ON THE CORRELATION OF A GROUP OF RANKINGS WITH AN EXTERNAL ORDERING RELATIVE TO THE INTERNAL CONCORDANCE

by

A. D. Palachek
W. R. Schucany

Technical Report No. 161
Department of Statistics ONR Contract
May 1982

Research sponsored by the Office of Naval Research
Contract No. N00014-82-K-0207
Project NR 042-479

Department of Statistics
Southern Methodist University
Dallas, Texas 75275

ON THE CORRELATION OF A GROUP OF RANKINGS WITH AN EXTERNAL ORDERING RELATIVE TO THE INTERNAL CONCORDANCE

by

Albert D. Palachek
Department of Quantitative Analysis, University of Cincinnati

and

William R. Schucany
Department of Statistics, Southern Methodist University

SUMMARY

A method is presented for comparing the strength of agreement of a group of rankings with an external ordering to the corresponding measure of concordance within the group. While the procedure is not model dependent, we illustrate the characteristics of interest using an existing model for a non-null distribution for a population of rankings. U-statistics and a jackknife with adjusted degrees of freedom are employed to set approximate confidence intervals on the contrast between the two measures of rank order agreement.

Some key words: Concordance; Jackknife; Rankings; U-Statistic.

1. INTRODUCTION

The general problem of concordance among a group of judges as to the preference ordering of a set of k objects can be extended from the classical problem of m rankings to the problem of detecting agreement between the rankings and a specified predicted ordering of the objects that is given by the external ranking $\underline{y} = (y_1,\ldots,y_k)'$. Tests for the problem of m rankings were proposed by Kendall & Babington Smith (1939) and Ehrenberg (1952). Tests for agreement between the judges and an external ranking were proposed by Jonckheere (1954), Lyerly (1952), and Page (1963). All of these tests are based on statistics that are distribution-free under the null hypothesis of random rankings; i.e., that there is no agreement among the judges in the population. However, it is often known in advance that there is some concordance among the judges. The question of interest then becomes one of whether the judges agree with the predicted ordering of the objects. This question should not be interpreted as one of perfect agreement. In other words, the issue is not whether every judge elects the ranking $\underline{y}$ with probability one but whether the consensus ranking has a strong positive rank correlation with $\underline{y}$.

This external ranking setting can also be viewed as a special case of the problem of two-group concordance where the second population assigns probability one to the ranking $\underline{y}$. Tests for two-group concordance have been given by Schucany & Frawley (1973), Hollander & Sethuraman (1978), and recently by Kraemer (1981).

2. U-STATISTICS FOR INTERNAL AND EXTERNAL AGREEMENT

Quade, in a 1972 Technical Report at the Mathematical Centre, University of Amsterdam, uses U-statistic theory to examine the concordance of a population of judges as to the ordering of k objects. Let $\underline{X}_i = (X_{i1},\ldots,X_{ik})'$,

$i=1,\ldots,n$, denote the rankings obtained from a sample of n judges, each of whom independently rank the k objects. The coefficient of rank correlation between $\underline{X}_i$ and $\underline{X}_j$ will be denoted by $R(\underline{X}_i, \underline{X}_j)$. This coefficient may, for example, be taken to be the Spearman (1904) or Kendall (1938) rank correlation coefficient.

Quade's measure of concordance is given by

$$\rho = E\{R(\underline{X}_i, \underline{X}_j)\}, \qquad i \neq j,$$

where the expectation is with respect to the multinomial probability distribution over the population of k! rankings. This measure of concordance is referred to as the internal rank correlation, and $\rho = 0$ under the null hypothesis of random rankings. Furthermore, most investigators interpret $\rho > 0$ to be concordance among the judges.

The external rank correlation, which is a measure of the agreement between the judges and the external ranking, will be defined as

$$\rho_1 = E\{R(\underline{X}_i, \underline{y})\}.$$

This external rank correlation is positive if there is agreement between the judges and the predicted ordering.

