American Journal of Theoretical and Applied Statistics
2013; 2(6): 273-281
Published online January 10, 2014 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20130206.26

Intrinsically ties adjusted Tau (C-Tat) correlation coefficient

OYEKA CYPRIL ANENE, OSUJI GEORGE AMAEZE*, NWANKWO CHRISTIAN CHUKWUEMEKA

Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria

Email address:
george.osuji99@yahoo.com(O. G. AMAEZE)

To cite this article:
OYEKA CYPRIL ANENE, OSUJI GEORGE AMAEZE, NWANKWO CHRISTIAN CHUKWUEMEKA. Intrinsically Ties Adjusted Tau (C-Tat) Correlation Coefficient. *American Journal of Theoretical and Applied Statistics.* Vol. 2, No. 6, 2013, pp. 273-281.
doi: 10.11648/j.ajtas.20130206.26

Abstract: This paper proposes a method for correcting and adjusting the usual or regular estimates of Tau correlation coefficients for the possibility of ties within and between observations in the population being correlated. The index here called C-Tat for 'ties adjusted Tau correlation coefficient' is formulated to intrinsically and structurally adjust and correct the estimated Tau correlation coefficient for the possible presence of tied observations in the sampled populations and for the fact that the estimates obtained are often dependent on, that is, vary depending on which of the two populations under study has its assigned ranks arranged in their natural order and which has its assigned ranks arranged in their natural order and which has its assigned ranks tagged along. The proposed method is illustrated with some sample data and shown to yield more reliable and efficient estimates of tau correlation coefficients than the usual method which is able to give the same estimates only if there are no tied observations what-so-ever in the sampled populations.

Keywords: Tau Correlation Coefficient, C-Tat, Tied Observations and Ties Adjusted

1. Introduction

The Kendall's tau, correction coefficient ρ_k is the non-parametric equivalence of the well known Pearson's moment correlation coefficient ρ_p commonly used in parametric statistics to measure the strength of association between two continuous normally distributed populations. Here the populations of interest X and Y need not be continuous and normally distributed but only need to be measured on at least the ordinal scale. If random sample observations drawn from these populations are each ranked in the usual way, the Kendall's sample tau correlation coefficient r_k will give a measure of the degree of association or correlation between the two sets of ranks. To calculate an estimate of the Kendall's tau correlation coefficient we first rank the n sample observations drawn from population X either from the smallest to the largest or from –the largest to the smallest. The n sample observations drawn from populations Y are similarly ranked. Tied observations in each sample are as usual assigned their mean ranks.

The ranks assigned to one of the samples, for example those assigned to observations from X are now arranged in their natural order from '1' through 'n' together with the labels of the subjects or respondents with these ranks. The ranks assigned to the observations in the sample drawn from the second population Y say, are then juxtaposed against the naturally arranged ranks for the corresponding subjects in the sample drawn from the first population X. Interest is now to determine the degree of agreement between the rankings of observations from X with these from Y. since the ranks of observations from X are already arranged in their natural order; interest is then actually, in determining how many pairs of ranks of observations form population Y are also in their natural order relative to each other.

Now the maximum possible total number of agreements between the ranks assigned to observations from population X and Y would be obtained if the rankings are in perfect agreement, in this case when the rankings of observations from Y are also in their natural order as those for observations from populations X. This is the rationale for Kendall's tau correlation coefficient between two populations X and Y which is estimated as the ratio of the actual observed total agreement score to the total maximum possible score under perfect agreement, where the total

maximum possible agreement score is

$$S\max = \binom{n}{2} = \frac{n(n-1)}{2} \tag{1}$$

This is the basis for the estimation of Kendall's tau correlation coefficient between populations X and Y as

$$r_k = rxy = \frac{Sa}{S\max} \tag{2}$$

Where Sa is the total sum of 1s (or+) and –1s (or-) obtained by comparing members of each pair of the ranks assigned to observations from one of the variables Y say in relation. To each other when these observations are arranged in accordance with the naturally ordered ranks of the observations from the other variable X say. Thus Kendall's tau correlation coefficient between two populations is estimated as the ratio of the actual observed total agreement score to the total maximum possible score under perfect agreement. However the approach in Equation 2 does not immediately provide for the possible presence of ties in either the X or Y variable or in both. Although an alternative formulae exists for the estimation of tau when ties occur in the data, this formulae is however often tedious to use in practice. The estimate of the corresponding standard deviation is also cumbersome to evaluate. Moreover, the estimates obtained are not always independent of which of the two samples has its ranks arranged in their natural order and which of the samples has its ranks tagged along. Different estimates are often obtained.

We here propose a more formatted and generalized method that covers situations in which there are no tied observations in any of the two populations of interest when only one of the populations has tied observations, as well as when there are equal and unequal number of tied observations in the two populations.

2. The Proposed Method

Let x_i and y_i be the ith observation in random samples of size 'n' drawn from populations X and Y respectively which may be measurements on at least the ordinal scale, for i =1, 2,…,n. suppose the 'n' observations x_i drawn from population X are ranked from the smallest to the largest say, and the 'n' observations y_i drawn from population Y are similarly ranked. Tied observations in each sample are as usual assigned their mean ranks.

