall (1975, Section 5.9) notes that no recursive method is known for constructing the null distribution of D. For the case of no ties in either ranking, Kendall (1975, Appendix 2) has tabulated the null distributions of D for $n = 4, \ldots, 13$. Franklin (1987a, 1987b) has presented complete tables for $n = 12, \ldots, 18$. So far as we know, no work has been done on the problem of enumerating the null distribution of D when there are ties in either one or both of the rankings.

In Section 2.1, the recursion formula for computing the null distribution of S in the absence of tied ranks, as given by Kendall (1975), is extended to incorporate the case with ties in one ranking for any n and for any combination of ties. The extension is a generalization of the method which Sillitto (1947) used in solving the problem for tied pairs. An algorithm for computing the exact null distribution of S, or of D, when both rankings are tied is then developed in Section 2.2. If there are ties in only one of the rankings, the algorithm of Section 2.1 is computationally more efficient for enumerating the null distribution of S. Section 2.3 concludes with some illustrative examples and timings on a microcomputer.

2.1 The distribution of S with ties in one ranking

2.1.1 Recursion formula for the distribution of S

Consider m pairs of observations, x_i and y_i; $i = 1, \ldots, m$ on two variables X and Y which are, respectively, continuous and discrete random variables and also are, under the null hypothesis, assumed to be independent. Next, consider the addition of r pairs of observations where the added r observations, on the second variable, are all tied but are distinct from the previous m observations. A set of r such tied observations leads to an r-tuplet in the ranking R_2 of the observations y_i. Since the null distribution of S depends, Sillitto (1947), only on the partitioning of the second ranking R_2 into t_1 singles, t_2 pairs, t_3 triplets, $\ldots$, t_g g-tuplets and not on the actual magnitudes of the ranks, we may, without loss of generality, set the added r-tuplet to be greater than any of the previous m ranks of R_2. Similarly, it may be assumed that R_1 is in ascending order. Following Kendall (1975), $U(m, S; v_1, \ldots, v_k)$ denotes the number of ways of obtaining a score S with m pairs of observations where there are ties, in R_2, of extent v_j; $j = 1, \ldots, k$ and $\sum_{j=1}^{k} v_j = m$. If $U(m, S; v_1, \ldots, v_k, v_{k+1})$ is used to denote the number of ways of obtaining a score S after addition of an r-tuplet, it then follows that $v_{k+1} = r$.

To develop an appropriate recursion formula, note that the following positions, and corresponding changes in scores, may be occupied by the added r-tuplet:

(i) All r entries occupy the last r positions. The increase in the score is mr and the added r-tuplet is completely uninverted (zero inversions). An entry gives rise to j inversions if it is positioned such that there are j ranks of smaller magnitude on its right.

(ii) All r entries occupy the first r positions. The increase in the score is $-mr$ and the added r-tuplet is completely inverted (mr inversions).

(iii) The r entries occupy positions between the above two extremes with a total of ℓ inversions, $0 \leq \ell \leq mr$. The increase in the score is $mr - 2\ell$; i.e. the score is first increased by mr as in (i) above and then reduced by 2 for every inversion of each entry.

Let S' be the score for some permutation of R_2 based on m observations. Then for r entries of an r-tuplet with ℓ inversions the new score S'' is obtained as $S'' = S' + mr - 2\ell$. Hence a new score S, based on $m + r$ observations, may be generated from an old score, based on m observations, by the insertion of an r-tuplet with an appropriate number of inversions. This leads to the recursion formula

$$U(m+r, S; v_1, \ldots, v_k, r) = \sum_{\ell=0}^{mr} U(m, S - mr + 2\ell; v_1, \ldots, v_k) \times C(\ell; v_1, \ldots, v_k, r) \tag{2.3}$$

where ℓ represents the total number of inversions due to insertion of an r-tuplet and $C(\ell; v_1, \ldots, v_k, r)$ is equal to the number of distinct sets of r positions corresponding to a given value of ℓ. Eqn. (2.3) generalizes the result of Kendall (1975, eqn. (5.1)) to the case of ties in one ranking.

The recursion formula, eqn. (2.3), may be used to generate the distribution of S given

(i) A starting point; eg. $U(1, 0; 1) = 1$ defines the distribution of S with $m = 1$.

(ii) The values $t_1, t_2, \ldots, t_g$ corresponding to the breakdown of the second ranking into numbers of r-tuplets, $r = 1, \ldots, g$.

(iii) The coefficients, $C(\ell; v_1, \ldots, v_k, r)$.

To actually implement the recursion only non-negative values of S need be considered since the distribution of S is symmetric about zero, i.e. $U(m, -S; v_1, \ldots, v_k) = U(m, S; v_1, \ldots, v_k)$. Let $M(v_1, \ldots, v_k)$ be the maximum score for m ranks. Then it follows from our previous argument that $M(v_1, \ldots, v_k, r) = M(v_1, \ldots, v_k) + mr$. Since each added inversion reduces the score by 2, the number of non-negative values of S is obtained as $[(M(v_1, \ldots, v_k, r) + 2)/2]$ where [ ] denotes the integer part of. When taken in conjunction with the fact that adjacent scores differ by 2, this completely specifies the set of values of S to be considered in step (vi) of the algorithm, Section 2.1.3.

2.1.2 Calculation of $C(\ell, v_1, \ldots, v_k, r)$

Consider $m + 1$ positions (position zero through position m) which each of the r new ranks may occupy. Within this context, several of the added ranks may now occupy the same position. When the added ranks are completely inverted they have all assumed position zero. When the added ranks are completely uninverted they have all assumed position m. Since the r new ranks are identical care must be taken to avoid permutations of the added ranks. Conceptually, this may be done by labelling each of the r new ranks as $p_1, p_2, \ldots, p_r$ and imposing the restriction that $\text{pos}(p_r) \leq \text{pos}(p_{r-1}) \leq \cdots \leq \text{pos}(p_1)$. An exhaustive search of all the positions, $\text{pos}(p_i)$; $i = 1, \ldots, r$ which the added r-tuplet can assume subject to the specified restriction is then needed. Such an exhaustive search of an ordered sequence of available positions may be efficiently implemented by using a generalized backtrack algorithm, Reingold, Nievergelt and Deo (1977, Section 4.1). For each combination of positions, the number of inversions is easily determined.

The algorithm used is

(i) If $r = 1$ set $C(\ell; v_1, \ldots, v_k, r) = 1$ for all ℓ and return. Otherwise proceed to step (ii).

(ii) Initialize $j = 1$.

(iii) Let $j_c = j$. Set $\text{pos}(p_j) = 0;\ j_c \leq j \leq r$.

(iv) Set $j = r$.

(v) Compute the number of inversions corresponding to the current position vector.

(vi) If $\text{pos}(p_r) = m$ return. Since $\text{pos}(p_r)$ is the smallest position.

(vii) If $\text{pos}(p_j) = \text{pos}(p_{j-1});\ 2 \leq j \leq r$ then set $j = 1$. Otherwise, let j be the largest integer such that $\text{pos}(p_j) \neq \text{pos}(p_{j-1})$.

(viii) Let $\text{pos}(p_j) = \text{pos}(p_j) + 1$.

(ix) If $j = r$ go to step (v). Otherwise, set $j = j + 1$ and go to step (iii).

2.1.3 Algorithm for implementing the recursion

The above development leads to the following algorithm for determining the distribution of S for $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. It is assumed that the X data series contains untied values.

(i) Sort the Y data series into ascending order.

(ii) Use the sorted series to determine $t_1, t_2, \ldots, t_{M_t}$ where M_t is the size of the largest tuplet in the data set.

(iii) Set $m = M_t$ and $U(m, 0; M_t) = 1$. This defines the distribution of S with $m = M_t$ for M_t tied ranks.

(iv) Add the largest tuplet available, say an r-tuplet.

(v) Determine the coefficients, $C(\ell; v_1, \ldots, v_k, r)$, for the current value of m and the current tuplet size, r.

(vi) Use eqn. (2.3) to find $U(m + r, S; v_1, \ldots, v_k, r)$ for all S.

(vii) Set $m = m + r$. If $m = n$ stop. Otherwise go to (iv).

Theoretically, the order of addition of the tuplets is irrelevant. However, in the given algorithm, step (v) requires the most computer time because the number of different ways of inserting an r-tuplet into $m+r$ positions grows rapidly with mr. Consequently, it is most efficient to cascade downwards (largest first and smallest last) through the tuplets, thereby minimizing the value of m for the larger values of r.

The above algorithm, which has been coded in Fortran, is preferable to that of Panneton and Robillard (1972) since the use of high precision integer arithmetic which is generally not available in most computing environments is not needed. In fact, ordinary double precision arithmetic suffices.

2.1.4 Illustrative example

Lamb (1970) gives data on the worldwide volcanic dust veil while Manley (1974) provides data on the average monthly temperatures for Central England. It is of interest to ascertain whether or not there appears to be any relationship between the levels of volcanic dust and temperature. The data used, comprising 10-year averages of the original data sets, are listed in Table 2.1.

Table 2.1: Mean Temperature of Central England (TEMP) and Volcanic Dust Veil Index (DVI)

Years	DVI	TEMP	*Years*	DVI	TEMP
1750 – 1759	113	9.043	1860 – 1869	68	9.308
1760 – 1769	24.5	9.059	1870 – 1879	42	9.084
1770 – 1779	30.5	9.230	1880 – 1889	145.5	8.871
1780 – 1789	142	8.861	1890 – 1899	23.5	9.184
1790 – 1799	52	9.108	1900 – 1909	60.5	9.113
1800 – 1809	18	9.159	1910 – 1919	16.5	9.282
1810 – 1819	238.5	8.798	1920 – 1929	0	9.374
1820 – 1829	80.5	9.350	1930 – 1939	0	9.611
1830 – 1839	237.5	9.218	1940 – 1949	0	9.672
1840 – 1849	79	9.090	1950 – 1959	0	9.488
1850 – 1859	35	9.163	1960 – 1969	40	9.278

The data are all untied except for the four zero values of the DVI covering the period 1920-1959. The observed Kendall's score for this data set is given by $S = -119$ and the normal approximation yields a significance level of 8.4×10^{-4} which implies a strong negative relationship between DVI and TEMP.

Since there are ties in only one ranking it is of interest to check the normal approximation with our exact algorithm. Note that $n = 22$ is greater than 15 which precludes the use of Panneton's and Robillard's algorithm. Enumeration of the exact distribution takes 0.4 seconds on a microcomputer and produces an exact two-sided significance level of 5.1×10^{-4}. In this case the normal approximation is very good. However, if the ties were more extensive, the normal approximation could become inadequate.

2.2 The distributions when there are ties in both rankings

Certain essential features, which permit the application of a recursive technique to the problem of determining the distribution of S when at most one of the two rankings contains ties, are now missing, viz.:

(i) If both rankings are arranged into ascending order, then a single inversion within one ranking no longer changes the score by a predetermined amount of -2. The actual change in the score depends on the configuration of ties within the two rankings at the points where the inversion takes place.

(ii) For a given ordered subset k of the n lower ranks, the score obtained is no longer independent of the actual positions occupied by the k ranks since the relative configuration of the upper set of ranks to the lower set of ranks now varies with position.

This suggests that any attempt to develop a recursive approach for solving the problem of enumerating the null distribution of S will be futile. As mentioned previously, a recursive technique is unavailable for enumerating the null distribution of D even in the absence of ties. Consequently, an approach based on enumeration, and evaluation of the associated score S or D, of all permutations of one ranking relative to another is used. This requires a systematic method of enumerating these permutations. To do this, results developed by MacMahon (1915, Chapter 2) are utilized.

It is helpful to establish an overview of the fundamental problem to be solved with regard to a specific example which then serves to motivate the application of MacMahon's results. Consider, therefore, observations $(x_i, y_i) = (7,5)$, (9,3), (7,4), (7,4), (9,3), (9,6), (7,3) on X and Y. There are, then, ties of extent $H_x(i)$; $i = 1,\ldots,N_x$ and $H_y(j)$; $j = 1,\ldots,N_y$ in the corresponding rankings, R_1 and R_2, so that

$$\sum_{i=1}^{N_x} H_x(i) = \sum_{j=1}^{N_y} H_y(j) = n \tag{2.4}$$

where, currently, $H_x = (4,3)$, $H_y = (3,2,1,1)$, $N_x = 2$, $N_y = 4$ and $n = 7$. Our problem is to enumerate all possible permutations of R_1 relative to R_2.

The original data may be represented as a table:

Table 2.2: Frequency table showing the rankings of an example data set

1	2	1	0	4
2	0	0	1	3
3	2	1	1	7

A possible permutation of the data is given by the pairs (7,3), (9,5), (7,4), (7,4), (9,3), (9,6), (7,3) which in tabular form is:

Table 2.3: Frequency table showing a permutation of Table 2.2

2	2	0	0	4
1	0	1	1	3
3	2	1	1	7

In general, any such permutation may be represented as a frequency table showing the rankings of n members:

Table 2.4: Frequency table showing the rankings of n members

a_{11}	a_{12}	$\cdots$	a_{1N_y}	$H_x(1)$	Ranking R_1
a_{21}	a_{22}	$\cdots$	a_{2N_y}	$H_x(2)$	
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	
$a_{N_x 1}$	$a_{N_x 2}$	$\cdots$	$a_{N_x N_y}$	$H_x(N_x)$	
$H_y(1)$	$H_y(2)$	$\cdots$	$H_y(N_y)$	n	

Ranking R_2

Define a configuration of Table 2.4 by the variable A. Our problem then becomes that of enumerating all possible outcomes of A.

2.2.1 Theory and application of homogeneous product sums

MacMahon (1915, Chapter 1) defines a symmetric function of k quantities $\alpha_1, \ldots, \alpha_k$ as a function which remains unchanged however these k quantities may be interchanged or permuted. Then,

$$(x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_k) = x^k - a_1x^{k-1} + a_2x^{k-2} - \cdots + (-1)^k a_k \qquad (2.5)$$

where

$$\begin{aligned} a_1 &= \sum \alpha_1 = \alpha_1 + \alpha_2 + \cdots + \alpha_k, \\ a_2 &= \sum \alpha_1\alpha_2 = \alpha_1\alpha_2 + \alpha_1\alpha_3 + \cdots + \alpha_{k-1}\alpha_k, \\ &\vdots \\ a_k &= \sum \alpha_1\alpha_2\cdots\alpha_k = \alpha_1\alpha_2\cdots\alpha_k \ . \end{aligned}$$

Also,

$$\begin{aligned} (1-\alpha_1x)(1-\alpha_2x)(1-\alpha_3x)\cdots &= 1 - a_1x + a_2x^2 - a_3x^3 + \cdots \\ &= \frac{1}{1 + h_1x + h_2x^2 + h_3x^3 + \cdots} \end{aligned} \qquad (2.6)$$

where,

$$h_1 = \sum \alpha_1 = (1),$$

$$h_2 = \sum \alpha_1^2 + \sum \alpha_1\alpha_2 = (2) + (1^2),$$

$$h_3 = \sum \alpha_1^3 + \sum \alpha_1^2\alpha_2 + \sum \alpha_1\alpha_2\alpha_3 = (3) + (21) + (1^3),$$

$$h_4 = \sum \alpha_1^4 + \sum \alpha_1^3\alpha_2 + \sum \alpha_1^2\alpha_2^2 + \sum \alpha_1^2\alpha_2\alpha_3 + \sum \alpha_1\alpha_2\alpha_3\alpha_4$$

$$= (4) + (31) + (22) + (21^2) + (1^4),$$

$$\vdots \qquad \vdots$$

It may be noted that h_s is the sum of a number of symmetric functions, each of which is denoted by a partition of the integer s. In fact h_s is the sum of the whole of such symmetric functions. The h_s are the homogeneous product sums of weight s of the quantities $\alpha_1, \ldots, \alpha_k$.

For k finitely large, each term of h_s may be written as a vector $(p_1, p_2, \ldots, p_k)$ where each p_i is equal to the power of α_i appearing in the particular term. Hence $0 \leq p_i \leq s$ and $\sum_{i=1}^k p_i = s$. With this representation, it may be observed that the homogeneous product sum of weight s of the k quantities $\alpha_1, \alpha_2, \ldots, \alpha_k$ corresponds identically to a k-part composition of the integer s. A k-part composition of s is defined, Reingold, Nievergelt and Deo (1977, Section 5.3), as follows. Consider the generation of partitions of a positive integer s into a sequence of non-negative integers $p_1, p_2, \ldots, p_k$ so that $\sum_{i=1}^k p_i = s$. If the order of the p_i is important then $(p_1, p_2, \ldots, p_k)$ is called a composition of s; if the value of k is fixed and $p_i = 0$ is allowed then these compositions are called a k-part composition of s. For example, suppose $k = 3$. Then $h_2 \equiv (2,0,0), (0,2,0), (0,0,2), (1,1,0), (1,0,1), (0,1,1)$.

Suppose there are n pairs of observations which lead to rankings R_1 and R_2. A permutation of R_1 versus R_2 corresponds to a distribution of the objects of R_1 into the parcels of R_2. MacMahon (1915, Chapter 2) illustrates, by the following example, how all possible permutations of R_1 versus R_2 may be enumerated.

Consider a product h_4h_3 pertaining to the four terms $\alpha_1, \alpha_2, \alpha_3, \alpha_4$. This may be written as

$$\alpha_1^4 + \alpha_2^4 + \alpha_3^4 + \alpha_4^4$$
$$+\alpha_1^3\alpha_2 + \alpha_1\alpha_2^3 + \alpha_1^3\alpha_3 + \alpha_1\alpha_3^3 + \alpha_1^3\alpha_4 + \alpha_1\alpha_4^3 + \alpha_2^3\alpha_3 + \alpha_2\alpha_3^3 + \alpha_2^3\alpha_4$$
$$+ \alpha_2\alpha_4^3 + \alpha_3^3\alpha_4 + \alpha_3\alpha_4^3$$
$$+\alpha_1^2\alpha_2^2 + \alpha_1^2\alpha_3^2 + \alpha_1^2\alpha_4^2 + \alpha_2^2\alpha_3^2 + \alpha_2^2\alpha_4^2 + \alpha_3^2\alpha_4^2$$
$$+\alpha_1^2\alpha_2\alpha_3 + \alpha_1^2\alpha_2\alpha_4 + \alpha_1^2\alpha_3\alpha_4 + \alpha_2^2\alpha_1\alpha_3 + \alpha_2^2\alpha_1\alpha_4 + \alpha_2^2\alpha_3\alpha_4 + \alpha_3^2\alpha_1\alpha_2 + \alpha_3^2\alpha_1\alpha_4$$
$$+ \alpha_3^2\alpha_2\alpha_4 + \alpha_4^2\alpha_1\alpha_2 + \alpha_4^2\alpha_1\alpha_3 + \alpha_4^2\alpha_2\alpha_3$$
$$+\alpha_1\alpha_2\alpha_3\alpha_4$$

multiplied by

$$\alpha_1^3 + \alpha_2^3 + \alpha_3^3 + \alpha_4^3$$
$$+\alpha_1^2\alpha_2 + \alpha_1\alpha_2^2 + \alpha_1^2\alpha_3 + \alpha_1\alpha_3^2 + \alpha_1^2\alpha_4 + \alpha_1\alpha_4^2 + \alpha_2^2\alpha_3 + \alpha_2\alpha_3^2 + \alpha_2^2\alpha_4$$
$$+ \alpha_2\alpha_4^2 + \alpha_3^2\alpha_4 + \alpha_3\alpha_4^2$$
$$+\alpha_1\alpha_2\alpha_3 + \alpha_1\alpha_2\alpha_4 + \alpha_1\alpha_3\alpha_4 + \alpha_2\alpha_3\alpha_4 \; .$$

Seeking the coefficients of $\alpha_1^3\alpha_2^2\alpha_3\alpha_4$, a term that arises when the multiplication is carried out, it is found that the term is formed in eleven different ways, viz.

$\alpha_1^3\alpha_2 \cdot \alpha_2\alpha_3\alpha_4$	$\alpha_1^3\alpha_3 \cdot \alpha_2^2\alpha_4$	$\alpha_1^3\alpha_4 \cdot \alpha_2^2\alpha_3$	$\alpha_1^2\alpha_2^2 \cdot \alpha_1\alpha_3\alpha_4$
$\alpha_1^2\alpha_2\alpha_3 \cdot \alpha_1\alpha_2\alpha_4$	$\alpha_1^2\alpha_2\alpha_4 \cdot \alpha_1\alpha_2\alpha_3$	$\alpha_1^2\alpha_3\alpha_4 \cdot \alpha_1\alpha_2^2$	$\alpha_2^2\alpha_1\alpha_3 \cdot \alpha_1^2\alpha_4$
$\alpha_2^2\alpha_1\alpha_4 \cdot \alpha_1^2\alpha_3$	$\alpha_2^2\alpha_3\alpha_4 \cdot \alpha_1^3$	$\alpha_1\alpha_2\alpha_3\alpha_4 \cdot \alpha_1^2\alpha_2$	

This is clearly a distribution of the seven quantities $\alpha_1, \alpha_1, \alpha_1, \alpha_2, \alpha_2, \alpha_3, \alpha_4$ into seven parcels, four of which are of one kind and the remaining three of another. In fact the eleven terms enumerated above represent all the possible outcomes of A

for the specific observations on X and Y given earlier. Each of the homogeneous product sums, h_4 and h_3, specifies the entire set of possible solutions, for a row of Table 2.4, given the row total (4 or 3) and the number of columns (4); these solutions disregarding the column totals. Multiplication of the homogeneous product sums and extraction of the coefficients of $\alpha_1^3\alpha_2^2\alpha_3\alpha_4$ corresponds to selecting all combinations of row solutions which satisfy the column totals $H_y = (3,2,1,1)$. It follows, therefore, that multiplication of h_4 by h_3 generates a set of terms of which the enumerations of interest form a subset.

Furthermore, it is convenient to also fully illustrate the dual problem of interchanging the rows and columns of Table 2.4 and applying the preceding method of solution. This leads to consideration of the product $h_3h_2h_1h_1$, pertaining to the two terms β_1 and β_2, which may be written as

$$(\beta_1^3 + \beta_2^3 + \beta_1^2\beta_2 + \beta_2^2\beta_1) \times (\beta_1^2 + \beta_2^2 + \beta_1\beta_2) \times (\beta_1 + \beta_2)^2 .$$

Seeking the coefficients of $\beta_1^4\beta_2^3$ yields eleven different enumerations of the term, viz.

$\beta_1^3 \cdot \beta_1\beta_2 \cdot \beta_2 \cdot \beta_2$	$\beta_1^3 \cdot \beta_2^2 \cdot \beta_1 \cdot \beta_2$	$\beta_1^3 \cdot \beta_2^2 \cdot \beta_2 \cdot \beta_1$
$\beta_2^3 \cdot \beta_1^2 \cdot \beta_1 \cdot \beta_1$	$\beta_1^2\beta_2 \cdot \beta_1^2 \cdot \beta_2 \cdot \beta_2$	$\beta_1^2\beta_2 \cdot \beta_2^2 \cdot \beta_1 \cdot \beta_1$
$\beta_1^2\beta_2 \cdot \beta_1\beta_2 \cdot \beta_1 \cdot \beta_2$	$\beta_1^2\beta_2 \cdot \beta_1\beta_2 \cdot \beta_2 \cdot \beta_1$	$\beta_2^2\beta_1 \cdot \beta_1^2 \cdot \beta_1 \cdot \beta_2$
$\beta_2^2\beta_1 \cdot \beta_1^2 \cdot \beta_2 \cdot \beta_1$	$\beta_2^2\beta_1 \cdot \beta_1\beta_2 \cdot \beta_1 \cdot \beta_1$	

Once again, ordered multiplication of the appropriate homogeneous product sums generates, amongst other things, the enumerations of interest. An immediate point of interest, from a computational perspective, is that the complexity of the h functions to be multiplied is reduced with

(i) reduction of the number of different kinds of objects being permuted

(ii) reduction of the maximum tuplet of the parcels where (i) appears to be the predominant factor. Note that both (i) and (ii) will often be simultaneously satisfied.

2.2.2 Algorithm for computing the distribution of S or of D

The distribution of S or of D may now be computed as follows:

(i) Sort each of X and Y into ascending order.

(ii) Use the sorted data set to

(a) count the number of different ties in each ranking, N_x and N_y

(b) determine the maximum tuplets, M_x and M_y

(c) determine the order of tuplets appearing in (i). Store these as $H_x(i)$; $i = 1, \ldots, N_x$ and $H_y(j); j = 1, \ldots, N_y$ where $H_x(i) = r$ if the i th tie occurs as an r-tuplet in the x-ranking and similarly for $H_y(j)$.

(iii) If $N_x < N_y$ then assign X to objects O and Y to parcels P etc.; the rest of the assignment statement conforming to our earlier observation that the complexity of the h functions to be multiplied may be reduced by appropriate choice of which ranking is to be permuted.

(iv) Generate the homogeneous product sums of weight $H_p(i)$; $i = 1, \ldots, N_p$ as the set of N_o-part compositions of $H_p(i)$.

(v) Multiply the homogeneous product sums to obtain terms $\alpha_1^{p_1} \cdots \alpha_{N_o}^{p_{N_o}}$. Retain only those products for which $p_j = H_o(j)$; $j = 1, \ldots, N_o$.

(vi) For each retained product use its component terms to generate a permutation vector.

(vii) Compute the appropriate score, S or D, for this permutation vector and the parcels vector. Record the score. Increment the probability of obtaining a score with this value by the probability of obtaining this particular configuration, A_m say, of the two rankings. This probability may be obtained by a generalization of eqn. (4.5), Burr (1960). Let a_{ij} denote the ij th element of Table 2.4. The desired probability is then obtained as

$$\Pr(A_m) = \frac{\prod_i H_o(i)! \prod_j H_p(j)!}{n! \prod_i \prod_j a_{ij}!} . \qquad (2.7)$$

For example, $\Pr(A_m) = 3/35$ for both Tables 2.2 and 2.3.

(viii) Determine the next permutation vector and go to (vii). If all permutation vectors have been exhausted output the distribution of S.

(ix) If there are no ties in the rankings, a more efficient approach to enumerating the distribution of D is to use an efficient permutation generator and record the associated scores. Reingold, Nievergelt and Deo (1977, Section 5.1) state that their algorithm 5.3 is one of the most efficient for generating permutations. This algorithm is therefore used whenever $N_x = N_y = n$.

2.2.3 Generation of a k-part composition of n

Following Feller (1968, Section 2.5), the k-part compositions of n may be regarded as an occupancy problem of placing n balls into k cells where empty cells are permissible. Then, as proven by Feller, there are $\binom{k+n-1}{k-1}$ such compositions. Consequently, as Ehrlich (1973) has noted, a sequence of all compositions of n to k nonnegative terms can be received from a sequence of all combinations of $k-1$ out of $n+k-1$ ($k-1$ ones and n zeroes), by relating to each combination C_b a composition, $C[1:k]$ defined by: $C(i) =$ the number of zeroes between the $(i-1)^{\text{th}}$ 1 of C_b and the i^{th} 1 of C_b where it is assumed that $C_b(0) = C_b(n+k) = 1$.

Our current problem reduces to generating all combinations of $N_o - 1$ out of $H_p(i) + N_o - 1$. Algorithm 7 of Ehrlich (1973) which generates combinations in minimal change order is used to generate the combinations.

2.2.4 Multiplication of the homogeneous product sums

Since the product terms of $\prod_{i=1}^{N_p} h_{H_p(i)}$ result from multiplying polynomials in $\alpha_1, \alpha_2, \ldots, \alpha_{N_o}$, where each term of each polynomial has coefficient 1, it suffices to add the powers of like α_i terms in the product component terms in order to arrive at the product terms. Hence only sums of the vector representations of the component terms of $h_{H_p(i)}$ need be generated. Let $C(i,\cdot)$ be an N_o-part composition of $H_p(i)$ and $N_c(i)$ be the number of such compositions. Let $(C(1,\cdot), C(2,\cdot), \ldots, C(N_p,\cdot))$ denote a set of vectors such that the constraint $\sum_{i=1}^{N_p} C(i,\cdot) = H_o$, where $C(i,\cdot)$ and H_o are both vectors, is satisfied. Each $C(i,\cdot)$ is a member of a finite, linearly ordered set $h_{H_p(i)}$. Thus an exhaustive search, which considers the elements of $h_{H_p(1)} \times h_{H_p(2)} \times \cdots \times h_{H_p(N_p)}$ as potential solutions, is required. A backtrack algorithm is an efficient method of implementing such a search, Reingold, Nievergelt and Deo (1977, Section 4.1).

There are $\prod_{i=1}^{N_p} N_c(i)$ different sums which may be generated. Preclusion of a great many of these sums may be obtained by observing that once any element of $\sum_{i=1}^{m} C(i,\cdot)$, $m < N_p$ exceeds the corresponding element of H_o then all the subtrees associated with the current track may be removed since they will all yield sums which violate the constraint. The algorithm thus developed is

(i) Initialize $j = 1$.

(ii) Set $\text{loc}(j) = 1$.

(iii) If $j = N_p$ go to step (v). Otherwise go to step (iv).

(iv) Add the current composition vector to the partial sum vector. If the partial sum is admissible set $j = j + 1$ and go to step (ii). Otherwise go to step (vii); thus excluding unnecessary subtrees.

(v) Call a routine to implement addition of the final composition vector. This routine returns a permutation vector if the resulting sum is admissible.

(vi) If a permutation vector is returned go to step (ix). Otherwise go to step (vii).

(vii) If $\text{loc}(j) = N_c(j)$ go to step (viii). Otherwise set $\text{loc}(j) = \text{loc}(j) + 1$ and go to step (iii).

(viii) Set $j = j - 1$. If $j = 0$ exit. Otherwise go to step (vii).

(ix) Compute the score and its probability of occurrence. Increment the probability vector. Go to step (viii); since there is at most one composition for $j = N_p$ which can lead to an admissible sum.

Specification of a permutation vector is effected by scanning the product component terms from left to right and from $C(1, \cdot)$ through $C(N_p, \cdot)$. Whenever a non-zero entry of size k is encountered in position i; $0 \leq i \leq N_o$, then k $r_m(i)$'s are placed into the permutation vector; $r_m(i)$ being the midrank corresponding to $H_o(i)$. Lehmann (1975, (A.14)) gives $r_m(i)$ as $r_m(i) = \frac{1}{2}(H_o(i) + 1) + \sum_{j=0}^{i-1} H_o(j)$ where $H_o(0) = 0$.

2.3 Further examples and conclusion

Fortran computer programs, which implement the algorithms of Sections 2.1 and 2.2, have been developed. These programs for enumerating the distributions of Kendall's S and Spearman's D were run on a VAX 11/785 minicomputer with several examples given by Burr (1960), Klotz (1966) and Kendall (1975). A modified version, of the combined programs, which computes upper tail probabilities is listed in the Appendix to Chapters 2 and 3. The reported times are the CPU times taken to implement steps (iii) through (ix) of the algorithm given in Section 2.2.2; these being computed in subroutine MULTHPS.

Example 2.1: Let $n = 10$, $H_x = (3,4,3)$ and $H_y = (2,3,3,2)$. In the array whose score is S_1, the highest score, Burr (1960, example 8.2) shows that there are precisely four transpositions which reduce the score by 5. Thus there are five arrays with scores $S_1 = 29$ (one array) and $S_2 = 24$ (four arrays). Burr then states, "Perhaps less obviously, there are precisely four ways of obtaining the score 20 (in each case by a further transposition), from which we can show that $S_3 = 20$, $P_2(S_3) = 6 + 6 + 6 + 6 = 24$ and the significance level of S_3 is 81/4200".

Our algorithm, which takes 0.3 seconds to generate the distribution of S, yields a significance level of 69/4200 which indicates that $P_2(S_3) = 6 + 6 = 12$, this corresponding to two further transpositions. Burr's error occurs because a further transposition, of each of the four S_2 arrays, leads to only two distinct arrays. A feature of this problem is that successive transpositions will, sooner or later, lead to duplication. Therefore, unless time consuming checks are made to protect against such duplications, erroneous significance levels will result.

Example 2.2: Let $n = 10$, $H_x = (1,2,2,2,1,2)$ and $H_y = (1,1,4,3,1)$. Kendall (1975, example 3.1) discussed this example for which both N_x and N_y are fairly large. The distribution of S was generated in 5.6 seconds.

In the notation of Section 2.2, let $N_y = 2$ so that there is a dichotomy in the Y ranking. Then Kendall's S test is equivalent to the Wilcoxon rank sum and the Mann-Whitney two-sample tests.

Example 2.3: For the case where $N_x = n$ so that there are no ties in the X ranking, the algorithm of Section 2.1 may be used to rapidly find the null distribution of S for $n \leq 20$. Let M_y denote the number of tied elements in the larger of the two tuplets. For $M_y = (10, 11, 12, 13, 14, 15, 16, 17, 18, 19)$ and $n = 20$ the corresponding machine times, in seconds, are (12.7, 10.3, 7.0, 4.0, 1.8, 0.7, 0.2, 0.1, 0.02, 0.01). The distribution obtained for the case where $M_y = 10$ and $n = 20$ is identical to that given by Lehmann (1975, Table B) who gives the probabilities to four digit accuracy. Fellingham ar .oker (1964) reported that an Edgeworth approximation almost coincides with the exact distribution when $n = 20$. However, they neglected to specify the value of M_y used. We investigated their claim and found it to be true providing $M_y \leq 15$. For larger values of M_y the approximation, while still reasonably satisfactory, starts to break down.

Example 2.4: Klotz (1966) developed an algorithm for computing the exact null distribution of the Wilcoxon test statistic when $N_x < n$ so that there are ties in the X ranking. Klotz computed eight exact distributions for values of n ranging from 10 to 22. Table 2.5 shows the machine time, in seconds, taken for each of these eight cases.

Table 5: Machine time for 8 selected distributions

n	H_x	H_y	*Time*
10	(1,1,2,1,1,2,1,1)	(5,5)	0.17
10	(1,1,1,2,1,1,2,1)	(5,5)	0.17
10	(1,3,1,2,1,1,1)	(5,5)	0.11
14	(1,2,1,1,1,1,2,2,1,2)	(6,8)	1.81
22	(3,3,4,3,4,5)	(10,12)	4.26
21	(1,2,3,4,5,6)	(11,10)	2.07
21	(6,5,4,3,2,1)	(11,10)	2.05
16	(2,2,2,2,2,2,2,2)	(8,8)	2.70

Example 2.5: Let $n = 12$, $H_x = (7,5)$ and $H_y = (6,3,3)$. Burr (1960, example 9.3) found that the normal approximation is not very good for this distribution as well as for several other distributions with $n = 9, 10, 12$. He then suggested using the method which led to the erroneous significance level as discussed in Example 2.1 above. The algorithm of Section 2.2, which takes 0.02 seconds to generate the specified distribution, is recommended.

Example 2.6: Let $n = 17$, $H_x = (12,5)$ and $H_y = (4,1,2,5,3,1,1)$. Burr (1960, example 9.4), Kendall (1975, examples 3.4 and 3.5) and Whitfield (1947) have all discussed this example. It is included here since n is fairly large. Our algorithm takes 0.5 seconds to generate the distribution of S.

When there are ties in both rankings, the time taken for enumerating the null distributions of D is less than that taken for S since computation of the score corresponding to each enumeration requires $O(n)$ and $O(n^2)$ steps, respectively. For the case where there are ties in only one of the rankings, the algorithm of Section 2.2 may be used if n is fairly small.

Example 2.7: Let $n = 11$, $N_x = n$ and $H_y = (3,3,3,2)$. The machine time taken to generate the distribution of D is 1.6 minutes. If n increases to 12, with $H_y = (3,3,3,3)$, the time taken increases to 6.7 minutes.

Given the complexity of the algorithm, Section 2.2.2, it is reassuring to note that it possesses a self-checking feature. Since the probability of each admissible permutation is calculated separately, then any error in the application of the algorithm should result in a distribution whose probabilities do not sum to one. Given the advent of supercomputers with parallel processors it is anticipated that the above algorithms will be capable of computing the exact distributions of S or of ρ_s, in the presence of ties, for moderately large n. However, enumeration of the distribution of S for example 4.3 of Kendall (1975) which has $n = 12$ and $N_x = N_y = 8$ currently takes three hours of machine time from which it follows that a measure of the time, T, which it takes to implement the enumeration process is necessary. This Chapter thus concludes with the development of an iterative method for estimating the remaining time required to complete the enumeration.

The tree pruning technique, used during multiplication of the homogeneous product sums, guarantees that the multiplication process is always converging toward an admissible solution and, therefore, it is anticipated that the time taken to implement the multiplication will be approximately proportional to the number of permutations, N_{pe}, of X against Y. For each admissible permutation, the associated score must then be computed so that T will also depend on

(i) The number of computations, a function of N_x and N_y, required to generate a permutation vector and to calculate the product $\prod a_{ij}!$.

(ii) The number of computations, $O(n)$ and $O(n^2)$ for D and S, respectively, required to then evaluate the score.

(iii) A constant number of computations associated with the enumeration.

The relationship between T, N_{pe}, N_x, N_y and n is subsequently considered.

Data were obtained by randomly generating partitions of n, for $n = 10(5)30$, corresponding to supplied values of N_x and N_y. N_x and N_y ordered data points were then assigned via an index vector, obtained as a random permutation of the first n natural integers, to data vectors X and Y. Values of (N_x, N_y) used range from (2, 2) to the largest values which permit computation of the scores in less than two minutes. Runs for which $N_{pe} < 18$ or $T > 110$ seconds were discarded.

Since computation of S requires $O(n^2)$ steps, the time requirement in item (ii) will dominate that of item (i) which may therefore be incorporated into item (iii). It then follows that a model of the form

$$T_k = \beta_1 N_{pe} n^2 + \beta_2 N_{pe} \tag{2.8}$$

may be expected to explain most of the variation in T_k when enumerating the null distribution of S. Figure 2.1 shows a plot of T_k versus N_{pe} which confirms that, for fixed n, T_k is roughly proportional to N_{pe}. Regression analysis confirmed the adequacy of the model given in eqn. (2.8) with the two explanatory variables accounting for 99.24% of the variation in T_k. Thus an estimated value of N_{pe} allows an estimate of T_k to be obtained.

For enumeration of the null distribution of D, Figure 2.2 reveals a much less clear-cut relationship between T_s, N_{pe} and n. This is not surprising since the factors N_x and N_y become more important as only n computations are required to evaluate the score D. However, as is evident from Figure 2.3, the relationship between T_k and T_s is roughly linear for fixed n. In fact, a model of the form

$$T_s = \beta_1 T_k / n \tag{2.9}$$

accounts for 98.9% of the variation in T_s and this suggests that a model of the form

$$T_s = \beta_1 N_{pe} n + \beta_2 N_{pe} \ , \tag{2.10}$$

Figure 2.1: Plot of Time Taken vs Number of Permutations

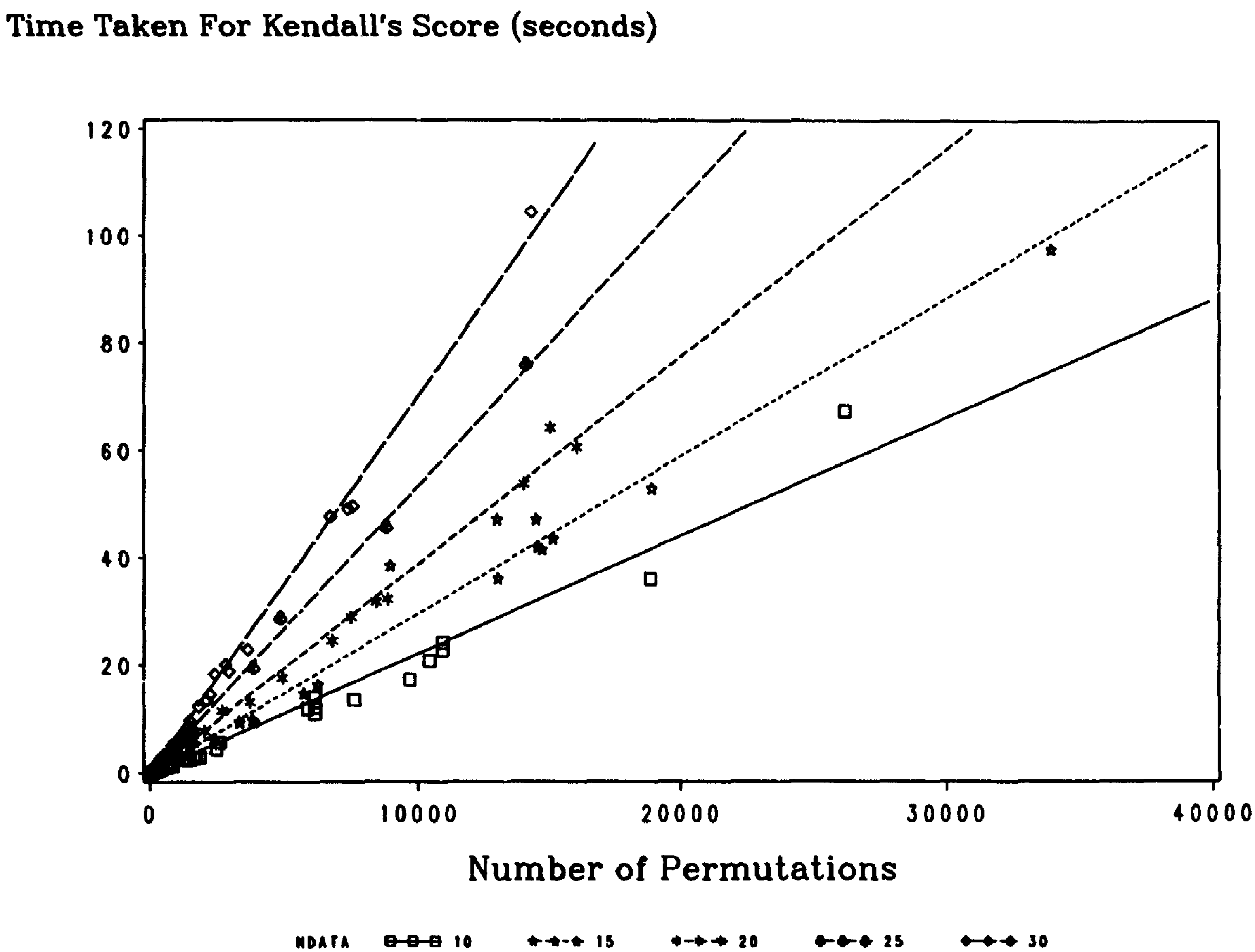


Figure 2.2: Plot of Time Taken vs Number of Permutations

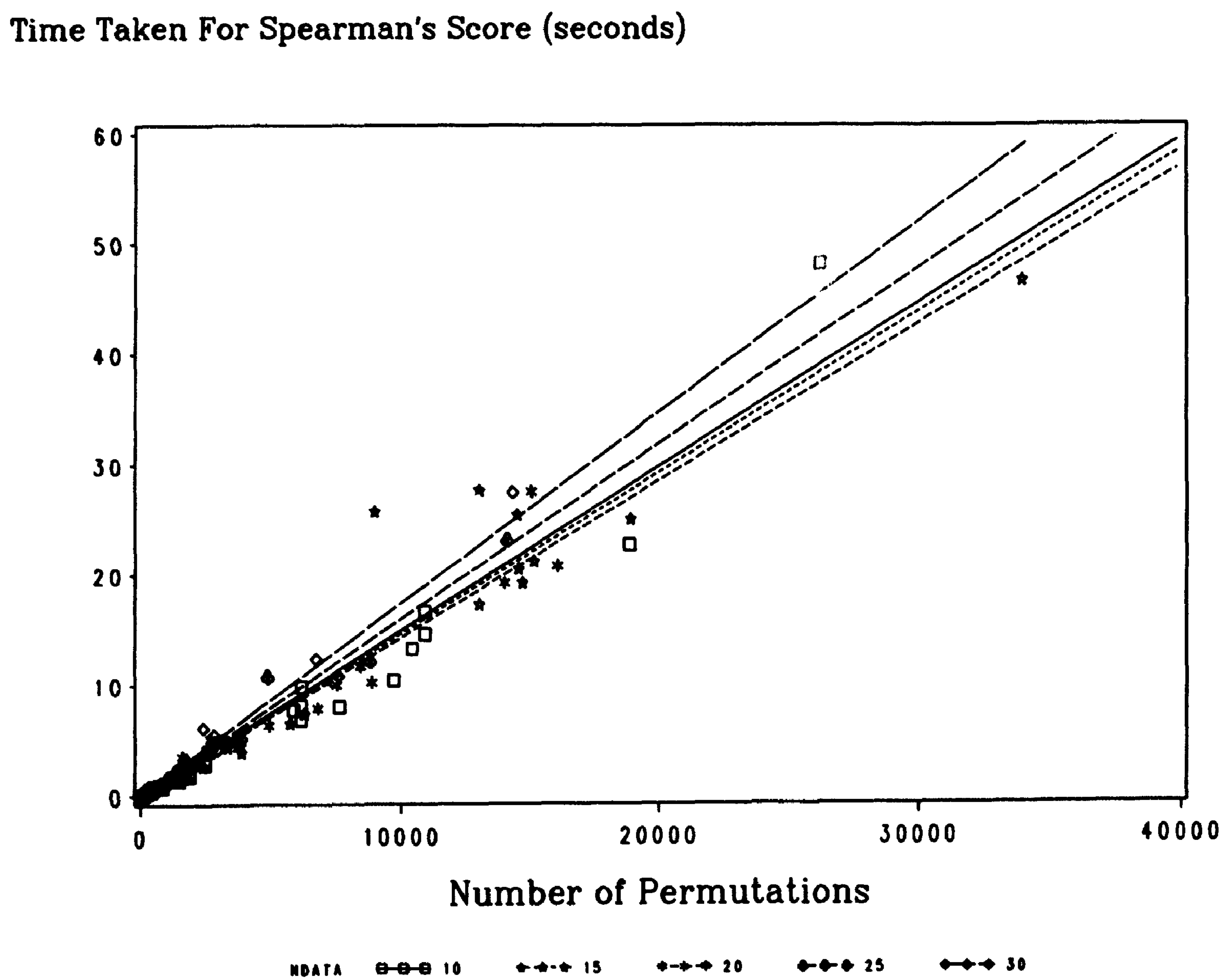
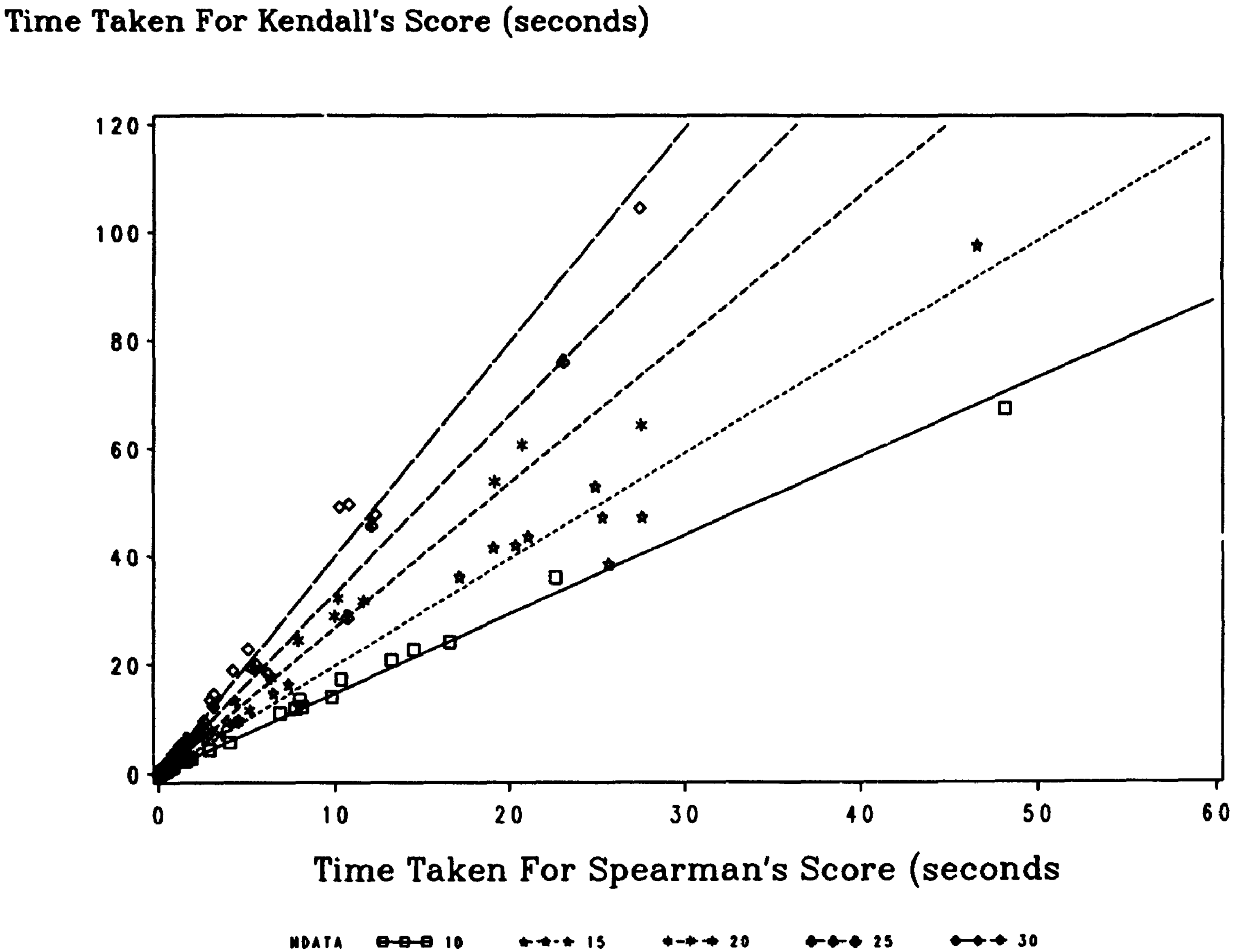


Figure 2.3: Plot of Time Taken For S vs Time Taken for D



where item (i) above is once again incorporated into item (iii), will be adequate for prediction of T_s. The model given in eqn. (2.10) accounts for 96.1% of the variation in T_s so that, as for T_k, an estimated value of N_{pe} allows an estimate of T_s to be obtained. It therefore remains to obtain an estimate of N_{pe}.

