Research Journal of Mathematics and Statistics 4(4): 94-97, 2012
ISSN: 2040-7505

Submitted: September 28, 2012 Accepted: November 23, 2012 Published: December 25, 2012

An Efficient Alternative to the Spearman's Rank Correlation Coefficient for Tied Observations

[1]Aliyu Usman and [2]Mohammed Samaila
[1]Department of Mathematics, Statistics and Computer Science, Kaduna Polytechnic, Nigeria
[2]Department of Mathematics and Statistics, Nuhu Bamalli Polytechnic, Nigeria

Abstract: This study has introduced an efficient rank correlation formula similar to the method earlier derived by Spearman. The formula is based on the sum of the ranks rather than their difference as in Spearman's formula. The formula was derived and tested with life data. Hence, if there are no ties in the data, the result shows that the formula gives the same result with the Spearman's formula. When there are ties, it gives better result by reducing the influence of ties. The formula is easy to use devoid of any negative sign in the process of computation. The result obtained shows that the formula is similar to the Spearman's formula when there are no ties but better when there are ties. Also the formula can be related directly to the Kendall's Coefficient of Concordance. It is recommended that this formula be used when there are ties.

Keywords: Concordance, cross-products, nonparametric correlation, ranks, sum of squares, tied observations

INTRODUCTION

Correlation is a general measure that describes whether variables are associated. It is used in a variety of context to indicate the degree and direction of linear relationship between two or more quantitative variables (Younger, 1995). If statistical inference is to be used, the variables must be random variables with a specified probability model (Timm, 2002). For *n* pairs of observations (x_i, y_i), the Karl-Pearson product-moment correlation is as follows:

$$\rho = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 (Y_i - \bar{Y})^2}} = \frac{Cov(X_i, Y_i)}{\sigma_X \sigma_Y}$$

Measures of correlation range from *+1*, indicating perfect positive agreement to *-1*, indicating perfect negative agreement. If the underlying distribution of (x_i, y_i) is bivariate normal, with correlation coefficient ρ, its unbiased estimate in large samples is given by:

$$\hat{\rho} = 2Sin(\tfrac{1}{6}\pi\rho_S) = 2Sin(30\rho_S)^0$$

Kruskal and Tanur (1978) highlighted other rank correlation coefficients include, Kendall's τ and the Fisher-Yates correlation coefficient obtained by replacing (x_i, y_i) by their normal scores. Spearman (1904) introduced hitherto the most popular rank correlation method using the ranks rather than the actual observations. The result is Spearman's coefficient of rank correlation, more readily computed from the formula below:

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}$$

where, $D_i = r_{xi} - r_{yi}$, is the difference in rank between the pairs of observations (x_i, y_i). Norman (2009) highlighted that if there are many tied observations where average ranks are usually assigned to such tied observations, it is best to make corrections and to calculate the rank correlation using the formula below:

$$r_S = \frac{S_x + S_y - \sum_{i=1}^{n} d_i^2}{2\sqrt{S_x \cdot S_y}}$$

where,

$$S_x = \frac{n(n^2 - 1) - \sum_{i=1}^{g}(t_i^3 - t_i)}{12}$$

With *g* the number of groups with average ranks and t_i the size of group *i* for the *X* sample and:

$$S_y = \frac{n(n^2 - 1) - \sum_{j=1}^{h}(t_j^3 - t_j)}{12}$$

Corresponding Author: Aliyu Usman, Department of Mathematics, Statistics and Computer Science, Kaduna Polytechnic, Nigeria

With *h* the number of groups with average ranks and t_j the size of group *j* for the *Y* sample. If there are no ties, the usual rank correlation formula is used. However, it is mathematically the same to use the formula above untied samples. In that case, the observations are seen as groups of size *1*. Meaning that *g = h = n and* $t_j = t_j = 1$ *for i,j = 1,2,…,n* and:

$$S_x = S_y = \frac{n(n^2-1)}{12}$$

The Spearman's rank correlation coefficient is the most popular nonparametric correlation method but affected by multiple ties. The issue that arises for this study is the need to have a similar but more efficient rank correlation coefficient especially for data with tied observations. Hence, the need for a formula which could give closer value to the product-moment correlation coefficient especially when there are ties. The objective is to emphasize a rank correlation coefficient that neutralizes the influence of ties in the data and gives closer approximation to the product-moment correlation coefficient. The rank correlation coefficient emphasized in this study will particularly neutralize the influence of ties in the data as well as give better results for data with tied observations.