Furthermore suppose x_i is assigned the ranks of rix and y_i the rank of riy for i =1,2,…,n. To correct or to adjust for the fact that the estimated correlation coefficient is not always independent of which of the two populations has the ranks assigned to the observations from it as the ones that are naturally ordered and which has the ranks assigned to its observations as the ones that are tagged along, these two sets of ranks would each here be required to alternately play each of these two roles.

Hence, first suppose that the xi observations from population X have been arranged in their natural order from the smallest to the largest so that rix = i, i =1,2,…,n, and that the ranks riy for observations y_i from Y have each been arranged to correspond with the now naturally ordered ranks i for their sister observation xi from population X.

Define

$$ujk; y.x = \begin{cases} 1, \; if \;\; rjy < nky \\ 0, \; if \;\; rjy = rky \\ -1, \; if \;\; rjy > rky \end{cases} \tag{3}$$

for j =1,2,…,n-1; k = 2,3,…,n; j < k. That is provided that the rank assigned to the kth observation from population Y comes after, that is succeeds the rank assigned to the jth observation from the same population when these observations are arranged in accordance with the natural ordering or ranking of the corresponding observations from population X $(j < k)$

Let

$$\begin{aligned} \overset{+}{\prod}_y &= P(ujk; y.x = 1); \\ \overset{0}{\prod}_y &= P(ujk; y.x = 0); \\ \overset{-}{\prod}_y &= P(ujk; y.x = -1) \end{aligned} \tag{4}$$

Where

$$\overset{+}{\prod}_y + \overset{0}{\prod}_y + \overset{-}{\prod}_y = 1 \tag{5}$$

and $\overset{0}{\prod}_y = 0$, if there are no ties in population Y.

Note that by their specifications Equations 3-5 have provided adjustments for the possible presence of ties in population Y.

Now let

$$Sy = \sum_{\substack{j=1 \\ j<k}}^{n-1} \sum_{k=2}^{n} ujk; y.x. \tag{6}$$

Now from Equations (3) and (4) we have that

$$\begin{aligned} E(ujk; y.x) &= \overset{+}{\prod}_y - \overset{-}{\prod}_y; \\ Var(ujk; y.x) &= \overset{+}{\prod}_y + \overset{-}{\prod}_y - \left(\overset{+}{\prod}_y - \overset{-}{\prod}_y \right)^2 \end{aligned} \tag{7}$$

Similarly,

$$E(Sy)=\sum_{\substack{j=1\\ j<k}}^{n-1}\sum_{k=2}^{n}E(ujk;y.x)$$

That is:

$$\begin{aligned}E(Sy)&=\tfrac{1}{2}n(n^2-n)\left(\prod_y^{+}-\prod_y^{-}\right)\\&=\tfrac{1}{2}n(n-1)\left(\prod_y^{+}-\prod_y^{-}\right)\end{aligned}\tag{8}$$

Note that $\prod_y^{+},\prod_y^{0}$ *and* $\prod_y^{-}$ are respectively the probabilities that the rank assigned to the jth observation from population Y is less than, equal to or greater than the rank assigned to the kth observation from the same population if the rank assigned to the kth observation succeeds the rank assigned to the jth observation from Y when these observations are arranged in accordance with the natural ordering of the ranks assigned to their sister observations from population X these probabilities are estimated as

$$\begin{aligned}\overset{\cap}{\prod}_y^{+}&=\frac{f_y^{+}}{\binom{n}{2}}=2\frac{f_y^{+}}{n(n-1)};\\ \overset{\cap}{\prod}_y^{0}&=2\frac{f_y^{0}}{n(n-1)};\\ \overset{\cap}{\prod}_y^{-}&=2\frac{f_y^{-}}{n(n-1)}\end{aligned}\tag{9}$$

where f_y^{+}, f_y^{o} and f_y^{-} are respectively the number of 1s, 0s and -1s in the frequency distribution of the ½ n(n-1) values of these number in ujk;y.x, j = 1,2,…,n-1; k=2,3,…,n: j < k.

Hence the sample estimate of the observed total number of times the rankings of observations from population Y are in their natural order and consistent with the natural ordering of the ranks of observations from population X less the number of times they are out of order is from Equation8

$$Sy=\tfrac{1}{2}n(n-1)\left(\prod_y^{+}-\prod_y^{-}\right)=f_y^{+}-f_y^{-}\tag{10}$$

As noted above if these rankings are in their natural order, then the maximum possible total number of arrangement or scores is $Smax=\binom{n}{2}$, (see Eqn (1)). Hence the Kendall's tau correlation coefficient between X and Y may be estimated using Equation 10 in Equation 2 as

$$rxy=xry=\frac{S_y}{S\max}=\frac{\tfrac{1}{2}n(n-1)\left(\overset{\cap}{\prod}_y^{+}-\overset{\cap}{\prod}_y^{-}\right)}{\tfrac{1}{2}n(n-1)}$$

