A GENERALIZED CORRELATION COEFFICIENT: A STATISTIC FOR TESTING SIMILARITY BETWEEN LINKED OBSERVATIONS IN TWO SAMPLES

D. S. Burdick and H. H. Winsborough, Duke University

1. Frequently, in samples of observations, certain pairs of observations are suspected to be more (or less) nearly alike than is generally true for the population. For example, when the observations are made in sequence over a period of time, it may happen that successive observations resemble each other more than do observations which are not adjacent in time. Statistics which measure the relation between successive observations are Von Neumann's ratio and the serial correlation coefficient (1, 8).

Geary (3) proposed an obvious generalization of Von Neumann's ratio, which he called the contiguity ratio, to measure similarity of linked observations in a single sample when the linkage pattern is arbitrary. Geary used the contiguity ratio to test whether or not adjacent counties show similarity with respect to a certain attribute. Here an observation of the attribute for one county is linked with an observation for another county if the two counties are contiguous geographically.

The contiguity ratio is also useful in certain sociological situations. If observations are made on each person in a group of people, observations on persons who know each other might be expected to show a greater degree of similarity than observations on persons who are not acquainted. Winsborough, Quarantelli, and Yutzy (9) have discussed some of the applications of the contiguity ratio to sociology.

The case of two samples with links across the samples can also arise in sociological situations. The group of people may be divided into two categories, e.g., men and women. It may be desired to test whether or not men and women who are acquainted exhibit similar characteristics. A statistic for making such tests is introduced in this paper. This statistic is a generalization of the correlation coefficient in the same way that Geary's contiguity ratio is a generalization of the Von Neumann ratio.

In the next section the generalized correlation coefficient r will be introduced, and its mean and variance will be computed under the assumption that for each of the two samples the joint distribution of the observations is symmetric. Section 3 will contain examples of the application of r.

2. Let $x_1, \ldots, x_{n_1}$ be a sample from some population, and let $y_1, \ldots, y_{n_2}$ be another sample from another population. It is assumed that the joint distribution of the x_i's is symmetric and similarly for the y_j's. That is to say, different arrangements of the x_i's (y_j's) have the same likelihood of occurrence. Suppose further that there is a pattern of links between the x_i's and the y_j's. Let L be the number of these links. The generalized correlation coefficient r is defined by

$$(2.1) \qquad r = \frac{1}{L} \sum_{i \sim j} \sum u_i v_j$$

where

$$u_i = (x_i - \overline{x})/s_x, \quad v_j = (y_j - \overline{y})/s_y, \quad n_1\overline{x} = \sum_{i=1}^{n_1} x_1, \quad n_2\overline{y} = \sum_{j=1}^{n_2} y_j, \quad (n_1 - 1)s_x^2 = \sum_{i=1}^{n_1} (x_i - \overline{x})^2, \quad (n_2 - 1)s_y^2 = \sum_{j=1}^{n_2} (y_j - \overline{y})^2 .$$

The notation $\sum_{i \sim j} \sum$ indicates that the summation extends over all pairs (i, j) for which x_i and y_j are linked.

The statistic r is not quite an exact generalization of the standard product-moment correlation coefficient. A sample from a bivariate distribution is expressable in the form of the previous paragraph with $n_1 = n_2 = L$, where each observation x_i is linked with the corresponding observation y_i. In this case the product moment correlation coefficient is usually defined as $\frac{1}{L-1} \sum_{i \sim j} \sum u_i v_j$ instead of $\frac{1}{L} \sum_{i \sim j} \sum u_i v_j$. The difference is a trivial one, but the definition of r that we use permits a somewhat simpler expression for the variance of r.

It is worth mentioning that the range of possible values for r is unlimited. Of course, if the linkage structure is that described in the preceding paragraph, then r must satisfy $-(L-1)/L < r < (L-1)/L$, but for certain other linkage structures any real number is a possible value for r.

