BIOMETRICS 43, 409–416
June 1987

Kendall's and Spearman's Correlation Coefficients in the Presence of a Blocking Variable

Jeremy M. G. Taylor
Division of Biostatistics,
School of Public Health and Jonsson Comprehensive Cancer Center,
University of California at Los Angeles, Los Angeles, California, 90024, U.S.A.

SUMMARY

The use of a weighted sum of Kendall's taus or a weighted sum of Spearman's rhos for testing association in the presence of a blocking variable is discussed. In a Monte Carlo study the two are shown to have essentially the same power with the optimal choice of weights. In the presence of ties, the weighted sum of Spearman's rhos is preferred because its variance has a much simpler form. An example is given from the field of radiation therapy.

1. Introduction

Kendall's tau (Kendall, 1938) and Spearman's rho (Spearman, 1904) are two commonly used nonparametric methods of detecting associations between two variables. Their use is usually restricted to a single block. However, there are circumstances where the joint distribution of the two variables of interest (X and Y) is affected by the value of a third variable, called a blocking variable. In this paper, we restrict attention to the situation where there are multiple (X, Y) pairs for each value of the third variable. There are many ways to address this problem; for example, one could use a linear model approach and test for associations by examining the estimated regression coefficients. However, it may be desirable to use a less model-dependent approach. An example of where the underlying assumptions of the linear model are probably inappropriate is when the X and Y variables are ordered categorical, rather than continuous variables.

Elfving and Whitlock (1950), Ury (1968), Quade (1974), Reynolds (1974), and Korn (1984) have suggested using a weighted sum of Kendall's tau, across the blocks, to test for associations. Different choices in the weighting scheme and their merits are discussed by Korn. Spearman's rho is an alternative to Kendall's tau which can be used for testing for association. Kendall (1970, §1.2.4) claims that for many practical and most theoretical points of view, tau is preferable to rho. One reason for preferring tau, when estimating a correlation, is that the population parameter being estimated has a simpler interpretation. For hypothesis testing the interpretation of the test statistic is less important; the power properties of the respective tests are more relevant. For the case when one of the variables has two possible values, Lehmann (1975, p. 135) considers a weighted sum of Wilcoxon statistics.

Tau and rho are closely related; they are both functions of the ranks. Hettmansperger (1984, p. 203) gives insight into the relationship by showing that, assuming no ties,

$$\text{tau} = 1 - \frac{4}{n(n-1)} \sum\sum_{1 \le i < j \le n} s(R_i - R_j)$$

Key words: Conditional independence; Kendall's tau; Rank correlation; Spearman's rho.

and

$$\text{rho} = 1 - \frac{12}{n(n^2 - 1)} \sum\sum_{1 \leq i < j \leq n} (j - i)s(R_i - R_j),$$

where n is the number of pairs, $s(x) = 1$ if $x > 0$ and $s(x) = 0$ otherwise, and R_i is the rank of Y_i after the X's have been arranged in the order $X_1 < \cdots < X_n$. Kendall and Stuart (1973, p. 481) show that, if X and Y are independent, the correlation between tau and rho is at least .98 and tends to 1 as $n \rightarrow \infty$.

Van der Waerden (1969, p. 338) and Woodworth (1970) compare tau and rho for a bivariate normal distribution with correlation equal to c. By approximating the power, van der Waerden shows that rho is slightly more powerful than tau, for testing the hypothesis that c equals 0. Woodworth shows that the Bahadur efficiency of tau to rho appears to be between 1.0 and 1.05 for all values of c. For ordered contingency tables, Simon (1978) shows that tau and rho have the same efficacy; see also Proctor (1973) and Brown and Benedetti (1977).

We would expect the very similar power properties of tau and rho to extend to weighted averages of tau and rho, across many blocks. In this article we discuss what weighting scheme to use with tau and rho, both in the situation where there are no ties and when there are many ties. Based on a small Monte Carlo study, we conclude that optimally weighted sums of tau and rho are essentially equivalent in terms of power. However, the calculation of the significance level of the test is appreciably simpler for rho, if there are ties. Section 4 gives an example of the use of the weighted sum of rhos for some data from radiation treatment of cancer.

