Biometrika (1975), 62, 2, *p.* 417
Printed in Great Britain

Some properties of a test for concordance of two groups of rankings

BY LORETTA LI
Bishop College, Dallas, Texas
AND WILLIAM R. SCHUCANY
Department of Statistics, Southern Methodist University, Dallas, Texas

SUMMARY

The statistic $\mathscr{L}$, introduced to test for concordance within and between two groups of rankings of k objects is shown to be related to several measures of internal rank correlation. It has been previously shown that Kendall's W is proportional to the average of the rank correlations of all pairs of rankings in a single group and that Page's L is essentially the average of the rank correlations of an external ranking with each ranking in a group suggested by Lyerly. Similarly $\mathscr{L}$ is here shown to be directly related to the average of all rank correlations between a ranking from one group with a ranking from the other group. This statistic is also equal to the total of the L statistics calculated for each judge in one group with all judges in the other. The $\mathscr{L}$ statistic is shown to be uncorrelated with Kendall's W for concordance within either group. The asymptotic normality of $\mathscr{L}$ is established. A modification for ties is reported.

Some key words: Asymptotic normality; Concordance; Friedman statistic; Interaction; Internal rank correlation; Rank test; Tied rankings.

1. INTRODUCTION

The solution to the problem of testing for agreement among m sets of rankings of k objects has several approaches. Based on Spearman's coefficient ρ, Kendall & Babington Smith (1939) proposed the coefficient of concordance W. This statistic is intimately related to Friedman's χ^2 (1937) for two-way analysis of variance by ranks. Ehrenberg (1952) proposed a statistic u based on Kendall's correlation coefficient τ, and D. Quade in an unpublished University of Amsterdam report summarized Kendall's W and Ehrenberg's u as a special case of an average internal correlation which is a U-statistic.

Suppose that m rankings of k objects are given by m male judges. Also, n female judges are asked to rank the same k objects. We might be interested in whether the male judges and the female judges agree on the same ordering of the k objects. A special case occurs when either $m = 1$ or $n = 1$. This particular case was studied by Lyerly (1952) and Page (1963). Lyerly considered the average of each Spearman's ρ between each ranking in the group and the other independent ranking. Page introduced a statistic L which is related to Lyerly's average ρ. Schucany (1971) generalized these ideas and introduced the statistic $\mathscr{L}$ to be used to test the hypothesis of agreement of judges on the rankings of objects within each group and between the two groups.

2. DEFINITION AND PROPERTIES OF $\mathscr{L}$ AND $\mathscr{W}$

Suppose that one group has m rankings with ranks R_{ij} $(i = 1, \ldots, m; j = 1, \ldots, k)$ and that the other group has n rankings with ranks R'_{ij} $(i = 1, \ldots, n; j = 1, \ldots, k)$. Let

$$S_j = \sum_{i=1}^{m} R_{ij}, \quad T_j = \sum_{i=1}^{n} R'_{ij};$$

then the $\mathscr{L}$ statistic is defined to be

$$\mathscr{L} = \sum_{j=1}^{k} S_j T_j$$

(Schucany & Frawley, 1973). Under the null hypothesis that all row permutations are equally likely, the mean and variance of $\mathscr{L}$ as given by Schucany & Frawley (1973) may be used to define a standardized statistic, $\mathscr{L}^*$.

Ties occur if an observer cannot express a preference between two or more objects. If the midrank method is employed, the mean of $\mathscr{L}$ is unaltered but the variance is reduced. Considering the rankings in group I, let u_{ip} denote the number of objects involved in the pth set of ties for the ith ranking. Then if we define

$$U = \sum_{i=1}^{m} u_i,$$

where $u_i = \frac{1}{12}\Sigma_p u_{ip}(u_{ip}-1)(u_{ip}+1)$, it can be shown that

$$\operatorname{var}(S_j) = \frac{mk(k^2-1)-12U}{12k},$$

$$\operatorname{cov}(S_j, S_{j'}) = -\frac{1}{k-1}\operatorname{var}(S_j) \quad (j \neq j').$$

For group II let V be defined in a fashion similar to U. Then the variance of $\mathscr{L}$ in the presence of ties is less than the uncorrected variance by

$$C_t = \{k(k^2-1)(nU+mV)-12UV\}/\{12(k-1)\}.$$

Hence when midranks are used for tied rankings the standardized statistic $\mathscr{L}^*$ is given by

$$\{\mathscr{L}-E(\mathscr{L})\}/\{\operatorname{var}(\mathscr{L} \mid \text{ties})\}^{\frac{1}{2}}.$$

The distribution of $\mathscr{L}^*$ can be approximated by the standard normal. Large values of $\mathscr{L}^*$ indicate rejection of the null hypothesis in favour of the specific alternative of agreement within and between groups. If there is discordance in either group, $\mathscr{L}^*$ will assume values near zero and if there is concordance within each group on rankings which are not in agreement between groups, $\mathscr{L}^*$ will take values in the extreme left tail of its distribution.

