Cumulant Generating Function and Tail Probability Approximations for Kendall's Score with Tied Rankings

Paul D. Valz
SACDA Inc.
343 Dundas St.
London, Ontario N6B 1V5
Canada

A. Ian McLeod
Department of Statistical and Actuarial Sciences
University of Western Ontario
London, Ontario N6A 5B7
Canada

Mary E. Thompson
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada

April 18, 1994

Short Title: Approximations for Kendall's Score

Key words and phrases: Cumulant generating function of Kendall's score, Hypergeometric distribution, Kendall's rank correlation with ties in both rankings, Asymptotic normality, Normal, Edgeworth and Pearson curve approximations.

AMS 1991 subject classifications: Primary 62G10; secondary 60-04, 60C05, 60E10, 62E20, 62G20

Abstract

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 and it is used to provide a direct proof of the asymptotic normality of the standardized score, $S/\sqrt{\mathrm{Var}\,(S)}$, when both rankings are tied. The third cumulant of S is derived and an expression for exact evaluation of the fourth cumulant is given. Significance testing in the general case of tied rankings via a Pearson Type I curve and an Edgeworth approximation to the null distribution of S is investigated and compared with results obtained under the standard Normal approximation as well as the exact distribution obtained by enumeration.

1 Introduction

Kendall's score may be written

$$S = \sum_{i<j}^{n} \mathrm{sign}((X_j - X_i)\,(Y_j - Y_i)) \tag{1}$$

where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are n independent replications of the random variables (X, Y); the score S permitting a nonparametric test of independence between X and Y. This test and the computation of its significance level are included in many statistical computer packages. In practice, ties in both of the rankings often arise due to the discreteness of the random variables. Even when conceptually the random variables are continuous, numerous ties may be present due to censoring or multiple detection levels. Hipel and McLeod (1994, §23.5 and §24.3) present several case studies involving trend tests of water quality variables in which multiple independent measurements were taken at the same time which gives ties in the time variable and due to the discreteness of the measurements and also due to detection level effects, the water quality parameter may exhibit a number of different tied values. So ties in both rankings are of interest in trend testing. The results of this paper may be used to develop an improved algorithm for the computation of the significance level of S in the case where ties are present in both rankings.

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 of 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 extents of observed ties are denoted by u_i $(i = 1, \ldots, k)$ and v_j $(j = 1, \ldots, \ell)$, respectively, and are given by

$$u_i = \sum_{j=1}^{\ell} a_{ij} \qquad \text{and} \qquad v_j = \sum_{i=1}^{k} a_{ij} \ . \tag{2}$$

Notice that $\sum_i u_i = \sum_j v_j = n$. 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 $\underset{\sim}{u} = (u_1, \ldots, u_k)$ and $\underset{\sim}{v} = (v_1, \ldots, v_\ell)$, which can be computed from the distribution of A given $\underset{\sim}{u}$, $\underset{\sim}{v}$ and the assumption of independence. Let $S_{\underset{\sim}{u},\underset{\sim}{v}}$ denote the random variable with this distribution and let $S_{\underset{\sim}{u},\underset{\sim}{v},A}$ denote the degenerate random variable obtained by conditioning upon A. In the case of no ties, $u_i = 1 \quad (i = 1, \ldots, k)$ and $v_j = 1 \quad (j = 1, \ldots, \ell)$ and $k = \ell = n$, the random variable $S_{\underset{\sim}{u},\underset{\sim}{v}}$ may be denoted by S_n. If there are ties in only the X-ranking, the random variable may be denoted by $S_{\underset{\sim}{u}}$. Similarly for $S_{\underset{\sim}{v}}$. In summary, $S_{\underset{\sim}{u}} = S_{\underset{\sim}{u},\underset{\sim}{1}}$, $S_{\underset{\sim}{v}} = S_{\underset{\sim}{v},\underset{\sim}{1}}$ and $S_n = S_{\underset{\sim}{1},\underset{\sim}{1}}$, where $\underset{\sim}{1}$ is a vector of n ones. Notice that for clarity, all vectors are indicated by a wavy underline. Finally, the cumulant generating functions (cgf's) of S_n, $S_{\underset{\sim}{u}}$ and $S_{\underset{\sim}{v}}$ are denoted by $K_n(t)$, $K_{\underset{\sim}{u}}(t)$ and $K_{\underset{\sim}{v}}(t)$ respectively.

2 The Cumulant Generating Function of $S_{\underset{\sim}{u},\underset{\sim}{v}}$

An explicit expression for the cgf of the score S_n has been derived by Silverstone (1950) and by David et al. (1951). All odd order cumulants are zero while even order cumulants are explicit polynomials in n of one degree higher than the order of the cumulant. Taking the known result in conjunction with the relation $S_n = S_{\underset{\sim}{u}} + \sum_{i=1}^{k} S_{u_i}$, where $S_{\underset{\sim}{u}}, S_{u_1}, \ldots, S_{u_k}$ are independent realizations of Kendall's scores, Robillard (1972) obtained an expression for the cgf of $S_{\underset{\sim}{u}}$. In the more general case of ties in both rankings the distribution or moments of $S_{\underset{\sim}{u},\underset{\sim}{v}}$ are determined from the fact that under the null hypothesis of independence the joint distribution of the matrix A given $\underset{\sim}{u}$, $\underset{\sim}{v}$ is central hypergeometric. The variable transformation developed below allows us to obtain explicit expressions for the moments of $S_{\underset{\sim}{u},\underset{\sim}{v}}$ and its cgf.

2.1 A fundamental variable transformation relationship

Let R_x and R_y denote the vectors of the ranks of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ respectively. Consider a specific permutation of the rankings R_x and R_y where there are ties of extent $u_i \quad (i = 1, \ldots, k)$ and $v_j \quad (j = 1, \ldots, \ell)$, respectively, in R_x and R_y. Let $A = (a_{ij})$ be the associated matrix. The following theorem introduces the fundamental variable transformation relationship upon which subsequent results are based.

Theorem 1

A score $S_{\underset{\sim}{u},\underset{\sim}{v},A}$ computed from a fixed matrix A may be related to a score $S_{n,A}$, corresponding to two untied rankings of size n, by the equation

$$S_{n,A} = S_{\underset{\sim}{u},\underset{\sim}{v},A} + \sum_{j=1}^{\ell} S_{\underset{\sim}{a}_j} + \sum_{i=1}^{k} S_{u_i} \ . \tag{3}$$

Proof: Let the rankings R_x and R_y be expressed by replacing observations by their midranks so that a tie of length v_j represents the repetition of the mean of v_j consecutive integers. Arbitrarily, replacing these v_j identical ranks in R_y by the corresponding integers increments the score from $S_{\underset{\sim}{u},\underset{\sim}{v},A}$ to $S_{\underset{\sim}{u},\underset{\sim}{v},A} + S_{\underset{\sim}{a}_j}$ where the incremental score $S_{\underset{\sim}{a}_j}$ 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. This follows from the definition, eqn. (1), of the score since all original contributions to the score are unchanged and, whereas previously $\text{sign}(Y_g - Y_h) = 0$ for Y_g, Y_h belonging to the set $\{v_j\}$ of tied ranks, it now holds that $\text{sign}(Y_g - Y_h) \neq 0$ for all Y_g, $Y_h \in \{v_j\}$. There are, however, still ties of extent $\underset{\sim}{a}_j = (a_{1j}, \ldots, a_{kj})$ in R_x, corresponding to the set $\{v_j\}$ in R_y, and thus untying the v_j tied ranks generates $S_{\underset{\sim}{a}_j}$. Repeated application of this procedure, firstly to ties in R_y and then to ties in R_x, yields eqn. (3). Note that when the u_i tied ranks in R_x are untied these generate an incremental score S_{u_i} since the ranking R_y is now completely untied.

