The Annals of Statistics
1995, Vol. 23, No. 1, 144–160

CUMULANT GENERATING FUNCTION AND TAIL PROBABILITY APPROXIMATIONS FOR KENDALL'S SCORE WITH TIED RANKINGS

BY PAUL D. VALZ, A. IAN MCLEOD AND MARY E. THOMPSON

SACDA Inc., University of Western Ontario and University of Waterloo

Robillard's 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

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

where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are n independent replications of the random variables (X, Y); the score S permits 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), Sections 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

Received May 1992; revised April 1994.

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

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.

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, $j = 1, \ldots, l$, 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, $j = 1, \ldots, l$, respectively, and are given by

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

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

2. The cumulant generating function of $S_{\mathbf{u},\mathbf{v}}$. An explicit expression for the cgf of the score S_n has been derived by Silverstone (1950) and by David, Kendall and Stuart (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 $S_n = S_{\mathbf{u}} + \sum_{i=1}^{k} S_{u_i}$, where $S_{\mathbf{u}}, S_{u_1}, \ldots, S_{u_k}$ are independent realizations of Kendall's scores, Robillard (1972) obtained an expression for the cgf of $S_{\mathbf{u}}$. In the more general case of ties in both rankings, the distribution or moments of $S_{\mathbf{u},\mathbf{v}}$ are determined from the fact that under the null hypothesis of independence the joint distribution of the matrix A given $\mathbf{u}, \mathbf{v}$ is central hypergeometric. The variable transformation developed below allows us to obtain explicit expressions for the moments of $S_{\mathbf{u},\mathbf{v}}$ 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, $i = 1, \ldots, k$, and v_j, $j = 1, \ldots, l$, 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 $S_{\mathbf{u},\mathbf{v},A}$ 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*

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

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 $S_{\mathbf{u},\mathbf{v},A}$ to $S_{\mathbf{u},\mathbf{v},A} + S_{\mathbf{a}_j}$, where the incremental score $S_{\mathbf{a}_j}$ 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 [equation (1)] of the score since all original contributions to the score are unchanged and, whereas previously $\operatorname{sign}(Y_g - Y_h) = 0$ for Y_g and Y_h belonging to the set $\{v_j\}$ of tied ranks, it now holds that $\operatorname{sign}(Y_g - Y_h) \neq 0$ for all $Y_g, Y_h \in \{v_j\}$. There are, however, still ties of extent $\mathbf{a}_j = (a_{1j},\ldots,a_{kj})$ in R_x, corresponding to the set $\{v_j\}$ in R_y, and thus untying the v_j tied ranks generates $S_{\mathbf{a}_j}$. Repeated application of this procedure, first to ties in R_y and then to ties in R_x, yields (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. □

The Robillard (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 (3) to yield a further reduction to

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

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 nonprobabilistic and is true for any fixed matrix A. Under the null hypothesis of independence, the $k + l$ untyings in (3) can be chosen to be independent of each other and of A. It follows that all scores on the right-hand side of (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 over the null distribution of A on $\mathbf{u},\mathbf{v}$, 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 $\mathbf{u}$, the difference being due to the fact that Robillard begins his construction with the unconditional random variable $S_{\mathbf{u}}$ while our construction begins with the conditional random variable $S_{\mathbf{u},\mathbf{v},A}$.

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

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

Hence,

$$\sum_{\substack{A \\ \Sigma_j a_{ij} = u_i \\ \Sigma_i a_{ij} = v_j}} \frac{1}{\Pi_i \Pi_j a_{ij}!} = \frac{n!}{\Pi_i u_i! \Pi_j v_j!}. \tag{6}$$

2.3. *Cumulant generating function of* $S_{\mathbf{u},\mathbf{v}}$. Under the null hypothesis of independence, the characteristic function (cf) of the random variable $S_{n,A}$ is

$$\begin{aligned} &E(\exp(itS_{n,A})|A) \\ &\quad = \exp(itS_{\mathbf{u},\mathbf{v},A}) E\left(\exp\left(it \sum_{j=1}^{l} S_{\mathbf{a}_j}\right)\middle| A\right) E\left(\exp\left(it \sum_{i=1}^{k} S_{u_i}\right)\right). \end{aligned} \tag{7}$$

