IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 10, Issue 5 Ver. III (Sep-Oct. 2014), PP 09-17
www.iosrjournals.org

Ties Adjusted Rank Correlation Coefficient

Oyeka I. C. A. and Nwankwo Chike, H.
Department of Statistics Nnamdi Azikiwe University, Awka Nigeria

Abstract: *This paper proposes, develops and presents a method that may be used to adjust and correct sample estimates of the usual spearman rank correlation coefficient for the possible presence of tied observation in the sampled populations. Without these adjustments the sums of squares of ranks in the denominator of the usual expression for the estimation of spearman rank correlation coefficients are often over estimated if there are ties in the populations and no effort is made to correct and adjust for them especially if these ties are not few, resulting in inefficient, unreliable and often misleading estimates. The proposed method is shows to be easily modified and applied to estimate ties adjusted spearman partial rank correlation coefficient between any two populations holding observations from a third population at constant levels with all the populations adjusted for the possible presence of tied observations. The usual statistical tests for significance are also easily modified using the ties adjusted rank correlation coefficients in place of the unadjusted and usual rank correlation coefficients. The proposed method is illustrated with some sample data and shows to produce more efficient estimates than the usual ties unadjusted uncorrected spearman rank correlation coefficients.*

Keywords: *Rank, Ties, Adjusted, Correlation, Partial, Efficient, Ordinal.*

I. Introduction

In correlation analysis to determine the strength of association between two variables that do not satisfy the usual assumption of normality for a valid and proper use of parametric methods, the use of non parametric techniques is indicated and preferable. A frequently used measure of association in these cases is the well known spearman simple rank correlation coefficient which uses only the ranks rather than the raw scores themselves (Gibbons, 1971; Hollander and Wolfe, 1973; Siegel, 1956; Maritz, 1981).

However a necessary assumption for the use of spearman rank correlation coefficient is that the populations of interest be continuous. This would in effect mean that the probability that any two observations from each of the sampled populations have exactly the same value is at least theoretically zero.

But in practice ties do occur and their effects especially if these ties are not few, may seriously affect the results obtained in the analysis, sometimes resulting in misleading conclusions.

Now if these are no tied observations in any of the populations of interest being correlated, then the usual expression for the estimations of spearman rank correlation coefficient may be applied using the ranks assigned separately to the sampled observations from each of the populations.

However when some of the observations from the sampled population are tied and hence assigned their mean ranks, then even though the ranks assigned to the observations from each population have a unique sum whether or not there are tied ranks, the sum of squares of these ranks depends on whether the observations are tied and hence assigned mean ranks or whether each observation in the sample has a unique rank assigned to it.

The sum of squares of assigned ranks in the presence of ties, especially if these ties are not few, is often smaller in value than the corresponding sum of squares of assigned ranks in the absence of ties between observations in the population.

Thus specifically the value of the sums of squares in the denominator of the regular or usual expression for the estimation of spearman simple rank correlation coefficient or not there are tied observations in these populations. If some of these observations are tied in value and hence assigned their mean ranks, then without proper adjustments to reflect these ties in ranks, then use of the regular expression for the estimation of spearman simple rank correlation coefficient would yield rather larger values of the sum of squares in the denominator of the expression and may hence result in the eventual underestimation of the required rank correlation coefficient. Hence these sums of squares of rank need to be modified to adjust or correct for the possible effect of tied observations in the calculation of the correlation coefficient.

We in this paper propose, develop and present a method that would enable one modify and adjust the usual expression for the estimation of spearman rank correlation coefficient that factors in and reflects the possible presence of tied observations in the sampled populations in the calculation of sums of squares of assigned ranks and hence enable one obtain more efficient estimate of the rank correlation coefficient.

II. The Proposed Method

Let x_i be the ith observation in a random sample of size 'n' drawn from population X, and y_i be the ith observation in a random sample also of size n drawn from population Y, for i = 1, 2,…,n. Populations X and Y may be either related or independent populations. If X and Y are normally distributed and continuous, then the strength of association between the two populations may be measured using any of the usual parametric methods such as familiar Pearsons coefficient of correlation ρ_p, estimated as

$$\hat{\rho}_p = r_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} = \frac{\sum_i x_i y_i - \frac{\sum x_i . \sum y_i}{n}}{\sqrt{\left(\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^{\,2} - \frac{\left(\sum y_i\right)^2}{n}\right)}} \qquad (1)$$

However, if populations X and Y are not normally distributed, then parametric methods can not validly and properly be used to analyze these data and determine any measure of association between the two populations.