It is shown in the development of eqn. (3.30) in Chapter 3 that

$$\sum_A \frac{1}{\prod_i \prod_j a_{ij}!} = \frac{n!}{\prod_i H_x(i)! \prod_j H_y(j)!} \tag{2.11}$$

so that

$$N_{pe} = \frac{1}{\bar{a}} \cdot \frac{n!}{\prod_i H_x(i)! \prod_j H_y(j)!} \tag{2.12}$$

where $\bar{a}$ is the mean value of $\prod_{ij} a_{ij}!$. It is easy to maintain an estimate of $\bar{a}$ based on the number of permutations already enumerated so that an estimate of N_{pe} and, therefore, an estimate of the time remaining to completion may be obtained after, for example, every 5000 enumerations. This information then enables the user to either allow the enumeration to continue or to abort the enumeration and switch to an approximation technique for obtaining the significance levels of observed data.

REFERENCES

Burr, E.J. (1960). The distribution of Kendall's score S for a pair of tied rankings. *Biometrika*, **47**, 151–171.

Ehrlich, G. (1973). Loopless algorithms for generating permutations, combinations and other combinatorial configurations. *Journal of the Association for Computing Machinery*, **20**, 500–513.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (Vol. 1). Wiley.

Fellingham, S.A. and Stoker, D.J. (1964). An approximation for the exact distribution of the Wilcoxon test for symmetry. *Journal of the American Statistical Association*, **59**, 899–905.

Franklin, L.A. (1987a). The complete exact null distribution of Spearman's rho for $n = 12(1)16$. *Proceedings of the 19th Symposium on the Interface between Statistics and Computer Science*, 1987, 337–342.

Franklin, L.A. (1987b). Approximations, convergence and exact tables for Spearman's rank correlation coefficient. *Proceedings of the Statistical Computing Section of the American Statistical Association*, 1987, 244–247.

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

Kendall, M.G. (1975). *Rank Correlation Methods* (4th ed). Griffin and Co. Ltd.

Klotz, J.H. (1966). The Wilcoxon, ties, and the computer. *Journal of the American Statistical Association*, **61**, 772–787.

Lamb, H.H. (1970). Volcanic dust in the atmosphere; with a chronological assessment of its meteorological significance. *Philosophical Transactions of the Royal Society of London*, **266**, 425–527.

Lehmann, E.L. (1975). *Nonparametrics: Statistical methods based on ranks.* Holden-Day.

MacMahon, P.A. (1915). *Combinatory Analysis* (Vol. 1). Cambridge University Press.

Manley, G. (1974). Central England temperatures. *Quarterly Journal of the Royal Meteorological Society*, **100**, 679–681.

Mann, H.B. (1945). Non-parametric tests against trend. *Econometrica*, **13**, 245–259.

Panneton, M. and Robillard, P. (1972). Algorithm AS 54. Kendall's S frequency distribution. *Applied Statistics*, **21**, 345–348.

Reingold, E.M., Nievergelt, J. and Deo, N. (1977). *Combinatorial Algorithms: Theory and Practice.* Prentice Hall.

Sillitto, G.P. (1947). The distribution of Kendall's τ coefficient of rank correlation in rankings containing ties. *Biometrika*, **34**, 36–40.

Whitfield, J.W. (1947). Rank correlation between two variables, one of which is ranked, the other dichotomous. *Biometrika*, **34**, 292–296.

Chapter 3

SOME NEW RESULTS ON THE DISTRIBUTION OF KENDALL'S SCORE

Robillard's (1972) approach to obtaining an expression for the cumulant generating function of the null distribution of Kendall's S statistic, when one ranking is tied, is extended to the general case where both rankings are tied. An expression is obtained for the cumulant generating function which is used to provide a simple proof of the asymptotic normality of the score S when both rankings are tied. A new approach to deriving the variance of S, in the absence of ties, is presented. The method used to obtain an expression for the cgf of S is then applied to derive the mean and variance of S for the general case of ties in both rankings. The third cumulant of S is derived and an approximation to the fourth cumulant is obtained. The usefulness of an Edgeworth approximation to the null distribution of S, in the general case of tied rankings, is investigated and compared with the standard normal approximation.

3.1 The asymptotic normality of Kendall's score

Kendall's score may be written

$$S = \sum_{i<j}^{n} \operatorname{sign}((X_j - X_i)(Y_j - Y_i)) \tag{3.1}$$

where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are n independent replications of the random variable (X, Y). In many situations due to discreteness of the actual distribution or due to censoring, the distributions of X and Y are non-continuous. Let k and ℓ denote the number of distinct values assumed in a particular realization of the random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let α_i $(i = 1, \ldots, k)$ and β_j $(j = 1, \ldots, \ell)$ denote the ordered distinct values for the X's and Y's, respectively. Then, as shown by Burr (1960), the observed Kendall score, S, is equal to the sum of all second-order determinants of the matrix $A = (a_{ij})$, where a_{ij} is the number of times that $(x_g, y_g) = (\alpha_i, \beta_j)$. The extent of the observed ties for the X's and Y's are denoted by u_i $(i = 1, \ldots, k)$ and v_j $(j = 1, \ldots, \ell)$, respectively, and are given by

$$u_i = \sum_{j=1}^{\ell} a_{ij} \qquad \text{and} \qquad v_j = \sum_{i=1}^{k} a_{ij}. \tag{3.2}$$

Notice that $\sum_i u_i = \sum_j v_j = n$. Given an observed matrix A, the null distribution of S for testing that X and Y are independent is the distribution of S conditional on the observed row and column totals $u = (u_1, \ldots, u_k)$ and $v = (v_1, \ldots, v_\ell)$. Let $S_{u,v}$ denote the random variable with this distribution. In the case of no ties, $u_i = 1$ $(i = 1, \ldots, k)$ and $v_j = 1$ $(j = 1, \ldots, \ell)$, the random variable is denoted by S_n. If there are ties in only one ranking, say the X-ranking, the random variable is denoted by S_u.

Robillard (1972) obtained an expression for the cumulant generating function (cgf) of the score S_u by expressing the score S_n as a sum of the score S_u and scores S_{u_i} $(i = 1, \ldots, k)$. This approach, when extended to a consideration of the score $S_{u,v}$, yields an expression for the cumulant generating function of $S_{u,v}$ which is subsequently used to provide a simple proof of the asymptotic normality of $S_{u,v}$ under some trivial conditions on the relative growth rates of n and the maximum extent of a tie in either ranking. Kendall (1975, chapter 5) has noted that a simple proof of the normality of $S_{u,v}$, which follows as a consequence of general results obtained by Hoeffding (1948), is not easy to give.

3.1.1 Two fundamental variable transformation relationships

Consider the rankings R_x and R_y of $(X_1, Y_1), \ldots, (X_n, Y_n)$. A score $S_n | A$, corresponding to two untied rankings of size n, may be obtained by

(i) Computing $S_{u,v}|A$.

(ii) Untying the tied ranks v_1 through v_ℓ to generate a score $\sum_{j=1}^{\ell} S_{s_j}$. The score S_{s_j} may be regarded as a score obtained on v_j observations when there are ties of extent $a_{1j}, \ldots, a_{kj}$ in only one ranking.

(iii) Next untying the tied ranks u_1 through u_k to generate a score $\sum_{i=1}^{k} S_{u_i}$.

It immediately follows that

$$S_n|A = S_{u,v}|A + \sum_{j=1}^{\ell} S_{s_j} + \sum_{i=1}^{k} S_{u_i} \tag{3.3}$$

where the $(k+\ell+1)$ variables on the right of eqn. (3.3) are independent. Using Robillard's (1972) relationship for the scores when only one ranking is tied gives

$$S_{v_j} = S_{s_j} + \sum_{i=1}^{k} S_{a_{ij}} \tag{3.4}$$

where the $(k+1)$ variables on the right hand side of eqn. (3.4) are independent.

Let the rankings R_x and R_y be obtained by replacing observations by their midranks. Then a tie of length v_j represents the repetition of the mean of v_j consecutive integers. Randomly replacing these v_j identical ranks by the corresponding integers changes the score in (i) to $S_{u,v}|A + S_{s_j}$ where there are still ties in R_x to be accounted for. Repeated application of this procedure, firstly to ties in R_y and then to ties in R_x, yields eqn. (3.3). Starting with ties in only one ranking and using the same argument leads to eqn. (3.4).

3.1.2 Proof of the asymptotic normality of $S_{u,v}$

The characteristic function of the random variable $S_n|A$ is

$$\mathcal{E}(e^{itS_n|A}) = \mathcal{E}(e^{itS_{u,v}|A})\mathcal{E}(e^{it\sum_{j=1}^{\ell} S_{s_j}})\mathcal{E}(e^{it\sum_{i=1}^{k} S_{u_i}}) . \tag{3.5}$$

Using eqn. (3.4) to obtain an expression for the characteristic function of S_{v_j}, solving the resulting expression for $\mathcal{E}(e^{it\sum_{j=1}^{\ell} S_{s_j}})$ and substituting into eqn. (3.5), then gives

$$\mathcal{E}(e^{itS_{u,v}|A}) = \frac{\mathcal{E}(e^{itS_n|A})\mathcal{E}(e^{it\sum_i\sum_j S_{a_{ij}}})}{\prod_i \mathcal{E}(e^{itS_{u_i}})\prod_j \mathcal{E}(e^{itS_{v_j}})} . \tag{3.6}$$

Applying the fact that

$$\mathcal{E}(e^{itS_{u,v}}) = \mathcal{E}_A\big(\mathcal{E}(e^{itS_{u,v}|A})\big)$$

to eqn. (3.6) and taking logs yields

$$\log \mathcal{E}(e^{itS_{u,v}}) = \log \mathcal{E}_A\Big\{\mathcal{E}(e^{itS_n|A})\mathcal{E}(e^{it\sum_i\sum_j S_{a_{ij}}})\Big\} - \sum_i K_{u_i}(t) - \sum_j K_{v_j}(t) \tag{3.7}$$

where $K_m(t)$ is the cgf of the score for two untied rankings of m elements. Rewriting the cumulant generating function for $S_{u,v}$ in terms of the standard deviation $\sqrt{\kappa_2}$ as unit then gives

$$\log \mathcal{E}(e^{it\frac{S_{u,v}}{\sqrt{\kappa_2}}}) = \log \mathcal{E}_A\Big\{\mathcal{E}(e^{it\frac{S_n|A}{\sqrt{\kappa_2}}})\mathcal{E}(e^{it\sum_i\sum_j \frac{S_{a_{ij}}}{\sqrt{\kappa_2}}})\Big\} - \sum_i K_{u_i}\Big(\frac{t}{\sqrt{\kappa_2}}\Big) - \sum_j K_{v_j}\Big(\frac{t}{\sqrt{\kappa_2}}\Big) . \tag{3.8}$$

Billingsley (1986, eqn. (26.5)) gives the inequality

$$\Big|\mathcal{E}(e^{itx}) - \sum_{g=0}^{m}\frac{(it)^g}{g!}\mathcal{E}(x^g)\Big| \le \mathcal{E}\Big[\min\Big\{\frac{|\,tx\,|^{m+1}}{(m+1)!}, \frac{2|\,tx\,|^m}{m!}\Big\}\Big] \tag{3.9}$$

so that

$$\Big|\mathcal{E}(e^{it\sum_i\sum_j\frac{S_{a_{ij}}}{\sqrt{\kappa_2}}}) - \sum_{g=0}^{1}\frac{(it)^g}{g!}\mathcal{E}\Big\{(\sum_i\sum_j\frac{S_{a_{ij}}}{\sqrt{\kappa_2}})^g\Big\}\Big| \le \frac{t^2}{2!\,\kappa_2}\mathcal{E}\Big\{(\sum_i\sum_j S_{a_{ij}})^2\Big\} \tag{3.10}$$

which converges to zero as $n \to \infty$ provided that $\mathcal{V}ar(\sum_i\sum_j S_{a_{ij}})$ is of lower order in n than is $\mathcal{V}ar(S_{u,v})$. Therefore, since $\mathcal{E}(\sum_i\sum_j S_{a_{ij}}) = 0$ and assuming that the convergence condition is satisfied, Billingsley's inequality yields

$$\mathcal{E}(e^{it\sum_i\sum_j\frac{S_{a_{ij}}}{\sqrt{\kappa_2}}}) \longrightarrow 1 . \tag{3.11}$$

Substituting from eqn. (3.11) into eqn. (3.8) then yields the result that, for large n,

$$\log \mathcal{E}(e^{it\frac{S_{u,v}}{\sqrt{\kappa_2}}}) \longrightarrow K_n\left(\frac{t}{\sqrt{\kappa_2}}\right) - \sum_i K_{u_i}\left(\frac{t}{\sqrt{\kappa_2}}\right) - \sum_j K_{v_j}\left(\frac{t}{\sqrt{\kappa_2}}\right) . \tag{3.12}$$

Let $M_u = \max(u_i)$ and $M_v = \max(v_j)$. Then for large n,

$$\begin{aligned} Var\left(\sum_i \sum_j S_{a_{ij}}\right) &< \frac{n^2}{M_u M_v}\left(\frac{1}{9}\sqrt{M_u^3 M_v^3} + \frac{1}{2} M_u M_v\right) \\ &= n^2\left(\frac{1}{2} + \frac{1}{9}\sqrt{M_u M_v}\right) \end{aligned} \tag{3.13}$$

and

$$\begin{aligned} Var(S_{u,v}) &> \frac{1}{9}\left(n^3 - \frac{n}{M_u}\cdot M_u^3 - \frac{n}{M_v}\cdot M_v^3\right) \\ &= \frac{n}{9}(n^2 - M_u^2 - M_v^2) , \end{aligned} \tag{3.14}$$

so that eqn. (3.11) holds if both M_u and M_v are each of order less than $n^{2/3}$. Note that $\max(a_{ij}) \leq \sqrt{M_u M_v}$. Robillard (1972) has shown that a sufficient condition for $K_n\left(\frac{t}{\sqrt{\kappa_2}}\right) - \sum_i K_{u_i}\left(\frac{t}{\sqrt{\kappa_2}}\right)$, where κ_2 is now based on $v_j = 1$ for all j, to converge to $-\frac{1}{2}t^2$ is that the value of M_u does not increase at a rate greater than or equal to the order of $n^{2/3}$. It therefore follows that a sufficient condition for the asymptotic normality of the distribution of $S_{u,v}$ is that both M_u and M_v be of lower order than $n^{2/3}$.

The expression, eqn. (3.8), obtained for the cumulant generating function of $S_{u,v}$ indicates that improvement to the normal approximation via an Edgeworth series expansion such as is possible for the cases of no ties, David et al. (1951), and ties in only one ranking, Robillard (1972), requires direct evaluation of the moments of $S_{u,v}$.

3.2 A simplified derivation of the variance of Kendall's S

Let R_1 and R_2 be the rankings of n individuals with respect to two criteria and assume, initially, that there are no ties in either ranking. Then, without loss of generality, it may also be assumed that R_2 is in its natural order so that $R_2 = (1, 2, \cdots, n)$. Let $R_1 = (r_1, r_2, \cdots, r_n)$. Then the negative score, Q, is given by

$$Q = \sum_{i>j} I_{(0,\infty)}(r_j - r_i), \tag{3.15}$$

where $I_{(0,\infty)}(\bullet)$ denotes the indicator function on $(0, \infty)$. Kendall's score S (Kendall 1975, eqns. 1.3 and 1.5) is then given by,

$$S = \frac{1}{2}n(n-1) - 2Q. \tag{3.16}$$

A rather lengthy derivation of the variance of S was given by Kendall (1975, chapter 5) for the general case of tied ranks. Noether (1967, chapter 10) presented a more concise approach. However, for the case when the two criteria are assumed to be independent and continuous, the derivation given in Section 3.2.1 is more direct than these other approaches. In Section 3.2.2, the derivation is extended to the case where there are ties in both R_1 and R_2.

The notion of an inversion vector provides the basis for our derivation. Reingold, Nievergelt and Deo (1977, section 5.1.2) defined an inversion vector, $I_k = (i_1, i_2, \cdots, i_k)$, as follows.

Let $X = (x_1, x_2, \cdots, x_k)$ be a sequence of numbers. A pair (x_ℓ, x_j) is called an inversion of X if $\ell < j$ and $x_\ell > x_j$. The inversion vector of X is the sequence of integers $i_1, i_2, \cdots, i_k$ obtained by letting i_j be the number of x_ℓ such that (x_ℓ, x_j) is an inversion. Hence i_j is the number of elements greater than x_j and to its left in the sequence. Note that $0 \le i_j \le j - 1$. For example, the inversion vector for the permutation $P = (4, 3, 5, 2, 1, 7, 8, 6, 9)$ is $I = (0, 1, 0, 3, 4, 0, 0, 2, 0)$. It may be proven by induction that each inversion vector uniquely represents a permutation of the first k natural numbers.

3.2.1 Derivation of the variance

Let I_n be the inversion vector corresponding to the ranking R_1 so that

$$I_n = (0, i_2, i_3, \cdots, i_n), \qquad 0 \le i_j \le j-1.$$

It follows from the definitions of Q and I_n that

$$Q = \sum_{j=1}^{n} i_j \, . \tag{3.17}$$

Under the assumption of independent rankings, inversion vectors are equiprobable. Since the set of $n!$ inversion vectors may be divided into $(n!/j)$ subsets of j inversion vectors so that members of the same subset differ only on the j^{th} element, it then follows that each of the j possible values $(0, 1, \cdots, j-1)$ of i_j have probability j^{-1}. Hence

$$\mathcal{E}(i_j) = (j-1)/2 \tag{3.18}$$

and, consequently,

$$\mathcal{E}(Q) = \sum_{j=1}^{n} \mathcal{E}(i_j) = \frac{1}{2}\sum_{j=1}^{n}(j-1) = \frac{1}{2}\binom{n}{2}. \tag{3.19}$$

Similarly,

$$\mathcal{E}(i_j^2) = \sum_{i_j} {i_j}^2 \Pr(i_j) = (j-1)(2j-1)/6 \tag{3.20}$$

and, from the independence of i_j and i_ℓ,

$$\begin{aligned}\sum_{j \ne \ell}^{n} \mathcal{E}(i_j i_\ell) &= \frac{1}{4}\sum_{j \ne \ell}^{n}(\ell-1)(j-1) \\ &= \left(\frac{1}{2}\sum_{j=1}^{n}(j-1)\right)^2 - \sum_{j=1}^{n}\frac{1}{4}(j-1)^2 \, .\end{aligned} \tag{3.21}$$

Consequently,

$$\mathcal{E}(Q^2) = \mathcal{E}\left(\sum_{j=1}^{n}\sum_{\ell=1}^{n} i_j i_\ell\right)$$

$$= \left(\frac{1}{2}\binom{n}{2}\right)^2 - \frac{1}{4}\sum_{j=1}^{n}(j^2 - 2j + 1) + \frac{1}{6}\sum_{j=1}^{n}(2j^2 - 3j + 1)$$

$$= \left(\frac{1}{2}\binom{n}{2}\right)^2 + \frac{n}{72}(n-1)(2n+5) \; . \qquad (3.22)$$

Hence

$$\mathcal{V}ar(Q) = n(n-1)(2n+5)/72 \qquad (3.23)$$

and

$$\mathcal{V}ar(S_n) = n(n-1)(2n+5)/18 \; . \qquad (3.24)$$

3.2.2 Extension to incorporate ties in both rankings

It is known that

$$\mathcal{V}ar(S_n) = \mathcal{E}_A\big(\mathcal{V}ar(S_n|A)\big) + \mathcal{V}ar_A\big(\mathcal{E}(S_n|A)\big) \qquad (3.25)$$

where, from eqns. (3.3) and (3.4),

$$\mathcal{E}(S_n|A) = S_{u,v}|A \qquad (3.26)$$

since there exists only one value of $S_{u,v}$ for each matrix A and, from independence,

$$\mathcal{V}ar(S_n|A) = \sum_{j=1}^{\ell} \mathcal{V}ar\big(S_{e_j}\big) + \sum_{i=1}^{k} \mathcal{V}ar\big(S_{u_i}\big) \; . \qquad (3.27)$$

Eqn. (3.26) suffices to show that $\mathcal{E}(S_{u,v}) = 0$. From eqn. (3.4) it follows that

$$\mathcal{V}ar\big(S_{e_j}\big) = \mathcal{V}ar(S_{v_j}) - \sum_{i=1}^{k} \mathcal{V}ar(S_{a_{ij}}) \; . \qquad (3.28)$$

Substituting from eqns. (3.26), (3.27) and (3.28) into eqn. (3.25) then yields

$$\mathcal{V}ar(S_n) = \mathcal{V}ar(S_{u,v}) + \sum_i \mathcal{V}ar(S_{u_i}) + \sum_j \mathcal{V}ar(S_{v_j}) - \mathcal{E}_A\Big\{\sum_i\sum_j \mathcal{V}ar(S_{a_{ij}})\Big\} \qquad (3.29)$$

so that $\mathcal{V}ar(S_{u,v})$ is immediately determined upon evaluation of $\mathcal{E}_A\big(\mathcal{V}ar(S_{a_{ij}})\big)$.

At this point it is important to note that

$$\Pr(A) \neq \frac{1}{\text{\# of different values of } A} .$$

It is necessary to allow for the fact that some configurations occur more frequently than others. Let n_a be the number of ways of untying A along R_2 so that $n_a = \prod_j (v_j!/\prod_i a_{ij}!)$ and $\sum_A n_a = n!/\prod_i u_i!$. Then the relative frequency of A is obtained as

$$\Pr(A) = \frac{n_a}{\sum_A n_a} = \frac{\prod_i u_i! \prod_j v_j!}{n! \prod_i \prod_j a_{ij}!} \tag{3.30}$$

which implies that

$$\sum_A \frac{1}{\prod_i \prod_j a_{ij}!} = \frac{n!}{\prod_i u_i! \prod_j v_j!} . \tag{3.31}$$

It is now shown that the particular forms of eqns. (3.30) and (3.31) allow exact determination of $\mathcal{E}_A(a_{ij}^{(r)})$ where the factorial polynomial $a_{ij}^{(r)}$ is defined as

$$a_{ij}^{(r)} = a_{ij}(a_{ij}-1)\ldots(a_{ij}-r+1) \; ; \qquad r \geq 0 . \tag{3.32}$$

For some fixed value of (i,j) let $\{A'\}$ be the subset of $\{A\}$ such that $a_{ij} > r-1$ for each $A \in \{A'\}$ and $a_{ij} \leq r-1$ for each $A \in \{A\} - \{A'\}$. Define

$$(a'_{ij}, u'_i, v'_j, n') = (a_{ij}-r, u_i-r, v_j-r, n-r)$$

and consider the set $\{A'\}$ with $(a'_{ij}, u'_i, v'_j, n')$ replacing (a_{ij}, u_i, v_j, n). Repeating the argument which led to eqn. (3.31) yields

$$\sum_{A'} \frac{1}{(\prod_{gh \neq ij} a_{gh}!)\, a'_{ij}!} = \frac{n'!}{(\prod_{g \neq i} u_g!)(\prod_{h \neq j} v_h!)\, u'_i!\, v'_j!} . \tag{3.33}$$

From eqns. (3.30) and (3.33), it then follows that

$$\begin{aligned} \mathcal{E}_A(a_{ij}^{(r)}) &= \frac{\prod_g u_g! \prod_h v_h!}{n!} \sum_A \frac{a_{ij}^{(r)}}{\prod_g \prod_h a_{gh}!} \\ &= \frac{\prod_g u_g! \prod_h v_h!}{n!} \sum_{A'} \frac{1}{(\prod_{gh \neq ij} a_{gh}!)\, a'_{ij}!} \\ &= \frac{u_i^{(r)} v_j^{(r)}}{n^{(r)}} . \end{aligned} \tag{3.34}$$

Consequently,

$$\sum_i \sum_j \mathcal{E}_A\big(\mathcal{V}ar(S_{a_{ij}})\big) = \sum_i \sum_j \mathcal{E}_A\Big(\frac{1}{9}a_{ij}^{(3)} + \frac{1}{2}a_{ij}^{(2)}\Big)$$

$$= \sum_i \sum_j \Big(\frac{u_i^{(3)}v_j^{(3)}}{9n^{(3)}} + \frac{u_i^{(2)}v_j^{(2)}}{2n^{(2)}}\Big) \tag{3.35}$$

so that

$$\mathcal{V}ar(S_{u,v}) = \mathcal{V}ar(S_n) - \sum_{i=1}^{k} \mathcal{V}ar(S_{u_i}) - \sum_{j=1}^{\ell} \mathcal{V}ar(S_{v_j}) +$$

$$\frac{1}{9n(n-1)(n-2)} \sum_{i=1}^{k} u_i(u_i-1)(u_i-2) \sum_{j=1}^{\ell} v_j(v_j-1)(v_j-2) +$$

$$\frac{1}{2n(n-1)} \sum_{i=1}^{k} u_i(u_i-1) \sum_{j=1}^{\ell} v_j(v_j-1) \,. \tag{3.36}$$

3.3 The third and fourth cumulants of $S_{u,v}$

From eqn. (3.3),

$$\mathcal{E}(S_n^3|A) = \mathcal{E}(S_{u,v}^3|A) + 3\mathcal{E}(S_{u,v}|A)\Big\{\mathcal{E}\Big(\sum_j S_{v_j}\Big)^2 + \mathcal{E}\Big(\sum_i S_{u_i}\Big)^2\Big\} \tag{3.37a}$$

and

$$\mathcal{E}(S_n^4|A) = \mathcal{E}(S_{u,v}^4|A) + \mathcal{E}\Big(\sum_j S_{v_j}\Big)^4 + \mathcal{E}\Big(\sum_i S_{u_i}\Big)^4 + 6\mathcal{E}\Big(\sum_j S_{v_j}\Big)^2 \mathcal{E}\Big(\sum_i S_{u_i}\Big)^2$$

$$+ 6\mathcal{E}(S_{u,v}^2|A)\Big\{\mathcal{E}\Big(\sum_j S_{v_j}\Big)^2 + \mathcal{E}\Big(\sum_i S_{u_i}\Big)^2\Big\} \,, \tag{3.37b}$$

omitting terms whose expectations yield zero. Since $\mathcal{E}_A\mathcal{E}(S_{u,v}|A) = 0$, it then follows from eqns. (3.4) and (3.28) that

$$\mathcal{E}(S_{u,v}^3) = 3\mathcal{E}_A\Big(S_{u,v}|A \sum_i \sum_j \mathcal{E}(S_{a_{ij}}^2)\Big) \tag{3.38a}$$

where $S_{u,v}|A$, the sum of all second order determinants in the matrix A, may be expressed as $S_{u,v}|A = \sum_{i=1}^{k-1}\sum_{j=1}^{\ell-1}\sum_{g>i}\sum_{h>j}(a_{ij}a_{gh} - a_{gj}a_{ih})$. Also, it follows that

$$\mathcal{E}(S_{u,v}^4) = \mathcal{E}(S_n^4) - \mathcal{E}\Big(\sum_i S_{u_i}\Big)^4 - \mathcal{E}_A\Big[\mathcal{E}\Big(\sum_j S_{s_j}\Big)^4 + 6\sum_j \mathcal{E}(S_{s_j}^2)\sum_i \mathcal{E}(S_{u_i}^2)$$
$$+ 6\mathcal{E}(S_{u,v}^2|A)\Big\{\sum_j \mathcal{E}(S_{s_j}^2) + \sum_i \mathcal{E}(S_{u_i}^2)\Big\}\Big] . \tag{3.38b}$$

Similarly,

$$\mathcal{E}(S_{u,v}^2) = \mathcal{E}(S_n^2) - \sum_i \mathcal{E}(S_{u_i}^2) - \mathcal{E}_A \sum_j \mathcal{E}(S_{s_j}^2) . \tag{3.38c}$$

Let b_{wz} be the polynomial in a_{wz} given by $\mathcal{E}(S_{a_{wz}}^2)$. Modifying the argument which led to eqn. (3.34) shows that

$$\mathcal{E}_A(a_{ij}^{(r)}a_{ih}^{(s)}) = \frac{u_i^{(r+s)}v_j^{(r)}v_h^{(s)}}{n^{(r+s)}} \tag{3.39}$$

whence it is easily seen that $\mathcal{E}_A(a_{ij}a_{gh} - a_{gj}a_{ih}) = 0$. It follows, in an analogous manner, that $\mathcal{E}_A((a_{ij}a_{gh} - a_{gj}a_{ih})b_{wz}) = 0$ for $wz \neq ij, gh, gj$ or ih. Consequently, eqn. (3.38a) yields

$$\mathcal{E}(S_{u,v}^3) = 3\mathcal{E}_A \sum_{i=1}^{k-1}\sum_{j=1}^{\ell-1}\sum_{g>i}\sum_{h>j}((a_{ij}a_{gh} - a_{gj}a_{ih})(b_{ij} + b_{gh} + b_{gj} + b_{ih})) . \tag{3.40}$$

Now $b_{ij} = (2a_{ij}^{(3)} + 9a_{ij}^{(2)})/18$ and $a_{ij}b_{ij} = (2a_{ij}^{(4)} + 15a_{ij}^{(3)} + 18a_{ij}^{(2)})/18$ so that

$$\mathcal{E}_A(b_{ij}a_{ij}a_{gh}) = \frac{1}{18}\Big(\frac{2u_i^{(4)}v_j^{(4)}u_gv_h}{n^{(5)}} + \frac{15u_i^{(3)}v_j^{(3)}u_gv_h}{n^{(4)}} + \frac{18u_i^{(2)}v_j^{(2)}u_gv_h}{n^{(3)}}\Big) \tag{3.41a}$$

and

$$\mathcal{E}_A(b_{ij}a_{gj}a_{ih}) = \frac{1}{18}\Big(\frac{2u_i^{(4)}v_j^{(4)}u_gv_h}{n^{(5)}} + \frac{9u_i^{(3)}v_j^{(3)}u_gv_h}{n^{(4)}}\Big) \tag{3.41b}$$

from which

$$\mathcal{E}_A(b_{ij}a_{ij}a_{gh} - b_{ij}a_{gj}a_{ih}) = u_gv_h\Big(\frac{u_i^{(3)}v_j^{(3)}}{3n^{(4)}} + \frac{u_i^{(2)}v_j^{(2)}}{n^{(3)}}\Big) . \tag{3.42}$$

Therefore, since $\mathcal{E}(S_{u,v}) = 0$,

$$\kappa_3 = \sum_{i=1}^{k-1}\sum_{j=1}^{\ell-1}\sum_{g>i}\sum_{h>j}\Big[\frac{1}{n^{(4)}}\Big\{u_i^{(3)}u_g(v_j^{(3)}v_h - v_jv_h^{(3)}) + u_iu_g^{(3)}(v_jv_h^{(3)} - v_j^{(3)}v_h)\Big\}$$
$$+ \frac{3}{n^{(3)}}\Big\{u_i^{(2)}u_g(v_j^{(2)}v_h - v_jv_h^{(2)}) + u_iu_g^{(2)}(v_jv_h^{(2)} - v_j^{(2)}v_h)\Big\}\Big] . \tag{3.43}$$

Note that Stirling numbers are used to convert polynomials in a_{ij} to polynomials in $a_{ij}^{(r)}$ and vice versa.

Substituting from eqns. (3.38b) and (3.38c) into the relationship $\kappa_4 = \mathcal{E}(S_{u,v}^4) - 3\big(\mathcal{E}(S_{u,v}^2)\big)^2$ yields

$$\kappa_4 = \kappa_4(n) - \sum_{i=1}^{k} \kappa_4(u_i) - \mathcal{E}_A \mathcal{E}\Big(\sum_j S_{a_j}\Big)^4 - 6\mathcal{E}_A\Big(S_{u,v}^2 | A \sum_j \mathcal{E}(S_{a_j}^2)\Big)$$
$$- 3\Big(\mathcal{E}_A \sum_j \mathcal{E}(S_{a_j}^2)\Big)^2 + 6\Big(\mathcal{E}(S_n^2) - \sum_i \mathcal{E}(S_{u_i}^2)\Big)\Big(\mathcal{E}_A \sum_j \mathcal{E}(S_{a_j}^2)\Big) \quad (3.44),$$

where the fact that $K\big(\sum_i x_i\big) = \sum_i K(x_i)$, for x_i independent of x_j and $K(x_i)$ the cgf of x_i, has been used. Eqn. (3.4), taken in conjunction with the mutual independence of S_{v_j} and S_{v_i}, yields

$$\mathcal{E}_A \sum_j \mathcal{E}(S_{a_j}^2) = \sum_j \mathcal{E}(S_{v_j}^2) - \mathcal{E}_A\Big(\sum_i \sum_j \mathcal{E}(S_{a_{ij}}^2)\Big) \qquad (3.45a)$$

and

$$\mathcal{E}_A \mathcal{E}\Big(\sum_j S_{a_j}\Big)^4 = \mathcal{E}\Big(\sum_j S_{v_j}\Big)^4 - \mathcal{E}_A \mathcal{E}\Big(\sum_i \sum_j S_{a_{ij}}\Big)^4$$
$$- 6\mathcal{E}_A\Big(\sum_j \mathcal{E}(S_{a_j}^2) \sum_i \sum_j \mathcal{E}(S_{a_{ij}}^2)\Big) \ . \quad (3.45b)$$

Substituting from eqns. (3.45a) and (3.45b) into eqn. (3.44) and simplifying then gives

$$\kappa_4 = \kappa_4(n) - \sum_{i=1}^{k} \kappa_4(u_i) - \sum_{j=1}^{\ell} \kappa_4(v_j) + \sum_{i=1}^{k} \sum_{j=1}^{\ell} \mathcal{E}_A\big(\kappa_4(a_{ij})\big) -$$
$$3\,\mathcal{V}ar_A\Big(\sum_i \sum_j \mathcal{E}(S_{a_{ij}}^2)\Big) + 6\,\mathcal{C}ov_A\Big(S_{u,v}^2 | A, \sum_i \sum_j \mathcal{E}(S_{a_{ij}}^2)\Big) \ . \qquad (3.46)$$

To date, the last term of eqn. (3.46) has not yielded to simplification. However, the result eqn. (3.12), suggests that the first three terms of eqn. (3.46) will give a reasonable approximation to κ_4 for moderately large values of n.

3.4 Edgeworth approximation to the distribution of $S_{u,v}$

Bickel and Doksum (1977, section 1.5D) have observed that the normal approximation to the distribution of a random variable utilizes only the first two moments of the random variable. They noted that it is sometimes possible to improve on the normal approximation by also utilizing the third and fourth moments. The preceding results, when combined with the algorithm of Section 2.3, facilitate investigation into the usefulness of an Edgeworth approximation for determining the significance levels of $S_{u,v}$.

David et al. (1951) and Silverstone (1950) have shown that, in the absence of ties, an Edgeworth expansion of the distribution of S_n results in substantially more accurate significance levels than those obtained from the normal approximation. Their results have been used by Best and Gipps (1974) to develop an algorithm which yields one-sided significance levels for S_n with a maximum error of 0.0004. Robillard (1972) has demonstrated a similar result for the case where one ranking is tied. His results, when taken in conjunction with the algorithm of Section 2.2, have been used to develop an algorithm which will give one-sided significance levels for S_u with a maximum error of 0.0004 provided that the maximum extent of a single tie is not greater than $0.8n$; the algorithm is listed in the Appendix to Chapters 2 and 3 as the function PRKST1.

The distribution of $S_{u,v}$ possesses two features which serve to inhibit substantial improvement over the accuracy of significance levels obtained from the normal approximation: Firstly, the spacing between adjacent scores is not constant, the irregularity being pronounced in the tails of the distribution. However, as will be seen below, the adjacent scores differ by one over most of the distribution provided that the ties are not too extensive. For the special case wherein one ranking is a dichotomy, which occurs for $k = 2$, Burr (1960) has recommended that one-half of the highest common factor of the numbers $v_1 + v_2$, $v_2 + v_3$, ..., $v_{\ell-1} + v_\ell$ be used as the correction for continuity. It generally follows that as soon as $v_j = 1$, for some j, then the recommended correction for continuity is one-half. Secondly, the distributions display serrated profiles which clearly limit the ability of a smooth curve to accurately approximate the true distribution. This factor is exacerbated as the extent of the ties increases.

Table 3.1: Cumulative probability distribution of $S_{u,v}$
Example 8.2 of Burr (1960)

Score	Cumulative Probability			Differences	
	Exact	Normal	Edgewo	N.diff	E.diff
-29	0.0021	0.0024	0.0010	-0.0002	0.0011
-24	0.0136	0.0099	0.0080	0.0036	0.0056
-20	0.0164	0.0267	0.0255	-0.0102	-0.0091
-19	0.0457	0.0334	0.0328	0.0123	0.0130
-18	0.0486	0.0415	0.0415	0.0071	0.0071
-15	0.0829	0.0754	0.0779	0.0074	0.0049
-14	0.0981	0.0905	0.0940	0.0076	0.0041
13	0.9019	0.9095	0.9060	-0.0076	-0.0041
14	0.9171	0.9246	0.9221	-0.0074	-0.0049
15	0.9514	0.9377	0.9361	0.0137	0.0153
18	0.9543	0.9666	0.9672	-0.0123	-0.0130
19	0.9836	0.9733	0.9745	0.0102	0.0091
20	0.9864	0.9789	0.9804	0.0075	0.0060
24	0.9979	0.9924	0.9943	0.0055	0.0035
29	1.0000	0.9983	0.9995	0.0017	0.0005
The approx [1] standardized cumulants are:				0.0000	-0.3160
The exact [2] standardized cumulants are:				0.0000	-0.2922
Time Taken:					0.43 seconds

[1] based on eqns. (3.43) and (3.46)
[2] based on the frequency distribution of $S_{u,v}$

For the case of a dichotomy in one ranking, Klotz (1966) found that the Edgeworth approximation offered little improvement over the normal approximation. As Tables 3.1 and 3.2, which compare the normal and Edgeworth approximations for two selected examples taken from Burr (1960) and Kendall (1975) indicate, the latter approximation appears to result in appreciable improvement over the former whenever the ties are not too extensive. However, the improvement is not uniform over either of the distributions; this being caused by the serrated profile of the distributions. Most adjacent scores differ by 1 and therefore the continuity correction

Table 3.2: Cumulative probability distribution of $S_{u,v}$
Example 3.1 of Kendall (1975)

Score	Cumulative Probability			Differences	
	Exact	Normal	Edgewo	N.diff	E.diff
-30	0.0011	0.0024	0.0012	-0.0013	-0.0001
-29	0.0017	0.0032	0.0019	-0.0015	-0.0001
-28	0.0024	0.0043	0.0028	-0.0019	-0.0004
-27	0.0034	0.0057	0.0040	-0.0022	-0.0006
-26	0.0060	0.0074	0.0057	-0.0014	0.0002
-25	0.0070	0.0096	0.0079	-0.0026	-0.0009
-24	0.0103	0.0123	0.0107	-0.0020	-0.0003
-23	0.0137	0.0158	0.0142	-0.0020	-0.0004
-22	0.0187	0.0199	0.0185	-0.0013	0.0001
-21	0.0224	0.0250	0.0239	-0.0027	-0.0015
-20	0.0319	0.0312	0.0305	0.0007	0.0014
-19	0.0363	0.0385	0.0383	-0.0022	-0.0019
-18	0.0490	0.0472	0.0476	0.0018	0.0015
-17	0.0578	0.0574	0.0585	0.0004	-0.0007
-16	0.0716	0.0692	0.0711	0.0023	0.0005
-15	0.0863	0.0829	0.0855	0.0035	0.0008
-14	0.1016	0.0985	0.1019	0.0031	-0.0003
13	0.8969	0.9015	0.8981	-0.0046	-0.0012
14	0.9150	0.9171	0.9145	-0.0021	0.0005
15	0.9283	0.9308	0.9289	-0.0025	-0.0007
16	0.9422	0.9426	0.9415	-0.0004	0.0007
17	0.9513	0.9528	0.9524	-0.0014	-0.0011
18	0.9625	0.9615	0.9617	0.0010	0.0008
19	0.9700	0.9688	0.9695	0.0012	0.0005
20	0.9757	0.9750	0.9761	0.0007	-0.0004
21	0.9824	0.9801	0.9815	0.0023	0.0009
22	0.9859	0.9842	0.9858	0.0016	0.0000
23	0.9894	0.9877	0.9893	0.0017	0.0000
24	0.9933	0.9904	0.9921	0.0029	0.0012
25	0.9937	0.9926	0.9943	0.0011	-0.0006
26	0.9968	0.9943	0.9959	0.0025	0.0009
27	0.9975	0.9957	0.9972	0.0017	0.0003
28	0.9981	0.9968	0.9981	0.0013	0.0000
29	0.9990	0.9976	0.9988	0.0015	0.0002
30	0.9994	0.9982	0.9993	0.0011	0.0001
31	0.9997	0.9987	0.9996	0.0010	0.0001
33	1.0000	0.9993	1.0000	0.0007	0.0000
The approx [1] standardized cumulants are:				8.7313E-05	-0.2762
The exact [2] standardized cumulants are:				8.7313E-05	-0.2716
Time Taken:				10.15 seconds	

[1] based on eqns. (3.43) and (3.46)

[2] based on the frequency distribution of $S_{u,v}$

of 1 used by Kendall (1975, example 4.2) is not to be recommended. The true significance level for this example is 0.015 while the normal and Edgeworth approximations, with a continuity correction of 1/2, yield 0.019 and 0.016, respectively.

Kendall obtained a significance level of 0.021 by using the normal approximation with a continuity correction of 1. Enumerating the distribution for this particular example, which has $n = 12$ and $k = \ell = 8$, or $N_x = N_y = 8$, took 3 hours of machine time.

The significance levels shown for the Edgeworth approximation were computed using eqns. (3.43) and (3.46). As the tables clearly indicate, the error introduced by the approximation to the fourth cumulant is negligible whenever the ties are not too extensive and n is moderately large. Furthermore, the errors of approximation obtained in both tables suggest that errors associated with significance levels obtained by the Edgeworth approximation will be negligible for $n > 20$ providing that the ties are not too extensive. Thus the Edgeworth approximation is recommended for practical problems with large n. In the notation of this chapter, $u = (3,4,3)$ and $v = (2,3,3,2)$ for the example used in Table 3.1, and $u = (1,2,2,2,1,2)$ and $v = (1,1,4,3,1)$ for the example used in Table 3.2.

3.5 Approximations to the distribution of D

The irregularity in the spacing of $S_{u,v}$ and the serrated nature of its distribution are much more pronounced for Spearman's D even when only one ranking contains ties. Consequently, an Edgeworth approximation to the null distribution of D, in the presence of ties, was not considered. Kendall, Kendall and Babington Smith (1938) have noted that the distribution of D is serrated even in the absence of ties. For this case, they demonstrated that a transformation suggested by Pitman (1937) improves over the normal approximation. Glasser and Winter (1961) have shown that an Edgeworth approximation due to David et al. (1951) is consistently more accurate than the approximation of Pitman. Franklin (1987) has shown that a Pearson type II curve recommended by Olds (1938), and utilized by Zar (1972) to create approximate critical values for D for $n = 4(1)100$, is clearly superior to

Pitman's approximation. Direct comparison of the Pearson type II curve and the Edgeworth approximation has, as yet, not been implemented. The function PRSPD listed in the Appendix to Chapters 2 and 3 computes probabilities for D, in the absence of ties, using an Edgeworth approximation as discussed by David et al. For $n = (9, 10, 11)$ the maximum absolute errors times 10^4 are $(10.60, 5.92, 3.42)$ while Franklin obtained $(14.44, 9.69, 6.94)$ with the Pearson type II curve. Consequently, the Edgeworth approximation is recommended as being superior to the Pearson type II curve.

REFERENCES

Best, D.J. and Gipps, P.G. (1974). The upper tail probabilities of Kendall's tau. *Applied Statistics*, **23**, 98–100.

Bickel, P.J. and Doksum, K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden–Day.

Billingsley, P. (1986). *Probability and Measure* (2nd ed). John Wiley and Sons.

Burr, E.J. (1960). The distribution of Kendall's score S for a pair of tied rankings. *Biometrika*, **47**, 151–171.

David, S.T., Kendall, M.G. and Stuart, A. (1951). Some questions of distribution in the theory of rank correlation. *Biometrika*, **38**, 131–140.

Franklin, L.A. (1987). Approximations, convergence and exact tables for Spearman's rank correlation coefficient. *Proceedings of the Statistical Computing Section of the American Statistical Association*, 1987, 244–247.

Glasser, G.J. and Winter, R.F. (1961). Critical values of the coefficient of rank correlation for testing the hypothesis of independence. *Biometrika*, **48**, 444–448.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, **19**, 293–325.

Kendall, M.G. (1975). *Rank Correlation Methods* (4th ed). Griffin and Co. Ltd.

Kendall, M.G., Kendall S.F.H. and Babington Smith, B. (1938). The distribution of Spearman's rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika*, **30**, 251–273.

Klotz, J.H. (1966). The Wilcoxon, ties, and the computer. *Journal of the American Statistical Association*, **61**, 772–787.

Noether, G.E. (1967). *Elements of Nonparametric Statistics.* New York: Wiley.

Olds, E.G. (1938). Distributions of sums of squares of rank differences for small numbers of individuals. *Annals of Mathematical Statistics*, **9**, 133–148.

Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any populations. II. The correlation coefficient test. *Journal of the Royal Statistical Society, Suppl.*, **4**, 225–232.

Robillard, P. (1972). Kendall's S distribution with ties in one ranking. *Journal of the American Statistical Association*, **67**, 453–455.

Reingold, E.M., Nievergelt, J. and Deo, N. (1977). *Combinatorial Algorithms: Theory and Practice.* Prentice-Hall. New Jersey.

Silverstone, H. (1950). A note on the cumulants of Kendall's S-distribution. *Biometrika*, **37**, 231–235.

Zar, J.H. (1972). Significance testing of the Spearman rank correlation coefficient. *Journal of the American Statistical Association*, **67**, 578–580.