METHODOLOGY

Recall the Karl-Pearson's product-moment correlation coefficient given by the formula:

$$\rho = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2(Y_i-\overline{Y})^2}} = \frac{Cov(X_i,Y_i)}{\sigma_X\sigma_Y}$$

Using the computational form of the product-moment correlation coefficient below, we can derive a rank correlation that is efficient even for data with tied observations:

$$\rho = \frac{n\sum_{i=1}^{n}X_iY_i - \sum_{i=1}^{n}X_i\sum_{i=1}^{n}Y_i}{\sqrt{\left[n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right]\left[n\sum_{i=1}^{n}Y_i^2-\left(\sum_{i=1}^{n}Y_i\right)^2\right]}} \quad (1)$$

Hence we shall particularly derive; from the product-moment correlation coefficient formula, a rank correlation similar to the Spearman's method. Whereas Spearman used the difference of the ranks; in this study we shall use the sum of the ranks for another approach to rank correlation formula. Thus the Spearman's rank correlation has the formula:

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n}D_i^2}{n(n^2-1)} \quad (2)$$

In this case $D_i = r_{xi} - r_{yi}$ is the difference of the ranks. Instead of using differences between ranks of pairs of *X* and *Y*, We may use the sums of the ranks for each pair, where $S_i = r_{xi} + r_{yi}$ (Meddis, 1984). To derive the formula, in Eq. (2) the sum of squares and the cross-products are replaced by their respective ranks (assuming no ties). Thomas (1989) suggested that using the sums of the ranks for each pair is an equivalent of using differences between ranks of pairs of *X* and *Y*. Hence, the method for the sums of the ranks goes as follows:

$$\sum_{i=1}^{n}X_i = \sum_{i=1}^{n}r_{X_i} = 1+2+\ldots+n = \frac{n(n+1)}{2} \quad (3)$$

Similarly,

$$\sum_{i=1}^{n}Y_i = \sum_{i=1}^{n}r_{Y_i} = 1+2+\ldots+n = \frac{n(n+1)}{2} \quad (4)$$

And

$$\sum_{i=1}^{n}X_i^2 = \sum_{i=1}^{n}r_{X_i}^2 = 1^2+2^2+\ldots+n^2 = \frac{n(n+1)(2n+1)}{6} \quad (5)$$

Similarly,

$$\sum_{i=1}^{n}Y_i^2 = \sum_{i=1}^{n}r_{Y_i}^2 = 1^2+2^2+\ldots+n^2 = \frac{n(n+1)(2n+1)}{6} \quad (6)$$

From above, it is established that Eq. (1) = (2) and (3) = (4) and *n* = number of paired observations:

$$\sum_{i=1}^{n}S_i = \sum_{i=1}^{n}(r_{X_i}+r_{Y_i})^2 = \sum_{i=1}^{n}r_{X_i}^2 + \sum_{i=1}^{n}r_{Y_i}^2 + 2\sum_{i=1}^{n}r_{X_i}r_{Y_i} \quad (7)$$

From Eq. (5), the following equations are hereby obtained:

$$\sum_{i=1}^{n}r_{X_i}r_{Y_i} = \frac{1}{2}\left[\sum_{i=1}^{n}\left(r_{X_i}+r_{Y_i}\right)^2 - \sum_{i-1}^{n}r_{X_i}^2 - \sum_{i=1}^{n}r_{Y_i}^2\right]$$

$$\sum_{i=1}^{n}S_i^2 - \frac{n}{3}(n+1)(2n+1)\Bigg]$$

$$\therefore \sum_{i=1}^{n}X_iY_i = \sum_{i=1}^{n}r_{X_i}r_{Y_i} = \frac{1}{2}\left[\sum_{i=1}^{n}S_i^2 - \frac{n}{3}(n+1)(2n+1)\right] \quad (8)$$