Robillard's (1972) relationship for the scores when only one ranking is tied can be applied separately to each of the ℓ terms of the middle summation in eqn. (3) to yield a further reduction to

$$S_{v_j} = S_{\underset{\sim}{a}_j} + \sum_{i=1}^{k} S_{a_{ij}} \tag{4}$$

Note that this application of the Robillard reduction is not part of the algebraic transformation, but is applied separately at a later stage of the probabilistic construction.

2.2 Probabilistic Behaviour Under the Null Hypothesis

The construction of the preceding subsection is non-probabilistic and is true for any fixed matrix A. Under the null hypothesis of independence, the $k + \ell$ untyings in eqn. (3) can be chosen to be independent of each other and of A. It follows that all scores on the right of eqn. (3) can be taken as independently distributed Kendall scores of the type indicated by their respective subscripts. The scores S_{u_i} and S_{v_j} arrived at are independent of A and therefore their distributions are independent of A. The pair of untied rankings and its score $S_{n,A}$ are dependent on all the others. The conditional distribution of $S_{n,A}$, when averaged out over the null distribution of A on $\underset{\sim}{u}$, $\underset{\sim}{v}$, leads to the marginal distribution of a score S_n for two untied rankings of size n. In direct contrast, the distribution of S_n obtained with Robillard's construction, is independent of $\underset{\sim}{u}$; the difference being due to the fact that Robillard begins his construction with the unconditional random variable $S_{\underset{\sim}{u}}$ while our construction begins with the conditional random variable $S_{\underset{\sim}{u},\underset{\sim}{v},A}$.

As previously noted the null distribution of A conditional on $\underset{\sim}{u}$, $\underset{\sim}{v}$ follows a multivariate hypergeometric distribution,

$$\Pr(A) = \frac{\prod_i u_i! \prod_j v_j!}{n! \prod_i \prod_j a_{ij}!} \ . \tag{5}$$

Hence,

$$\sum_{\substack{A \\ \sum_j a_{ij}=u_i \\ \sum_i a_{ij}=v_j}} \frac{1}{\prod_i \prod_j a_{ij}!} = \frac{n!}{\prod_i u_i! \prod_j v_j!} \ . \tag{6}$$

2.3 Cumulant Generating Function of $S_{\underset{\sim}{u},\underset{\sim}{v}}$

Under the null hypothesis of independence, the characteristic function (cf) of the random variable $S_{n,A}$ is

$$E\left(e^{itS_{n,A}}|A\right) = e^{itS_{\underset{\sim}{u},\underset{\sim}{v},A}} E\left(e^{it\sum_{j=1}^{\ell} S_{\underset{\sim}{a}_j}}|A\right) E\left(e^{it\sum_{i=1}^{k} S_{u_i}}\right) \ . \tag{7}$$

Using eqn. (4) to obtain an expression for the cf of S_{v_j}, solving the resulting expression for $E\left(e^{it\sum_{j=1}^{\ell} S_{\underset{\sim}{a}_j}}|A\right)$ and substituting into eqn. (7), then gives

$$E\left(e^{itS_{n,A}}|A\right) = \frac{e^{itS_{\underset{\sim}{u},\underset{\sim}{v},A}} \prod_i E\left(e^{itS_{u_i}}\right) \prod_j E\left(e^{itS_{v_j}}\right)}{E\left(e^{it\sum_i \sum_j S_{a_{ij}}}|A\right)} \ . \tag{8}$$

Applying the fact that

$$E_A\left(E\left(e^{itS_{n,A}}|A\right)\right) = E\left(e^{itS_n}\right)$$

to eqn. (8) and taking logs yields

$$K_n(t) - \sum_i K_{u_i}(t) - \sum_j K_{v_j}(t) = \log E_A \left\{ e^{itS_{\underset{\sim}{u},\underset{\sim}{v},A}} e^{-\sum_i \sum_j K_{a_{ij}}(t)} \right\} \ , \tag{9}$$

where $K_m(t)$ is the cgf of a score for two untied rankings of m elements, and E_A is expectation with respect to the distribution of A conditional on $\underset{\sim}{u}$, $\underset{\sim}{v}$. Silverstone (1950) showed that

$$K_m(t) = \log \prod_{r=1}^{m} \frac{\sin rt}{r \sin(t)} \ , \qquad 0 < tm < \pi \ . \tag{10}$$

Let

$$x = E_A \left(e^{itS_{\underset{\sim}{u},\underset{\sim}{v},A}} e^{-\sum_i \sum_j K_{a_{ij}}(t)} \right) . \tag{11}$$

From eqns. (9) and (10), it follows that x is real-valued and positive for sufficiently small t. Using the fact that $K_{\underset{\sim}{u}}(t) = K_n(t) - \sum K_{u_i}(t)$ and $K_{\underset{\sim}{v}}(t) = K_n(t) - \sum K_{v_j}(t)$, it follows that $\log(x) = K_{\underset{\sim}{u}}(t) + K_{\underset{\sim}{v}}(t) - K_n(t)$. Note that in the degenerate case where all X's and Y's are tied, $K_{\underset{\sim}{u}}(t) = K_{\underset{\sim}{v}}(t) = 0$ so that $\log(x) = -K_n(t)$. From Robillard (1972, eqn. 1.4), it follows that $K_{\underset{\sim}{u}}(t) \leq 0$ and $K_{\underset{\sim}{v}}(t) \leq 0$. Hence $\log(x) \leq -K_n(t)$. Silverstone (1950) established that

$$-K_n(t) \leq \frac{1}{2}\sigma_n^2 t^2 + \frac{1}{n-1}\sigma_n^4 t^4 \ , \tag{12}$$

where $\sigma_n^2 = \mathrm{Var}\,(S_n)$ has been previously derived by Kendall (1975) and is given below by eqn. (23). Let $\Delta_1 > 0$ be the real positive solution to

$$\frac{1}{2}\sigma_n^2 \Delta_1^2 + \frac{1}{n-1}\sigma_n^4 \Delta_1^4 = \log(2) \ . \tag{13}$$

Then for $0 < t < \Delta_1$, we have $0 < x < 2$ and hence,

$$\log(x) = \sum_{\ell=1}^{\infty} (-1)^{\ell+1} \frac{(x-1)^\ell}{\ell} \ . \tag{14}$$

Now we can write

$$\begin{aligned} x &= E_A \left(e^{itS_{\underset{\sim}{u},\underset{\sim}{v},A}} \right) + E_A \left\{ \sum_{j=2}^{\infty} a_j \frac{(it)^j}{j!} \right\} \\ &= y + z \ , \end{aligned} \tag{15}$$

where a_j is the coefficient of $(it)^j/j!$ in the expansion of $\left(e^{itS_{\underset{\sim}{u},\underset{\sim}{v},A}}\right)\left(\sum_{h=1}^{\infty}\frac{\left(-\sum_i\sum_j K_{a_{ij}}(t)\right)^h}{h!}\right)$.