Appendix to Chapters 2 and 3

THE UPPER TAIL PROBABILITIES OF KENDALL'S TAU AND OF SPEARMAN'S RHO

Keywords: Kendall's tau; Spearman's rho; Edgeworth series approximation

Language

ANSI Standard Fortran 1977

Description and Purpose

The Function PRRANK computes the upper tail probability for either Kendall's score S or Spearman's score D for n pairs of observations on two random variables. The statistics S and D are widely used as measures of trend and as measures of the monotonic relationship between two variables.

Numerical Method

First consider Kendall's S. If there are no ties then the algorithm AS 71 developed by Best and Gipps (1974) is used. For ties in only one ranking then the algorithm of Section 2.2 ($n \leq 20$) and an Edgeworth approximation ($n > 20$), with the cumulants given by Robillard (1972), are used in the Function PRKST1. If there are ties in both rankings then the algorithm of Section 2.3 and an Edgeworth approximation, as developed in Chapter 3, are used in the Function PRKST2. Now consider Spearman's D. If there are no ties in either ranking, and $n \leq 10$, an exact enumeration is used in the Function PRSPD. For $n > 10$ an Edgeworth approximation, as developed by David et al (1951), is used in PRSPD. For ties in at least one ranking, the algorithm of Section 2.3 and the Normal approximation are used in the Function PRSPDT. If $n > 30$, both PRSPDT and PRKST2 use an approximation. Otherwise, they both allow the user, after inspecting the ties, to determine whether an approximation or exact enumeration is to be used. If exact enumeration is chosen, the user is still allowed, based on the estimated total number of enumerations required, to subsequently switch to an approximation.

Structure

FUNCTION PRRANK (X,Y,N,IW,NIW,W,NW,LGUI,LGUO,ITYPE,IFAULT)

Formal Parameters

X	Real array(*N*)	input:	a data vector
Y	Real array(*N*)	input:	a data vector
N	Integer	input:	the size of the data vectors
IW	Integer array (*NIW*)	input:	work vector
NIW	Integer	input:	dimension of *IW*. Set to $30000 + 6n$.
W	Real*8 array (*NW*)	input:	work vector
NW	Integer	input:	dimension of *W*. Set to $30000 + 6n$.
LGUI	Integer	input:	Fortran code for terminal input
LGUO	Integer	input:	Fortran code for terminal output
ITYPE	Integer	input:	respectively, 1 or 2, for S or D
IFAULT	Integer	output:	a failure indicator, equal to: 1 if $n \leq 1$, 2 if either $Y(i)$ or $X(i)$ is constant for all i, 3 if the score passed to any Function is inconsistent with n, 4 if the error tolerance in cumulative sum of probabilities is exceeded, 5 if a dimension of the array *ICOMP* is inadequate

Auxiliary Algorithms

The following auxiliary subroutines and functions are called:

FUNCTION ALNORM (X, UPPER)—Algorithm AS 66 (Hill, 1973).

FUNCTION PRTAUS (IS, N, IFAULT)—Algorithm AS 71 (Best and Gipps, 1974)

SUBROUTINE INDEXX (N, ARRIN, INDX)—(Press, Fla·...ery, Teukolsky and Vetterling, 1986)

Accuracy

The maximum error is less than 0.0004 when ties are absent. If $n.le.20$, or the largest tie is less than $0.8n$ for $n > 20$, and there are ties in only one ranking, the maximum error is less than 0.0004 for Kendall's score. For ties in both of the rankings, or ties in one ranking for Spearman's score, the errors are comparatively larger and there is no way of specifying the maximum error. However, for moderately small n the error is 0.0. For values of n which are too large to facilitate exact enumeration, the comparison, shown in Tables 3.1 and 3.2, of the probabilities obtained by exact enumeration with those obtained from the Edgeworth series approximation suggests that the maximum error, in the outer 10% of the tails of the distribution, will be of order 10^{-3} for Kendall's score.

References

Best, D.J. and Gipps, P.G. (1974). The upper tail probabilities of Kendall's tau. *Applied Statistics*, **23**, 98–100.

David, S.T., Kendall, M.G. and Stuart, A. (1951). Some questions of distributions in the theory of rank correlation. *Biometrika*, **38**, 131–140.

Hill, I.D. (1973). Algorithm AS 66. The normal integral. *Applied Statis...s*, **22**, 424–427.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986). *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press.

Robillard, P. (1972). Kendall's S distribution with ties in one ranking. *Journal of the American Statistical Association*, **67**, 453–455.

```
      FUNCTION PRRANK(X,Y,N,IW,NIW,W,NW,LGUI,LGUO,ITYPE,IFAULT)

C     GIVEN N PAIRS OF OBSERVATIONS ON X AND Y, THE FUNCTION COMPUTES
C     THE PROBABILITY OF OBTAINING A VALUE GREATER THAN, OR EQUAL TO,
C     EITHER THE ASSOCIATED KENDALL'S SCORE, IS, OR THE ASSOCIATED
C     SPEARMAN'S SCORE, D, UNDER THE NULL HYPOTHESIS OF INDEPENDENCE.

C     .. ARRAY ARGUMENTS ..
      DIMENSION   X(N), Y(N), IW(NIW)
      DOUBLE PRECISION     W(NW)
C     .. LOCAL SCALARS ..
      LOGICAL     SWX, SWY
      CHARACTER*1 RESP

C     ETA - ERROR TOLERANCE IN CUMULATIVE SUM OF ALL PROBABILITIES.
      PARAMETER (ETA=1.0E-8)

C     ENSURE THAT THE NUMBER OF OBSERVATIONS EXCEEDS 1 AND THAT NEITHER
C     X NOR Y IS SINGULAR (I.E  A SINGLE VALUE REPEATED N TIMES).

      IFAULT = 1
      IF (N .LE. 1) RETURN
      SWX = .TRUE.
      SWY = .TRUE.
      DO 10 I = 2,N
         IF (X(I) .NE. X(I-1)) SWX = .FALSE.
         IF (Y(I) .NE. Y(I-1)) SWY = .FALSE.
   10 CONTINUE
      IFAULT = 2
      IF (SWX .OR. SWY) RETURN
      N1 = N + 1
      N2 = N1 + N
      N3 = N2 + N
      N4 = N3 + N

C     OBTAIN INDEX AND RANK TABLES FOR THE GIVEN DATA.

      CALL INDEXX(N, X, IW(1))
      CALL RANK(X, IW(1), W(1), N, SUMX)
      CALL INDEXX(N, Y, IW(N1))
      CALL RANK(Y, IW(N1), W(N1), N, SUMY)

C     COMPUTE THE APPROPRIATE SCORE FOR THE GIVEN DATA.

      GO TO (20,30), ITYPE
   20 IS = ISCORE(W(1), W(N1), N, ITYPE)
      WRITE (7,99999) IS
99999 FORMAT(//, 1X, 'KENDALL''S SCORE IS:', I6)
      GO TO 40
   30 ID = ISCORE(W(1), W(N1), N, ITYPE)
      D = ID/4.0
      WRITE (7,99998) D
99998 FORMAT(//, 1X, 'SPEARMAN''S SCORE IS:', 1PE12.6)


C     SUMX AND SUMY ARE USED TO DETERMINE THE APPROPRIATE ANALYSIS.

   40 IF (SUMX .EQ. 0.0 .AND. SUMY .EQ. 0.0) THEN
         GO TO (50,60), ITYPE
      ELSE IF (SUMX .EQ. 0.0 .OR. SUMY .EQ. 0.0) THEN
         GO TO (70,80), ITYPE
      ELSE
         GO TO 80
      END IF

C     NO TIES IN DATA. KENDALL'S SCORE.

   50 PRRANK = PRTAUS(IS, N, IFAULT)
      IF (IFAULT .NE. 0) IFAULT = 3
      RETURN
```

```
C     NO TIES IN DATA. SPEARMAN'S SCORE.

   60 NIW = NIW - 1
      ID = ID/4
      PRRANK = PRSPD(ID, IW, NIW, N, ETA, IFAULT)
      RETURN

C     TIES IN ONE RANKING. KENDALL'S SCORE.

   70 NW = 190
      NW2 = NW + 2
      NIW = NIW - 4*N
      IF (SUMX .EQ. 0.0) THEN
         PRRANK = PRKST1(IS, Y, IW(N1), IW(N2), IW(N3), N, IW(N4),
     *                       NIW, W(1), W(NW2), NW, ETA, IFAULT)
         RETURN
      ELSE
         PRRANK = PRKST1(IS, X, IW(1), IW(N2), IW(N3), N, IW(N4),
     *                      NIW, W(1), W(NW2), NW, ETA, IFAULT)
         RETURN
      END IF

C     DETERMINE THE TUPLETS FOR X AND Y VECTORS.

   80 N5 = N4 + N
      N6 = N5 + N
      CALL SCAN (X, IW(1), IW(N2), IW(N4), N, MXX, NX)
      CALL SCAN (Y, IW(N1), IW(N3), IW(N5), N, MXY, NY)
      NIW = NIW - 6*N - 1
      NW = NW - 1
      IAL = 2
      IF (N .GT. 30) GO TO 110

C     IF N .LE. 30 THEN THE USER DECIDES WHETHER OR NOT TO
C     ATTEMPT AN EXACT ENUMERATION.

      WRITE (LGUO,*) 'THE X TUPLETS ARE:'
      NMXX = N2 + MXX - 1
      DO 90 I = N2,NMXX
         IF (IW(I) .EQ. 0) GO TO 90
         I1 = I - N2 + 1
         WRITE (LGUO,*) (I1, J=1,IW(I))
   90 CONTINUE
      WRITE (LGUO,*) 'THE Y TUPLETS ARE:'
      NMXY = N3 + MXY - 1
      DO 100 I = N3,NMXY
         IF (IW(I) .EQ. 0) GO TO 100
         I1 = I - N3 + 1
         WRITE (LGUO,*) (I1, J=1,IW(I))
  100 CONTINUE
      PRINT *, 'DO YOU WISH TO COMPUTE THE EXACT PROBABILITY? (Y/N)'
      READ (LGUI,99999) RESP
      99999 FORMAT (A1)
      IAL = 1
      IF (RESP .EQ. 'N' .OR. RESP .EQ. 'n') IAL = 2
  110 GO TO (120,130), ITYPE

C     TIES IN BOTH RANKINGS. KENDALL'S SCORE.

  120 PRRANK = PRKST2(IS, IW(N2), IW(N3), IW(N4), IW(N5), MXX, MXY, NX,
     *                   NY, IW(N6), NIW, W, NW, N, IAL, ETA, IFAULT)
      RETURN

C     TIES IN ONE OR BOTH OF THE RANKINGS. SPEARMAN'S SCORE.

  130 ID = IFIX(4*D)
      PRRANK = PRSPDT(ID, IW(N2), IW(N3), IW(N4), IW(N5), MXX, MXY, NX,
     *         NY, IW(N6), NIW, W, NW, N, IAL, SUMX, SUMY, ETA, IFAULT)
      RETURN
      END
```

```
      SUBROUTINE RANK(Z, IND, R, N, SUMT)

C     OBTAINS RANK FROM ORDER PERMUTATION GIVEN BY THE SUBROUTINE
C     INDEXX. TIED VALUES ARE AVERAGED. ALSO RETURNS THE SUM OVER
C     T OF T*(T**2-1) WHERE T IS THE EXTENT OF A TIE.

C     INPUT PARAMETERS
C        Z:      DATA VECTOR
C        IND:    THE INDEX TABLE FOR Z
C        R:      THE RANK TABLE FOR Z

      DIMENSION  Z(N), IND(N), R(N)
      LOGICAL    SW
      SUMT = 0.0
      I = 1
      R(IND(1)) = 1.0
      SW = .TRUE.
   10 I = I+1
      IF (I .GT. N) GOTO 40
      R(IND(I)) = FLOAT(I)
      IF (SW)  K = 1
      IF (SW)  AV = R(IND(I-1))
      IF (Z(IND(I)) .NE. Z(IND(I-1))) GOTO 30
      K = K+1
      AV = AV + (R(IND(I))-AV)/FLOAT(K)
      DO 20 J = 1, K
   20 R(IND(I+1-J)) = AV
      SW = .FALSE.
      GOTO 10
   30 SW = .TRUE.
      SUMT = SUMT + FLOAT(K*(K*K-1))
      GOTO 10
   40 IF (.NOT. SW) SUMT = SUMT + FLOAT(K*(K*K-1))
      RETURN
      END

      SUBROUTINE SCAN(Z, INDX, IPZ, IHZ, N, MXZ, NZ)

C     DETERMINES THE NUMBER OF EACH R-TUPLET OF TIED RANKS AND THE
C     MAXIMUM VALUE OF R. ALSO DETERMINES THE ORDER OF TUPLETS FOR
C     BOTH SETS OF RANKS WHERE THE RANKS ARE ARRANGED IN ASCENDING
C     ORDER.

C     INPUT PARAMETERS
C        Z:        DATA VECTOR WITH TIES
C        INDX:     INDEX TABLE FOR Z

C     OUTPUT PARAMETERS
C        IPZ:   VECTOR CONTAINING NUMBER OF TIMES EACH TUPLET OCCURS
C        MXZ:   THE LARGEST TUPLET FOR THE GIVEN DATA VECTOR
C        IHZ:   VECTOR GIVING THE ORDER OF OCCURRENCE OF THE TUPLETS
C               IN Z (I.E. IHZ(J) = R IF THE JTH SPECIES IS AN R-TUPLET)
C        NZ:    THE NUMBER OF DIFFFERENT SPECIES IN Z

C     .. ARRAY ARGUMENTS ..
      DIMENSION Z(N), INDX(N), IPZ(N), IHZ(N)

C     DETERMINE IPZ, IHZ, MXZ AND NZ.

      DO 10 J = 1,N
         IPZ(J) = 0
   10 CONTINUE
      J = 2
      INDEX = 1
      MXZ = 1
   20 ICOUNT = 1
   30 IF (Z(INDX(J)) .GT  Z(INDX(J-1))) GO TO 40
      ICOUNT = ICOUNT + 1
      IF (J .EQ. N) GO TO 40
      J = J + 1
      GO TO 30
```

```
   40 IPZ(ICOUNT) = IPZ(ICOUNT) + 1
      IF (ICOUNT .GT. MXZ) MXZ = ICOUNT
      IHZ(INDEX) = ICOUNT
      INDEX = INDEX + 1
      J = J + 1
      IF (J .GT. N) GO TO 50
      GO TO 20
   50 IF (Z(INDX(N)) .GT. Z(INDX(N-1))) IPZ(1) = IPZ(1) + 1
      IF (Z(INDX(N)) .GT. Z(INDX(N-1))) IHZ(INDEX) = 1
      NZ = INDEX - 1
      IF (Z(INDX(N)) .GT. Z(INDX(N-1))) NZ = INDEX
      RETURN
      END

      FUNCTION PRSPD(ID, NSCO, NIW, N, ETA, IER)

C     GIVEN A VALUE OF ID CALCULATED FROM TWO RANKINGS (WITHOUT TIES)
C     OF N OBJECTS, THE FUNCTION COMPUTES THE PROBABILITY OF GETTING
C     A VALUE GREATER THAN OR EQUAL TO ID.

C     INPUT PARAMETERS
C        ETA:   ETA TOLERANCE IN CUMULATIVE SUM OF PROBABILITIES

C     .. ARRAY ARGUMENTS ..
      INTEGER NSCO(0:NIW)
C     .. LOCAL SCALARS ..
      DOUBLE PRECISION RNFACT, PSCO, SUMPR1, SUMPR2
C     .. LOCAL ARRAYS ..
      INTEGER PI(0:12), P(12), E(12)
      REAL HE(11)

C     CHECK ON THE VALIDITY OF ID AND N VALUES.

      PRSPD = 1.0
      IER = 3
      MAXSCO = N*(N**2-1)/3
      IF (ID .GT. MAXSCO .OR. ID .LT. 0) RETURN
      IER = 0
      IF (ID .EQ. 0) RETURN
      IF (N .GT. 10) GO TO 70
      IER = 4

C     A MINIMAL CHANGE ORDER PERMUTATION GENERATOR, REINGOLD ET AL
C     (1977) IS USED TO GENERATE ALL PERMUTATIONS.

      RNFACT = 1.0D0
      DO 10 I=1,N
         PI(I) = I
         P(I) = I
         E(I) = -1
         RNFACT = RNFACT*I
   10 CONTINUE
      E(1) = 0
      PI(0) = N+1
      PI(N+1) = N+1

C     NSCO(I) STORES THE FREQUENCY WITH WHICH S OCCURS.
C     INITIALIZE NSCO TO ZERO WITHIN A KNOWN RANGE FOR S.

      MPSCO = MAXSCO/2
      DO 20 I = 0,MPSCO
         NSCO(I) = 0
   20 CONTINUE
      ISCO = 0

C     START PERMUTATION GENERATOR.

   30 NSCO(ISCO) = NSCO(ISCO) + 1
      M = N
   40 IF (PI(P(M)+E(M)) .LE. M) GO TO 50
      E(M) = -E(M)
```

```
      M = M-1
      GO TO 40
   50 IPITEMP = PI(P(M))
      PI(P(M)) = PI(P(M)+E(M))
      PI(P(M)+E(M)) = IPITEMP
      PE = P(M) + E(M)
      IF (P(M) .LT. PE) THEN
         ISCO = ISCO + PI(P(M)) - PI(PE)
      ELSE
         ISCO = ISCO + PI(PE) - PI(P(M))
      ENDIF
      IPTEMP = P(PI(P(M)))
      P(PI(P(M))) = P(M)
      P(M) = IPTEMP
      IF (M .NE. 1) GO TO 30
      SUMPR1 = 0.0
      SUMPR2 = 0.0
      DO 60 I = 0,MPSCO
         PSCO = NSCO(I)/RNFACT
         I1 = 2*I
         SUMPR1 = SUMPR1 + PSCO
         IF (I1 .GE. ID) SUMPR2 = SUMPR2 + PSCO
   60 CONTINUE
      IF (ABS(SUMPR1-1.0D0) .GT. ETA) RETURN
      PRSPD = SUMPR2
      IER = 0
      RETURN

C     DETERMINE PROBABILITIES USING AN EDGEWORTH SERIES EXPANSION
C     WITH TERMS UP TO ORDER N**-3. USE A SHEPPARD'S CORRECTION
C     FOR THE CUMULANTS.

   70 DEN = FLOAT(N*(N**2-1))
      H = 2*6/DEN
      SK2 = 1.0/(N-1) - H**2/12.0
      SK4 = -6*((19*N+5.)*N-36.)/(25*DEN) + H**4/120.0
      SK6 = 48*(((((583*N+723.)*N-2603)*N-2637)*N+4054)*N+2760)*N
     2      -1800)/(245*SK2*DEN**3) - H**6/252.0
      SK8 = 144*((((((((-41939*N-83709.)*N+304254)*N+578442)*N
     2   - 1012323)*N-1690125)*N+1800776)*N+2358048)*N-1616688)*N
     3   - 1080567)*N+846720)/(875*SK2**2*DEN**5) + H**8/240.0

C     COMPUTE THE STANDARDIZED SCORE CORRECTED FOR CONTINUITY, THE
C     RELEVANT HERMITE POLYNOMIALS AND THE DESIRED PROBABILITY

      X = (1.0 - 6*(ID-1)/DEN)/(SK2**0.5)
      HE(1) = X
      HE(2) = X**2 - 1.0
      DO 80 L = 3,11
         HE(L) = X*HE(L-1) - (L-1)*HE(L-2)
   80 CONTINUE
      HERCUM = SK4*HE(3)/24 + SK6*HE(5)/720 + HE(7)*SK4**2/1152
     2       + SK8*HE(7)/40320 + SK6*SK4*HE(9)/17280
     3       + HE(11)*SK4**3/82944
      CORRFAC = HERCUM*0.39894228*EXP(-0.5*X**2)
      PRSPD = ALNORM(X, .FALSE.) - CORRFAC
      IF (PRSPD .LT. 0.0) PRSPD = 0.0
      IF (PRSPD .GT. 1.0) PRSPD = 1.0
      RETURN
      END

      FUNCTION PRKST1(IS, Z, INDX, IHZ, IP, N, ICOEF, NIW, USCO1,
     *                                       USCO2, NW, ETA, IER)

C     GIVEN A VALUE OF IS CALCULATED FROM TWO RANKINGS OF N OBJECTS,
C     WITH TIES IN ONLY ONE RANKING, THE FUNCTION COMPUTES THE
C     PROBABILITY OF OBTAINING A VALUE GREATER THAN, OR EQUAL TO, IS.

C     INPUT PARAMETERS
C        Z:       DATA VECTOR WITH TIES IN THE DATA
C        INDX:    INDEX TABLE FOR Z
```

```
C        ETA:      ERROR TOLERANCE IN CUMULATIVE SUM OF PROBABILITIES

C     .. ARRAY ARGUMENTS ..
      DIMENSION Z(N), INDX(N), IHZ(N), IP(N), ICOEF(N!W)
      DOUBLE PRECISION USCO1(0:NW), USCO2(0:NW)
C     .. LOCAL SCALARS ..
      DOUBLE PRECISION RNTOT, PSCO, SUMPR1, SUMPR2
C     .. LOCAL ARRAYS ..
      INTEGER MAXSCO(2)
      REAL  HE(15)

C     DETERMINE THE TUPLETS FOR Z.

      CALL SCAN(Z, INDX, IP, IHZ, N, MAXTUP, NKIND)

C     CHECK ON THE VALIDITY OF IS AND IP VALUES.

      IER = 3
      NPRE = 0
      DO 20 I = 1,MAXTUP
         IF (IP(I) .EQ. 0) GO TO 20
         DO 10 J = 1,IP(I)
            MAXSCO(2) = MAXSCO(1) + NPRE*I
            MAXSCO(1) = MAXSCO(2)
            NPRE = NPRE + I
   10    CONTINUE
   20 CONTINUE
      IF (IABS(IS) .GT. MAXSCO(2)) RETURN
      IER = 0
      IF (N .GT. 20) GO TO 110

C     IMPLEMENT THE RECURSION USED TO DETERMINE THE SCORE DISTRIBUTION.
C     INITIALIZE APPROPRIATE PARAMETERS PRIOR TO ENTERING THE LOOP
C     WHICH DETERMINES THE SCORE DISTRIBUTION. THE OUTERMOST TWO LOOPS
C     ARE USED TO ENTER TUPLETS INTO THE CURRENT SET OF RANKS. NOTE
C     THAT THE DISTRIBUTION OF KENDALL'S SCORE WITH TIES IN ONE RANKING
C     IS SYMMETRICAL ABOUT ZERO AND THEREFORE ONLY POSITIVE SCORES NEED
C     TO BE CONSIDERED. ISCO = NEW SCORE. IOSCO = OLD SCORE.

      IER = 4
      NPRE = 0
      USCO1(0) = 1.0D0
      MAXSCO(1) = 0
      DO 80 I = 1,MAXTUP
         I1 = MAXTUP + 1 - I
         IF (IP(I1) .EQ. 0) GO TO 80
         DO 70 J = 1,IP(I1)
            IF (NPRE .EQ. 0) GO TO 60
            NINV = NPRE*I1 + 1
            N1 = NINV + 1
            CALL COEFF(NPRE, ICOEF(1), NINV, ICOEF(N1), I1)
            MAXSCO(2) = MAXSCO(1) + NPRE*I1
            NNEGS = IFIX((FLOAT(MAXSCO(2))+2.0)/2.0)
            DO 40 K = 1,NNEGS
               ISCO = MAXSCO(2) - 2*(K-1)
               USCO2(ISCO) = 0.0D0
               DO 30 L = 1,NINV
                  M = L - 1
                  IOSCO = IABS(ISCO + 1 - NINV + (2*M))
                  IF (IOSCO .GT. MAXSCO(1)) GO TO 30
                  USCO2(ISCO) = USCO2(ISCO) + ICOEF(L)*USCO1(IOSCO)
   30          CONTINUE
   40       CONTINUE
            DO 50 K = 1,NNEGS
               ISCO = MAXSCO(2) - 2*(K-1)
               USCO1(ISCO) = USCO2(ISCO)
   50       CONTINUE
            MAXSCO(1) = MAXSCO(2)
   60       NPRE = NPRE + I1
   70    CONTINUE
```

```
   80 CONTINUE

C     DETERMINE THE DESIRED PROBABILITY

      RNTOT = 0.0D0
      M2SCO = MAXSCO(2)
      M1SCO = -M2SCO
      DO 90 I = M1SCO,M2SCO,2
         RNTOT = RNTOT + USCO2(IABS(I))
   90 CONTINUE
      SUMPR1 = 0.0D0
      SUMPR2 = 0.0D0
      DO 100 I = M1SCO,M2SCO,2
         PSCO = USCO2(IABS(I))/RNTOT
         SUMPR1 = SUMPR1 + PSCO
         IF (I .GE. IS) SUMPR2 = SUMPR2 + PSCO
  100 CONTINUE
      IF (ABS(SUMPR1-1.0D0) .GT. ETA) RETURN
      IER = 0
      PRKST1 = SUMPR2
      RETURN

C     DETERMINE PROBABILITIES USING AN EDGEWORTH SERIES EXPANSION
C     WITH TERMS UP TO ORDER N**-4 (ORDER BASED ON ABSENCE OF TIES).
C     COMPUTE THE CUMULANTS, IGNORING TIES, WITH ALLOWANCE FOR
C     SHEPPARD'S CORRECTION TO THE CUMULANTS.

  110 IH = 2
      RK2 = N*(N-1)*(5.+2*N)/18.0 - IH**2/12.0
      RK4 = -N*(((6*N+15.)*N+10)*N**2-31)/225.0 + IH**4/120.0
      RK6 = 8*N*((((6*N+21.)*N+21)*N**2-7)*N**2-41)/1323.0-IH**6/252.0
      RK8 = -8*N*(((((10*N+45.)*N+60)*N**2-42)*N**2+20)*N**2-93)/675.0
     2      +IH**8/240.0
      RK10 =128*N*((((((6*N+33.)*N+55)*N**2-66)*N**2+66)*N**2-33)*N**2
     2      -61)/1089.0 - IH**10/132.0

C     CORRECT THE CUMULANTS FOR TIES

      DO 120 L = 1,MAXTUP
         IF (IP(L) .EQ. 0) GO TO 120
         RK2 = RK2 - IP(L)*L*(L-1)*(5.+2*L)/18.0
         RK4 = RK4 + IP(L)*L*(((6*L+15.)*L+10)*L**2-31)/225.0
         RK6 = RK6-IP(L)*8*L*((((6*L+21.)*L+21)*L**2-7)*L**2-41)/1323.
         RK8 = RK8 + IP(L)*8*L*(((((10*L+45.)*L+60)*L**2-42)*L**2+20)*
     2         L**2-93)/675.0
         RK10 =RK10-IP(L)*128*L*((((((6*L+33.)*L+55)*L**2-66)*L**2+66)
     2         *L**2-33)*L**2-61)/1089.0
  120 CONTINUE

C     STANDARDIZE THE CUMULANTS.

      RK4 = RK4/RK2**2
      RK6 = RK6/RK2**3
      RK8 = RK8/RK2**4
      RK10 = RK10/RK2**5

C     COMPUTE THE STANDARDIZED SCORE CORRECTED FOR CONTINUITY, THE
C     RELEVANT HERMITE POLYNOMIALS AND FINALLY THE DESIRED PROBABILITY

      X = (IS-1)/RK2**0.5
      HE(1) = X
      HE(2) = X**2 - 1.0
      DO 130 L = 3,15
         HE(L) = X*HE(L-1) - (L-1)*HE(L-2)
 130  CONTINUE
      HERCUM = RK4*HE(3)/24 + RK6*HE(5)/720 + HE(7)*RK4**2/1152
     2       + RK8*HE(7)/40320+RK6*RK4*HE(9)/17280+HE(11)*RK4**3/82944
     3       + RK10*HE(9)/3628800 + RK4*RK8*HE(11)/967680
     4       + HE(11)*RK6**2/1036800 + HE(13)*RK6*RK4**2/829440
     5       + HE(15)*RK4**4/7962624
      CORRFAC = HERCUM*0.39894228*EXP(-0.5*X**2)
```

```
      PRKST1 = ALNORM(X, .TRUE.) + CORRFAC
      IF (PRKST1 .LT. 0.0) PRKST1 = 0.0
      IF (PRKST1 .GT. 1.0) PRKST1 = 1.0
      RETURN
      END

      SUBROUTINE COEFF(NPRE, ICOEF, NINV, IW, ITUPLE)

C     COMPUTES THE COEFFICIENTS FOR THE RECURRENCE FORMULA USED TO
C     DETERMINE THE DISTRIBUTION OF KENDALL'S SCORE.

C     INPUT PARAMETERS
C        NPRE:     THE PREVIOUS NUMBER OF RANKS USED IN DETERMINING THE
C                  DISTRIBUTION OF S
C        ITUPLE:   THE SIZE OF THE TUPLET BEING ADDED TO NPRE
C        NINV:     NPRE*ITUPLE + 1
C        IW:       A WORK VECTOR

C     OUTPUT PARAMETERS
C        ICOEF:    CONTAINS THE COEFFICIENTS C(L;V1,...,VK,R)

C     .. ARRAY ARGUMENTS ..
      DIMENSION ICOEF(NINV), IW(ITUPLE)

C     INITIALIZE ICOEF TO ZERO. IF ITUPLE = 1 THEN SET ICOEF EQUAL TO
C     THE UNIT VECTOR AND RETURN.

      DO 10 J = 1,NINV
         ICOEF(J) = 0
         IF (ITUPLE .NE. 1) GO TO 10
         ICOEF(J) = 1
   10 CONTINUE
      IF (ITUPLE .EQ. 1) RETURN

C     BEGIN BACKTRACK ALGORITHM

      J = 1
   20 IW(J) = 0
      IF (J .EQ. ITUPLE) GO TO 30
      J = J + 1
      GO TO 20

C     DETERMINE THE NUMBER OF INVERSIONS FOR CURRENT POSITION VECTOR.

   30 NUM = 0
      DO 40 K = 1,ITUPLE
         NUM = NUM + IW(K)
   40 CONTINUE
      ICOEF(NUM+1) = ICOEF(NUM+1) + 1

C     IF ALL PERMISSIBLE POSITIONS HAVE BEEN EXHAUSTED RETURN.

      IF (IW(ITUPLE) .EQ. NPRE) RETURN

C     COMMENCE BACKTRACKING UNTIL A SUITABLE VALUE OF J IS FOUND FOR
C     WHICH TO ADVANCE. ADVANCE THE POSITION FOR THAT J AND THEN RESET
C     ALL POSITIONS FOR LARGER J'S TO ZERO.

   50 IF (IW(J) .EQ. IW(J-1)) GO TO 70
   60 IW(J) = IW(J) + 1
      IF (J .EQ. ITUPLE) GO TO 30
      J = J + 1
      GO TO 20
   70 J = J - 1
      IF (J .EQ. 1) GO TO 60
      GO TO 50
      END

      FUNCTION PRKST2(IS, IPX, IPY, IHX, IHY, MXX, MXY, NX, NY,
     *                ISOBS, NIW, PSCO, NW, N, IAL, ETA, IER)
```

```
C     GIVEN A VALUE IS OF KENDALL'S SCORE FOR TWO RANKINGS OF N
C     OBJECTS, BOTH CONTAINING TIES, THIS FUCTION COMPUTES THE
C     PROBABILITY OF OBTAINING A VALUE GREATER THAN, OR EQUAL TO,
C     IS.

C     INPUT PARAMETERS
C        IAL:  SPECIFIES APPROXIMATION OR EXACT ENUMERATION
C        ETA:  ERROR TOLERANCE FOR SUM OF CUMULATIVE PROBABILITIES
C        IPX, IPY, IHX, IHY, MXX, MXY, NX, NY

C     .. ARRAY ARGUMENTS ..
      DIMENSION IPX(MXX), IPY(MXY), IHX(NX), IHY(NY), ISOBS(0:NIW)
      DOUBLE PRECISION PSCO(0:NW)
C     .. LOCAL SCALARS ..
      DOUBLE PRECISION SUMPR1, SUMPR2
C     .. LOCAL ARRAYS ..
      REAL HE(7)

      IER = 3
      MXSCO = (N*(N-1))/2
      IF (IABS(IS) .GT. MXSCO) RETURN
      GO TO (10,50), IAL

C     PSCO(L) STORES THE PROBABILITY OF OBTAINING A SCORE L.
C     INITIALIZE PSCO TO ZERO WITHIN A KNOWN RANGE FOR L.

   10 MXSCO1 = MXSCO*2
      DO 20 I = 0,MXSCO1
         PSCO(I) = 0.0D0
         ISOBS(I) = 0
   20 CONTINUE

C     COMPUTE THE NULL DISTRIBUTION OF KENDALL'S SCORE.

      ITYPE = 1
      CALL MULTHPS(IPX, IPY, IHX, IHY, MXX, MXY, NX, NY, PSCO,
     *                          ISOBS, MXSCO1, N, ITYPE, IER)
      IF (IER .EQ. 1) THEN
         PRINT *, 'AN APPROXIMATION WILL SUFFICE SINCE NX OR NY',
     *            ' EXCEEDS MAXI1=10'
         GO TO 50
      ELSE IF (IER .EQ. 2) THEN
         PRINT *, 'OKAY: I AM SWITCHING TO AN APPROXIMATION'
         GO TO 50
      ELSE IF (IER .NE. 0) THEN
         IER = 5
         RETURN
      END IF
      IER = 4
      SUMPR1 = 0.0D0
      SUMPR2 = 0.0D0
      MXSCOP1 = MXSCO + 1
      DO 30 K = MXSCOP1,MXSCO1
         K1 = MXSCOP1 + MXSCO1 - K
         IF (ISOBS(K1) .EQ. 0) GO TO 30
         ISCO = MXSCO - K1
         SUMPR1 = SUMPR1 + PSCO(K1)
         IF (ISCO .GE. IS) SUMPR2 = SUMPR2 + PSCO(K1)
   30 CONTINUE
      DO 40 K = 0,MXSCO
         IF (ISOBS(K) .EQ. 0) GO TO 40
         SUMPR1 = SUMPR1 + PSCO(K)
         IF (K .GE. IS) SUMPR2 = SUMPR2 + PSCO(K)
   40 CONTINUE
      IF (ABS(SUMPR1-1.0D0) .GT. ETA) RETURN
      IER = 0
      PRKST2 = SUMPR2
      RETURN

C     PROBABILITY ASCERTAINED VIA AN EDGEWORTH SERIES EXPANSION.
C     COMPUTE THE CUMULANTS, IGNORING TIES, WITH ALLOWANCE FOR
```

```
C     SHEPPARD'S CORRECTION TO THE CUMULANTS BASED ON AN INTERVAL
C     WIDTH OF ONE.

   50 IH = 1
      RK2 = N*(N-1)*(5.0+2*N)/18.0 - IH**2/12.0
      RK4 = -N*(((6*N+15.)*N+10)*N**2-31)/225.0 + IH**4/120.0

C     CORRECT THE CUMULANTS FOR TIES

      C1VARX = 0.0
      C2VARX = 0.0
      DO 60 L = 1,MXX
         IF (IPX(L) .EQ. 0) GO TO 60
         RK2 = RK2 - IPX(L)*L*(L-1)*(5.0+2*L)/18.0
         RK4 = RK4 + IPX(L)*L*(((6*L+15.)*L+10)*L**2-31)/225.0
         C1VARX = C1VARX + IPX(L)*L*(L-1.0)*(L-2.0)
         C2VARX = C2VARX + IPX(L)*L*(L-1.0)
   60 CONTINUE
      C1VARY = 0.0
      C2VARY = 0.0
      DO 70 L = 1,MXY
         IF (IPY(L) .EQ. 0) GO TO 70
         RK2 = RK2 - IPY(L)*L*(L-1)*(5.0+2*L)/18.0
         RK4 = RK4 + IPY(L)*L*(((6*L+15.)*L+10)*L**2-31)/225.0
         C1VARY = C1VARY + IPY(L)*L*(L-1.0)*(L-2.0)
         C2VARY = C2VARY + IPY(L)*L*(L-1.0)
   70 CONTINUE
      CVAR = (C1VARX*C1VARY/(9*(N-2)) + C2VARX*C2VARY/2)/(N*(N-1))
      RK2 = RK2 + CVAR
      RK3 = 0.0
      D1 = FLOAT(N*(N-1)*(N-2)*(N-3))
      D2 = FLOAT(N*(N-1)*(N-2))
      DO 190 I = 1,NX
         C1X = FLOAT(IHX(I)*(IHX(I)-1)*(IHX(I)-2))
         C2X = FLOAT(IHX(I)*(IHX(I)-1))
         IP1 = I + 1
         IM1 = I - 1
         DO 180 J = 1,NY
            C1Y = FLOAT(IHY(J)*(IHY(J)-1)*(IHY(J)-2))
            C2Y = FLOAT(IHY(J)*(IHY(J)-1))
            R = C1X*C1Y/D1 + 3*C2X*C2Y/D2
            IF (R .EQ. 0.0) GO TO 180
            SUMS = 0.0
            JP1 = J + 1
            JM1 = J - 1
            IF (I .EQ. NX) GO TO 120
            DO 110 I1 = IP1,NX
               IF (J .EQ. NY) GO TO 90
               DO 80 J1 = JP1,NY
                  SUMS = SUMS + IHX(I1)*IHY(J1)
   80          CONTINUE
   90          IF (J .EQ. 1) GO TO 110
               DO 100 J1 = 1,JM1
                  SUMS = SUMS - IHX(I1)*IHY(J1)
  100          CONTINUE
  110       CONTINUE
  120       IF (I .EQ. 1) GO TO 170
            DO 160 I1 = 1,IM1
               IF (J .EQ. NY) GO TO 140
               DO 130 J1 = JP1,NY
                  SUMS = SUMS - IHX(I1)*IHY(J1)
  130          CONTINUE
  140          IF (J .EQ. 1) GO TO 160
               DO 150 J1 = 1,JM1
                  SUMS = SUMS + IHX(I1)*IHY(J1)
  150          CONTINUE
  160       CONTINUE
  170       RK3 = RK3 + R*SUMS
  180    CONTINUE
  190 CONTINUE
```

```
C     STANDARDIZE THE CUMULANTS

      RK3 = RK3/SQRT(RK2**3)
      RK4 = RK4/RK2**2

C     COMPUTE THE APPROPRIATE CORRECTION FOR CONTINUITY AS
C     RECOMMENDED BY BURR (1960, EQ. 9.2)

      IF (NX .GT. 2 .AND. NY .GT. 2) THEN
         CF = 0.5
         GO TO 220
      ELSE IF (NX .EQ. 2) THEN
         NYM1 = NY-1
         DO 200 J = 1,NYM1
            IHY(J) = IHY(J) + IHY(J+1)
  200    CONTINUE
         CF = 0.5*IHCF(IHY,NYM1)
         GO TO 220
      ELSE IF (NY .EQ. 2) THEN
         NXM1 = NX-1
         DO 210 J = 1,NXM1
            IHX(J) = IHX(J) + IHX(J+1)
  210    CONTINUE
         CF = 0.5*IHCF(IHX,NXM1)
      END IF

C     FIND THE DESIRED PROBABILITY

  220 X = (IS-CF)/SQRT(RK2)
      HE(1) = X
      HE(2) = X**2 - 1.0
      DO 230 L = 3,5
         HE(L) = X*HE(L-1) - (L-1)*HE(L-2)
  230 CONTINUE
      HERCUM = RK3*HE(2)/6 + RK4*HE(3)/24 + RK3**2*HE(5)/72
      CORRFAC = HERCUM*0.39894228*EXP(-0.5*X**2)
      PRKST2 = ALNORM(X, .TRUE.) + CORRFAC
      IF(PRKST2 .LT. 0.0) PRKST2 = 0.0
      IF(PRKST2 .GT. 1.0) PRKST2 = 1.0
      IER = 0
      RETURN
      END

      FUNCTION PRSPDT(ID, IPX, IPY, IHX, IHY, MXX, MXY, NX, NY, ISOBS,
     *                          NIW, PSCO, NW, N, IAL, U, T, ETA, IER)

C     GIVEN A VALUE ID OF SPEARMAN'S SCORE (*4) FOR TWO RANKINGS,
C     WITH ONE OR BOTH CONTAINING TIES, THIS FUCTION COMPUTES THE
C     PROBABILITY OF OBTAINING A VALUE GREATER THAN, OR EQUAL TO, ID.

C     INPUT PARAMETERS
C        IAL:  SPECIFIES APPROXIMATION OR EXACT ENUMERATION
C        ETA:  ERROR TOLERANCE FOR SUM OF CUMULATIVE PROBABILITIES
C        IPX, IPY, IHX, IHY, MXX, MXY, NX, NY, U, T

C     .. ARRAY ARGUMENTS ..
      DIMENSION IPX(N), IPY(N), IHX(N), IHY(N), ISOBS(0:NIW)
      DOUBLE PRECISION PSCO(0:NW)
C     .. LOCAL SCALARS ..
      DOUBLE PRECISION SUMPR1, SUMPR2
C     .. LOCAL ARRAYS ..

      IER = 3
      MXSCO = 4*N*(N**2-1)/3
      IF (ID .GT. MXSCO .OR. ID .LT. 0) RETURN
      WRITE (7,*) 'THE CORRECTION FACTOR IS:', CF
      GO TO (10,40), IAL

C     PSCO(L) STORES THE PROBABILITY OF OBTAINING A SCORE L.
C     INITIALIZE PSCO TO ZERO WITHIN A KNOWN RANGE FOR L.
```

```
   10 DO 20 I = 0,MXSCO
         PSCO(I) = 0.0D0
         ISOBS(I) = 0
   20 CONTINUE
C     COMPUTE THE NULL DISTRIBUTION OF SPEARMAN'S SCORE.
      ITYPE = 2
      CALL MULTHPS(IPX, IPY, IHX, IHY, MXX, MXY, NX, NY, PSCO,
     *                              ISOBS, MXSCO, N, ITYPE, IER)
      IF (IER .EQ. 1) THEN
         PRINT *, 'AN APPROXIMATION WILL SUFFICE SINCE NX OR NY',
     *            ' EXCEEDS MAXI1=10'
         GO TO 40
      ELSE IF (IER .EQ. 2) THEN
         PRINT *, 'OKAY: I AM SWITCHING TO AN APPROXIMATION'
         GO TO 40
      ELSE IF (IER .NE. 0) THEN
         IER = 5
         RETURN
      END IF
      IER = 4
      SUMPR1 = 0.0D0
      SUMPR2 = 0.0D0
      DO 30 I = 0,MXSCO
         IF (ISOBS(I) .EQ. 0) GO TO 30
         SUMPR1 = SUMPR1 + PSCO(I)
         IF (I .GE. ID) SUMPR2 = SUMPR2 + PSCO(I)
   30 CONTINUE
      IF (ABS(SUMPR1-1.0D0) .GT. ETA) RETURN
      IER = 0
      PRSPDT = SUMPR2
      RETURN
C     PROBABILITY ASCERTAINED VIA NORMAL APPROXIMATION WITH A
C     CORRECTION FOR CONTINUITY AS RECOMMENDED BY VALZ (1990).
   40 SHX = 0.0
      NXM1 = NX-1
      DO 50 J = 1,NX
         IF (J .GT. 1) RX1 = RX
         IF (J .GT. 1) SHX = SHX + FLOAT(IHX(J-1))
         RX = 0.5*(IHX(J)+1) + SHX
         IF (J .EQ. 1) GO TO 50
         IHX(J-1) = IFIX(2*(RX - RX1))
   50 CONTINUE
      SHY = 0.0
      NYM1 = NY-1
      DO 60 J = 1,NY
         IF (J .GT. 1) RY1 = RY
         IF (J .GT. 1) SHY = SHY + FLOAT(IHY(J-1))
         RY = 0.5*(IHY(J)+1) + SHY
         IF (J .EQ. 1) GO TO 60
         IHY(J-1) = IFIX(2*(RY - RY1))
   60 CONTINUE
      CF = 0.25*IHCF(IHX,NXM1)*IHCF(IHY,NYM1)
      U = U/12
      T = T/12
      C = FLOAT(N*(N**2-1))/6
      D = ID/4.0
      SRHO = (C-D-T-U+CF)/SQRT((C-2*T)*(C-2*U))
      X = SRHO*SQRT(N-1.0)
      PRSPDT = ALNORM(X, .FALSE.)
      IF (PRSPDT .LT. 0.0) PRSPDT = 0.0
      IF (PRSPDT .GT. 1.0) PRSPDT = 1.0
      IER = 0
      RETURN
      END

      SUBROUTINE MULTHPS(IPX, IPY, IHX, IHY, MXX, MXY, NX, NY,
     *                   PSCO, ISOBS, MXSCO1, N, ITYPE, IER)
```

```
C     DETERMINES AND MULTIPLIES THE HOMOGENEOUS PRODUCT SUMS.
C     ASCERTAINS THE CORRESPONDING KENDALL'S OR SPEARMAN'S
C     SCORE AND THE PROBABILITY OF OCCURRENCE FOR EACH
C     ADMISSIBLE PERMUTATION.

C     .. ARRAY ARGUMENTS ..
      DIMENSION IPX(MXX), IPY(MXY), IHX(NX), IHY(NY), ISOBS(0:MXSCO1)
      DOUBLE PRECISION PSCO(0:MXSCO1)
C     .. LOCAL SCALARS ..
      DOUBLE PRECISION RNFACT, PFACT, OFACT, AIJFACT, RATFACT, SUM
      CHARACTER*1 RESP
C     .. LOCAL ARRAYS ..
      INTEGER IHO(10), IHP(10), IAD(10), LOC(10), NCOMP(10),
     *        ICUMSUM(0:10,10)
      REAL RFIX(30), RPERM(30), SHO(10), SHP(10)
C     .. ARRAYS IN COMMON ..
      INTEGER*2 ICOMP(7,10,500)

      COMMON MAXI1, MAXI2, MAXI3, ICOMP

C     MAXI1, MAXI2, MAXI3 LIMITS FOR ICOMP

      IER = 1
      MAXI1=7
      MAXI2=10
      MAXI3=500

C     IF EITHER NX OR NY EXCEEDS MAXI2 RETURN AND USE APPROXIMATION

      IF (NX .GT. MAXI2 .OR. NY .GT. MAXI2) RETURN
      IER = 0
      NPERM = 0
      SUM = 0.0D0
      RNFACT = 1.0D0
      DO 40 I = 2,N
         RNFACT = RNFACT*I
   40 CONTINUE
      MXSCO = MXSCO1/2

C     ASSIGN X OR Y TO PARCELS, THE OTHER TO OBJECTS, SO AS TO
C     MAXIMIZE COMPUTATIONAL EFFICIENCY. DETERMINE THE COMPOSITION
C     VECTORS AND STORE THEM IN A COMMON ARRAY FOR EASY ACCESS

      NK = NY
      IF (NX .GT. NY) NK = NX
      CALL ASSIGN(IHX, NX, IHY, NY, IHO, NO, IHP, NP, MXX, MXY, NK)
      CALL COMPO(IHP, NP, IHO, NO, IAD, NAD, NCOMP, IER)
      IF (IER .NE. 0) RETURN

C     DETERMINE THE FIXED RANKING VECTOR, RFIX. INITIALIZE THE
C     VECTOR ICUMSUM, WHICH STORES THE PARTIAL PRODUCTS, TO ZERO.

      INDEX = 1
      PFACT = 1.0D0
      SHP(1) = 0.0
      DO 60 I = 1,NP
         IF (I .GT. 1) SHP(I) = SHP(I-1) + FLOAT(IHP(I-1))
         DO 50 J = 1,IHP(I)
            RFIX(INDEX) = 0.5*(IHP(I)+1) + SHP(I)
            INDEX = INDEX + 1
            PFACT = PFACT*J
   50    CONTINUE
   60 CONTINUE
      OFACT = 1.0D0
      DO 80 I = 1,NO
         ICUMSUM(0,I) = 0
         DO 70 J = 1,IHO(I)
            OFACT = OFACT*J
   70    CONTINUE
   80 CONTINUE
      RATFACT = (PFACT*OFACT)/RNFACT
```

```
C     MULTIPLY THE CUMPOSITIONS.

      J = 1
   90 LOC(J) = 1
  100 DO 110 I = 1,NO
         ICUMSUM(J,I) = ICUMSUM(J-1,I) + ICOMP(IAD(J),I,LOC(J))
         IF (ICUMSUM(J,I) .GT. IHO(I)) GO TO 160
         IF (J .LT. NP) GO TO 110
         IF (ICUMSUM(J,I) .LT. IHO(I)) GO TO 160
  110 CONTINUE
      IF (J .EQ. NP) GO TO 120
      J = J + 1
      GO TO 90

C     THE MULTIPLICATION HAS YIELDED A SUCCESSFUL RESULT.
C     DETERMINE THE CORRESPONDING PERMUTATION VECTOR AND AIJFACT

  120 INDEX = 1
      AIJFACT = 1.0D0
      SHO(1) = 0.0
      DO 150 J = 1,NP
         DO 140 I = 1,NO
            IF (J.EQ.1 .AND. I.GT.1) SHO(I)=SHO(I-1)+FLOAT(IHO(I-1))
            IF (ICOMP(IAD(J),I,LOC(J)) .EQ. 0) GO TO 140
            DO 130 K = 1,ICOMP(IAD(J),I,LOC(J))
               RPERM(INDEX) = 0.5*(IHO(I)+1) + SHO(I)
               INDEX = INDEX + 1
               AIJFACT = AIJFACT*K
  130       CONTINUE
  140    CONTINUE
  150 CONTINUE
      GO TO 180
  160 IF (LOC(J) .EQ. NCOMP(J)) GO TO 170
      LOC(J) = LOC(J) + 1
      GO TO 100
  170 J = J - 1
      IF (J .EQ. 0) RETURN
      GO TO 160

C     IF AN ADMISSIBLE CONFIGURATION IS OBTAINED THEN COMPUTE THE
C     ASSOCIATED SCORE. INCREMENT PSCO(SCORE) BY THE PROBABILITY
C     OF OBTAINING THE PERMUTATION.

  180 IDSCO = ISCORE(RFIX, RP[illegible], ITYPE)
      IF (IDSCO .LT. 0) IDS[illegible]O - IDSCO
      PSCO(IDSCO) = PSCO(IDSC[illegible]RA..ACT/AIJFACT
      ISOBS(IDSCO) = 1
      NPERM = NPERM + 1
      SUM = SUM + 1.0D0/AIJFACT
      IF (MOD(NPERM,5000) .EQ. 0) THEN
         RPNPERM = NPERM/(RATFACT*SUM)
         PRINT *, 'ESTIMATED # OF PERMUTATIONS AFTER   ', NPERM,
     *            'ENUMERATIONS IS:   ', RPNPERM
      ENDIF

C     IF ENUMERATION IS TAKING TOO LONG RETURN AND USE APPROXIMATION

      IF (MOD(NPERM,15000) .EQ. 0) THEN
         PRINT *, 'DO YOU WISH TO CONTINUE? (Y/N)'
         READ (*,99999) RESP
         IF (RESP .EQ. 'N' .OR. RESP .EQ. 'n') THEN
            IER = 2
            RETURN
         END IF
      END IF
99999 FORMAT (A1)
      GO TO 170
      END

      SUBROUTINE ASSIGN(IHX,NX,IHY,NY,IHO,NO,IHP,NP,MXX,MXY,NK)
```

```
C     ASSIGNS THE APPROPRIATE RANKS AND ASSOCIATED PARAMETERS TO BE
C     THE SET OF PARCELS (FIXED RANKS) AND THE SET OF OBJECTS
C     (PERMUTED RANKS), RESPECTIVELY

C     INPUT PARAMETERS
C        IHX, IHY, NX, NY, MXX, MXY, NK

C     OUTPUT PARAMETERS
C        IHO, IHP, NO, NP

C     .. ARRAY ARGUMENTS ..
      DIMENSION IHX(NX), IHY(NY), IHO(NK), IHP(NK)

      OBJECT = 1
      IF (NY .EQ. NX) GO TO 10
      IF (NY .LT. NX) OBJECT = 2
      GOTO (20, 50), OBJECT
   10 IF (MXX .LT. MXY) OBJECT = 2
      GOTO (20, 50), OBJECT
   20 NO = NX
      NP = NY
      DO 30 I = 1,NX
         IHO(I) = IHX(I)
   30 CONTINUE
      DO 40 I = 1,NY
         IHP(I) = IHY(I)
   40 CONTINUE
      GO TO 80
   50 NO = NY
      NP = NX
      DO 60 I = 1,NX
         IHP(I) = IHX(I)
   60 CONTINUE
      DO 70 I = 1,NY
         IHO(I) = IHY(I)
   70 CONTINUE
   80 RETURN
      END

      SUBROUTINE COMPO(IHP, NP, IHO, NO, IAD, NAD, NCOMP, IER)

C     SERVES AS A CONTROL FOR SUBROUTINE COMBIN WHICH DETERMINES THE
C     NO-PART COMPOSITIONS OF IHP(I), I = 1, ... , NP. FOR IHP(J) =
C     IHP(I), J>I, THE COMPOSITIONS ARE IDENTICAL AND HENCE AN ADDRESS
C     VECTOR, USED IN CONJUNCTION WITH A COMPOSITION ARRAY FOR DISTINCT
C     VALUES OF IHP(I), DETERMINES THE SET OF REQUIRED COMPOSITIONS.
C     ANY COMPOSITION SUCH THAT ICOMP(J) > IHO(J) MAY BE REGARDED AS AN
C     INADMISSIBLE COMPOSITION AND HENCE IT IS IGNORED

C     INPUT PARAMETERS
C        IHP, IHO, NP, NO

C     OUTPUT PARAMETERS
C        IAD:      AN ADDRESS VECTOR
C        NAD:      NUMBER OF DIFFERENT ADDRESSES
C        NCOMP:    VECTOR WITH # OF ADMISSIBLE COMPOSITIONS FOR EACH
C                  IHP(I)

C     .. ARRAY ARGUMENTS ..
      DIMENSION IHP(NP), IHO(NO), IAD(NP), NCOMP(NP)

      ICOUNT = 1
      DO 30 I = 1,NP
         IF (I .EQ. 1) GO TO 20
         J = I
   10    J = J - 1
         IF (IHP(I) .NE. IHP(J) .AND. J .GT. 1) GO TO 10
         IF (IHP(I) .NE. IHP(J)) GO TO 20
         IAD(I) = IAD(J)
         NCOMP(I) = NCOMP(J)
         GO TO 30
```

```
   20    CALL COMBIN(ICOUNT, IHP(I), NCOMP(I), IHO, NO, IER)
         IF (IER .NE. 0) RETURN
         NCOMP(I) = NCOMP(I) - 1
         IAD(I) = ICOUNT
         ICOUNT = ICOUNT + 1
   30 CONTINUE
      NAD = ICOUNT - 1
      RETURN
      END

      SUBROUTINE COMBIN(IPL, IHPI, NCOMPI, IHO, NO, IFAULT)

C     CALLED BY SUBROUTINE COMPO
C     GENERATES ALL N(C)M WAYS OF PLACING M+1 BALLS INTO N-M CELLS.
C     THESE ARE THEN USED TO SPECIFY THE SET OF M+1 PART COMPOSITIONS OF
C     N-M. THE BASIC ALGORITHM USED IS DUE TO EHRLICH

C     INPUT PARAMETERS
C        IPL:      1ST ARGUMENT OF ICOMP ARRAY
C        IHPI:     VALUE OF IHP(I)
C        IHO, NO

C     OUTPUT PARAMETERS
C        ICOMP:   ARRAY OF COMPOSITIONS. STORED IN COMMON
C        NCOMPI:  NUMBER OF ADMISSIBLE COMPOSITIONS + 1

C     .. ARRAY ARGUMENTS ..
      DIMENSION IHO(NO)
C     .. LOCAL ARRAYS ..
      INTEGER IBAR(30), ICOMB(30)
C     .. ARRAYS IN COMMON ..
      INTEGER*2 ICOMP(7,10,500)
      COMMON MAXI1, MAXI2, MAXI3, ICOMP
C     ..
      IFAULT = 3
      IF (IPL .GT. MAXI1) RETURN
      IFAULT = 0
      N = NO + IHPI - 1
      M = NO - 1

C     INITIALIZE IBAR AND DETERMINE THE INITIAL COMBINATION VECTOR

      IBAR(N) = 0
      NCOMPI = 1
      DO 10 J = 1,N
         ICOMB(J) = 1
         IF (J .GT. M) ICOMB(J) = 0
   10 CONTINUE
      I = 0
      GO TO 80

C     COMPUTE THE COMPOSITION VECTOR. IF ADMISSIBLE STORE AND INCREMENT
C     NCOMPI BY 1

   20 INDEX = 0
      ITRACK = 1
      DO 40 J = 1,N
         IF (ICOMB(J) .EQ. 0) GO TO 30
         ICOMP(IPL,ITRACK,NCOMPI) = INDEX
         IF (ICOMP(IPL,ITRACK,NCOMPI) .GT. IHO(ITRACK)) GO TO 50
         ITRACK = ITRACK + 1
         IF (ITRACK .GT. MAXI2) GOTO 100
         INDEX = 0
         GO TO 40
   30    INDEX = INDEX + 1
   40 CONTINUE
      ICOMP(IPL,ITRACK,NCOMPI) = INDEX
      IF (ICOMP(IPL,ITRACK,NCOMPI) .GT. IHO(ITRACK)) GO TO 50
      NCOMPI = NCOMPI + 1
      IF (NCOMPI .GT. MAXI3) GOTO 110
```

```
C     DETERMINE THE NEXT COMBINATION VECTOR

   50 I = N
   60 I = I - 1
      IF (IBAR(I) .EQ. 0 .AND. I .GE. 2) GO TO 60
      IF (I .EQ. 1 .AND. IBAR(I) .EQ. 0) RETURN
      IBAR(I) = 0
      J = I
   70 J = J + 1
      IF (ICOMB(J) .EQ. ICOMB(I)) GO TO 70
      INT = ICOMB(J)
      ICOMB(J) = ICOMB(I)
      ICOMB(I) = INT
   80 L1 = I + 1
      L2 = N - 1
      DO 90 J = L1,L2
         K = L2 + L1 - J
         IF (I .EQ. 0) IBAR(K) = 0
         IF (ICOMB(K) .NE. ICOMB(K+1) .OR. IBAR(K+1) .EQ. 1) IBAR(K) = 1
   90 CONTINUE
      GO TO 20
  100 IFAULT = 4
      RETURN
  110 IFAULT = 5
      RETURN
      END

      INTEGER FUNCTION ISCORE(RFIX, RPERM, N, ITYPE)

C     COMPUTES KENDALL'S SCORE OR SPEARMAN'S SCORE*4 FOR EACH
C     PERMUTATION OF THE TWO SETS OF RANKS

C     INPUT PARAMETERS
C        RFIX:   THE FIXED RANKING
C        RPERM:  THE PERMUTED RANKING
C        N:      THE # OF RANKS IN EACH RANKING
C        ITYPE:  SPECIFIES SPEARMAN'S OR KENDALL'S SCORE

C     .. ARRAY ARGUMENTS ..
      DIMENSION RFIX(N), RPERM(N)

      ISCORE = 0
      GO TO (10,40), ITYPE
   10 DO 30 I = 2,N
         IM1 = I - 1
         DO 20 J = 1,IM1
            DIFIX = RFIX(I) - RFIX(J)
            DIPERM = RPERM(I) - RPERM(J)
            IF (DIFIX .EQ. 0.0 .OR. DIPERM .EQ. 0.0) GO TO 20
            ISGNFIX = IFIX(DIFIX/ABS(DIFIX))
            ISGNPER = IFIX(DIPERM/ABS(DIPERM))
            ISCORE = ISCORE + ISGNFIX*ISGNPER
   20    CONTINUE
   30 CONTINUE
      RETURN
   40 DO 50 I = 1,N
         ISCORE = ISCORE + IFIX(4*(RFIX(I)-RPERM(I))**2)
   50 CONTINUE
      RETURN
      END

      INTEGER FUNCTION IHCF(IX, N)

C     COMPUTES THE HIGHEST COMMON FACTOR OF THE NUMBERS IN X)

C     .. ARRAY ARGUMENTS ..
      DIMENSION IX(N)

      IF (IX(1) .EQ. 1) THEN
         IHCF = 1
```

```
      END IF
      IHCF = IX(1)
      DO 10 I = 2,N
         IHCF = IGCF(IHCF,IX(I))
         IF (IHCF .EQ. 1) GO TO 20
   10 CONTINUE
   20 RETURN
      END

      INTEGER FUNCTION IGCF(IH, IX)

C     THE EUCLIDEAN ALGORITHM IS USED TO FIND THE GCD OF 2 NUMBERS

      IR1 = IX
      IR2 = IH
   10 IRTEMP = IR2
      IR2 = MOD(IR1,IR2)
      IR1 = IRTEMP
      IF (IR2 .EQ. 0) THEN
         IGCF = IRTEMP
         RETURN
      END IF
      GO TO 10
      END
```