Or

$$rxy=xry=\overset{\cap}{\prod}_y^{+}-\overset{\cap}{\prod}_y^{-}=\frac{2\left(f_y^{+}-f_y^{-}\right)}{n(n-1)}\tag{11}$$

Notes that the variance of S_y is from Equation 6

$\mathrm{Var(Sy)}=\sum_{\substack{j=1\\ j<k}}^{n-1}\sum_{k=2}^{n}\mathrm{Var(ujk;y.x)}$, which from Eqn 7 is

$$Var(Sy)=\tfrac{1}{2}n(n-1)\left(\prod_y^{+}+\prod_y^{-}\right)-\left(\prod_y^{+}-\prod_y^{-}\right)^2\tag{12}$$

A sample estimate of this variance is

$$Var(Sy)=\tfrac{1}{2}n(n-1)\left(\overset{\cap}{\prod}_y^{+}+\overset{\cap}{\prod}_y^{-}\right)-\left(\overset{\cap}{\prod}_y^{+}-\overset{\cap}{\prod}_y^{-}\right)^2\tag{13}$$

Note that Equations 10 and 13 have been adjusted for the possibility of ties in population Y.

The estimated variance of rxy = xry is

$$\begin{aligned}&Var(rxy)=Var(xry)=Var\left(S_y/\binom{n}{2}\right)\\&=Var\frac{Sy}{\left(n(n-1)\right)/4}=\frac{4Var(Sy)}{n^2(n-1)^2}\\&which\ \ from\ \ Eqn\ 13\ 1s\\&Var(rxy)=Var(xry)\\&=\frac{2(\overset{\cap}{\prod}_y^{+}+\overset{\cap}{\prod}_y^{-}-(\overset{\cap}{\prod}_y^{+}-\overset{\cap}{\prod}_y^{-})^2)}{n(n-1)}\end{aligned}\tag{14}$$

The null hypothesis

$$Ho;\rho_{xy}=\rho_o Versus\ H_i:\rho_{xy}>\rho_{o\ say}\quad(-1\le\rho_o\le1)\tag{15}$$

May be tested using the test statistic

$\chi^2=\frac{(\mathrm{rxy}-\rho_0)^2}{\mathrm{Var(rxy)}}$ which from Equations (11) and (14) is

$$\chi^2=\frac{n(n-1)(\left(\overset{\cap}{\prod}_y^{+}-\overset{\cap}{\prod}_y^{-}\right)-\rho_o)^2}{2\left(\left(\overset{\cap}{\prod}_y^{+}+\overset{\cap}{\prod}_y^{-}\right)-\left(\overset{\cap}{\prod}_y^{+}-\overset{\cap}{\prod}_y^{-}\right)^2\right)}\tag{16}$$

Which has approximately a chi- square distribution with I degree of freedom for sufficiently large. $n\,(n\ge10)$

The null hypothesis H_0 is rejected at the α level of significance if

$$\chi^2 \geq \chi^2_{1-\alpha;1} \tag{17}$$

Otherwise H_0 is accepted

Equation 11 and 16 can not be correctly used to estimate tau correlation coefficient and test appropriate null hypothesis unless there are no tied observations in the data, that is unless there are no tired observations in both populations X and Y. this is because even though S_y of Equation 11 by specifications has been adjusted, for the possible presence of ties in population Y, its denominator Smax has not been so adjusted. Further more as noted above rxy is not always independent of which sample has its assigned ranks arranged in their natural order and which sample has its assigned ranks tagged along. Eqns 11 and 16 therefore need to be appropriately modified. Thus Smax of Eqn 11 needs to be modified to reflect the possible presence of tied observations in population Y. this is done by subtracting f_y^o, the observed number of ties in population y from Smax, thereby obtaining.

$$S\max.C = S\max - f_y^o = \binom{n}{2} - f_y^o \tag{18}$$

Now using Eqn 18 in Eqn 11 yields an estimates of the ties corrected or adjusted tau correlation coefficient between population x and y, when there are ties observations in Y as

$$\text{xry.c} = \frac{S_y}{\text{Smax.c}} = \frac{f_y^+ - f_y^-}{\binom{n}{2} - f_y^0} = \frac{\left(f_y^+ - f_y^-\right)/\binom{n}{2}}{1 - f_y^o/\binom{n}{2}}$$

That is:

$$xry.c = \frac{xry}{1-\overset{\cap}{\underset{o}{\prod}}_y} = \frac{\overset{\cap}{\underset{+}{\prod}}_y - \overset{\cap}{\underset{+}{\prod}}_y}{1-\overset{\cap}{\underset{o}{\prod}}_y} \tag{19}$$

Note from Eqn 19 that failure to adjust xry for the presence of ties in Y would lead to an underestimation of the true Tau correlation coefficient, a bias that is seen to increase with the number of tried observations in Y, that is with $\overset{o}{\prod}_y$ .The variance of xry.c is estimated as

$$\text{Var}(\text{xry.c}) = \text{Var}\left(\text{xry}/\left(1-\overset{\cap}{\underset{o}{\prod}}_y\right)\right) = \frac{\text{Var}(\text{xry})}{\left(1-\overset{\cap}{\underset{o}{\prod}}_y\right)^2}$$