We now wish to derive the mean and variance of r under the assumption that each x_i is independent of each y_j regardless of whether or not x_i and y_j are linked. Then each u_i is independent of each v_j and therefore

$$(2.2)\quad E(r) = \frac{1}{L} \sum_{i \to j} \sum E(u_i v_j) = \frac{1}{L} \sum_{i \to j} \sum E(u_i)\, E(v_j)$$

$$(2.3)\quad E(r^2) = E\left(\left(\frac{1}{L} \sum_{i \to j} \sum u_i v_j\right)^2 \right) =$$

$$\frac{1}{L^2} \Bigg(\sum_{i \to j} \sum E(u_i^2)E(v_j^2) + \sum_{i \to j_1, j_2} \sum \sum_{j_1 \neq j_2} E(u_i^2)E(v_{j_1} v_{j_2}) + \sum_{i_1, i_2 \to j} \sum \sum_{i_1 \neq i_2} E(u_{i_1} u_{i_2})E(v_j^2)$$

$$+ \sum_{i_1 \to j_1, i_2 \to j_2} \sum \sum \sum_{\substack{i_1 \neq i_2 \\ j_1 \neq j_2}} E(u_{i_1} u_{i_2})E(v_{j_1} v_{j_2}) \Bigg) .$$

The summations above with unequal subscripts extend over all ordered pairs of the subscript which satisfy the linkage conditions. Thus if u_1 is linked to v_1 and to v_2, the term $u_1 v_1 v_2$ will occur twice: once as $u_1 v_1 v_2$ and again as $u_1 v_2 v_1$.

In order to evaluate (2.2) and (2.3) we evaluate $E(u_i)$, $E(u_i^2)$, $E(u_{i_1} u_{i_2})$, and $E(v_j)$, $E(v_j^2)$, $E(v_{j_1} v_{j_2})$. By definition of the u_i's and v_j's we have

$$\sum_{i=1}^{n_1} u_i = O = \sum_{j=1}^{n_2} v_j, \quad \sum_{i=1}^{n_1} u_i^2 = n_1 - 1, \quad \sum_{j=1}^{n_2} v_j^2 = n_2 - 1.$$

Since $\sum_{i=1}^{n_1} u_i = O$, we have $O = E(O) = E(\sum_{i=1}^{n_1} u_i) = \sum_{i=1}^{n_1} E(u_i)$. But since the joint distribution of the x_i's is symmetric, the joint distribution of the u_i's is also symmetric, and therefore $E(u_i)$ is the same for each i. Thus $O = \sum_{i=1}^{n_1} E(u_i) = n_1 E(u_i)$ which implies that $E(u_i) = O$. In a similar manner it can be shown that $E(v_j) = O$, $E(u_i^2) = (n_1 - 1)/n_1$, and $E(v_j^2) = (n_2 - 1)/n_2$. To evaluate $E(u_{i_1} u_{i_2})$ observe $\sum_{i=1}^{n_1} u_i^2 + \sum_{i_1 \neq i_2} \sum u_{i_1} u_{i_2} = (\sum_{i=1}^{n_1} u_i)^2 = O$.

Taking expectations and again making use of the symmetry of the joint distribution, we have $n_1(n_1 - 1)E(u_{i_1} u_{i_2}) = -(n_1 - 1)$ or $E(u_{i_1} u_{i_2}) = -\frac{1}{n_1}$.

Similarly, $E(v_{j_1} v_{j_2}) = -\frac{1}{n_2}$.

Substituting the results of the preceding paragraph into (2.2) and (2.3) yields

$$(2.4)\quad E(r) = \frac{1}{L} \sum_{i \to j} \sum E(u_i)E(v_j) = O$$

$$(2.5)\quad E(r^2) = \mathrm{var}(r) = \frac{1}{L^2} \Bigg(\sum_{i \to j} \sum (n_1 - 1)(n_2 - 1)/n_1 n_2 + \sum_{i \to j_1, j_2} \sum \sum_{j_1 \neq j_2} -(n_1 - 1)/n_1 n_2 + \sum_{i_1, i_2 \to j} \sum \sum_{i_1 \neq i_2} -(n_2 - 1)/n_1 n_2 +$$

$$\sum_{i_1 \to j_1, i_2 \to j_2} \sum \sum \sum_{\substack{i_1 \neq i_2 \\ j_1 \neq j_2}} 1/n_1 n_2 \Bigg) .$$