Substituting Eq. (3), (4), (5), (6) and (8) into Eq. (1) we have the following results:

$$\rho_n = \frac{\frac{n}{2}\left[\sum_{i=1}^{n}S_i^2 - \frac{n}{3}(n+1)(2n+1)\right] - \frac{n^2}{4}(n+1)^2}{\sqrt{\left[\frac{n}{6}(n+1)(2n+1) - \frac{n^2}{4}(n+1)^2\right]^2}} \quad (9)$$

$$= \frac{\frac{n}{6}\left[3\sum_{i=1}^{n} S_i^2 - n(n+1)(2n+1)\right] - \frac{n^2}{4}(n+1)^2}{\frac{n^2}{6}(n+1)(2n+1) - \frac{n^2}{4}(n+1)^2} \quad (10)$$

Simplifying Eq. (10) will give the following equation:

$$\rho_n = \frac{2n\left[3\sum_{i=1}^{n} S_i^2 - n(n+1)(2n+1)\right] - 3n^2(n+1)^2}{2n^2(n+1)(2n+1) - 3n^2(n+1)^2} \quad (11)$$

$$= \frac{2\left[3\sum_{i=1}^{n} S_i^2 - n(n+1)(2n+1)\right] - 3n(n+1)^2}{2n(n+1)(2n+1) - 3n(n+1)^2} \quad (12)$$

$$= \frac{6\sum_{i=1}^{n} S_i^2 - 2n(n+1)(2n+1) - 3n(n+1)^2}{n(n+1)[2(2n+1) - 3(n+1)]} \quad (13)$$

$$= \frac{6\sum_{i=1}^{n} S_i^2 - n(n+1)[2(2n+1) + 3(n+1)]}{n(n+1)(n-1)} \quad (14)$$

$$= \frac{6\sum_{i=1}^{n} S_i^2 - n(n+1)(7n+5)}{n(n+1)(n-1)} \quad (15)$$

Finally, separating the fraction in Eq. (15) gives us the desired rank correlation formula in Eq. (16) as follows:

$$\rho_n = \frac{6\sum_{i=1}^{n} S_i^2}{n(n^2-1)} - \frac{7n+5}{n-1} \quad (16)$$

In this case $S_i = r_{xi} + r_{yi}$ is the difference of the ranks. The formula essentially gives the same result as the Spearman's formula when there are no ties in the data and better result for data with tied observations.

RESULTS

The three correlation formulae were used on 800 simulated sets of bivariate data all with tied observations. In all the 800 data sets, this formula gives values closer to the product-moment correlation coefficient. Table 1 shows the percentage scores, in Mathematics (*X*) and Statistics (*Y*), obtained from a random sample of 20 fresh Statistics students in 2006. When there are no ties, we shall demonstrate the equality of both formulae. Using the Spearman's formula the following results were obtained:

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} = 1 - \frac{6(386)}{20(20^2-1)} = 0.7098$$

Using the method derived by this study the following results were obtained:

$$\rho_n = \frac{6\sum_{i=1}^{n} S_i^2}{n(n^2-1)} - \frac{7n+5}{n-1} = \frac{6(11094)}{20(20^2-1)} - \frac{7(20)+5}{20-1} = 0.7098$$

Hence, when there are no ties, both formulae will always produce the same results. Similarly Table 2 shows a similar data set from another set of a random sample of 20 fresh Statistics students. When there are ties, we shall demonstrate the superiority of the method introduced in this study as follows:

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)} = 1 - \frac{6(989)}{20(20^2-1)} = 0.2564$$

Using the method derived by this study the following result is obtained:

Table 1: Rank correlation without ties

S No	*X*	*Y*	r_{xi}	r_{yi}	D_i	D_i^2	S_i	S_i^2
1	49	64	6	8	-2	4	14	196
2	52	66	8	10	-2	4	18	324
3	40	44	1	2	-1	1	3	9
4	58	53	10	6	4	16	16	256
5	69	70	16	14	2	4	30	900
6	68	76	15	19	-4	16	34	1156
7	46	74	4	17	-13	169	21	441
8	67	67	14	11	3	9	25	625
9	51	51	7	4	3	9	11	121
10	75	77	19	20	-1	1	39	1521
11	90	75	20	18	2	4	38	1444
12	60	68	12	12	0	0	24	576
13	59	52	11	5	6	36	16	256
14	41	50	2	3	-1	1	5	25
15	53	42	9	1	8	64	10	100
16	63	69	13	13	0	0	26	676
17	48	54	5	7	-2	4	12	144
18	45	65	3	9	-6	36	12	144
19	72	71	17	15	2	4	32	1024
20	73	73	18	16	2	4	34	1156
Total						386		11094