2. Testing for Association

2.1 *Kendall's Tau*

Using the notation of Korn, suppose we observe (X_{ik}, Y_{ik}), $i = 1, \ldots, n_k$; $k = 1, \ldots, K$; where n_k is the number of pairs in the kth block.

Let

$$T_k = \text{Kendall's tau for the } k\text{th block}$$

$$= \frac{2}{n_k(n_k - 1)} \sum\sum_{i<j} \text{sign}[(X_i - X_j)(Y_i - Y_j)],$$

where sign(a) equals 1, 0, or -1 if a is positive, zero, or negative, respectively.

T_k is an estimator of the probability of concordance minus the probability of discordance in block k; denote this by $\tau_{XY.k}$. Under the hypothesis H_0, of conditional independence between X and Y given the block, the mean and variance of T_k are given by

$$E(T_k) = 0 \quad \text{and} \quad \sigma_0^2(T_k) = \frac{2(2n_k + 5)}{9n_k(n_k - 1)}. \tag{2.1}$$

Let $T = \sum W_k T_k$ be the weighted sum of the T_k's; it is a suitable statistic for testing conditional independence. Different choices of weights can be used. A naive approach would be to use $W_k^{(1)} = 1$. Elfving and Whitlock (1950), Ury (1968), Quade (1974), and Reynolds (1974) suggest using $W_k^{(2)} = \frac{1}{2}n_k(n_k - 1)$. Korn suggests using $W_k^{(3)} = 1/\sigma_0^2(T_k)$. He notes that weighting by the inverse of the variance is optimal under alternatives in which $\tau_{XY.k}$ is constant across blocks. In practice, one may not know that $\tau_{XY.k}$ is constant across blocks, and even if it is not the case one may still be interested in testing for conditional independence.

If ties occur in the observations, then equation (2.1) is incorrect. Under the independence hypothesis the variance of T_k, conditional on the set of observed ranks, is (Noether, 1967, p. 77)

$$\begin{aligned}\sigma_{00}^2(T_k) = {} & \frac{2}{9n_k^2(n_k-1)^2}[n_k(n_k-1)(2n_k+5) - \sum f(f-1)(2f+5) - \sum g(g-1)(2g+5)] \\ & + \frac{4}{9n_k^3(n_k-1)^3(n_k-2)}[\sum f(f-1)(f-2)][\sum g(g-1)(g-2)] \\ & + \frac{2}{n_k^3(n_k-1)^3}[\sum f(f-1)][\sum g(g-1)],\end{aligned} \tag{2.2}$$

where f and g denote the multiplicity of ties of the X and Y ranks in block k, and the summations are over all sets of ties. Note that (2.2) reduces to (2.1) when there are no ties, and that $\sigma_{00}^2(T_k)$ is approximately equal to $\sigma_0^2(T_k)$ unless there is a vast number of ties. In particular, only if the observations are heavily grouped with a substantial proportion of the observations tied at a few values will $\sigma_0^2(T_k)$ and $\sigma_{00}^2(T_k)$ be significantly different.

A fourth possible weighting scheme is $W_k^{(4)} = 1/\sigma_{00}^2(T_k)$. Not only is $W_k^{(4)}$ practically unattractive, but it also has the disadvantage of depending on the data.

To test H_0, the null hypothesis of conditional independence, one compares $Z_1 = T[\sum W_k^2\sigma_{00}^2(T_k)]^{-1/2}$ with the standard normal distribution. Alternatively, one could use $Z_2 = T[\sum W_k^2\sigma_0^2(T_k)]^{-1/2}$, and hope that the effect of the ties was negligible.

The relative merits of using $W^{(1)}$, $W^{(2)}$, $W^{(3)}$, and $W^{(4)}$ for small sample sizes are addressed in Section 3, in the two cases where there are no ties and many ties.