Using (4) to obtain an expression for the cf of S_{v_j}, solving the resulting expression for $E(\exp(it\Sigma_{j=1}^{l} S_{\mathbf{a}_j})|A)$ and substituting into (7), then gives

$$E(\exp(itS_{n,A})|A) = \frac{\exp(itS_{\mathbf{u},\mathbf{v},A}) \Pi_i E\left(\exp(itS_{u_i})\right) \Pi_j E\left(\exp(itS_{v_j})\right)}{E\left(\exp(it\Sigma_i \Sigma_j S_{a_{ij}})|A\right)}. \tag{8}$$

Applying the fact that

$$E_A(E(\exp(itS_{n,A})|A)) = E(\exp(itS_n))$$

to (8) and taking logs yields

$$\begin{aligned} &K_n(t) - \sum_i K_{u_i}(t) - \sum_j K_{v_j}(t) \\ &\quad = \log E_A\left\{\exp(itS_{\mathbf{u},\mathbf{v},A}) \exp\left(-\sum_i \sum_j K_{a_{ij}}(t)\right)\right\}, \end{aligned} \tag{9}$$

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 $\mathbf{u},\mathbf{v}$. 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

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

From (9) and (10), it follows that x is real-valued and positive for sufficiently small t. Using the fact that $K_{\mathbf{u}}(t) = K_n(t) - \Sigma K_{u_i}(t)$ and $K_{\mathbf{v}}(t) = K_n(t) - \Sigma K_{v_j}(t)$, it follows that $\log(x) = K_{\mathbf{u}}(t) + K_{\mathbf{v}}(t) - K_n(t)$. Note that in the degenerate case where all X's and Y's are tied, $K_{\mathbf{u}}(t) = K_{\mathbf{v}}(t) = 0$ so that $\log(x) = -K_n(t)$. From Robillard [(1972), equation 1.4], it follows that $K_{\mathbf{u}}(t) \leq 0$ and $K_{\mathbf{v}}(t) \leq 0$. 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 (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,

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

Now we can write

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

where a_j is the coefficient of $(it)^j/j!$ in the expansion of

$$(\exp(itS_{\mathbf{u},\mathbf{v},A}))\left(\sum_{h=1}^{\infty} \frac{\left(-\Sigma_i\Sigma_j K_{a_{ij}}(t)\right)^h}{h!}\right).$$

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

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

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

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

where b_j is the coefficient of $(it)^j/j!$ in

$$\sum_{l=1}^{\infty} \frac{(-1)^{l+1}}{l} \sum_{k=1}^{l} \binom{l}{k} (y-1)^{l-k} z^k.$$

Now

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

since (3) yields $E(S_{n,A}|A) = S_{\mathbf{u},\mathbf{v},A}$ so that $E_A(S_{\mathbf{u},\mathbf{v},A}) = 0$. It follows that the 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

$$(y' - 1) = E_A\left(S_{\mathbf{u},\mathbf{v},A}^2\right) \frac{(it)^2}{2!}$$

and

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

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

$$\begin{aligned} b_2 &= -E_A\left(\sum_i \sum_j k_2(a_{ij})\right), \\ b_3 &= -3E_A\left(S_{\mathbf{u},\mathbf{v},A} \sum_i \sum_j k_2(a_{ij})\right), \\ b_4 &= -E_A\left(\sum_i \sum_j k_4(a_{ij})\right) - 6\,\mathrm{Cov}_A\left(S_{\mathbf{u},\mathbf{v},A}^2, \sum_i \sum_j k_2(a_{ij})\right) \\ &\quad + 3\,\mathrm{Var}_A\left(\sum_i \sum_j k_2(a_{ij})\right). \end{aligned} \tag{19}$$

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

$$\begin{aligned} &\log E(\exp(itS_{\mathbf{u},\mathbf{v}})) \\ &\quad = K_n(t) - \sum_{i=1}^{k} K_{u_i}(t) - \sum_{j=1}^{l} K_{v_j}(t) - \sum_{j=2}^{\infty} b_j \frac{(it)^j}{j!}, \end{aligned} \tag{20}$$