In these cases use of non-parametric methods is now indicated and preferable. A non-parametric method that readily suggests its self for this purpose because of its relatively easy estimation is the Spearman simple rank correlation coefficient ρ_s.

To estimate Spearman simple rank correlation coefficient between two populations X and Y which may now be measurements on as low as the ordinal scale and need not be continuous or even numbers, the observations drawn from each of these populations are each converted into ranks and then used in subsequent analyses.

Thus suppose r_{ix} is the rank assigned to x_i, the ith observation drawn from population X and r_{iy} is the rank assigned to y_i, the ith observation drawn from population Y for i = 1,2,…,n.

Then the Spearman simple rank correlation coefficient ρ_s between populations X and Y is estimated as

$$\hat{\rho}_p = r_{xy} = \frac{\sum_{i=1}^{n}\left(r_{ix} - \bar{r}\right)\left(r_{iy} - \bar{r}\right)}{\sqrt{\sum_{i=1}^{n}\left(r_{ix} - \bar{r}\right)^2\left(r_{iy} - \bar{r}\right)^2}} = \frac{\sum r_{ix} r_{iy} - \frac{n(n+1)^2}{4}}{\sqrt{\left(\sum_{i=1}^{n} r_{ix}^2 - \frac{n(n+1)^2}{4}\right)\left(\sum r_{iy}^{\,2} - \frac{n(n+1)^2}{4}\right)}} \qquad - - - \ (2)$$

(Siegel, 1956; Kendall, 1971; Hollander & Wolfe, 1973)

Since $\bar{r} = \frac{n+1}{2}$

Now the values of the sums of squares and cross products of the ranks and hence the value of $\rho_s = r_{xy}$ of Equation 2 strictly speaking depends on whether or not there are ties in the sampled populations X and Y and are often underestimated if ties exist.

However under the usual assumption of continuity often required for populations X and Y for a valid and proper use of Spearman simple correlation coefficient, then the probability that any two observations from population X having exactly equal values and also the probability that any two observations from population Y having exactly equal values are theoretically zero (Maritz, 1981). Hence under this assumption, since the ranks assigned to the sample observations from the populations are values of the first set of positive integers, we would have that the sum of squares of the ranks are

$$\sum_{i=1}^{n} r_{ix}^2 = \sum_{i=1}^{n} r_{iy}^2 = \frac{n(n+1)(2n+1)}{6} \qquad - \quad - \quad - \quad (3)$$

Hence using Equation 3 in Equation 2 we obtain the sample estimate of Spearman simple rank correlation coefficient between populations X and Y as

$$\hat{\rho}_s = r_{xy} = \frac{\sum_{i=1}^{n}\left(r_{ix} - \frac{n+1}{2}\right)\left(r_{iy} - \frac{n+1}{2}\right)}{\frac{n(n^2-1)}{12}} = \frac{\sum_{i=1}^{n} r_{ix} . r_{iy} - \frac{n(n+1)^2}{4}}{\frac{n(n^2-1)}{12}} \qquad - \quad - \quad - \quad (4)$$

OR when expressed in its more familiar form we have

$$\hat{\rho}_s = r_{xy} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2-1\right)} \quad - \quad - \quad - \quad (5)$$

Where

$$d_i = r_{ix} - r_{iy} \quad - \quad - \quad - \quad (6)$$

is the difference between the rank r_{ix} assigned to the ith observation from population X, and the rank r_{iy} assigned to the corresponding ith observation from population Y for i = 1, 2, …,n.

As noted above if there are no ties in the sampled populations X and Y, then Equations 4 or Equation 5 may equivalently be used to estimate Spearman simple rank correlation coefficient between populations X and Y. However in reality, ties do occur and the presence of tied observations, especially if they are not few, if not corrected and adjusted for in the analysis, may seriously affect the results of such analysis and may often invalidate or put in doubt any conclusions reached from the analysis.

Hence in the presence of tied observations in the sampled populations, to obtain more reliable results the simple rank correlation coefficient between populations X and Y is preferably estimated using Eqn. 2.

Now the sums of squares of the ranks r_{ix} and r_{iy} in the denominator of Eqn. 2 depend, as noted above, on whether or not there are tied observations in populations X and Y and are often underestimated if in fact there are tied observations in the populations.

Hence to adjust for this possibility we may let r_{ix} be the rank that would have been expected to be assigned to x_i, the ith observation from population X if x_i were not tied with any other observation from the population, and let r'_{ix} be the mean rank actually assigned to x_i if this observation were tied with some other observations from population X. Similarly let r_{iy} and r'_{iy} be respectively the corresponding rank or mean rank for y_i, the ith observations from population Y, for i = 1, 2, …,n.