Since y is the cf of $S_{\underset{\sim}{u},\underset{\sim}{v}}$ for which all moments exist and since $S_{\underset{\sim}{u},\underset{\sim}{v}}$ has mean zero and variance $\sigma^2_{\underset{\sim}{u},\underset{\sim}{v}}$ given below in eqn. (29) as well as in Kendall (1975), we obtain using a well-known expansion for cf's (Loève, 1963, p.200),

$$y = 1 - \frac{1}{2}\sigma^2_{\underset{\sim}{u},\underset{\sim}{v}}t^2 + o(t^2)\ , \qquad \text{as} \quad t \to 0. \tag{16}$$

Hence there is a $\Delta_2 > 0$ such that $|y-1| < 1$ when $0 < t < \Delta_2$. Taking $0 < t < \min(\Delta_1, \Delta_2)$ we have $|x-1| < 1$ and $|y-1| < 1$ so that

$$\begin{aligned} \log(x) &= \sum_{\ell=1}^{\infty}(-1)^{\ell+1}\frac{((y-1)+z)^{\ell}}{\ell} \\ &= \log y + \sum_{j=2}^{\infty} b_j \frac{(it)^j}{j!}\ , \end{aligned} \tag{17}$$

where b_j is the coefficient of $(it)^j/j!$ in $\sum\limits_{\ell=1}^{\infty}\frac{(-1)^{\ell+1}}{\ell}\sum_{k=1}^{\ell}\binom{\ell}{k}(y-1)^{\ell-k}z^k$.

Now

$$y - 1 = \sum_{k=2}^{\infty} E_A\left(S^k_{\underset{\sim}{u},\underset{\sim}{v},A}\right)\frac{(it)^k}{k!} \tag{18}$$

since eqn. (3) yields $E(S_{n,A}|A) = S_{\underset{\sim}{u},\underset{\sim}{v},A}$ so that $E_A(S_{\underset{\sim}{u},\underset{\sim}{v},A}) = 0$. It follows that b_j for $j = 2,\ 3$ and 4 are obtained as the coefficients of $(it)^j/j!$ in $z' - (y'-1)z' - z'^2/2$ where

$$(y'-1) = E_A\left(S^2_{\underset{\sim}{u},\underset{\sim}{v},A}\right)(it)^2/2!$$

and

$$\begin{aligned} z' &= E_A\left[\left\{1 + itS_{\underset{\sim}{u},\underset{\sim}{v},A} + \frac{1}{2}(itS_{\underset{\sim}{u},\underset{\sim}{v},A})^2\right\}\left\{\left(\sum_i\sum_j k_2(a_{ij})\right)^2\frac{(it)^4}{8}\right.\right. \\ &\quad \left.\left. - \sum_i\sum_j\left(k_2(a_{ij})\frac{(it)^2}{2!} + k_4(a_{ij})\frac{(it)^4}{4!}\right)\right\}\right]\ . \end{aligned}$$

Note that $k_m(a_{ij})$ is the mth cumulant of the cgf $K_{a_{ij}}(t)$. One easily obtains

$$b_2 \;=\; -E_A\left(\sum_i\sum_j k_2(a_{ij})\right)$$

$$
\begin{aligned}
b_3 &= -3E_A\left(S_{\underset{\sim}{u},\underset{\sim}{v},A}\sum_i\sum_j k_2(a_{ij})\right)\\
b_4 &= -E_A\left(\sum_i\sum_j k_4(a_{ij})\right) - 6\ \mathrm{Cov}_A\left(S^2_{\underset{\sim}{u},\underset{\sim}{v},A},\sum_i\sum_j k_2(a_{ij})\right)\\
&\quad +3\ \mathrm{Var}_A\left(\sum_i\sum_j k_2(a_{ij})\right)\ . \qquad (19)
\end{aligned}
$$

Substituting from eqn. (17) into eqn. (9) then yields the expression

$$
\log E\left(e^{itS_{\underset{\sim}{u},\underset{\sim}{v}}}\right) = K_n(t) - \sum_{i=1}^{k} K_{u_i}(t) - \sum_{j=1}^{\ell} K_{v_j}(t) - \sum_{j=2}^{\infty} b_j\frac{(it)^j}{j!}\ , \qquad (20)
$$

where b_2, b_3 and b_4 are specified by eqn. (19), for the cgf of $S_{\underset{\sim}{u},\underset{\sim}{v}}$.

Let α denote $\sum_i\sum_j K_{a_{ij}}(t)$. In eqn. (15), the exponential $e^{-\alpha}$ is expanded about zero since this allows the coefficients b_2, b_3 and b_4 to be most efficiently extracted. Later, it is appropriate to also expand about $E_A(\alpha)$ whence eqn. (9) becomes

$$
K_n(t)-\sum_i K_{u_i}(t)-\sum_j K_{v_j}(t)+E_A(\alpha) = \log E_A\left\{e^{itS_{\underset{\sim}{u},\underset{\sim}{v},A}}\left(\sum_{h=0}^{\infty}(-1)^h\frac{(\alpha-E_A(\alpha))^h}{h!}\right)\right\}\ . \qquad (21)
$$

Applying the argument which led from eqn. (9) to eqn. (20) shows that

$$
\log E\left(e^{itS_{\underset{\sim}{u},\underset{\sim}{v}}}\right) = K_n(t) - \sum_{i=1}^{k} K_{u_i}(t) - \sum_{j=1}^{\ell} K_{v_j}(t) + E_A\left(\sum_i\sum_j K_{a_{ij}}(t)\right) - \sum_{j=3}^{\infty} d_j\frac{(it)^j}{j!}\ , \qquad (22)
$$

where d_j is the coefficient of $(it)^j/j!$ in $\sum\limits_{\ell=1}^{\infty}\frac{(-1)^{\ell+1}}{\ell}\sum_{k=1}^{\ell}\binom{\ell}{k}(y-1)^{\ell-k}w^k$, $y = E_A\left(e^{itS_{\underset{\sim}{u},\underset{\sim}{v},A}}\right)$ and $w = E_A\left\{\left(e^{itS_{\underset{\sim}{u},\underset{\sim}{v},A}}\right)\left(\sum_{h=1}^{\infty}(-1)^h\frac{(\alpha-E_A(\alpha))^h}{h!}\right)\right\}$.