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

Let α denote $\sum_i \sum_j K_{a_{ij}}(t)$. In (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 also to expand about $E_A(\alpha)$, whence (9) becomes

$$K_n(t) - \sum_i K_{u_i}(t) - \sum_j K_{v_j}(t) + E_A(\alpha)$$

$$= \log E_A\left\{\exp(itS_{\mathbf{u},\mathbf{v},A})\left(\sum_{h=0}^{\infty}(-1)^h\frac{(\alpha - E_A(\alpha))^h}{h!}\right)\right\}. \tag{21}$$

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

$$\log E(\exp(itS_{\mathbf{u},\mathbf{v}})) = K_n(t) - \sum_{i=1}^{k} K_{u_i}(t) - \sum_{j=1}^{l} K_{v_j}(t)$$

$$+ E_A\left(\sum_i \sum_j K_{a_{ij}}(t)\right) - \sum_{j=3}^{\infty} d_j \frac{(it)^j}{j!}, \tag{22}$$

where d_j is the coefficient of $(it)^j/j!$ in

$$\sum_{l=1}^{\infty}\frac{(-1)^{l+1}}{l}\sum_{k=1}^{l}\binom{l}{k}(y-1)^{l-k}w^k, \qquad y = E_A(\exp(itS_{\mathbf{u},\mathbf{v},A}))$$

and

$$w = E_A\left\{(\exp(itS_{\mathbf{u},\mathbf{v},A}))\left(\sum_{h=1}^{\infty}(-1)^h\frac{(\alpha - E_A(\alpha))^h}{h!}\right)\right\}.$$

3. Asymptotic normality. The expression (22) obtained for the cgf is used to provide a simple proof of the asymptotic normality of $S_{\mathbf{u},\mathbf{v}}/\sqrt{\mathrm{Var}(S_{\mathbf{u},\mathbf{v}})}$ 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), page 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 $S_{\mathbf{u},\mathbf{v}}$ are evaluated and expressions for exact computation of the fourth cumulant are provided.

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

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

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

$$\begin{aligned}\operatorname{Var}(S_{\mathbf{u},\mathbf{v}}) &= \operatorname{Var}(S_n) - \sum_i \operatorname{Var}(S_{u_i}) \\ &\quad - \sum_j \operatorname{Var}(S_{v_j}) + \sum_i \sum_j E_A\Big(\operatorname{Var}\big(S_{a_{ij}}|A\big)\Big),\end{aligned} \tag{24}$$

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

It is now shown that the particular forms of (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

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

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

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

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

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

Consequently,

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

so that

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

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

It follows from (19) and (20) that

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

where $S_{\mathbf{u},\mathbf{v},A}$, the sum of all second-order determinants in the matrix A, may be expressed as $S_{\mathbf{u},\mathbf{v},A} = \sum_{i=1}^{k-1}\sum_{j=1}^{l-1}\sum_{g>i}\sum_{h>j}(a_{ij}a_{gh} - a_{gj}a_{ih})$. Let b_{wz} be the polynomial in a_{wz} given by $E(S^2_{a_{wz}})$. Modifying the argument which led to (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 $E_A(a_{ij}a_{gh} - a_{gj}a_{ih}) = 0$. It follows, in an analogous manner, that $E_A((a_{ij}a_{gh} - a_{gj}a_{ih})b_{wz}) = 0$ for $wz \neq ij$, gh, gj or ih.

Consequently, (30) yields

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

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

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

and

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

from which

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

Substituting into (32) then yields

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

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, Kendall and Stuart (1951) showed that $k_4(n) = -n(6n^4 + 15n^3 + 10n^2 - 31)/225$, from which it is readily shown that

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

Squaring $\Sigma_i\Sigma_j \mathrm{Var}(S_{a_{ij}}|A)$, taking expectations and reducing gives

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

where $u_i^{(6)} = u_i^{(3)}(u_i^{(3)} + c_{u_i,3})$, $u_i^{(5)} = u_i^{(3)}(u_i^{(2)} + c_{u_i,23})$, $u_i^{(4)} = u_i^{(2)}(u_i^{(2)} + c_{u_i,2})$ 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 $c_{v_j,3}$ and so on, in $f(v_j)$. The notation $f(v_j)$ in equation (37) above, and in equation (38) below, is used to designate the preceding component summed over j and evaluated at v instead of u.