Then in the presence of ties in the sampled populations X and Y, the Spearman simple rank correlation coefficient between the two populations is estimated as

$$\hat{\rho}_s = r_{xy} = \frac{Cov\left(r'_{ix}, r'_{iy}\right)}{\sqrt{Var\left(r'_{ix}\right)Var\left(r'_{iy}\right)}} = \frac{\sum_{i=1}^{n} r'_{ix} r'_{iy} - \frac{n(n+1)^2}{4}}{\sqrt{\left(\sum_{i=1}^{n}\left(r'_{ix}\right)^2 - \frac{n(n+1)^2}{4}\right)\left(\sum_{i=1}^{n}\left(r'_{iy}\right)^2 - \frac{n(n+1)^2}{4}\right)}} \quad - \; - \; - \quad (7)$$

Now to calculate the sums of squares in the denominator of Eqn. 7 we may let

$$d_{ix} = r_{ix} = r'_{ix}; \quad d_{iy} = r_{iy} - r'_{iy} \quad - \quad - \quad - \quad (8)$$

where d_{ix} is the difference between r_{ix}, the rank that would have been expected to be assigned to x_i when not tied and r'_{ix} the rank that would have been actually assigned to x_i when x_i is tied with other observations from population X. The difference d_{iy} is similarly defined for y_i, the ith observation from population Y.

Hence

$$r'_{ix} = r_{ix} - d_{ix}; \quad r'_{iy} = r_{iy} - d_{iy} \quad - \quad - \quad - \quad (9)$$

Using the results of Eqn. 9 in the denominator of Eqn 7. We have that the sum of the ranks r_{ix} of observations from population X which reflects the presence of tied observations in X is

$$\sum_{i=1}^{n}\left(r'_{ix}\right)^2 = \sum_{i=1}^{n}\left(r_{ix} - d_{ix}\right)^2 = r_{ix}^2 - 2\sum_{i=1}^{n}\left(r_{ix}d_{ix} - \frac{d_{ix}^2}{2}\right) \qquad \text{That is}$$

$$\sum_{i=1}^{n}\left(r'_{ix}\right)^2 = \frac{n(n+1)(2n+1)}{6} - 2\sum_{i=1}^{n}\left(r_{ix}d_{ix} - \frac{d_{ix}^2}{2}\right)^2 = \frac{n(n+1)(2n+1)}{6} - 2\sum_{i=1}^{n} z_{ix} \quad - \; - \; - \quad (10)$$

Similarly, the sum of squares of the ranks r'_{iy} of observations from population Y in the presence of ties is

$$\sum_{i=1}^{n}\left(r'_{iy}\right)^2 = \sum_{i=1}^{n}\left(r_{iy} - d_{iy}\right)^2 = \sum_{i=1}^{n} r_{iy}^{\,2} - 2\sum_{i=1}^{n}\left(r_{iy}d_{iy} - \frac{d_{iy}^2}{2}\right)$$

That is

$$\sum_{i=1}^{n}\left(r'_{iy}\right)^2=\frac{n(n+1)(2n+1)}{6}-2\sum_{i=1}^{n}\left(\hat{r}_{iy}d_{iy}-\frac{d_{iy}^{\,2}}{2}\right)=\frac{n(n+1)(2n+1)}{6}-2\sum_{i=1}^{n}z_{iy} \quad - \quad - \quad - \ (11)$$

Where

$$z_{ix}=r_{ix}d_{ix}-\frac{d_{ix}^{\,2}}{2}, z_{iy}=\left(r_{iy}.d_{iy}-\frac{d_{iy}^{\,2}}{2}\right) \qquad - \qquad - \qquad - \qquad (12)$$

Now $z_{ix}=r_{ix}d_{ix}-\frac{d_{ix}^{\,2}}{2}$ of Eqn 10 assumes non-zero values only if d_{ix} of Eqn 8 is non-zero for some i = 1, 2,…,n. In other words if the ith observation from population X is tied in value with some other observations from population X, and assumes the value zero in the absence of ties, that is if x_i is not a tied observation from population X.

The value $z_{iy}=r_{iy}d_{iy}-\frac{d_{iy}^{\,2}}{2}$ for y_i, the ith observation from population Y, is similarly interpreted for some i = 1, 2,…,n

Hence to adjust the sum of squares of the ranks of Equation 10 and correct this sum of squares in the denominator of Eqn. 7 for the possible presence of tied observations in population X we may let

$$u_{ix}=\begin{cases}1, if\ z_{ix}\neq 0\\ 0, if\ \ z_{ix}=0\end{cases}$$

In other words u_{ix} assumes the values 1 if and only if the i^{th} observation from population X is tied in value with some other observation(s) from X, and is hence assigned the mean rank r'_{ix} otherwise u_{ix} assumes the value 0, for some i = 1, 2,…,n.