Is given in Equation 14 Hence,

$$Var(xry.c) = \frac{2\left(\overset{\cap}{\underset{+}{\prod}}_y + \overset{\cap}{\underset{-}{\prod}}_y - \left(\overset{\cap}{\underset{+}{\prod}}_y - \overset{\cap}{\underset{-}{\prod}}_y\right)^2\right)}{n(n-1)\left(1-\overset{\cap}{\underset{o}{\prod}}_y\right)^2} \tag{20}$$

To further adjust the estimate of the Tau correlation coefficient obtained in Equation 19 for the fact that in the presence of ties, its value depends on which of the two populations has its assigned ranks arranged in their natural order and which has its assigned ranks tagged along, we now interchange the rolls of X and Y so that observations drawn from Y now have their assigned ranks arranged in their natural order, and observations drawn from X have their ranks tagged along. We then define.

$$Ujk;x.y = \begin{cases} 1, & if \quad rjx < rkx \\ 0, & if \quad rjx = rkx \\ -1, & if \quad rjx > rkx \end{cases} \tag{21}$$

for j = 1, 2,..., n-1; k=2,3,…,n. j < k. In other words provided that the rank assigned to the kth observation from population X comes after, that is succeeds the rank assigned to the jth observation from the same populations When these observations are arranged in accordance with the natural ordering or ranking of the corresponding observations from population

Y (j < k), let

$$\begin{aligned} \overset{+}{\prod}_x &= P(ujk;x.y=1); \\ \overset{o}{\prod}_x &= P(ujk;x.y=0); \\ \overset{-}{\prod}_x &= P(ujk;x.y=-1) \end{aligned} \tag{22}$$

$$where \ \overset{+}{\prod}_x + \overset{o}{\prod}_x + \overset{-}{\prod}_x = 1 \tag{23}$$

and $\overset{o}{\prod}_x = 0,$ if there are no ties in X Also define

$$S_x = \sum_{\substack{j=1 \\ j<k}}^{n-1} \sum_{k=2}^{n} ujk;x.y \tag{24}$$

Now as before

$$\begin{aligned} E(ujk;x.y) &= \overset{+}{\prod}_x - \overset{-}{\prod}_x; \\ Var(ujk;x.y) &= \overset{+}{\prod}_x + \overset{-}{\prod}_x - \left(\overset{+}{\prod}_x - \overset{-}{\prod}_x\right)^2 \end{aligned} \tag{25}$$

$$And\ E(S_x)=\frac{1}{2}n(n-1)\left(\prod_x^+-\prod_x^-\right) \tag{26}$$

Note that $\prod_x^+, \prod_x^o$ *and* $\prod_x^-$ are respectively the probabilities that the rank assigned to the jth observation from population X is less than, equal to or greater than the rank assigned to the kth observations from the same population if the rank assigned to the jth observation succeeds the rank assigned to the jth observation from X when these observations are arranged in accordance with the natural ordering of the ranks assigned to the observations from population Y. these probabilities are estimated as

$$\hat{\prod}_x^+=\frac{2f_x^+}{n(n-1)};\ \hat{\prod}_x^0=\frac{2f_x^o}{n(n-1)};\ \hat{\prod}_x^-=\frac{2f_x^-}{n(n-1)} \tag{27}$$

where f_x^+, $f_x^+ - f_x^o$ are respectively the number of 1s, Os and –1s in the frequency distribution of the ½ n(n - 1) values of these numbers in ujk; x.y, f or j = 1,2,… ,n-1, k= 2,3, ...,n-1; j < k.

+1hence the sample estimate of the total number of times the rankings of observations from population x are in their natural order and consistent with the natural ordering of the ranks of observations from population Y less the number of times they are out of order is from Eqn (26)

$$S_x=\tfrac{1}{2}n(n-1)\left(\hat{\prod}_x^+-\hat{\prod}_x^-\right)=f_x^+-f_x^- \tag{28}$$

Therefore as before an estimate of Kendall's Tan correlation coefficient not yet corrected for ties in X is

$$yrx=\frac{S_x}{S\max}=\hat{\prod}_x^+-\hat{\prod}_x^-=\frac{2\left(f_x^+-f_x^-\right)}{n(n-1)} \tag{29}$$

The corresponding estimate of the variance is

$$Var(yrx)=\frac{2\left(\hat{\prod}_x^++\hat{\prod}_x^-\left(\hat{\prod}_x^+-\hat{\prod}_x^-\right)^2\right)}{n(n-1)} \tag{30}$$

Hence the estimated ties adjusted or corrected tau correlation coefficient between X and Y when there are ties in X is obtained using Eqn 18 in Eqn 29 as

$$\text{yrx.c}=\frac{S_x}{S\max.c}=\frac{S_x}{S\max-f_x^o}=\frac{S_x/\binom{n}{2}}{1-f_x^o/\binom{n}{2}}$$

That is:

$$yrx.c=\frac{yrx}{1-\hat{\prod}_x^o}=\frac{\hat{\prod}_x^+-\hat{\prod}_x^-}{1-\hat{\prod}_x^o} \tag{31}$$

