PSYCHOMETRIKA—VOL. 21, NO. 3
SEPTEMBER, 1956

RANK-BISERIAL CORRELATION

EDWARD E. CURETON

UNIVERSITY OF TENNESSEE

A formula is developed for the correlation between a ranking (possibly including ties) and a dichotomy, with limits which are always ± 1. This formula is shown to be equivalent both to Kendall's τ and Spearman's ρ.

Suppose we have two correlated variables, one represented by a ranking (possibly including ties) and the other by a dichotomy. The dichotomy may be considered a ranking concentrated into two multiple ties; its ties, however, do not represent equal measurements (or judgments of equality) on a continuous (or at least a many-step) variable. Rather, the ties represent a broad grouping of the data into two categories, or possibly an actual two-point distribution (sex, e.g.). Since the number of distinct ranks in the ranked variable will always be much greater than 2 and will equal N in the untied case, exact rank agreement of the two variables, pair by pair for each individual, is impossible. In this situation we desire a coefficient which will still have attainable limits ± 1 in all circumstances. It should be $+1$ when all ranks in the "higher" category of the dichotomy exceed all ranks in the "lower" category, and -1 when all ranks in the "lower" category exceed all ranks in the "higher" category. It should be strictly non-parametric, i.e., defined wholly in terms of inversions and agreements between pairs of rank-pairs, without use of such concepts as mean, variance, covariance, or regression. Finally, it should resemble the usual rank correlation coefficients in some reasonable sense.

Let R_x represent the dichotomy, with categories R_x+ and R_x-, and let R_y represent the ranked variable. Ties in R_y are to be handled by the mid-rank method. We then arrange the ranks R_y in as nearly as possible the natural order (N, $N-1$, $\cdots$, 1), with rank N "high" and rank 1 "low," and allocate them to the categories R_x+ and R_x- as in the following example:

$R_x +$	$R_x -$	*Inv.*	*Agr.*
9.5			4
9.5			4
8			4
6.5			3
	6.5	2	
4.5			3
4.5			3
	2.5		
	2.5		
	1		
$N_1 = 6$	$N_2 = 4$	$Q = 2$	$P = 21$

(1)

No two R_y ranks may be in the same row, but in case of a tie in R_y with one member falling under $R_x +$ and the other under $R_x -$, the relation between the row and column allocations is immaterial. Thus, in (1), the first 6.5 might as well have been allocated to $R_x -$ and the second to $R_x +$.

With this arrangement, there is an *inversion* at any given number under $R_x -$ for every smaller number under $R_x +$. Thus, at 6.5 in $R_x -$ we have two inversions, one for each of the values 4.5 under $R_x +$. There is also an *agreement* at any given number under $R_x +$ for every smaller number under $R_x -$. Let Q be the total number of inversions, and let P be the total number of agreements.

With this method of allocation to rows and columns, perfect positive correlation would require that all numbers under $R_x +$ should be larger than all numbers under $R_x -$, and in this case we should find that $Q = 0$ and $P = P_{\max}$. Perfect negative correlation would require that all numbers under $R_x +$ should be smaller than all numbers under $R_x -$, and in this case we should find that $P = 0$ and $Q = Q_{\max}$. Also, $P_{\max} = Q_{\max}$, since the two result merely from an interchange of the sets of numbers under $R_x +$ and $R_x -$. Our coefficient may therefore be of the form

$$r_{RB} = (P - Q)/P_{\max} . \tag{2}$$

It will be $+1$ if $Q = 0$ and $P = P_{\max}$, -1 if $P = 0$ and $Q = Q_{\max} = P_{\max}$, and 0 if $P = Q$.

To determine $P_{\max}$, we note first that in the situation in which the coefficient is $+1$, there will be N_2 agreements for every number under $R_x +$, or N_1N_2 in all. There is one case, however, so far passed over, in which $P_{\max}$ cannot be as great as N_1N_2. This case is illustrated in our example. If we set up explicitly the situation for $P = P_{\max}$ with these data, we have:

$R_x +$	$R_x -$	*Inv.*	*Agr.*
9.5			4
9.5			4
8			4
6.5			4
6.5			4
4.5			3
	4.5		
	2.5		
	2.5		
	1		
$N_1 = 6$	$N_2 = 4$	$Q = 0$	$P = 23$
			$N_1N_2 = 24$

(3)

One agreement is lost because the lowest rank under $R_x +$ is tied with the highest under $R_x -$. In other cases there might be a triple or multiple tie at the point of dichotomy. We shall term a tie at this point a *bracket tie.* For any bracket tie, the value of P_{max} will be reduced from N_1N_2 by unity for every *pair* of members of this tie one of which is under $R_x +$ and the other under $R_x -$, after R_y has been rearranged to be as nearly as possible in the natural order and allocation under $R_x +$ and $R_x -$ is made in such a manner as to preserve the original values of N_1 and N_2. If t_1 is the number under $R_x +$ participating in the bracket tie, and t_2 the number under $R_x -$, $P_{max} = N_1N_2 - t_1t_2$, and our formula becomes

$$r_{RB} = \frac{P - Q}{N_1N_2 - t_1t_2}. \tag{4}$$

Physically, it is not necessary to rearrange the original data in order to compute t_1t_2. We merely draw a horizontal line across columns $R_x +$ and $R_x -$ in (1), at a level which leaves N_1 cases above the line and N_2 below it. Since the original arrangement in (1) was with R_y in as nearly as possible the natural order, a bracket tie will then consist of any group of identical numbers, some immediately above and some immediately below this line. The number above is t_1 and the number below is t_2. For the example of (1), we find by (4):

$$r_{RB} = \frac{21 - 2}{(6)(4) - (1)(1)} = .826.$$

Clearly r_{RB} is a Kendall-type coefficient, since Q and P are the numbers of unweighted inversions and agreements, respectively (**2**). But it is also a Spearman-type coefficient. Durbin and Stuart (**1**) have shown that, in the untied case, Spearman's coefficient is given by $(U - V)/(U - V)_{max}$, where

V is the number of inversions and U the number of agreements, each weighted by the difference between the two ranks concerned. It is easily shown that the difference which supplies the weight may come from either R_y or R_x, and it is also easy to find $(U - V)_{max}$ for the cases corresponding to Kendall's ρ_a and ρ_b. The writer has not been able to prove in these cases that the values given by $(U - V)/(U - V)_{max}$ are necessarily equal in general to those given by the corresponding formulas based on Σd^2, but he has verified each of them on several sets of numerical data.

In the present case, we need merely note that all R_x values bracketed under $R_x +$ would have one mid-rank value, and all those bracketed under $R_x -$ another. If, then, we weight each inversion and agreement by the corresponding rank-difference in R_x, all weights will be equal (and equal to the difference between the two mid-rank values), and it follows at once that r_{RB} is a Spearman-type coefficient.

The hypothesis that r_{RB} differs only by chance from $\rho_{RB} = 0$ may be tested by the Mann-Whitney extension of the Wilcoxon test (3).

REFERENCES

1. Durbin, J. and Stuart, A. Inversions and rank correlation coefficients. *J. roy. statist. Soc.*, 1951, 13, 303-9.
2. Kendall, M. G. Rank correlation methods. London, Charles Griffin and Co., 1948, Ch. 3.
3. Mann, H. B. and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. math. Statist.*, 1947, 18, 50-60.

Manuscript received 11/14/55