It now remains to evaluate $\mathrm{Cov}_A(S^2_{\mathbf{u},\mathbf{v},A}, \Sigma_i\Sigma_j E(S^2_{a_{ij}}|A))$, which requires evaluation of $E_A(S^2_{\mathbf{u},\mathbf{v},A}\Sigma_i\Sigma_j E(S^2_{a_{ij}}|A))$. This is an extremely tedious process which is too lengthy to be reproduced here. However, a sketch of the derivation is presented. Consider $(\Sigma_i\Sigma_{g>i})^2$. This gives the following terms: (i) $\Sigma_i\Sigma_{g>i}$ with $i_1 = i_2 \cap g_1 = g_2$; (ii) $2\Sigma_i\Sigma_{g_1>i}\Sigma_{g_2>g_1}$ with $i_1 = i_2 \cap g_1 \neq g_2$; (iii) $2\Sigma_{i_1}\Sigma_{i_2>i_1}\Sigma_{g>i_2}$ with $i_1 \neq i_2 \cap g_1 = g_2$; (iv) $2\Sigma_{i_1}\Sigma_{i_2>i_1}\Sigma_{g>i_2}$ with $i_1 \neq i_2 \cap g_1 \neq g_2 \cap g_1 = i_2$; and (v) $6\Sigma_{i_1}\Sigma_{i_2>i_1}\Sigma_{g_1>i_1}\Sigma_{g_2>i_2}$ with no tied subscripts, from which it follows that there are 25 terms to be considered in $S^2_{\mathbf{u},\mathbf{v}}$. 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 16 terms are then multiplied by $\Sigma_i \Sigma_j E(S^2_{a_{ij}}|A)$ prior to evaluation of the appropriate expectations, which are then summed and reduced. This yields, for $E_A(S^2_{\mathbf{u},\mathbf{v},A} \Sigma_i \Sigma_j E(S^2_{a_{ij}}|A)) = Z$,

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

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

3.2. *Proof of asymptotic normality.*

THEOREM 2. *The distribution of* $S_{\mathbf{u},\mathbf{v}}/\sqrt{\operatorname{Var}(S_{\mathbf{u},\mathbf{v}})}$ *converges to the standard normal distribution provided that* M_u/n *and* M_v/n *are bounded away from* 1 *as* $n \to \infty$, *where* $M_u = \max(u_i)$ *and* $M_v = \max(v_j)$.

PROOF. Rewriting the cgf for $S_{\mathbf{u},\mathbf{v}}$ in terms of the standard deviation $\sqrt{\kappa_2}$ as unit yields

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

where $k_m(n')$ denotes the mth-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 $K_{\mathbf{u},\mathbf{v}}(t/\sqrt{\kappa_2})$ are seen to converge to those of the standard normal distribution. The application of the converse of the second limit theorem [Kendall and Stuart (1963), Volume 1, Section 4.30] secures the desired result.

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

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

since the ratio is seen to be [applying (41) below] of order n^{1-g}. Equation (40) establishes that the first four terms of the coefficient of $(it)^m/m!$ in (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 (35) that $d_3/\sqrt{\kappa_2}^3 \to 0$ since κ_3 is seen to be of at most order n^4. Now $\Sigma_i \Sigma_j a_{ij}^{2g+1} < (\Sigma_i \Sigma_j a_{ij})^{2g+1} = n^{2g+1}$, so that $\Sigma_i \Sigma_j |k_{2g}(a_{ij})|$ 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 (41) into $\Sigma_i\Sigma_j K_{a_{ij}}(t/\sqrt{\kappa_2})$, ignoring terms of order n^{1-g} for $g \geq 2$ and applying the definitions of $(y-1)$ and w in (22), yields each coefficient $d_m/\sqrt{\kappa_2}^m$ as a finite sum of terms of the form

$$C\big(E_A(S'^{c_1}_{\mathbf{u},\mathbf{v},A})\big)^{l-k}\big(E_A(K^{c_2}_{A,n}S'^{c_3}_{\mathbf{u},\mathbf{v},A})\big)^k,$$