Chapter 4

THE DISTRIBUTION OF KENDALL'S PARTIAL TAU UNDER THE COMPLETE NULL HYPOTHESIS

Let R_1, R_2 and R_3 be the rankings of n individuals with respect to three criteria. Kendall (1942) defined a partial rank correlation coefficient $t_{12.3}$ so that

$$t_{12.3} = \frac{t_{12} - t_{13}t_{23}}{\sqrt{(1-t_{13}^2)(1-t_{23}^2)}} , \tag{4.1}$$

which expresses partial tau in terms of the coefficients t_{ij} between the original pairs of rankings. Kendall argued that this coefficient provides a measure of the correlation between R_1 and R_2 independently of the influence of R_3.

Moran (1951) considered the distribution of partial tau and was unable to find an expression for $Var(t_{12.3})$. Hoflund (1963) simulated the distributions of $t_{12.3}$ for $n = 4, 5, \ldots, 10$. Maghsoodloo (1975) generated the exact sampling distributions of $t_{12.3}$ for $n = 3, 4, \ldots, 7$ and used large scale Monte Carlo sampling to estimate the quantiles for $n = 8(1)20$ and $n = 25, 30$. Maghsoodloo and Pallos (1981) subsequently estimated the quantiles for $n = 35(5)50(10)90$ and the variance for $n = 8(1)20(5)50(10)90$. Hoeffding (1948) concluded that if $t_{13} = t_{23} = 0$, then $\sqrt{n}(t_{12.3} - \tau_{12.3})$ has the same limiting distribution as $\sqrt{n}(t_{12} - \tau_{12})$. Kendall (1975, section 8.7) stated that no tests of significance are yet known for partial tau.

In the sequel, distributional results are developed for use in significance testing under the hypothesis that all rankings are equiprobable which, Hoflund (1963), is referred to as the complete null hypothesis. An algorithm for enumerating the exact distribution of $t_{12.3}$ is developed via application of the concept of an inversion vector which was introduced in Section 3.2. Next, an expression for $Var(t_{12.3})$, based on the approximation $\mathcal{E}(\text{ratio}) \approx \text{ratio(expectations)}$ and the notion of a relative inversion vector, is derived. Upper and lower bounds for $\mathcal{V}ar(t_{12.3})$ are established and a proof of the asymptotic normality of $t_{12.3}$ is given. It is assumed that there are no ties in the rankings.

4.1 Algorithm for enumerating the distribution of $t_{12.3}$

Let R_3, the fixed ranking, be in the natural order. Since, for n observations per ranking,

$$t_{ij} = 1 - \frac{4Q_{ij}}{n(n-1)} \tag{4.2}$$

where Q_{ij} is the negative score for rankings R_i and R_j, it follows from eqn. (4.1) that $t_{12.3}$ is a function of $(n, Q_{12}, Q_{13}, Q_{23})$. Hence determination of the distribution of $t_{12.3}$ requires determination of the conditional distribution of Q_{12} given Q_{13}, Q_{23} and n which is subsequently written as $Q_{12}|Q_{13}, Q_{23}, n$.

4.1.1 Inversion vectors: a useful tool

It is now shown that the inversion vector is an ideal tool to work with. Let $I_n = (i_1, i_2, \ldots, i_n)$ be an inversion vector. Then $n+1$ new inversion vectors $((i_1, \ldots, i_n, 0), (i_1, \ldots, i_n, 1), \ldots, (i_1, \ldots, i_n, n))$ may be generated by adding $i_{n+1} = 0, 1, \ldots, n$ respectively, to I_n. Applying this procedure to each of the original $n!$ inversion vectors generates $(n+1)!$ inversion vectors corresponding to the set of $(n+1)!$ permutations of the integers $1, 2, \ldots, n+1$. Since

$$Q_{ij,n+1} = \sum_{\ell=1}^{n+1} i_\ell = Q_{ij,n} + i_{n+1} \tag{4.3}$$

it follows that if $U(n+1, Q_{ij})$ is the number of I_{n+1} vectors with Q_{ij} inversions, then

$$U(n+1, Q_{ij}) = U(n, Q_{ij}) + U(n, Q_{ij}-1) + \cdots + U(n, Q_{ij}-n)\,. \tag{4.4}$$

Eqn. (4.4), which is analogous to Kendall's (1938) recursion formula for determining the distribution of the total score S, is the equation which Kendall and

Stuart (1973, Vol. 2, Chapter 31) used for developing distributional results for t_{ij}. It is thus seen that use of the inversion vector provides a formal framework for establishing the recursion relationship.

4.1.2 Problem specification using inversion vectors

Since specifying an inversion vector for R_1 automatically determines the value of Q_{13}, and similarly for R_2, only the conditional distribution $Q_{12}|I^1, I^2, n$ needs to be considered. Let $R_1 = (x_1, \ldots, x_n)$ and $R_2 = (y_1, \ldots, y_n)$ where $1 \le x_i, y_i \le n$; $x_i \ne x_j$ and $y_i \ne y_j$. The negative score Q is obtained as

$$Q_{12} = \sum_{i<j}^{n} I_{(-\infty,0)}\big((x_i - x_j)(y_i - y_j)\big) \tag{4.5}$$

where $I_{(\bullet)}$ is the indicator function on $(-\infty, 0)$.

Determine the entries in eqn. (4.5) according to the following pairing of elements of R_1

$$(x_1, x_2) \quad (x_1, x_3), (x_2, x_3) \quad (x_1, x_4), \ldots, (x_3, x_4) \quad \ldots \quad (x_1, x_n), \ldots, (x_{n-1}, x_n)$$

which gives $n-1$ entries with the j^{th} entry having j components. Replacing each pair of elements, (x_i, x_j) say, by signum$(x_j - x_i)$ yields a sequence of plus and minus ones which is uniquely determined by I_n^1. Repeating the preceding for R_2, to generate a second sequence of plus and minus ones, taking the pairwise products (same position in both sequences) and summing over the minus ones occurring in the product sequence then gives Q_{12} as defined in eqn. (4.5).

Addition of i_{n+1}^1 to I_n^1 to get I_{n+1}^1 leaves the prior sequence unchanged. An n^{th} entry of plus and minus ones is added. Our interest centers on the pairwise products obtained from two such additions of an n^{th} entry resulting from additions of i_{n+1}^1 and i_{n+1}^2 to I_n^1 and I_n^2, respectively. Since the prior sequences of plus and minus ones, and therefore the prior product sequence, are unchanged, knowledge of the additional pairwise products and of $Q_{12}|I^1, I^2, n$ enables determination of $Q_{12}|I^1, I^2, n+1$. The contribution of the n additional pairwise products to $Q_{12}|I^1, I^2, n+1$ from prior information and the current values of i^1_{n+1} and i^2_{n+1} is determined below.

4.1.3 Determination of the n^{th} column of plus and minus ones

Let $I_n = (i_1, i_2, \ldots, i_n)$ be an inversion vector. Then $0 \le i_n \le n-1$. Let the $(n-1)!$ subset, of the $n!$ inversion vectors, for which i_n is constant form a block. Refer to the resulting n blocks as block 0 through block $n-1$ according to the value of i_n. Inversion vectors are ordered within and between blocks as follows:

1. Set $I_1 = (0)$ to initialize step 2.
2. Given the set of ordered inversion vectors for $n = k$ generate the set of ordered inversion vectors for $n = k+1$ by adding j to the entire set of I_k vectors to obtain block j; $j = 0, 1, \ldots, k$. Order the blocks as block 0, block 1, ..., block k.
3. Set $k = k+1$ and go to step 2. Stop when n is as large as is desired.

Consider addition of i_{n+1} to I_n where I_n is at row position j of block k. Let the permutation corresponding to the resulting I_{n+1} vector be $(x_1, \ldots, x_{n+1})$. Then x_{n+1} has i_{n+1} greater elements on its left while x_n has k greater elements on its left, since $i_n = k$. Therefore

$$\begin{array}{ll} x_n > x_{n+1} & \text{if } i_{n+1} > k \\ x_n < x_{n+1} & \text{if } i_{n+1} \le k\,. \end{array} \tag{4.6}$$

Consequently, the n^{th} component of the n^{th} entry of plus and minus ones is readily determined.

Suppose that the permutation associated with I_n is $P_n = (u_1, \ldots, u_n)$. Then $P_{n+1} = (x_1, \ldots, x_{n+1})$ may be obtained from P_n and i_{n+1} as follows:

$$x_{n+1} = n + 1 - i_{n+1} \tag{4.7a}$$

and

$$x_i = \begin{cases} u_i, & \text{if } u_i < x_{n+1} \\ u_i + 1, & \text{otherwise} \end{cases} ; \qquad i = 1, \ldots n. \tag{4.7b}$$

Eqn. (4.7a) follows from the definition of an inversion vector. Eqn. (4.7b) follows from eqn. (4.7a) and the fact that the first n elements of I_{n+1} are identical to the n elements of I_n.

Now consider the vector $(x_1, \ldots, x_{n-1}, x_{n+1})$ and the associated inversion vector I'_n. If $i_{n+1} > k$, implying that $x_n > x_{n+1}$, then x_{n+1} now has $i_{n+1} - 1$ larger elements on its left. Otherwise x_{n+1} has i_{n+1} larger elements on its left. Consequently, in terms of the n blocks for inversion vectors of size n,

$$\begin{array}{ll} I'_n \in \text{block } i_{n+1} - 1 & \text{if } i_{n+1} > k \\ I'_n \in \text{block } i_{n+1} & \text{if } i_{n+1} \leq k . \end{array} \tag{4.8}$$

The first $n - 1$ elements of I'_n are the same as the first $n - 1$ elements of I_{n+1} since these both depend only upon the sequence $x_1, \ldots, x_{n-1}$. Therefore these $n - 1$ elements are identical to the first $n - 1$ elements of I_n. By construction, the first $n - 1$ elements uniquely determine how far down each block an inversion vector is located. That is, if two inversion vectors in two different blocks occupy the same row positions within their respective blocks then their first $n - 1$ elements are identical. The converse also applies. Therefore I'_n is at row position j of block $i_{n+1} - 1$ or of block i_{n+1} and the first $n - 1$ components of the n^{th} entry of plus and minus ones for I_{n+1} are the $n - 1$ components of the $(n-1)^{\text{th}}$ or last entry for I_n at row position j of block $i_{n+1} - 1$ or of block i_{n+1}, according as $i_{n+1} > k$ or $i_{n+1} \leq k$.

Having thus established that the original row position of I_n in block k is preserved with regard to I'_n in block $i_{n+1} - 1$ or block i_{n+1}, the above argument may be repeated for each I_n in block k to obtain the result that if i_{n+1} is added to block k then the n^{th} column of entries is determined as follows:

$$\left.\begin{array}{ll} \text{add a } - \text{ to each row of the } (n-1)^{\text{th}} \text{ column for block } i_{n+1} - 1 & \text{if } i_{n+1} > k \\ \text{add a } + \text{ to each row of the } (n-1)^{\text{th}} \text{ column for block } i_{n+1} & \text{if } i_{n+1} \leq k . \end{array}\right\} \tag{4.9}$$

Table 4.1 lists permutations and their associated inversion vectors, which are ordered within and between blocks as above, as well as their associated sequences of plus and minus ones, for $n = 1, 2, 3$ and 4. Examination of the permutations, inversion vectors and blocks will demonstrate how eqns. (4.6) through (4.8) work.

Table 4.1: Permutations and inversion vectors

n=1

P	*I*
1	0

n=2

P	*I*	Signs
12	00	+
21	01	−

n=3

P	*I*	Signs
123	000	+ ++
213	010	− ++
132	001	+ +−
312	011	− −+
231	002	+ −−
321	012	− −−

n=4

P	*I*	Signs
1234	0000	+ ++ +++
2134	0100	− ++ +++
1324	0010	+ +− +++
3124	0110	− −+ +++
2314	0020	+ −− +++
3214	0120	− −− +++
1243	0001	+ ++ ++−
2143	0101	− ++ ++−
1423	0011	+ +− +−+
4123	0111	− −+ −++
2413	0021	+ −− +−+
4213	0121	− −− −++
1342	0002	+ ++ +−−
3142	0102	− ++ −+−
1432	0012	+ +− +−−
4132	0112	− −+ −+−
3412	0022	+ −− −−+
4312	0122	− −− −−+
2341	0003	+ ++ −−−
3241	0103	− ++ −−−
2431	0013	+ +− −−−
4231	0113	− −+ −−−
3421	0023	+ −− −−−
4321	0123	− −− −−−

To further illustrate the above arguments, suppose that 3 is added to 0111 to obtain the inversion vector 01113. Since $3 > 1$ and 0111 is the fourth vector in block 1 go to the fourth vector in block 2. Take the last three signs and add a minus sign to get $-+--$ as the $n=4$ components of the 4^{th} entry. Since $P_4 = 4123$ then eqn. (4.7) gives $P_5 = 51342$ which is consistent with $I_5 = 01113$ and $-+--$ as the components of the 4^{th} entry.

4.1.4 Recursive determination of the conditional distribution $Q_{12}|Q_{13}, Q_{23}$

Suppose that the $(n-1)!^2$ matrix arrays of the conditional distributions, $\{Q_{12}|I^1, I^2, n-1\}$ and $\{Q_{12}|I^1, I^2, B_p, B_q, n\}$, where the notation indicates that in the latter case Q_{12} for interactions between the inversion vectors in block p of I^1 and those in block q of I^2 is being considered, are known. It follows from the above arguments that the difference in elements between these two matrix arrays is due to the effect of interactions between the $(n-1)^{th}$ columns of plus and minus ones for blocks p and q. Refer to this difference as the pq block effect ($BE = pq$ in subsequent notation) as one goes from $n-1$ to n. It is proposed that the matrix array $\{Q_{12}|I^1, I^2, n\}$, eqns. (4.6), (4.8) and (4.9), and the pq block effects; $p = 0, 1, \ldots, n-1$; $q = 0, 1, \ldots, n-1$ be used to determine the conditional distribution $\{Q_{12}|I^1, I^2, n+1\}$.

Consider the addition of $i_{n+1} = g, h$ to the current $n!$ inversion vectors thus giving blocks g_{n+1} and h_{n+1}. Each of these two new blocks is composed of n subsets where a subset corresponds to an original block. The $n!^2$ matrix array of the conditional distribution $\{Q_{12}|I^1, I^2, B_g, B_h, n+1\}$ is required. This array comprises n^2 matrix subarrays where each subarray corresponds to multiplication of an $(n-1)!$ subset of block g_{n+1} with an $(n-1)!$ subset of block h_{n+1}. Suppose, therefore, that g is added to current block p and h is added to current block q resulting in subsets pg and qh. Let M be a matrix array of size $(n-1)!^2$ where each element of M is given by

$$m_{ij} = \begin{cases} 0 & \text{if } ((g > p) \cap (h > q)) \cup ((g \leq p) \cap (h \leq q)) \\ 1 & \text{otherwise} \end{cases} \tag{4.10}$$

the matrix M being determined in accordance with eqn. (4.6). An appropriate

matrix of block effects is selected by applying eqn. (4.8) to the pairs g, p and h, q. It then follows from eqn. (4.9) that

$$\{Q_{12}|I_1, I_2, B_g, B_h, S_{pg}, S_{qh}, n+1\} = \{Q_{12}|I_1, I_2, B_p, B_q, n\} + M + BE \quad (4.11)$$

where S_{pg} refers to subset pg.

As an example of how eqn. (4.11) works, consider the distributions for $n = 2$ and $n = 3$; the distribution for $n = 2$ being the 2×2 submatrix in the upper left hand corner of the distribution for $n = 3$ which is shown in table 4.2.

Table 4.2: Distribution of Q_{12} for $n = 3$

$I^1 \backslash I^2$	000	010	001	011	002	012
000	0	1	1	2	2	3
010	1	0	2	1	3	2
001	1	2	0	3	1	2
011	2	1	3	0	2	1
002	2	3	1	2	0	1
012	3	2	2	1	1	0

The block effect matrices are obtained by subtracting the 2×2 matrix obtained for $n-1$ from each of the 2×2 submatrices which comprise the distribution for $n = 3$. This gives

Table 4.3: Block Effect Matrices

block	0	1	2
0	0 0	1 1	2 2
	0 0	1 1	2 2
1	1 1	0 2	1 1
	1 1	2 0	1 1
2	2 2	1 1	0 0
	2 2	1 1	0 0

To compute the distribution $\{Q_{12}|I_1, I_2, B_1, B_2, n+1\}$ observe that $g = 1$

and $h = 2$ which, in conjunction with eqns. (8), (9) and (10), give the appropriate BE and M matrices as

	$B_0 + 2$	$B_1 + 2$	$B_2 + 2$
$B_0 + 1$	$m_{ij} = 0$ $BE = 01$	$m_{ij} = 0$ $BE = 01$	$m_{ij} = 1$ $BE = 02$
$B_1 + 1$	$m_{ij} = 1$ $BE = 11$	$m_{ij} = 1$ $BE = 11$	$m_{ij} = 0$ $BE = 12$
$B_2 + 1$	$m_{ij} = 1$ $BE = 11$	$m_{ij} = 1$ $BE = 11$	$m_{ij} = 0$ $BE = 12$

Combining the block effect matrices, the M matrices and the conditional distribution $\{Q_{12}|I_1, I_2, n\}$ gives the required distribution as

Table 4.4: $\{Q_{12}|I^1, I^2, B_1, B_2, n+1\}$

$I^1 \backslash I^2$	0002	0102	0012	0112	0022	0122
0001	1	2	2	3	5	6
0101	2	1	3	2	6	5
0011	2	5	1	6	2	3
0111	5	2	6	1	3	2
0021	3	6	2	5	1	2
0121	6	3	5	2	2	1

Note that the new 12 block effect matrix is the incremental matrix which was added to $\{Q_{12}|I_1, I_2, n\}$ in order to obtain $\{Q_{12}|I_1, I_2, B_1, B_2, n+1\}$.

The above theoretical development leads to a very simple computer program for deriving the exact distribution of $t_{12.3}$ for $n \leq 6$ - see the Appendix to Chapter 4. A more complex program is used to derive the distribution for $n = 7$. However, the exponential growth of both the machine time and storage requirements limit practical application to $n \leq 7$. Maghsoodloo's algorithm determines all permutations of the integers 1 through n and then calculates $t_{12.3}$ for the appropriate $(n! - 2)^2$ pairwise combinations of these permutations. The essential difference between the two algorithms is that determination of $t_{12.3}$ in Maghsoodloo's algorithm takes $\binom{n}{2}$ steps while determination of $t_{12.3}$ in our algorithm is accomplished in roughly 2 steps via use of the block effect matrices.

4.2 Conjugate rankings and $\mathcal{E}(t_{12.3})$

Kendall (1975, Section 1.18) referred to the concept of conjugate rankings. Corresponding to any two rankings R_i and R_j with correlation t there are two rankings R_i^c and R_j with correlation $-t$. Kendall demonstrated this fact via the reordering of R_i into the natural order. Moran (1951) incorrectly specified that reversing either R_i or R_j reverses the sign of t; from which it would then follow that reversing a ranking yields its conjugate. Reversing R_i does reverse the correlation t when R_j is in its natural order. However, this result does not extend generally to all other rankings, R_j. A definition of the conjugate ranking R_i^c is now given. Let $R_i = (x_1, x_2, \ldots, x_n)$ and $R_j = (y_1, y_2, \ldots, y_n)$ be two rankings of the first n natural numbers. Then

$$R_i^c = (x_1^c, x_2^c, \ldots, x_n^c); \qquad x_i^c = n + 1 - x_i; \qquad i = 1, 2, \ldots, n \tag{4.12}$$

is the conjugate of R_i. The total score S is given by

$$S = \sum_{i<j}^{n} \text{signum}\{(x_i - x_j)(y_i - y_j)\} \tag{4.13}$$

while the conjugate score S^c, based on R_i^c and R_j, is given by

$$\begin{aligned} S^c &= \sum_{i<j}^{n} \text{signum}\{(x_i^c - x_j^c)(y_i - y_j)\} \\ &= -S \end{aligned} \tag{4.14}$$

since $(x_i^c - x_j^c) = -(x_i - x_j)$. It follows that $t^c = -t$ as required.

Corresponding to R_i^c is the conjugate of an inversion vector I_n which may be defined as

$$I_n^c = (0, 1, \ldots, n-1) - I_n \tag{4.15}$$

where I_n^c uniquely defines R_i^c given that I_n defines R_i. It is immediately obvious that any $n!$ set of rankings may be divided into two sets such that the rankings in one set are the conjugates of the rankings in the other set. Moran (1951) has used this property, taken in conjunction with eqn. (4.1), to show that $t_{12.3}$ is symmetrically distributed about zero so that $\mathcal{E}(t_{12.3}) = 0$.

4.3 Derivation of an asymptotic variance estimator for $t_{12.3}$

Kendall (1942) defines $t_{12.3}$ in terms of the entries of a two-way table, Table 4.5, as

$$t_{12.3} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} . \tag{4.16}$$

Table 4.5: Agreements of rankings R_1 and R_2 with R_3

		Ranking R_1		
		Pairs + (agreeing with R_3)	Pairs – (disagreeing with R_3)	Totals
Ranking R_2	Pairs + (agreeing with R_3)	a	b	$a + b = \binom{n}{2} - Q_{23}$
	Pairs – (disagreeing with R_3)	c	d	$c + d = Q_{23}$
	Totals	$a + c = \binom{n}{2} - Q_{13}$	$b + d = Q_{13}$	$a + b + c + d = \binom{n}{2}$

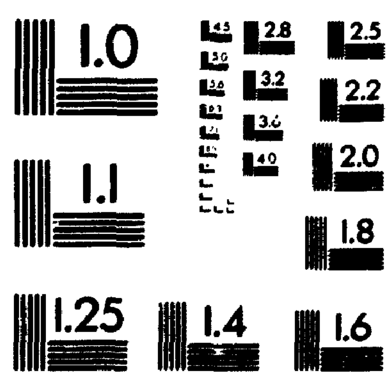
It is evident that $b+c = Q_{12}$ and therefore it follows that $d = (Q_{13} + Q_{23} - Q_{12})/2$, $c = (Q_{23} + Q_{12} - Q_{13})/2$, $b = (Q_{13} + Q_{12} - Q_{23})/2$ and $a = \binom{n}{2} - (Q_{13} + Q_{23} + Q_{12})/2$. Substituting into eqn. (4.16) then yields

$$t_{12.3} = \frac{\binom{n}{2}(Q_{13} + Q_{23} - Q_{12})/2 - Q_{13}Q_{23}}{\sqrt{Q_{13}Q_{23}\left(\binom{n}{2} - Q_{13}\right)\left(\binom{n}{2} - Q_{23}\right)}} \tag{4.17}$$

where

$$Q_{12} \geq \text{abs}(Q_{13} - Q_{23}) \tag{4.18a}$$

2 of/de 2



$$Q_{12} \le Q_{13} + Q_{23},\ 2\binom{n}{2} - (Q_{13} + Q_{23}) \tag{4.18b}$$

$$Q_{13} \ne 0, \binom{n}{2} \quad \text{and} \quad Q_{23} \ne 0, \binom{n}{2} . \tag{4.18c}$$

If $Q_{12} = 0$ then eqn. (4.18a) gives $Q_{13} = Q_{23}$ which, upon substitution into eqn. (4.17), shows that $(t_{12.3}|Q_{12} = 0) = 1$. Also, if $Q_{12} = \binom{n}{2}$ then eqn. (4.18b) gives $Q_{13} + Q_{23} = \binom{n}{2}$ which, upon substitution into eqn. (4.17), shows that $\left(t_{12.3}|Q_{12} = \binom{n}{2}\right) = -1$. These two situations correspond, respectively, to the cases $b = c = 0$ and $a = d = 0$.

Since $\mathcal{E}(t_{12.3}) = 0$ it follows that $\mathcal{V}ar(t_{12.3}) = \mathcal{E}(t_{12.3}^2) = \mathcal{E}(U/V)$ where $U = Q_{13}^2Q_{23}^2 - \binom{n}{2}(Q_{13}^2Q_{23} + Q_{13}Q_{23}^2 - Q_{12}Q_{13}Q_{23}) + \binom{n}{2}^2(Q_{13}^2 + Q_{23}^2 + Q_{12}^2 + 2Q_{13}Q_{23} - 2Q_{12}Q_{13} - 2Q_{12}Q_{23})/4$, and $V = Q_{13}Q_{23}\left(\binom{n}{2} - Q_{13}\right)\left(\binom{n}{2} - Q_{23}\right)$. It is proposed that $\mathcal{E}(t_{12.3}^2)$ be approximated by $\mathcal{E}(U)/\mathcal{E}(V)$, an approximation which is often valid for moderately large n. The difficulty in evaluating $\mathcal{E}(U)/\mathcal{E}(V)$ lies in evaluation of the term $\mathcal{E}(Q_{12}Q_{13}Q_{23})$ since only Q_{13} and Q_{23} are independent, the value of Q_{12} then depending on the particular ranking configurations which resulted in the values of Q_{13} and Q_{23}. To address this problem, the concept of a relative inversion vector, $I_n^{12} = (i_{12,1}, i_{12,2}, \ldots, i_{12,n})$ for two rankings R_1 and R_2 in the presence of a third ranking R_3 which is in the natural order, is introduced.

4.3.1 Relative inversion vectors

With R_3 in the natural order, inversion vectors I_n^1 and I_n^2 corresponding to R_1 and R_2, respectively, are obtained. Let R_2 be temporarily rearranged into the natural order and R_1 be correspondingly rearranged. This new configuration of R_1 defines, uniquely, an inversion vector I_n^{12} which is referred to as the inversion vector of R_1 relative to R_2.

Lemma 4.1: Given a ranking $R_2 = (y_1, \ldots, y_n)$ and its associated inversion vector $I_n^2 = (i_{21}, i_{22}, \ldots, i_{2n})$ the following algorithm will rearrange $(y_1, \ldots, y_n)$

into ascending order.

1. Initialize $j = 0$.
2. Set $j = j + 1$. If $j > n$ exit. Otherwise go to step 3.
3. If $i_{2j} = 0$ go to 2. Otherwise move y_j to the left of, and adjacent to, $y_{j-i_{2j}}$. Rename the new sequence as $(y_1, \ldots, y_n)$. Go to step 2.

Proof: The proof is inductive. Suppose that at step 2 of the algorithm, with $j = p$, the corresponding inversion vector is $(0, 0, \ldots, 0, i_{2p}, i_{2,p+1}, \ldots, i_{2n})$. It follows that $y_1 < y_2 < \cdots < y_{p-1}$ and that $y_{p-1}, y_{p-2}, \ldots, y_{p-i_{2p}}$ are all greater than y_p while $y_{p-i_{2p}-1}, \ldots, y_1$ are all smaller than y_p. Therefore, after y_p is moved i_{2p} places to the left and the new sequence is renamed, as in step 3, it must hold that $y_1 < y_2 < \cdots < y_p$ and the associated inversion vector is then $(0, 0, \ldots, 0, i_{2,p+1}, \ldots, i_{2n})$. The induction is completed by noting that every inversion vector has $i_{2j} = 0$ for $j = 1$ and that when $i_{2j} = 0$; $j = 1, \ldots, n$ the corresponding ranking is in the natural order.

Since, whenever R_2 is rearranged into its natural order, R_1 is to be correspondingly rearranged, the algorithm of lemma 4.1 may be applied to the elements of R_1 to obtain its rearrangement. This yields

Lemma 4.2: Given two inversion vectors I_n^1 and I_n^2 the following algorithm produces a relative inversion vector I_n^{12}. Note that here, and subsequently, I_n^2 is used as a basis for modifying I_n^1 to produce I_n^{12}.

1. Initialize $j = 0$.
2. Set $j = j + 1$. If $j > n$ exit. Otherwise go to step 3.
3. If $i_{2j} = 0$ set $i_{12,j} = i_{1j}$ and go to step 2. Otherwise do

    3.1 Set $k = 0$

    3.2 Let

$$(i_{12,j-k-1}, i_{12,j-k}) = \begin{cases} (i_{1,j-k} - 1, i_{1,j-k-1}) & \text{if } i_{1,j-k} > i_{1,j-k-1} \\ (i_{1,j-k}, i_{1,j-k-1} + 1) & \text{otherwise} \end{cases}$$

3.3 Let $(i_{1,j-k-1}, i_{1,j-k}) = (i_{12,j-k-1}, i_{12,j-k})$

3.4 Set $k = k+1$

3.5 If $k = i_{2j}$ go to step 2. Otherwise go to step 3.2

Proof: Apply the algorithm of lemma 4.1 to the elements of R_1. Suppose that in step 3 of that algorithm $i_{2j} \neq 0$. Then x_j is moved to the left, a single move at a time, for a total of i_{2j} such moves. Consider the first such move whereby $k = 0$. It holds that

$$\begin{array}{ll} x_{j-1} > x_j & \text{if } i_{1j} > i_{1,j-1} \\ x_{j-1} < x_j & \text{otherwise} . \end{array} \tag{4.19}$$

Hence interchanging the positions of x_{j-1} and x_j and renaming them according to their new positions yields

$$\begin{array}{ll} x_{j-1} < x_j & \text{if } i_{1j} > i_{1,j-1} \\ x_{j-1} > x_j & \text{otherwise} . \end{array} \tag{4.20}$$

Consequently, there are now $i_{1j} - 1$ or i_{1j} greater elements to the left of x_{j-1} while there are now $i_{1,j-1}$ or $i_{1,j-1}+1$ greater elements to the left of x_j, according as $i_{1j} > i_{1,j-1}$ or $i_{1j} \leq i_{1,j-1}$. The same argument applies to each value of k; this result being embodied in step 3.2 of the current algorithm.

It follows, as before, that $Q_{12,n} = \sum_{\ell=1}^{n} i_{12,\ell}$ and therefore

$$\begin{aligned} \mathcal{E}(Q_{13}Q_{23}Q_{12}) &= \mathcal{E}\left\{\left(\sum_{j=1}^{n} i_{1j}\right)\left(\sum_{k=1}^{n} i_{2k}\right)\left(\sum_{\ell=1}^{n} i_{12,\ell}\right)\right\} \\ &= \sum_{j=1}^{n}\sum_{k\neq j}^{n} \mathcal{E}\left\{i_{1j}i_{2k}\sum_{\ell=1}^{n}\mathcal{E}(i_{12,\ell}|i_{1j},i_{2k})\right\} + \\ &\quad \sum_{j=1}^{n}\mathcal{E}\left\{i_{1j}i_{2j}\mathcal{E}(Q_{12}|i_{1j},i_{2j})\right\} . \end{aligned} \tag{4.21}$$

4.3.2 Evaluation of $\mathcal{E}(i_{12,\ell}|i_{1j}, i_{2k})$, $j \neq k$

Two lemmas and a theorem pertaining to $\mathcal{E}(i_{12,\ell}|i_{1j}, i_{2k})$, $j \neq k$ are now stated and proven.

Lemma 4.3: Let $\{I_n^1\}$ be the entire $n!$ set of I_n^1 vectors and let I_n^2 be any arbitrarily chosen inversion vector. Then the $n!$ set of I_n^{12} vectors obtained by applying the algorithm of lemma 4.2 to $\{I_n^1\}$ and I_n^2 is identical to the set $\{I_n^1\}$.

Proof: Since I_n^2 is constant each time the algorithm is applied to a member of $\{I_n^1\}$ it follows that the rearrangements of two non-identical I_n^1 vectors must lead to two non-identical I_n^{12} vectors. Hence an $n!$ set of distinct I_n^1 vectors leads to an $n!$ set of distinct I_n^{12} vectors. The result follows since there are exactly $n!$ distinct inversion vectors of size n.

Corollary: Let $\{I_n^1|i_{1j}\}$ be an $n!/j$ set of I_n^1 vectors with i_{1j} fixed and let I_n^2 be arbitrary. Then the $n!/j$ set of intermediate inversion vectors obtained, by applying the algorithm of lemma 4.2 to $\{I_n^1|i_{1j}\}$ and I_n^2 up to and including iteration $j-1$ of steps 2 and 3, is identical to the set $\{I_n^1|i_{1j}\}$. Hence the first $j-1$ elements of I_n^2 may be replaced by zero prior to applying the algorithm.

Proof: The set $\{I_n^1|i_{1j}\}$ may be divided into $(j+1)(j+2)\cdots(n-1)n$ subsets each of size $(j-1)!$ so that, between any two vectors belonging to the same subset, the first $(j-1)$ pairs of elements contain at least one distinct pair while the remaining pairs of elements are all pairwise identical. Applying lemma 4.3 to each of these subsets, with $j-1$ replacing the n of the lemma, yields the result.

Lemma 4.4: Let $\{I_n^1|i_{1n}\}$ be an $(n-1)!$ set of I_n^1 vectors with i_{1n} fixed and let $\{I_n^2|i_{2\ell}=0;\ell=1,\ldots,n-1\}$ be a set $\{(0,\ldots,0,0),(0,\ldots,0,1),\ldots,(0,\ldots,0,n-1)\}$ of I_n^2 vectors. Let $\{I_n^{12}\}$ be the $n!$ set of inversion vectors formed by applying the algorithm of lemma 4.2 to $\{I_n^1|i_{1n}\}$ and $\{I_n^2|i_{2\ell}=0;\ell=1,\ldots,n-1\}$. Then $\{I_n^{12}\}$ is identical to the $n!$ set of distinct I_n^1 vectors.

Proof: Consider application of the algorithm to the set $\{I_n^1|i_{1n}\}$ and a fixed member of $\{I_n^2|i_{2\ell}=0;\ell=1,\ldots,n-1\}$. As in the proof of lemma 4.3, the resulting set of I_n^{12} vectors is a distinct set. Next consider the permutations corresponding to

$\{I_n^1|i_{1n}\}$. Each of these permutations has the value $n - i_{1n}$ for its last element. It therefore follows that applying the algorithm to a fixed member of $\{I_n^1|i_{1n}\}$ and the set $\{I_n^2|i_{2\ell}=0; \ell=1,\ldots,n-1\}$ will yield n distinct I_n^{12} vectors. Combining these two cases shows that all the $n!$ relative inversion vectors are distinct as required.

Corollary: Let $\{I_n^1|i_{1j}\}$ be an $n!/j$ set of I_n^1 vectors with i_{1j} fixed and let $\{I_n^2|i_{2\ell}=0; \ell=1,\ldots,j-1\}$ be an $n!/(j-1)!$ set of I_n^2 vectors. Then the $n!^2/j!$ set of intermediate vectors obtained, by applying the algorithm of lemma 4.2 to the given sets up to and including iteration j of steps 2 and 3, is comprised of $n!/j!$ sets of $\{I_n^1\}$ vectors.

Proof: The set $\{I_n^2|i_{2\ell}=0; \ell=1,\ldots,j-1\}$ may be divided into $n!/j!$ subsets each of size j, so that i_{2j} assumes each of the values $0, 1, \ldots, j-1$ in each subset. Applying lemma 4.4 to $\{I_n^1|i_{1j}\}$ and each of these subsets, with j replacing the n of the lemma, yields the result.

Theorem 4.1: $\mathcal{E}(i_{12,\ell}|i_{1j}, i_{2k})$, $j \neq k$ is independent of i_{1j} and of i_{2k}. Hence $\mathcal{E}(i_{12,\ell}|i_{1j}, i_{2k}) = (\ell-1)/2$.

