Behavior Research Methods, Instruments, & Computers
1986, 18 (1), 45-46

An algorithm and program for calculation of Kendall's rank correlation coefficient

ALFRED L. BROPHY
Behavioral Science Associates, West Chester, Pennsylvania

Although less widely used than Spearman's rho, Kendall's (1938, 1975) rank correlation coefficient (tau) possesses advantages that may make it a preferred statistic: Its distribution under the null hypothesis is approximately normal even when the sample size (n) is fairly small[1]; it allows determination of confidence interval bounds; and it can be applied to partial correlation (Kendall, 1975). Nevertheless, many statistics texts do not discuss tau, and some that do offer only fragmentary information. A few texts (e.g., McCall, 1980; Walker & Lev, 1953) present rho in considerable detail, mention the advantages of tau, and then ironically say nothing more about the use or calculation of tau.

Kendall (1975) described the calculation of tau as involving comparison of every pair of ranks within each of the two distributions being studied. For a given pair of ranks x_i and x_j, where $i < j$, a score of +1 is assigned if $x_i < x_j$; a score of −1 is assigned if $x_i > x_j$; and a score of 0 is assigned if $x_i = x_j$. The statistic S, which is linearly related to tau, is then obtained by summing the products of the resulting scores for each corresponding pair of ranks in the two distributions. After a little simplification, the calculations can be summarized by the equation

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{sgn}\,[(x_j - x_i)(y_j - y_i)], \qquad (1)$$

where x and y represent ranks in the first and second distributions, respectively. Note that x and y can represent raw scores other than ranks and that this procedure requires neither sorting of the data nor assignment of ranks. Equation 1 comprises the main part of an algorithm for calculation of tau. [Tau = S/D, where D is the maximum possible value of S for a given n. When there are no tied ranks, $D = n(n-1)/2$. Ties decrease D, and thus increase tau for a given value of S.]

Kendall (1938, 1975) proposed shortcut methods for calculating S, one of which appears to be the most common current method. It requires the sorting into natural order of ranks in one distribution, together with ranks for corresponding subjects in the second distribution. A score of +1 or −1 is assigned to each pair of ranks in the second distribution, depending on whether the rank for the second member of the pair is greater than or less than the rank for the first member. A score of zero is assigned if there is a tie in either or both rankings. S is then obtained by summing the scores. In this method, raw scores other than ranks can be used, but data in one of the distributions must be sorted.

There is some confusion about the use of the latter method with tied ranks. For example, Siegel's (1956) influential manual does not explain that a zero score should be assigned when there is a tie in *either of* the two rankings.[2] Schaeffer and Levitt (1956) noted the same oversight in another source. In any case, the method is not particularly efficient in a computer implementation.

A Program. Table 1 shows a BASIC program for calculation of S and tau that applies the algorithm of Equation 1. The program compares favorably in features and efficiency with other published procedures (e.g., Böttcher & Posthoff, 1973; Knight, 1966; Stuart, 1977). The routine in Lines 60-100 was selected from several similar routines because it is relatively fast. Data can be entered either from the keyboard or from data statements.

The program counts the number of tied pairs in each distribution to correct tau for ties according to Equation 3.3 in Kendall (1975, p. 35). The program also calculates the variance of S for use in the normal approximation to the distribution of S so that the significance of S, and consequently tau, can be tested. The method of calculating the variance, which is valid for rankings with or without ties, is derived from Equation 14 in Kendall (1947). This equation, which is seldom used, is simpler to adapt for the program than equivalent formulas given by Kendall (1947, 1975, p. 55). Because the equation is for the variance of $2S$, it is first divided by 4 to obtain the equation for the variance of S:

$$\begin{aligned} s^2 = {} & [n(n-1)(n-2)/3 - \Sigma t(t-1)(t-2)/3] \\ & [n(n-1)(n-2)/3 - \Sigma u(u-1)(u-2)/3] \\ & /[n(n-1)(n-2)] + [n(n-1) - \Sigma t(t-1)] \\ & [n(n-1) - \Sigma u(u-1)]/[2n(n-1)], \end{aligned} \qquad (2)$$

where t is the number of scores in each set of ties in the first distribution, and u is the number of scores in each set of ties in the second distribution. Without tied scores, Equation 2 reduces to $n(n-1)(2n+5)/18$, the variance of S with no ties. Kendall (1947) showed that his Equation 14 also reduces to the formulas for the variance of $2S$ for all possible combinations of distributions containing ties, no ties, and dichotomies.

A correction for continuity of one unit is employed in the normal approximation, as recommended by Kendall (1975). The correction and the normal approximation itself appear satisfactory in most cases, unless there are numerous or lengthy ties.[3] (Burr, 1960, recommended a different correction when both distributions contain ties, and both Burr and Kendall, 1975, suggested other corrections for dichotomous distributions.)

The author's mailing address is: Behavioral Science Associates, P.O. Box 748, West Chester, PA 19381.