2.2 *Spearman's Rho*

Let

$$r_k = \frac{\sum (u_i - \bar{u})(v_i - \bar{v})}{[\sum (u_i - \bar{u})^2 \sum (v_i - \bar{v})^2]^{1/2}},$$

where u_i (v_i) is the rank of X_i (Y_i); average ranks are used if there are ties. r_k is Spearman's rho in block k, the rank analogue of the product-moment correlation coefficient.

Under H_0, the hypothesis of conditional independence between X and Y, $E(r_k) = 0$ and $\text{var}(r_k) = (n_k - 1)^{-1}$. r_k has the nice property that the variance conditional on the set of ranks is $(n_k - 1)^{-1}$, even when there are ties in the data; see Kendall and Stuart (1973, p. 477).

Let $R = \sum W_k r_k$ be the weighted sum of the rhos. We consider two weighting schemes: (i) the naive choice $W_k^{(1)} = 1$, and (ii) $W_k^{(2)} = n_k - 1$. If the underlying population values of r_k are constant across blocks, then $W_k^{(2)}$ maximises the efficacy of R for testing H_0.

To test H_0, one compares $P = R[\sum W_k^2(n_k - 1)^{-1}]^{-1/2}$ with the standard normal distribution. Other approximations to the distribution of R could be used; however, we will use this simple approach.

3. Monte Carlo Comparison

To investigate the relative merits of tau and rho, and different weighting schemes, a small Monte Carlo simulation was performed. The study was not designed to cover all possible alternatives, but rather to provide information on three specific questions: (i) Is T more powerful than R or vice versa, in the two situations where there are no ties and many ties? The theoretical results would suggest that they are essentially equivalent. (ii) Which weights are best? Theoretically, $W^{(3)}$ and $W^{(4)}$ for T and $W^{(2)}$ for R are the optima, but are they significantly better than the other weights? (iii) If there are many ties in the observations, is it satisfactory to use the simple incorrect variance (2.1) for T, i.e., to use Z_2 rather than Z_1?

Four designs were considered: (a) $K = 1$, $n_1 = 10$; (b) $K = 1$, $n_1 = 40$; (c) $K = 2$, $(n_1, n_2) = (10, 40)$; and (d) $K = 5$, $(n_1, n_2, n_3, n_4, n_5) = (10, 10, 20, 40, 40)$. For each design 2000 bivariate normal pairs were generated, using the IMSL subroutine GGNSM. The pairs were from the model

$$\begin{pmatrix} X_{ik} \\ Y_{ik} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_k \\ \nu_k \end{pmatrix}, \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}\right),$$

where $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\nu} = (0, -.5, .5, -1.0, 1.0)$. When ties were desired, both X and Y were transformed by the function f where

$$f(a) = \begin{cases} 0 & \text{if } a < -1 \\ 1 & \text{if } -1 \leq a < 0 \\ 2 & \text{if } 0 \leq a < 1 \\ 3 & \text{if } 1 \leq a \end{cases}.$$

The only reason for the nonzero choice of $\boldsymbol{\nu}$ was to make the distribution of the margins in the 4×4 table, in the case of ties, not equal across all blocks.

For each fixed design and fixed c the same set of random (X, Y) samples was used to calculate the power of all the test statistics.

Table 1 shows the Monte Carlo powers for different values of c, for designs (a) and (b) with one block. The weighting scheme is irrelevant for one block. The three test statistics considered are Z_1 (tau with the correct tied variance), Z_2 (tau with the incorrect untied variance), and P (rho). The statistics Z_1 and Z_2 are identical when there are no ties. Notice that Z_1 and P attain the correct significance level at $c = 0$, but that Z_2 is too conservative

Table 1
Simulated rejection probabilities;
Monte Carlo results: 1 *block*, 2000 *simulations*