The U-statistic estimators of ρ and ρ_1 are

$$\overline{R} = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \sum R(\underline{X}_i, \underline{X}_j)$$

and

$$R_1 = n^{-1} \sum_{i=1}^{n} R(\underline{X}_i, \underline{y}),$$

respectively. The tests for concordance introduced by Kendall & Babington Smith and Ehrenberg are based on $\overline{R}$. The external ranking tests due to Jonckheere, Lyerly, and Page are based on R_1 and test the null hypothesis of random rankings against the alternative that $\rho_1 > 0$.

3. COMPARISON OF INTERNAL AND EXTERNAL RANK CORRELATION

Although $\rho_1 > 0$ indicates that there is some agreement between the population of judges and the external ranking, there are situations in which there are marked differences between the consensus of the judges and the external ranking even though ρ_1 is positive. To illustrate this, consider a model introduced by Mallows (1957) and later studied by Feigin & Cohen (1978). For simplicity we will only consider rankings of the objects that contain no ties. Let $\underline{x}_0$ be a fixed vector denoting one of the k! possible orderings of the objects, and let $d(\underline{x}_0, \underline{x})$ denote a distance (in a rank correlation sense) between the orderings $\underline{x}_0$ and $\underline{x}$. A model which assigns equal probability to all rankings, $\underline{x}$, with the same value of $d(\underline{x}_0, \underline{x})$ is then

$$P(\underline{x}) = C(\theta)\theta^{d(\underline{x}_0, \underline{x})}, \quad 0 \leq \theta \leq 1.$$

Consider this model when k=4, $\underline{x}_0 = (1,2,3,4)'$, and the distance measure, $d(\underline{x}_0, \underline{x})$, is taken to be the number of discordant pairs of objects between $\underline{x}_0$ and $\underline{x}$. Further, restrict attention to the case in which $R(\cdot,\cdot)$ is the Spearman rank correlation coefficient and the external ranking is $\underline{y} = (1,4,2,3)'$. Table 1 presents some values of ρ_1 and ρ as functions of θ for this example. For the specific case of $\theta = .2$ the value $\rho_1 = .326$ is certainly large enough for the Page test to have reasonably good power at a moderate sample size. However, the expected ranking from the model is $\underline{\mu} = (1.24, 2.06, 2.94, 3.76)'$, which differs from $\underline{y}$ on the ordering of object 2 relative to objects 3 and 4.

TABLE 1

External and Internal Rank Correlation Coefficients for Mallows' Model

θ	ρ_1	ρ
0	.400	1.000
.1	.363	.865
.2	.326	.709
.3	.287	.547
.4	.246	.394
.5	.202	.263
.6	.158	.158
.7	.114	.083
.8	.073	.034
.9	.035	.008
1.0	0.000	0.000

Notice in this situation that the internal rank correlation is $\rho = .709$, which is larger than ρ_1. This indicates that there is stronger agreement within the population concerning some consensus ranking than there is with this particular external ranking. This clearly implies that the consensus of the judges is not the external ranking. If the external ranking had been chosen to be $\underline{y} = (1,2,3,4)'$, then the external rank correlation would be $\rho_1 = .842$, which is larger than ρ. So it appears that a comparison of ρ_1 and ρ should be made to determine "substantial" agreement with the external ranking.

Kraemer's two-group procedure is based on estimating a parameter that involves both the inter- and intra-group concordance. Using this approach in the external ranking setting would correspond to estimating

$$\rho^* = \frac{1}{2} + \frac{\rho_1}{\rho+1},$$

where ρ_1 and ρ are based on the Spearman $R(\cdot,\cdot)$. Extending Kraemer's two-group definition of "complete concordance" would require that $\rho^* = 1$. How-

ever, this can only occur when $\underline{\mu} = \underline{y}$, which means that each judge in the population assigns the ranking $\underline{y}$ with probability one. While it is appealing to relate the external ranking setting to the two-group problem, the condition that $\rho^* = 1$ is too stringent to be used as a definition of agreement.