3 Asymptotic Normality

The expression, eqn. (22), obtained for the cgf is used to provide a simple proof of the asymptotic normality of $S_{\underset{\sim}{u},\underset{\sim}{v}}/\sqrt{\mathrm{Var}\,(S_{\underset{\sim}{u},\underset{\sim}{v}})}$ under a trivial bound 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 asymptotic normality, which follows as a consequence of general results obtained by Hoeffding (1948), is not easy to give. Lehmann (1975, p.294) establishes the asymptotic normality of the standardized Spearman rank correlation in the case of tied ranks under equivalent conditions on the relative growth rates of n and the maximum extent of a tie in either ranking.

As a starting point, the second and third cumulants of $S_{\underset{\sim}{u},\underset{\sim}{v}}$ are evaluated and expressions for exact computation of the fourth cumulant are provided.

3.1 The Cumulants of $S_{\underset{\sim}{u},\underset{\sim}{v}}$

Let $\kappa_i,\ i = 1, 2, \ldots$ denote the i-th order cumulant of $S_{\underset{\sim}{u},\underset{\sim}{v}}$. It follows immediately from the absence of an $(it)^1$ term in eqn. (20), that $\kappa_1 = E(S_{\underset{\sim}{u},\underset{\sim}{v}}) = 0$. Under the null hypothesis of independent rankings, Noether (1967), Kendall (1975) and Valz and McLeod (1990) have shown that

$$\operatorname{Var}(S_n) = \frac{n(n-1)(2n+5)}{18} = \frac{n^{(3)}}{9} + \frac{n^{(2)}}{2}, \tag{23}$$

where $n^{(2)} = n(n-1)$ and $n^{(3)} = n(n-1)(n-2)$. Eqn. (20) yields

$$\operatorname{Var}(S_{\underset{\sim}{u},\underset{\sim}{v}}) = \operatorname{Var}(S_n) - \sum_i \operatorname{Var}(S_{u_i}) - \sum_j \operatorname{Var}(S_{v_j}) + \sum_i \sum_j E_A\Big(\operatorname{Var}(S_{a_{ij}}|A)\Big) \tag{24}$$

so that $\operatorname{Var}(S_{\underset{\sim}{u},\underset{\sim}{v}})$ is immediately determined upon evaluation of $E_A\Big(\operatorname{Var}(S_{a_{ij}}|A)\Big)$.

It is now shown that the particular forms of eqns. (5) and (6) allow exact determination of $E_A(a_{ij}^{(r)})$ for $r \geq 1$ where the factorial polynomial $a_{ij}^{(r)}$ is defined as

$$a_{ij}^{(r)} = a_{ij}(a_{ij}-1)\cdots(a_{ij}-r+1). \tag{25}$$

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

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

and consider the set $\{A'\}$ with $(a'_{ij}, u'_i, v'_j, n')$ replacing (a_{ij}, u_i, v_j, n) in eqn. (6). This yields

$$\sum_{A'} \frac{1}{\left(\prod_{gh\neq ij} a_{gh}!\right) a'_{ij}!} = \frac{n'!}{\left(\prod_{g\neq i} u_g!\right)\left(\prod_{h\neq j} v_h!\right) u'_i! v'_j!} \ . \tag{26}$$

From eqns. (5) and (26), it then follows that

$$\begin{aligned} E_A\left(a_{ij}^{(r)}\right) &= \frac{\prod_g u_g! \prod_h v_h!}{n!} \sum_A \frac{a_{ij}^{(r)}}{\prod_g \prod_h a_{gh}!} \\ &= \frac{\prod_g u_g! \prod_h v_h!}{n!} \sum_{A'} \frac{1}{\left(\prod_{gh\neq ij} a_{gh}!\right) a'_{ij}!} \\ &= \frac{u_i^{(r)} v_j^{(r)}}{n^{(r)}} \ . \end{aligned} \tag{27}$$

Consequently,

$$\begin{aligned} \sum_i \sum_j E_A\left(\operatorname{Var}\left(S_{a_{ij}}|A\right)\right) &= \sum_i \sum_j E_A\left(\frac{1}{9} a_{ij}^{(3)} + \frac{1}{2} a_{ij}^{(2)}\right) \\ &= \sum_i \sum_j \left(\frac{u_i^{(3)} v_j^{(3)}}{9n^{(3)}} + \frac{u_i^{(2)} v_j^{(2)}}{2n^{(2)}}\right) \end{aligned} \tag{28}$$

so that

$$\begin{aligned} \kappa_2 = \operatorname{Var}\left(S_{\underset{\sim}{u},\underset{\sim}{v}}\right) &= \frac{1}{9n^{(3)}}\left(n^{(3)} - \sum_i u_i^{(3)}\right)\left(n^{(3)} - \sum_j v_j^{(3)}\right) \\ &\quad + \frac{1}{2n^{(2)}}\left(n^{(2)} - \sum_i u_i^{(2)}\right)\left(n^{(2)} - \sum_j v_j^{(2)}\right) \ , \end{aligned} \tag{29}$$

a result which is consistent with that obtained by Kendall (1975) and Noether (1967), both of whom used different approaches.

It follows from eqns. (19) and (20) that

$$\kappa_3 = 3E_A\left(S_{\underset{\sim}{u},\underset{\sim}{v},A} \sum_i \sum_j k_2(a_{ij})\right) \tag{30}$$

where $S_{\underset{\sim}{u},\underset{\sim}{v},A}$, the sum of all second order determinants in the matrix A, may be expressed as $S_{\underset{\sim}{u},\underset{\sim}{v},A} = \sum_{i=1}^{k-1} \sum_{j=1}^{\ell-1} \sum_{g>i} \sum_{h>j} \left(a_{ij}a_{gh} - a_{gj}a_{ih}\right)$. Let b_{wz} be the polynomial in a_{wz} given by $E(S^2_{a_{wz}})$. Modifying the argument which led to eqn. (27) shows that

$$E_A\left(a_{ij}^{(r)} a_{ih}^{(s)}\right) = \frac{u_i^{(r+s)} v_j^{(r)} v_h^{(s)}}{n^{(r+s)}} \tag{31}$$

whence it is easily seen that $E_A\left(a_{ij}a_{gh} - a_{gj}a_{ih}\right) = 0$. It follows, in an analogous manner, that $E_A\left(\left(a_{ij}a_{gh} - a_{gj}a_{ih}\right)b_{wz}\right) = 0$ for $wz \neq ij,\ gh,\ gj$ or ih.