Also in Equation 20 the estimated variance of yrx.c is using Equation 30

$$Var(yrx.c)=\frac{2\left(\hat{\prod}_x^++\hat{\prod}_x^--\left(\hat{\prod}_x^+-\hat{\prod}_x^+\right)^2\right)}{n(n-1)\left(1-\hat{\prod}_x^o\right)^2} \tag{32}$$

Now a sample estimate of the ties-adjusted or corrected tau (C-TAT) correlation coefficient between X and Y is the weighted average of the ties-adjusted tau correlation coefficient for tied observations in X and Y, where the weights are functions of the proportions of tied observations in each population namely

$$\text{rxy.c}=\left(1-\hat{\prod}_x^o\right)\text{yrx.c}+\frac{\left(1-\hat{\prod}_y^0\right)\text{xry.c}}{2-\hat{\prod}_x^o-\hat{\prod}_y^o}$$

That is:

$$rxy.c=\frac{yrx+xry}{2-\hat{\prod}_x^o-\hat{\prod}_y^o}=\frac{(\hat{\prod}_x^+-\hat{\prod}_x^-)+(\hat{\prod}_y^+-\hat{\prod}_y^-)}{2-\hat{\prod}_x^o-\hat{\prod}_y^o} \tag{33}$$

Note that if there are no tied observations in populations X and Y, that is

$$\prod_x^o=\prod_y^o=0,\qquad \text{then}$$

rcy.c=rxy=xry.c=yrx.c=yrx=xry;that is $rxy=\hat{\prod}_y^+-\hat{\prod}_y^-=\hat{\prod}_x^+-\hat{\prod}_x^-$. Hence, here, there is no need for adjustments. The estimated ties adjusted tau correlation coefficient remains unchanged no matter which of the two sampled populations has its assigned ranks arranged in their natural order and which has its assigned ranks tagged along.

If the two populations have equal number of tied observations, that is if

$$\overset{o}{\prod}_x = \overset{o}{\prod}_y, \text{ then } rxy.c = \frac{xry + yrx.c}{2\left(1 - \overset{\frown}{\overset{o}{\prod}}_y\right)};$$

$$\text{that is } rxy.c = \frac{\left(\overset{\frown}{\overset{+}{\prod}}_y - \overset{\frown}{\overset{-}{\prod}}_y\right) + \left(\overset{\frown}{\overset{+}{\prod}}_x - \overset{\frown}{\overset{-}{\prod}}_x\right)}{2\left(1 - \overset{\frown}{\overset{o}{\prod}}_y\right) - \overset{\frown}{\overset{o}{\prod}}_x}$$

In other cases in which the two populations have unequal number of tied Observations, that is when $\overset{o}{\prod}_x \neq \overset{o}{\prod}_y$, the estimate obtained also depends on which of the sampled populations has its assigned ranks arranged in their in their natural order and which has its assigned rank tagged along. The estimated ties adjusted tau correlation coefficient is now taken as a weighted average of the two estimates (Eqn 33)

To estimate the variance of rxy.c of Equation 33 we have that

$$Var(rxy.c) = \frac{Var\left((xry + yrx)/\left(2 - \overset{\frown}{\overset{o}{\prod}}_x \overset{\frown}{\overset{o}{\prod}}_y\right)\right) = Var(xry + yrx)}{\left(2 - \overset{\frown}{\overset{o}{\prod}}_x - \overset{\frown}{\overset{o}{\prod}}_y\right)^2}$$

$$= \frac{\text{var}(xry) + Var(yrx) + 2Cov(xry, yrx)}{\left(2 - \overset{\frown}{\overset{o}{\prod}}_x - \overset{\frown}{\overset{o}{\prod}}_y\right)^2}$$

Where $\text{Var}(\text{yrx})$ and $\text{Var}(\text{xry})$ are given in Equations 30 and 14 respectively it is easily shown that $\text{cov}(\text{yrx}, \text{xry}) = 0$. To do this is sufficient to show that $\text{Cov}\,(\text{Sx}, \text{Sy}) = 0$ we here note that $\text{Cov}\,(\text{S}_\text{x}, \text{S}_\text{y}) = \text{E}(\text{S}_\text{x}\,\text{S}_\text{y}) - \text{E}(\text{S}_\text{x})\text{E}(\text{S}_\text{y})$ where $\text{E}(\text{S}_\text{x})$ and $\text{E}(\text{S}_\text{y})$ are given in Equations 8 and 26 respectively, and

$$E(SxSy) = E\left(\sum_{\substack{j=1 \\ j<k}}^{n-1} \sum_{k=2}^{n} ujk; y.x\right)\left(\sum_{\substack{j=1 \\ j<k}}^{n-1} \sum_{k=2}^{n} ujk; x.y\right)$$

$$= \sum_{\substack{r,q=1 \\ r<s}}^{n-1} \sum_{\substack{s,h=2 \\ q<h}}^{n} E(urs; y.x\ ugh; x.y)$$

The three possible values urs; y.x and ugh;x.y can assume are 1, 0 and -1. It assumes the values 1 if urs; y.x and ugh;x.y both assume the value 1 or