To complete the evaluation we must count the number of terms in each of the sums in (2.5). The number of terms in the first sum is clearly L. The number of terms in the other sums can be expressed in terms of the following quantities: m_i = the number of links involving x_i, f_j = the number of links involving y_j. The number of ordered pairs of links to x_i is then $m_i(m_i - 1)$. Thus for each i the number of terms which occur in the second sum in (2.5) is $m_i(m_i - 1)$. The total number of terms in the second sum is therefore $\sum_{i=1}^{n_1} m_i(m_i - 1) = \sum_{i=1}^{n_1} m_i^2 - L$ since $\sum_{i=1}^{n_1} m_i = L$. Similarly, the total number of terms in the third sum is $\sum_{j=1}^{n_2} f_j(f_j - 1) = \sum_{j=1}^{n_2} f_j^2 - L$.

To obtain an expression for the number of terms in the fourth sum in (2.5) observe that if x_i and y_j are linked, the number of links which involve either x_i or y_j is $m_i + f_j - 1$. The number of links which involve neither x_i nor y_j is therefore $L - m_i - f_j + 1$. The total number of terms in the fourth sum is then $\sum_{i \to j} \sum (L - m_i - f_j + 1)$

$= L^2 + L - \sum_{i \neq j} \sum m_i - \sum_{i \neq j} \sum f_j$. In the sum $\sum_{i \neq j} \sum m_i$ the term m_i appears once for each link involving x_i . Since there are m_i such links in all, we have $\sum_{i \neq j} \sum m_i = \sum_{i=1}^{n_1} m_i^2$. Similarly, $\sum_{i \neq j} \sum f_j = \sum_{j=1}^{n_2} f_j^2$. Thus, the number of terms in the fourth sum can be expressed as $L^2 + L - \sum_{i=1}^{n_1} m_i^2 - \sum_{j=1}^{n_2} f_j^2$. As a check on our results the total number of terms in all the sums should be L^2. We have $(L) + (\sum_{i=1}^{n_1} m_i^2 - L) + (\sum_{j=1}^{n_2} f_j^2 - L) + (L^2 + L - \sum_{i=1}^{n_1} m_i^2 - \sum_{j=1}^{n_2} f_j^2) = L^2$ as expected.

Substituting these results into (2.5) yields

$$E(r^2) = \frac{1}{n_1 n_2 L^2} \left((n_1 - 1)(n_2 - 1)L - (n_1 - 1)(\sum_{i=1}^{n_1} m_i^2 - L) - (n_2 - 1)(\sum_{j=1}^{n_2} f_j^2 - L) + L^2 + L - \sum_{i=1}^{n_1} m_i^2 - \sum_{j=1}^{n_2} f_j^2 \right) = \frac{1}{L^2} \left(L + L^2/n_1 n_2 - \frac{1}{n_1} \sum_{j=1}^{n_2} f_j^2 - \frac{1}{n_2} \sum_{i=1}^{n_1} m_i^2 \right)$$. We can summarize the results of this section by

(2.6) $E(r) = 0$.

(2.7) $\mathrm{Var}(r) = \frac{1}{L^2} \left(L + L^2/n_1 n_2 - \frac{1}{n_2} \sum_{i=1}^{n_1} m_i^2 - \frac{1}{n_1} \sum_{j=1}^{n_2} f_j^2 \right)$

The generalization of the assumptions made in deriving the mean and variance of r is worth emphasizing. The assumption of symmetry of the joint distributions is satisfied whenever the two samples are random samples from any two populations. This assumption is also satisfied when all within sample permutations of the observations are considered equally likely to have occurred. It is understood that the permutations in question do not affect the linkage structure, i.e., if x_i and y_j are linked originally and after permuting x_i becomes x_i' and y_j becomes y_j', then x_i' and y_j' will be considered linked for the purpose of computing r.