The indicator u_{iy} is similarly defined for $z_{iy}=r_{iy}d_{iy}-\frac{d_{iy}^{\,2}}{2}$ for y_i the i^{th} observation from populations Y, for some i = 1, 2,…,n.

Now let

$$\prod_x=p(U_{ix}=1) \qquad - \qquad - \qquad - \qquad (14)$$

And

$$W_x=\sum_{i=1}^{n}u_{ix}z_{ix}=\sum_{i=1}^{n}u_{ix}\left(r_{ix}d_{ix}-\frac{d_{ix}^{\,2}}{2}\right) \qquad - \qquad - \qquad - \qquad (15)$$

Now

$$E(u_{ix})=\prod_x \ and\ Var(u_{ix})=\prod_x.(1-\prod_x) \qquad - \qquad - \qquad - \qquad (16)$$

Also from Eqn 15 we have that

$$E(W_x)=\sum_{i=1}^{n}\left(r'_{ix}d_{ix}-\frac{d_{ix}^{\,2}}{2}\right)E(u_{ix})=\prod_x\sum_{i=1}^{n}\left(r_{ix}d_{ix}-\frac{d_{ix}^{\,2}}{2}\right) \quad - \quad - \quad - \quad (17)$$

Now $\prod_x$ is the proportion of observations from population X that are tied in value with some other observations from the population, its sample estimate is

$$\hat{\prod}_x=\frac{f_x}{n} \qquad - \qquad - \qquad - \qquad (18)$$

Where 'f_x' is the number of sample observations from population X that are tied in values with some other observations from the same population and are hence assigned some mean ranks, so that the corresponding differences d_{ix} are non-zero for some i = 1, 2,…,n. In other words 'f_x' is the number of 1's in the frequency distribution of the 'n' values of 1's and 0's in u_{ix}, for i = 1, 2, . . . ,n.

Using Eqn (18) in Eqn (10) we have that the sample estimate of the sum of squares of the ranks of observations from population X adjusted for ties in this population is

$$\sum_{i=1}^{n}r'^{2}_{ix}=\frac{n(n+1)(2n+1)}{6}-2\hat{\prod}_x\sum_{i=1}^{n}\left(r_{ix}d_{ix}-\frac{d_{ix}^{\,2}}{2}\right)$$

$$=\frac{n(n+1)(2n+1)}{6}-2\frac{f_x}{n}\sum_{i=1}^{n}\left(r_{ix}d_{ix}-\frac{d_{ix}^{\,2}}{2}\right) \qquad - \qquad - \qquad - \qquad (19)$$

Similar approach as above yields the sample estimate of the sum of squares of the ranks of observations from population Y in the presence of tied observations in this population using Eqn11 as

$$\sum_{i=1}^{n} r_{iy}'^{2} = \frac{n(n+1)(2n+1)}{6} - 2\hat{\prod}_{y} \sum_{i=1}^{n}\left(r_{iy} d_{iy} - \frac{d_{iy}^{2}}{2} \right)$$

$$= \frac{n(n+1)(2n+1)}{6} - 2\frac{f_y}{n} \sum_{i=1}^{n}\left(r_{iy} d_{iy} - \frac{d_{iy}^{2}}{2} \right) \quad - \quad - \quad - \quad (20)$$

Now using Eqn 19, we have that the estimated variance of population X in terms of assigned ranks when there are tied observations in this population, that is when the variance is adjusted or corrected for possible presence of tied observations is

$$Var(r_{ix}) = \frac{n(n+1)(2n+1)}{6} - 2\hat{\prod}_{x} \sum_{i=1}^{n}\left(r_{ix} d_{ix} - \frac{d_{ix}^{2}}{2} \right) - \frac{n(n+1)^{2}}{4} \quad \text{OR}$$

$$Var(r_{ix}') = \frac{n(n^{2}-1)}{12} - 2\hat{\prod}_{x} \sum_{i=1}^{n}\left(r d_{ix} - \frac{d_{ix}^{2}}{2} \right) \quad - \quad - \quad - \quad (21)$$