A statistic whose range is the closed interval $[-1, 1]$ and may be viewed as a generalized concordance coefficient is given by

$$\mathscr{W} = \frac{\mathscr{L}-E(\mathscr{L})}{\max(\mathscr{L})-E(\mathscr{L})} = \frac{12\mathscr{L}-3mnk(k+1)^2}{mn(k^3-k)} = \{mn(k-1)\}^{-\frac{1}{2}}\mathscr{L}^*.$$

An interesting and useful property of $\mathscr{W}$ is given by the following theorem relating it to an average of all Spearman rank correlation coefficients of rankings from separate groups.

THEOREM 1. *Let ρ_{ij} denote the Spearman's ρ between the ith ranking of the first group and the jth ranking of the second group, then*

$$\mathscr{W} = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\rho_{ij}.$$

Proof. Let $\bar{\rho}_j = \Sigma\rho_{ij}/m$. From Lyerly (1952),

$$\bar{\rho}_j = 1 - \frac{2(2k+1)}{k-1} + \frac{12\sum_{i=1}^{k} R'_{ij}S_i}{m(k^3-k)},$$

and thus

$$\frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\rho_{ij} = 1 - \frac{2(2k+1)}{k-1} + \frac{12\sum_{i=1}^{k} S_iT_i}{mn(k^3-k)} = \mathscr{W}.$$

When $m = n = 1$, $\mathscr{W}$ becomes Spearman's ρ between the two rankings. When $m = 1$ or $n = 1$, $\mathscr{W}$ becomes Lyerley's average ρ and $\mathscr{L}$ becomes Page's L.

Thus, in fact, the intuitively appealing average of average ρ's is the statistic $\mathscr{W}$.

3. PREVIOUS WORK RELATED TO $\mathscr{L}$

Linhart (1960), Hays (1960) and Quade have different definitions for agreement between groups. Linhart requires the coefficient of concordance, Kendall's W, to be identical in both populations. His statistic is based on k statistics and is related to the difference between the Kendall's W of the two groups. Hays defined his average agreement within a single group as the average Kendall's τ. Let $T = m+n$ and $\bar{\tau}_i$ and $\bar{\tau}_T$ denote the average Kendall's τ between pairs of rankings in the ith group ($i = 1, 2$) and in the combined group, respectively. Then $\bar{\tau}_T$ can be partitioned into components as

$$\binom{T}{2}\bar{\tau}_T = \binom{m}{2}\bar{\tau}_1 + \binom{n}{2}\bar{\tau}_2 + mn\bar{\tau}_{12},$$

where $\bar{\tau}_{12}$ is defined as the average agreement between pairs drawn from group I and group II. If we adopt our average agreement within a single group as the average Spearman's ρ, we can also partition the total agreement in the combined group into agreement between pairs within the same group and agreement between pairs in different groups. Using a notation similar to Hays's, we have

$$\binom{T}{2}\bar{\rho}_T = \binom{m}{2}\bar{\rho}_1 + \binom{n}{2}\bar{\rho}_2 + mn\mathscr{W}.$$

Hence the average agreement between pairs in different groups is $\mathscr{W}$, as shown in Theorem 1.

Quade's definition for agreement between groups is that the expected rank correlation between any two rankings in each group have the same value, whereas Hays employs the narrower definition that the probability distribution is the same in each group of rankings. Both Hays and Quade pool the two groups and analyze the agreement in the combined group. This is not appropriate when m and n are extremely different as illustrated by the following example.

Example 1. Suppose $m = 2$, $n = 10$ and $k = 3$. Suppose the two judges in the first group are in perfect agreement on the ordering $(1, 2, 3)$, so that $S' = (2, 4, 6)$. Also, suppose the

ten judges in the second group all prefer the ordering (3, 2, 1) so that $T' = (30, 20, 10)$; hence $\mathscr{L} = 2 \times 30 + 4 \times 20 + 6 \times 10 = 200$. This is the minimum value of $\mathscr{L}$, so that $\mathscr{W} = -1$. The statistic indicates the agreement within and disagreement between. However, if the two Friedman's groups are pooled, $S' + T' = (32, 24, 16)$ and $m = 12$. The usual Friedman's test gives $\chi_r^2 = 10{\cdot}66$. Under the null hypothesis, Friedman's test is approximately chi-squared with two degrees of freedom. The critical value for a chi-squared distribution with two degrees of freedom at the 0·005 level is given by 10·6. Hence, contrary to the analysis using $\mathscr{W}$, the conclusion is to reject the null in favour of the alternative hypothesis that there is a substantial amount of agreement among all judges at most usual levels.