Consequently, eqn. (30) yields

$$\kappa_3 = 3E_A \sum_{i=1}^{k-1}\sum_{j=1}^{\ell-1}\sum_{g>i}\sum_{h>j}\left(a_{ij}a_{gh} - a_{gj}a_{ih}\right)\left(b_{ij} + b_{gh} + b_{gj} + b_{ih}\right) \ . \tag{32}$$

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

$$E_A\left(b_{ij}a_{ij}a_{gh}\right) = \frac{1}{18}\left(\frac{2u_i^{(4)}v_j^{(4)}u_g v_h}{n^{(5)}} + \frac{15u_i^{(3)}v_j^{(3)}u_g v_h}{n^{(4)}} + \frac{18u_i^{(2)}v_j^{(2)}u_g v_h}{n^{(3)}}\right) \tag{33}$$

and

$$E_A\left(b_{ij}a_{gj}a_{ih}\right) = \frac{1}{18}\left(\frac{2u_i^{(4)}v_j^{(4)}u_g v_h}{n^{(5)}} + \frac{9u_i^{(3)}v_j^{(3)}u_g v_h}{n^{(4)}}\right) \tag{34}$$

from which

$$E_A\left(b_{ij}a_{ij}a_{gh} - b_{ij}a_{gj}a_{ih}\right) = u_g v_h\left(\frac{u_i^{(3)}v_j^{(3)}}{3n^{(4)}} + \frac{u_i^{(2)}v_j^{(2)}}{n^{(3)}}\right) \ .$$

Substituting into eqn. (32) then yields

$$\begin{aligned}\kappa_3 \ &= \ \sum_{i=1}^{k-1}\sum_{j=1}^{\ell-1}\sum_{g>i}\sum_{h>j}\left[\frac{1}{n^{(4)}}\left(u_i^{(3)}u_g - u_i u_g^{(3)}\right)\left(v_j^{(3)}v_h - v_j v_h^{(3)}\right)\right. \\ &\quad \left. + \frac{3}{n^{(3)}}\left(u_i^{(2)}u_g - u_i u_g^{(2)}\right)\left(v_j^{(2)}v_h - v_j v_h^{(2)}\right)\right] \ .\end{aligned} \tag{35}$$

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

We now proceed to evaluate each term of b_4. Silverstone (1950) and David et al. (1951) showed that $k_4(n) = -n(6n^4 + 15n^3 + 10n^2 - 31)/225$ from which it is readily shown that

$$\begin{aligned}\sum_i\sum_j E_A\left(k_4(a_{ij})\right) \ &= \ -\frac{1}{225}\left(\frac{6}{n^{(5)}}\sum_i u_i^{(5)}\sum_j v_j^{(5)} + \frac{75}{n^{(4)}}\sum_i u_i^{(4)}\sum_j v_j^{(4)}\right. \\ &\quad \left. + \frac{250}{n^{(3)}}\sum_i u_i^{(3)}\sum_j v_j^{(3)} + \frac{225}{n^{(2)}}\sum_i u_i^{(2)}\sum_j v_j^{(2)}\right) \ .\end{aligned} \tag{36}$$

Squaring $\sum_i \sum_j$ Var $\left(S_{a_{ij}}|A\right)$, taking expectations and reducing gives

$$
\begin{aligned}
\text{Var}_A\left(\sum_i\sum_j k_2(a_{ij})\right) \;=\;& \left[\frac{1}{81n^{(6)}}\left\{\left(\sum_i u_i^{(3)}\right)^2 + \sum_i c_{u_i,3}u_i^{(3)}\right\}\{f(v_j)\}\right.\\
&+\frac{1}{9n^{(5)}}\left\{\left(\sum_i u_i^{(3)}\right)\left(\sum_i u_i^{(2)}\right) + \sum_i c_{u_i,23}u_i^{(3)}\right\}\{f(v_j)\}\\
&+\frac{1}{4n^{(4)}}\left\{\left(\sum_i u_i^{(2)}\right)^2 + \sum_i c_{u_i,2}u_i^{(2)}\right\}\{f(v_j)\}\\
&+\frac{1}{324}\left(\frac{36}{n^{(5)}}\sum_i u_i^{(5)}\sum_j v_j^{(5)} + \frac{288}{n^{(4)}}\sum_i u_i^{(4)}\sum_j v_j^{(4)}\right.\\
&\left.\left.+\frac{564}{n^{(3)}}\sum_i u_i^{(3)}\sum_j v_j^{(3)} + \frac{162}{n^{(2)}}\sum_i u_i^{(2)}\sum_j v_j^{(2)}\right)\right]\\
&-\left[\frac{1}{9n^{(3)}}\sum_i u_i^{(3)}\sum_j v_j^{(3)} + \frac{1}{2n^{(2)}}\sum_i u_i^{(2)}\sum_j v_j^{(2)}\right]^2 \qquad (37)
\end{aligned}
$$

where $u_i^{(6)} = u_i^{(3)}\left(u_i^{(3)} + c_{u_i,3}\right)$, $u_i^{(5)} = u_i^{(3)}\left(u_i^{(2)} + c_{u_i,23}\right)$, $u_i^{(4)} = u_i^{(2)}\left(u_i^{(2)} + c_{u_i,2}\right)$ so that $c_{u_i,3} = -9u_i^2 + 45u_i - 60$, $c_{u_i,23} = -6u_i + 12$, $c_{u_i,2} = -4u_i + 6$ and similarly for $c_{v_j,3}$ etc. in $f(v_j)$. The notation $f(v_j)$ in eqn. (37) above, and in eqn. (38) below, is used to designate the preceding component summed over j and evaluated at v instead of u.

It now remains to evaluate Cov $_A\left(S^2_{\underset{\sim}{\text{u}},\underset{\sim}{\text{v}},A}, \sum_i\sum_j E(S^2_{a_{ij}}|A)\right)$ which requires evaluation of $E_A\left(S^2_{\underset{\sim}{\text{u}},\underset{\sim}{\text{v}},A}\sum_i\sum_j E(S^2_{a_{ij}}|A)\right)$. This is an extremely tedious process which is too lengthy to be reproduced here. However, a sketch of the derivation is presented. Consider $\left(\sum_i\sum_{g>i}\right)^2$. This gives the terms: (i) $\sum_i\sum_{g>i}$ with $i_1 = i_2 \cap g_1 = g_2$, (ii) $2\sum_i\sum_{g_1>i}\sum_{g_2>g_1}$ with $i_1 = i_2 \cap g_1 \neq g_2$, (iii) $2\sum_{i_1}\sum_{i_2>i_1}\sum_{g>i_2}$ with $i_1 \neq i_2 \cap g_1 = g_2$, (iv) $2\sum_{i_1}\sum_{i_2>i_1}\sum_{g>i_2}$ with $i_1 \neq i_2 \cap g_1 \neq g_2 \cap g_1 = i_2$ and (v) $6\sum_{i_1}\sum_{i_2>i_1}\sum_{g_1>i_1}\sum_{g_2>i_2}$ with no tied subscripts, from which it follows that there are twenty-five terms to be considered in $S^2_{\underset{\sim}{\text{u}},\underset{\sim}{\text{v}}}$. It is easily shown that the nine terms with no tied subscripts in either one or both of i_1, i_2, g_1, g_2 and j_1, j_2, h_1, h_2 may be ignored. The remaining sixteen terms are then multiplied by $\sum_i\sum_j E(S^2_{a_{ij}}|A)$ prior to evaluation of the appropriate expectations which are then summed and reduced. This yields, for $E_A\left(S^2_{\underset{\sim}{u},\underset{\sim}{v},A}\sum_i\sum_j E(S^2_{a_{ij}}|A)\right)=Z$,