Proof: Consider the set of all I_n^1 and I_n^2 vectors where i_{1j} and i_{2k} are fixed. Since the members of $\{I_n^1|i_{1j}\}$ are equiprobable and the members of $\{I_n^2|i_{2k}\}$ are also equiprobable then the members of the resulting set of I^{12} vectors are equiprobable. It thus suffices to show that this set of I^{12} vectors is a multiple of the $n!$ set of distinct I_n inversion vectors.

Apply the corollary, lemma 4.3, to obtain the original set of I_n^1 vectors and a set of I_n^2 vectors whose first $j-1$ elements are all zero. Next apply the corollary, lemma 4.4, to obtain a set $\{I_n^1\}$ of $n!$ distinct I_n^1 vectors as intermediates in conjunction with a set of I_n^2 vectors whose first j elements are all zero. Finally, apply lemma 4.3 to obtain the desired result.

4.3.3 Evaluation of $\mathcal{E}(Q_{12}|i_{1j}, i_{2j})$

Four lemmas and a theorem pertaining to $\mathcal{E}(Q_{12}|i_{1j}, i_{2j})$ are now stated and proven.

Lemma 4.5: $\mathcal{E}(Q_{13}|i_{1j}) = \binom{n}{2}/2 + i_{1j} - (j-1)/2$.

Proof: By definition $Q_{13} = \sum_{\ell=1}^{n} i_{1\ell}$ so that

$$\mathcal{E}(Q_{13}|i_{1j}) = \sum_{\ell=1}^{n} \mathcal{E}(i_{1\ell}) + i_{1j} - \mathcal{E}(i_{1j})$$
$$= \frac{1}{2}\binom{n}{2} + i_{1j} - (j-1)/2 \ . \tag{4.22}$$

Lemma 4.6: Let $\{I_n^1|i_{1j}\}$ be an $n!/j$ set of I_n^1 vectors with i_{1j} fixed. Let $I_n^2 = (0, \ldots, 0, i_{2j}, 0, \ldots, 0)$ be an inversion vector with every element other than the j^{th} having a zero value. Then

$$\mathcal{E}(Q_{12}|i_{1j}, I_n^2) = \mathcal{E}(Q_{13}|i_{1j}) - \left(\frac{2i_{1j} - (j-1)}{j-1}\right) i_{2j} \ . \tag{4.23}$$

Proof: Consider $\{P_n^1|i_{1j}\}$ the set of permutations corresponding to $\{I_n^1|i_{1j}\}$. Apply the algorithm of lemma 4.1 to $\{P_n^1|i_{1j}\}$ and I_n^2 to obtain $\{P_n^{12}|i_{1j}, I_n^2\}$ the set of permutations corresponding to $\{I_n^{12}|i_{1j}, I_n^2\}$, the resulting set of relative inversion vectors. Let $(\Delta Q|i_{1j}, I_n^2)$ be the decrease in the negative score associated with changing $P_n^1|i_{1j}$ to $P_n^{12}|i_{1j}, I_n^2$. It follows that

$$Q_{12}|i_{1j}, I_n^2 = Q_{13}|i_{1j} - \Delta Q|i_{1j}, I_n^2 \tag{4.24}$$

and therefore that

$$\mathcal{E}(Q_{12}|i_{1j}, I_n^2) = \mathcal{E}(Q_{13}|i_{1j}) - \mathcal{E}(\Delta Q|i_{1j}, I_n^2) \ . \tag{4.25}$$

Let $(x_1, \ldots, x_n)$ be one of the permutations belonging to $\{P_n^1|i_{1j}\}$. Of the $j-1$ elements preceding x_j there are, therefore, i_{1j} larger and $j-1-i_{1j}$ smaller elements than x_j. Each of these elements occupies any one of the first $j-1$

positions with probability $\frac{1}{j-1}$. Hence if x_j is moved one position to its left the expected decrease in the negative score is $(i_{1j} - (j-1-i_{1j}))/(j-1)$ since each time x_j passes a larger element the negative score is decreased by one while each time x_j passes a smaller element the negative score is increased by one. With i_{2j} moves of x_j to the left it follows that

$$\mathcal{E}(\Delta Q | i_{1j}, I_n^2) = \left(\frac{i_{1j} - (j-1-i_{1j})}{j-1}\right) i_{2j} \tag{4.26}$$

from which the result follows.

Lemma 4.7: Let $\{I_n^{12} | i_{1j}, I_n^2\}$ be the set of relative inversion vectors obtained in lemma 4.6. Let $\{I_n^2 | i_{2\ell} = 0, \ell \neq j+1; i_{2,j+1} = 0, 1, \ldots, n\}$ be a set of $(j+1)$ I_n^2 vectors. The expected value of the additional decrease in the negative score resulting from applying the algorithm of lemma 1 to $\{I_n^{12} | i_{1j}, I_n^2\}$ and $\{I_n^2 | i_{2\ell} = 0, \ell \neq j + 1; i_{2,j+1} = 0, 1, \ldots, n\}$ is

$$\mathcal{E}(\Delta Q_{\text{additional}}) = \frac{j}{j-1}\left(\frac{1}{2} - \frac{j - i_{2j}}{j+1}\right)\left(1 - \frac{2(j - i_{1j})}{j+1}\right) . \tag{4.27}$$

Proof: Consider the permutation $(x_1, \ldots, x_n)$ of lemma 4.6 and the permutation $(v_1, \ldots, v_j, x_{j+1}, \ldots, x_n)$ obtained by moving x_j to the left through i_{2j} positions, as in lemma 6. Now move x_{j+1}. There are four situations to consider, viz: (i) $x_{j+1} > x_j$ or $i_{1,j+1} \leq i_{1j}$, (ii) $x_{j+1} < x_j$ or $i_{1,j+1} > i_{1j}$, (iii) $i_{2,j+1} \leq i_{2j}$ and (iv) $i_{2,j+1} > i_{2j}$. These four situations combine to yield the four mutually exclusive events (i) and (iii), (i) and (iv), (ii) and (iii) and finally, (ii) and (iv). Let $(\Delta Q | i_{1,j+1}, i_{2,j+1})$ be the additional decrease in the negative score due to moving x_{j+1} in the permutation $(v_1, \ldots, v_j, x_{j+1}, \ldots, x_n)$. It then follows that

Case: (i) and (iii)

Since $i_{2,j+1} \leq i_{2j}$ then x_{j+1} lies to the right of x_j after both elements are moved. There are $i_{1,j+1}$ larger and $(j - i_{1,j+1} - 1)$ smaller elements, excluding

x_j, on the left of x_{j+1} since $x_{j+1} > x_j$. Consequently, the expected decrease in the negative score is

$$\mathcal{E}(\Delta Q | i_{1,j+1}, i_{2,j+1}) = \left(\frac{i_{1,j+1} - (j - i_{1,j+1} - 1)}{j-1} \right) i_{2,j+1} . \qquad (4.28)$$

Case: (i) and (iv)

Since $i_{2,j+1} > i_{2j}$ then x_{j+1} lies to the left of x_j after both elements are moved. Consequently, the expected decrease in the negative score is

$$\mathcal{E}(\Delta Q | i_{1,j+1}, i_{2,j+1}) = \left(\frac{i_{1,j+1} - (j - i_{1,j+1} - 1)}{j-1} \right) (i_{2,j+1} - 1) - 1 . \qquad (4.29)$$

Case: (ii) and (iii)

Since $x_{j+1} < x_j$ there are $i_{1,j+1} - 1$ larger elements and $j - i_{1,j+1}$ smaller elements, excluding x_j, to the left of x_{j+1}. Analogously to eqn. (4.28), it follows that

$$\mathcal{E}(\Delta Q | i_{1,j+1}, i_{2,j+1}) = \left(\frac{i_{1,j+1} - 1 - (j - i_{1,j+1})}{j-1} \right) i_{2,j+1} . \qquad (4.30)$$

Case: (ii) and (iv)

It is readily seen that

$$\mathcal{E}(\Delta Q | i_{1,j+1}, i_{2,j+1}) = \left(\frac{i_{1,j+1} - 1 - (j - i_{1,j+1})}{j-1} \right) (i_{2,j+1} - 1) + 1 . \qquad (4.31)$$

There are $j+1$ values of $i_{1,j+1}$ and $j+1$ values of $i_{2,j+1}$ each with equal probability of occurring. Therefore it follows that

$$\mathcal{E}(\Delta Q_{\text{additional}}) = \frac{1}{(j+1)^2} \Bigg[\sum_{i_{1,j+1}=0}^{i_{1j}} \Bigg\{ \sum_{i_{2,j+1}=0}^{i_{2j}} \left(\frac{2i_{1,j+1} - (j-1)}{j-1} \right) i_{2,j+1} +$$
$$\sum_{i_{2,j+1}=i_{2j}+1}^{j} \left(\left(\frac{2i_{1,j+1} - (j-1)}{j-1} \right) (i_{2,j+1} - 1) - 1 \right) \Bigg\} +$$
$$\sum_{i_{1,j+1}=i_{1j}+1}^{j} \Bigg\{ \sum_{i_{2,j+1}=0}^{i_{2j}} \left(\frac{2i_{1,j+1} - (j+1)}{j-1} \right) i_{2,j+1} +$$

$$\sum_{i_{2,j+1}=i_{2j}+1}^{j} \left(\left(\frac{2i_{1,j+1}-(j+1)}{j-1} \right) (i_{2,j+1}-1)+1 \right) \Bigg\} \Bigg]$$

$$= \frac{1}{(j+1)^2} \Bigg[\sum_{i_{1,j+1}=0}^{i_{1j}} \Bigg\{ \sum_{i_{2,j+1}=0}^{j} \left(\frac{2i_{1,j+1}-(j-1)}{j-1} \right) i_{2,j+1} -$$

$$(j-i_{2j}) \left(\frac{2i_{1,j+1}}{j-1} \right) \Bigg\} +$$

$$\sum_{i_{1,j+1}=i_{1j}+1}^{j} \Bigg\{ \sum_{i_{2,j+1}=0}^{j} \left(\frac{2i_{1,j+1}-(j-1)-2}{j-1} \right) i_{2,j+1} -$$

$$(j-i_{2j}) \left(\frac{2i_{1,j+1}-2j}{j-1} \right) \Bigg\} \Bigg]$$

$$= \frac{1}{(j+1)^2} \Bigg[\sum_{i_{1,j+1}=0}^{j} \Bigg\{ \left(\frac{2i_{1,j+1}-(j-1)}{j-1} \right) \left(\frac{j(j+1)}{2} \right) -$$

$$(j-i_{2j}) \left(\frac{2i_{1,j+1}}{j-1} \right) \Bigg\} + (j-i_{1j}) \Bigg\{ \frac{2j(j-i_{2j})}{j-1} - \frac{j(j+1)}{j-1} \Bigg\} \Bigg]$$

$$= \frac{j}{(j-1)(j+1)^2} \Bigg[\frac{(j+1)^2}{2} \Bigg\{ 1 - \frac{2(j-i_{2j})}{j+1} \Bigg\} -$$

$$(j-i_{1j})(j+1) \Bigg\{ 1 - \frac{2(j-i_{2j})}{j+1} \Bigg\} \Bigg]$$

$$= \frac{j}{2(j-1)} \left(1 - \frac{2(j-i_{2j})}{j+1} \right) \left(1 - \frac{2(j-i_{1j})}{j+1} \right) \tag{4.32}$$

which is the result stated in eqn. (4.27).

Lemma 4.8: The expected value of the additional decrease in the negative score is constant with each application of the algorithm of lemma 4.1 for iterations $j+1, j+2, \ldots, n$ of steps 2 and 3.

Proof: Consider the results for steps $j+\ell$ and $j+\ell+1$. From the symmetrical nature of permutations it follows that the set of permutations of $R_1|i_{1j}$ may be divided into two equal subsets typified by members $(x_{1s_1}, x_{2s_1}, \ldots, x_{ns_1})$ and $(x_{1s_2}, x_{2s_2}, \ldots, x_{ns_2})$ such that $x_{j+\ell,s_1} = x_{j+\ell+1,s_2}$ and $x_{j+\ell,s_2} = x_{j+\ell+1,s_1}$. Let $i_{2,j+\ell}$ be fixed and equal to q.

For $i_{2,j+\ell+1} \leq q$ the change in score due to moving $x_{j+\ell+1,s_1}$ through

$i_{2,j+\ell+1}$ steps to the left is the same as that previously obtained by moving $x_{j+\ell,s_2}$ through $i_{2,j+\ell+1}$ steps to the left. For $i_{2,j+\ell+1} > q$ the change in score due to moving $x_{j+\ell+1,s_1}$ through $i_{2,j+\ell+1}$ steps to the left is the same as that previously obtained by moving $x_{j+\ell,s_2}$ through $i_{2,j+\ell+1} - 1$ steps to the left since movements of $x_{j+\ell+1}$ across $x_{j+\ell}$ result in a net change of zero as is evident from the division into two subsets as specified above. It follows that

$$\begin{aligned}(\Delta Q | i_{2,j+\ell+1} = \{0,1,\ldots,j+\ell\}, i_{2,j+\ell} = q) &= (\Delta Q | i_{2,j+\ell} = \{0,1,\ldots,j+\ell-1\}) \\ &\quad + (\Delta Q | i_{2,j+\ell} = q)\,. \qquad (4.33)\end{aligned}$$

Now

$$\begin{aligned}\mathcal{E}(\Delta Q_{\text{iteration } j+\ell+1}) &= \mathcal{E}_{i_{2,j+\ell}}\left(\mathcal{E}(\Delta Q_{\text{iteration } j+\ell+1} | i_{2,j+\ell})\right) \\ &= \mathcal{E}_{i_{2,j+\ell}}\left(\frac{\Delta Q | i_{2,j+\ell+1} = \{0,1,\ldots,j+\ell\}, i_{2,j+\ell}}{j+\ell+1}\right) \qquad (4.34)\end{aligned}$$

and

$$\mathcal{E}(\Delta Q_{\text{iteration } j+\ell}) = \frac{\Delta Q | i_{2,j+\ell} = \{0,1,\ldots,j+\ell-1\}}{j+\ell} \qquad (4.35)$$

so that

$$\begin{aligned}\mathcal{E}(\Delta Q_{\text{iteration } j+\ell+1}) &= \mathcal{E}_{i_{2,j+\ell}}\left\{\frac{(j+\ell)\mathcal{E}(\Delta Q_{\text{iteration } j+\ell})}{j+\ell+1} + \frac{\Delta Q | i_{2,j+\ell}}{j+\ell+1}\right\} \\ &= \mathcal{E}(\Delta Q_{\text{iteration } j+\ell}) \qquad (4.36)\end{aligned}$$

since $\mathcal{E}_{i_{2,j+\ell}}(\Delta Q | i_{2,j+\ell}) = \mathcal{E}(\Delta Q_{\text{iteration } j+\ell})$.

Theorem 4.2: The expected value of the negative score given i_{1j}, i_{2j} is

$$\begin{aligned}\mathcal{E}(Q_{12} | i_{1j}, i_{2j}) &= \frac{1}{2}\binom{n}{2} + i_{1j} - (j-1)/2 - \left(\frac{2i_{1j} - (j-1)}{j-1}\right) i_{2j} \\ &\quad - \frac{(n-j)j}{2(j-1)}\left(1 - \frac{2(j-i_{1j})}{j+1}\right)\left(1 - \frac{2(j-i_{2j})}{j+1}\right). \qquad (4.37)\end{aligned}$$

Proof: The result follows from the corollary, lemma 4.3, used in conjunction with lemmas 4.5, 4.6, 4.7 and 4.8.

4.3.4 Evaluation of $\mathcal{E}(Q_{12}Q_{13}Q_{23})$

Substituting the results from theorems 4.1 and 4.2 into eqn. (4.21) yields

$$\mathcal{E}(Q_{12}Q_{13}Q_{23}) = \left(\frac{1}{2}\binom{n}{2}\right)^3 + \sum_{j=2}^{n}\left[\frac{-2}{j-1}\left\{\mathcal{E}(i_{1j}^2 - i_{1j}(j-1)/2)\right\}^2 - \frac{j(n-j)}{2(j-1)}\left\{\mathcal{E}\left(i_{1j} - \frac{2(ji_{1j} - i_{1j}^2)}{j+1}\right)\right\}^2\right]$$

$$= \left(\frac{1}{2}\binom{n}{2}\right)^3 + \sum_{j=2}^{n}\left[\frac{-2}{j-1}\left\{\frac{(j-1)(2j-1)}{6} - \left(\frac{j-1}{2}\right)^2\right\}^2 - \frac{j(n-j)}{2(j-1)}\left\{(j-1)/2 - \frac{j(j-1)-(j-1)(2j-1)/3}{j+1}\right\}^2\right]$$

$$= \left(\frac{1}{2}\binom{n}{2}\right)^3 - \frac{1}{72}\sum_{j=2}^{n}\left\{(j-1)(j+1)^2 + (j-1)j(n-j)\right\}$$

$$= \left(\frac{1}{2}\binom{n}{2}\right)^3 - \frac{1}{216}(2n^2+6n+7)\binom{n}{2}\ . \tag{4.38}$$

Under the hypothesis of equiprobable rankings, it follows that Q_{12}, Q_{13} and Q_{23} are pairwise independent and therefore $\mathcal{E}(Q_{12}Q_{13}) = \mathcal{E}(Q_{12}Q_{23}) = (\mathcal{E}(Q_{23}))^2$.

It is now possible to evaluate $\mathcal{E}(U)/\mathcal{E}(V)$. However, it is necessary to correct for the constraints, eqn. (4.18c), placed on Q_{13} and on Q_{23}.

4.3.5 Correction of expectations for the constraints on Q_{13} and on Q_{23}

A prime is used to denote a constrained random variable. Since Q_{12} is implicitly constrained it is not primed on the right side of equations. Now,

$$\mathcal{E}(Q'_{13}) = \sum_{Q'_{13}} Q'_{13}\,\Pr(Q'_{13})$$

$$= \sum_{Q_{13}} Q_{13}\left(\frac{n!}{n!-2}\right)\Pr(Q_{13}) - \binom{n}{2}\left(\frac{n!}{n!-2}\right)\Pr\left(Q_{13} = \binom{n}{2}\right)$$

$$= \left(\frac{n!}{n!-2}\right)\mathcal{E}(Q_{13}) - \binom{n}{2}\left(\frac{1}{n!-2}\right) \tag{4.39a}$$

since there are $n!$ values of Q_{13} but only $n!-2$ values of Q'_{13}. Similarly,

$$\mathcal{E}({Q'_{13}}^2) = \left(\frac{n!}{n!-2}\right)\mathcal{E}(Q_{13}^2) - \binom{n}{2}^2\left(\frac{1}{n!-2}\right)\ . \tag{4.39b}$$

Substituting for $\mathcal{E}(Q_{13})$ and $\mathcal{E}(Q_{13}^2)$ in eqns. (4.39a) and (4.39b) yields

$$\mathcal{E}(Q'_{13}) = \mathcal{E}(Q_{13}) \tag{4.40a}$$

and

$$\mathcal{E}({Q'_{13}}^2) = \mathcal{E}(Q_{13}^2) - \frac{\frac{1}{2}\binom{n}{2}^2}{(n!-2)} + \frac{2\binom{n}{2}(2n+5)}{36(n!-2)} \tag{4.40b}$$

since $\mathcal{E}(Q_{13}) = \frac{1}{2}\binom{n}{2}$ and $\mathcal{E}(Q_{13}^2) = \frac{1}{4}\binom{n}{2}^2 + \binom{n}{2}\left(\frac{2n+5}{36}\right)$. Next,

$$\begin{aligned}
\mathcal{E}(Q'_{12}Q'_{13}) &= \sum_{Q'_{13}}\sum_{Q_{12}} Q'_{13}Q_{12}\Pr(Q'_{13})\Pr(Q_{12}|Q'_{13}) \\
&= \sum_{Q_{13}}\sum_{Q_{12}} Q_{13}Q_{12}\left(\frac{n!}{n!-2}\right)\Pr(Q_{13})\Pr(Q_{12}|Q_{13}) - \\
&\quad \sum_{Q_{12}}\binom{n}{2}Q_{12}\left(\frac{n!}{n!-2}\right)\Pr\left(Q_{13}=\binom{n}{2}\right)\Pr\left(Q_{12}|Q_{13}=\binom{n}{2}\right) \\
&= \left(\frac{n!}{n!-2}\right)\mathcal{E}(Q_{12}Q_{13}) - \binom{n}{2}\left(\frac{1}{n!-2}\right)\mathcal{E}(Q_{12}) \\
&= \left(\frac{1}{2}\binom{n}{2}\right)^2
\end{aligned} \tag{4.41}$$

since $\Pr(Q_{12}|Q'_{13}) = \Pr(Q_{12}|Q_{13})$ and $\Pr\left(Q_{12}|Q_{13}=\binom{n}{2}\right) = \Pr(Q_{12})$. Similarly,

$$\begin{aligned}
\mathcal{E}({Q'_{12}}^2) &= \sum_{Q'_{13}}\sum_{Q'_{23}} Q_{12}^2\Pr(Q'_{13},Q'_{23}) \\
&= \left(\frac{n!}{n!-2}\right)^2\Bigg[\sum_{Q_{13}}\sum_{Q_{23}} Q_{12}^2\Pr(Q_{13},Q_{23}) - \\
&\quad 2\sum_{Q_{13}} Q_{12}^2\left\{\Pr(Q_{13},Q_{23}=0) + \Pr\left(Q_{13},Q_{23}=\binom{n}{2}\right)\right\} + \\
&\quad Q_{12}^2\left\{\Pr(0,0) + \Pr\left(\binom{n}{2},\binom{n}{2}\right) + 2\Pr\left(0,\binom{n}{2}\right)\right\}\Bigg] \\
&= \left(\frac{n!}{n!-2}\right)^2\Bigg[\mathcal{E}(Q_{12}^2) - 2\sum_{Q_{13}}\left\{Q_{13}^2 + \left(\binom{n}{2} - Q_{13}\right)^2\right\}\frac{\Pr(Q_{13})}{n!} + \\
&\quad \frac{2}{n!^2}\binom{n}{2}^2\Bigg]
\end{aligned}$$

$$= \left(\frac{n!}{n!-2}\right)^2 \left[\mathcal{E}(Q_{12}^2) - \frac{4}{n!}\mathcal{E}(Q_{13}^2) + \frac{2}{n!^2}\binom{n}{2}^2\right]$$

$$= \mathcal{E}(Q_{12}^2) + \frac{\binom{n}{2}^2}{(n!-2)^2} - \frac{4\binom{n}{2}(2n+5)}{36(n!-2)^2} \tag{4.42}$$

since $(Q_{12}|Q_{23}=0) = Q_{13}$ and $(Q_{12}|Q_{23}=\binom{n}{2}) = \binom{n}{2} - Q_{13}$. Finally,

$$\begin{aligned}\mathcal{E}(Q'_{12}Q'_{13}Q'_{23}) &= \sum_{Q'_{13}}\sum_{Q'_{23}} Q_{12}Q'_{13}Q'_{23}\,\Pr(Q'_{13},Q'_{23}) \\ &= \left(\frac{n!}{n!-2}\right)^2 \left[\sum_{Q_{13}}\sum_{Q_{23}} Q_{12}Q_{13}Q_{23}\,\Pr(Q_{13},Q_{23}) - \right. \\ &\quad \left. 2\sum_{Q_{13}} Q_{13}\left(\binom{n}{2} - Q_{13}\right)\binom{n}{2}\frac{\Pr(Q_{13})}{n!}\right] \\ &= \left(\frac{n!}{n!-2}\right)^2 \mathcal{E}(Q_{12}Q_{13}Q_{23}) \\ &\quad - 2\binom{n}{2}\frac{n!}{(n!-2)^2}\left\{\binom{n}{2}\mathcal{E}(Q_{13}) - \mathcal{E}(Q_{13}^2)\right\} .\end{aligned} \tag{4.43}$$

4.3.6 Ratio(expectations): corrected for constraints

Since Q'_{13} is independent of Q'_{23} it follows that

$$\mathcal{E}\left\{Q'_{13}Q'_{23}\left(\binom{n}{2} - Q'_{13}\right)\left(\binom{n}{2} - Q'_{23}\right)\right\} = \left\{\binom{n}{2}\mathcal{E}(Q'_{13})\right\}^2 + (\mathcal{E}(Q'^2_{13}))^2 - 2\binom{n}{2}\mathcal{E}(Q'_{13})\mathcal{E}(Q'^2_{13}) . \tag{4.44}$$

Substituting from eqns. (4.40a) and (4.40b) into eqn. (4.44) and reducing then gives

$$\mathcal{E}(V) = \frac{\binom{n}{2}^2}{36^2}\left\{81\binom{n}{2}^2 - 18\binom{n}{2}(2n+5) + (2n+5)^2\right\}\left(1 + \frac{4}{n!-2} + \frac{4}{(n!-2)^2}\right) \tag{4.45}$$

where the factor $\left(1 + \frac{4}{n!-2} + \frac{4}{(n!-2)^2}\right)$ represents the correction factor for the constraints on Q_{13} and on Q_{23} . Next, since Q'_{12} is independent of Q'_{23} and it has been shown that $\mathcal{E}(Q'_{12}Q'_{13}) = \left(\mathcal{E}(Q'_{13})\right)^2$, it follows that

$$\mathcal{E}(U) = \frac{1}{4}\binom{n}{2}^2 \mathcal{E}(2Q'^{\,2}_{13} + Q'^{\,2}_{12} - 2Q'_{13}Q'_{23}) + \left(\mathcal{E}(Q'^{\,2}_{13})\right)^2 + \binom{n}{2}\mathcal{E}(Q'_{12}Q'_{13}Q'_{23}) - 2\binom{n}{2}\mathcal{E}(Q'_{13})\mathcal{E}(Q'^{\,2}_{13}) \,. \tag{4.46}$$

Substituting from eqns. (4.40a) through (4.43) into eqn. (4.46) and reducing then yields

$$\mathcal{E}(U) = \frac{\binom{n}{2}^2}{36^2}\left\{9\binom{n}{2}(2n+5) + (2n+5)^2 - 6(2n^2+6n+7)\right\} \left(1 + \frac{4}{n!-2} + \frac{4}{(n!-2)^2}\right) \,. \tag{4.47}$$

Taking the ratio of the RHS of eqn (4.47) to the RHS of eqn. (4.45) then gives an approximation to $\mathcal{V}ar(t_{12.3})$ as

$$\mathcal{V}ar(t_{12.3}) = \frac{(2n+5)}{\left\{9\binom{n}{2} - (2n+5)\right\}} - \frac{4(n-2)(n+1)}{\left\{9\binom{n}{2} - (2n+5)\right\}^2} \tag{4.48}$$

where the correction factor for the constraints on Q_{13} and on Q_{23} has cancelled out. The variance est.. ıtor given by eqn. (4.48) is subsequently referred to as the ratio variance estimator.

For large n, it follows from eqn. (4.48) that

$$\left(\mathcal{V}ar(t_{12.3})\right)^{-1} = \frac{9\binom{n}{2}}{2n+5} - 1 \,. \tag{4.49}$$

Moran (1951) speculated that, approximately,

$$\left(\mathcal{V}ar(t_{12.3})\right)^{-1} = \frac{9\binom{n}{2}}{2n+5} \tag{4.50}$$

which gives the same asymptotic estimate of $\mathcal{V}ar(\sqrt{n}t_{12.3})$ as the expression in eqn. (4.48). Under the hypothesis of equiprobable rankings, Hoeffding's (1948) work shows that the result, eqn. (4.5:'), is valid. This variance estimator is subsequently referred tc as the pairwise τ variance estimator.

4.4 Upper and lower bounds for the variance of $t_{12.3}$

Table 4.6 shows the variance of $t_{12.3}$ and its two estimators, eqns. (4.48) and (4.50), for $n = 3, \dots, 7$. It is interesting to observe that, in all five cases, the true variance lies between the two estimators. This suggests that the two estimators provide an upper and a lower bound for $\mathcal{V}ar(t_{12.3})$ and that a measure of the maximum error obtained by using one-half the sum of the two estimators is given by one-half the interval width between the estimators. The maximum error of approximation then declines rapidly to 0.3% at $n = 20$ and to 0.05% at $n = 50$. Theorem 4.3 provides a formal proof of the suggested result.

Table 4.6: Variance of $t_{12.3}$ for $n = 3, \dots, 7$

n	Exact	Ratio [1]	Ratio [2]	Pairwise τ [3]	Mean [4]
3	0.6250	0.6250	0.6250	0.4074	0.5162
4	0.2829	0.2933	0.2933	0.2407	0.2670
5	0.1799	0.1872	0.1872	0.1667	0.1765
6	0.1318	0.1360	0.1360	0.1259	0.1310
7	0.1038	0.1062	0.1062	0.1005	0.1034

[1] $\mathcal{E}(U)/\mathcal{E}(V)$ obtained by enumeration
[2] $\mathcal{E}(U)/\mathcal{E}(V)$ obtained from eqn. (4.48)
[3] Estimator obtained from eqn. (4.50)
[4] Mean of ratio and pairwise τ estimators

Theorem 4.3: The ratio and the pairwise τ variance estimators provide upper and lower bounds, respectively, for $\mathcal{V}ar(t_{12.3})$. Furthermore, the error obtained from using either estimator is $O(n^{-3})$.

Proof: From the properties of a geometric series, it follows that

$$\begin{aligned}\frac{U}{V} &= (t_{12}^2 - 2t_{12}t_{13}t_{23} + t_{13}^2t_{23}^2)\Big(\sum_{k=0}^{\infty} t_{13}^{2k}\Big)\Big(\sum_{k=0}^{\infty} t_{23}^{2k}\Big) \\ &= (t_{12}^2 - 2t_{12}t_{13}t_{23} + t_{13}^2t_{23}^2)(1 + t_{13}^2 + t_{23}^2 + t_{13}^2t_{23}^2 + t_{13}^4 + t_{23}^4 + \cdots). \quad (4.51)\end{aligned}$$

Consequently,

$$\mathcal{E}(U/V) = \mu_2 + 3\mu_2^2 + 4\mu_2\mu_4 - 2\mathcal{E}(t_{12}t_{13}t_{23}) - 4\mathcal{E}(t_{12}t_{13}t_{23}^2)$$

$$+ \mathcal{E}(t_{12}^2 t_{13}^2 t_{23}^2) + O(n^{-4}) \tag{4.52}$$

where μ_{2r} is the 2^{rth} central moment of t_{ij}.

Let $t_\ell = \binom{n}{2}^{-1} \sum_{i<j} \text{signum}((x_{1i} - x_{1j})(x_{2i} - x_{2j}))$ where x_1 and x_2 are observations on the random variables X_1 and X_2 appropriate to t_ℓ. Then, in an abbreviated notation,

$$\prod_{\ell=1}^{k} t_\ell \stackrel{\text{def}}{=} \binom{n}{2}^{-k} \sum_{i_1<j_1} \sum_{i_2<j_2} \cdots \sum_{i_k<j_k} .$$

Expanding the product and taking expectations shows that if $\mathcal{E}(t_\ell) = 0$; $\ell = 1, \ldots, k$ then any term which has at least one pair of subscripts i_g, j_g distinct from all the other subscripts i_ℓ, j_ℓ; $\ell \neq g$ yields a zero expectation. Therefore the dominant terms in the expectation are those which contain a minimum of $k/2$, or $(k+1)/2$ for odd k, tied subscripts and no distinct pairs i_g, j_g. Consider, therefore,

$$\mathcal{E}(t_{12}^2 t_{13}^2 t_{23}^2) = \binom{n}{2}^{-6} \sum_{g_1<h_1} \sum_{g_2<h_2} \sum_{i_1<j_1} \sum_{i_2<j_2} \sum_{k_1<\ell_1} \sum_{k_2<\ell_2}$$

where subscripts g, h are for t_{12}, subscripts i, j are for t_{13} and subscripts k, ℓ are for t_{23}. It follows from the preceding argument that the dominant terms contain three ties. Of all of these terms, the pairwise independence of t_{12}, t_{13} and t_{23} then shows that the only ones which yield nonzero expectations are those for which one of g_1, h_1 is tied with one of g_2, h_2, one of i_1, j_1 is tied with one of i_2, j_2, and one of k_1, ℓ_1 is tied with one of k_2, ℓ_2. The expectation of a typical dominant term is then

$$\binom{n}{2}^{-6} \mathcal{E}\Big(\sum_{g_1 \neq h_1 \neq h_2} \sum_{i_1 \neq j_1 \neq j_2} \sum_{k_1 \neq \ell_1 \neq \ell_2} \Big) = \binom{n}{2}^{-6} \Big(\mathcal{E} \sum_{g_1 \neq h_1 \neq h_2}\Big)\Big(\mathcal{E} \sum_{i_1 \neq j_1 \neq j_2}\Big)\Big(\mathcal{E} \sum_{k_1 \neq \ell_1 \neq \ell_2}\Big)$$

from which it follows that

$$\mathcal{E}(t_{12}^2 t_{13}^2 t_{23}^2) = \mu_2^3 + O(n^{-4}) . \tag{4.53}$$

Turning to

$$\mathcal{E}(t_{12}t_{13}t_{23}^3) = \binom{n}{2}^{-5} \sum_{g<h}\sum_{i<j}\sum_{k_1<\ell_1}\sum_{k_2<\ell_2}\sum_{k_3<\ell_3} ,$$

the pairwise independence requires that for a typical dominant term, one of g,h is tied with one of i,j is tied with one of k_1,ℓ_1 or k_2,ℓ_2 or k_3,ℓ_3 – a double tie. Note that the double tie may occur at g,h or at k,ℓ and not necessarily at i,j as in the foregoing. The third tie must then occur between the two currently untied k,ℓ pairs - a single tie. Taking expectations over the sums with a double tie yields $3\mathcal{E}(t_{12}t_{13}t_{23})$ since there are three ways of choosing the tied k,ℓ pair. Taking expectations over the sums with a single tie yields $\mathcal{E}(t_{23}^2)$. Combining these two results shows that

$$\mathcal{E}(t_{12}t_{13}t_{23}^3) = 3\mu_2\mathcal{E}(t_{12}t_{13}t_{23}) + O(n^{-4}) . \tag{4.54}$$

Expanding $t_{12}t_{13}t_{23}$ in terms of Q_{12}, Q_{13} and Q_{23}, and substituting from eqn. (4.38) into the result shows that

$$\mathcal{E}(t_{12}t_{13}t_{23}) = \frac{2n^2+6n+7}{27\binom{n}{2}^2} . \tag{4.55}$$

Substituting from eqns. (4.53) to (4.55) into eqn. (4.52), and noting that $\mu_4 = 3\mu_2^2$, then yields

$$\begin{aligned}\mathcal{E}(U/V) &= \mu_2 + \frac{8n}{27\binom{n}{2}^2} + 13\mu_2^3 - 12\mu_2\mathcal{E}(t_{12}t_{13}t_{23}) + O(n^{-4}) \\ &> \mu_2 .\end{aligned} \tag{4.56}$$

Since $\mathcal{E}(V) = (1-\mu_2)^2$ it follows that

$$\begin{aligned}\frac{\mathcal{E}(U)}{\mathcal{E}(V)} &= \left(\mu_2 - 2\mathcal{E}(t_{12}t_{13}t_{23}) + \mu_2^2\right)\Big(\sum_{k=0}^{\infty}\mu_2^k\Big)^2 \\ &= \mathcal{E}(U/V) - 8\mu_2^3 + 8\mu_2\mathcal{E}(t_{12}t_{13}t_{23}) + O(n^{-4}) \\ &> \mathcal{E}(U/V) .\end{aligned} \tag{4.57}$$

Eqns. (4.56) and (4.57) establish the theorem.

4.5 The asymptotic normality of $t_{12.3}$

Hoflund (1963) concluded that the normal distribution gives a reasonable approximation to the distribution of $t_{12.3}$ for $n > 10$. Maghsoodloo and Pallos (1981) concluded that the normal distribution provides adequate estimates of the quantiles for $n \geq 50$. Examination of Hoflund's Table 1 and of Maghsoodloo's and Pallos's Tables 2 and 6 leads to the conclusion that the latter were much too conservative in their stated value of n and that the normal distribution certainly does give a reasonable approximation for, at least, $n \geq 20$.

The asymptotic normality of $t_{12.3}$ may be formally established by using the same method which Kendall (1975, Sect. 5.8) used to prove the asymptotic normality of the total score S. Since the distribution of t is asymptotically normal, it follows from the Second Limit Theorem that

$$\mu_{2r} = \frac{(2r)!}{2^r r!}(\mu_2)^r \tag{4.58}$$

where μ_{2r} is the $2^{r\text{th}}$ central moment of t_{ij}. Let λ_{2r} be the $2^{r\text{th}}$ central moment of $t_{12.3}$. Then eqn (4.49) shows that, for large n, $\lambda_2 = \mu_2$. Since the distribution of $t_{12.3}$ is symmetrical about zero it suffices to show that, for large n, $\lambda_{2r} = \mu_{2r}$ in order to demonstrate, via the converse of the Second Limit Theorem, that $t_{12.3}$ is asymptotically normal.

For large n the constraints on t_{13} and on t_{23} may be ignored. Approximating $\mathcal{E}(U^r/V^r)$ by $\mathcal{E}(U^r)/\mathcal{E}(V^r)$ gives

$$\lambda_{2r} = \frac{\mathcal{E}\{(t_{12} - t_{13}t_{23})^{2r}\}}{\mathcal{E}\{(1 - t_{13}^2)^r(1 - t_{23}^2)^r\}}\,. \tag{4.59}$$

Now

$$\mathcal{E}\{(1 - t_{13}^2)^r(1 - t_{23}^2)^r\} = \left(\sum_{k=0}^{r} E\binom{r}{k}(-t_{13}^2)^{r-k}\right)^2$$

$$= \left(\sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} \frac{(2r-2k)!}{2^{r-k}(r-k)!} (\mu_2)^{r-k} \right)^2$$

$$= \left(1 + \sum_{k=0}^{r-1} (O(n^{-1}))^{r-k} \right)^2 , \qquad (4.60)$$

from which it follows that $\mathcal{V}ar(V^r) = O(n^{-1})$; this sufficing, via a Taylor linearization to quadratic terms, to establish the validity of the approximation used in eqn. (4.59) above. Also

$$\mathcal{E}\{(t_{12} - t_{13}t_{23})^{2r}\} = \sum_{k=0}^{2r} \mathcal{E}\binom{2r}{k} t_{12}^k (-t_{13}t_{23})^{2r-k}$$

$$= \mu_{2r} + \sum_{k=0}^{2r-1} \mathcal{E}\binom{2r}{k} t_{12}^k (-t_{13}t_{23})^{2r-k} . \qquad (4.61)$$

It remains to be shown that each term of $\sum_{k=0}^{2r-1} \mathcal{E}\binom{2r}{k} t_{12}^k (-t_{13}t_{23})^{2r-k}$ is of lower order in n than μ_{2r} . The Cauchy-Schwarz inequality gives

$$\left| \mathcal{E}\left(t_{12}^k (t_{13}t_{23})^{2r-k}\right) \right| \leq \left\{ \mathcal{E}(t_{12}^k) \mathcal{E}\left((t_{12}t_{23})^{4r-2k}\right) \right\}^{1/2}$$

$$= \left\{ \mu_{2k} (\mu_{4r-2k})^2 \right\}^{1/2}$$

$$\propto (\mu_2)^{2r - \frac{1}{2}k} \qquad (4.62)$$

which is of lower order in n than μ_{2r} for $k < 2r$. Therefore $\mathcal{E}\{(t_{12} - t_{13}t_{23})^{2r}\} = \mu_{2r}$ which completes the proof.

4.6 Assessment of the complete null hypothesis

Based on Kendall's (1942) arguments, the appropriate null hypothesis to test for the presence or absence of a monotonic relationship between X_1 and X_2 , independently of the influence of a third variable X_3 , is the hypothesis $H_o : \tau_{12} = \tau_{13}\tau_{23}$. Moran (1951), Hoflund (1963), Maghsoodloo (1975), and Maghsoodloo and Pallos (1981), all considered the distribution of $t_{12.3}$ under the complete null hypothesis. This hypothesis, while facilitating the development of both theoretical and simulation results which enable a statistical test of the hypothesis to be carried out, assumes independence between X_3 and X_1 and between X_3 and X_2. It therefore fails to directly address the fundamental objective of removing the influence of X_3 since it assumes, a priori, that no such influence exists.

For the Pearson product moment correlation coefficient ρ, it is well known that the distribution of $r_{12.3}$, under the general null hypothesis that $\rho_{12} = \rho_{13}\rho_{23}$, is independent of the underlying values of ρ_{12}, ρ_{13} and ρ_{23}. Kendall and Stuart (1973, Vol 2, Ch. 27) demonstrate that this follows as a consequence of specific geometric properties, and the independence of the n observations on each of X_1, X_2 and X_3, of the product moment correlation coefficients. It follows that the distribution of $r_{12.3}$ which pertains under the restricted hypothesis of pairwise independence is the distribution which pertains under the general null hypothesis. If this situation were to extend itself to the Kendall rank correlation coefficients, then results developed under the complete null hypothesis would immediately be applicable to H_o. Regarding this, consider Maghsoodloo's analysis of the data of Table 1, Maghsoodloo (1975). For this data, $t_{xz} = -0.7143$, $t_{yz} = -0.9048$ and $t_{xy} = 0.8095$ so that $t_{xy.z} = 0.548$. Maghsoodloo used the approximate quantiles of $t_{xy.z}$, obtained by Monte Carlo simulation, to test the independence of X and Y given that Z is fixed. However, the quantiles used are those appropriate to the complete null hypothesis and, therefore, the extension of results from the complete null hypothesis to H_o, discussed above, is implicit in Maghsoodloo's data analysis.

Moran (1950) states, "while it is clear ... that t can be regarded as a product moment correlation coefficient between two sets of scores, and thus as the cosine of an angle, it is nevertheless very difficult to visualize this angle or to use this fact to construct a sampling theory". Furthermore, he notes that the scores are not independent among themselves. It would therefore appear that there is no fundamental basis for assuming that the aforementioned extension from ρ to τ

is valid. The validity of the extension may be examined in more detail by using simulation studies. These require simulation of data with known parental rank correlation coefficients. It is therefore appropriate, at this juncture, to consider two probability models which are capable of achieving this objective. One model is based on the equation,

$$\mathcal{E}(t) = \frac{2}{\pi} \sin^{-1} \rho \,, \tag{4.63}$$

connecting ρ and $\mathcal{E}(t)$ for the samples from a bivariate normal probability distribution. The other model is developed in the succeeding chapter.

REFERENCES

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, **19**, 293–325.

Hoflund, O. (1963). Simulated distributions for small n of Kendall's partial rank correlation coefficient. *Biometrika*, **50**, 520–522.

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

Kendall, M.G. (1942). Partial rank correlation. *Biometrika*, **32**, 277–283.

Kendall, M.G. (1975). *Rank Correlation Methods* (4th ed). Griffin and Co. Ltd.

Kendall, M.G. and Stuart, A. (1973). *The Advanced Theory of Statistics* (3rd ed). Griffin and Co. Ltd.

Maghsoodloo, S. (1975). Estimates of the quantiles of Kendall's partial rank correlation coefficient. *Journal of Statistical Computation and Simulation*, **4**, 155–164.

Maghsoodloo, S. and Pallos, L.L. (1981). Asymptotic behaviour of Kendall's partial rank correlation coefficient and additional quantile estimates. *Journal of Statistical Computation and Simulation*, **13**, 41–48.

Moran, P.A.P. (1950). Recent developments in ranking theory. *Journal of the Royal Statistical Society*, B, **12**, 153–162.

Moran, P.A.P. (1951). Partial and multiple rank correlation. *Biometrika*, **38**, 26–32.

Appendix to Chapter 4

ENUMERATING THE DISTRIBUTION OF KENDALL'S PARTIAL TAU

Keywords: Kendall's partial tau

Language

(i) APL (ii) ANSI Standard Fortran 1977

Description and Purpose

Each program computes the exact distribution and variance of Kendall's partial rank correlation coefficient for $n \leq 6$.

Method of Enumeration

The programs implement a methodology, for enumerating the distribution of $t_{12.3}$, which was developed in Chapter 4. The APL program affords an example of the economy of programming which this language allows. Use is made of the relatively new concept of nested arrays. The Fortran program executes much more rapidly than the APL version.

