Metodološki zvezki, Vol. 8, No. 1, 2011, 17-37

A Comparison of the Most Commonly Used Measures of Association for Doubly Ordered Square Contingency Tables via Simulation

Atila Göktaş[1] and Öznur İşçi[2]

Abstract

Spearman and Pearson correlation coefficient, Gamma coefficient, Kendall's tau-b, Kendall's tau-c, and Somers' d are the most commonly used measures of association for doubly ordered contingency tables. So far there has been no study expressing a priority on those measures of association. The aim of this study is to compare those measures of association for several types and different sample sizes of generated squared doubly ordered contingency tables and determine which measures of association are more efficient. It is found that both the sample sizes and the dimension of the doubly ordered contingency tables play a significant role on the effect of those measures of association.

1 Introduction

When categorical measures have a natural order (ex., strongly agree to strongly disagree; high, medium, low), additional information may be presented beside nominal variables. When there are two categorical variables that are both naturally ordered, a variety of effect size measures have been proposed for such ordinal data, including Gamma coefficient, Kendall's tau-b, Kendall's tau-c, and Somers' d (Garson, 2008).

An ordinal variable is also a type of a categorical variable. The only difference between the two is that there is a clear ordering of the ordinal variables, whereas there is no such ordering for ordinary categorical variables. For example, suppose you have a variable, patient's status, with three categories (worse, no difference and much better). In addition to being able to classify patients into these three categories, you can order the categories as worse, no difference and much better. Now think of a variable like educational background (with levels

1 University of Mugla, Faculty of Sciences, Department of Statistics, Mugla, Turkey; gatilla@mu.edu.tr

2 University of Mugla, Faculty of Sciences, Department of Statistics, Mugla, Turkey; oznur.isci@mu.edu.tr

such as elementary school graduate, high school graduate, some college and university graduate). These also can be ordered as elementary school, high school, some college, and university graduate.

Even though the levels are ordered from lowest to highest, the distance between the levels need not to be the same across the levels of the variables. Suppose we assign scores for the levels of educational experience as 1, 2, 3 and 4 respectively and we compare the difference in education between levels one and two with the difference in educational experience between levels two and three, or the difference between levels three and four. The difference between levels one and two (elementary and high school) is perhaps much larger than the difference between categories two and three (high school and some college). In this example, we can order the people in level of educational experience but the size of the difference between levels is inconsistent (because the distance between levels one and two is larger than levels two and three) i.e the level of measuring is ordinal not interval (Ucla, 2007).

A doubly ordered categorical data or doubly ordered contingency tables are data with two variables that are both naturally ordered and cross tabulated. The most commonly and widely used measures of association for doubly ordered categorical data are measures of differences between probabilities of concordant and discordant pairs. Examples of these are Kendall's tau-b, Stuart's tau-c, Goodman-Kruskal's gamma, and Somers'd (Svensson, 2000). The difference among these measures lies in the power of overcoming of ties. One of the most well known non-parametric measures of association is called the Spearman rank-correlation ρ_s. Another famous measure of association is Kendall's tau which may be formulated as a Pearson product-moment correlation between signed indicators of X's and Y's, and Spearman's rank-correlation is the special case of the Pearson product-moment using the ranks instead of the actual variates the correlation with (Kruskal, 1958 and Hoeffding, 1948).

Kendall's tau which does not need to specify the ranking scores for both row and column and Somers' d coefficients are alternatives to Pearson's product-moment correlation coefficient and Spearman's rank-order correlation coefficient for ordinal data (Cyrus and Nitin, 1995).

2 The most commonly used measures of association

Spearman's rank-order correlation coefficient and Pearson's product-moment correlation coefficient, Goodman-Kruskal's gamma coefficient, Kendall's tau-b, Kendall's tau-c, and Somers' d are the most commonly used measures of association for doubly ordered contingency tables. This study was performed for the square doubly ordered contingency tables. What square term actually means is that the number of row categories equals to the number of column categories.

