# GROUP-RANK CORRELATION COEFFICIENT

*By* S. V. SIMHADRI

(*Department of Electrical Communication Engineering Indian Institution of Science Bangalore-12, India*)

[Received: October 30, 1972]

## ABSTRACT

*A new correlation coefficient, the group rank correlation coefficient $r_g$ is defined and compared with the product-moment and rank correlation coefficients. Formula for $r_g$ is specifically derived for the case of 5 group ranks under the assumption of normal populations. A problem is worked out to illustrate the advantages of the group rank coefficient.*

## INTRODUCTION

A number of correlation coefficients have been proposed in Statistical Theory [1]. The most basic of them is the Pearson product-moment correlation cofficient. Second in vouge is the Spearman rank correlation coefficient. It has advantages over the Pearson coefficient in the case of calculation, in case of non existence of quantitative data, and in the case where only correlation between ranks is required. Its disadvantages lie in the nonutilisation of all the data available, in the resulting of errors in case of ties among ranks, and in excessive labour in the case of a large population. The proposed group correlation coefficient takes into account the distributions of the two variables in contrast to the Spearman coefficient which is a nonparametric quantity; this results in the elimination of the disadvantages of the Spearman coefficient while retaining all its advantages. The distributions are assumed to be normal and it is justified in practice if a fairly large population is concerned. Also, when the data of the two variables are corrupted with noise, the group correlation coefficient might be expected to give more reliable results than the other two coefficients.

The group coefficient is similar to the rank coefficient in that the original scores are replaced by ranks. If the population is $n$, in the calculation of rank coefficient original scores are replaced by numbers from 1 to $n$, while in the calculation of group coefficient original scores are replaced by numbers 1 to $r$ where $r$ is very small compared to $n$. By definition, the ranks

1,2, 3. . .$r$ have a normal distribution, that is, the number of original scores replaced by 1 and $r$ are equal, the number of original scores replaeed by 2 and $r$-1 are equal etc., and these numbers are in turn distributed normally.

## DEFINITION

The group correlation coefficient is defined as

$$r_g = \frac{\text{cov}\ (x, y)}{\sigma_x \cdot \sigma_y} \quad \cdots \qquad [1]$$

$$x, y = 1, 2, 3 \cdots r \ (r << n)$$

where cov $(x, y)$ = the covariance between $x$ and $y$

$\sigma_x$ = standard deviation of $x$

$\sigma_y$ = standard deviation of $y$

$n$ = the number of pairs of $x$ and $y$

and $r$ = the number of group ranks chosen.

## SPECIMEN CALCULATIONS OF $r_g$

Let $r = 5$

Let the individuals $a_1, a_2, \cdots a_i \cdots a_n$ have $x$ and $y$ scores as $x_1, x_2 \cdots x_i, \cdots x_n$, and $y_1, y_2, \cdots y_i \cdots y_n$ respectively where $x_i$'s and $y_i$'s are the numbers 1, 2, 3, 4 and 5. If perfect correlation were to exist between $x$ and $y$, then $x_i = y_i$ for all $i$'s. If not, let us write

$$d_i = x_i - y_i \quad \cdots \qquad [2]$$

Squaring $d_i$, and summing over all $n$, it can be seen after some rearrangement that

$$\sum_i x_i y_i = \frac{1}{2} \sum_i x_i^2 + \frac{1}{2} \sum_i y_i^2 - \frac{1}{2} \sum_i d_i^2$$

Assuming identical normal disributions for $x$ and $y$ it can be shown that (see Appendix)

$$\sum_i x_i^2 = \sum_i y_i^2 \simeq 10n \qquad [3]$$

$$\frac{1}{n} \sum_i x_i = \frac{1}{n} \sum_i y_i \simeq 3 \qquad [4]$$

and $$\sigma_x = \sigma_y \simeq 1 \qquad [5]$$

Now, by definition (1)

$$\text{cov}\ (x, y) = \frac{1}{n}\sum_i x_i . y_i - \frac{1}{n_2}\sum_i x_i \cdot \sum_i y_i \qquad [6]$$