Auxiliary Algorithms

The following auxiliary subroutine is called:

SUBROUTINE QCKSRT (*N, ARR*)—(Press, Flannery, Teukolsky and Vetterling, 1986)

References

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986). *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press.

```
∇PTAUDISTN1 [⎕] ∇
∇Z←PTAUDISTN1 N;Q12;BE;J
  ⍝ MAIN FUNCTION. ACTS MOSTLY AS A CONTROL.
  ⍝ INITIALIZE VARIABLES
  ⎕IO←0
  Q12←2 2ρ0 1 1 0
  BE←BLKEFFINIT
  J←3 ◊ J ◊ →FIRST
  LOOPJ: →(((J>N)∧N≤5),(J>5)∧N>5)/PASTJ,CHANGEJ
  BE←J BLKEFFCUR BE
  FIRST: Q12←(J (!(J−1))) UNNEST BE + (J Jρ1)∘.×⊂Q12
  Z←J PTDIS1 Q12
  J←J+1
  J
  →LOOPJ
  PASTJ: →0
  CHANGEJ: Z←N PTAUDISTN2 (Q12 BE)
∇
∇BLKEFFINIT [⎕] ∇
∇Z←BLKEFFINIT;X1;X2
  ⍝ RETURNS THE INITIAL BLOCK EFFECTS MATRICES OBTAINED
  ⍝ FROM THE DISTRIBUTIONS FOR N = 2,3
  X1←2 2ρ0 2 2 0
  X2←2 2ρ1
  Z←3 3ρ(X2−1) X2 (X2+1) X2 X1 X2 (X2+1) X2 (X2−1)
∇
∇BLKEFFCUR [⎕] ∇
∇Z←N BLKEFFCUR BE;BECNTRL;M;NP;T
  ⍝ DETERMINES THE CURRENT BLOCK EFFECTS MATRICES
  NP←N−1
  BECNTRL←(0 2 1 3)⍉T∘.,T←((⍳N)∘.×NPρ1)−(⍳N)∘.>⍳NP
  M←(0 2 1 3)⍉T∘.≠T←(⍳N)∘.>⍳NP
  Z←(N Nρ⊂NP (!(NP−1))) UNNEST ¨⊂[2 3] M+(⊂¨BECNTRL)⊃¨⊂BE
∇
∇UNNEST [⎕] ∇
∇Z←LK UNNEST R;DLK
  ⍝ RETURNS AN UNNESTED MATRIX OF DIMENSION
  ⍝ (L×K),(L×K) FROM AN (L,L) MATRIX R WITH
  ⍝ NESTED SUBARRAYS OF SIZE (K,K).
  Z←(DLK DLK←×/LK)ρ,(0 2 1 3)⍉⊃R
∇
∇PTDIS1 [⎕] ∇
∇Z←N PTDIS1 Q12;T;T12;T13;T123;DT123;VT123;VTA123
  ⍝ USES THE MATRIX ARRAY Q12 TO COMPUTE THE
  ⍝ DISTRIBUTION OF TAU12.3.
  T12←1−(¯1 ¯1↓1 1↓Q12)×2÷(2!N)
  T13←1−(¯1↓1↓Q12[0;])×2÷(2!N)
  T123←,(T12−T13∘.×T13)÷T∘.×T←(1−T13*2)*0.5
  →(N<5)/CONT
  T123←|T123
  CONT: DT123← FREQ1 T123
  VT123←(+/DT123[;1]×DT123[;0]*2)÷+/DT123[;1]
```

```
[10]  T←(9×(2!N))-5+2×N
[11]  VTA123←((2×N)+5-(4×(N-2)×(N+1)÷T)+36×(2!N)÷(!N)*2)÷T
[12]  Z←DT123, [0] (VT123,VTA123)
    ∇
     ∇FREQ1 [⎕] ∇
    ∇Y←FREQ1 X;O
[1]   Y←O, [0.5+⎕IO] +/(O←O [⍋O←∪X] )∘.=X
    ∇


     ⍝ THE ABOVE FUNCTIONS SUFFICE FOR N .LE. 5
     ⍝ TO CONSERVE WORKSPACE REQUIREMENTS, THE FUNCTIONS
     ⍝ PTAUDISTN2, PTDIS2 AND FREQ2, WHICH UTILIZE A
     ⍝ SYMMETRY PROPERTY OF T12.3, ARE USED FOR N .EQ. 6


     ∇PTAUDISTN2 [⎕] ∇
    ∇Z←N PTAUDISTN2 R;J;JP;G;H;Q12;IQ12;BE;IBE;DT;DT123;T
[1]   J←6 ◊ G←0 ◊ JP←J-1
[2]   DT←((T+1) (1+T←⌊(J-1)÷2))ρ0
[3]   J ◊ DT123←2ρ0
[4]   Q12←⊃R [0]  ◊ BE←⊃R [1]
[5]   LOOPG: →(G>T)/PASTG
[6]   H←0
[7]   LOOPH: →((G≠0)∨H>T)/NOTADDH
[8]   IQ12←Q12+H
[9]   DT [G;H] ←⊂(J,G,H) PTDIS2 (IQ12 Q12)
[10]  →INC
[11]  NOTADDH: →((H<G),H>T)/TRANSPOSE,PASTH
[12]  IBE←((G>⍳JP)∘.≠(H>⍳JP)) + (⊂¨(G-G>⍳JP)∘.,(H-H>⍳JP))⊃¨⊂BE
[13]  IQ12←Q12 + (JP (!(JP-1))) UNNEST IBE
[14]  DT [G;H] ←⊂(J,G,H) PTDIS2 (IQ12 Q12)
[15]  →INC
[16]  TRANSPOSE: DT [G;H] ←DT [H;G]
[17]  INC: DT123←DT123, [0] ⊃DT [G;H]
[18]  H←H+1 ◊ →LOOPH
[19]  PASTH: G←G+1 ◊ →LOOPG
[20]  PASTG: Z←(0,0,0) PTDIS2 DT123
    ∇
     ∇PTDIS2 [⎕] ∇
    ∇Z←L PTDIS2 R;Q12;IQ12;T;T12;T13;T23;T123;DT123;VT123;VTA123
[1]   ⍝ USES THE MATRIX ARRAY Q12 TO COMPUTE THE
[2]   ⍝ DISTRIBUTION OF TAU12.3.
[3]   N←L [0]  ◊ →(0=+/L)/FINAL
[4]   IQ12←⊃R [0]  ◊ Q12←⊃R [1]
[5]   →(((L [1] =0)∧L [2] =0),((L [1] =0)∧L [2] ≠0),L [1] ≠0)/L1,L2,L3
[6]   L1: T12←1-(1 1↓IQ12)×2÷(2!N)
[7]   T13←1-(1↓IQ12 [0;] )×2÷(2!N)
[8]   T123←|,(T12-T13∘.×T13)÷T∘.×T←(1-T13*2)*0.5
[9]   →CONT
```

```
[10]   L2: T12←1-(1 0↓IQ12)×2÷(2!N)
[11]   T13←1-(1↓Q12[0;])×2÷(2!N)
[12]   T23←1-IQ12[0;]×2÷(2!N)
[13]   T123←|,(T12-T13∘.×T23)÷((1-T13*2)*0.5)∘.×(1-T23*2)*0.5
[14]   →CONT
[15]   L3: T12←1-IQ12×2÷(2!N)
[16]   T13←1-((Q12[0;])+L[1])×2÷(2!N)
[17]   T23←1-((Q12[0;])+L[2])×2÷(2!N)
[18]   T123←|,(T12-T13∘.×T23)÷((1-T13*2)*0.5)∘.×(1-T23*2)*0.5
[19]   CONT: Z← FREQ1 T123
[20]   →RETURN
[21]   FINAL: N←6 ⋄ DT123← FREQ2 1↓[0] R
[22]   VT123←(+/DT123[;1]×DT123[;0]*2)÷+/DT123[;1]
[23]   T←(9×(2!N))-5+2×N
[24]   VTA123←((2×N)+5-(4×(N-2)×(N+1)÷T)+36×(2!N)÷(!N)*2)÷T
[25]   Z←DT123,[0](VT123,VTA123)
[26]   RETURN: →0
      ∇
       ∇FREQ2[⎕]∇
      ∇Y←FREQ2 X;O
[1]    Y←O,[0.5+⎕IO]((O←O[⍋O←∪X[;0]])∘.=X[;0])+.×X[;1]
      ∇
       )OUTPUT
```

```
      PROGRAM PRANKF
C     COMPUTES THE EXACT DISTRIBUTION OF T12.3 FOR N<=6

C     .. LOCAL SCALARS ..
      DOUBLE PRECISION DEN, NUM, SUMN, SUMD, SUMS, TAU12, TAU13,
     *                 TAU23, TAU123, DERR
      REAL VARRAT, VARTAU, VAREQN, VART12
      INTEGER FIN, ARG1, ARG2, ADDRESS, COUNT
      LOGICAL DISTIN
C     .. LOCAL ARRAYS ..
      REAL TAUDIS(10000), RLOC(10000)
      DOUBLE PRECISION X(0:10000,2)
      INTEGER*2 IQ12(720,720), IBE(0:5,0:5,120,120), TAUFREQ(10000),
     * IBECNTRL(0:5,0:5), M(0:5,0:5,0:5,0:5), IBEO(0:4,0:4,120,120)

      COMMON IBE, IBEO

      DATA X/20002*0.0D0/
C     INITIALIZE IQ12 FOR N=2

      OPEN(6, FILE = 'PTDIST.DAT', STATUS = 'NEW')
      N = 6
      IQ12(1,1) = 0
      IQ12(1,2) = 1
      IQ12(2,1) = 1
      IQ12(2,2) = 0
      NFAC = 2
      J=3
10    IF (J .GT. N) GO TO 100
      JP = J - 1
      NFACP = NFAC
      NFAC = NFACP*J
      COUNT = 0
      DISTIN = .TRUE.

      SUMN = 0.0D0
      SUMD = 0.0D0
      SUMS = 0.0D0
      DO 50 I1 = 1,J
         K1 = I1 - 1
         DO 40 I2 = 1,J
            K2 = I2 - 1
            CALL BLKEFF(K1,K2,JP,NFACP,IBECNTRL,M)
            DO 30 L1 = 1,NFACP
               ARG1 = L1 + NFACP*K1
               DO 20 L2 = 1,NFACP
                  ARG2 = L2 + NFACP*K2
                  IQ12(ARG1,ARG2) = IQ12(L1,L2) + IBE(K1,K2,L1,L2)
                  TAU13 = 1.0D0 - IQ12(1,ARG1)*4.0D0/(J*JP)
                  TAU23 = 1.0D0 - IQ12(1,ARG2)*4.0D0/(J*JP)
                  DEN = (1.0D0-TAU13**2)*(1.0D0-TAU23**2)
                  IF (DEN .EQ. 0.0D0) GO TO 20
                  TAU12 = 1.0D0 - IQ12(ARG1,ARG2)*4.0D0/(J*JP)
                  NUM = TAU12-TAU13*TAU23
                  SUMN = SUMN + NUM**2
                  SUMD = SUMD + DEN
                  SUMS = SUMS + NUM**2/DEN
                  TAU123 = NUM/DSQRT(DEN)
                  IF (TAU123 .LT. -0.1E-06) GO TO 20
                  MTA123 = ANINT(TAU123*10000)
                  IF (X(MTA123,2) .EQ. 0.0D0) THEN
                     COUNT = COUNT + 1
                     X(MTA123,2) = 1.0
                     X(MTA123,1) = TAU123
                     RLOC(COUNT) = REAL(MTA123)
                  ELSE
                     X(MTA123,2) = X(MTA123,2) + 1.0
                     DERR = ABS(X(MTA123,1) - TAU123)
                     IF (DERR .GT. 1.0E-06) DISTIN = .FALSE.
                  END IF
20             CONTINUE
```

```
30          CONTINUE
40        CONTINUE
50      CONTINUE
        CALL QCKSRT(COUNT,RLOC)
        DO 60 I = 1,COUNT
          LOC = INT(RLOC(I))
          TAUDIS(I) = X(LOC,1)
          TAUFREQ(I) = X(LOC,2)
          X(LOC,1) = 0.0
          X(LOC,2) = 0.0
60      CONTINUE
        VARRAT = SUMN/SUMD
        VARTAU = SUMS/(NFAC-2)**2
        DEND = 9*J*(J-1.0)/2 - (2*J+5)
        VAREQN = (2*J+5)/DEND - 4*(J-2)*(J+1)/DEND**2
        VART12 = (2*J+5)/(9*J*(J-1.0)/2)

C       OUTPUT THE DISTRIBUTION OF TAU12.3

        WRITE (6,99999) J
        WRITE (6,99998)
        WRITE (6,99997)
        WRITE (6,99996) VARTAU, VARRAT, VAREQN, VART12
        WRITE (6,99995)
        CUMP = 0.0
        DO 70 I = 1,COUNT
          K = COUNT + 1 - I
          CUMP = CUMP + FLOAT(TAUFREQ(K))/(NFAC-2)**2
          WRITE (6,99994) TAUDIS(K), TAUFREQ(K), CUMP
70      CONTINUE
        WRITE (6,99993) COUNT
        IF (DISTIN) THEN
          WRITE (6,99992)
        ELSE
          WRITE (6,99991)
        ENDIF
        IF (J .EQ. N) GO TO 100

C       MAKE A COPY IBEO OF IBE FOR USE IN THE NEXT ITERATION.

        DO 90 I1 = 1,J
          K1 = I1 - 1
          DO 80 I2 = 1,J
            K2 = I2 - 1
            CALL COPYBLK(NFACP,K1,K2)
80        CONTINUE
90      CONTINUE
        J=J+1
        GO TO 10
100     CLOSE (6)
99999 FORMAT(5(/),1X,'CURRENT SIZE OF RANKING IS:  ',I2)
99998 FORMAT(2(/),1X,'VARIANCE OF TAU12.3:   EXACT   RATIO',
     *       ' APPROX   RATIO APPROX   VARTAU12')
99997 FORMAT(33X,'(COMPUTED)',5X,'(EQUATION)',5X,'(EXACT)')
99996 FORMAT(/,24X,F8.4,5X,F6.4,9X,F6.4,7X,F6.4)
99995 FORMAT(2(/),1X,'TAU12.3    FREQUENCY   CUMULATIVE PROBABILITY')
99994 FORMAT(/,1X,F6.4,5X,I6,14X,F5.4)
99993 FORMAT(/,1X,'THERE ARE',I6,' DIFFERENT VALUES OF TAU12.3',
     *       ' IN THE ABOVE FREQUENCY DISTRIBUTION')
99992 FORMAT(/,1X,'THE VALUES OF TAU12.3 IN THE ABOVE FREQ.',
     *       ' DISTRIBUTION ARE DISTINCT')
99991 FORMAT(/,1X,'THE VALUES OF TAU12.3 IN THE ABOVE FREQ.',
     *       ' DISTRIBUTION ARE NOT DISTINCT')
        STOP
        END

        SUBROUTINE BLKEFF(K1,K2,JP,NFACP,IBECNTRL,M)

C       CALLED BY MAIN PROGRAM. COMPUTES THE REQUIRED BLOCK EFFECTS
C       MATRIX. ON THE FIRST CALL, INITIALIZES THE BLOCK EFFECTS
C       MATRICES FOR THE DISTRIBUTIONS CORRESPONDING TO N=1,2.
```

```
C     .. ARRAY ARGUMENTS ..
      INTEGER*2 IBECNTRL(0:5,0:5), M(0:5,0:5,0:5,0:5)
C     .. LOCAL SCALARS ..
      INTEGER ADDRESS, ARG1, ARG2, ARG3, ARG4
C     .. LOCAL ARRAYS ..
      INTEGER*2 IB(0:5,0:5)
C     .. ARRAYS IN COMMON ..
      INTEGER*2 IBE(0:5,0:5,120,120), IBEO(0:4,0:4,120,120)

      COMMON IBE, IBEO

      IF (NFACP .NE. 2) GO TO 30
      IBEO(0,0,1,1) = 0
      IBEO(0,1,1,1) = 1
      IBEO(1,0,1,1) = 1
      IBEO(1,1,1,1) = 0
      DO 20 I2 = 0,5
         DO 10 I1 = 0,5
            IBECNTRL(I1,I2) = -1
   10    CONTINUE
   20 CONTINUE
   30 DO 90 I1 = 1,JP
         J1 = I1-1
         IF (IBECNTRL(K1,J1) .NE. -1) GO TO 40
         IF (K1 .GT. J1) THEN
            IBECNTRL(K1,J1) = K1-1
            IB(K1,J1) = 1
         ELSE
            IBECNTRL(K1,J1) = K1
            IB(K1,J1) = 0
         ENDIF
   40    ARG3 = IBECNTRL(K1,J1)
         DO 80 I2 = 1,JP
            J2 = I2-1
            IF (IBECNTRL(K2,J2) .NE. -1) GO TO 50
            IF (K2 .GT. J2) THEN
               IBECNTRL(K2,J2) = K2-1
               IB(K2,J2) = 1
            ELSE
               IBECNTRL(K2,J2) = K2
               IB(K2,J2) = 0
            ENDIF
   50       IF (IB(K1,J1) .EQ. IB(K2,J2)) THEN
               M(K1,K2,J1,J2) = 0
            ELSE
               M(K1,K2,J1,J2) = 1
            ENDIF
            ARG4 = IBECNTRL(K2,J2)
            DO 70 L1 = 1,(NFACP/JP)
               ARG1 = (NFACP/JP)*J1 + L1
               DO 60 L2 = 1,(NFACP/JP)
                  ARG2 = (NFACP/JP)*J2 + L2
                  IBE(K1,K2,ARG1,ARG2) = IBEO(ARG3,ARG4,L1,L2)
     *                                   + M(K1,K2,J1,J2)
   60          CONTINUE
   70       CONTINUE
   80    CONTINUE
   90 CONTINUE
      RETURN
      END

      SUBROUTINE COPYBLK(NFACP,K1,K2)

C     CALLED BY MAIN PROGRAM. MAKES A COPY OF THE PREVIOUS BLOCK
C     EFFECTS MATRICES SINCE THEIR VALUES ARE DESTROYED WHEN THE
C     CURRENT BLOCK EFFECTS MATRICES ARE BEING COMPUTED

C     .. SCALAR ARGUMENTS ..
      INTEGER NFACP, K1, K2
C     .. ARRAYS IN COMMON ..
      INTEGER*2 IBE(0:5,0:5,120,120), IBEO(0:4,0:4,120,120)
```

```
      COMMON IBE, IBEO

      DO 20 I1 = 1,NFACP
         DO 10 I2 = 1,NFACP
            IBEO(K1,K2,I1,I2) = IBE(K1,K2,I1,I2)
10       CONTINUE
20    CONTINUE
      RETURN
      END
```

Chapter 5

A PROBABILITY MODEL FOR THE NON-NULL DISTRIBUTION OF KENDALL'S TAU

The fact that $\mathcal{E}(t) = \tau$, for random samples taken from a population with rank correlation τ is used as a basis for developing a probability distribution for the elements of an inversion vector. It is shown that this distribution leads to an exact variance, for t, which is identical to the upper limit, in large samples, of the variance estimator which pertains when the variates are drawn from a bivariate normal distribution. An application of the results to hypothesis testing is presented and some potential applications of the probability model are suggested.

The developed probability distribution is conceptualized as resulting from an unknown sampling mechanism for samples of size n taken from an infinite population of order statistics, $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$. As Barnett (1977) states, "a generating mechanism for the sample will yield a particular set of possible samples each of which occurs with a probability determined from the nature of the generating mechanism". Consider a sample $x_1, x_2, \ldots, x_n$ and the associated value of t. By definition

$$t = 1 - \frac{2q_n}{\binom{n}{2}} \qquad \text{and} \qquad q_n = \sum_{j=1}^{n} i_j \tag{5.1}$$

where i_j is the j^{th} element of the inversion vector corresponding to $x_1, x_2, \ldots, x_n$ and q_n is the negative score. For the null case, it has been shown, Section 3.2.1, how the first and second moments of t can be readily obtained once the probability distribution of i_j is known. The sequel will

(i) Show that the use of a random sampling mechanism leads to the correct probability distribution of i_j in the null case.

(ii) Obtain a probability distribution for i_j, in the non-null case, which is consistent with the requirement that $\mathcal{E}(t) = \tau$.

(iii) Derive $\mathcal{V}ar(t)$ based on the probability distribution of i_j obtained in (ii).

(iv) Present an application of $\mathcal{V}ar(t)$ to one-sample tests of τ.

(v) Conclude with a discussion of, and suggested applications for, the developed probability model.

5.1 Application of the concept of a generating mechanism: null case

Draw a simple random sample without replacement of size j from the population of order statistics; this corresponding to a sampling mechanism which chooses observations one at a time at random and without replacement. Since the elements of the population form a partition of the sample space it follows that

$$\Pr(i_j = k) = \sum_{m=1}^{N} \Pr(x_j = X_{(m)}) \times \Pr(i_j = k | x_j = X_{(m)}) \qquad (5.2)$$

where, from the properties of a random sample,

$$\Pr(x_j = X_{(m)}) = \frac{1}{N}$$

and

$$\Pr(i_j = k | x_j = X_{(m)}) = \frac{\binom{m-1}{j-k-1}\binom{N-m}{k}}{\binom{N-1}{j-1}}, \qquad 0 \le i_j \le j-1 .$$

Therefore,

$$\Pr(i_j = k) = \sum_{m=1}^{N} \frac{1}{N} \times \frac{\binom{m-1}{j-k-1}\binom{N-m}{k}}{\binom{N-1}{j-1}}$$

$$= \binom{j-1}{k} \sum_{m=1}^{N} \frac{1}{N} \left(\frac{m}{N}\right)^{j-k-1} \left(1 - \frac{m}{N}\right)^{k} + R_N \qquad (5.3)$$

where the term R_N is a sum of terms all of which are $O(1/N)$ or of lower order in N. Consequently,

$$\lim_{N\to\infty} \Pr(i_j = k) = \binom{j-1}{k} \int_0^1 x^{j-k-1}(1-x)^k \, dx . \qquad (5.4)$$

The CRC Standard Mathematical Tables (1979) gives

$$\int x^m(a+bx)^n\,dx = \frac{x^{m+1}(a+bx)^n}{m+n+1} + \frac{an}{m+n+1}\int x^m(a+bx)^{n-1}\,dx$$

from which it follows that

$$\begin{aligned}\int_0^1 x^{j-k-1}(1-x)^k\,dx &= \frac{k}{j}\int_0^1 x^{j-k-1}(1-x)^{k-1}\,dx \\ &= \frac{k!}{j(j-1)\cdots(j-k+1)}\int_0^1 x^{j-k-1}\,dx \\ &= \frac{k!}{j(j-1)\cdots(j-k)}\,. \end{aligned} \tag{5.5}$$

Substituting from eqn. (5.5) into eqn. (5.4) then yields

$$\lim_{N\to\infty}\Pr(i_j = k) = \frac{1}{j} \tag{5.6}$$

which is a completely satisfactory result since, under the null hypothesis of equiprobable rankings, the set of permutations of the first n natural integers possesses the property that $\Pr(i_j = k) = \frac{1}{j}$ for $j = 1,\ldots,n$ and $0 \le k \le j-1$.

5.2 A probability distribution for i_j: non-null case

In order to derive an appropriate probability distribution for i_j it is necessary to first consider some desirable properties of such a distribution. Taking expectations in eqn. (5.1) yields $\mathcal{E}(q_n) = \frac{1}{2}\binom{n}{2}(1-\mathcal{E}(t))$. Subtracting $\mathcal{E}(q_n)$ from $\mathcal{E}(q_{n+1})$, and substituting for q_n in the result, then yields $\mathcal{E}(i_{n+1}) = \frac{n}{2}(1-\tau)$ or, more generally,

$$\mathcal{E}(i_j) = \frac{1}{2}(j-1)(1-\tau)\,. \tag{5.7}$$

If the restriction is imposed that, for $\tau = 0$, the ensuing probability distribution must correspond to the sampling mechanism of Section 5.1 then, in the absence of any further information, $\tau = 0$ is synonymous with an absence of rank correlation.

The probability distribution must then possess the property that

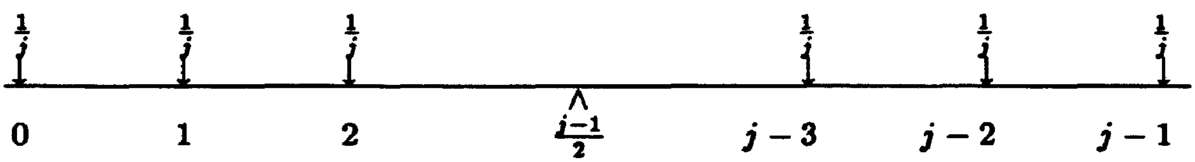
$$\tau = 0 \;\Rightarrow\; \Pr(i_j = k) = \frac{1}{j}$$

$$\tau = 1 \;\Rightarrow\; \Pr(i_j = k) = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\tau = -1 \;\Rightarrow\; \Pr(i_j = k) = \begin{cases} 1 & \text{if } k = j-1 \\ 0 & \text{otherwise.} \end{cases} \tag{5.8}$$

Having elucidated, in eqns. (5.7) and (5.8), the desirable properties of an appropriate probability distribution the search for such a distribution is now begun. Consider a weightless beam, supported at position $\frac{1}{2}(j-1)$ and with point masses of $\frac{1}{j}$ placed at positions $0, 1, \ldots, j-1$ as shown below:



The beam as shown is perfectly balanced. Now consider displacing the beam's support to the left or to the right, according as $\tau \geq 0$ or $\tau \leq 0$, to position $\frac{1}{2}(j-1)(1-\tau)$. Our objective is to balance the beam by reducing the point masses at some locations while increasing those at other locations so that the net mass remains constant. The cases $j = 2$ and $j = 3$ are considered prior to defining a general formula for redistributing the point masses subject to eqn. (5.8). This general formula will then define a probability distribution subject to eqns. (5.7) and (5.8).

Case: $j = 2$

Let the point masses be $p_0 = \frac{1}{2} + a$ and $p_1 = \frac{1}{2} - b$ at locations denoted by the subscripts on p. Since $p_0 + p_1 = 1$ then $a = b$. Taking moments about the support gives

$$\left(\frac{1}{2} + a\right)\left(\frac{1-\tau}{2}\right) = \left(\frac{1}{2} - a\right)\left(1 - \frac{1-\tau}{2}\right)$$

which reduces to $a = \frac{1}{2}\tau$. Therefore $p_0 = \frac{1}{2}(1+\tau)$ and $p_1 = \frac{1}{2}(1-\tau)$.

Case: $j = 3$

Let the point masses be $p_0 = \frac{1}{3} + a$, $p_1 = \frac{1}{3} - b$ and $p_2 = \frac{1}{3} - c$ so that $a = c + b$. Taking moments about the support then gives

$$\left(\frac{1}{3} + b + c\right)(1-\tau) = \left(\frac{1}{3} - b\right)\tau + \left(\frac{1}{3} - c\right)(1+\tau)$$

which reduces to $b + 2c = \tau$. Now eqn. (5.8) requires that $p_1 = 0$ for $\tau = \pm 1$. Hence b is chosen so that $\frac{1}{3} - b = \frac{1}{3}(1-\tau)(1+\tau)$ which yields $b = \frac{1}{3}\tau^2$. It then follows that $c = \frac{1}{6}(3\tau - \tau^2)$ and that $a = \frac{1}{6}(3\tau + \tau^2)$. Therefore $p_0 = \frac{1}{6}(1+\tau)(2+\tau)$, $p_1 = \frac{2}{6}(1+\tau)(1-\tau)$ and $p_2 = \frac{1}{6}(1-\tau)(2-\tau)$.

The results for the case $j = 3$ are highly suggestive and lead to our defining, as the required probability distribution, the function

$$p_{k,j} = \begin{cases} \frac{1}{j!}\binom{j-1}{k}(1-\tau)\cdots(k-\tau)(1+\tau)\cdots(j-k-1+\tau) & \text{if } 0 \le k \le j-1 \\ 0 & \text{otherwise.} \end{cases} \tag{5.9}$$

Two theorems, which establish that the function $p_{k,j}$ does define a probability function whose mean is given by eqn. (5.7), are now proven.

Theorem 5.1: The function $p_{k,j}$ as defined by eqn. (5.9) is a probability function.

Proof: The theorem is obviously true for $j = 2$ and for $j = 3$. Assume that it is true for arbitrary $j > 0$. It suffices to establish that this assumption implies that the result is true for $j + 1$.

Decompose each element $p_{k,j}$ into two parts by multiplying by $\frac{1}{j+1}(j-k+\tau)$ and $\frac{1}{j+1}(k+1-\tau)$ and then sum the decompositions. Now consider the sum

$$p_{k,j} \times \frac{1}{j+1}(j-k+\tau) \ + \ p_{k-1,j} \times \frac{1}{j+1}((k-1)+1-\tau)$$

which is equal to

$$\frac{1}{(j+1)!}(1-\tau)\cdots(k-\tau)(1+\tau)\cdots(j-k+\tau)\left\{\binom{j-1}{k} + \binom{j-1}{k-1}\right\}$$

$$= \frac{1}{(j+1)!}\binom{j}{k}(1-\tau)\cdots(k-\tau)(1+\tau)\cdots(j-k+\tau)$$
$$= p_{k,j+1}\,. \qquad (5.10)$$

If $p_{-1,j} = 0$ and $p_{j,j} = 0$ it is easily seen that eqn. (5.10) holds for $0 \le k \le j$. Hence summing the decompositions of $p_{k,j}$ as in eqn. (5.10) generates the set $p_{k,j+1}$ and therefore $\sum_{k=0}^{j} p_{k,j+1} = 1$ which establishes the result.

Theorem 5.2: The probability function $p_{k,j}$ generates a probability distribution with mean given by $\mathcal{E}(i_j) = \frac{1}{2}(j-1)(1-\tau)$.

Proof: The theorem is true for $j = 2$ and for $j = 3$. It is therefore shown that if the theorem is true for arbitrary $j > 0$ then this implies that $\mathcal{E}(i_{j+1}) = \frac{1}{2}j(1-\tau)$. Now $\mathcal{E}(i_{j+1}) = \sum_{k=1}^{j} kp_{k,j+1}$ which yields, upon substitution from eqn. (5.10),

$$\mathcal{E}(i_{j+1}) = \frac{1}{j+1}\sum_{k=1}^{j}\Big(k(j-k+\tau)p_{k,j} + k(k-\tau)p_{k-1,j}\Big)$$
$$= \frac{1}{j+1}\Big\{(1-\tau) + \sum_{k=1}^{j-1}\Big(k(j-k+\tau) + (k+1)(k+1-\tau) - (1-\tau)\Big)p_{k,j}\Big\} \qquad (5.11)$$

by virtue of the fact that $p_{j,j} = 0$ and $\sum_{k=0}^{j-1} p_{k,j} = 1$ from which $(1-\tau)p_{0,j} = (1-\tau)\big(1 - \sum_{k=1}^{j-1} p_{k,j}\big)$. The coefficient of $p_{k,j}$ in eqn. (5.11) reduces to $(j+2)k$ so that

$$\mathcal{E}(i_{j+1}) = \frac{1}{j+1}\Big\{(1-\tau) + (j+2)\sum_{k=1}^{j-1} kp_{k,j}\Big\}$$
$$= \frac{1}{2}j(1-\tau) \qquad (5.12)$$

where, by assumption, $\mathcal{E}(i_j) = \frac{1}{2}(j-1)(1-\tau)$. Eqn. (5.12) establishes the theorem.

5.3 Derivation of the variance of t

By definition,

$$\mathcal{E}(i_{j+1}^2) = \sum_{k=1}^{j} k^2 p_{k,j+1}$$

$$= \frac{1}{j+1}\left\{(1-\tau)+\sum_{k=1}^{j-1}\Big(k^2(j-k+\tau)+(k+1)^2(k+1-\tau)-(1-\tau)\Big)p_{k,j}\right\}. \tag{5.13}$$

The coefficient of $p_{k,j}$ reduces to $(j+3)k^2+(3-2\tau)k$ and therefore

$$\mathcal{E}(i_{j+1}^2) = \frac{1-\tau}{j+1}\left\{1+\frac{1}{2}(j-1)(3-2\tau)+\left(\frac{j+3}{1-\tau}\right)\mathcal{E}(i_j^2)\right\}. \tag{5.14}$$

Solving directly for $\mathcal{E}(i_2^2)$ gives $\frac{1-\tau}{2}$. Substituting recursively into eqn. (5.14) and solving then gives,

j	$\mathcal{E}(i_j^2)$	C_{1j}	C_{2j}
2	$\frac{1}{2}(1-\tau)$	1	0
3	$\frac{1}{3}(1-\tau)(5-\tau)$	5	1
4	$\frac{1}{4}(1-\tau)(14-4\tau)$	14	4
5	$\frac{1}{5}(1-\tau)(30-10\tau)$	30	10
6	$\frac{1}{6}(1-\tau)(55-20\tau)$	55	20

where C_{1j} and C_{2j} are such that $\mathcal{E}(i_j^2) = \frac{1}{j}(1-\tau)(C_{1j}-C_{2j}\tau)$. The sequence C_{1j} is recognizable as the series of partial sums of the square of the first $(j-1)$ natural integers while the sequence $(C_{2j}-C_{2,j-1})$ is recognizable as the series of partial sums of the first $(j-2)$ natural integers. Consequently, $C_{1j} = \sum_{k=1}^{j-1} k^2$ and $C_{2j} = \sum_{k=1}^{j-2}\sum_{\ell=1}^{k}\ell$ so that $C_{1j} = \frac{1}{6}(j-1)j(2j-1)$ and $C_{2j} = \frac{1}{6}\{(j-1)j(2j-1)-j(j^2-1)\}$. It now follows that

$$\begin{aligned}\mathcal{E}(i_j^2) &= \frac{1}{6}(2j^2-3j+1)(1-\tau)^2+\frac{1}{6}\tau(1-\tau)(j^2-1)\\ &= (1-\tau)^2\mathcal{E}(i_j^2|\tau=0)+\frac{1}{6}\tau(1-\tau)(j^2-1)\end{aligned} \tag{5.15}$$

a result which satisfies eqn. (5.14). A feature of the probability model is that it assumes independence between the elements of the inversion vector. Consequently,

$$\begin{aligned}\mathcal{E}(Q^2) &= \sum_{j=1}^{n}\mathcal{E}(i_j^2)+\sum_{j\neq\ell}^{n}\mathcal{E}(i_j i_\ell)\\ &= (1-\tau)^2\mathcal{E}(Q^2|\tau=0)+\sum_{j=1}^{n}\frac{1}{6}\tau(1-\tau)(j^2-1)\end{aligned} \tag{5.16}$$

from which, since $\mathcal{E}(Q) = (1-\tau)\mathcal{E}(Q|\tau=0)$, it follows that

$$\begin{aligned} \mathcal{V}ar(Q) &= (1-\tau)^2\,\mathcal{V}ar(Q|\tau=0) + \frac{1}{36}\tau(1-\tau)n(n-1)(2n+5) \\ &= \frac{1}{36}(1-\tau^2)(2n+5)\binom{n}{2}\,. \end{aligned} \tag{5.17}$$

Hence, from eqns. (5.1) and (5.17),

$$\begin{aligned} \mathcal{V}ar(t) &= \frac{(1-\tau^2)(2n+5)}{9\binom{n}{2}} \\ &= (1-\tau^2)\,\mathcal{V}ar(t|\tau=0)\,. \end{aligned} \tag{5.18}$$

5.4 Application to one-sample tests of tau

Schemper (1987) applied bootstrapping and jackknife techniques to the problem of one and two-sample tests of τ. Her results for the one-sample test are compared to those obtained under the developed probability model which directly yields an estimator for the variance of t under the null hypothesis. Tables 5.1 and 5.2 are identical to those reported by Schemper except that they have been extended to incorporate results obtained by using eqn. (5.18). Each result is based on 1000 simulations of a bivariate normal distribution with correlation coefficient given by $\rho = \sin(\pi\tau_a/2)$, where τ_a and τ_{H_0} designate the actual and hypothesized values of tau. The mean value of t for each set of 1000 simulations with $n = 15$ is also shown in Table 5.2. These confirm that the simulation process is producing samples with the desired underlying parent value of τ_a.

Table 5.1: Size of one-sample tests by permutational variance/eqn. (5.18)/ bootstrap techniques/Edgeworth-corrected bootstrap at $\alpha=0.05$

	$\tau=0.0$	$\tau=0.4$	$\tau=0.8$	$\tau=0.9$
$n=15$	5/4.8/6/8	3/4.5/6/8	0/1.1/2/4	0/0.9/1/2
$n=30$	5/4.5/5/6	3/3.8/5/6	0/0.6/2/3	0/0.3/6/3

The power of the test for $\tau_a = 0.9$ and $\tau_{H_0} = 0.8$ is conservative when compared with that for $\tau_a = 0.8$ and $\tau_{H_0} = 0.9$. This demonstrates the fact that hypothesis tests based on eqn. (5.18) will be conservative if $|\tau_a| > |\tau_{H_o}|$, since the sample variance computed under H_o will overestimate the population variance. Conversely, for $|\tau_a| < |\tau_{H_o}|$ the computed sample variance, under H_o, will underestimate the population variance thus leading to a more sensitive test. Apart from this one case where the power of the test is low, the power of tests based on eqn. (5.18) compares favorably with the power of the bootstrap tests as reported by Schemper and, in some instances, outperforms them (see $\tau_a = 0.4$ and $\tau_{H_o} = 0.8, 0.9$ for example). The size of our test is low for high values of τ; this reflecting the fact that, for sampling from a bivariate normal distribution, the variance given by eqn. (5.18) is a conservative value of, or an upper limit for, the true variance of t - see eqn. (5.19) below. However, the test is very simple, is easily implemented, and the Tables indicate that it performs reasonably well when compared with the much more complex and computationally demanding bootstrap techniques.

Table 5.2: Power of one-sample tests by permutational variance/eqn. (5.18)/ bootstrap techniques/Edgeworth-corrected bootstrap at α=0.05

τ_a	mean t (n=15)	τ_{H_o}	n=15	n=30
0.0	0.0014	0.4	53/62/46/46	88/91/85/84
0.4	0.3991	0.0	54/59/57/62	89/90/89/90
		0.8	55/86/62/52	93/99/98/97
		0.9	76/99/89/79	99/100/100/100
0.8	0.7969	0.4	48/80/87/92	97/100/99/100
		0.9	1/20/5/1	3/38/55/39
0.9	[illegible]19	0.4	83/99/97/99	100/100/100/100
		0.8	0/0/17/29	0/2/34/48

5.5 Discussion and conclusion

Consider the variance of t when it is known that random samples are drawn from a bivariate normal population. Kendall (1975) established that for large samples,

$$Var(t) \leq \frac{(1-t^2)(2n+5)}{9\binom{n}{2}} \,. \tag{5.19}$$

(Note that while Kendall used a slightly different inequality than that given by eqn. (5.19), the above inequality follows directly upon substitution from eqn. (9.16) into eqn. (9.15) of Kendall's text). This is indeed a gratifying result since the right hand side of eqn. (5.19) is equivalent to the variance obtained in eqn. (5.18). This result also allows an interpretation of the absence of any additional information, beyond the value of t, about the population which is being sampled. In such a circumstance a possible approach is to estimate the variance of t by the upper limit attainable under the assumption that sampling is from a bivariate normal distribution. A more accurate estimator of the variance of t may be obtained by estimating a second population parameter. Daniels and Kendall (1947) and Hoeffding (1947) have both developed the necessary results. However, this estimator yields an estimate of $Var(t)$ based on the observed sample values of τ and the second population parameter, whereas eqn. (5.18) allows for an estimate of $Var(t)$ under H_o.

The developed probability model provides a basis for studying the sampling distribution of t in the non-null case. Furthermore, the probability distribution is indexed only by the parameter τ and thus distributional results are functions only of τ and not of an array of unknown parameters. Sundrum (1953) has discussed the problem of obtaining higher order moments of t, and the diverse array of unknown parameters which must be dealt with, for the general case where parental correlation exists. Additionally, the developed probability model may be used in simulation studies since eqn. (5.9) used in conjunction with $U(0,1)$ random variables permits the simulation of permutations, of the first n natural integers, with the desired characteristics – see the Appendix to Chapter 5. Confidence intervals for t may be readily obtained and hypothesis testing easily implemented. Also, since permutations are directly generated, it is to be expected that the results of simulation studies using this probability model will more accurately reflect the behaviour of ordinal data than would the results of simulation studies using the normal probability model.

The situation, whereby the distribution of t depends not only on τ but also on other population parameters pertaining to the actual arrangement of order statistics in the parent population, also applies to the case $\tau = 0$. All that is often known about the parent population is that τ appears, on the basis of repeated sampling, to be zero or non-zero. The restriction that $\tau = 0$ is synonymous with independence, or with an absence of rank correlation, therefore appears to be quite reasonable.

REFERENCES

Barnett, V. (1974). *Elements of Sampling Theory.* The English University Press Ltd.

Beyer, W.H. ed. (1979). *CRC Standard Mathematical Tables* (25th ed). CRC Press, Inc.

Daniels, H.E. and Kendall, M.G. (1947). The significance of rank correlations where parental correlation exists. *Biometrika*, **34**, 197–208.

Hoeffding, W. (1947). On the distribution of the rank correlation coefficient τ when the variates are not independent. *Biometrika*, **34**, 183–196.

Kendall, M.G. (1975). *Rank Correlation Methods* (4th ed). Griffin and Co. Ltd.

Schemper, M. (1987). One- and two-sample tests of Kendall's τ. *Biometrical Journal*, **29**, 1003–1009.

Sundrum, R.M. (1953). Moments of the rank correlation coefficient τ in the general case. *Biometrika*, **40**, 409–420.

Appendix to Chapter 5

THE SIMULATION OF PERMUTATIONS WITH A PARENTAL RANK CORRELATION OF TAU

Keywords: Kendall's tau; inversion vector; rank correlation

Language

ANSI Standard Fortran 1977

Description and Purpose

Generates permutations of the first n natural numbers which, when correlated against the natural order permutation, have a parent correlation of τ. The permutations thus generated possess the property that the distribution of t is determined entirely by τ. The statistic t is widely used as a measure of trend and as a measure of monotonic relationship between two variables.

Method of Simulation

In Chapter 5 we developed a probability distribution for the elements of an inversion vector so that $\mathcal{E}(t) = \tau$ for the corresponding permutation and the natural order permutation. Independent $U(0,1)$ random variates are used to simulate an inversion vector with the desired probability distribution and the permutation corresponding to this inversion vector is then determined.

Structure

SUBROUTINE TAUSIM (TAU, N, ISEED, JSEED, TOBS, IP, INV, IW)

Formal Parameters

TAU	Real	input:	the population parameter
N	Integer	input:	the size of the permutation
ISEED	Integer	input:	seed for uniform random number generator
JSEED	Integer	input:	seed for uniform random number generator
TOBS	Real	output:	the sample value of Kendall's tau
IP	Integer array (*N*)	output:	the desired permutation
INV	Integer array (*N*)	output:	the associated inversion vector
IW	Integer array (*N*)	output:	work vector. Keeps a record of which integers have already been assigned to *IP*

Auxiliary Algorithms

The random number generator Rsuper-duper (Marsaglia, 1976) is used to generate $U(0,1)$ variates.

References

Marsaglia, G. (1976). Random number generation. *Encyclopedia of Computer Science*, ed. A. Ralston, pp. 1192–1197. New York: Petrocelli and Charter.

```
      SUBROUTINE TAUGEN(TAU, IP, INV, IW, N, TOBS, ISEED, JSEED)
C     THE PROBABILITY MODEL DEVELOPED IN CHAPTER 5 IS USED TO
C     SIMULATE DATA WITH A PARENTAL RANK CORRELATION OF TAU.
C     THE SUBROUTINE RETURNS A PERMUTATION IP AND THE OBSERVED
C     VALUE OF TAU FOR THIS PERMUTATION CORRELATED AGAINST THE
C     NATURAL ORDER PERMUTATION
C     .. ARRAY ARGUMENTS ..
      DIMENSION IP(N), INV(N), IW(N)
C     GENERATE THE INVERSION VECTOR
      INV(1) = 0
      QINV = 0.0
      DO 60 I = 1,N
         IW(I) = 0
         J = N + 1 - I
         IF (J .EQ. 1) GO TO 60
         JM1 = J-1
         SUMPR = 0.0
         RAND = RSUPER(ISEED,JSEED)
         DO 50 K = 0,JM1
            PKJ = 1.0/J
            IF (K .EQ. 0) GO TO 20
            DO 10 IG = 1,K
               PKJ = PKJ*(IG - TAU)/IG
10          CONTINUE
20          JK1 = J-K-1
            IF (JK1 .EQ. 0) GO TO 40
            DO 30 IH = 1,JK1
               PKJ = PKJ*(IH + TAU)/IH
30          CONTINUE
40          SUMPR = SUMPR + PKJ
            IF (RAND .LT. SUMPR) THEN
               INV(J) = K
               QINV = QINV + INV(J)
               GO TO 60
            ENDIF
50       CONTINUE
60    CONTINUE
      TOBS = 1.0 - 4*QINV/(N*(N-1.0))
C     COMPUTE THE CORRESPONDING PERMUTATION
      DO 80 I = 1,N
         J = N+1-I
         L = 0
         IF (J .EQ. N) THEN
            IP(J) = N - INV(J)
            IW(IP(J)) = 1
            GO TO 80
         END IF
         K = N
70       IF (IW(K) .EQ. 0) L = L+1
         IF (L .GT. INV(J)) THEN
            IP(J) = K
            IW(K) = 1
         ELSE
            K = K-1
            GO TO 70
         END IF
80    CONTINUE
      RETURN
      END
```

Chapter 6

THE ASYMPTOTIC NULL DISTRIBUTION OF PARTIAL TAU WHEN PARENTAL RANK CORRELATION EXISTS

Monte Carlo simulations reveal that the variance of $t_{12.3}$, under $H_o : \tau_{12} = \tau_{13}\tau_{23}$, varies with the underlying values of τ_{13} and τ_{23}. This result leads to a consideration, in Sections 6.2 and 6.3, of some properties of the distributions of both t and $t_{12.3}$, when parental rank correlation exists. Explicit use of the indicator random variable allows straightforward derivations of the non-null variance of t and the covariance of t_{12} and t_{13}. An asymptotic variance estimator for $t_{12.3}$ is derived and the asymptotic normality of $t_{12.3}$, under H_o and for the general case of variates with underlying parental correlation, is established. Finally, the suitability of $t_{12.3}$ as a statistic for measuring the monotonic correlation between X_1 and X_2, independently of the influence of X_3, is addressed. It is shown, in Section 6.4, that $t_{12.3}$ is not a suitable test statistic when the magnitudes of τ_{13} and τ_{23} are both large.

6.1 Data Simulation with $\tau_{12} = \tau_{13}\tau_{23}$

Eqn. (4.63) suggests that the relationship $\rho_{ij} = \sin(\pi\tau_{ij}/2)$ may be used as a basis for simulating data with a parental rank correlation of τ_{ij}. It then suffices to simulate data from a trivariate normal distribution with variance-covariance matrix, V, equal to the correlation matrix, $\rho_{ij} = \sin(\pi\tau_{ij}/2)$ for $ij = 12, 13, 23$, and $\tau_{12} = \tau_{13}\tau_{23}$. Kennedy and Gentle (1980, Section 6.5.9) note that this may be done by generating z_1, z_2, z_3 as independent $N_1(0,1)$ variates, forming the vector z distributed as $N_3(\mathbf{o}, \mathbf{I})$ and then obtaining x as $x = Az$, where A is such that

$\boldsymbol{AA'} = \boldsymbol{V}$. $\boldsymbol{A}$ is obtained from a Cholesky decomposition of $\boldsymbol{V}$, Kennedy and Gentle (1980, Section 7.4).

Table 6.1 shows the variance of $t_{12.3}$, for varying values of τ_{13} and τ_{23}, obtained by using 100 simulations of 100 observations on each of X_1, X_2 and X_3.

Table 6.1: $Var(t_{12.3})$ for varying τ_{13} and τ_{23} [1]

τ_{13}	τ_{23}	t_{13}	t_{23}	t_{12}	$t_{12.3}$	$Var(t_{12.3})$
-0.8	-0.8	-0.80	-0.80	0.64	-0.0020	0.0009
-0.8	-0.6	-0.80	-0.60	0.48	0.0002	0.0012
-0.8	-0.4	-0.80	-0.40	0.32	-0.0023	0.0013
-0.8	-0.2	-0.80	-0.20	0.15	-0.0076	0.0016
-0.8	0.0	-0.80	0.01	0.00	0.0043	0.0019
-0.8	0.2	-0.80	0.19	-0.16	0.0004	0.0019
-0.8	0.4	-0.80	0.40	-0.31	0.0061	0.0018
-0.8	0.6	-0.80	0.60	-0.48	0.0046	0.0015
-0.8	0.8	-0.80	0.80	-0.64	-0.0003	0.0008
-0.6	-0.6	-0.60	-0.60	0.36	-0.0006	0.0021
-0.6	-0.4	-0.60	-0.40	0.24	-0.0051	0.0025
-0.6	-0.2	-0.61	-0.20	0.12	-0.0041	0.0023
-0.6	0.0	-0.60	0.00	0.01	0.012	0.0022
-0.6	0.2	-0.59	0.19	-0.11	0.0042	0.0025
-0.6	0.4	-0.60	0.39	-0.23	0.0037	0.0026
-0.6	0.6	-0.60	0.60	-0.36	0.0035	0.0025
-0.6	0.8	-0.60	0.80	-0.48	0.0026	0.0015
-0.4	-0.4	-0.41	-0.39	0.15	-0.0077	[illegible]
-0.4	-0.2	-0.39	-0.20	0.07	-0.0077	0.0041
-0.4	0.0	-0.38	0.01	0.00	-0.0018	0.0035
-0.4	0.2	-0.40	0.20	-0.08	-0.0014	0.0046
-0.4	0.4	-0.40	0.40	-0.16	-0.0039	0.0036
-0.4	0.6	-0.41	0.60	-0.25	-0.0024	0.0026
-0.4	0.8	-0.40	0.80	-0.32	0.0010	0.0015
-0.2	-0.2	-0.20	-0.19	0.03	-0.012	0.0037
-0.2	0.0	-0.20	-0.01	0.00	-0.0029	0.0047
-0.2	0.2	-0.20	0.18	-0.03	0.0070	0.0036
-0.2	0.4	-0.19	0.40	-0.09	-0.012	0.0050
-0.2	0.6	-0.19	0.60	-0.11	0.0012	0.0026
-0.2	0.8	-0.21	0.80	-0.17	-0.0056	0.0014
0.0	0.0	0.00	-0.01	-0.01	-0.0056	0.0042
0.0	0.2	-0.01	0.21	-0.01	-0.0099	0.0044
0.0	0.4	-0.01	0.40	-0.01	-0.0069	0.0039
0.0	0.6	0.00	0.60	0.005	0.0054	0.0035
0.0	0.8	-0.01	0.80	-0.01	-0.0014	0.0018
0.2	0.2	0.19	0.19	0.03	-0.010	0.0048
0.2	0.4	0.19	0.39	0.09	0.011	0.0050
0.2	0.6	0.21	0.60	0.12	-0.0014	0.0030
0.2	0.8	0.20	0.80	0.16	-0.0011	0.0022
0.4	0.4	0.39	0.40	0.15	-0.0076	0.0028
0.4	0.6	0.40	0.60	0.24	0.0086	0.0024
0.4	0.8	0.40	0.80	0.32	-0.0039	0.0013
0.6	0.6	0.60	0.60	0.36	0.0031	0.0021
0.6	0.8	0.60	0.80	0.48	0.0024	0.0013
0.8	0.8	0.80	0.80	0.64	0.0047	0.0011

[1] sample values $t_{13}, \ldots, t_{12.3}$ are means of 100 simulations

The sample means for t_{ij} clearly show that the data simulation procedure does produce data with the desired parental rank correlation coefficients. Since the variance of $t_{12.3}$ decreases as the magnitudes of both t_{13} and t_{23} increase away from zero, this effect being particularly noticeable whenever $|\tau| \geq 0.6$, it follows that distributional results obtained under the complete null hypothesis do not extend to the general null hypothesis.

