A NOTE ON SPEARMAN'S RANK-ORDER CORRELATION AND MAXIMUM LIKELIHOOD RELIABILITY COEFFICIENTS

Harry E. Anderson, Jr.
University of Georgia

Introduction

It is well known that Spearman's rank-order correlation coefficient, r*, is numerically equal in terms of ranked data to the usual Pearsonian product-moment correlation coefficient, and that the latter coefficient is the maximum likelihood solution to the problem of association between two independent,normally distributed variables under the assumption that there are differences between the two population means and between the two population variances. Kendall's (10) tau coefficient is another rank-order correlation coefficient, and, though it is not the rank-order analogue of the Pearsonian coefficient, it does have certain advantages over the Spearman coefficient, such as in the computation of partial correlations. Correlation coefficients are quite useful in many kinds of association problems involving several types of assumptions, such as in reliability problems. The latter problems are quite common in the behavioral sciences, and in instances of ranking, there is a question of which rank-order coefficient to use with the data. The present paper shows that the Spearman rank-order correlation coefficient is, quite generally, the rank-order analogue of maximum likelihood reliability coefficients as well as the Pearsonian correlation.

The Spearman Rank-Order Correlation

The development of the Spearman rank-order correlation coefficient, r*, is well known among behavioral scientists (3, pp. 193-195; 16, pp. 202-213) so that only the major characteristics will be presented here for easy reference. The statistic, r*, is defined (6) for two ranked distributions, X and Y, using small letters for deviation-from-the-mean values,

$$(1) \qquad r^* = \frac{\Sigma xy}{\sqrt{(\Sigma x^2)(\Sigma y^2)}}$$

which can also be written (4)

$$(2) \qquad r^* = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma d^2}{2\sqrt{(\Sigma x^2)\ (\Sigma y^2)}}$$

where d is the difference between a given pair of ranks in X and Y. Now, the sum of N ranks is

$$\Sigma X = 1 + 2 + \ldots\ldots + N, \text{ or}$$

$$(3) \qquad \Sigma X = \frac{N\ (N+1)}{2} .$$

The sum of squares of the N ranks,

$$\Sigma X^2 = 1^2 + 2^2 + \ldots\ldots + N^2, \text{ is}$$

$$(4) \qquad \Sigma X^2 = \frac{N(N + 1)\ (2N + 1)}{6} .$$

The sum of squared deviations for ranks, using equations (3) and (4), is

$$(5) \qquad \Sigma x^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = \frac{N^3 - N}{12} .$$

Since for the N pairs of ranks in the two distributions, $\Sigma x^2 = \Sigma y^2$, substitution of equation (5) in equation (2) gives, after simplification,

$$(6) \qquad r^* = 1 - \frac{6\ \Sigma d^2}{N\ (N^2 - 1)}$$

which is the general computational form for r*. Olds (12, 13) provided various significance levels for Σd^2, but for N > 10 the statistic

$$t = r^* \sqrt{\frac{N - 2}{1 - r^{*2}}}$$

is distributed approximately as Student's t (10, pp. 47-48) and, with N-2 degrees of freedom, may be used to test the significance of an obtained r* under the null hypothesis.

Maximum Likelihood Association Solutions, Reliability, and r*

If two variables, X and Y, are normally distributed, the simultaneous probability distribution, $P(X_1, X_2, \ldots, X_N; Y_1, Y_2, \ldots, Y_N)$, for all N pairs of values is

(7) $P(X_1, X_2, \ldots, X_N; Y_1, Y_2, \ldots, Y_N) = c\, e^{\mu}$, where

$$c = (2\pi \sigma_x \sigma_y \sqrt{1 - \rho^2})^{-N}, \text{ and}$$

$$\mu = \frac{1}{2(1-\rho^2)} \sum_N \left[\frac{(X - \mu_x)^2}{\sigma_x^2} + \frac{(Y - \mu_y)^2}{\sigma_y^2} - \frac{2\rho (X - \mu_x)(Y - \mu_y)}{\sigma_x \sigma_y} \right]$$

and where σ_x and σ_y are the population standard deviations in X and Y, respectively; μ_x and μ_y are the population means in X and Y, respectively; and ρ is the population correlation, or index of association, between the two variables. Assuming $\sigma_x \neq 0$, $\sigma_y \neq 0$, and $\rho \neq 1$, the population parameters are estimable from the sample values, in terms of least squares, by the method of maximum likelihood which consists of taking partial derivatives of $P(X_1, X_2, \ldots, X_N; Y_1, Y_2, \ldots, Y_N)$ with respect to each of the five parameters, setting each of the resulting linear equations equal to zero, and solving the equations as a set of simultaneous equations for each of the five parameters. The maximum likelihood solutions (9) are

$$\mu_x = (\Sigma X)/N \doteq \bar{X},$$

$$\mu_y = (\Sigma Y)/N \doteq \bar{Y},$$

$$\sigma_x^2 = (\Sigma x^2)/N \doteq S_x,$$

$$\sigma_y^2 = (\Sigma y^2)/N \equiv Sy, \text{ and}$$

$$(8) \qquad \rho_5 = (\Sigma x\, y)/\sqrt{\Sigma x^2\, \Sigma y^2} \equiv r_5 \quad .$$