Notation

The following notations are used throughout this study:

X_i	Row variable arranged in ascending order: $X_1 < X_2 < \ldots < X_R$
Y_j	Column variable arranged in ascending order: $Y_1 < Y_2 < \ldots < Y_C$
f_{ij}	Frequency in row category i and column category j
c_j	$\sum_{i=1}^{R} f_{ij}$ - the subtotal of *j*-th column
r_i	$\sum_{j=1}^{C} f_{ij}$ - the subtotal of *i*-th row
W	$\sum_{j=1}^{C} c_j = \sum_{i=1}^{R} r_i$ - the general total of the sample size

2.1 Pearson correlation coefficient

Pearson Product Moment Correlation is the most widely and common used measure of correlation also called Pearson's correlation for short. The Pearson Product Moment correlation is represented by the Greek letter ρ (rho) when calculated from a population, whereas it is represented by the letter "r" if it is computed from a sample that is sometimes called "Pearson's r". Pearson's correlation reflects the degree of linear relationship between two variables. It varies from -1 to +1. A positive correlation means that as X and Y increases in the same direction. A correlation of +1 means that there is a perfect positive linear relationship between variables that is the degree of increment in X is proportional to the degree of increment of Y (Lane, 1997). A reverse explanation may be given for a -1 correlation.

Some assumptions are required and given below for the calculation of Pearson's product moment correlation r:

- Significant linear relationship between X and Y variables
- X and Y are continuous random variables
- Both variables must be normally distributed

There is a relationship between simple linear regression and Pearson’s correlation coefficient. The main difference is that the variables used for the calculatiaons are treated as response and explanatory for simple linear regression whereas there is no such discriminaton for the Pearson’s correlation. The square of r is called the goodness of fit or coefficient of determination and denotes the portion of total variance explained by the simple linear regression model.

The formula of Pearson's product moment correlation r may be given as in the definition (2.1),

$$\mathbf{r = \frac{cov(X,Y)}{\sqrt{S(X)S(Y)}} \equiv \frac{S}{T}} \tag{2.1}$$

where $\mathbf{cov(X,Y)}$ which is given below in equation (2.2) is also called the covariance of X and Y

$$\mathbf{cov(X,Y) = \sum_{i,j} X_i Y_j f_{ij} - \left(\sum_{i=1}^{R} X_i r_i\right)\left(\sum_{j=1}^{C} Y_j c_j\right) / W} \tag{2.2}$$

$\mathbf{S(X)}$ which is presented in equation (2.3) is also called the variance of X

$$\mathbf{S(X) = \sum_{i=1}^{R} X_i^2 r_i - \left(\sum_{i=1}^{R} X_i r_i\right)^2 / W} \tag{2.3}$$

and $\mathbf{S(Y)}$ in (2.4) is the variance of Y

$$\mathbf{S(Y) = \sum_{j=1}^{C} Y_j^2 c_j - \left(\sum_{j=1}^{C} Y_j c_j\right)^2 / W} \tag{2.4}$$

The variance of *r* is

$$\mathbf{var_1 = \frac{1}{T^4} \sum_{i,j} f_{ij} \left\{ T(X_i - \bar{X})(Y_j - \bar{Y}) - \frac{S}{2T}[(X_i - \bar{X})^2 S(Y) + (Y_j - \bar{Y})^2 S(X)] \right\}^2} \tag{2.5}$$