From (2), (3) and (4)

$$\text{cov}\ (x, y) = 1 - \frac{1}{2n}\sum_i d_i^2 \qquad [7]$$

Therefore from (1), (5) and (7),

$$r_g = 1 - \frac{1}{2n}.\sum_i d_i^2 \qquad [8]$$

It can be seen that $r_g = 1$ or perfect correlation exists between $x$ any $y$ when $d_i^2 = 0$ i.e. $x_i = y_i$. Similarly it can be seen that $(r_g)_{min} = -1$

## General Formula for Calculating $r_g$

The general formula for obtaining $r_g$ when the number of group ranks is $g$ is given by

$$r_g = 1 - \frac{1}{2n\,k_g^2}\sum_i d_i^2 \qquad [9]$$

where $r_g$ denotes the group rank correlation co-efficient when two identical gaussian populations have each been divided into $g$ ranks, and $k_g^2$ is a constant. Table 1 gives the values of $k_g^2$ as a function of $g$.

## A Worked Example

Table 2 gives the Matriculation marks ($x$), the B.Sc., marks ($y$), and the Selection examination marks ($z$) of 93 students who appeared for selection for the B.E. degree course in Electrical Communication Engineering in the Indian Institute of Science in 1969. The product moment correlation co-efficicnts between $x$ and $y$, between $y$ and $z$, and between $z$ and $x$ are as follows :

$$r_{xy} = 0.08 \qquad r_{yz} = 0.20 \qquad r_{zx} = 0.56$$

The group-rank correlation coefficients have also been calculated when the number of group ranks chosen is from 2 to 10. The ranks have been allotted to the scores on the basis of normal distribution. For example when $r_g = 5$, the number of scores which have been given in 1st and 5th ranks are approximately 6 per cent each of the total population, those given 2nd and 6th ranks are approximately 25 per cent each and those given the

3rd rank are approximately 38 per cent of the total population. As the mean ($m$) and standard deviation ($\sigma$) of the scores have been known, a second method of grade ranking also has been adopted and found to give almost the same ranks as the above procedure. Table 3 indicates the method of allotting ranks. Table 4 shows the group rank correlation coefficients between $x$ and $y$, between $y$ and $z$, and between $z$ and $x$ for the ranks 2 to 10.

## Discussion of the Results

It can be seen that the group-correlation coefficients when the number of group ranks chosen is 6, 7, 8. or 9 are not much different from the product moment correlation coefficients. But the two coefficients do not match when the number of ranks is reduced below 6 or increased above 10. Thus when the population is about 100, $r$ can be one of 6, 7, 8 or 9. If the population is much less than 100, $r$ can be one of 3, 4, 5 or 6. Similarly if the population is about 200, $r$ can be chosen to be 8, 9 or 10. But it can also be seen from the Table 4, that even when the number of group ranks is reduced to 2, the group correlation coefficients are not far from the product moment coefficients. This result can be used to estimate the Pearson

TABLE 1

Variation of $k_g^2$ with the number of group ranks chosen.

| $g$ | $k_g^2$ | $g$ | $k_g^2$ |
|---|---|---|---|
| 2 | 0.25 | 7 | 1.38 |
| 3 | 0.32 | 8 | 1.83 |
| 4 | 0 49 | 9 | 2.21 |
| 5 | 1.00 | 10 | 2.82 |
| 6 | 1.07 | | |

TABLE 2

The matriculation ($x$), the B.Sc., ($y$) and the selection ($z$) examinations marks in percentage of 93 students.