Aspects of association in reliability problems (viz., test/re-test, equivalent forms, and internal consistency) have received wide-spread attention throughout the behavioral sciences. Most recent approaches, such as used by Rajaratnam (14) utilize a variance ratio based on a linear model for the data, and the developments of Horst (5) and Cronbach (2) seem most signal. Horst presented a generalized coefficient of which the Spearman-Brown correction (for length) and the Kuder-Richardson 21 formulas are shown to be special cases. Cronbach's alpha coefficient, which is a more detailed, generalized development of Hoyt's (7) coefficient through specification that items can be scored other than 0 or 1, was shown to be the mean of all possible split-half (Pearsonian) correlation coefficients; moreover, the Kuder-Richardson formula 20 was shown to be a special case of alpha. For purposes of rank-order correlation in reliability problems, however, where N is small, the form of the distributions completely unknown,or merely for interim calculations in large studies, it is convenient to consider the maximum likelihood solutions for estimating the population parameters.

In the test/re-test and equivalent forms methods of reliability, we assume that the means of the two distributions may be different, because of practice or learning effects, but that the variances will be the same. The maximum likelihood solution (8) for the values in the probability function $P(X_1, X_2, \ldots, X_N; Y_1, Y_2, \ldots, Y_N)$, therefore, involves the specification of <u>four</u> parameters and their sample estimates; viz., ρ, μ_x, μ_y, and $\sigma_x = \sigma_y = \sigma$. In this type of problem, the association parameter results in the form of an intraclass correlation coefficient,

$$(9) \qquad {}_4 = \frac{\Sigma xy}{\frac{1}{2}(\Sigma x^2 + \Sigma y^2)} \equiv r_4 \quad ,$$

where the subscript 4 stands for the number of specified parameters. As noted previously for ranked distributions, $\Sigma x^2 = \Sigma y^2$, and the equivalence of equations (9) and (8) is readily evident; but then equation (9) is equivalent also to equation (1) so that r* is the rank-order analogue of the reliability coefficient, r_4.

A second reliability problem is exemplified in the split-half form where the underlying assumptions are that the means and variances of the two distributions are respectively the same. The maximum likelihood solution for the values in the probability function $P(X_1, X_2, \ldots, X_N; Y_1, Y_2, \ldots, Y_N)$, therefore involves the specification of <u>three</u> parameters and their sample estimates; viz., ρ, $\mu_x = \mu_y = u$, and $\sigma_x = \sigma_y = \sigma$. Here, the association parameter is

(10)
$$\rho_3 = \frac{2\Sigma XY - \frac{(\Sigma X + \Sigma Y)^2}{2N}}{\Sigma X^2 + \Sigma Y^2 - \frac{(\Sigma X + \Sigma Y)^2}{2N}} \equiv r_3$$

Using equations (3) and (4) in (10), for ranked distributions,

$$r_3 = \frac{12\Sigma XY - 3N^3 - 6N^2 - 3N}{N3 - N} \quad ,$$

but,

$$\Sigma XY = \Sigma xy - \frac{(\Sigma X)(\Sigma Y)}{N} \quad ,$$

$$= \frac{4N^3 + 6N^2 + 2N - 6\Sigma d^2}{12} \quad ,$$

so that

$$r_3 = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)} = r^* \quad .$$

The above result shows that r* is the rank correlation analogue of the maximum likelihood solution for the three-parameter reliability problem.

Most writers (e.g., 15) present the Spearman rank-order correlation coefficient as the non-parametric equivalent of the Pearsonian product-moment correlation coefficient since, indeed, it was defined on that basis. The results herein, however, show that r* is also the non-parametric analogue of two maximum likelihood solutions to reliability problems. The other major rank correlation,

tau, has some advantages over r* in that tau can be generalized to the problem of partial correlation (10, chap. 8). Both r* and tau, however, have equal power in rejecting the null hypothesis, and ".... neither (r*) nor tau will effectively adjust for errors arising from broad grouping or censoring" (1, p. 361). Both statistics are about 90 per cent as efficient as the Pearsonian correlation in detecting a relationship between two variables (6,p. 43). The results herein, however, suggest r* to be the more generally applicable statistic in reliability problems. Moreover, the coefficient of concordance (11), being a linear function of the average of all of the possible r* coefficients between m ranked distributions, is a kind of non-parametric analogue of Cronbach's alpha coefficient and therefore should be useful in reliability studies.