6.2 $\mathcal{E}(t)$, $\mathcal{V}ar(t)$ and $\mathcal{C}ov(t_{ij}, t_{jk})$: non-null case

The mean and the variance of t have been addressed by several authors including Daniels and Kendall (1947), Hoeffding (1947) and Noether (1967). However, it is helpful to develop the results by making explicit use of the indicator random variable. Consider a fixed population of size N and let $s_{g_1g_2}$ denote the sign of $(x_{1g_1} - x_{1g_2})(x_{2g_1} - x_{2g_2})$. Draw a random sample, without replacement, of size n. By definition,

$$t = \binom{n}{2}^{-1} \sum_{g<h}^{n} s_{gh} = \binom{n}{2}^{-1} \sum_{g<h}^{N} s_{gh} I_{gh} , \tag{6.1}$$

where

$$I_{g_1g_2\ldots g_m} = \begin{cases} 1 & \text{if } \boldsymbol{x}_{g_1}, \boldsymbol{x}_{g_2}, \ldots, \boldsymbol{x}_{g_m} \in S_a \\ 0 & \text{otherwise} \end{cases} , \tag{6.2}$$

$\boldsymbol{x}$ denotes the pair (x_1, x_2) and S_a denotes the sample. Now, under random sampling, $\mathcal{E}(I_{g_1g_2\ldots g_m}) = \binom{N-m}{n-m}/\binom{N}{n} = n^{(m)}/N^{(m)}$ where, for example, $n^{(m)} = n(n-1)\ldots(n-m+1)$. It follows that

$$\begin{aligned} \mathcal{E}(t) &= \binom{n}{2}^{-1} \sum_{g<h}^{N} s_{gh} \mathcal{E}(I_{gh}) \\ &= \binom{N}{2}^{-1} \sum_{g<h}^{N} s_{gh} = \tau . \end{aligned} \tag{6.3}$$

Similarly, $\mathcal{E}(t^2)$ is obtained as,

$$\mathcal{E}(t^2) = \mathcal{E}\left(\binom{n}{2}^{-2} \sum_{g<h}^{N} s_{gh} I_{gh} \sum_{i<j}^{N} s_{ij} I_{ij} \right)$$

$$= \frac{1}{\binom{n}{2}^2}\left(\sum_{g<h}^{N} s_{gh}^2 \mathcal{E}(I_{gh}) + \sum_{g\neq h\neq j}^{N} s_{gh}s_{gj}\mathcal{E}(I_{ghj}) + \sum_{g<h}^{N}\sum_{i'<j'}^{N} s_{gh}s_{i'j'}\mathcal{E}(I_{ghi'j'})\right)$$

$$= \frac{1}{\binom{n}{2}}\left(\frac{1}{\binom{N}{2}}\sum_{g<h}^{N} s_{gh}^2 + \frac{2(n-2)}{N^{(3)}}\sum_{g\neq h\neq j}^{N} s_{gh}s_{gj} + \frac{\binom{n-2}{2}}{\binom{N}{2}\binom{N-2}{2}}\sum_{g<h}^{N}\sum_{i'<j'}^{N} s_{gh}s_{i'j'}\right) \tag{6.4}$$

where i' and j' are both distinct from either g or h. Let G_g and H_g be the number of x_h in the population which are concordant and discordant, respectively, with x_g. Then $\sum_{g\neq h}^{N} s_{gh} = \sum_{g}^{N}(G_g - H_g)$ and $\sum_{g\neq h}^{N} s_{gh}^2 = \sum_{g}^{N}(G_g + H_g)$ so that

$$\binom{N}{2}^{-1}\sum_{g<h}^{N} s_{gh}^2 = \frac{1}{N^{(2)}}\sum_{g}^{N}(G_g + H_g)$$
$$= \pi_c + \pi_d \tag{6.5}$$

where (π_c, π_d) are the probabilities of drawing a (concordant, discordant) pair from the population. Also,

$$\frac{1}{N^{(3)}}\sum_{g\neq h\neq j}^{N} s_{gh}s_{gj} = \frac{1}{N^{(3)}}\sum_{g\neq h}^{N}\left(s_{gh}\sum_{j\neq h,g} s_{gj}\right) = \frac{1}{N^{(3)}}\left(\sum_{g}^{N}(G_g - H_g)^2 - \sum_{g\neq h}^{N} s_{gh}^2\right)$$
$$= \frac{1}{N^{(3)}}\sum_{g}^{N}\Big(G_g(G_g-1) + H_g(H_g-1) - G_gH_g - H_gG_g\Big)$$
$$= \pi_{cc} + \pi_{dd} - \pi_{cd} - \pi_{dc} \tag{6.6}$$

where, for example, π_{cc} is the probability that among three observations x_g, x_h and x_j, x_g is concordant with both x_h and x_j. Use of $\pi_{cc}, \ldots, \pi_{dc}$ follows Noether's (1967) notation; the parameter π_{cc} being the parameter k used by Hoeffding (1947). Noting that

$$\binom{N}{2}^{-1}\binom{N-2}{2}^{-1}\sum_{g<h}^{N}\sum_{i'<j'}^{N} s_{gh}s_{i'j'} = \left(\binom{N}{2}^{-1}\sum_{g<h}^{N} s_{gh}\right)^2 + O(N^{-1}), \tag{6.7}$$

letting N approach infinity, and substituting from eqns. (6.5) to (6.7) into eqn. (6.4), yields the result that

$$\mathcal{E}(t^2) = \frac{1}{\binom{n}{2}}\left((\pi_c + \pi_d) + 2(n-2)(\pi_{cc} + \pi_{dd} - \pi_{cd} - \pi_{dc}) + \binom{n-2}{2}\tau^2\right)$$

so that

$$Var(t) = \frac{1}{\binom{n}{2}}\Big((\pi_c + \pi_d) + 2(n-2)(\pi_{cc} + \pi_{dd} - \pi_{cd} - \pi_{dc}) - (2n-3)\tau^2\Big) \quad (6.8)$$

which is equivalent to eqn. (10.8) of Noether (1967). If ties are disallowed, eqn. (6.8) may be further simplified as is done by Noether.

Let unbiased estimators of $(\pi_c + \pi_d)$ and of $(\pi_{cc} + \pi_{dd} - \pi_{cd} - \pi_{dc})$ be given by $(p_c + p_d)$ and $(p_{cc} + p_{dd} - p_{cd} - p_{dc})$. Then $(p_c + p_d)$ and $(p_{cc} + p_{dd} - p_{cd} - p_{dc})$ are obtained by redefining G_g and H_g with respect to the sample and replacing N by n in eqns. (6.5) and (6.6). Consider the unbiased variance estimator

$$V_u(t) = a_1(p_c + p_d) + a_2(p_{cc} + p_{dd} - p_{cd} - p_{dc}) + a_3 t^2 .$$

Taking expectations and solving for a_1, a_2 and a_3 shows that an unbiased estimator of $Var(t)$ is obtained as

$$V_u(t) = \frac{1}{\binom{n-2}{2}}\Big((p_c + p_d) + 2(n-2)(p_{cc} + p_{dd} - p_{cd} - p_{dc}) - (2n-3)t^2\Big) \quad (6.9)$$

a result which is equivalent to eqn. (5.65) of Kendall (1975), when ties are absent. With the advent of high speed digital computers, computation of eqn. (6.9) in small amounts of computer time is feasible, thus obviating the need to utilize the inequality for $Var(t)$ as given by Daniels and Kendall (1947, eqn. (9.1)). Schemper (1987) investigated the performance of jackknife and bootstrap techniques, in one- and two-sample tests of τ, and concluded that the use of Daniel's and Kendall's inequality is the worst choice.

For simplicity of exposition it is now assumed that ties are disallowed so that $s^{12}_{gh}s^{13}_{gh} = s^{23}_{gh}$. Taking $\mathcal{E}(t_{12}t_{13})$ for illustration,

$$\mathcal{E}(t_{12}t_{13}) = \frac{1}{\binom{n}{2}^2}\Big(\sum_{g<h}^{N} s^{12}_{gh}s^{13}_{gh}\mathcal{E}(I_{gh}) + \sum_{g\neq h\neq j}^{N} s^{12}_{gh}s^{13}_{gj}\mathcal{E}(I_{ghj}) + \sum_{g<h}^{N}\sum_{i'<j'}^{N} s^{12}_{gh}s^{13}_{i'j'}\mathcal{E}(I_{ghi'j'})\Big)$$

$$= \frac{1}{\binom{n}{2}}\Big(\frac{1}{\binom{N}{2}}\sum_{g<h}^{N} s^{23}_{gh} + \frac{2(n-2)}{N^{(3)}}\sum_{g\neq h\neq j}^{N} s^{12}_{gh}s^{13}_{gj} + \frac{\binom{n-2}{2}}{\binom{N}{2}\binom{N-2}{2}}\sum_{g<h}^{N}\sum_{i'<j'}^{N} s^{12}_{gh}s^{13}_{i'j'}\Big) . \quad (6.10)$$

Reducing eqn. (6.10) after the manner in which eqn. (6.4) was reduced then yields

$$\mathcal{E}(t_{12}t_{13}) = \frac{1}{\binom{n}{2}}\left(\tau_{23} + 2(n-2)(\pi'_{cc} + \pi'_{dd} - \pi'_{cd} - \pi'_{dc}) + \binom{n-2}{2}\tau_{12}\tau_{13}\right)$$

so that

$$\mathcal{C}ov(t_{12}, t_{13}) = \frac{1}{\binom{n}{2}}\left(\tau_{23} + 2(n-2)(\pi'_{cc} + \pi'_{dd} - \pi'_{cd} - \pi'_{dc}) - (2n-3)\tau_{12}\tau_{13}\right) \tag{6.11}$$

where, $'$ denotes ^{12,13} and, for example, $\pi^{12,13}_{cc}$ is the probability that among four observations $x^{12}_g, x^{12}_h, x^{13}_g$ and x^{13}_j, x^{12}_g is concordant with x^{12}_h and x^{13}_g is concordant with x^{13}_j. By analogy with eqn. (6.9) an unbiased estimator of $\mathcal{C}ov(t_{12}, t_{13})$ is obtained as

$$\mathcal{C}_u(t_{12}, t_{13}) = \frac{1}{\binom{n-2}{2}}\left(t_{23} + 2(n-2)(p'_{cc} + p'_{dd} - p'_{cd} - p'_{dc}) - (2n-3)t_{12}t_{13}\right) \tag{6.12}$$

where, for example, p'_{cc} is an unbiased estimator of π'_{cc}.

6.3 The Asymptotic Normality of $t_{12.3}$ under $H_o : \tau_{12} = \tau_{13}\tau_{23}$

It follows from eqns. (6.8) and (6.11) that the variances and covariances of t_{ij} are all $O(n^{-1})$. Consequently, a Taylor series expansion of $t_{12.3}$ about τ_{12}, τ_{13} and τ_{23}, under H_o, yields

$$t_{12.3} = C\big((t_{12} - \tau_{12}) - \tau_{23}(t_{13} - \tau_{13}) - \tau_{13}(t_{23} - \tau_{23})\big) + R_n \tag{6.13}$$

so that

$$\mathcal{E}(t_{12.3}) = \mathcal{E}(R_n) = O(n^{-1}) \tag{6.14}$$

and

$$\begin{aligned} \mathcal{V}ar(t_{12.3}) = C^2\big(\mathcal{V}ar(t_{12}) + \tau^2_{23}\,\mathcal{V}ar(t_{13}) + \tau^2_{13}\,\mathcal{V}ar(t_{23}) - 2\tau_{23}\mathcal{C}ov(t_{12}, t_{13}) \\ - 2\tau_{13}\mathcal{C}ov(t_{12}, t_{23}) + 2\tau_{13}\tau_{23}\mathcal{C}ov(t_{13}, t_{23})\big)\,, \end{aligned} \tag{6.15}$$

where $C = \left(\sqrt{(1-\tau^2_{13})(1-\tau^2_{23})}\right)^{-1}$. Let $W = (t_{12.3} - R_n)/C$ and λ_{2r} be the

2^{rth} central moment of $t_{12.3}$. Eqn. (6.13) then yields

$$\lambda_{2r} = \sum_{j=0}^{2r} \sum_{\ell=0}^{2r-j} \binom{2r}{j} \binom{2r-j}{\ell} (-O(n^{-1}))^{\ell} C^j \mathcal{E}(W^j R_n^{2r-j-\ell}) . \tag{6.16}$$

Eqn. (6.16) comprises terms such as $\mathcal{E}\big((t_{12}-\tau_{12})^{a_1}(t_{13}-\tau_{13})^{a_2}(t_{23}-\tau_{23})^{a_3}\big)$ where $a_1 + a_2 + a_3 = j + 2(2r - j - \ell)$. Consider the expansion of any such term in a manner analogous to the expansion of $\mathcal{E}(t^2)$ in eqn. (6.4). The result, eqn. (6.7), which shows that, for large N, the expectation of a product of two summations over distinct subscripts is equal to the product of the expectations, is found to apply more generally to products involving more than two summations provided that the summations are over distinct sets of subscripts. Suppose, therefore, that in the expansion of $\mathcal{E}\big((t_{12} - \tau_{12})^{a_1}(t_{13} - \tau_{13})^{a_2}(t_{23} - \tau_{23})^{a_3}\big)$ a product term with $\sum_{g<h}(s_{gh} - \tau)$ as one element of the product occurs. Since the expectation of this element is zero it follows that the expectation of the product term is zero. This leads to the result, stated in the proof of Theorem 4.3, that for large N, the non-zero terms of $\mathcal{E}\big((t_{12} - \tau_{12})^{a_1}(t_{13} - \tau_{13})^{a_2}(t_{23} - \tau_{23})^{a_3}\big)$ must contain a minimum of $k/2$, or $(k + 1)/2$ for odd k where $k = a_1 + a_2 + a_3$, tied subscripts. It follows that $\mathcal{E}\big((t_{12} - \tau_{12})^{a_1}(t_{13} - \tau_{13})^{a_2}(t_{23} - \tau_{23})^{a_3}\big)$ is $O(n^{-k'/2})$, where k' is equal to k or $k+1$ according as k is even or odd, so that

$$O(n^{-\ell})\mathcal{E}(W^j R_n^{2r-j-\ell}) = O(n^{-2r+j'/2}) \tag{6.17}$$

where j' is equal to j or $j-1$ according as j is even or odd. Consequently,

$$\lambda_{2r} = C^{2r}\mathcal{E}(W^{2r}) + O(n^{-(r+1)}) \tag{6.18a}$$

where $\mathcal{E}(W^{2r})$ is $O(n^{-r})$, and

$$\lambda_{2r+1} = O(n^{-(r+1)}) . \tag{6.18b}$$

Thus λ_{2r+1} is of order $n^{-1/2}$ in comparison to λ_{2r} and therefore it suffices to show that $\lambda_{2r} = \big((2r)!/(2^r r!)\big)\lambda_2^r$ in order to establish the asymptotic normality of $t_{12.3}$.

For notational simplicity, let $(t_1, t_2, t_3) = (t_{12}, t_{13}, t_{23})$ and $\mu_2^{i,j}$ denote $\mathcal{C}ov(t_i, t_j)$ for $i = 1,2$ or 3 and $j = 1,2$ or 3. Then

$$\lambda_2 = C^2\big(\mu_2^{1,1} + \tau_3^2\mu_2^{2,2} + \tau_2^2\mu_2^{3,3} - 2\tau_3\mu_2^{1,2} - 2\tau_2\mu_2^{1,3} + 2\tau_2\tau_3\mu_2^{2,3}\big) \qquad (6.19a)$$

so that

$$\lambda_2^r = \sum_{b_1+\cdots+b_6=r} (-1)^{b_4+b_5} 2^{b_4+b_5+b_6} \tau_3^{2b_2+b_4+b_6} \tau_2^{2b_3+b_5+b_6} C^{2r} \frac{r!}{\prod_{i=1}^6 b_i!} \times$$
$$\mu_2^{1,1^{b_1}} \mu_2^{2,2^{b_2}} \mu_2^{3,3^{b_3}} \mu_2^{1,2^{b_4}} \mu_2^{1,3^{b_5}} \mu_2^{2,3^{b_6}} . \qquad (6.19b)$$

Also,

$$\lambda_{2r} = \mathcal{E}\left[\sum_{c_1+c_2+c_3=2r} (-1)^{c_2+c_3} \tau_3^{c_2} \tau_2^{c_3} C^{2r} \frac{(2r)!}{\prod_{i=1}^3 c_i!} (t_1 - \tau_1)^{c_1} (t_2 - \tau_2)^{c_2} (t_3 - \tau_3)^{c_3}\right] . \qquad (6.20)$$

The expectation of any term of λ_{2r} involves an expectation of a product of $2r$ summations, of the form $\sum_{g<h}(s_{gh} - \tau)$. Upon expanding the product, the dominant terms of its expectation are those obtained by grouping the summations into pairs and tying a subscript from one summation to a subscript from the other summation, of the pair. Taking the expectation of such a product pair yields a variance or covariance term to $O(n^{-2})$. For example,

$$\mathcal{E}\Big(\frac{1}{\binom{n}{2}^2} \sum_{g\neq h\neq j}^{n} (s_{gh} - \tau_1)(s_{gj} - \tau_2)\Big) = \mathcal{C}ov(t_1, t_2) + O(n^{-2}) \approx \mu_2^{1,2} . \qquad (6.21)$$

Now consider the term in λ_{2r} such that

$$c_1 = 2b_1 + b_4 + b_5, \quad c_2 = 2b_2 + b_4 + b_6 \quad \text{and} \quad c_3 = 2b_3 + b_5 + b_6 \qquad (6.22)$$

and determine the number of ways of selecting pairs so that

$$\mathcal{E}\big((t_1 - \tau_1)^{c_1} (t_2 - \tau_2)^{c_2} (t_3 - \tau_3)^{c_3}\big) = \mu_2^{1,1^{b_1}} \mu_2^{2,2^{b_2}} \mu_2^{3,3^{b_3}} \mu_2^{1,2^{b_4}} \mu_2^{1,3^{b_5}} \mu_2^{2,3^{b_6}} . \quad (6.23)$$

There are $\binom{c_1}{2} \cdots \binom{c_1-2b_1+2}{2}/b_1!$, $\binom{c_2}{2} \cdots \binom{c_2-2b_2+2}{2}/b_2!$, and $\binom{c_3}{2} \cdots \binom{c_3-2b_3+2}{2}/b_3!$ ways of selecting pairs which yield variance terms. These chosen, there are then

$(c_1 - 2b_1)\cdots(c_1 - 2b_1 - b_4 + 1)(c_2 - 2b_2)\cdots(c_2 - 2b_2 - b_4 + 1)/b_4!$, $(c_1 - 2b_1 - b_4)\cdots(1)(c_3 - 2b_3)\cdots(c_3 - 2b_3 - b_5 + 1)/b_5!$, and $(c_2 - 2b_2 - b_4)\cdots(1)(c_3 - 2b_3 - b_5)\cdots(1)/b_6!$ ways of selecting pairs which yield covariance terms. Consequently, there is a total of $\prod_{i=1}^{3} c_i!/(2^{b_1+b_2+b_3}\prod_{i=1}^{6} b_i!)$ ways of selecting pairs and thus λ_{2r} contains a term

$$\frac{(-1)^{b_4+b_5}}{2^{b_1+b_2+b_3}}\tau_3^{2b_2+b_4+b_6}\tau_2^{2b_3+b_5+b_6}C^{2r}\frac{(2r)!}{\prod_{i=1}^{6} b_i!}\mu_2^{1,1^{b_1}}\mu_2^{2,2^{b_2}}\mu_2^{3,3^{b_3}}\mu_2^{1,2^{b_4}}\mu_2^{1,3^{b_5}}\mu_2^{2,3^{b_6}} .$$

Comparison of this term with eqn. (6.19b) shows that $\lambda_{2r} = \big((2r)!/(2^r r!)\big)\lambda_2^r$ which establishes the desired result.

The preceding proof uses the same approach as that used by Kendall (1975, Section 5.21) to establish the asymptotic normality of t when parental rank correlation exists. Kendall notes that an essential condition is that $1 - \tau^2$ be of order 1 so that the tendency to normality may break down for high correlations; a statement which obviously applies to the above result.

The variance estimator given by eqn. (6.15) is dependent on too many unknown parameters. Note that while H_o specifies a relationship among the τ 's and hence the π_c 's and π_d 's, it does not specify a relationship among the π_{cc} 's and the π'_{cc} 's etc. Consequently, H_o does not facilitate significant simplification of the expression in eqn. (6.15) and does nothing to alleviate its parameter dependency. Nevertheless, for large n, the asymptotic normality of $t_{12.3}$ and eqns. (6.14) and (6.15) provide a basis for assessing the behaviour of $t_{12.3}$ under H_o. Table 6.2 shows the fraction, α, of rejects for 300 simulations of 50 observations on each of X_1, X_2 and X_3 at a significance level of 0.05; the data being generated as described in Section 6.1. It is evident that the observed rejection rates are compatible with the used significance level.

6.4 Assessment of the hypothesis $H_o : \tau_{12} = \tau_{13}\tau_{23}$

Agresti (1977) has suggested that the term partial association be used to

Table 6.2: Fraction, α, of rejects, under H_o, for varying τ_{13} and τ_{23} [1]

τ_{13}	τ_{23}	t_{13}	t_{23}	t_{12}	$t_{12.3}$	$Var(t_{12.3})$	$\hat{\sigma}^2(t_{12.3})$	α
-0.8	-0.8	-0.80	-0.80	0.64	-0.001	0.0020	0.0022	0.057
-0.8	-0.6	-0.80	-0.61	0.49	0.005	0.0028	0.0031	0.033
-0.8	-0.4	-0.80	-0.41	0.32	-0.002	0.003[illegible]	0.0035	0.070
-0.8	-0.2	-0.80	-0.20	0.16	-0.001	0.0037	0.0038	0.030
-0.8	0.0	-0.80	0.01	-0.01	-0.002	0.0037	0.0039	0.043
-0.8	0.2	-0.80	0.20	-0.16	-0.004	0.0035	0.0038	0.053
-0.8	0.4	-0.80	0.41	-0.33	-0.002	0.0035	0.0035	0.043
-0.8	0.6	-0.80	0.61	-0.48	0.002	0.0032	0.0029	0.070
-0.8	0.8	-0.80	0.80	-0.64	-0.004	0.0021	0.0022	0.050
-0.6	-0.6	-0.60	-0.60	0.35	-0.003	0.0049	0.0048	0.060
-0.6	-0.4	-0.60	-0.41	0.25	0.006	0.0051	0.0055	0.023
-0.6	-0.2	-0.60	-0.21	0.14	0.012	0.0057	0.0061	0.040
-0.6	0.0	-0.60	0.00	0.00	-0.007	0.0076	0.0062	0.090
-0.6	0.2	-0.60	0.20	-0.12	0.004	0.0059	0.0061	0.057
-0.6	0.4	-0.60	0.40	-0.24	-0.005	0.0060	0.0056	0.043
-0.6	0.6	-0.60	0.60	-0.37	-0.004	0.0047	0.0047	0.050
-0.6	0.8	-0.60	0.80	-0.47	0.003	0.0029	0.0030	0.063
-0.4	-0.4	-0.40	-0.40	0.16	0.003	0.0075	0.0070	0.060
-0.4	-0.2	-0.40	-0.19	0.07	-0.009	0.0074	0.0076	0.047
-0.4	0.0	-0.40	-0.01	0.02	0.016	0.0082	0.0081	0.050
-0.4	0.2	-0.40	0.20	-0.08	-0.001	0.0085	0.0078	0.067
-0.4	0.4	-0.40	0.40	-0.16	-0.001	0.0074	0.0069	0.060
-0.4	0.6	-0.39	0.60	-0.24	-0.004	0.0057	0.0055	0.070
-0.4	0.8	-0.41	0.81	-0.33	-0.004	0.0031	0.0034	0.050
-0.2	-0.2	-0.20	-0.20	0.05	0.009	0.0095	0.0089	0.063
-0.2	0.0	-0.20	-0.01	0.01	0.012	0.0085	0.0090	0.067
-0.2	0.2	-0.21	0.20	-0.04	-0.002	0.0088	0.0087	0.083
-0.2	0.4	-0.20	0.40	-0.07	0.006	0.0079	0.0078	0.057
-0.2	0.6	-0.20	0.59	-0.12	-0.004	0.0059	0.0062	0.040
-0.2	0.8	-0.20	0.80	-0.16	-0.004	0.0035	0.0037	0.067
0.0	0.0	-0.01	0.00	0.00	-0.001	0.0098	0.0094	0.060
0.0	0.2	-0.01	0.20	0.00	-0.001	0.0098	0.0091	0.057
0.0	0.4	0.00	0.39	-0.01	-0.005	0.0087	0.0077	0.060
0.0	0.6	0.01	0.60	0.01	-0.001	0.0059	0.0061	0.043
0.0	0.8	-0.01	0.80	-0.01	0.003	0.0044	0.0039	0.053
0.2	0.2	0.20	0.20	0.04	0.002	0.0092	0.0088	0.063
0.2	0.4	0.21	0.40	0.09	0.009	0.0085	0.0076	0.063
0.2	0.6	0.21	0.60	0.12	-0.004	0.0068	0.0061	0.053
0.2	0.8	0.20	0.80	0.16	-0.001	0.0031	0.0038	0.040
0.4	0.4	0.39	0.41	0.17	0.007	0.0067	0.0071	0.040
0.4	0.6	0.40	0.59	0.24	-0.003	0.0056	0.0057	0.057
0.4	0.8	0.40	0.80	0.33	0.005	0.0032	0.0035	0.037
0.6	0.6	0.60	0.60	0.36	0.003	0.0049	0.0046	0.060
0.6	0.8	0.60	0.80	0.48	-0.002	0.0028	0.0030	0.053
0.8	0.8	0.80	0.80	0.64	0.002	0.0022	0.0022	0.050

[1] sample values $t_{13}, \ldots, t_{12.3}$ are means of 300 simulations
$\hat{\sigma}^2(t_{12.3})$ is the mean of 300 sample estimates obtained from eqn. (6.15)

refer to a descriptive measure of the degree of association between two variables X_1 and X_2, controlling for a third variable X_3. Then, as Quade (1974) has shown, four different concepts of control may be distinguished. Of these, only the first three are relevant to our purposes. They are

(i) Control by holding X_3 constant. This leads to weighted averages of bivariate correlation coefficients, obtained according to the X_3 classification or blocking variable.

(ii) Adjusting for X_3. This requires obtaining two residual processes, $X_{1r} = X_1 - f(X_3)$ and $X_{2r} = X_2 - g(X_3)$ where f and g are functions used for predicting X_1 and X_2 from X_3 and are such that the residuals are concentrated about zero as closely as possible. A product-moment bivariate correlation coefficient is then computed using the residuals. Moran (1950), eqn. (7), has shown that $t_{12.3}$ may be rewritten in terms of two residual processes so that $t_{12.3}$ belongs to category (ii).

(iii) Using a fourfold table, as in Table 4.5, and applying Kendall's (1942) argument to define independence in the table. It is evident that $\tau_{12.3}$ also belongs to this category.

6.4.1 The trivariate normal distribution and $\tau_{12.3}$

Agresti (1977) has drawn attention to the fact that for the trivariate normal distribution, $\tau_{12.3} \neq 0$ when $\rho_{12.3} = 0$, except for the trivial case where $\rho_{13} = 0$ or $\rho_{23} = 0$ (Agresti also included ρ_{13} or ρ_{23} equals 1 or -1; but $\tau_{12.3}$ is undefined for these values). Korn (1984) has calculated that for the trivariate normal distribution and $\rho_{12.3} = 0$, the values of $\tau_{12.3}$ move between a range of ± 0.293. It is thus clear that $H_o' : X_1$ and X_2 are conditionally, given X_3, independent of each other, does not imply $H_o : \tau_{12} = \tau_{13}\tau_{23}$. Confusion over this point led Shirahata (1977) to claim that the numerator of $t_{12.3}$ is a consistent estimator of zero under H_o' providing that the random variables are continuous. This is clearly incorrect as the results for the trivariate normal distribution show. The claim is accurate, as demonstrated by eqn. (6.14), if H_o' is replaced by H_o.

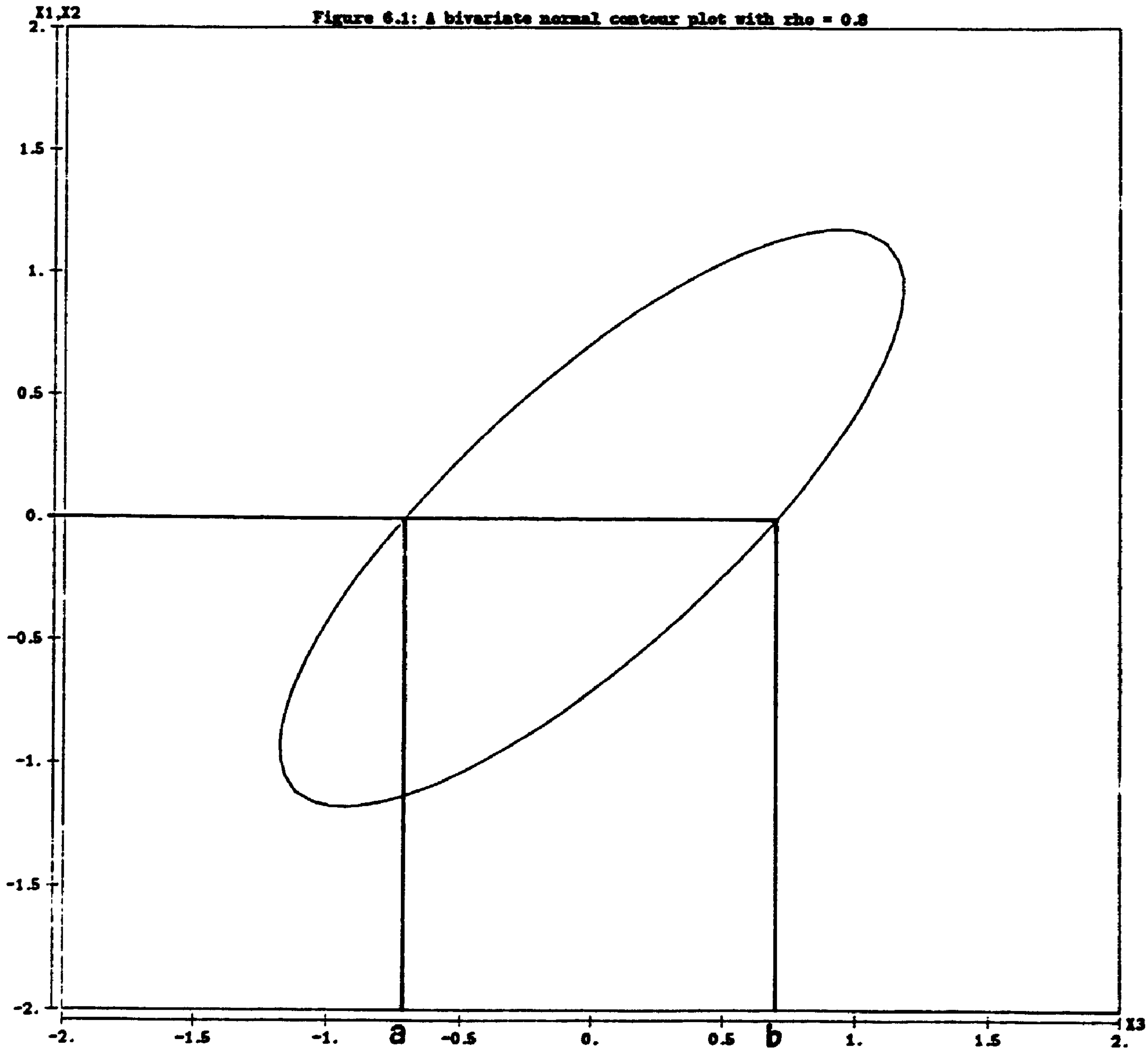


Figure 6.1: A bivariate normal contour plot with rho = 0.8
X1,X2
2.
1.5
1.
0.5
0.
-0.5
-1.
-1.5
-2.
-2.
-1.5
-1.
a
-0.5
0.
0.5
b
1.
1.5
2.
X3

With reference to Table 4.5, Kendall's argument requires that $a/b = c/d$ which implies that $\Pr(s_{gh}^{13} = \pm | s_{gh}^{23}) = \Pr(s_{gh}^{13} = \pm)$. Now suppose that (X_1, X_2, X_3) have a trivariate normal distribution with zero mean and such that the bivariate marginal distributions of (X_1, X_3) and (X_2, X_3) are identical. Consider the set of observations falling within the contour shown in Figure 6.1. Note that observations fall within this contour with equal probability. It follows from the figure that if $x_{3g} < a$ and $x_{3h} > b$ or if $x_{3h} < a$ and $x_{3g} > b$ then $s_{gh}^{13} = s_{gh}^{23} = s_{gh}^{12} = 1$ so that s_{gh}^{13} is perfectly correlated with s_{gh}^{23}. For Figure 6.1 and $\rho_{12.3} = 0$, it is found from eqn. (4.63) that $\tau_{12.3} > 0$ so that there is an excess of positive s_{gh}^{12} scores, above those predicted by Kendall's argument, which concurs with what is to be intuitively expected given the region of perfect correlation between s_{gh}^{13} and s_{gh}^{23}. More generally, for the trivariate normal distribution and $\rho_{12.3} = 0$, $\tau_{12.3} > 0$ for $\rho_{13}/\rho_{23} > 0$ and $\tau_{12.3} < 0$ for $\rho_{13}/\rho_{23} < 0$ and the corresponding excess of positive or negative s_{gh}^{12} scores always concurs with the value of s_{gh}^{12} associated with the region of perfect correlation between s_{gh}^{13} and s_{gh}^{23}. This is a feature of the trivariate normal distribution which relates to the concept of magnitude and which Kendall's method of assigning scores does not take into account.

Considering $\tau_{12.3}$ within the framework of category (ii) further demonstrates this aspect of H_o. The same residual *fractional scores* are obtained for $x_{1g} < x_{1h}$ regardless of $|x_{1g} - x_{1h}|$. As Wilson (1974) in his critique of proposals for multivariate analysis with ordinal variables based on analogies with product moment formulae has stated, "But it is not at all clear what such fractional pair scores might mean empirically in the ordinal case, nor even how they might be related to an empirical interpretation". For normal random variables, which are interval scale variables, it is quite clear that these residual fractional scores provide an inadequate basis for testing H_o', the relevant null hypothesis.

Table 6.3 shows, the fraction of rejects, under H_o', for 300 simulations of 50 observations on each of X_1, X_2 and X_3 at a significance level of 0.05. The data was generated as described in Section 6.1 except that ρ_{12} was set equal to $\rho_{13}\rho_{23}$. Note that the observed value of α is always close to 0.05 for ρ_{13} or ρ_{23} equal to zero, which concurs with an absence of a region of perfect correlation between s_{gh}^{13}

Table 6.3: Fraction, α, of rejects, under H_o', for varying τ_{13} and τ_{23} [1]

τ_{13}	τ_{23}	t_{13}	t_{23}	t_{12}	$t_{12.3}$	$Var(t_{12.3})$	$\hat{\sigma}^2(t_{12.3})$	α
-0.8	-0.8	-0.80	-0.80	0.72	0.22	0.0047	0.0058	0.910
-0.8	-0.6	-0.80	-0.60	0.56	0.17	0.0045	0.0049	0.720
-0.8	-0.4	-0.80	-0.41	0.38	0.11	0.0038	0.0041	0.383
-0.8	-0.2	-0.80	-0.19	0.19	0.050	0.0038	0.0039	0.140
-0.8	0.0	-0.80	0.01	-0.01	-0.002	0.0037	0.0039	0.043
-0.8	0.2	-0.80	0.20	-0.19	-0.054	0.0037	0.0039	0.133
-0.8	0.4	[illegible]	0.41	-0.38	-0.11	0.0043	0.0041	0.360
-0.8	0.6	-0.80	0.61	-0.56	-0.16	0.0049	0.0045	0.703
-0.8	0.8	-0.80	0.80	-0.72	-0.23	0.0050	0.0060	0.933
-0.6	-0.6	-0.60	-0.60	0.45	0.14	0.0060	0.0059	0.443
-0.6	-0.4	-0.60	-0.41	0.32	0.11	0.0054	0.0060	0.280
-0.6	-0.2	-0.60	-0.21	0.18	0.064	0.0059	0.0062	0.113
-0.6	0.0	-0.60	0.00	0.00	-0.007	0.0076	0.0062	0.090
-0.6	0.2	-0.60	0.20	-0.16	-0.056	0.0061	0.0062	0.100
-0.6	0.4	-0.60	0.40	-0.32	-0.11	0.0067	0.0061	0.290
-0.6	0.6	-0.60	0.61	-0.46	-0.15	0.0061	0.0059	0.493
-0.6	0.8	-0.60	0.80	-0.55	-0.16	0.0043	0.0048	0.677
-0.4	-0.4	-0.40	-0.40	0.23	0.080	0.0079	0.0073	0.177
-0.4	-0.2	-0.40	-0.19	0.10	-0.031	0.0074	0.0076	0.050
-0.4	0.0	-0.40	-0.01	0.02	0.016	0.0082	0.0081	0.050
-0.4	0.2	-0.41	0.20	-0.12	-0.041	0.0087	0.0079	0.097
-0.4	0.4	-0.40	0.40	-0.23	-0.077	0.0078	0.0071	0.180
-0.4	0.6	-0.39	0.60	-0.31	-0.11	0.0062	0.0059	0.253
-0.4	0.8	-0.41	0.81	-0.39	-0.11	0.0037	0.0040	0.410
-0.2	-0.2	-0.20	-0.20	0.07	0.031	0.0094	0.0089	0.090
-0.2	0.0	-0.21	0.00	0.01	0.012	0.0085	0.0090	0.067
-0.2	0.2	-0.21	0.20	-0.07	-0.024	0.0089	0.0087	0.073
-0.2	0.4	-0.20	0.40	-0.11	-0.035	0.0081	0.0078	0.077
-0.2	0.6	-0.20	0.59	-0.16	-0.055	0.0058	0.0064	0.107
-0.2	0.8	-0.20	0.80	-0.19	-0.054	0.0038	0.0038	0.147
0.0	0.0	-0.01	0.00	0.00	-0.001	0.0098	0.0094	0.060
0.0	0.2	-0.01	0.20	0.00	-0.001	0.0098	0.0091	0.057
0.0	0.4	0.00	0.39	-0.01	-0.005	0.0087	0.0077	0.060
0.0	0.6	0.01	0.60	0.01	-0.001	0.0059	0.0061	0.043
0.0	0.8	-0.01	0.80	-0.01	0.003	0.0044	0.0039	0.053
0.2	0.2	0.20	0.21	0.07	0.024	0.0092	0.0088	0.060
0.2	0.4	0.21	0.40	0.13	0.049	0.0085	0.0077	0.113
0.2	0.6	0.21	0.60	0.16	0.048	0.0071	0.0062	0.107
0.2	0.8	0.20	0.80	0.19	0.050	0.0033	0.0039	0.093
0.4	0.4	0.39	0.41	0.23	0.084	0.0068	0.0074	0.143
0.4	0.6	0.40	0.59	0.31	0.099	0.0061	0.0061	0.233
0.4	0.8	0.40	0.80	0.38	0.11	0.0039	0.0040	0.407
0.6	0.6	0.60	0.60	0.45	0.15	0.0064	0.0058	0.500
0.6	0.8	0.60	0.80	0.56	0.16	0.0046	0.0046	0.673
0.8	0.3	0.80	0.80	0.72	0.22	0.0053	0.0058	0.920

[1] sample values $t_{13}, \ldots, t_{12.3}$ are means of 300 simulations
$\hat{\sigma}^2(t_{12.3})$ is the mean of 300 sample estimates obtained from eqn. (6.15)

and s_{gk}^{23}

6.4.2 Ordinal data and $\tau_{12.3}$

Proponents of the extension of product moment correlation and regression analysis to ordinal measurements sought to provide a framework for the multivariate analysis of ordinal data. Hawkes (1971) noted that, "None of the operations of arithmetic except those of equality, greater than, and less than apply to them;" while Ploch (1974) stated, "Association and prediction are between the direction of difference of two observations. This point cannot be overstressed". It is therefore appropriate to address the question of how well H_o represents H_o' for such data.

Since the probability model developed in Chapter 5 represents a probability model on the permutations of the first n natural integers it provides a perfect tool for our purposes. Let R_3 be arranged in the natural order, generate two sets of inversion vectors and use these to obtain R_1 and R_2. If the $U(0,1)$ random variables used in the generation of R_1 are distinct from those used in the generation of R_2, it follows that R_1 and R_2 are conditionally independent of each other. Consequently, the data generated conform to H_o' and, unlike data generated from eqn. (4.63), do not possess interval scale characteristics. A potential limitation is that ties are excluded.

Table 6.4 displays the same information as does Table 6.3 except that the data are simulated as described above. The results, although much more moderate than those obtained in Table 6.3, once again suggest that, for large underlying correlations, $t_{12.3}$ is not an appropriate statistic for testing H_o'.

6.5 Conclusion

Having demonstrated that $t_{12.3}$ is not a consistent estimator of zero under conditional independence for the trivariate normal distribution, Agresti (1977) concluded that weighted average type measures would be considered superior to $t_{12.3}$

Table 6.4: Fraction, α, of rejects, under H_o', for varying τ_{13} and τ_{23} [1]

τ_{13}	τ_{23}	t_{13}	t_{23}	t_{12}	$t_{12.3}$	$Var(t_{12.3})$	$\hat{\sigma}^2(t_{12.3})$	α
-0.8	-0.8	-0.80	-0.80	0.68	0.11	0.0075	0.0083	0.157
-0.8	-0.6	-0.80	-0.60	0.53	0.11	0.0069	0.0069	0.217
-0.8	-0.4	-0.80	-0.40	0.37	0.089	0.0059	0.0062	0.147
-0.8	-0.2	-0.80	-0.20	0.18	0.038	0.0051	0.0053	0.063
-0.8	0.0	-0.81	-0.01	0.01	-0.007	0.0057	0.0049	0.067
-0.8	0.2	-0.80	0.20	-0.19	-0.039	0.0060	0.0053	0.080
-0.8	0.4	-0.80	0.40	-0.36	-0.078	0.0054	0.0060	0.157
-0.8	0.6	-0.80	0.60	-0.53	-0.11	0.0067	0.0070	0.227
-0.8	0.8	-0.80	0.80	-0.68	-0.10	0.0085	0.0080	0.157
-0.6	-0.6	-0.61	-0.60	0.44	0.11	0.0077	0.0073	0.217
-0.6	-0.4	-0.61	-0.40	0.30	0.084	0.0058	0.0068	0.153
-0.6	-0.2	-0.60	-0.20	0.16	0.054	0.0073	0.0068	0.110
-0.6	0.0	-0.60	0.01	-0.01	0.001	0.0067	0.0066	0.063
-0.6	0.2	-0.60	0.20	-0.16	-0.052	0.0070	0.0069	0.107
-0.6	0.4	-0.60	0.40	-0.31	-0.089	0.0062	0.0072	0.153
-0.6	0.6	-0.61	0.61	-0.44	-0.11	0.0084	0.0073	0.257
-0.6	0.8	-0.59	0.80	-0.52	-0.10	0.0069	0.0069	0.207
-0.4	-0.4	-0.41	-0.41	0.23	0.077	0.0085	0.0078	0.150
-0.4	-0.2	-0.40	-0.19	0.12	0.044	0.0080	0.0081	0.113
-0.4	0.0	-0.41	0.00	0.00	-0.005	0.0079	0.0080	0.050
-0.4	0.2	-0.39	0.20	-0.12	-0.047	0.0079	0.0080	0.093
-0.4	0.4	-0.40	0.40	-0.22	-0.074	0.0076	0.0077	0.127
-0.4	0.6	-0.41	0.60	-0.32	-0.10	0.0075	0.0070	0.230
-0.4	0.8	-0.40	0.80	-0.36	-0.079	0.0060	0.0058	0.153
-0.2	-0.2	-0.21	-0.20	0.07	0.027	0.0087	0.0088	0.073
-0.2	0.0	-0.21	0.00	0.00	0.002	0.010	0.0089	0.080
-0.2	0.2	-0.21	0.20	-0.07	-0.034	0.0091	0.0086	0.077
-0.2	0.4	-0.20	0.39	-0.12	-0.048	0.0092	0.0081	0.127
-0.2	0.6	-0.20	0.60	-0.16	-0.054	0.0063	0.0069	0.077
-0.2	0.8	-0.20	0.80	-0.19	-0.040	0.0055	0.0053	0.070
0.0	0.0	-0.01	0.00	-0.01	-0.004	0.011	0.0093	0.077
0.0	0.2	0.00	0.20	0.00	0.000	0.0082	0.0089	0.063
0.0	0.4	0.00	0.40	0.01	0.013	0.0094	0.0082	0.073
0.0	0.6	0.00	0.61	0.00	0.003	0.0060	0.0066	0.053
0.0	0.8	0.00	0.80	0.00	0.002	0.0056	0.0050	0.053
0.2	0.2	0.20	0.20	0.07	0.027	0.0085	0.0089	0.063
0.2	0.4	0.19	0.39	0.11	0.045	0.0077	0.0081	0.080
0.2	0.6	0.20	0.60	0.16	0.053	0.0068	0.0070	0.100
0.2	0.8	0.19	0.80	0.18	0.044	0.0056	0.0053	0.083
0.4	0.4	0.39	0.40	0.22	0.070	0.0068	0.0076	0.117
0.4	0.6	0.41	0.60	0.30	0.088	0.0074	0.0070	0.190
0.4	0.3	0.41	0.80	0.37	0.086	0.0061	0.0059	0.177
0.6	0.6	0.60	0.60	0.42	0.11	0.0066	0.0070	0.203
0.6	0.8	0.60	0.79	0.53	0.010	0.0067	0.0069	0.153
0.8	0.8	0.80	0.80	0.67	0.092	0.0073	0.0079	0.113

[1] sample values $t_{13}, \ldots, t_{12.3}$ are means of 300 simulations
$\hat{\sigma}^2(t_{12.3})$ is the mean of 300 sample estimates obtained from eqn. (6.15)

in terms of measuring the bivariate ordinal association between X_1 and X_2, after removing the influence of X_3. Korn (1984) considered different weighting schemes and their relative merits for a weighted sum of Kendall's tau across blocks; a block being determined by replications of the X_3 variable whose effect is to be partialled out. Taylor (1987) compared the use of a weighted sum of Kendall's τ's with a weighted sum of Spearman's ρ's and concluded, from a Monte Carlo study, that the two have essentially the same power with the optimal choice of weights.

These weighted average type measures possess three limitations. Firstly, they require that X_3 be a categorical variable. Secondly, they discard potentially valuable information because they make no data comparisons across blocks. Thirdly, and perhaps more importantly, they do not facilitate the development of a framework for the multivariate analysis of ordinal data.

It is clear from Section 6.4 that a fundamental limitation of partial tau is the fact that the method of scoring does not incorporate the concept of magnitude. Spearman's rho provides a nonparametric correlation coefficient which overcomes this deficiency and therefore the use of a partial Spearman's rho as an alternative to partial tau is next considered in Chapter 7.

REFERENCES

Agresti, A. (1977). Considerations in measuring partial association for ordinal categorical data. *Journal of the American Statistical Association*, **72**, 37–45.

Daniels, H.E. and Kendall, M.G. (1947). The significance of rank correlations where parental correlation exists. *Biometrika*, **34**, 197–208.

Hawkes, R.K. (1971). The multivariate analysis of ordinal measures. *American Journal of Sociology*, **76**, 908–926.

Hoeffding, W. (1947). On the distribution of the rank correlation coefficient τ when the variates are not independent. *Biometrika*, **34**, 183–196.

Kendall, M.G. (1942). Partial rank correlation. *Biometrika*, **32**, 277-283.

Kendall, M.G. (1975). *Rank Correlation Methods* (4th ed). Griffin and Co. Ltd.