The value -1 with probability $\overset{+}{\prod}_x \overset{+}{\prod}_y + \overset{-}{\prod}_x \overset{-}{\prod}_y$; it assumes the value 0 if one of its two factors urs;y.x or ugh;x.y assumes the value 0 no matter the value assumed by the other factor with probability $\overset{o}{\prod}_x\left(\overset{+}{\prod}_y + \overset{-}{\prod}_y\right) + \overset{o}{\prod}_y\left(\overset{+}{\prod}_x + \overset{-}{\prod}_x\right)$; and assumes the value -1 if urs;y.x assumes the value 1 and ugh;x.y assumes the Value -1 or vice versa with probability $\overset{+}{\prod}_x \overset{-}{\prod}_y + \overset{-}{\prod}_x \overset{+}{\prod}_y$. Hence $\text{E}(\text{urs; y.x Ugh; x.y})\overset{+}{\prod}_x \overset{-}{\prod}_y + \overset{+}{\prod}_x \overset{-}{\prod}_y - \left(\overset{+}{\prod}_x \overset{-}{\prod}_y + \overset{=}{\prod}_x \overset{+}{\prod}_y\right)$; Therefore

$$Cov(Sx, Sy) = \left(\tfrac{1}{2}n(n-1)\right)^2 \left(\begin{array}{l} \overset{+}{\Pi}_x \overset{+}{\Pi}_y + \overset{-}{\Pi}_x \overset{-}{\Pi}_y - \\ \left(\overset{+}{\Pi}_x \overset{-}{\Pi}_y + \overset{-}{\Pi}_x \overset{+}{\Pi}_y\right)\left(\overset{+}{\Pi}_x - \overset{-}{\Pi}_x\right)\left(\overset{+}{\Pi}_y - \overset{-}{\Pi}_y\right) \end{array}\right) = 0$$

Collecting terms we have that the variance of the estimated tau-correlation coefficient between X and Y when each of the two populations has ties observations is estimated as

$$\text{Var}(\text{rxy.c}) = \text{Var}(\text{yrx.c} + \text{xry.c}) = \frac{\text{Var}(\text{xry}) + \text{Var}(\text{yrx})}{\left(2 - \overset{\frown}{\overset{o}{\prod}}_x - \overset{\frown}{\overset{o}{\prod}}_y\right)^2}$$

Which from Equation 14 and 33 is

$$Var(rxy.c) = \frac{2\left(\overset{\frown}{\overset{+}{\prod}}_x + \overset{\frown}{\overset{-}{\prod}}_x\left(\overset{\frown}{\overset{+}{\prod}}_x - \overset{\frown}{\overset{-}{\prod}}_x\right)^2 + \overset{\frown}{\overset{+}{\prod}}_y + \overset{\frown}{\overset{-}{\prod}}_y - \left(\overset{\frown}{\overset{+}{\prod}}_y - \overset{\frown}{\overset{-}{\prod}}_y\right)^2\right)}{n(n-1)\left(2 - \overset{\frown}{\overset{o}{\prod}}_x - \overset{\frown}{\overset{o}{\prod}}_y\right)^2} \quad 34)$$

or equivalently

$$\text{Var (rxy.c)} = \frac{\left(1 - \overset{\frown}{\overset{0}{\prod}}_y\right)^2 Var(xry.c) + \left(1 - \overset{\frown}{\overset{o}{\prod}}_x\right)^2 Var(yrx.c)}{\left(2 - \overset{\frown}{\overset{o}{\prod}}_x - \overset{\frown}{\overset{o}{\prod}}_y\right)^2} \quad (35)$$

Where Var (xry.c) and Var (yrx.c) are given in Eqns 20 and 32 respectively to test the null hypothesis of Equation 15 we may use the test statistic

$$\chi_o^2 = \frac{(rxy.c - \ell o)^2}{Var(rxy.c)} \quad (36)$$

Which has approximately a chi-square distribution with 1 degree of freedom for sufficiently large n (n≥10) where rxy.c and Var (rxy.c) are given in Equation 33 and 35 respectively? H_0 is rejected at the α level of significance if Equation 17 is satisfied; otherwise H_0 is accepted.

3. Illustrative Example

Let us use the following letter grades which are measurements on the ordinal scale earned by a random sample of students in two courses in statistics to illustrate

the estimation of ties adjusted tau (C-TAT) correlation coefficient when there are ties in the data.