Correction for Ties

The formula for r* in equation (6) is quite general and is the one for use in reliability estimates as well as estimates for the product-moment correlation. If ties exist in either distribution, however, and a correction for ties is applied to the data, the resulting values estimating r_5 and r_3 or r_4 will, in general, be different. Kendall (10, pp. 25-36) shows that the effect of <u>each</u> set of tied ranks is a reduction in the sum of squared deviations in equation (5) by a factor

$$T = \frac{t^3 - t}{12} ,$$

where <u>t</u> is the number of tied observations in a given set. When the correction is applied for all sets of tied ranks, equation (5) becomes

$$\Sigma x^2 - \Sigma T_x = \frac{N^3 - N}{12} - \Sigma T_x .$$

When the correction term is applied to r* in estimating r_5, equation (2) becomes

(11)

$$r_5^* = \frac{2\left(\frac{N^3 - N}{12} - \sum T_x - \sum T_y - \sum d^2\right)}{2\sqrt{\left[\frac{N^3 - N}{12} - \sum T_x\right]\left[\frac{N^3 - N}{12} - \sum T_y\right]}}$$, or

(12)

$$= \frac{N^3 - N - 6\sum T_x - 6\sum T_y - 6\sum d^2}{12\sqrt{\frac{N^3 - N}{12}\left[\frac{N^3 - N}{12} - (\sum T_x + \sum T_y)\right] + \sum T_x \sum T_y}} .$$

Computationally, it is probably easier to work with equation (11) than with (12).

Similarly, when r* is used as a reliability estimator of r_4, equation (9) becomes

(13)

$$r_4^* = \frac{2\left(\frac{N^3 - N}{12} - \sum T_x - \sum T_y - \sum d^2\right)}{2\left(\frac{N^3 - N}{12} - \sum T_x - \sum T_y\right)}$$, or

(14)

$$= 1 - \frac{6\sum d^2}{N(N^2 - 1) - 6\sum T_x - 6\sum T_y} .$$

In this instance, it is probably easier to work with equation (14) than with (13). The same result is obtained when r* is used as an estimator of r_3 so that equation (14) is the general "corrected" formula for r* in reliability estimates. Only when $\sum T_x = \sum T_y$ will equations (12) and (14) produce the same results.

Summary

The developments in this paper show that the Spearman rank-order correlation coefficient is not only the non-parametric analogue of the Pearson product-moment correlation, but also of two maximum likelihood reliability coefficients. A convenient computational formula is presented when corrections are to be applied for tied ranks.

References

1. Carrol, J. B. The nature of the data, or how to choose a correlation coefficient. Psychometrika, 26 (1961), 347-372.

2. Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 16 (1951), 297-334.

3. Edwards, A. L. Statistical Methods for the Behavioral Sciences. New York: Rinehart, 1954.

4. Fruchter, B. and Anderson, H. E., Jr. Geometrical representation of two methods of linear least squares multiple correlation. Psychometrika, 26 (1961), 433-442.

5. Horst, P. A generalized expression for the reliability of measures. Psychometrika, 14 (1949), 21-31.

6. Hotelling, H. and Pabst, M. R. Rank correlation and tests of significance involving no assumption of normality. Ann. Math. Statis., 7 (1936), 29-43.

7. Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 6 (1941), 153-160.

8. Jackson, R. W. B. and Ferguson, G. A. Studies on the Reliability of Tests. Toronto: Department of Educational Research, University of Toronto, 1941, pp. 107-112.

9. Johnson, P. O. Statistical Methods in Research, New York: Prentice-Hall, 1949.

10. Kendall, M. G. Rank Correlation Methods. London: Griffin, 1948.

11. Kendall, M. G. and Smith, B. B. The problem of m rankings. Ann. Math. Statis., 10 (1939), 275-287.

12. Olds, E. G. Distributions of sums of squares of rank differences for small numbers of individuals. Ann. Math. Statis., 9 (1938), 133-149.

13. Olds, E. G. The 5% significance levels for sums of squares of rank differences and a correction. Ann. Math. Statis., 20, (1949), 117-118.

14. Rajaratnam, N. Reliability formulas for independent decision data when reliability data are matched. Psychometrika, 25 (1960) 261-271.

15. Scheffé, H. Statistical influence in the non-parametric case. Ann. Math. Statis., 14 (1943), 305-332.

16. Siegel, S. Non-parametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956.