If the null hypothesis which is "$\mathbf{H_0 : \rho = 0}$" against the alternative hypothesis which is either "$\mathbf{H_1 : \rho \neq 0}$" or "$\mathbf{H_1 : \rho > 0}$" or "$\mathbf{H_1 : \rho < 0}$" is true, the variance of r may be presented as in (2.6),

$$\mathbf{var_0 = \frac{\sum_{i,j} f_{ij} X_i^2 Y_j^2 - \left(\sum_{i,j} f_{ij} X_i Y_j\right)^2 / W}{\left(\sum_i r_i X_i^2\right)\left(\sum_i c_j Y_j^2\right)}} \tag{2.6}$$

where $\mathbf{\bar{X} = \sum_{i=1}^{R} X_i r_i / W}$ and $\mathbf{\bar{Y} = \sum_{j=1}^{C} Y_j c_j / W}$ are the mean of X and the mean of Y respectively. Under the null hypothesis that there is no correlation,

$$\mathbf{t_{calculated} = \frac{r\sqrt{W-2}}{\sqrt{1-r^2}}} \tag{2.7}$$

statistics has a t distribution with W - 2 degrees of freedom.

2.2 Spearman rank correlation coefficient

Calculating the Pearson's correlation coefficient needs the assumption that the two samples are normally distributed. If the assumption of normality is violated, Pearson's correlation coefficient will produce unreliable results. Hence a very best alternative for Pearson's correlation coefficient may be the use of Spearman's rank correlation r_s which can be calculated under the first assumption of Pearson's product moment correlation (Lohninger, 1999). There is no need of satisfaction of the second and third assumptions of the Pearson's product moment correlations for the use of Spearman rank correlation. Dependency of the ordinal variables is denoted as a rank correlation and their intensity is expressed by correlation coefficients. One of the most used ordinal coefficients is Spearman's correlation coefficient (Rezankova, 2009). The Spearman's rank correlation coefficient r_s is computed by using rank scores R_i for X_i and rank scores C_j for Y_j. These rank scores are defined as follows:

$$\mathbf{R_i = \sum_{k<i} r_k + (r_i + 1)/2} \text{ for i = 1, 2, …, R} \tag{2.8}$$

$$\mathbf{C_j = \sum_{h<j} c_h + (c_j + 1)/2} \quad \text{for j = 1, 2, …, C} \tag{2.9}$$

The formulas for r_s can be obtained from the Pearson formula given in (2.1) by substituting R_i and C_j for X_i and Y_j, respectively. And its asymptotic variance of the Spearman correlation can be obtained under the null hypothesis of no correlation from the formula presented in (2.6) by substituting R_i and C_j for X_i and Y_j, respectively.

$$\mathbf{r_S = \frac{cov(R,C)}{\sqrt{S(R)S(C)}} \equiv \frac{S}{T}} \tag{2.10}$$

If there are no ties, another simple formula for obtaining Spearman's rank correlation is given in (2.11) as follows:

$$\mathbf{r_S = 1 - \frac{6\sum d_i^2}{W(W^2-1)}} \tag{2.11}$$

Where d_i in Spearman's rank correlation coefficient represents the difference in the ranks assigned to the values of the variable for each item of the certain data. When W is fairly small, the computation of the formula is very straightforward. In case of numerically equal observations an arithmetic average of the rank numbers associated with the ties are assigned to the values of the variables. This formula of Spearman's rank correlation coefficient is applied in cases when there are no tied ranks. When there are tied ranks the formula in (2.11) is not algebraically equivalent to the formula in (2.10). However, when there are a reasonable number of ties in the pairs of values of the variables, this approximation of Spearman's rank correlation coefficient is often used as fairly good approximations.

The Spearman's rank correlation coefficient may be used to test for association between both ordinals and continues variables. The underlying relationship between variables must be monotonic. In other words, generally speaking, the variables should either increase in values together, or when one gets increased, and then the other should get decreased.

Some difficulties of calculating Spearman's rank correlation coefficient arise, when the sample is large. For large data it can be hard to rank the data for both variables and consequently it is time consuming to perform Spearman's rank correlation coefficient test.