Similarly the sample estimate of the variance of population Y in terms of assigned ranks, adjusted or corrected for possible tied observations in population Y is obtained using Eqn (20) as

$$Var(r_{iy}') = \frac{n(n^{2}-1)}{12} - 2\hat{\prod}_{y} \sum_{i=1}^{n}\left(r_{iy} d_{iy} - \frac{d_{iy}^{2}}{2} \right) \quad - \quad - \quad - \quad (22)$$

Now using Eqns 21 and 22 in Eqn 7 we obtain $r_{xy.c}$, the sample estimate of ties adjusted or corrected Spearman simple rank correlation coefficient between populations X and Y adjusted or corrected for tied observations in these populations as

$$r_{xy.c} = \frac{\sum_{i=1}^{n} r_{ix}' r_{iy'}' - \frac{n(n+1)^{2}}{4}}{\sqrt{\left(\frac{n(n^{2}-1)}{12} - 2\hat{\prod}_{x} \sum_{i=1}^{n}\left(r_{ix} d_{ix} - \frac{d_{ix}^{2}}{2} \right) \right)\left(\frac{n(n^{2}-1)}{12} - 2\hat{\prod}_{y} \sum_{i=1}^{n}\left(r_{iy} d_{iy} - \frac{d_{iy}^{2}}{2} \right) \right)}} \quad - \quad - \quad - \quad (23)$$

Equation 23 enables the estimation of Spearman simple rank correlation coefficient between two populations that may be measurements on as low as the ordinal scale which is adjusted or corrected for the possible presence of tied observations in the two populations.

Note that if there are no tied observations in any of the two populations, then $d_{ix} = d_{iy} = 0$ for all i = 1, 2, . . .,n, so that $r_{ix}' = r_{ix}$, $r_{iy}' = r_{iy}; and \prod_{x} = \prod_{y} = 0$. Hence in these situations we obtain the usual or regular sample estimate of Spearman simple rank correlation coefficient as

$$\rho_{s} = r_{xy.c} = r_{xy} = \frac{\sum_{i=1}^{n} r_{ix} r_{iy} - \frac{n(n+1)^{2}}{4}}{\frac{n(n^{2}-1)}{12}} \quad - \quad - \quad - \quad (24)$$

which is the same expression as Eqn 4 or Eqn 5 when there are no tied observations in the sampled population. If there are tied observations in only one population, population Y say, but not in population X say, then we would have from Eqn 23 that $r_{ix} = r_{ix}; r_{iy}' = r_{iy}$ for some i, $d_{ix} = 0$, for all i, $d_{iy} \neq 0$ for some i = 1, 2, . . . ,n; so that $\prod_{x} = 0 \; and \prod_{y} \neq 0$. With these results we again obtain a sample estimate of Spearman simple rank correlation coefficient between populations X and Y when there are tied observations in population Y as

$$r_{xy.}c = \frac{\sum_{i=1}^{n} r_{ix} r_{iy}' - \frac{n(n+2)^{2}}{4}}{\sqrt{\frac{n(n^{2}-1)}{12}\left(\frac{n(n^{2}-1)}{12} - 2\hat{\prod}_{y} \sum_{i=1}^{n}\left(r_{iy}' d_{iy} - \frac{d_{iy}^{2}}{2} \right) \right)}} \quad - \quad - \quad - \quad (25)$$

Note that if there are actually tied observations in the sampled populations which are not adjusted for or corrected before estimating the required Spearman simple rank correlation coefficient then the expression that is usually used in the estimation is, strictly speaking, from Eqns 4 or 5

$$r_{xy} = \hat{\rho}_s = \frac{\sum_{i=1}^{n} r'_{ix}.r'_{iy} - \frac{n(n+1)^2}{4}}{\frac{n(n^2-1)}{12}} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad - \quad - \quad - \quad (26)$$

Now since Eqns 23 and 26 have the same numerator it is easy to show that the ties adjusted or corrected estimate of Spearman simple rank correlation coefficient $r_{xy.c}$ of Eqn 23 is relatively more efficient than its ties unadjusted or uncorrected counterpart, r_{xy} of Equation 26. To show this we note that the efficiency of $r_{xy.c}$ relative to r_{xy} for the same sample size 'n' from Eqns 23 and 26 is

$$RE(r_{xy.c}; r_{xy}) = \frac{Var(r_{xy})}{Var(r_{xy.c})}$$, that is

$$RE(r_{xy.c}; r_{xy}) = \frac{\left(\frac{n(n^2-1)}{12}\right)\left(\frac{n(n^2-1)}{12}\right)}{\left(\frac{n(n^2-1)}{12} - 2\prod_x \sum_{i=1}^{n}\left(r'_{ix}d_{ix} - \frac{d_{ix}^2}{2}\right)\right)\left(\frac{n(n^2-1)}{12} - 2\prod_y \sum_{i=1}^{n}\left(r_{iy}d_{iy} - \frac{d_{iy}^2}{2}\right)\right)} \geq 1 \quad \ldots(27)$$