| $x$ | $y$ | $z$ | $x$ | $y$ | $z$ |
|---|---|---|---|---|---|
| 64.5 | 61.4 | 26.7 | 56.6 | 69 6 | 53.3 |
| 52.5 | 62.6 | 41.3 | 55.0 | 64.6 | 49.3 |
| 55.4 | 65.2 | 36.0 | 57.6 | 70.4 | 37.3 |
| 61.0 | 60.0 | 48.0 | 41.2 | 68.0 | 41.3 |
| 76.7 | 63 0 | 68 0 | 70.0 | 69 7 | 52.0 |
| 66.4 | 61.9 | 29.3 | 54.0 | 61.4 | 38.7 |
| 67.3 | 60.8 | 42.7 | 73.0 | 63.2 | 38.7 |
| 66.2 | 71.4 | 61.3 | 73.0 | 71 4 | 60.0 |
| 59.1 | 62.6 | 42.7 | 74.0 | 68.6 | 53.3 |
| 64.7 | 73.3 | 54.7 | 69.0 | 65.9 | 61.3 |
| 57.0 | 64.9 | 38.7 | 51.0 | 66 0 | 48.0 |
| 51.1 | 64.6 | 33.3 | 66.0 | 76.2 | 54.7 |
| 73.0 | 66.2 | 44.0 | 62.0 | 60 0 | 41.3 |
| 72.2 | 68.0 | 53.3 | 50.0 | 64 5 | 42.7 |
| 56.0 | 64.1 | 44.0 | 69.0 | 60 0 | 40.0 |
| 58.0 | 68.7 | 48.0 | 67.0 | 63 4 | 53.3 |
| 68.1 | 61.0 | 52.0 | 62.0 | 62 8 | 60.0 |
| 63.0 | 74.8 | 61.3 | 62.0 | 82.6 | 46.7 |
| 63.4 | 70.9 | 56.0 | 59.0 | 82.0 | 30.7 |
| 57.0 | 64.3 | 46.7 | 60.5 | 74.0 | 41.3 |
| 67.5 | 70.3 | 56.0 | 76.0 | 84.0 | 69 3 |
| 62.3 | 70.2 | 54.7 | 78.0 | 62 6 | 78.7 |
| 64.2 | 70.0 | 45.3 | 67.6 | 70.0 | 37.3 |
| 62.0 | 63.6 | 41.3 | 64.0 | 69.7 | 49.3 |
| 57.8 | 61.5 | 36.0 | 73.0 | 78.9 | 50.7 |
| 59.8 | 67.4 | 54.7 | 69.6 | 70.0 | 65.3 |

TABLE 2—(*contd*)

| *x* | *y* | *z* | *x* | *y* | *z* |
|---|---|---|---|---|---|
| 45.0 | 70 0 | 42.7 | 60.0 | 70.0 | 36.0 |
| 58.0 | 60 8 | 33.3 | 61.0 | 80.0 | 37.3 |
| 62.0 | 61.9 | 45.3 | 77.0 | 80.0 | 60.0 |
| 64 8 | 75.0 | 56.0 | 59.0 | 62.5 | 42.7 |
| 74.4 | 61.6 | 53.3 | 64.0 | 70.0 | 38.7 |
| 63.0 | 67 0 | 49.3 | 75.0 | 70.0 | 54.7 |
| 58 0 | 70 0 | 37.3 | 65 0 | 62.5 | 38.7 |
| 47.0 | 70 0 | 33.3 | 59.0 | 62.5 | 44.0 |
| 58.0 | 62 5 | 41.3 | 62.0 | 62.5 | 50.7 |
| 65.0 | 62.5 | 40.0 | 55.5 | 62.5 | 42.7 |
| 63 0 | 62 5 | 38.7 | 75.0 | 62 5 | 48.0 |
| 57.0 | 70 0 | 50.7 | 72.0 | 70.0 | 36 0 |
| 66 0 | 62 5 | 29.3 | 45.0 | 74.7 | 29.3 |
| 70 0 | 65 9 | 53.3 | 52.0 | 66.0 | 49.3 |
| 63 2 | 66 6 | 46.7 | 46.3 | 66.6 | 44.0 |
| 55.0 | 62 5 | 56.0 | 62.6 | 65 2 | 41.3 |
| 67.8 | 60 0 | 40.0 | 67.2 | 64.9 | 44.0 |
| 69.4 | 64.3 | 60.0 | 73.0 | 62.0 | 41.0 |
| 75 0 | 68 0 | 61.3 | 64,6 | 68.2 | 36.0 |
| 68 5 | 64 8 | 69.3 | 65.0 | 73.0 | 54.7 |
| 58.0 | 68 0 | 40.0 | | | |