The following example considered by both Hays (1960) and Quade shows that pooling the two groups is not appropriate even when $m = n$.

Example 2. The rankings of $k = 6$ objects by two groups of $m = n = 16$ judges are analyzed for concordance. Hays calculated the average τ in each group and the combined group. He was unable to state any conclusion with respect to the comparison of agreement. Quade's average correlation coefficient could either be average ρ or average τ. In both cases, he calculated the difference between two average correlation coefficients for each group and then obtained the corresponding normal deviates. His conclusion was that the differences between the within-group agreements were not significant at the usual levels. This does not address the question of agreement between the two consensus rankings. The rank rotals are $S' = (48, 26, 57, 48, 79, 78)$ and $T' = (83, 51, 55, 32, 54, 61)$. Simple calculation yields $\mathscr{L}^* = 1{\cdot}509$ and $\mathscr{W} = 0{\cdot}042$.

The value of $\mathscr{L}^*$ is not significant at the 0·05 level, which provides still more insight. It may be noted that in the combined group, the Friedman test is significant at the usual levels. This should mean that there is a consensus of the judges in the combined group on the ordering of the 6 objects. Actually, as pointed out by Hays, a rank order which is a best fit to group I judges is 3 1 4 2 6 5 and to group II judges is 6 2 3 1 5 4. A conclusion that group I judges agree and group II judges agree but the two groups do not agree on the same ordering is a better conclusion than saying that the 32 judges agree on a particular ordering of the 6 objects. Therefore, the $\mathscr{L}$ statistic should be used instead of pooling the two groups together and analyzing the combined group.

4. Correlation with Friedman's test

The concordance within each group is represented by the Friedman statistic, or equivalently Kendall's W or $\bar{\rho}_i$. An interesting result, that the Friedman statistic for either group and the $\mathscr{L}$ statistic are uncorrelated, is given by the following theorem.

Theorem 2. *Under the null hypothesis that all row permutations are equally likely,*

$$\operatorname{cov}\left(\mathscr{L}, \sum_{j=1}^{k} S_j^2\right) = \operatorname{cov}\left(\mathscr{L}, \sum_{j=1}^{k} T_j^2\right) = 0.$$

Proof. It is clear that

$$\operatorname{cov}\left(\mathscr{L}, \sum_{j=1}^{k} S_j^2\right) = k \operatorname{cov}(S_j^2, S_j T_j) + k(k-1) \operatorname{cov}(S_j^2, S_{j'} T_{j'}).$$

But

$$\operatorname{cov}(S_j^2, S_j T_j) = \frac{m^2 n(k-1)(k+1)^3}{24},$$

$$\operatorname{cov}(S_j^2, S_{j'} T_{j'}) = -\frac{m^2 n(k+1)^3}{24} \quad (j \neq j').$$

Therefore,

$$\operatorname{cov}\left(\mathscr{L}, \sum_{j=1}^{k} S_j^2\right) = 0, \quad \operatorname{cov}\left(\mathscr{L}, \sum_{j=1}^{k} T_j^2\right) = 0.$$

The zero correlation and the asymptotic normality of $\mathscr{L}$, which will be discussed in § 5, suggest that $\mathscr{L}$ and Friedman's test might be asymptotically independent. Even though this has not been rigorously established, it raises the possibility of a procedure whereby the Friedman test and the $\mathscr{L}$ statistic could be used jointly to examine a second-stage hypothesis in the event that $\mathscr{L}$ does not significantly indicate the original alternative. In other words, if $\mathscr{L}^*$ is not significant we would like to examine the Friedman χ^2 within one of the groups and be able to treat the joint inference as if the two were independent.