Kennedy, W.J. and Gentle, J.E. (1980). *Statistical Computing.* Dekker.

Korn, E.L. (1984). The ranges of limiting values of some partial correlations under conditional independence. *The American Statistican*, **38**, 61–62.

Korn, E.L. (1984). Kendall's tau with a blocking variable. *Biometrics*, **40**, 209–214.

Moran, P.A.P. (1950). Recent developments in ranking theory. *Journal of the Royal Statistical Society*, B, **12**, 153–162.

Noether, G.E. (1967). *Elements of Nonparametric Statistics.* New York: Wiley.

Ploch, D.R. (1974). Ordinal measures of association and the general linear model. *Measurement in the Social Sciences*, ed. H.M. Blalock, Ch. 12. Aldine.

Quade, D. (1974). Nonparametric partial correlation. *Measurement in the Social Sciences*, ed. H.M. Blalock, Ch. 13. Aldine.

Schemper, M. (1987). One- and two-sample tests of Kendall's τ. *Biometrical Journal*, **29**, 1003–1009.

Shirahata, S. (1977). Tests of partial correlation in a linear model. *Biometrika*, **64**, 162–164.

Taylor, J.M.G. (1987). Kendall's and Spearman's correlation coefficients in the presence of a blocking variable. *Biometrics*, **43**, 409–416.

Wilson, T.P. (1974). On interpreting ordinal analogies to multiple regression and path analysis. *Social Forces*, **53**, 196–199.

Chapter 7

A PARTIAL RANK CORRELATION TEST BASED ON SPEARMAN'S RHO

Somers (1959) demonstrated that a generalized partial correlation coefficient can be defined as a simple extension of Daniels's (1944) generalized correlation coefficient. Now Kendall's (1942) argument, which defines partial τ in connection with independence in a fourfold table, appears to allow a more fundamental interpretation of partial τ. However, it has been noted in Section 4.6 that Kendall's argument leads to a test of $H_o : \tau_{12} = \tau_{13}\tau_{23}$ as the null hypothesis. That H_o is inadequate for measuring the bivariate correlation between X_1 and X_2, independently of the influence of X_3, has been established in Section 6.4. Thus there is no reason why analysis should not be based on partial Spearman's ρ_s.

Consider some potential advantages associated with using partial Spearman's ρ_s, which is denoted by $\rho_{s,12.3}$. Firstly, Durbin and Stuart (1951) have shown that Spearman's ρ_s, by using the sum of weighted inversions with the weight being the numerical difference between the inverted ranks, incorporates the concept of magnitude. In contrast, Kendall's τ assigns equal weight to all inversions regardless of the numerical difference between inverted ranks. Secondly, for the trivariate normal distribution and $\rho_{12.3} = 0$, Korn (1984) has calculated that $\rho_{s,12.3}$ moves between a range of ± 0.012, as opposed to a range of ± 0.293 for $t_{12.3}$. Consequently, although $r_{s,12.3}$ is not a consistent estimator of zero, it is a consistent estimator of a quantity close to zero and therefore it may be expected to yield reasonable inferences regardless of the magnitude of the underlying bivariate correlations. Thirdly, Kendall (1975) has noted that the product-moment correlation coefficient between variate values and their associated ranks is generally quite high. He has suggested that in virtue of the fairly close relationship between ranks and variates it may be expected that, if variate values were replaced by ranks and then the latter operated on as if they were the primary variates, in many cases the same conclusions should be drawn. Reinforcing this notion, Conover and Iman (1981) have presented many of the more useful and powerful nonparametric procedures in a unified manner by treating them as rank transformation procedures. They concluded that this technique of applying parametric procedures to ranks instead of to the original data should be viewed as a useful tool for developing nonparametric procedures to solve new problems. The use of a partial rank correlation procedure based on Spearman's ρ_s follows naturally from these arguments.

7.1 Hypothesis tests with $r_{s,12.3}$

It is reasonable to propose that the asymptotic distribution of $r_{s,12.3}$, under the complete null hypothesis, is the same as that of r_s under independence. However, we are again faced with the situation that distributional results obtained under the complete null hypothesis do not necessarily apply to the general null hypothesis that $\rho_{s,12.3} = 0$. It is shown that for continuous random variables, $\mathcal{E}(r_{s,12.3}) = 0 + O(n^{-1})$ under $H_o : \rho_{s,12} = \rho_{s,13}\rho_{s,23}$. The form of calculations involved in obtaining this result suggests that $r_{s,12.3}$ is asymptotically normally distributed. Derivation of a variance estimator poses an immediate problem. The asymptotic variance of $r_{s,12.3}$ under the complete null hypothesis and sample estimates of the variance based on 1000 simulations are used in our simulation studies. A comparison of the results obtained using these two variance estimates then permits an assessment of the severity of the effect due to using the complete null hypothesis instead of H_o.

Daniels (1951) presented a straightforward derivation of $E(r_s)$ based on the defining equation,

$$r_s = \frac{3}{n(n^2-1)}\left(\sum_{i\neq j}^{n} a_{ij}b_{ij} + \sum_{i\neq j\neq k}^{n} a_{ij}b_{ik}\right) \tag{7.1}$$

where $a_{ij} = \text{sign}(x_{1i} - x_{1j})$ and $b_{ij} = \text{sign}(x_{2i} - x_{2j})$. Thus

$$r_s = \frac{3}{(n+1)}t + \frac{(n-2)}{(n+1)}u \tag{7.2}$$

where $u = \left(3\sum_{i\neq j\neq k}^{n} a_{ij}b_{ik}\right)/n(n-1)(n-2)$. It is trivially shown, via use of indicator random variables as in Chapter 6, that $\mathcal{E}(u) = \rho_s$ for large N, the population size. Therefore,

$$\mathcal{E}(r_s) = \frac{3}{(n+1)}\tau + \frac{(n-2)}{(n+1)}\rho_s \tag{7.3a}$$

and

$$\mathcal{V}ar(r_s) = \frac{9}{(n+1)^2}\mathcal{V}ar(t) + \frac{(n-2)^2}{(n+1)^2}\mathcal{V}ar(u) + \frac{6(n-2)}{(n+1)^2}\mathcal{C}ov(t,u)\ . \tag{7.3b}$$

Expanding any of the variance or covariance terms on the right hand side of eqn. (7.3b) and taking its expectation leads to an order (1/n) term so that $\mathcal{V}ar(r_s) = O(n^{-1})$. Similarly, $\mathcal{C}ov(r_s^{12}, r_s^{13}) = O(n^{-1})$ and therefore the result that $\mathcal{E}(r_{s,12.3}) = 0 + O(n^{-1})$, under H_o, is obtained in exactly the same way in which the corresponding result, eqn. (6.14), was obtained for $t_{12.3}$.

The dominant terms of the moments of r_s are determined by the moments of u. Apart from the fact that u incorporates a sum over three subscripts in its definition while t uses a sum over two subscripts, the algebra of obtaining the moments of u is the same as the algebra used to obtain the moments of t. Therefore, although no attempt has been made at verification, it is assumed that $r_{s,12.3}$ is asymptotically normally distributed. Consequently, hypothesis tests of H_o are implemented by referring $r_{s,12.3}$, appropriately standardized, to the $N(0,1)$ distribution. Table 7.1 shows the fractions, α_a and α_b, of rejects for 1000 simulations of 50 observations on each of X_1, X_2 and X_3 at a significance level of 0.05. The data are generated from a trivariate normal distribution, α_a is obtained using the complete null hypothesis asymptotic variance estimate of $(n-1)^{-1}$, and α_b is obtained using the sample estimate of the variance as computed from the 1000 simulated

Table 7.1: Fraction of rejects, under H_o, for varying $\rho_{s,13}$ and $\rho_{s,23}$ [1]

$\rho_{s,13}$	$\rho_{s,23}$	$r_{s,13}$	$r_{s,23}$	$r_{s,12.3}$	$Var(r_{s,12.3})$	α_a	α_b
-0.8	-0.8	-0.78	-0.78	0.023	0.023	0.043	0.035
-0.8	-0.6	-0.78	-0.58	0.012	0.033	0.054	0.018
-0.8	-0.4	-0.78	-0.38	0.007	0.030	0.064	0.022
-0.8	-0.2	-0.78	-0.21	0.000	0.029	0.043	0.020
-0.8	0.0	-0.79	0.00	0.000	0.025	0.051	0.030
-0.8	0.2	-0.78	0.19	-0.009	0.032	0.053	0.023
-0.8	0.4	-0.78	0.39	-0.004	0.025	0.072	0.056
-0.8	0.6	-0.78	0.58	-0.011	0.026	0.065	0.039
-0.8	0.8	-0.77	0.78	-0.018	0.027	0.072	0.037
-0.6	-0.6	-0.58	-0.58	0.002	0.030	0.048	0.018
-0.6	-0.4	-0.58	-0.39	0.005	0.032	0.060	0.025
-0.6	-0.2	-0.59	-0.19	0.005	0.032	0.060	0.023
-0.6	0.0	-0.58	0.00	0.000	0.025	0.049	0.039
-0.6	0.2	-0.58	0.19	0.004	0.023	0.056	0.049
-0.6	0.4	-0.59	0.39	-0.005	0.029	0.059	0.026
-0.6	0.6	-0.58	0.58	-0.008	0.025	0.051	0.037
-0.6	0.8	-0.58	0.77	-0.003	0.030	0.055	0.032
-0.4	-0.4	-0.39	-0.39	0.005	0.028	0.048	0.027
-0.4	-0.2	-0.38	-0.20	0.010	0.036	0.046	0.017
-0.4	0.0	-0.39	0.01	-0.001	0.030	0.062	0.027
-0.4	0.2	-0.39	0.20	0.010	0.027	0.047	0.031
-0.4	0.4	-0.38	0.39	-0.006	0.032	0.044	0.017
-0.4	0.6	-0.39	0.58	-0.007	0.033	0.054	0.018
-0.4	0.8	-0.39	0.78	-0.009	0.031	0.062	0.022
-0.2	-0.2	-0.19	-0.20	0.000	0.026	0.035	0.025
-0.2	0.0	-0.21	0.00	-0.006	0.030	0.055	0.022
-0.2	0.2	-0.20	0.20	0.002	0.032	0.046	0.021
-0.2	0.4	-0.19	0.39	-0.002	0.027	0.054	0.025
-0.2	0.6	-0.20	0.58	0.002	0.026	0.051	0.035
-0.2	0.8	-0.19	0.78	0.000	0.030	0.045	0.020
0.0	0.0	0.00	-0.01	-0.004	0.026	0.056	0.034
0.0	0.2	0.00	0.19	0.001	0.025	0.063	0.045
0.0	0.4	-0.01	0.39	-0.001	0.036	0.063	0.022
0.0	0.6	0.01	0.59	-0.007	0.028	0.036	0.022
0.0	0.8	0.00	0.78	0.007	0.037	0.051	0.015
0.2	0.2	0.20	0.19	-0.005	0.028	0.039	0.017
0.2	0.4	0.19	0.39	0.002	0.027	0.046	0.033
0.2	0.6	0.19	0.59	-0.002	0.032	0.057	0.020
0.2	0.8	0.18	0.78	0.009	0.033	0.041	0.021
0.4	0.4	0.39	0.39	0.006	0.033	0.043	0.017
0.4	0.6	0.38	0.59	-0.002	0.039	0.061	0.014
0.4	0.8	0.39	0.78	0.006	0.025	0.051	0.036
0.6	0.6	0.59	0.59	0.008	0.029	0.061	0.033
0.6	0.8	0.58	0.78	0.013	0.027	0.051	0.033
0.8	0.8	0.78	0.78	0.022	0.029	0.079	0.036

[1] sample values $r_{s,13}, \ldots, r_{s,12.3}$ are means of 1000 simulations

values of $r_{s,12.3}$. Eqn. (4.63), which relates τ to ρ for a bivariate normal distribution, is replaced with the relation $\rho = 2\sin(\pi\rho_s/6)$ which connects ρ_s to ρ. Table 7.2 shows the same information as Table 7.1 except that the data are generated under H_o' so that $\rho_{12} = \rho_{13}\rho_{23}$.

Table 7.2: Fraction of rejects, under H_o', for varying $\rho_{s,13}$ and $\rho_{s,23}$ [1]

$\rho_{s,13}$	$\rho_{s,23}$	$r_{s,13}$	$r_{s,23}$	$r_{s,12.3}$	$Var(r_{s,12.3})$	α_a	α_b
-0.8	-0.8	-0.78	-0.78	0.033	0.023	0.059	0.044
-0.8	-0.6	-0.78	-0.58	0.023	0.033	0.063	0.018
-0.8	-0.4	-0.78	-0.38	0.015	0.030	0.064	0.024
-0.8	-0.2	-0.78	-0.21	0.004	0.029	0.045	0.019
-0.8	0.0	-0.79	0.00	0.000	0.025	0.051	0.030
-0.8	0.2	-0.78	0.19	-0.013	0.032	0.053	0.023
-0.8	0.4	-0.78	0.39	-0.012	0.025	0.071	0.054
-0.8	0.6	-0.78	0.58	-0.022	0.026	0.068	0.045
-0.8	0.8	-0.77	0.78	-0.029	0.027	0.071	0.040
-0.6	-0.6	-0.58	-0.58	0.012	0.030	0.047	0.013
-0.6	-0.4	-0.58	-0.39	0.013	0.032	0.061	0.023
-0.6	-0.2	-0.59	-0.19	0.009	0.033	0.062	0.025
-0.6	0.0	-0.58	0.00	0.000	0.025	0.049	0.039
-0.6	0.2	-0.58	0.19	0.000	0.023	0.056	0.049
-0.6	0.4	-0.59	0.39	-0.013	0.029	0.063	0.024
-0.6	0.6	-0.58	0.58	-0.018	0.025	0.051	0.038
-0.6	0.8	-0.58	0.77	-0.013	0.030	0.058	0.031
-0.4	-0.4	-0.39	-0.39	0.011	0.028	0.050	0.028
-0.4	-0.2	-0.38	-0.20	0.014	0.036	0.047	0.018
-0.4	0.0	-0.39	0.01	-0.001	0.030	0.062	0.027
-0.4	0.2	-0.39	0.20	0.007	0.028	0.047	0.032
-0.4	0.4	-0.38	0.39	-0.012	0.032	0.049	0.017
-0.4	0.6	-0.39	0.58	-0.014	0.034	0.050	0.020
-0.4	0.8	-0.39	0.78	-0.017	0.031	0.061	0.024
-0.2	-0.2	-0.19	-0.20	0.002	0.026	0.037	0.025
-0.2	0.0	-0.21	0.00	-0.006	0.030	0.055	0.022
-0.2	0.2	-0.20	0.20	0.000	0.032	0.044	0.021
-0.2	0.4	-0.19	0.39	-0.005	0.027	0.054	0.026
-0.2	0.6	-0.20	0.58	-0.002	0.026	0.050	0.036
-0.2	0.8	-0.19	0.78	-0.004	0.030	0.042	0.021
0.0	0.0	0.00	-0.01	-0.004	0.026	0.056	0.034
0.0	0.2	0.00	0.19	0.001	0.025	0.063	0.045
0.0	0.4	-0.01	0.39	-0.001	0.036	0.063	0.022
0.0	0.6	0.01	0.59	-0.007	0.028	0.036	0.022
0.0	0.8	0.00	0.78	0.007	0.037	0.051	0.015
0.2	0.2	0.20	0.19	-0.004	0.028	0.037	0.018
0.2	0.4	0.19	0.39	0.006	0.027	0.043	0.032
0.2	0.6	0.19	0.59	0.003	0.032	0.057	0.021
0.2	0.8	0.18	0.78	0.013	0.033	0.041	0.021
0.4	0.4	0.39	0.39	0.013	0.033	0.043	0.016
0.4	0.6	0.38	0.59	0.006	0.040	0.063	0.016
0.4	0.8	0.39	0.78	0.014	0.025	0.053	0.038
0.6	0.6	0.59	0.59	0.019	0.029	0.063	0.035
0.6	0.8	0.58	0.78	0.024	0.027	0.049	0.030
0.8	0.8	0.78	0.78	0.032	0.029	0.084	0.035

[1] sample values $r_{s,13}, \ldots, r_{s,12.3}$ are means of 1000 simulations

As discussed in Section 6.4, the probability model of Chapter 5 may be used

to generate data which conform to H_o'. It is assumed that, for a fixed value of τ, data generated using eqn. (5.9) correspond to a fixed value of ρ_s. Therefore, although the parent Spearman correlation coefficients are unknown, simulations

Table 7.3: Fraction of rejects, under H_o', for varying τ_{13} and τ_{23} [1]

τ_{13}	τ_{23}	$r_{s,13}$	$r_{s,23}$	$r_{s,12.3}$	$Var(r_{s,12.3})$	α_a	α_b
-0.8	-0.8	-0.90	-0.90	0.027	0.027	0.087	0.057
-0.8	-0.6	-0.90	-0.75	0.034	0.028	0.096	0.060
-0.8	-0.4	-0.90	-0.54	0.026	0.024	0.079	0.055
-0.8	-0.2	-0.90	-0.28	0.014	0.021	0.048	0.046
-0.8	0.0	-0.90	0.01	0.004	0.022	0.055	0.047
-0.8	0.2	-0.90	0.28	-0.020	0.022	0.061	0.047
-0.8	0.4	-0.90	0.55	-0.028	0.024	0.079	0.064
-0.8	0.6	-0.90	0.75	-0.039	0.028	0.094	0.056
-0.8	0.8	-0.90	0.90	-0.026	0.028	0.098	0.056
-0.6	-0.6	-0.76	-0.75	0.054	0.025	0.083	0.056
-0.6	-0.4	-0.75	-0.54	0.048	0.022	0.078	0.070
-0.6	-0.2	-0.76	-0.29	0.016	0.021	0.058	0.051
-0.6	0.0	-0.75	0.00	-0.005	0.021	0.047	0.045
-0.6	0.2	-0.75	0.28	-0.035	0.023	0.072	0.052
-0.6	0.4	-0.75	0.54	-0.040	0.024	0.082	0.060
-0.6	0.6	-0.75	0.75	-0.047	0.027	0.104	0.048
-0.6	0.8	-0.75	0.90	-0.029	0.027	0.090	0.057
-0.4	-0.4	-0.54	-0.54	0.029	0.021	0.061	0.058
-0.4	-0.2	-0.54	-0.28	0.018	0.022	0.061	0.050
-0.4	0.0	-0.54	-0.01	-0.010	0.020	0.042	0.042
-0.4	0.2	-0.55	0.29	-0.031	0.022	0.066	0.055
-0.4	0.4	-0.54	0.54	-0.040	0.024	0.073	0.049
-0.4	0.6	-0.54	0.75	-0.038	0.023	0.071	0.060
-0.4	0.8	-0.54	0.90	-0.028	0.025	0.074	0.045
-0.2	-0.2	-0.28	-0.28	0.021	0.021	0.052	0.050
-0.2	0.0	-0.28	0.01	-0.001	0.020	0.039	0.045
-0.2	0.2	-0.29	0.29	-0.013	0.020	0.044	0.046
-0.2	0.4	-0.29	0.54	-0.017	0.023	0.061	0.050
-0.2	0.6	-0.29	0.75	-0.024	0.023	0.061	0.047
-0.2	0.8	-0.28	0.90	-0.020	0.022	0.063	0.056
0.0	0.0	0.00	0.00	-0.002	0.022	0.058	0.051
0.0	0.2	0.00	0.28	0.002	0.022	0.055	0.048
0.0	0.4	0.01	0.54	-0.005	0.021	0.053	0.048
0.0	0.6	-0.01	0.75	0.002	0.022	0.059	0.049
0.0	0.8	-0.01	0.90	0.005	0.021	0.040	0.040
0.2	0.2	0.28	0.28	0.022	0.019	0.050	0.056
0.2	0.4	0.28	0.54	0.025	0.020	0.056	0.058
0.2	0.6	0.29	0.75	0.027	0.021	0.062	0.054
0.2	0.8	0.29	0.90	0.011	0.020	0.045	0.047
0.4	0.4	0.55	0.54	0.040	0.023	0.068	0.054
0.4	0.6	0.54	0.75	0.036	0.024	0.079	0.057
0.4	0.8	0.54	0.90	0.026	0.024	0.063	0.045
0.6	0.6	0.75	0.76	0.043	0.024	0.078	0.057
0.6	0.8	0.75	0.90	0.038	0.027	0.100	0.058
0.8	0.8	0.90	0.90	0.025	0.030	0.104	0.053

[1] sample values $r_{s,13}, \ldots, r_{s,12.3}$ are means of 1000 simulations

based on eqn. (5.9) provide a basis for further assessing the behaviour of $r_{s,12.3}$ under H_o'. Table 7.3 shows rejection rates for data generated using eqn. (5.9).

The sample values of $r_{s,13}$ and $r_{s,23}$ strongly indicate that, as assumed, there is a one to one correspondence between τ and ρ_s for data generated using eqn. (5.9). Discussion of the rejection rates shown in Tables 7.1 to 7.3 is deferred until the next section.

Table 7.4: Fraction of rejects, under H_o, with varying $\rho_{s,13}$ and $\rho_{s,23}$ [1]

$\rho_{s,13}$	$\rho_{s,23}$	$r'_{s,13}$	$r'_{s,23}$	$r'_{s,12.3}$	$Var(r'_{s,12.3})$	α_a	α_b
-0.8	-0.8	-0.79	-0.79	0.005	0.032	0.045	0.019
-0.8	-0.6	-0.79	-0.59	-0.006	0.028	0.049	0.036
-0.8	-0.4	-0.79	-0.38	0.000	0.035	0.058	0.023
-0.8	-0.2	-0.79	-0.21	0.002	0.031	0.039	0.021
-0.8	0.0	-0.80	0.00	-0.004	0.046	0.050	0.010
-0.8	0.2	-0.79	0.20	0.001	0.028	0.050	0.040
-0.8	0.4	-0.79	0.40	0.001	0.032	0.075	0.042
-0.8	0.6	-0.79	0.59	0.003	0.035	0.061	0.017
-0.8	0.8	-0.78	0.79	0.003	0.027	0.072	0.050
-0.6	-0.6	-0.59	-0.59	-0.002	0.030	0.041	0.024
-0.6	-0.4	-0.59	-0.40	-0.003	0.045	0.055	0.017
-0.6	-0.2	-0.60	-0.19	0.005	0.029	0.055	0.032
-0.6	0.0	-0.59	0.00	0.004	0.028	0.043	0.029
-0.6	0.2	-0.59	0.20	0.007	0.027	0.049	0.038
-0.6	0.4	-0.60	0.39	0.009	0.031	0.049	0.024
-0.6	0.6	-0.59	0.59	0.006	0.046	0.051	0.012
-0.6	0.8	-0.59	0.78	0.009	0.031	0.058	0.031
-0.4	-0.4	-0.39	-0.39	0.003	0.025	0.041	0.037
-0.4	-0.2	-0.39	-0.21	0.001	0.026	0.039	0.041
-0.4	0.0	-0.39	0.01	0.003	0.035	0.052	0.022
-0.4	0.2	-0.40	0.20	0.008	0.027	0.039	0.035
-0.4	0.4	-0.39	0.39	0.000	0.029	0.037	0.027
-0.4	0.6	-0.40	0.59	0.002	0.031	0.045	0.026
-0.4	0.8	-0.40	0.79	0.004	0.027	0.060	0.044
-0.2	-0.2	-0.19	-0.20	-0.001	0.028	0.032	0.023
-0.2	0.0	-0.21	0.00	-0.008	0.036	0.042	0.017
-0.2	0.2	-0.21	0.20	0.006	0.031	0.039	0.027
-0.2	0.4	-0.19	0.39	-0.005	0.025	0.049	0.037
-0.2	0.6	-0.20	0.59	0.008	0.025	0.042	0.040
-0.2	0.8	-0.20	0.79	0.008	0.031	0.042	0.027
0.0	0.0	0.00	-0.01	-0.010	0.040	0.045	0.010
0.0	0.2	0.00	0.20	0.002	0.030	0.052	0.028
0.0	0.4	-0.01	0.40	-0.002	0.026	0.057	0.049
0.0	0.6	0.01	0.60	-0.002	0.028	0.032	0.028
0.0	0.8	0.00	0.79	0.006	0.037	0.048	0.015
0.2	0.2	0.20	0.20	-0.006	0.029	[illegible].31	0.019
0.2	0.4	0.20	0.40	0.002	0.030	0.039	0.027
0.2	0.6	0.20	0.60	0.002	0.034	0.048	0.019
0.2	0.8	0.19	0.79	0.001	0.028	0.039	0.033
0.4	0.4	0.39	0.40	-0.005	0.033	0.035	0.020
0.4	0.6	0.39	0.60	0.002	0.035	0.055	0.027
0.4	0.8	0.40	0.79	-0.001	0.028	0.049	0.036
0.6	0.6	0.60	0.60	-0.001	0.030	0.058	0.035
0.6	0.8	0.59	0.79	0.001	0.026	0.050	0.045
0.8	0.8	0.79	0.79	0.004	0.046	0.070	0.014

[1] sample values $r'_{s,13}, \ldots, r'_{s,12.3}$ are means of 1000 simulations

7.2 Hypothesis tests with $r'_{s,12.3}$

Eqn. (7.3a) shows that r_s is a biased estimator of ρ_s. To compensate for this bias it is suggested that the test statistic $r'_{s,12.3}$, where

$$r'_{s,12.3} = \frac{r'_{s,12} - r'_{s,13}r'_{s,23}}{\sqrt{(1 - {r'_{s,13}}^2)(1 - {r'_{s,23}}^2)}} \tag{7.4a}$$

and

$$r'_{s,ij} = \frac{(n+1)}{(n-2)}\left(r_{s,ij} - \frac{3}{(n+1)}\tau_{ij}\right), \tag{7.4b}$$

be conside. ed. It follows as before that $\mathcal{E}(r'_{s,12.3}) = 0 + O(n^{-1})$ under H_o. It also follows, from eqn. (7.4b), that the asymptotic variance estimate under the complete null hypothesis is $(n+1)^2/((n-1)(n-2)^2)$. Tables 7.4 to 7.6 display rejection rates for $r'_{s,12.3}$, corresponding to those shown in Tables 7.1 to 7.3, where the asymptotic variance $(n+1)^2/((n-1)(n-2)^2)$ is used to obtain α_a.

Our discussion of the rejection rates displayed in Tables 7.1 to 7.6 begins with a consideration of results obtained for data generated from the trivariate normal distribution. Of primary importance is the fact that, except for the first two rows and the last row of Tables 7.1 and 7.2, the rejection rates obtained under H_o do not appreciably differ from those obtained under H'_o. With the exception of only the first row of Tables 7.4 and 7.5, a similar result holds for these tables. Surprisingly, it is immediately evident that the values of α_a are more consistent with a significance level of 0.05 than are the values of α_b which, on the whole, are systematically lower than 0.05. This indicates that either the sample estimates of the variance are biased upward or that the normality assumption is erroneous.

It is difficult to recommend one of either $r_{s,12.3}$ or $r'_{s,12.3}$ as being superior to the other. While the α_a rates in the middle of Tables 7.1 and 7.2 are closer to 0.05 than are the corresponding rates in Tables 7.4 and 7.5, the rates at both ends of Tables 7.4 and 7.5 are closer to 0.05 than are the corresponding rates in Tables 7.1 and 7.2. However, it is quite clear from all four tables that either of the statistics

Table 7.5: Fraction of rejects, under H_o, with varying $\rho_{s,13}$ and $\rho_{s,23}$ [1]

$\rho_{s,13}$	$\rho_{s,23}$	$r'_{s,13}$	$r'_{s,23}$	$r'_{s,12.3}$	$Var(r'_{s,12.3})$	α_a	α_b
-0.8	-0.8	-0.79	-0.79	0.017	0.032	0.054	0.023
-0.8	-0.6	-0.79	-0.59	0.006	0.028	0.050	0.035
-0.8	-0.4	-0.79	-0.38	0.008	0.035	0.060	0.023
-0.8	-0.2	-0.79	-0.21	0.007	0.031	0.037	0.020
-0.8	0.0	-0.80	0.00	-0.004	0.046	0.050	0.010
-0.8	0.2	-0.79	0.20	-0.004	0.028	0.047	0.039
-0.8	0.4	-0.79	0.40	-0.008	0.031	0.072	0.043
-0.8	0.6	-0.79	0.59	-0.008	0.035	0.060	0.019
-0.8	0.8	-0.78	0.79	-0.009	0.027	0.069	0.048
-0.6	-0.6	-0.59	-0.59	0.009	0.030	0.043	0.023
-0.6	-0.4	-0.59	-0.40	0.005	0.045	0.051	0.017
-0.6	-0.2	-0.60	-0.19	0.009	0.029	0.053	0.032
-0.6	0.0	-0.59	0.00	0.004	0.028	0.043	0.029
-0.6	0.2	-0.59	0.20	0.003	0.027	0.051	0.040
-0.6	0.4	-0.60	0.39	0.004	0.031	0.051	0.023
-0.6	0.6	-0.59	0.59	-0.005	0.046	0.047	0.012
-0.6	0.8	-0.59	0.78	-0.002	0.031	0.056	0.031
-0.4	-0.4	-0.39	-0.39	0.010	0.025	0.042	0.036
-0.4	-0.2	-0.39	-0.21	0.004	0.026	0.039	0.041
-0.4	0.0	-0.39	0.01	0.003	0.035	0.052	0.022
-0.4	0.2	-0.40	0.20	0.004	0.027	0.039	0.034
-0.4	0.4	-0.39	0.39	-0.006	0.028	0.038	0.027
-0.4	0.6	-0.40	0.59	-0.006	0.031	0.043	0.026
-0.4	0.8	-0.40	0.79	-0.004	0.027	0.059	0.043
-0.2	-0.2	-0.19	-0.20	0.001	0.028	0.029	0.023
-0.2	0.0	-0.21	0.00	-0.008	0.036	0.042	0.017
-0.2	0.2	-0.21	0.20	0.004	0.031	0.042	0.026
-0.2	0.4	-0.19	0.39	-0.008	0.025	0.047	0.038
-0.2	0.6	-0.20	0.59	0.003	0.025	0.042	0.040
-0.2	0.8	-0.20	0.79	0.003	0.031	0.039	0.024
0.0	0.0	0.00	-0.01	-0.010	0.040	0.045	0.010
0.0	0.2	0.00	0.20	0.002	0.030	0.052	0.028
0.0	0.4	-0.01	0.40	-0.002	0.026	0.057	0.049
0.0	0.6	0.01	0.60	-0.002	0.028	0.032	0.028
0.0	0.8	0.00	0.79	0.006	0.037	0.048	0.015
0.2	0.2	0.20	0.20	-0.005	0.029	0.032	0.020
0.2	0.4	0.20	0.40	0.005	0.030	0.039	0.077
0.2	0.6	0.20	0.60	0.006	0.034	0.048	0.020
0.2	0.8	0.19	0.79	0.006	0.028	0.040	0.034
0.4	0.4	0.39	0.40	0.001	0.033	0.037	0.020
0.4	0.6	0.39	0.60	0.011	0.035	0.058	0.027
0.4	0.8	0.40	0.79	0.008	0.028	0.049	0.035
0.6	0.6	0.60	0.60	0.010	0.030	0.056	0.036
0.6	0.8	0.59	0.79	0.011	0.026	0.052	0.044
0.8	0.8	0.79	0.79	0.015	0.046	0.073	0.014

[1] sample values $r'_{s,13}, \ldots, r'_{s,12.3}$ are means of 1000 simulations

$r_{s,12.3}$ or $r'_{s,12.3}$ taken in conjunction with its asymptotic distribution, as obtained under the complete null hypothesis, provides an adequate basis for hypothesis tests of H'_o when the data are from a trivariate normal distribution.

Table 7.6: Fraction of rejects, under H_o', for varying τ_{13} and τ_{23} [1]

τ_{13}	τ_{23}	$r'_{s,13}$	$r'_{s,23}$	$r'_{s,12.3}$	$Var(r'_{s,12.3})$	α_a	α_b
-0.8	-0.8	-0.91	-0.91	0.009	0.032	0.089	0.048
-0.8	-0.6	-0.91	-0.76	0.020	0.032	0.104	0.051
-0.8	-0.4	-0.91	-0.55	0.017	0.028	0.077	0.052
-0.8	-0.2	-0.91	-0.28	0.010	0.024	0.050	0.045
-0.8	0.0	-0.91	0.01	0.004	0.025	0.058	0.045
-0.8	0.2	-0.91	0.29	-0.016	0.025	0.058	0.041
-0.8	0.4	-0.91	0.56	-0.019	0.028	0.084	0.059
-0.8	0.6	-0.91	0.76	-0.025	0.032	0.093	0.053
-0.8	0.8	-0.91	0.91	-0.009	0.033	0.096	0.054
-0.6	-0.6	-0.77	-0.76	0.043	0.028	0.070	0.050
-0.6	-0.4	-0.76	-0.55	0.041	0.025	0.072	0.060
-0.6	-0.2	-0.77	-0.29	0.012	0.023	0.051	0.051
-0.6	0.0	-0.76	0.00	-0.005	0.023	0.044	0.044
-0.6	0.2	-0.76	0.28	-0.032	0.025	0.067	0.049
-0.6	0.4	-0.76	0.55	-0.033	0.026	0.076	0.056
-0.6	0.6	-0.76	0.76	-0.036	0.030	0.090	0.040
-0.6	0.8	-0.76	0.91	-0.015	0.031	0.090	0.054
-0.4	-0.4	-0.54	-0.55	0.024	0.022	0.045	0.051
-0.4	-0.2	-0.55	-0.28	0.015	0.024	0.054	0.054
-0.4	0.0	-0.55	-0.01	-0.011	0.021	0.033	0.042
-0.4	0.2	-0.56	0.29	-0.029	0.023	0.054	0.053
-0.4	0.4	-0.55	0.55	-0.035	0.025	0.058	0.048
-0.4	0.6	-0.55	0.76	-0.031	0.025	0.068	0.055
-0.4	0.8	-0.55	0.91	-0.020	0.028	0.071	0.041
-0.2	-0.2	-0.29	-0.28	0.019	0.022	0.049	0.051
-0.2	0.0	-0.29	0.01	-0.001	0.021	0.029	0.047
-0.2	0.2	-0.29	0.29	-0.012	0.021	0.036	0.048
-0.2	0.4	-0.29	0.55	-0.014	0.024	0.054	0.049
-0.2	0.6	-0.29	0.76	-0.021	0.025	0.054	0.049
-0.2	0.8	-0.28	0.91	-0.016	0.026	0.067	0.057
0.0	0.0	0.00	0.00	-0.002	0.023	0.049	0.052
0.0	0.2	0.00	0.29	0.002	0.023	0.048	0.048
0.0	0.4	0.01	0.55	-0.006	0.022	0.046	0.046
0.0	0.6	-0.01	0.76	0.002	0.024	0.052	0.048
0.0	0.8	-0.01	0.91	0.005	0.023	0.041	0.041
0.2	0.2	0.29	0.29	0.021	0.020	0.041	0.056
0.2	0.4	0.29	0.55	0.023	0.021	0.049	0.059
0.2	0.6	0.29	0.76	0.024	0.023	0.056	0.055
0.2	0.8	0.29	0.91	0.007	0.023	0.047	0.048
0.4	0.4	0.56	0.55	0.035	0.025	0.059	0.053
0.4	0.6	0.55	0.76	0.028	0.026	0.069	0.056
0.4	0.8	0.55	0.91	0.017	0.028	0.064	0.046
0.6	0.6	0.76	0.77	0.031	0.027	0.071	0.051
0.6	0.8	0.76	0.91	0.025	0.031	0.101	0.055
0.8	0.8	0.91	0.91	0.007	0.035	0.106	0.050

[1] sample values $r'_{s,13}, \ldots, r'_{s,12.3}$ are means of 1000 simulations

Turning our attention to the rejection rates obtained for data generated using the probability model given in eqn. (5.9), it is found that the α_b values are more consistent with a significance level of 0.05 than are the α_a values. The α_b values in

Tables 7.3 and 7.6, with $0.04 \leq \alpha_b \leq 0.06$ for Table 7.6, are in excellent agreement with the desired value of 0.05 thus indicating that a test based on the assumption of normality for $r'_{s,12.3}$, with variance equal to the simulated variance, has close to its nominal size. It also suggests that data generated from eqn. (5.9) conform to H'_o. Hence the probability model, eqn. (5.9), provides a basis for further exploring the concept of conditionally independent rankings.

It is evident that the rejection rates, α_a, at either end of Table 7.3 or 7.6 are systematically higher than the desired value of 0.05, and that the discrepancy worsens as the magnitudes of both $r_{s,13}$ and ${}_{.3}$ increase. If attention is restricted to the rows for which at least one of $|r_{s,13}|$ and $|r_{s,23}|$ is less than 0.75, the rejection rates obtained in Table 7.6 are more compatible with a significance level of 0.05 than are those in Table 7.3. These rejection rates are also deemed to be satisfactory and therefore it is concluded that hypothesis tests of H'_o may be safely implemented by using a test of $r'_{s,12.3}$ under the complete null hypothesis subject to the proviso that at least one of $|r'_{s,13}|$ and $|r'_{s,23}|$ be less than or equal to 0.6. An obvious problem for further research is that of obtaining a variance estimator for $r'_{s,12.3}$ under H_o. Hopefully, this variance estimator will lead to better rejection rates across a wider range of $r_{s,13}$ and $r_{s,23}$ values, thereby leading to a less restrictive test of H'_o.

7.3 An application of $r'_{s,12.3}$ to trend analysis with seasonal data

In appraising the behaviour of $r'_{s,12.3}$ it is necessary to keep in mind two objectives in selecting a test for use in an exploratory study. Hirsch et al. (1982) describe these two objectives, assuming that a significance level α has already been selected, as: 1) The actual significance should be relatively close to α for stochastic processes relatively similar to the time series one expects to be testing, and 2) the power for detecting trends should be relatively high compared to some alternative tests for processes in which trend exists and which are thought to be similar to the time series one expects to be testing. The results of Sections 7.1 and 7.2 show that $r'_{s,12.3}$ will satisfy the first of these two objectives.

For our application of $r'_{s,12.3}$, the stochastic process with linear trend given by Hirsch et al. (1982, eqns. (14c) and (15)) is used to simulate data. The stochastic process comprises normal independent variates with a seasonal cycle and an additive linear trend,

$$v_{ij} = 0.5\epsilon_{ij} + A\sin\left(\frac{\pi}{3} + \frac{\pi}{6}\cdot i\right) + \beta\left(\frac{i}{12} + j\right) . \qquad (7.5)$$

The series is generated for $i = 1, 2, \ldots, 12$ and $j = 1, 2, \ldots, n$ so that the data extend over twelve seasons, one for each month of the year, and n years. Hirsch et al. used 500 repetitions at each of $n = 5, 10$ and 20 in their simulations. Nine different values of β are used for each value of n, these being $\beta = 0.0\ (0.05)\ 0.4$ for $n = 5$, $\beta = 0.0\ (0.02)\ 0.16$ for $n = 10$, and $\beta = 0.0\ (0.0065)\ 0.052$ for $n = 20$.

The parameter A is the amplitude of the seasonal effect and is used to study the robustness of $r'_{s,12.3}$ to varying magnitudes of the seasonal effect. Hirsch et al. used $A = 1$ and it is to be noted that their test is 100% robust against variations in A because no comparisons are made across seasons. In contrast, a test which makes comparisons across seasons will be sensitive to variations in A. However, it is essential that such a test be reasonably robust so as to satisfy the second requirement, stated above, that its power for detecting trends should be relatively high compared to alternative tests. For the stochastic process shown in eqn. (7.5), ϵ_{ij} is a normal random variable with zero mean and unit variance so that 95% of the error components fall within (-0.98, 0.98). The seasonal component varies between $(-A, A)$ so that large values of A are clearly impractical. Three different values of A are used for our study, these being $A = 1, 3$ and 5.

Eqn. (7.1) is a defining equation for r_s only for the case of continuous random variables. Consequently, eqn. (7.4b) is strictly valid only for untied rankings.

However, by analogy with the adjustment for bias made in the absence of ties, eqn. (7.4b) is used to define $r'_{s,ij}$ in the presence of ties and t is defined as in eqn. (6.1). Improvement upon this ad hoc approach to correcting for bias in the presence of ties provides a problem for further research.

Table 7.7: Percent of rejects at a significance level of 0.05

		α_{HSS}	$\alpha_{r'_{s,12.3}}$		
n	β	α_{HSS}	$A = 1$	$A = 3$	$A = 5$
5	0.00	4.6	4.4	6.6	8.6
	0.05	13.4	16.0	15.6	15.8
	0.10	42.0	49.2	50.6	50.2
	0.15	75.6	85.2	84.0	84.4
	0.20	95.6	98.4	98.4	98.4
	0.25	99.6	100	100	100
	0.30	100	100	100	100
10	0.00	5.6	4.2	3.8	3.4
	0.02	23.0	23.2	19.2	17.6
	0.04	57.4	60.2	54.6	49.2
	0.06	92.0	94.6	91.6	90.6
	0.08	99.2	99.6	99.2	99.2
	0.10	100	100	100	100
20	0.00	4.6	4.6	2.0	1.6
	0.0065	16.0	14.4	10.4	8.8
	0.0130	59.6	59.2	50.6	44.4
	0.0195	93.2	92.2	87.4	83.4
	0.0260	99.4	100	98.8	99.0
	0.0325	100	99.8	100	99.6
	0.0390	100	100	100	100

Table 7.7 displays the results of our simulation study. The rejection rates are 100% for those values of β not shown while the subscript HSS denotes the seasonal Kendall test, with no comparisons across seasons, developed by Hirsch, Slack and Smith. It is clear that, for small n, $r'_{s,12.3}$ is a better statistic than the seasonal Kendall test. On the other hand for large n and large A the robustness of the seasonal Kendall test results in a more powerful test statistic. For large n and small A both test statistics perform equally well; this indicating that for sufficiently large data sets, the additional information contained in inter-block comparisons is

insignificant. However, for small n or for cases with many missing observations, the additional information obtained from inter-block comparisons leads to a more powerful test. The sensitivity of $r'_{s,12.3}$ to the magnitude of the seasonal effect depends on the trend. For the larger trends corresponding to $n = 5$ the statistic is robust against A. For the smaller trends corresponding to $n = 20$, the statistic is much more sensitive to A. This sensitivity appears to result from the fact that the asymptotic variance estimate, as obtained under the complete null hypothesis, overestimates the sample variance which would pertain under H_o. For example, when $n = 20$ and $\beta = 0.0$, the rejection rates obtained using $r'_{s,12.3}$, and an estimate of the sample variance based on the 500 simulated values of $r'_{s,12.3}$, are $\alpha_{r'_{s,12.3}} = (5.0, 4.4, 5.2)$ for $A = (1, 3, 5)$.

It is clear that, unless the magnitude of the seasonal effect is quite large, the test statistic $r'_{s,12.3}$ competes favorably with the seasonal Kendall test. It would be of interest to ascertain whether or not an improved variance estimator for $Var(r'_{s,12.3})$ under H_o would lead to a more robust test statistic. That a considerably more robust statistic would result is indicated by the rejection rates obtained by using a sample estimate of the variance and as exemplified by the rates given at the end of the preceding paragraph.

REFERENCES

Conover, W.J. and Iman, R.L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistican*, **35**, 124–129.

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

Daniels, H.E. (1951). Note on Durbin and Stuart's formula for $\mathcal{E}(r_s)$. *Journal of the Royal Statistical Society*, B, **13**, 310.

Durbin, J. and Stuart, A. (1951). Inversions and rank correlation coefficients. *Journal of the Royal Statistical Society*, B, **13**, 303–309.

Hirsch, R.M., Slack, J.R. and Smith, R.A. (1982). Techniques of trend analysis for monthly water quality data. *Water Resources Research*, **18**, 107–121.

Kendall, M.G. (1942). Partial rank correlation. *Biometrika*, **32**, 277–283.

Kendall, M.G. (1975). *Rank Correlation Methods*. (4th ed). Griffin and Co. Ltd.

Korn, E.L. (1984). The ranges of limiting values of some partial correlations under conditional independence. *The American Statistican*, **38**, 61–62.

Somers, R.H. (1959). The rank analogue of product-moment partial correlation and regression, with application to manifold, ordered contingency tables. *Biometrika*, **46**, 241–246.

Chapter 8

CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER WORK

Following are the conclusions and recommendations for further work.

8.1 Conclusions

Algorithms which are capable of enumerating the exact null distributions of Kendall's S and Spearman's D statistics, when there are ties in one or both of the rankings, have been successfully developed. An expression for the cumulant generating function of S, when there are ties in both rankings, has been derived and successfully applied to the problem of obtaining a simple proof of the asymptotic normality of S. It has been shown that an Edgeworth approximation to the null distribution of S, in the general case of tied rankings, improves over the normal approximation providing that the ties are not too extensive. It has also been demonstrated that an Edgeworth approximation to the null distribution of D, in an absence of ties, is superior to a Pearson type II curve approximation when the maximum absolute error of approximation is used as the basis for comparison.

The distribution of Kendall's partial rank correlation statistic $t_{12.3}$, under the complete null hypothesis, has been enumerated for $n = 3, \ldots, 7$. Upper and lower bounds for the variance of $t_{12.3}$ have been established. A Monte Carlo experiment has shown that the distribution of $t_{12.3}$ under $H_o : \tau_{12.3} = 0$ depends on the underlying pairwise parental correlation coefficients.

A probability model, with the property that for the associated permutations $\mathcal{E}(t) = \tau$, has been developed for the elements of an inversion vector. It has been shown that this distribution leads to an exact variance, for t, which is identical to the upper limit, in large samples, of the variance which pertains when the variates are drawn from a bivariate normal distribution. It has been demonstrated that for one-sample hypothesis tests of τ and for data simulated from a bivariate normal distribution, a test using this exact variance performs reasonably well when compared to the more complex and computationally demanding bootstrap techniques. An algorithm for simulating rankings of size n, so that $\mathcal{E}(t) = \tau$ has been obtained and successfully applied to the problem of assessing the behaviour of Kendall's partial τ and of Spearman's partial ρ_s.

An asymptotic variance estimator for $t_{12.3}$ has been derived and the asymptotic normality of $t_{12.3}$ has been established, under H_o and for the general case of variates with underlying parental correlation. A Monte Carlo experiment has been used to show that when the magnitudes of t_{13} and t_{23} are both moderately large, $t_{12.3}$ is not a suitable statistic for testing the hypothesis of conditional independence. A simulation study of $r_{s,12.3}$ has been used to show that when corrected for bias in $r_{s,12}$ etc., $r_{s,12.3}$ provides a satisfactory statistic for testing the hypothesis of conditional independence. Finally, it has also been shown that the asymptotic distribution of the corrected test statistic $r'_{s,12.3}$, under $H_o : \rho_{s,12.3} = 0$, may be adequately approximated by its asymptotic distribution under the complete null hypothesis.

8.2 Recommendations for further work

It is recommended that further applications of the probability model developed in Chapter 5 be implemented with the objective of fully assessing its usefulness.

Research into an improved variance estimator for the variance of $r'_{s,12.3}$ under $H_o : \rho_{s,12.3} = 0$, is recommended. It is anticipated that an improved variance estimator will result in a test statistic which is more robust against the magnitude of the seasonal effect. Additional simulation experiments to further study the behaviour and the usefulness of $r'_{s,12.3}$ as a test statistic for the hypothesis of conditional independence, are also recommended.

Finally, research into the natural extension of the results of chapter 7 to the problem of partialling out two or more extraneous variables is recommended. Regarding this, MacNeill (1963) has developed algebraic formulae for a generalized partial rank correlation coefficient with an arbitrary number of extraneous variables, thus generalizing Somer's (1959) work on a generalized partial rank correlation coefficient for 3 variables. Such research, if successful, should result in a theoretical framework for the multivariate analysis of ordinal data.

REFERENCES

MacNeill, I.B. (1963). *Relationships between rank and product moment correlation coefficients.* Unpublished Master's Thesis, Queen's University, 1963.

Somers, R.H. (1959). The rank analogue of product-moment partial correlation and regression, with application to manifold, ordered contingency tables. *Biometrika*, **46**, 241–246.