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

$$|C|\big(E_A|S'_{\mathbf{u},\mathbf{v},A}{}^{c_1}|\big)^{l-k}\big(E_A|K^{c_2+c_3}_{A,n}|\big)^{c_2k/(c_2+c_3)}\big(E_A|S'_{\mathbf{u},\mathbf{v},A}{}^{c_2+c_3}|\big)^{c_3k/(c_2+c_3)},$$

and hence $d_m/\sqrt{\kappa_2}^m \to 0$ as $n \to \infty$ provided that both $E_A|S'_{\mathbf{u},\mathbf{v},A}{}^{c_1}|$ and $E_A|S'_{\mathbf{u},\mathbf{v},A}{}^{c_2+c_3}|$, where $c_1 \leq m-2$ and $c_2 + c_3 \leq m-1$, are bounded and that $E_A(K^r_{A,n}) \to 0$ 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 $S'_{\mathbf{u},\mathbf{v},A}$ to zero then implies the boundedness of its constituent moments. The proof is thus completed by showing that $E_A(K^r_{A,n}) \to 0$ as $n \to \infty$.

Now

$$E_A\left(\sum_i\sum_j\left(a^{(3)}_{ij} - \frac{u^{(3)}_i v^{(3)}_j}{n^{(3)}}\right)\right)^r = \sum_{s=0}^{r}\binom{r}{s}(-1)^{r-s}\left(\sum_i\sum_j\frac{u^{(3)}_i v^{(3)}_j}{n^{(3)}}\right)^{r-s}E_A\left\{\left(\sum_i\sum_j a^{(3)}_{ij}\right)^s\right\}, \tag{42}$$

with

$$\begin{aligned}
&E_A\left\{\left(\sum_i\sum_j a^{(3)}_{ij}\right)^s\right\}\\
&= \sum_{p,q}\sum_{i_1,\ldots,i_p}\sum_{j_1,\ldots,j_q}\sum_{\{t_{gh}:\,\Sigma_{g=1}^{p}\Sigma_{h=1}^{q}t_{gh}=s\}}\binom{s}{t_{11}\ \cdots\ t_{pq}}E_A\left(\prod_{g=1}^{p}\prod_{h=1}^{q}a^{(3)}_{i_g j_h}\right)^{t_{gh}}\\
&= \not{\sum}\binom{s}{t_{11}\ \cdots\ t_{pq}}E_A\left(\prod_{g=1}^{p}\prod_{h=1}^{q}a^{(3t_{gh})}_{i_g j_h}\right) + E_A(R_{1s})\\
&= \not{\sum}\binom{s}{t_{11}\ \cdots\ t_{pq}}\frac{1}{n^{(3s)}}\left(\prod_{g=1}^{p}u^{(3\Sigma_h t_{gh})}_{i_g}\right)\left(\prod_{h=1}^{q}v^{(3\Sigma_g t_{gh})}_{j_h}\right) + E_A(R_{1s})\\
&= \not{\sum}\binom{s}{t_{11}\ \cdots\ t_{pq}}\frac{1}{n^{(3s)}}\prod_{g=1}^{p}\prod_{h=1}^{q}\left(u^{(3)}_{i_g}v^{(3)}_{j_h}\right)^{t_{gh}} + E_A(R_{1s}) + R_{2s}\\
&= \frac{(n^{(3)})^s}{n^{(3s)}}\left(\sum_i\sum_j\frac{u^{(3)}_i v^{(3)}_j}{n^{(3)}}\right)^s + E_A(R_{1s}) + R_{2s},
\end{aligned} \tag{43}$$

where $\not{\sum}$ denotes the four summations indicated in the first line; $p \leq s$ and $q \leq s$, respectively, denote the number of distinct i and j subscripts in a

typical term of the expansion of $(\Sigma_i \Sigma_j)^s$; and R_{1s} and 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 $\sum_{g=1}^{p}\sum_{h=1}^{q} r_{gh} \leq 3s - 1$ so that R_{1s} and R_{2s} are both of order less than or equal to n^{3s-1}. Substituting from (42) into (43) then shows that