		Kendall's		Spearman's
c	Ties	Z_1	Z_2	P
		$n = 10$		
0	No	.050	.050	.048
0	Yes	.060	.025	.051
.3	No	.109	.109	.110
.3	Yes	.121	.053	.102
.6	No	.372	.372	.361
.6	Yes	.376	.218	.333
.9	No	.926	.926	.919
.9	Yes	.893	.711	.858
		$n = 40$		
0	No	.048	.048	.047
0	Yes	.052	.028	.049
.3	No	.418	.418	.419
.3	Yes	.380	.283	.369
.6	No	.976	.976	.977
.6	Yes	.948	.913	.946
.9	No	1.000	1.000	1.000
.9	Yes	1.000	1.000	1.000

when there are many ties, even for sample size equal to 40. There is an appreciable loss of power in using Z_2 rather than Z_1 where there are ties. When there are no ties, Z_1 and P appear to have equivalent power. When there are ties, for $n = 10$, Z_1 is slightly more powerful than P (less than 5% difference).

Table 2 shows the Monte Carlo powers for different values of c, for designs (c) and (d). For tau, four tests are considered, corresponding to the four weights $W_k^{(T_1)} = 1$, $W_k^{(T_2)} = \frac{1}{2}n_k(n_k - 1)$, $W_k^{(T_3)} = 1/\sigma_0^2(T_k)$, and $W_k^{(T_4)} = 1/\sigma_{00}^2(T_k)$. In all cases the correct variance $\sigma_{00}^2(T_k)$ is used to calculate Z. The statistics based on $W_k^{(T_3)}$ and $W_k^{(T_4)}$ are identical in the case of no ties. For rho, two tests are considered, corresponding to $W_k^{(S_1)} = 1$ and $W_k^{(S_2)} = (n_k - 1)$. We denote the estimators by the symbols given to the weights. Notice that all the tests achieve approximately the correct significance level when $c = 0$. For c not equal to zero, $W^{(T_1)}$ and $W^{(S_1)}$ are appreciably less powerful than the other tests. $W^{(T_3)}$, $W^{(T_4)}$, and $W^{(S_2)}$ are almost identical in power, with $W^{(T_2)}$ being slightly worse.

The conclusion to be drawn from this study is that using weights inversely proportional to the variance is slightly preferable to other choices. If there are many observations tied at a few values in the data, it is important to use the variance formula that incorporates the ties when calculating the significance level of the tau statistic. Using the simpler variance formula that ignores the ties results in too small a significance level and a loss of power. There is essentially no difference between using the optimally weighted sum of taus and the optimally weighted sum of rhos. From a practical point of view the weighted sum of Spearman's rhos is preferred because the weights and the resulting standardised test statistic take on a much simpler form than for the weighted sum of Kendall's taus; this is especially true if there are ties.

Table 2
Simulated rejection probabilities;
Monte Carlo results: 2 and 5 blocks, 2000 simulations

		Kendall's				Spearman's	
c	Ties	$W^{(T_1)}$	$W^{(T_2)}$	$W^{(T_3)}$	$W^{(T_4)}$	$W^{(S_1)}$	$W^{(S_2)}$
			n = 10, 40				
0	No	.050	.049	.050	.050	.051	.047
0	Yes	.057	.044	.052	.053	.050	.051
.2	No	.162	.250	.272	.272	.171	.267
.2	Yes	.175	.215	.235	.234	.167	.232
.3	No	.298	.447	.471	.471	.319	.477
.3	Yes	.282	.399	.422	.420	.266	.413
.4	No	.490	.713	.759	.759	.524	.763
.4	Yes	.481	.636	.687	.682	.467	.678
.6	No	.889	.987	.992	.992	.904	.992
.6	Yes	.868	.967	.980	.978	.857	.979
			n = 10, 10, 20, 40, 40				
0	No	.039	.047	.047	.047	.038	.041
0	Yes	.052	.049	.047	.048	.047	.044
.2	No	.342	.474	.510	.510	.361	.517
.2	Yes	.319	.397	.433	.431	.308	.423
.3	No	.630	.829	.856	.856	.663	.865
.3	Yes	.576	.731	.777	.771	.570	.768
.4	No	.901	.974	.986	.986	.913	.987
.4	Yes	.846	.924	.956	.955	.845	.953
.6	No	.999	1.000	1.000	1.000	.999	1.000
.6	Yes	.998	.999	1.000	1.000	.998	1.000