In the case where $L = n_1 = n_2 = n$, $m_i = f_j = 1$ the statistic r is just the product-moment correlation coefficient multiplied by $(n-1)/n$ as was mentioned earlier in this section. We have for this case

$$\mathrm{Var}\left(nr/(n-1) \right) = \frac{n^2}{(n-1)^2} \cdot \frac{1}{n^2} \left(n + \frac{n^2}{n^2} - \frac{n}{n} - \frac{n}{n} \right) = \frac{1}{n-1}$$. This result under the assumption that the x_i's and y_j's are independent continuous variates is given as an exercise on page 396 of the book by Kendall and Stuart (5) .

3. Potential uses of the generalized correlation coefficient in social research are easy to suggest. This section will describe several uses of the statistic and present two examples: one a reanalysis of data drawn from a classic sociological investigation and another using unpublished data.

In introducing the generalized correlation coefficient, we have alluded to its applications to the analysis of a sociometric matrix. A number of problems in social research seem formally identical to the problem of assessing the similarity or dissimilarity of two kinds of persons linked by friendship ties. In studies of formal organizations, for instance, one might be interested in the similarity in output between linked members of two strata within the bureaucracy. Substantively, links in this problem might be defined as ones of friendship, the flow of work, the pattern of consulting, or participation in the same chain of command.

Consider, for example, the well known Roethlisberger and Dickson investigations of the bank wiring room (7). This study investigated one work group involved in the assembly of telephone equipment. Within the work group were wiremen, who performed one kind of operation and solderers, who performed another. Among these men was a complex net of social relationships: some being friendly, some playing games together at lunch hour, some arguing about whether the window should be open or closed, some generally antagonistic to one another. A signal interpretation of this investigation was that the level and the quality of output of the men in the bank wiring room was related to their position in this net of social relationships. This interpretation, however, was intuitively derived from inspection of the relationships and the production scores -- a method not well suited to working out the complexities of what kinds of links are associated with what kinds of measures.

In a recent paper the contiguity ratio has been used to re-analyze some of this data with rather interesting results(9). There it seemed that sanctioning of deviant production levels operated less through refusal to interact (i.e., play games together at lunch), and more through the expression of sentiment (i.e., the expression of antagonism). That investigation, however, considered only the relationships between the wiremen and ignored the solderers. Using the generalized correlation coefficient, however, it is possible to investigate whether, for instance, wiremen and solderers playing games together have similar output. We can do this in spite of the fact that the mean levels of output and the variance of output for the two groups are different.

We undertook such a re-analysis, primarily as illustrative of the uses of the technique. A fairly full set of connections between solderers and wiremen exists when connection is defined as playing games together. Output quantities are not available for solderers but measures of the quality of work are available for solderers and wiremen. The generalized correlation for quality of output between solderers and wiremen linked by playing games together was found to be .39. When links are defined as who gets into arguments about windows, another fairly full set of connections, the generalized correlation is lower, .26.

It may very well be that the sample sizes in these cases and the number of connections are too small to presume that the generalized correlation is normally distributed under the null that $r = 0$. We have worked out the variances for these two examples according to formula (2.7) anyway and find that by the normal test the correlation for playing games can be regarded as significant at the one per cent level while the correlation for arguments cannot be regarded as significant even at the five per cent level.

This finding, tenuous though it may be, seems of some heuristic value. Within the group of wiremen it had been previously suggested that participation in games was not related to output. Between wiremen and solderers we have some indication of a relationship. The possibility that patterns of interaction may be differently related to output within and between occupational groups provides an interesting addition to Homans's discussion of the relationship between interaction and the variable he describes as "activity" (4).

In the preceding example the coefficient has been used in its most general form. A special case seems of enough importance to deserve separate comment. This is the case in which members of one sample are connected to several members of the second sample but members of of the second sample are connected to only one member of the first. This condition obtains in many hierarchical situations. It obtains in the relationship between aggregates and their parts. It also obtains in the relationship between families over a generation.