as required, for $\hat{\prod}_x, \hat{\prod}_y \neq 0$

Ties adjusted or corrected Spearman partial rank correlation coefficient between populations X and Y holding observations from say population Z at constant levels is similarly estimated. This is done by first estimating $r_{xy.c}$, the ties adjusted simple rank correlation coefficient between populations X and Y; $r_{xz.c}$, the ties adjusted simple rank correlation coefficient between populations X and Z; and $r_{yz.c}$, the ties adjusted simple rank correlation coefficient between populations Y and Z, using Eqn 23.

These ties adjusted or corrected simple rank correlation coefficients are now used to replace the corresponding ties unadjusted or uncorrected simple rank correlation coefficient r_{xy}, r_{xz} and r_{yz} respectively in the usual expression for the estimation of Spearman partial rank correlation coefficient (Gibbons, 1973) given as

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} \quad - \quad - \quad - \quad (28)$$

No further problems should arise in the analysis.

Furthermore all test statistics based on the usual, that is uncorrected or ties unadjusted Spearman rank correlation coefficients about strength of association between two populations remain essentially unchanged except that all ties unadjusted or uncorrected rank correlation coefficients in the expressions for the test statistics are now replaced with their ties adjusted or corrected counterparts.

III. Illustrative Example

The estimation of the proposed ties adjusted or corrected Spearman rank correlation coefficient is illustrated with the following data on the grades of a random sample of 16 students who took courses in Biology in two consecutive years in a University where one of the courses, course 1, is a prerequisite for the other course, course 2 with the following results.

Students serial no	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Course 1 (x_i)	E	C^+	A^+	A^+	C	C	A^-	B^+	C^-	C^+	C	A^-	A^+	C^-	F	F
Rank of course 1 (r_{ix})	14	7.5	2	2	10	10	4.5	6	12.5	7.5	10	4.5	2	12.5	15.5	15.5
Course 2	C^-	C^+	A^+	A^+	A^-	C	A^-	F	E	C^+	C^-	C	A^+	B^-	C	F
Rank of Course 2 (r_{iy})	2.5	7.5	2	2	4.5	10	4.5	15.5	14	7.5	12.5	10	2	6	10	15.5

To apply the proposed method we would first rank student grades in each course here from the highest grade A^+ assigned the highest rank to the lowest grade F assigned the lowest rank. All Tied grades in each course are as usual assigned their mean ranks. The ranks of the grades in each course are shown above under the grades, here considered the actual assigned ranks designated by r_i, in contrast to what is here referred to as the expected ranks, r'_i, that would be assigned if the observations were not tied in values.

Now applying Equation 13 to these ranks we obtain values of u_i and other statistics shown in Table 1 for i = 1, 2, … ,16.