TABLE 3

Allotting Ranks when $m$ and $\sigma$ known.

| Range | Below $m-\frac{3\sigma}{2}$ | $m-\frac{3\sigma}{2}$ to $m-\frac{\sigma}{2}$ | $m-\frac{\sigma}{2}$ to $m+\frac{\sigma}{2}$ | $m+\frac{\sigma}{2}$ to $m+\frac{3\sigma}{2}$ | Above $m+\frac{3\sigma}{2}$ |
|---|---|---|---|---|---|
| Rank | 5 | 4 | 3 | 2 | 1 |

TABLE 4

Group-correlation coefficients between each pair of the variables $x$, $y$ and $z$ for different group ranks.

| Number of ranks | $r_{g\,xy}$ | $r_{g\,yz}$ | $r_{g\,zx}$ |
|---|---|---|---|
| 2 | 0.16 | 0.33 | 0.40 |
| 3 | 0.29 | 0.29 | 0.40 |
| 4 | 0.03 | 0.21 | 0.40 |
| 5 | 0.19 | 0.33 | 0.50 |
| 6 | 0.01 | 0.22 | 0.53 |
| 7 | −0.04 | 0.21 | 0.51 |
| 8 | 0.05 | 0.23 | 0.58 |
| 9 | 0.05 | 0.23 | 0 60 |
| 10 | −0.06 | 0.15 | 0.43 |

coefficient in a simple and easy manner. All that is needed is the range of scores of a population ; this range can be divided into two or three equal subranges and ranks allotted to the original scores depending on which subrange they fall into. Thus time and/or effort spent on obtaining correlation coefficients by the use of computers or by hand calculation can be reduced.

## CONCLUSIONS

A new correlation coefficient is proposed and compared with Pearson and Spearman coefficients. It has been shown that when the population is about 100, the group coefficient yields comparable value of correlation when the number of ranks chosen is between 6 and 9. The assumption that the populations are normally distributed is, in general, valid when correlation between two variables for a large population is required. It is also shown that even if the number of ranks chosen is as low as 2, the group coefficients give good qualitative estimates of the Pearson coefficients. The correlation

coefficient itself is only a rough qualitative estimation and not much significance can be attached to its exact value. Thus the group rank coefficient gives all the required information about a population described by two variables without the considerable calculations involved in the determination of Spearman or Pearson coefficients.

## ACKNOWLEDGEMENTS

The author thanks his students V. J. S. Mohan, B. N. Sundar and H. N. Srinidhi for useful calculations. He is grateful to Dr. P. S. Moharir for many useful suggestions. He also thanks Prof. B. S. Ramakrishna for his constant encouragement.

## APPENDIX

It is assumed that three standard deviations on either side of the mean cover almost the entire area of the normal distribution graph. If then the graph is divided equally into five regions which are denoted by the group ranks 1, 2, 3, 4 and 5 starting from the right side of the graph, the normalized areas[1] they occupy are 0.0678, 0.2417, 0.3830, 0.2417 and 0.0678 respectively.

Therefore the mean value of the ranks is given by

$$\frac{1}{n}\sum_i x_i = \frac{1}{n}\,[0.0678n + 2(0.2417n) + 3(0.3830n) + 4(0.2417n) + 5(0.0678n)]$$

$$\simeq 3$$

Similarly

$$\sum_i x_i^2 = [(0.0678n + 4(0.2417n) + 9(0.3830n) + 16(0.2417n) + 25(0.678n)]$$

$$\simeq 10n$$

$$\text{Therefore } \sigma x^{(1)} = \sum_i \frac{x_1^2}{n} - \left(\sum_i \frac{x_i}{n}\right)^2 \simeq 1$$

Exactly similar results are obtained for the variable $y$ also.

## REFERENCE

1. Kenny, K. F. and Keeping, E. S. .. Mathematics of Statistics, Part One, Van Nostrand Company Inc., 1954.