$$\begin{aligned}
Z &= \frac{1}{81n^{(6)}}\left[\left\{\left(n^{(3)}-\sum_i u_i^{(3)}\right)-9\left(n^2-5n\right)\right\}\sum_i u_i^{(3)}+9\sum_i\left(u_i^{(5)}+2u_i^{(4)}-6u_i^{(3)}\right)\right][f(v_j)] \\
&+\frac{1}{18n^{(5)}}\left[\left\{\left(n^{(3)}-\sum_i u_i^{(3)}\right)-6\left(n^2-4n\right)\right\}\sum_i u_i^{(2)}+6\sum_i\left(u_i^{(4)}+u_i^{(3)}-4u_i^{(2)}\right)\right][f(v_j)] \\
&+\frac{1}{18n^{(5)}}\left[\left\{\left(n^{(2)}-\sum_i u_i^{(2)}\right)-6\left(n-3\right)\right\}\sum_i u_i^{(3)}+6\sum_i u_i^{(4)}\right][f(v_j)] \\
&+\frac{1}{4n^{(4)}}\left[\left\{\left(n^{(2)}-\sum_i u_i^{(2)}\right)-4\left(n-2\right)\right\}\sum_i u_i^{(2)}+4\sum_i u_i^{(3)}\right][f(v_j)] \\
&+\frac{1}{n^{(5)}}\left\{\sum_{i_1\neq i_2}u_{i_1}^{(2)}u_{i_2}^{(3)}+2\sum_{i_1}\sum_{i_2>i_1}\sum_{i_3>i_2}\left(u_{i_1}^{(3)}u_{i_2}u_{i_3}-u_{i_1}u_{i_2}^{(3)}u_{i_3}+u_{i_1}u_{i_2}u_{i_3}^{(3)}\right)\right\}\{f(v_j)\} \\
&+\frac{2}{3n^{(5)}}\left\{\sum_{i_1\neq i_2}u_{i_1}u_{i_2}^{(4)}+2\sum_{i_1}\sum_{i_2>i_1}\sum_{i_3>i_2}u_{i_1}u_{i_2}^{(3)}u_{i_3}\right\}\{f(v_j)\} \\
&+\frac{2}{n^{(4)}}\left\{\sum_{i_1\neq i_2}u_{i_1}^{(2)}u_{i_2}^{(2)}+2\sum_{i_1}\sum_{i_2>i_1}\sum_{i_3>i_2}\left(u_{i_1}^{(2)}u_{i_2}u_{i_3}-u_{i_1}u_{i_2}^{(2)}u_{i_3}+u_{i_1}u_{i_2}u_{i_3}^{(2)}\right)\right\}\{f(v_j)\} \\
&+\frac{2}{n^{(4)}}\left\{\sum_{i_1\neq i_2}u_{i_1}u_{i_2}^{(3)}+2\sum_{i_1}\sum_{i_2>i_1}\sum_{i_3>i_2}u_{i_1}u_{i_2}^{(2)}u_{i_3}\right\}\{f(v_j)\} \\
&+\frac{1}{n^{(4)}}\sum_{i_1\neq i_2}u_{i_1}u_{i_2}^{(3)}\sum_{j_1\neq j_2}v_{j_1}v_{j_2}^{(3)}+\frac{2}{n^{(3)}}\sum_{i_1\neq i_2}u_{i_1}u_{i_2}^{(2)}\sum_{j_1\neq j_2}v_{j_1}v_{j_2}^{(2)}\ .
\end{aligned} \tag{38}$$

Substituting from eqns. (28), (29) and (36) through (38), into eqn. (19) then yields κ_4. Note that this result has been verified via exact enumeration of the distribution of $S_{\underset{\sim}{u},\underset{\sim}{v}}$ as discussed in Section 4.

3.2 Proof of asymptotic normality

Theorem 2

The distribution of $S_{\underset{\sim}{u},\underset{\sim}{v}}/\sqrt{\mathrm{Var}\left(S_{\underset{\sim}{u},\underset{\sim}{v}}\right)}$ converges to the standard normal distribution provided that M_u/n and M_v/n are bounded away from one as $n\to\infty$, where $M_u=\max(u_i)$ and $M_v=\max(v_j)$.

Proof: Rewriting the cgf for $S_{\underset{\sim}{u},\underset{\sim}{v}}$ in terms of the standard deviation $\sqrt{\kappa_2}$ as unit yields

$$\begin{aligned} K_{\underset{\sim}{u},\underset{\sim}{v}}\left(\frac{t}{\sqrt{\kappa_2}}\right) &= K_n\left(\frac{t}{\sqrt{\kappa_2}}\right) - \sum_{i=1}^{k} K_{u_i}\left(\frac{t}{\sqrt{\kappa_2}}\right) - \sum_{j=1}^{\ell} K_{v_j}\left(\frac{t}{\sqrt{\kappa_2}}\right) \\ &\quad + \sum_i \sum_j E_A\left(K_{a_{ij}}\left(\frac{t}{\sqrt{\kappa_2}}\right)\right) - \sum_{m=3}^{\infty} d_m \frac{(it/\sqrt{\kappa_2})^m}{m!} \\ &= -\frac{1}{2}t^2 \\ &\quad + \sum_{m=3}^{\infty} \kappa_2^{-\frac{m}{2}} \left[k_m(n) - \sum_{i=1}^{k} k_m(u_i) - \sum_{j=1}^{\ell} k_m(v_j) + \sum_{i=1}^{k}\sum_{j=1}^{\ell} E_A\left(k_m(a_{ij})\right) - d_m\right] \\ &\quad \frac{(it)^m}{m!}, \end{aligned} \tag{39}$$

where $k_m(n')$ denotes the m-th order cumulant of $S_{n'}$ for $n' = n, u_i, v_j, a_{ij}$. It suffices to establish that, for $m \geq 3$, each coefficient of $(it)^m/m!$ converges to zero as $n \to \infty$ so that the cumulants of $K_{\underset{\sim}{u},\underset{\sim}{v}}\left(t/\sqrt{\kappa_2}\right)$ are seen to converge to those of the standard normal distribution. The application of the converse of the Second Limit Theorem, Kendall and Stuart (Vol. 1, Section 4.30, 1963) secures the desired result.

Given that both M_u/n and M_v/n are bounded away from one it follows from eqn. (29) that $\liminf \quad \kappa_2/n^3 > 0$, i.e. κ_2 grows as fast as n^3. Since $k_{2g}(n)$ is of order n^{2g+1} it then follows that for $g \geq 2$

$$\left(k_{2g}(n) - \sum_i k_{2g}(u_i) - \sum_j k_{2g}(v_j) + \sum_i \sum_j E_A\left(k_{2g}(a_{ij})\right)\right) / \sqrt{\kappa_2}^{2g} \to 0 \tag{40}$$

since the ratio is seen to be (applying eqn. (41) below) of order n^{1-g}. Eqn. (40) establishes that the first four terms of the coefficient of $(it)^m/m!$ in eqn. (39) tend to zero as $n \to \infty$ for $m \geq 3$.