Table 1
BASIC Program to Calculate Kendall's Tau

```
10 DEFINT I,J,N,S-U
20 INPUT"Sample size "; N: N1=N-1: DIM X(N),Y(N)
30 INPUT"Enter data from (D)ata statements or (K)eyboard "; Z$
40 IF Z$="D" THEN FOR I=1 TO N: READ X(I),Y(I): NEXT I: GOTO 60
50 FOR I=1 TO N: PRINT"Subject #" I: INPUT" X, Y "; X(I),Y(I):
   NEXT I
60 FOR I=1 TO N1: X=X(I): Y=Y(I): T=0: U=0: FOR J=I+1 TO N
70 A=(X(J)-X)*(Y(J)-Y): IF A THEN S=S+SGN(A): GOTO 100
80 IF X=X(J) THEN T=T+1: T1=T1+1
90 IF Y=Y(J) THEN U=U+1: U1=U1+1
100 NEXT J: T2=T2+T*(T-1): U2=U2+U*(U-1): NEXT I
110 K=N*N1/2: B=(K-T1)*(K-U1): R=S/SQR(B)
120 PRINT"tau =" R, "S =" S
130 L=N*N1*(N-2): V=(L/3-T2)*(L/3-U2)/L+B/K
140 PRINT"Normal approximation:": Z=(ABS(S)-1)/SQR(V)
150 X=Z*Z: P=.5-SQR(1-EXP(-X*(.6366198-X*(9.564224E-03-X*.0004
    ))))/2: IF Z<0 THEN P=1-P
160 PRINT"  z =" Z, "p (one-tailed) = " USING"#.###"; P
170 END
```

Finally, the program estimates the one-tailed probability (p) corresponding to the approximated normal deviate, using Brophy's (1983) modification of Cadwell's (1951) compact approximation. The estimated p is accurate to three decimals, which is sufficient for the normal approximation of S.

Language, Time, and Memory Requirements. The program is written in IBM Personal Computer BASIC, a version of Microsoft BASIC. Little or no modification is necessary for use with most other BASIC dialects. The DEFINT statement in Line 10 can be removed, if necessary, without affecting the results.

The program occupies 722 bytes of memory. On the IBM PC microcomputer, a 12-subject example from Kendall (1975, pp. 55-56) ran in 2 sec, and a 100-subject problem ran in 73 sec. With data stored in data statements, less than 2.5K bytes were required to execute the 100-subject problem.

Availability. A listing of the program can be obtained without charge from the author.

REFERENCES

BEST, D. J. (1973). Extended tables for Kendall's tau. *Biometrika*, **60**, 429-430.

BEST, D. J., & GIPPS, P. G. (1974). Algorithm AS 71: The upper tail probabilities of Kendall's tau. *Applied Statistics*, **23**, 98-100.

BÖTTCHER, H. F., & POSTHOFF, C. (1973). Die mathematische Behandlung der Rangkorrelation—eine vergleichende Betrachtung der Koeffizienten von Kendall und Spearman. *Zeitschrift für Psychologie*, **183**, 201-217.

BROPHY, A. L. (1983). Accuracy and speed of seven approximations of the normal distribution function. *Behavior Research Methods & Instrumentation*, **15**, 604-605.

BURR, E. J. (1960). The distribution of Kendall's score S for a pair of tied rankings. *Biometrika*, **47**, 151-171.

CADWELL, J. H. (1951). The bivariate normal integral. *Biometrika*, **38**, 475-479.

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

KENDALL, M. G. (1947). The variance of τ when both rankings contain ties. *Biometrika*, **34**, 297-298.

KENDALL, M. (1975). *Rank correlation methods* (4th ed.). London: Griffin.

KNIGHT, W. R. (1966). A computer method for calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association*, **61**, 436-439.

MCCALL, R. B. (1980). *Fundamental statistics for psychology* (3rd ed.). New York: Harcourt Brace Jovanovich.

SCHAEFFER, M. S., & LEVITT, E. E. (1956). Concerning Kendall's tau, a nonparametric correlation coefficient. *Psychological Bulletin*, **53**, 338-346.

SIEGEL, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

STUART, A. (1977). Spearman-like computation of Kendall's tau. *British Journal of Mathematical & Statistical Psychology*, **30**, 104-112.

WALKER, H. M., & LEV, J. (1953). *Statistical inference*. New York: Holt.

NOTES

1. The normal approximation to tau is usually considered applicable when $n > 10$, although it sometimes provides less than three-decimal accuracy even when $n = 50$ (Best, 1973; Best & Gipps, 1974).

2. The numerical example that Siegel (1956, pp. 218-219) gave for tied ranks has no ties in the first distribution, which is used as the basis of the sorting. Based on his computational procedure, S has a unique value, and the result is correct. However, if the second distribution, which has three two-way ties, is used for the sorting, Siegel's procedure yields S values ranging from 22 to 28, depending on the arbitrary order of ranks in the first distribution that are associated with tied ranks in the second distribution. Tau would range from .34 to .43, with one-tailed probabililties between .07 and .03.

3. An anonymous reviewer, noting current reservations about use of Yates's correction with the chi-square distribution, requested information on the effect of the correction for continuity on the normal approximation to the distribution of S. Accordingly, the one-tailed probability (p) was calculated for every possible value of S (with no tied scores) for $n = 11$ through 20, 25, 50, and 100 using the normal approximation with and without a correction for continuity of unity. Results were compared with p values yielded by the Best and Gipps (1974) computer program, which is accurate to at least three decimal places. With the correction for continuity, the maximum absolute error of the approximation of p was .004; without the correction, the maximum error was .032. The corrected approximation had a smaller maximum absolute error at every value of n tested. The uncorrected approximation was, however, more accurate than the corrected approximation at low levels of p ($p < .01$ or .02, depending on n), although the advantage was never more than one unit in the third decimal place. Evaluation of the approximations is complicated when ties are present, but Best (1973) offered some evidence that the corrected approximation is fairly accurate when there are only a few ties in one distribution.

(Revision accepted for publication October 17, 1985.)