Table 2: Rank correlation with ties

S No	*X*	*Y*	r_{xi}	r_{yi}	D_i	D_i^2	S_i	S_i^2
1	68	90	18	20	-2	4	38	1444
2	58	60	10.5	13	-2.5	6.25	23.5	552.25
3	44	59	1	12	-11	121	13	169
4	51	40	3.5	1	2.5	6.25	4.5	20.25
5	66	53	16.5	7.5	9	81	24	576
6	46	63	2	14	-12	144	16	256
7	52	48	5	5	0	0	10	100
8	60	45	12	3	9	81	15	225
9	65	72	15	16.5	-1.5	2.25	31.5	992.25
10	72	73	20	18	2	4	38	1444
11	64	41	14	2	12	144	16	256
12	57	69	8.5	15	-6.5	42.25	23.5	552.25
13	66	54	16.5	9.5	7	49	26	676
14	57	84	8.5	19	-10.5	110.25	27.5	756.25
15	61	58	13	11	2	4	24	576
16	58	53	10.5	7.5	3	9	18	324
17	56	72	7	16.5	-9.5	90.25	23.5	552.25
18	51	46	3.5	4	-0.5	0.250	7.50	56.250
19	70	54	19	9.5	9.5	90.25	28.5	812.25
20	55	49	6	6	0	0	12	144
Total						989		10484

$$\rho_s = \frac{6\sum_{i=1}^{n} S_i^2}{n(n^2-1)} - \frac{7n+5}{n-1} = \frac{6(10484)}{20(20^2-1)} - \frac{7(20)+5}{20-1} = 0.2511$$

Using the parametric formula, the Pearson's product-moment method, the following results were obtained:

$$\rho = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}}$$

$$= \frac{20(70138) - 1177(1183)}{\sqrt{\left[20(70407) - (1177)^2\right]\left[20(73605) - (1183)^2\right]}} = 0.2548$$

Similar results patterns were obtained for all the 800 data sets. That is, for data with tied observations, this formula gives closer values to the product-moment correlation coefficient.

CONCLUSION

From the results, it was established that the modified rank correlation coefficient by this study gives the same result with the Spearman's rank correlation coefficient when there are no ties. In case of tied data, it gives a closer approximation to the parametric Karl-Pearson's product-moment correlation coefficient. Furthermore, the introduced method could be related directly to the coefficient of concordance suggested by Kendall (1970), for more than two groups of ranks denoted by *W*. Hence, the following relation:

$$W = \tfrac{1}{2}(\rho_S + 1)$$

This formula is yet another improvement in nonparametric correlation analysis. It has both neutralizes the influence of ties as well as provides a closer approximation to the parametric Karl-Pearson's product-moment correlation coefficient. In addition, the rank correlation coefficient formula has a direct relationship with the Kendall's coefficient of concordance. It is therefore recommended that researchers and users alike should adopt this formula to use in various areas of nonparametric correlation analysis. This formula could be used as an alternative nonparametric correlation method whenever there are ties.

REFERENCES

Kendall, M.G., 1970. Rank Correlation Method. 4th Edn., London, Griffin.

Kruskal, W.H. and J.M. Tanur, 1978. International Encyclopedia of Statistics. Collier Macmillan, London.

Meddis, R., 1984. Statistics Using Ranks: A Unified Approach. Basil Blackwell, Oxford.

Norman, C., 2009. Analyzing Multivariate Data. 4th Edn., Academic Press, New York.

Thomas, G.E., 1989. A note on correcting for ties with Spearman's rho. J. Stat. Comput. Simul., 31: 37-40.

Timm, N.H., 2002. Applied Multivariate Analysis. Springer, New York.

Younger, M.S., 1995. A First Course in Linear Regression. 2nd Edn., Duxbury, Boston.