Thus it remains to be shown that $d_m/\sqrt{\kappa_2}^m \to 0$ for $m \geq 3$ and $n \to \infty$. It follows from eqn. (35) that $d_3/\sqrt{\kappa_2}^3 \to 0$ since κ_3 is seen to be of at most order n^4. Now $\sum\limits_i \sum\limits_j a_{ij}^{2g+1}$

$< \left(\sum_i \sum_j a_{ij}\right)^{2g+1} = n^{2g+1}$ so that $\sum_i \sum_j |k_{2g}(a_{ij})|$ is of at most order n^{2g+1} and therefore,

$$\sum_i \sum_j \frac{|k_{2g}(a_{ij})|}{\kappa_2^g} \leq O(n^{1-g}) \ . \tag{41}$$

Substituting from eqn. (41) into $\sum_i \sum_j K_{a_{ij}}\left(\frac{t}{\sqrt{\kappa_2}}\right)$, ignoring terms of order n^{1-g} for $g \geq 2$ and applying the definitions of $(y-1)$ and w in eqn. (22), yields each coefficient $d_m/\sqrt{\kappa_2}^m$ as a finite sum of terms of the form

$$C\left(E_A\left({S'_{\underset{\sim}{u},\underset{\sim}{v},A}}^{c_1}\right)\right)^{\ell-k}\left(E_A\left(K_{A,n}^{c_2}{S'_{\underset{\sim}{u},\underset{\sim}{v},A}}^{c_3}\right)\right)^k$$

where C is a constant, $S'_{\underset{\sim}{u},\underset{\sim}{v},A} = S_{\underset{\sim}{u},\underset{\sim}{v},A}/\sqrt{\kappa_2}$ and $K_{A,n} = \sum_i \sum_j \left(k_2(a_{ij}) - E_A k_2(a_{ij})\right)/\kappa_2$. Equating the exponent of $\sqrt{\kappa_2}$ to m gives $c_1(\ell-k) + (2c_2+c_3)k = m$ where $c_1 \geq 2$, $c_2 \geq 1$ and $\ell \geq k \geq 1$. Applying Holder's inequality shows that the absolute value of such a term is bounded above by

$$|C|\left(E_A\left|{S'_{\underset{\sim}{u},\underset{\sim}{v},A}}^{c_1}\right|\right)^{\ell-k}\left(E_A\left|K_{A,n}^{c_2+c_3}\right|\right)^{c_2k/(c_2+c_3)}\left(E_A\left|{S'_{\underset{\sim}{u},\underset{\sim}{v},A}}^{c_2+c_3}\right|\right)^{c_3k/(c_2+c_3)}$$

and hence $d_m/\sqrt{\kappa_2}^m \to 0$ as $n \to \infty$ provided that both $E_A|{S'_{\underset{\sim}{u},\underset{\sim}{v},A}}^{c_1}|$ and $E_A|{S'_{\underset{\sim}{u},\underset{\sim}{v},A}}^{c_2+c_3}|$, where $c_1 \leq m-2$ and $c_2+c_3 \leq m-1$, are bounded and that $E_A(K_{A,n}^r) \to 0$ as $n \to \infty$. The convergence of $d_m/\sqrt{\kappa_2}^m$ to zero, for $m > 3$, then follows by induction since the convergence of any cumulant of $S'_{\underset{\sim}{u},\underset{\sim}{v},A}$ to zero then implies the boundedness of its constituent moments. The proof is thus completed by showing that $E_A(K_{A,n}^r) \to 0$ as $n \to \infty$.

Now

$$E_A\left(\sum_i \sum_j \left(a_{ij}^{(3)} - \frac{u_i^{(3)}v_j^{(3)}}{n^{(3)}}\right)\right)^r = \sum_{s=0}^r \binom{r}{s}(-1)^{r-s}\left(\sum_i \sum_j \frac{u_i^{(3)}v_j^{(3)}}{n^{(3)}}\right)^{r-s} E_A\left\{\left(\sum_i \sum_j a_{ij}^{(3)}\right)^s\right\} \tag{42}$$

with

$$E_A\left\{\left(\sum_i \sum_j a_{ij}^{(3)}\right)^s\right\} = \sum_{p,q} \sum_{i_1,\ldots,i_p} \sum_{j_1,\cdots,j_q} \sum_{\left\{t_{gh}:\sum_{g=1}^p \sum_{h=1}^q t_{gh}=s\right\}} \binom{s}{t_{11}\ldots t_{pq}} E_A\left(\prod_{g=1}^p \prod_{h=1}^q a_{i_g j_h}^{(3)}\right)^{t_{gh}}$$

$$
\begin{aligned}
&= \not\Sigma \binom{s}{t_{11}\dots t_{pq}} E_A\left(\prod_{g=1}^{p}\prod_{h=1}^{q} a_{i_g j_h}^{(3t_{gh})}\right) + E_A(R_{1s}) \\
&= \not\Sigma \binom{s}{t_{11}\dots t_{pq}} \frac{1}{n^{(3s)}}\left(\prod_{g=1}^{p} u_{i_g}^{\left(3\sum_h t_{gh}\right)}\right)\left(\prod_{h=1}^{q} v_{j_h}^{\left(3\sum_g t_{gh}\right)}\right) + E_A(R_{1s}) \\
&= \not\Sigma \binom{s}{t_{11}\dots t_{pq}} \frac{1}{n^{(3s)}}\prod_{g=1}^{p}\prod_{h=1}^{q}\left(u_{i_g}^{(3)} v_{j_h}^{(3)}\right)^{t_{gh}} + E_A(R_{1s}) + R_{2s} \\
&= \frac{\left(n^{(3)}\right)^s}{n^{(3s)}}\left(\sum_i\sum_j \frac{u_i^{(3)} v_j^{(3)}}{n^{(3)}}\right)^s + E_A(R_{1s}) + R_{2s} \qquad (43)
\end{aligned}
$$

where $\not\Sigma$ denotes the four summations indicated in the first line, $p \le s$ and $q \le s$ respectively denote the number of distinct i and j subscripts in a typical term of the expansion of $\left(\sum_i \sum_j\right)^s$, and R_{1s}, R_{2s} respectively consist of terms $\prod_{g=1}^{p}\prod_{h=1}^{q} a_{i_g j_h}^{(r_{gh})}$ and $\prod_{g=1}^{p}\prod_{h=1}^{q}(u_{i_g} v_{j_h})^{(r_{gh})}$ for which $\sum_{g=1}^{p}\sum_{h=1}^{q} r_{gh} \le 3s-1$ so that R_{1s} and R_{2s} are both of order less than or equal to n^{3s-1}. Substituting from eqn. (42) into eqn. (43) then shows that

$$
\begin{aligned}
E_A\left(\sum_i\sum_j\left(a_{ij}^{(3)} - \frac{u_i^{(3)} v_j^{(3)}}{n^{(3)}}\right)\right)^r \sim & \left(\sum_i\sum_j u_i^{(3)} v_j^{(3)}\right)^r \sum_{s=0}^{r}\binom{r}{s}(-1)^{r-s}\left[\frac{\left(n^{(3)}\right)^s}{n^{(3s)}} - 1\right] \\
& + \sum_{s=1}^{r}\binom{r}{s}\left(E_A(R_{1s}) + R_{2s}\right)(-1)^{r-s}\left(\sum_i\sum_j \frac{u_i^{(3)} v_j^{(3)}}{n^{(3)}}\right)^{r-s} \qquad (44)
\end{aligned}
$$

which is of at most order n^{3r-1}. This establishes the desired result on $E_A(K_{A,n}^r)$ and thus completes the proof. Setting both c_3 and $\ell - k$ to zero in the terms comprising $d_m/\sqrt{\kappa_2}^m$ shows that this result is a necessary condition for asymptotic normality and that, for large M_u and M_v, the rate at which normality is approached can in fact be governed by the rate at which $E_A(K_{A,n}^r)$ approaches zero.