5. Approximate and asymptotic normality of $\mathscr{L}$

The normal approximation for $\mathscr{L}^*$ has been appraised empirically by Schucany & Frawley (1973). Here we reconsider this approximation from the standpoint of the fourth moment of $\mathscr{L}^*$. Under the null hypothesis that all row permutations are equally likely, the third moment is zero due to symmetry and

$$E(\mathscr{L}^{*4}) - 3 = \frac{12(3k^2+10k+18)}{25mn(k-1)k(k+1)} - \frac{6}{k-1}\left(\frac{1}{m}+\frac{1}{n}\right) + \frac{6}{k-1}.$$

The first three moments of $\mathscr{L}$ agree with the moments of a standardized normal variate, and for large m, n and k, $E(\mathscr{L}^{*4}) \doteq 3$.

Table 1 presents values of $E(\mathscr{L}^{*4})$ for selected values of m, n and k. Note that $E(\mathscr{L}^{*4})$ goes to 3 most quickly if either $m = 1$ or $n = 1$. The normal approximation obviously improves as k increases. If both groups of judges are large, k must be at least 6 before the normal approximation is recommended. However, as may be seen from (2) below, for $k = 3$ and 5 we have an interesting appearance of the double exponential distribution. For fixed values of k, the value of $E(\mathscr{L}^{*4})$ does not differ substantially for larger values of m or n than are tabulated.

Further support for the use of a normal approximation is provided by the asymptotic normality of $\mathscr{L}^*$. This can be established by considering the limiting characteristic function

Table 1. *Values of* $E(\mathscr{L}^{*4})$ *for selected values of* m, n *and* k

	$k=2$			$k=6$			$k=10$			$k=25$		
m \ n	5	15	25	5	15	25	5	15	25	5	15	25
1	2·60	2·87	2·92	2·85	2·95	2·97	2·91	2·97	2·98	2·96	2·99	2·99
5	6·76	7·45	7·59	3·74	3·89	3·92	3·41	3·49	3·51	3·15	3·18	3·19
15	7·45	8·22	8·37	3·89	4·04	4·07	3·49	3·58	3·60	3·18	3·22	3·22
25	7·59	8·37	8·53	3·92	4·07	4·10	3·51	3·60	3·61	3·19	3·22	3·23

of $\mathscr{L}^*$ as each of m, n and k tend to infinity. First let us denote two $(k-1)\times 1$ vectors, S^* and T^*, as follows:

$$S^{*\prime} = (S_1^*, \ldots, S_{k-1}^*), \quad T^{*\prime} = (T_1^*, \ldots, T_{k-1}^*),$$

where

$$S_j^* = \frac{S_j - E(S_j)}{\{\mathrm{var}\,(S_j)\}^{\frac{1}{2}}} = \frac{S_j - \frac{1}{2}m(k+1)}{\{\frac{1}{12}m(k-1)\,(k+1)\}^{\frac{1}{2}}} \quad (j = 1, \ldots, k),$$

and the T_j are similarly standardized to obtain the T_j^*. Friedman showed that for large m and n both vectors, S^* and T^* are distributed as multivariate normals with mean vector 0 and covariance matrix Σ, where $(k-1)\Sigma = kI - J$, I is the $(k-1)\times(k-1)$ identity matrix and J is a matrix of ones. We may now state the following result which holds asymptotically in m and n.

THEOREM 3. *For fixed k and large m and n,*

$$E(e^{it\mathscr{L}^*}) = \left(1 + \frac{t^2}{k-1}\right)^{-\frac{1}{2}(k-1)}.$$

Proof. Writing $\mathscr{L}^* = \{(k-1)/k^2\}^{\frac{1}{2}} S^{*\prime}(I+J)\,T^*$, it can be shown that

$$E_{S^*}\{E_{T^*|S^*}(e^{it\mathscr{L}^*}|\,S^*)\} = E_{S^*}\left[\exp\left\{-\frac{t^2}{2k}S^{*\prime}(I+J)\,S^*\right\}\right]. \tag{1}$$

Since $S^{*\prime}(I+J)\,S^*(k-1)/k$ is asymptotically distributed as chi-squared with $(k-1)$ degrees of freedom, the right-hand side of (1) is the moment generating function of chi-squared with $k-1$ degrees of freedom evaluated at $-\frac{1}{2}t^2/(k-1)$, and the result follows.

Note that the characteristic function of $(k-1)^{\frac{1}{2}}\mathscr{L}^*$, for large m and n, is therefore

$$E\left[\exp\{it(k-1)^{\frac{1}{2}}\mathscr{L}^*\}\right] = (1+t^2)^{-\frac{1}{2}(k-1)}. \tag{2}$$

Hence, for odd values of k, and large values of m and n, $(k-1)^{\frac{1}{2}}\mathscr{L}^*$ is distributed as the sum of $\frac{1}{2}(k-1)$ independent variables, each having density $\frac{1}{2}e^{-|x|}$. In particular, when $k = 3$ and m and n are large, $\sqrt{2}\mathscr{L}^*$ has a Laplace distribution.