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

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 $l - k$ 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 $S_{\mathbf{u},\mathbf{v}}$. Burr (1960) and Valz (1990) have developed algorithms for obtaining the null distribution of $S_{\mathbf{u},\mathbf{v}}$ for small n by enumeration. The results of Sections 2 and 3 on the cgf of $S_{\mathbf{u},\mathbf{v}}$, when taken in conjunction with this algorithm, facilitate investigation into the usefulness of approximations which incorporate information on the third and fourth moments of $S_{\mathbf{u},\mathbf{v}}$. 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 subfamily of the type I curve) to approximate tail probabilities of Spearman's ρ, 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 $S_{\mathbf{u},\mathbf{v}}$ [Johnson and Kotz (1970), Chapter 24], using their equations (13) through (16) with β_2 replaced by β_1 in (20) [see Elderton and Johnson (1969)]. Cumulative probabilities are then obtained from the incomplete beta distribution.

David, Kendall and Stuart (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 demon-

strated 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, that is, 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, Kendall and Stuart (1951) 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 $S_{\mathbf{u},\mathbf{v}}$ possesses two features which serve to inhibit improvement over the accuracy of significance levels obtained from a normal approximation. First, spacing between adjacent scores is not constant, the irregularity being pronounced in the tails of the distribution. However, the adjacent scores differ by 1 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 $v_1 + v_2, v_2 + v_3, \ldots, v_{l-1} + v_l$ be used as the correction for continuity. More generally, as soon as $v_j = 1$ for some j, the recommended continuity correction is $\frac{1}{2}$. Second, the distributions display serrated profiles which clearly limit the ability of a smooth curve to approxi-

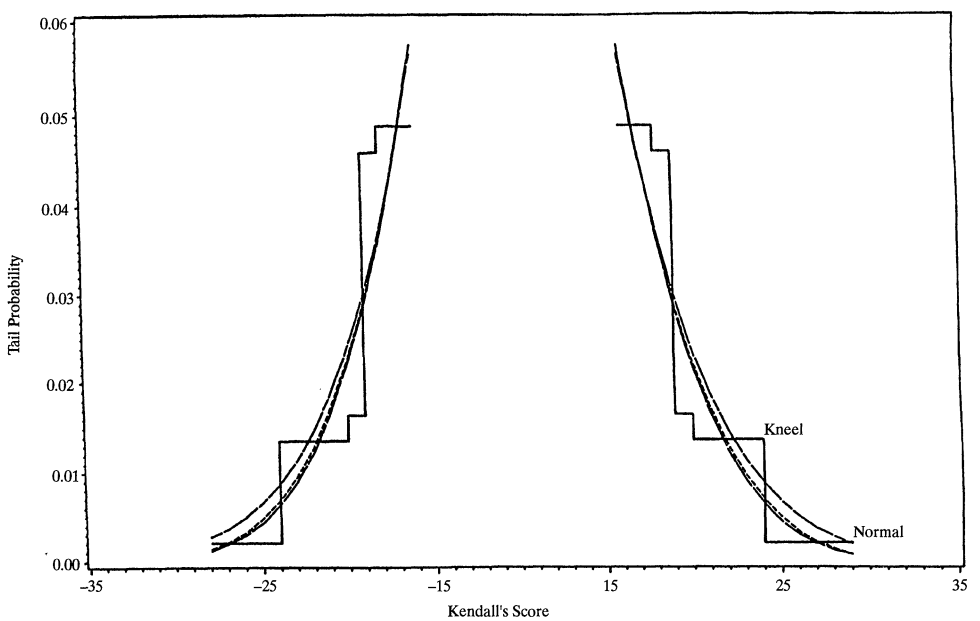

FIG. 1. *Plot of tail probability versus Kendall's score* [*Burr* (1960), *Example* 8.2]; *type I and Edgeworth curves are nearly coincident.*