4. Example

This example is from the field of radiation therapy. There are three main factors in the treatment of tumors over which the radiotherapist has control. One is the total dose given (D), the second is the size of each dose per fraction (d), and the third is the overall treatment time (T). A typical regimen is a total dose of 60 Gray given in 2 Gray fractions 5 days a week, with an overall treatment time of 40 days. One limitation on the amount of dose that can be given is the possibility of late side effects, i.e., side effects that manifest themselves a year or more after the treatment. These late effects may not be life threatening, but they are serious consequences of the radiation.

In simple terms, the effect of the radiation is to kill cells. The occurrence and the severity of the late effect is determined by the fraction of surviving cells. Increasing D reduces this fraction, as does increasing d for fixed D. The effect of T is not so well understood. One possible simple biological model is that the damaged and undamaged cells recognize the damage and start to repopulate during the course of the treatment. This would mean that increasing the overall treatment time would result in an increased surviving fraction and hence less severe late effects.

Although this is a sensible biological model, the magnitude of the regeneration may be small, and also the time taken to recognize damage may be longer than the overall treatment time. From a practical point of view, what the radiotherapist would like to know is whether extending the overall treatment time spares the late effects.

The data were the information on 268 patients treated for head and neck cancer. All the patients in this analysis survived at least 18 months after the treatment, long enough to manifest the late effect. For each patient we know D, d, T, and the severity of the late effect, measured on an ordinal scale from 0 (no late effect) to 4 (most severe). For 76% of the patients there was no late effect. In addition, there were four sites of the tumors and hence of the radiation—namely, oral cavity, buccal, tongue, and tonsil. The late effect occurs in the same sort of mucosal tissue at all these sites, so the biology of the response to the radiation is very similar. There was a wide range of D, d, and T. We wish to determine whether T is inversely correlated with the degree of late effect after controlling for the effect of D and d. The data were blocked according to the site of radiation and a combination of D and d. Five homogeneous groups were defined: group 1, $58 \leqslant D \leqslant 65$ and $1.9 \leqslant d \leqslant 2.3$; group 2, $65 \leqslant D < 72$ and $2.3 < d \leqslant 2.5$; group 3, $65 \leqslant D < 72$ and $2.5 < d \leqslant 2.7$; group 4, $59 < D < 65$ and $2.3 < d \leqslant 2.6$; and group 5, $51 \leqslant D \leqslant 59$ and $2.3 < d \leqslant 3.0$. These groups included 95% of all patients. The remainder have an extreme combination of D and d, and did not fall naturally into any homogeneous group.

Table 3 shows Spearman's rho between T and the late effect, and the sample size within each block. Some dose groups are not included because there were no patients in that group or because no patients in that group showed late effects. Notice that most of the values in the table are negative, as predicted by the biological model. Spearman's rho rather than Kendall's tau was used because there are many ties and rho is much easier to use in that case. Combining the rhos according to the formula $R = \sum W_k r_k$, where $W_k = n_k - 1$, gives $R = -41.57$ and $\text{var}(R) = \sum (n_k - 1) = 215$. The weighted average Spearman's rho given by $\sum w_k r_k / \sum w_k$ is $-.19$. Thus, a one-sided test of H_0: no association between overall treatment time and the degree of late effect, versus H_1: negative association between overall treatment time and the degree of late effect, gives $\text{P} = -2.84$, with P-value = .002. So it is established in a model-independent way that extending the overall treatment time does decrease the severity of the late effect.