4 Approximations to the null distribution of $S_{\underset{\sim}{u},\underset{\sim}{v}}$

Burr (1960) and Valz (1990) have developed algorithms for obtaining the null distribution of $S_{\underset{\sim}{u},\underset{\sim}{v}}$ for small n by enumeration. The results of Sections 2 and 3 on the cgf of $S_{\underset{\sim}{u},\underset{\sim}{v}}$, when taken in conjunction with this algorithm, facilitate investigation into the usefulness of approximations which incorporate information on the third and fourth moments of $S_{\underset{\sim}{u},\underset{\sim}{v}}$. To this end a Pearson Type I curve and an Edgeworth approximation will be considered.

Olds (1938), Zar (1972) and Franklin (1987) have all demonstrated that use of a Pearson Type II curve (the symmetric sub-family of the Type I curve) to approximate tail probabilities of Spearman's rho, in an absence of ties and under the null hypothesis, leads to considerable improvement over the Normal approximation. This suggests that a Pearson Type I curve be presently considered. Noting that this curve corresponds to a beta distribution, parameters of the Type I curve are obtained from the first four cumulants of $S_{\underset{\sim}{u},\underset{\sim}{v}}$, Johnson and Kotz (1970, Chapter 24), using their eqns. (13) through (16) with β_2 replaced by β_1 in eqn. (20) – see Elderton and Johnson (1969). Cumulative probabilities are then obtained from the incomplete beta distribution.

David et al. (1951) and Silverstone (1950) showed that, when ties are absent, 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. In both of these cases, S is a lattice random variable with a span of 2, i.e. S is distributed over a set of uniformly spaced points with an interval width of 2. It follows that Normal, or other continuous, approximations to tail probabilities for a score S should be evaluated at $S-1$ if S is positive and at $S+1$ if S is negative; this being a correction for continuity. However, David et al. further suggested that the variance used for scaling S, as well as the higher cumulants of S, should be adjusted by Sheppard's corrections. Kolassa and McCullagh (1990) justified the use of Sheppard-adjusted cumulants in the case of sums of independent lattice random variables. Robillard (1972) omitted Sheppard's corrections; the inclusion of which might perhaps, in view of Kolassa and McCullagh (1990), lead to a marginal improvement.

The distribution of $S_{\underset{\sim}{u},\underset{\sim}{v}}$ possesses two features which serve to inhibit improvement over the accuracy of significance levels obtained from a Normal approximation. Firstly, spacing between adjacent scores is not constant, the irregularity being pronounced in the tails of the distribution. However the adjacent scores differ by one over most of the distribution provided that the ties are not too extensive. For the special case where one ranking is a dichotomy, which occurs for $k = 2$, Burr (1960) recommends that one-half of the highest common denominator of the numbers, $v_1+v_2,\ v_2+v_3,\ \ldots,\ v_{\ell-1}+v_\ell$ be used as the correction for continuity. More generally, as soon as $v_j = 1$ for some j, the recommended continuity correction is $\frac{1}{2}$. 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 ties increases.

Figures 1 and 2 below compare the Normal, Pearson Type I curve and Edgeworth approximations to exact tail probabilities for two selected examples taken from Burr (1960) and Kendall (1975). In the notation of Section 1, $\underset{\sim}{u} = (3, 4, 3)$ and $\underset{\sim}{v} = (2, 3, 3, 2)$ for the example shown in Fig. 1, and $\underset{\sim}{u} = (1, 2, 2, 2, 1, 2)$ and $\underset{\sim}{v} = (1, 1, 4, 3, 1)$ for the example shown in Fig. 2. The first distribution is exactly symmetric while the second is very nearly symmetric with a standardized 3rd cumulant of 0.00009. No attempt has been made to correct for continuity in either plot, it being clear from Fig. 1 that no choice of continuity correction is very good. These plots clearly demonstrate deterioration in the performance of approximations as ties become more extensive. For the case of a dichotomy in one ranking, Klotz (1966) found that the Edgeworth approximation offered little improvement over the Normal approximation in the case of the Wilcoxon test. While some improvement is obtained in the extreme lower tail of Fig. 2, it is clear from Fig. 1 that as ties become more extensive at best marginal improvement, if any, is to be expected. With extensive ties, an enumeration technique may be used to obtain the exact distribution of $S_{\underset{\sim}{u},\underset{\sim}{v}}$.

Acknowledgment

The authors would like to thank three referees and an Associate Editor for constructive helpful comments which were incorporated in our article.

REFERENCES

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

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.

Elderton, W.P. and Johnson, N.L. (1969). *Systems of Frequency Curves.* London: Cambridge University Press.

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.

Hipel, K.W. and McLeod, A.I. (1994). *Time Series Modelling of Water Resources and Environmental Systems.* Amsterdam: Elsevier.

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

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions – 2.* New York: Wiley.

Kendall, M.G. (1975). *Rank Correlation Methods* (4th ed). London: Griffin.

Kendall, M.G. and Stuart, A. (1963). *The Advanced Theory of Statistics* (2nd ed). New York: Hafner.

Klotz, J.H. (1966). The Wilcoxon, ties, and the computer. *Journal of the American Statis-*

tical Association **61** 772–787.

Kolassa, J.E. and McCullagh, P. (1990). Edgeworth series for lattice distributions. *Annals of Statistics* **18** 981–985.

Lehmann, E.L. (1975). *Nonparametrics* San Francisco: Holden-Day.

Loève, M. (1963). *Probability Theory* (3rd ed). New York: Van Nostrand Reinhold.

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.

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

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

Valz, P.D. (1990). *Developments in Rank Correlation Procedures.* Ph.D. Thesis, Department of Statistical and Actuarial Sciences, University of Western Ontario.

Valz, P.D. and McLeod, A.I. (1990). A simplified derivation of the variance of Kendall's τ. *The American Statistician* **44** 39–40.

Zar, J.H. (1972). Significance Testing of the Spearman Rank Correlation Coefficient. *Journal of the American Statistical Association* **67** 578–580.