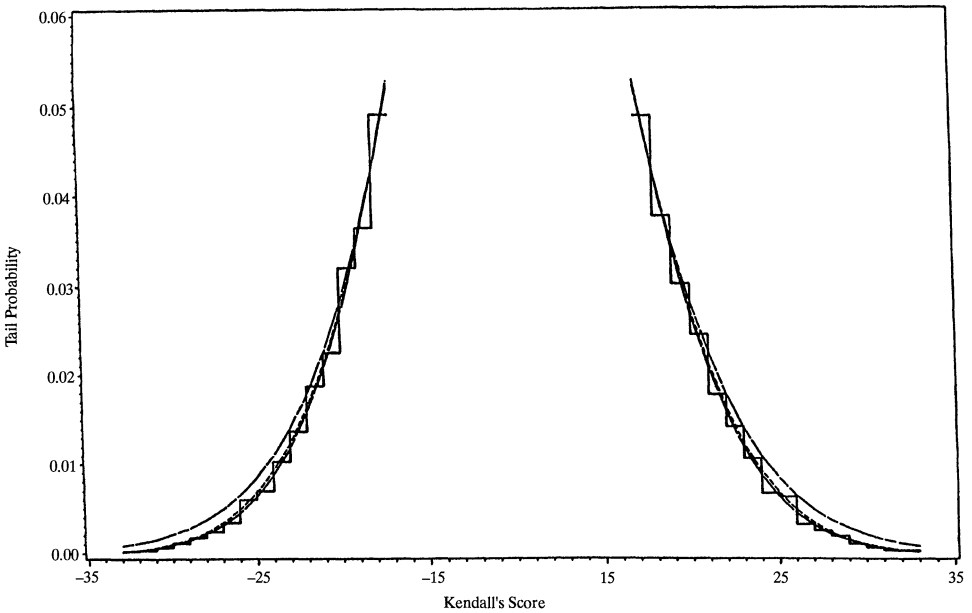

FIG. 2. *Plot of tail probability versus Kendall's score* [*Kendall* (1975), *Example* 3.1]; *type I and Edgeworth curves are nearly coincident.*

mate accurately the true distribution. This factor is exacerbated as the extent of ties increases.

Figures 1 and 2 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, $\mathbf{u} = (3, 4, 3)$ and $\mathbf{v} = (2, 3, 3, 2)$ for the example shown in Figure 1, and $\mathbf{u} = (1, 2, 2, 2, 1, 2)$ and $\mathbf{v} = (1, 1, 4, 3, 1)$ for the example shown in Figure 2. The first distribution is exactly symmetric, while the second is very nearly symmetric with a standardized third cumulant of 0.00009. No attempt has been made to correct for continuity in either plot, it being clear from Figure 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 Figure 2, it is clear from Figure 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 $S_{\mathbf{u},\mathbf{v}}$.

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. *J. Roy. Statist. Soc. Ser. C* **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*. Cambridge Univ. Press.

FRANKLIN, L. A. (1987). Approximations, convergence and exact tables for Spearman's rank correlation coefficient. In *Proceedings of the Statistical Computing Section* 244–247. Amer. Statist. Assoc., Alexandria, VA.

HIPEL, K. W. and MCLEOD, A. I. (1994). *Time Series Modelling of Water Resources and Environmental Systems*. North-Holland, Amsterdam.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

JOHNSON, N. L. and KOTZ, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions—2*. Wiley, New York.

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

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

KLOTZ, J. H. (1966). The Wilcoxon, ties, and the computer. *J. Amer. Statist. Assoc.* **61** 772–787.

KOLASSA, J. E. and MCCULLAGH, P. (1990). Edgeworth series for lattice distributions. *Ann. Statist.* **18** 981–985.

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

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

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

OLDS, E. G. (1938). Distributions of sums of squares of rank differences for small numbers of individuals. *Ann. Math. Statist.* **9** 133–148.

ROBILLARD, P. (1972). Kendall's S distribution with ties in one ranking. *J. Amer. Statist. Assoc.* **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. dissertation, Dept. Statistical and Actuarial Sciences, Univ. Western Ontario.

VALZ, P. D. and MCLEOD, A. I. (1990). A simplified derivation of the variance of Kendall's τ. *Amer. Statist.* **44** 39–40.

ZAR, J. H. (1972). Significance testing of the Spearman rank correlation coefficient. *J. Amer. Statist. Assoc.* **67** 578–580.

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