Since Spearman's rank correlation coefficient is a non parametric test, it does not depend upon the assumptions given for the Pearson's product moment correlation coefficient. Hence it is distribution free. It can be used to test whether there is a statistically significant association between variables. The null hypothesis we are testing is that there is no association between the variables under study. Thus, the main purpose of Spearman's rank correlation coefficient is to investigate the existence of any association in the underlying variables. To this end, the null hypothesis is constructed as having no rank correlation between the variables while using Spearman's rank correlation coefficient. Under the null hypothesis that there is no correlation,

$$\mathbf{t_{calculated}} = \frac{\mathbf{r_S}\sqrt{\mathbf{W-2}}}{\sqrt{\mathbf{1-r_s^2}}} \tag{2.12}$$

statistics has a t distribution with W - 2 degrees of freedom (Kendall and Stuart, 1973).

2.3 Goodman and Kruskal gamma (γ or G)

The Gamma (γ) statistics is proposed in a series of papers from 1954 to 1972 by Leo Goodman and William Kruskal. It is now mostly described just as Gamma that is used to investigate an association in a given doubly ordered contingency table.

The estimator of gamma uses only the number of concordant and discordant pairs of observations. It ignores tied pairs. In other words, pairs of observations that have equal values of *X* and equal values of *Y* are called tied pairs. Gamma can be calculated for only when both variables lie on an ordinal scale. It has the range $-1 \leq \gamma \leq 1$ just as Spearman's rank correlation coefficient. If there is no association between the two variables, then the estimator of gamma should be close to zero. The estimation of Gamma (γ) may be given as follows:

$$\gamma = \frac{\mathbf{P} - \mathbf{Q}}{\mathbf{P} + \mathbf{Q}} \tag{2.13}$$

where P has the form as $\mathbf{P} = \sum_{i,j} \mathbf{f_{ij}} \mathbf{C_{ij}}$ and it is the probability that a randomly selected pair of observations will place in the same order and Q has the form as $\mathbf{Q} = \sum_{i,j} \mathbf{f_{ij}} \mathbf{D_{ij}}$ and it is the probability that a randomly selected pair of observations will place in the opposite order, where $\mathbf{f_{ij}}$ is the frequency of *i*-th row and *j*-th column of the doubly order contingency table, C_{ij} is $\sum_{k>i} \sum_{l>j} \mathbf{f_{kl}} + \sum_{k<i} \sum_{l<j} \mathbf{f_{kl}}$ and D_{ij} is $\sum_{k>i} \sum_{l<j} \mathbf{f_{kl}} + \sum_{k<i} \sum_{l>j} \mathbf{f_{kl}}$. Its general standard error may be given as follows:

$$\mathbf{ASE_1} = \frac{\mathbf{4}}{\mathbf{(P+Q)^2}} \sqrt{\sum_{i,j} \mathbf{f_{ij}} \mathbf{(QC_{ij} - PD_{ij})^2}} \tag{2.14}$$

Under the null hypothesis of independence or no association, its standard error becomes as follows:

$$\mathbf{ASE_0} = \frac{\mathbf{2}}{\mathbf{(P+Q)}} \sqrt{\sum_{i,j} \mathbf{f_{ij}} \mathbf{(C_{ij} - D_{ij})^2} - \frac{\mathbf{1}}{\mathbf{W}} \mathbf{(P-Q)^2}} \tag{2.15}$$

For 2×2 tables, gamma is equivalent to Yule's *Q* which may be presented as follows (Goodman and Kruskal, 1979; Agresti, 2010; Brown and Benedetti, 1977b);

$$\mathbf{Q} = \frac{\mathbf{f_{11} f_{22} - f_{12} f_{21}}}{\mathbf{f_{11} f_{22} + f_{12} f_{21}}} \tag{2.16}$$

Gamma coefficient can also be calculated for even small or perhaps for zero frequency of a 2x2 table.

Suppose that we have a value of gamma to be .582. It can be inferred that knowing the independent variable reduces our errors in predicting the rank (not

value) of the dependent variable by 58.2%. Under statistical independence, gamma will be zero, but there are some other times in which gamma coefficient may be zero whenever the number of concordant equal to the number of discordant. Meanwhile, using gamma coefficient a perfect association is present whenever the number of discordant pairs is zero. Under the null hypothesis that there is no correlation,

$$\mathbf{Z_{calculated}} = \frac{\hat{\gamma}}{\mathbf{ASE_0}} \tag{2.17}$$

statistics has standard normal distribution.