Now letting $k\to\infty$ and applying Fubini's Theorem it can be seen that the asymptotic characteristic function of $\mathscr{L}^*$ is $e^{-\frac{1}{2}t^2}$, establishing the asymptotic normality of $\mathscr{L}$.

6. STATISTIC $\mathscr{L}$ AS A TEST FOR INTERACTION

Large values of $\mathscr{L}$ indicate that each group of judges exhibits concordance and both groups agree on the same ordering, while small values of $\mathscr{L}$ indicate a consensus within each group of judges but the two groups have different opinions on the ordering of the objects. If one group of judges consists of m men and the other n women, one may wish to examine the hypothesis that sex has no effect on the rankings of these k objects. From the viewpoint of analysis of variance, this is equivalent to the testing of interaction between objects and sex. If ranks are considered as observations, the sum of squares due to interaction will be given by

$$\begin{aligned} H &= \frac{1}{m}\sum_{j=1}^{k} S_j^2 + \frac{1}{n}\sum_{j=1}^{k} T_j^2 - \frac{1}{m+n}\sum_{j=1}^{k}(S_j+T_j)^2 \\ &= \frac{n}{m(m+n)}\sum_{j=1}^{k} S_j^2 + \frac{m}{n(m+n)}\sum_{j=1}^{k} T_j^2 - \frac{2}{m+n}\mathscr{L}. \end{aligned}$$

When $\mathscr{L}$ attains the maximum value, all the rankings in each group are in perfect agreement on the same ordering. Hence the sums of squares of the S_j and T_j also achieve their maxima. Therefore, $H = 0$, which is the minimum value of H.

When $\mathscr{L}$ attains its minimum, the rankings in each group are in exact agreement but the two consensus orderings are diametrically opposed. Here, both ΣS_j^2 and ΣT_j^2 again attain their maximum values and so the value of H is maximized. Thus, in the presence of main effects, the interaction can be tested by the $\mathscr{L}$ statistic.

In this setting the usual null and alternative hypotheses concerning interaction are interchanged. A significant value of $\mathscr{L}$ indicates that the null hypothesis should be rejected in favour of the alternative of no interaction. But when one group shows complete lack of concordance with all rank sums equal to their expected value, then $\mathscr{L} = E(\mathscr{L})$ regardless of what the other group does. If the other group also shows lack of concordance, we might wish to say that interaction does not exist. The fact that this special case of no interaction in the absence of a main effect is not detected by $\mathscr{L}$ is of little practical importance. However it again suggests the potential usefulness of the conditional distribution of Friedman's χ^2 given an insignificant $\mathscr{L}$.

7. Conclusion

The properties of $\mathscr{L}$ indicate it to be a very useful statistic for concordance in problems dealing with natural ranks. It must also be considered to be a reasonable nonparametric alternative in a classical analysis of variance setting. However, the usual relative efficiencies have not been obtained due to the difficulty of specifying a normal theory procedure which is appropriate for the same special hypothesis. The topics of multiple comparisons and secondary tests conditional on the value of $\mathscr{L}$ are still under investigation.

This research was supported in part by the U.S. Office of Naval Research.

References

Ehrenberg, A. S. C. (1952). On sampling from a population of rankings. *Biometrika* **39**, 82–7.

Friedman, M. (1937). The use of ranks to avoid assumptions of normality implicit in the analysis of variance. *J. Am. Statist. Assoc.* **32**, 675–701.

Hays, W. L. (1960). A note on average tau as a measure of concordance. *J. Am. Statist. Assoc.* **55** 331–41.

Kendall, M. G. & Babington Smith, B. (1939). The problem of m rankings. *Ann. Math. Statist.* **10**, 275–87.

Linhart, H. (1960). Approximate test for m rankings. *Biometrika* **47**, 476–80.

Lyerly, S. B. (1952). The average Spearman rank correlations coefficient. *Psychometrika* **17**, 421–8.

Page, E. B. (1963). Ordered hypothesis for multiple treatments: A significance test for linear ranks. *J. Am. Statist. Assoc.* **58**, 216–30.

Schucany, W. R. (1971). A rank test for two group concordance (Abstract). *Ann. Math. Statist.* **42**, 1146.

Schucany, W. R. & Frawley, W. H. (1973). A rank test for two group concordance. *Psychometrika* **38**, 249–58.

[*Received August* 1974. *Revised February* 1975]