This last observation suggests the potential applications of the statistic to the study of inter-generational social mobility. The generalized correlation coefficient provides the possibility of measuring the similarity between indicators of the socio-economic status of fathers and that of all their sons. In providing this measure, the coefficient controls differences in the mean and variance of socio-economic status between the two groups--differences which one would attribute to structural changes in the society rather than to the degree of openness of the society. A generalized correlation computed between socio-economic status of a sample of fathers and their sons would accomplish this standardization more accurately than many present techniques by using a more accurate estimate of the mean and variance of the variable in each generation.

Consider the following example. A recent study investigated retirement and pre-retirement problems of a non-random sample of older white couples living in the Piedmont region of North Carolina.[1] This study collected data on the educational level and the present or last occupation of men in the sample. The present occupational and educational levels of each of the sons in the labor force were also collected. Using a recently devised index of socio-economic status for occupations (6) and years of education as variables and combining both generalized correlations between fathers and sons and Pearsonian correlations between variables within generations, it is possible to to provide an interesting description of the mobility process within this sample. Table 1 provides these data. In that table .79 is the Pearsonian correlation between the level of education and the index of occupational status for fathers and .72 the Pearsonian correlation between the variables for sons. (These values are rather higher than similar correlations found

1. The study was financed through a grant made by the Ford Foundation for "Socio-Economic Studies of Aging." The data were collected between March 1960 and March 1961.

in other samples -- probably a result of over-representation of the tails of the occupational and educational distribution in the sample.) Other values in the table are generalized correlations.

The findings are interesting in that their pattern supports the model of inter-generational mobility suggested recently by Duncan and Hodge (2). In both generations the association between education and occupation is high while the intergenerational association between fathers' and sons' occupation, although significant, seems lower. It is also interesting that all generalized correlations are of about the same order of magnitude, with the association between fathers' occupation and sons' education being the highest and the association between fathers' education and sons' occupation being the lowest. This finding, perhaps, supports the notion that the major factor in the inheritance of status is related to the father's ability to purchase an education for his son.

One would be disinclined to push the analysis of these data farther because of the unsatisfactory nature of the sample. These findings, however, seem to indicate that further investigation of the use of the generalized correlation in the study of intergenerational mobility may be fruitful.

In summary, then, we feel that the generalized correlation coefficient is a statistic which should be a useful tool in the sociologist's repertoire -- one which deserves both empirical use and mathematical elaboration.

Table 1

Pearsonian and Generalized Correlations Among Level of Education and Occupational Socio-Economic Status of Fathers and All Their Sons[a]

	Fathers'		Sons'	
	Educa.	Occup.	Educa.	Occup.
Fathers'				
Education	--	.79	.55	.53
Occupation		--	.57	.56
Sons'				
Education			--	.72
Occupation				--

[a] All correlations are significantly different from zero at the .01 level.

REFERENCES

1. Dixon, Wilfred J., "Further Contributions to the Problem of Serial Correlation," Ann. Math. Stat., Vol. 15 (1944), pp. 119-144.
2. Duncan, Otis Dudley and Hodge, Robert W., "Education and Occupational Mobility: A Regression Analysis," Am. Jo. Sociology, Vol LXVIII (1963), pp. 629-644.
3. Geary, R. C., "The Contiguity Ratio and Statistical Mapping," Incorporated Statistician, Vol. 5 (1954), pp. 115-145.
4. Homan, George C., The Human Group. New York: Harcourt, Brace and Co., 1950, pp. 118-119.
5. Kendall, M. G. and Stuart, Alan, The Advanced Theory of Statistics Vol. 1, Three Volume Edition. London: Charles Griffin and Company, Ltd., 1958.
6. Reiss, Albert J., Jr., with collaborators, Occupations and Social Status. New York: The Free Press, 1961, Appendix B.
7. Roethlisberger, F. L., and Dickson, William J., Management and the Worker. Cambridge: Harvard University Press, 1946, pp. 379-548.
8. vonNeumann, J., "Distribution of the Ratio of the Mean Square Successive Difference to the Variance," Ann. Math. Stat., Vol. 12 (1941), pp. 367-395.
9. Winsborough, H. H., Quarantelli, E. L., and Yutzy, Daniel, "The Similarity of Connected Observations," Am. Sociological Rev., (forthcoming).