Table 1. *A random sample of students grades in two courses*

Students Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Grade in course 1(x_i)	B	C	B	C^-	F	F	A^+	F	C^+	C	A^-	C^-
Ranks of Grades in Course 1(rix)	9.5	6.5	9.5	4.5	2	2	12	2	8	6.5	11	4.5
Grade in course 2 (y_i)	E	A^-	C^+	C-	B	B	F	C^+	B^+	B^+	F	B^-
Rank of Grade in course 2 (riy)	3	12	5.5	4	8.5	8.5	1.5	5.5	10.5	10.5	1.5	7

The student grades in each course are ranked from the lowest (F) through the highest (A^+) assigning the rank of 1 to F, the rank of 2 to the next higher grades and finally the rank of 12 to the highest grade, A+. All tied grades in each course are as usual assigned their mean ranks. The results of the ranking are shown above, below each of the grades in the courses. To estimate ties adjusted tau correlation coefficient using these data we now arrange the ranks assigned to the grades in one of the courses, here course 1 (rix) in their natural order. The rank (rix) of the grade earned by each student in the second course is then written along side the naturally ordered rank of the corresponding grade by the

Student in course 1, the results are show in table 1

To estimate the ties adjusted tau correlation coefficient when there are ties in Y. we apply Eqn 3 to the ranks riy in the second column of table 1 which is here for greater clarity presented in a tabular form (table 2)

Table 2. *Naturally ordered ranks for grade in course 2 (Y) in course 1 (X) with corresponding ranking for grades in Course 2 (Y)*

Students Number	**5**	**6**	**8**	**4**	**12**	**2**	**10**	**9**	**1**	**3**	**11**	**7**
Natural Order of Ranks in Grade in course 1(rix)	2	2	2	4.5	4.5	6.5	6.5	8	9.5	9.5	11	12
Corresponding Ranks for Grades in Course 2(riy)	8.5	8.5	5.5	4	7	12	10.5	10.5	3	5.5	1.5	1.5

Table 3. *Calculation of ujk;y.x (Eqn.3) for the data of table 1*

Student No.		**5**	**6**	**8**	**4**	**12**	**2**	**10**	**9**	**1**	**3**	**11**	**7**
Rank for course 2(rky)		8.5	8.5	5.5	4	7	12	10.5	10.5	3	5.5	1.5	1.5
Student Number	Rank for course 2(rjy)												
5	8.5		0	-1	-1	-1	1	1	1	-1	-1	-1	-1
6	8.5			-1	-1	-1	1	1	1	-1	-1	-1	-1
8	5.5				-1	1	1	1	1	-1	0	-1	-1
4	4					1	1	1	1	-1	1	-1	-1
12	7						1	1	1	-1	-1	-1	-1
2	12							-1	-1	-1	-1	-1	-1
10	10.5								0	-1	-1	-1	-1
9	10.5									-1	-1	-1	-1
1	3										1	-1	-1
3	5.5											-1	-1
11	1.5												0
7	1.5												

Note that in table 2 we used the ranks assigned to the observations from X as the ones that are naturally ordered and the ranks assigned to the observations from Y as those that are tagged along. This enables the isolation and estimation of the effect of ties in Y using the results of the table 2. thus from table 2 we have that

$$f_y^+ = 19; f_y^o = 4; \text{ and } f_y^- = 43$$

Hence from Equation 9 we have that

$$\overset{\cap}{\prod}{}^{+}_{y} = \frac{19}{66} = 0.288;$$

$$\overset{\cap}{\prod}{}^{o}_{y} = \frac{4}{66} = 0.061;$$

$$and \ \overset{\cap}{\prod}{}^{-}_{y} = \frac{43}{66} = 0.652.$$

Therefore from Equation 11, we obtain an estimate of an uncorrected or unadjusted tau correlation coefficient as:

$$Xry=0.288-0.652= -0.364$$

Therefore an estimate of the Tau correlation coefficient adjusted for ties in Y is from Equation 19

$$Xry.c=\frac{0.288-0.652}{1-0.061}=-\frac{0.364}{0.939} = -0.388$$

Also from Eqn 20 we have that

$$Var\ (xry.c) = \frac{2(0.288+0.652-(0.288-0652)^2)}{12(12-1)\ (1-0.061)^2}$$

$$=\frac{1.616}{116.424} \quad =0.014$$

Having obtained an estimate of tau correlation coefficient between X and Y using the ranks assigned to the observations from X as the ones that are naturally ordered and the ranks assigned to the observations from Y as the ones that are tagged along, we now interchange the roles of the ranks assigned to observations from X and Y to obtain an estimate of the Tau correlation coefficient yrx using Eqn 22

Table 3 shows the ranks assigned to the observations from Y arranged in their natural order together with the ranks of the corresponding observation from X

Table 4. *Naturally Ordered ranks for Students Grades in course 2 (Y) with corresponding ranks in course 1 (X)*

Students Number	**7**	**11**	**1**	**4**	**3**	**8**	**12**	**5**	**6**	**9**	**10**	**2**
Natural Order of Ranks in Grade in course 2(riy)	1.5	1.5	3	4	5.5	5.5	7	8.5	8.5	10.5	10.5	12
Corresponding Ranks for Grades in Course 1(rix)	12	11	9.5	4.5	9.5	2	4.5	2	2	8	6.5	6.5

We now use the data of table 3 with Equation 22 to obtain an estimate of yrx. The results are presented in a tabular form (table 4)