What is needed is a definition of "substantial" agreement that is stronger than $\rho_1 > 0$ but not as stringent as $\rho^* = 1$. Since $\rho_1 < \rho$ indicates that the consensus is not the external ranking, then a reasonable definition of substantial agreement would be that $\rho_1 \geq \rho$. This indicates that the external rank correlation is substantial relative to the internal rank correlation.

The internal and external rank correlation can be compared by examining the parameter $\rho_d = \rho_1 - \rho$. This parameter is estimable of degree two with kernel

$$\phi(\underline{X}_i, \underline{X}_j) = \frac{1}{2}\{R(\underline{X}_i, \underline{y}) + R(\underline{X}_j, \underline{y})\} - R(\underline{X}_i, \underline{X}_j).$$

The U-statistic estimator of ρ_d is

$$R_d = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \phi(\underline{X}_i, \underline{X}_j) = R_1 - \overline{R}.$$

Since R_d is invariant to the jackknife procedure, we can estimate the variance of R_d using the jackknife, obtaining a multiple of a sample variance,

$$\hat{\sigma}_d^2 = \frac{4(n-1)}{n(n-2)^2} \sum_{i=1}^{n} (V_i - R_d)^2 = \frac{4(n-1)^2}{n(n-2)^2} S_d^2,$$

where

$$V_i = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} \phi(\underline{X}_i, \underline{X}_j).$$

Then the limiting distribution (as $n \to \infty$) of the studentized U-statistic $(R_d - \rho_d)/\hat{\sigma}_d$ is standard normal under mild regularity conditions; see Sen (1960). So $(R_d - \rho_d)/\hat{\sigma}_d$ can be used for approximate tests and confidence intervals.

In practice the sampling distribution of $(R_d - \rho_d)/\hat{\sigma}_d$ can be approximated by the Student-t distribution. Hinkley (1977) proposed a degrees of freedom

estimator for the t approximation of studentized jackknife estimators. Palachek & Schucany (1981) have shown that this procedure improves the coverage when estimating ρ using confidence intervals based on $\bar{R}$. This method is also useful for interval estimation of ρ_d.

The degrees of freedom estimator is given by

$$f_d = \frac{\frac{2}{n}(n-2)^2 s_d^4}{\frac{1}{n-1}\sum_{i=1}^{n}(V_i - R_d)^4 - \frac{n-1}{n}s_d^4}.$$

Thus an approximate $100(1-\alpha)\%$ confidence interval for ρ_d is given by

$$R_d - t_{\alpha/2}(f_d)\hat{\sigma}_d < \rho_d < R_d + t_{\alpha/2}(f_d)\hat{\sigma}_d,$$

where $t_\alpha(\nu)$ is the $(1-\alpha)$th quantile of the Student-t distribution on ν degrees of freedom. Some Monte Carlo evidence of the adequacy of the approximate confidence coefficients in the closely related one-group setting may be found in Palachek & Schucany (1981).

4. EXAMPLE

Consider the following hypothetical data set. Suppose that 20 judges have independently ranked 5 objects, leading to the rankings in Table 2. An investigator is interested in determining if the population consensus of the rankers agrees with the ordering given by $\underline{y} = (1,2,3,4,5)'$.

The Page test rejects H_0: "random rankings" in favor of an alternative that $\rho_1 > 0$. Moreover, the U-statistic approach (using the Spearman rank correlation coefficient) yields $R_1 = .585$, and the estimated variance of R_1 is

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\{R(\underline{X}_i, \underline{y}) - R_1\}^2 = .00423.$$