Table 1: Values of u_i and Other Statistics

Student S/No	Exp Rank of course 1 (r_{ix})	ACT Rank of course 1 (r'_{ix})	Diff ($dix=r_{ix}-r'_{ix}$)	Product ($r_{ix}d_{ix}$)	d^2ix_2	$r_{ix}d_{ix}-d^2_{ix}/_2$	$(r_{ix})^2$	u_{ix}	Exp Ranks of Course 2 (r_{iy})	ACT Rank of course 2 (r_{iy})	diff ($d_{iy}=r_{iy}-(r'_{iy})$)	Product ($r_{iy}.d_{iy}$)	$\frac{d^2_{iy}}{2}$	$r_{iy}d_{iy}-\frac{d^2_{iy}}{2}$	r^2_{iy}	U_{iy}	Product ($r'_{ix}r'_{iy}$)	$d'_i=r'_{ix}-r'_{iy}$	$d_i{}^{*2}$	Product ($r_{ix}r_{iy}$)	$d_i=r_{ix}-r_{iy}$	di^2
1	14	14	0	0	0	0	196.00	0	12	12.5	-0.5	-6.00	0.125	-6.125	156.25	1	175.00	1.5	2.25	168	2	4.00
2	7	7.5	-0.5	-3.5	0.125	-3.625	56.25	1	7	7.5	-0.5	-3.5	0.125	-3.625	56.25	1	56.25	0	0	49	0	0.00
3	1	2	-1	-1	0.50	-1.50	4.00	1	1	2	1	1.00	0.50	0.50	4.00	1	4.00	0	0	1.00	0	0.00
4	2	2	0	0	0	0	4.00	0	2	2	0	0	0.00	0.00	4.00	0	4.00	0	0	4.00	0	0.00
5	9	10	-1	-9	0.50	-9.50	100.00	1	4	4.5	-0.5	-2.00	0.125	-2.125	20.25	1	20.25	5.5	30.25	36	5	25.00
6	10	10	0	0	0	0.00	100.00	0	9	10	-1	-9.00	0.50	-9.50	100.00	1	100.00	0	0	90	1	1.00
7	4	4.5	-0.5	-2	0.125	-2.12	20.25	1	5	4.5	0.5	2.50	0.125	2.375	20.25	1	20.25	0	0	20.00	-1	1.00
8	6	6	0	0	0	0	36.00	0	15	15.5	-0.5	-7.50	0.125	-7.625	240.25	1	93.00	-9.5	90.25	90.0	-9	81.00
9	12	12.5	-0.5	-6.00	0.125	-6.125	156.25	1	14	14	0	0	0.00	0.00	196.00	0	175.00	-1.5	2.25	168	-2	4.00
10	8	7.5	0.5	4.0	0.125	3.875	56.25	1	8	7.5	0.5	4.00	0.125	3.875	56.25	1	56.25	0	0.00	64.00	0	0.00
11	11	10	1	11	0.50	10.50	100.00	1	13	12.5	0.5	6.50	0.125	6.375	156.25	1	125.25	-2.5	6.25	143	-2	4.00
12	5	4.5	0.5	2.5	0.125	2.375	20.25	1	10	10	0	0.00	0.00	0.00	100.00	0	45.00	-5.5	30.25	50.0	-5	25.00
13	3	2	1	3	0.50	2.50	4.00	1	3	2	1	3.00	0.50	2.50	4.00	1	4.00	0.0	0.00	9.00	0	0
14	13	12.5	0.5	6.5	0.125	6.375	156.25	1	6	6	0	0.00	0.00	0.00	36.00	0	75.00	6.5	42.25	78.0	7	49.00
15	15	15.5	-0.5	-7.5	0.125	-7.625	240.25	1	11	10	1	11.00	0.50	10.50	100.00	1	155.00	5.50	30.25	165.0	4	16.00
16	16	15.5	0.5	8.00	0.125	7.875	240.25	1	16	15.5	0.5	8.00	0.125	7.875	240.00	1	240.25	0	0.00	256.0	0	0.00
Total	136.00	136.00		6.00	5.00	3.00	1490.00	12	136.00	136.00		8.00	5.00	5.00	1490.00	12	1348.50		234.00	1391.00		210.00

From the columns of Table 1, we have that $f_x = 12$ and $f_y = 12$, so that from Equation 18 we have that

$$\hat{\Pi}_x = \frac{12}{16} = 0.750, \ \textit{and} \ \ \Pi_y = \frac{12}{16} = 0.750$$

Hence from Eqn 19 we have that the sample estimate of the sum of squares of the ranks assigned to student grades in course 1 adjusted for ties between student scores or grades in this course using the results of Table 1 is

$$\sum_{i=1}^{n} r'_{ix} = \frac{n(n+1)(2n+1)}{6} - 2\hat{\Pi}_x \sum_{i=1}^{n}\left(r_{ix}.d_{ix} - \frac{d_{ix}{}^2}{2}\right)$$

$$= \frac{16(17)(33)}{6} - 2(0.750)(6-3) = 1496.00 - (1.5)(3)$$

$$= 1496.00 - 4.5 = 1491.500$$

Similarly the sum of squares of the ranks assigned to student grades in course 2 adjusted for ties between these grades is from Eqn 20 and Table 1

$$\sum_{i=1}^{n} r'^2 ix = 1496.00 - 2(0.750)(8-3) = 1496.00 - (1.5)(5) = 1496.00 - 7.50 = 1488.500$$

Using these results we have that the estimated variances of student scores or grades in course 1 and 2, in terms of assigned ranks adjusted for ties between student scores in each course are respectively

$$Var(r'_{ix}) = \sum_{i=1}^{n} r'^2{}_{ix} - \frac{n(n+1)^2}{4} = 1491.500 - \frac{16(17)^2}{4} = 1491.500 - 1156.00 = 335.500$$

And

$$Var(r'_{iy}) = \sum_{i=1}^{n} r'^2{}_{iy} - \frac{n(n+1)^2}{4} = 1488.500 - \frac{16(17)^2}{4} = 1488.500 - 1156.00 = 332.500$$