As a check of homogeneity among the 14 population correlations underlying the 14 sample correlations, we can suppose that each r_k is roughly normally distributed as befits a sample mean on n_k^{-1} observations. This leads to a between-sum-of-squares test statistic,

Table 3
Spearman's rho (between overall treatment time and late effect) for many blocks

Radiation site	Dose group	Sample size (n_k)	Spearman's rho (r_k)
Oral cavity	2	35	−.0074
	3	24	.0096
	4	39	−.2098
Buccal	2	5	−.4125
	3	9	−.0188
	4	8	−.5410
	5	3	−.5000
Tongue	2	16	.0355
	3	28	−.5956
	4	21	−.3865
Tonsil	1	10	−.1785
	2	7	.4119
	3	3	−.5000
	4	21	−.1784

$\sum W_k(r_k - R)^2 = 11.89$ to be compared to the theoretical chi-squared distribution on $K - 1$ degrees of freedom. We conclude there is no evidence of departure from a single underlying correlation.

The next step in the analysis could be to try to quantify the effect of time by hypothesizing a mathematical model for the repopulation mechanism. One could combine this with a logistic or other link function, or possibly use the extension of logistic regression to multiple ordered categories (McCullagh, 1980).

The above analysis is designed to demonstrate the statistical method described in this paper. It was only part of the overall analysis of the data set.

ACKNOWLEDGEMENTS

The author thanks Dr Bogoslaw Maciejewski for making available the data, and Dr Ed Korn for some useful suggestions. This work was supported by National Cancer Institute Grant CA16042 to the Jonsson Comprehensive Cancer Center.

RÉSUMÉ

On discute de l'intérêt de sommes pondérées de différents taux de Kendall ou de différents rho de Spearman pour le test d'association en présence d'un facteur bloc. On montre, par une étude de Monte Carlo, que les deux tests ont pratiquement la même puissance quand les poids optimaux sont choisis. En présence d'ex-aequo, la somme pondérée des rho de Spearman est préférable, car sa variance a une forme plus simple. On donne un exemple dans le domaine de la radiothérapie.

REFERENCES

Brown, M. B. and Benedetti, J. K. (1977). Sampling behaviour of tests for correlation in two-way contingency tables. *Journal of the American Statistical Association* **72,** 309–315.

Elfving, G. and Whitlock, J. H. (1950). A simple trend test with application to erythrocyte size data. *Biometrics* **6,** 282–288.

Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks.* New York: Wiley.

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

Kendall, M. G. (1970). *Rank Correlation Methods,* 4th edition. London: Griffin.

Kendall, M. G. and Stuart, A. (1973). *The Advanced Theory of Statistics, Vol.* 2, 3rd edition. New York: Hafner.

Korn, E. L. (1984). Kendall's tau with a blocking variable. *Biometrics* **40,** 209–214.

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

McCullagh, P. (1980). Regression models for ordinal data (with Discussion). *Journal of the Royal Statistical Society, Series B* **42,** 109–142.

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

Proctor, C. H. (1973). Relative efficiencies of some measures of association for ordered two-way contingency tables under varying intervalness of measurement errors. In *Proceedings of the Social Statistics Section,* 372–379.

Quade, D. (1974). Nonparametric partial correlation. In *Measurement in the Social Sciences,* H. M. Blalock, Jr (ed.), 369–398. Chicago: Aldine.

Reynolds, H. T. (1974). Ordinal partial correlation and causal inferences. In *Measurement in the Social Sciences,* H. M. Blalock, Jr (ed.), 399–423. Chicago: Aldine.

Simon, G. A. (1978). Efficacies of measures of association for ordinal contingency tables. *Journal of the American Statistical Association* **73,** 545–551.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* **15,** 72–101.

Ury, H. (1968). Photochemical air pollution and automobile accidents in Los Angeles. *Archives of Environmental Health* **17,** 334–342.

Van der Waerden, B. L. (1969). *Mathematical Statistics.* Berlin: Springer-Verlag.

Woodworth, G. G. (1970). Large deviations and Bahadur efficiency of linear rank statistics. *Annals of Mathematical Statistics* **41,** 251–283.

Received March 1986; *revised November* 1986.