A tremendously simplified derivation of the variance of Kendall's τ

by

Paul D. Valz and A. I. M^c^Leod

Department of Statistical and Actuarial Sciences

University of Western Ontario

London, Ontario N6A 5B9

Canada

SUMMARY

Given two rankings R_1 and R_2 of the first n natural numbers, Kendall (1938) defines a statistic, τ, which provides a measure of the correlation between the two rankings. An expression for the variance of τ is given in Kendall (1970), whose derivation is exceedingly complex and lengthy. In this paper, we present a tremendously simplified derivation of the variance of τ.

KEYWORDS: Kendall's rank correlation coefficient; Inversion vector

1. INTRODUCTION

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

$$Q = \sum_{i>j} I_{(0,\infty)}(r_j - r_i), \tag{1}$$

where $I_{(0,\infty)}(\bullet)$ denotes the indicator function on $(0, \infty)$. Kendall's rank correlation coefficient (Kendall, 1970, equation 1.5) is then given by

$$\hat{\tau} = 1 - \frac{4Q}{n(n-1)}. \tag{2}$$

The variance of $\hat{\tau}$ when the two criteria are assumed to be independent is derived in Section 2. In Section 3, the derivation is extended to the case where there are ties in R_1.

The notion of an inversion vector provides the basis for our derivation. Reingold, Nievergelt and Deo (1977) define 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.

2. 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 of I_n that

$$Q = \sum_{j=1}^{n} i_j \ . \tag{3}$$

Since any of the set of $n!$ inversion vectors may be divided into $(\frac{n!}{j})$ subsets of j inversion vectors so that members of the same subset differ only on the jth element it follows that each of the j possible values $(0, 1, \cdots, j-1)$ of i_j, have probability j^{-1}. Hence,

$$E(i_j) = (j-1)/2 \tag{4}$$

and consequently

$$\begin{aligned} E(Q) &= \sum_{j=1}^{n} E(i_j) = \frac{1}{2}\sum_{j=1}^{n}(j-1) \\ &= \frac{1}{2}\binom{n}{2}. \end{aligned} \tag{6}$$

Similarly,

$$\begin{aligned} E(i_j^2) &= \sum_{i_j} {i_j}^2 \,\mathrm{Pr}(i_j) \\ &= (j-1)(2j-1)/6 \end{aligned} \tag{5}$$

and

$$\begin{aligned} \sum_{j\neq\ell}^{n} E(i_j i_\ell) &= \frac{1}{4}\sum_{j\neq\ell}^{n}(\ell-1)(j-1) \\ &= \left(\frac{1}{2}\sum_{j=1}^{n}(j-1)\right)^2 - \sum_{j=1}^{n}\frac{1}{4}(j-1)^2 \ . \end{aligned} \tag{8}$$

Consequently,

$$\begin{aligned}
E(Q^2) &= E\left(\sum_{j=1}^{n}\sum_{\ell=1}^{n} i_j i_\ell\right) \\
&= \left(\frac{1}{2}\binom{n}{2}\right)^2 - \frac{1}{4}\sum_{j=1}^{n}(j^2-2j+1) + \frac{1}{6}\sum_{j=1}^{n}(2j^2-3j+1) \\
&= \left(\frac{1}{2}\binom{n}{2}\right)^2 + \frac{n}{72}(n-1)(2n+5)\ . \qquad (9)
\end{aligned}$$

Hence,

$$Var(Q) = n(n-1)(2n+5)/72 \qquad (10)$$

and

$$Var(\tau) = \frac{2(2n+5)}{9n(n-1)}. \qquad (11)$$

3. EXTENSIONS AND CONCLUDING REMARK

In the event that there are m ties of length t_i, $1 \le t_i \le n$, $i = 1, 2, \cdots, m$; $(n = t_1 + t_2 + \cdots + t_m)$ in R_1, then following Robillard's (1972) argument and replacing the total score S by the negative score Q we have that

$$Q_n = Q^* + \sum_{i=1}^{m} Q_{t_i} \qquad (13)$$

where Q^* is the negative score obtained in the presence of ties and Q_{t_i} is the negative score obtained for two sets of untied observations on t_i objects. The $m+1$ negative scores on the right hand side of equation (13) are independent and therefore

$$\begin{aligned}
Var(Q^*) &= Var\ Q_n - \sum_{i=1}^{m} Var\ Q_{t_i} \\
&= \frac{1}{72}\left\{n(n-1)(2n+5) - \sum_{i=1}^{m} t_i(t_i-1)(2t_i+5)\right\} \qquad (14)
\end{aligned}$$

which is analogous to equation (4.4), of Kendall (1970), for the variance of the total score S.

The methodology presented in this article has been extended to obtain the variance of Kendall's partial rank correlation coefficient (Valz, 1988) and these results will be presented in a forthcoming article.

REFERENCES

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

Kendall, M.G. (1970). *Rank Correlation Methods* (4th ed). Griffin and Co. Ltd.

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

Robillard, P. (1972). Kendall's S distribution with ties in one ranking. *J. American Statistical Association*, **67**, 458.

Valz, P. (1988). Developments in Rank Correlation Procedures with Application to the Analysis of Water Quality Parameters. Ph.D. Thesis, University of Western Ontario.