Hence using these results in Eqn 23 we obtain $r_{xy.c}$, the sample estimate of ties adjusted or corrected Spearman simple rank correlation coefficient between student scores or grades in courses 1 and 2 as

$$r_{xy.c} = \frac{1348.25 - \frac{16(17^2)}{4}}{\sqrt{(335.500)(332.500)}} = \frac{1348.25 - 1156.00}{\sqrt{111553.75}}$$

$$= \frac{192.25}{333.997} = 0.576$$

If we had ignored adjusting for ties within each of the sampled populations, that is courses and simply apply the usual expression for the estimation of Spearman simple rank correlation coefficient between two sampled populations we would obtain

$$\hat{\rho}_s = r_{xy} = \frac{\sum_{i=1}^{n} r'_{ix}.r'_{iy} - \frac{n(n+1)^2}{4}}{n\left(n^2-1\right)/12}$$

$$= 1348.25 - \frac{\frac{16(16+1)^2}{4}}{16\left(16^2-1\right)/12} = \frac{192.25}{34.0} = 0.565$$

a value that is here only slightly less than $r_{xy.c} = 0.576$ when adjustments have been made for ties between grades or scores in the two courses, the sampled populations.

However the ties adjusted estimate $r_{xy.c}$ is relatively more efficient than the usual estimate r_{xy} that is unadjusted or uncorrected for ties. To show this we have from Eqn 27 that

R.E $(r_{xy.c}; r_{xy}) = \frac{\text{var}(r_{xy})}{\text{var}(r_{xy.c})} = \frac{(340)(340)}{(335:500)(332:500)}$

= 1.036>1, as required.

Note that if we had completely ignored adjusting for ties in the sampled populations and treated the grades on scores as if they are untied assigning them different ranks we would erroneously estimate the spearman rank correlation coefficient as

$$r_{xy} = \sum_{1=1}^{n} r_{ix} r_{iy} - \frac{\frac{n(n+1)^2}{4}}{4080/12} = \frac{1391 - \frac{16(17^2)}{4}}{340} = \frac{1391-1156}{340} = \frac{235}{340} = 0.691,$$

which is an over-estimate, resulting from the fact that the cross-product of remarks in the numerator is over estimated because of the effect of ties which is here ignored.

On the other hand if in the presence of tied observations in the populations, we had estimated the required spearman rank simple correlation coefficient from first principle using Egn 7 we would have that

$$\hat{\rho}_s = r_{xy} = \frac{1348.25 - \frac{16(17^2)}{4}}{\sqrt{\left((1490.00) - \frac{16(17^2)}{4}\right)\left((1490.00) - \frac{16(17^2)}{4}\right)}}$$

$$= \frac{1348.25 - 1156}{\sqrt{(1490.00-1156)(1490.00-1156)}}$$

$$= \frac{192.25}{\sqrt{(334.00)(334.00)}} = \frac{192.25}{334.00} = 0.576$$

which is exactly the same value as $r_{xy.c}$, the estimated ties adjusted simple rank correction coefficient for grades in the two courses.

IV. Summary And Conclusion

We have in the papers proposed, developed and presented a method for adjusting and correcting sample estimates of the usual spearman rank correlation coefficient between two populations to correct or adjust for the possible presence of tied observations in the sampled populations. Without these adjustments the sum of squares of ranks in the denominations of the usual expression for the estimation if spearman rank correlation coefficient are often over estimated of there are ties in the populations and no effort is made to correct and adjust for them, especially if the ties are not few. This would often result in the required rank correlation coefficient being under-estimated, inefficient, unreliable, thereby leading to misleading conclusions.

The proposed method is shown to produce more efficient estimates, with smaller variances.

The method is also shown to be easily modified and used to estimate ties adjusted spearman partial rank correlation coefficient between any two populations holding a third population at constant levels, with all the populations adjusted for the possible presence of ties.

The usual statistical tests for significance can also be readily applied using the ties adjusted rank correlation coefficients.

References

[1]. Gibbons, J. D. (1971): Non parametric Statistical Interference. McGraw-Hill Book Company, New York.
[2]. Siegel S. (1956): Nonparametric Statistics for the Behavioural Sciences. McGraw-Hill Kogakusha, Ltd. Tokyo
[3]. Maritz, J. S. (1981): Distribution-Free Statistical Methods. Chapman and Hall.
[4]. Kendall, M. G. (1970): Rank Correlation Methods (4th Edition). London: Griffin.
[5]. Hollander M, Wolfe D. A. (1973): Nonparametric Statistical Methods. New York; Wiley.