Using the t-approximation on $n-1 = 19$ degrees of freedom (since the $R(\underline{X}_i, \underline{y})$ are independent) leads to an approximate 95% confidence interval

$$.449 < \rho_1 < .721,$$

TABLE 2

Rankings of 5 Objects by 20 Judges

Judges	Objects A	B	C	D	E
1	1	3	2	4	5
2	1	5	2	4	3
3	1	5	2	3	4
4	1	4	2	3	5
5	1	3	2	4	5
6	1	3	2	4	5
7	1	4	3	2	5
8	3	4	2	1	5
9	1	4	2	3	5
10	1	5	2	4	3
11	1	2	3	4	5
12	1	4	2	3	5
13	1	4	2	3	5
14	2	5	1	3	4
15	1	5	2	3	4
16	1	4	3	2	5
17	1	4	2	3	5
18	1	3	2	4	5
19	1	5	4	3	2
20	1	4	2	3	5
Totals	23	80	44	63	90

which indicates that there is some agreement between the population of rankers and the predicted ordering.

The average internal rank correlation in this example is found to be $\bar{R} = .72$. Following the Palachek & Schucany approach, the jackknife variance estimator of $\bar{R}$ is found to be .00499, and the estimated degrees of freedom are $f_R = 8.21$. This leads to an approximate 95% confidence interval

$$.558 < \rho < .822.$$

Using the Bonferroni inequality these two intervals hold with an approximate confidence coefficient of .90. However, a sharp comparison between ρ_1 and ρ is not possible due to the overlap of the two intervals.

This problem can be circumvented by estimating ρ_d. The U-statistic obtained is $R_d = -.135$, and the jackknife variance estimator is $\hat{\sigma}_d^2 = .00339$. The estimated degrees of freedom are $f_d = 12.4$, which leads to an approximate 95% confidence interval

$$-.261 < \rho_d < -.009.$$

This interval is unambiguous in estimating that the external rank correlation is not substantial. In other words, the hypothesis that $\rho_d \geq 0$ is rejected at the .05 level in favor of an alternative that $\rho_d < 0$. Comparison of $\underline{y}$ with the average ranks

(1.15, 4.0, 2.2, 3.15, 4.5)

shows that the predicted ordering has misplaced object B relative to objects C and D.

5. CONCLUSIONS

The method presented here contrasts the internal and external rank correlation. This allows one to determine whether the agreement with a predicted order is "substantial" in light of the strength of the agreement within the population.

The degrees of freedom estimator should be used for small and moderate sized samples to avoid undercoverage of confidence intervals. However, this procedure is adaptive in that the estimated degrees of freedom are sometimes larger than n-1, leading to shorter, more precise intervals.

This work was partially supported by a contract with the Office of Naval Research and that sponsorship is gratefully acknowledged.

REFERENCES

Ehrenberg, A. S. C. (1952). On sampling from a population of rankers. Biometrika 39, 82-87.

Feigin, P. D. & Cohen, A. (1978). On a model for concordance between judges. J. Roy. Statist. Soc., Ser. B 40, 203-13.

Hinkley, D. V. (1977). Jackknife confidence limits using Student t approximations. Biometrika 64, 21-28.

Hollander, M. & Sethuraman, J. (1978). Testing for agreement between two groups of judges. Biometrika 65, 403-11.

Jonckheere, A. R. (1954). A test of significance for the relation between m rankings and k ranked categories. Brit. J. Statist. Psych. 7, 93-100.

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

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

Kraemer, H. C. (1981). Intergroup concordance: Definition and estimation. Biometrika 68, 641-46.

Lyerly, S. B. (1952). The average Spearman rank correlation coefficient. Psychometrika 17, 421-28.

Mallows, C. L. (1957). Non-null ranking models I. Biometrika 44, 114-30.

Page, E. B. (1963). Ordered hypotheses for multiple treatments: a significance test for linear ranks. J. Amer. Statist. Assoc. 58, 216-30.

Palachek, A. D. & Schucany, W. R. (1981). On approximate confidence intervals for measures of concordance. Technical Report #152, Department of Statistics, Southern Methodist University. (To appear in Psychometrika.)

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

Sen, P. K. (1960). On some convergence properties of U-statistics. Calcutta Statist. Assoc. Bull. 10, 1-18.

Spearman, C. (1904). The proof and measurement of association between two things. Amer. J. Psych. 15, 72-101.