Table 5. *Calculation of ujk;xy (Equation 22) for the data of table 3*

Student No.		**7**	**11**	**1**	**4**	**3**	**8**	**12**	**5**	**6**	**9**	**10**	**2**
rank in course 1 (rkx)		12	11	9.5	4.5	9.5	2	4.5	2	2	8	6.5	6.5
Student Number	Rank in course 1(rjx)												
7	12		-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
11	11			-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	9.5				-1	0	-1	-1	-1	-1	-1	-1	-1
4	4.5					1	-1	0	-1	-1	1	1	1
3	9.5						-1	-1	-1	-1	-1	-1	-1
8	2							1	0	0	1	1	1
12	4.5								-1	-1	1	1	1
5	2									0	1	1	1
6	2										1	1	1
9	8											-1	-1
2	6.5												0

From table 4 we have $f_x^+ = 17;\ f_x^0 = 6;\ f_x^- = 43$

Hence from Equation 28 we have that

$$\overset{\cap}{\prod}{}^{+}_{x} = \frac{17}{66} = 0.258; \overset{\cap}{\prod}{}^{o}_{x} = \frac{6}{66} = 0.091 \text{ and } \overset{\cap}{\prod}{}^{-}_{x} = \frac{43}{66} = 0.652$$

Therefore using Equation 30 we have that the uncorrected or unadjusted Tau correlation coefficient between X and Y based on the ranks assigned to the observations from X with ties is estimated as yrx=0.258-0.652=-0.394

Hence an estimate of ties adjusted tau correlation coefficient based on X is from Equation 31

$$yrx.c=\frac{0.258-0.652}{1-0.991}=-\frac{0.394}{0.909}=-0.433.$$

The corresponding variance estimate (Equation 34) is

$$\text{var}(yrx.c)=\frac{2(0.258+0.652-(0.258-0.652)^2)}{12(12-1)(1-0.091)^2}$$

$$=\frac{1.51}{109.032}=0.014$$

Finally from Equation 36, we obtain an estimate of ties correlated or adjusted Tau correlation coefficient between students grades in course 1 (x) and course 2 (Y) as

$$C\text{-}TAT=rxy.c=\frac{(0.288-0.652)+(0.258-0.652)}{2-0.061-0.091}$$

$$=-\frac{0.758}{1.848}=-0.410$$

The corresponding estimated variance is from Eqn 34

$$Var(rxy.c)=\frac{2(0.288+0.652-(0.288-0.652)2+0.258+0.652-(0.258-0.652)^2)}{12(12-1)(2-0.061-0.091)^2}$$

$$=\frac{2(1.563)}{450.78}=\frac{3.126}{450.78}=0.007$$

If interest is in testing the null hypothesis of Eqn 15 with $\rho_o = 0$ we have from Eqn 36 that.

$$\chi^2=\frac{(-0.410-0)^2}{0.007}=\frac{0.168}{0.007}=24.00\,(P\,value=0.0000)$$

Which with 1 degree of freedom is highly statistically significant indicating strong (and inverse) association between student performance in the two courses.

4. Conclusion

In this paper, we have proposed, developed and discussed a modified ties adjusted method for the estimation of Tau Correlation Coefficient here called C-TAT for ties adjusted tau Corrections Coefficient. C-TAT is developed and adjusted for the fact that often in the usual method used for the estimation of the regular Tau Correlations Coefficients; the value obtained depends on which of the two sample populations has its ranks arranged in their natural order and which has its assigned ranks tagged along. The value obtained also depends on whether or not there are ties within and between observations in the sampled populations. However, it is shown that proposed method gives essentially the same result as will be obtained with the usual Tau Correlation Coefficient when there are no ties and a better estimate of the Correction Coefficient when ties observations occurs in the samples. Furthermore, the proposed method is shown to be more robust than the parametric approach in that it can be used even when the usual assumption for the use of parametric methods are not satisfied by the data. More over, unlike the usual Tau Correlation Coefficient, the variance of the proposed C-TAT can be readily estimated.

However, the proposed statistics may be less powerful if the usual assumptions for the use of parametric methods are satisfied, one can easily fall back to the usual Tau Correlation Coefficient.

Finally, the proposed method is illustrated with some sample data and shown to yield more reliable Coefficient estimates of Tau Correction Coefficients than the usual method that is unadjusted and uncorrelated for these problems.

The value obtained also depends on whether or not there are ties within and between observations in the sampled populations.

The proposed method is illustrated with some sample data and shown to yield more reliable and efficient estimates of tau correlation coefficients than the usual method that is unadjusted and uncorrelated for these problems. The two methods yield the same estimates if and only if there are no tied observations what so ever in the sampled populations.

References

[1] Gibbion, J.D. (1973): Non parametric Statistical Inference "Mc Graw.Hills book Company, New York.

[2] Hollander, M. and Woife, D.A. (1999); Non parametric Statistical methods (2nd edition) Wiley-Inter Science, New York.

[3] Kendall, M. G. (1962); rank Correlation Methods Hafner Publishing Company, Inc., New York.

[4] Oyeka, C.A. etal (2009). A method of analyzing paired data intrinsically adjusted for tie Global Journal of maths and Statistics Vol.1, Pg 1-6.

[5] Oyeka, C.A. (1996). An introduction to applied Statistical Method , Nobern avocation Publishing Company, Enugu-Nigeria