2.4 Kendall's Tau-b

Kendall's tau-b (τ_b) is similar to gamma except that tau-b uses a correction for ties. The rule of both variables lie on an ordinal scale for calculation Tau-b is just the same as gamma coefficient. Tau-b has also the range $-1 \leq \tau_b \leq 1$ as both gamma and Spearman's rank correlation. It is estimated by,

$$\tau_b = \frac{P - Q}{\sqrt{D_r D_c}} \tag{2.18}$$

where D_r stands for $W^2 - \sum_{j=1}^{C} r_j^2$ and r_j is the total count or the total frequency of row *i* in the doubly ordered cross table, D_c stands for $W^2 - \sum_{j=1}^{C} c_j^2$ and c_j is the total count or the total frequency of column *j* in the doubly ordered cross table. Its general standard error may be obtained as follows:

$$ASE_1 = \frac{1}{(D_r D_c)} \sqrt{\sum_{i,j} f_{ij}(2\sqrt{D_r D_c}(C_{ij} - D_{ij}) + \tau_b v_{ij})^2 - W^3 \tau_b^2 (D_r + D_c)^2} \tag{2.19}$$

where v_{ij} is defined as $r_i D_r + c_j D_c$. Under the null hypothesis of independence or no association, the standard error takes its form as follows:

$$ASE_0 = 2\sqrt{\frac{\sum_{i,j} f_{ij}(C_{ij} - D_{ij})^2 - \frac{1}{W}(P - Q)^2}{D_r D_c}} \tag{2.20}$$

and under the null hypothesis of independence the asymptotic test statistics has standard normal distribution which is given as,

$$\mathbf{Z_{calculated} = \frac{\tau_b}{ASE_0}} \tag{2.21}$$

The test statistics given in (2.21) is used to test whether the degree of association of the cross tabulations when both variables are measured in ordinal scale is significant (Kendall, 1955; Brown and Benedetti, 1977a; SAS, 2010).

It adjusts the ties and is most appropriate for square tables what means that the number of row categories equals to the number of column categories. Value of −1 is 100% negative association or perfect inversion whereas value of +1 is 100% positive association, or perfect agreement. A value of zero indicates no association.

If $\tau_b = \pm 1$ then there is no ties and subjects from different cells form strict concordant and discordant pairs in these two extreme cases. When both $\tau_b = \pm 1$ and $\gamma = \pm 1$, it is generally concluded that τ_b is stronger than γ. If $\tau_b = 1$, then the table is diagonal and if $\tau_b = -1$, the table is skewed diagonal (Tu, 2007).

2.5 Kendall's Tau-c

Stuart's tau-c (τ_c) makes an adjustment for table size as well as a correction for ties. Tau-c is also appropriate only when both variables lie on an ordinal scale. Tau-c has the range $-1 \leq \tau_c \leq 1$ as well as Spearman's rank correlation, Gamma and Tau-b. It is estimated by

$$\tau_c = \frac{q(P-Q)}{W^2(q-1)} \tag{2.22}$$

where q is defined as min(R,C). Its general standard error may be written as follows:

$$ASE_1 = \frac{2q}{(q-1)W^2}\sqrt{\sum_{i,j} f_{ij}(C_{ij}-D_{ij})^2 - \frac{1}{W}(P-Q)^2} \tag{2.23}$$

Under the null hypothesis of no association ASE_1 is identical to ASE_0. Therefore the test statistics which may be used to investigate the degree of association for two ordinal variables under the null hypothesis of no association can be expressed as

$$\mathbf{Z_{calculated}} = \frac{\tau_c}{\mathbf{ASE_0}} \tag{2.24}$$

where $\mathbf{Z_{calculated}}$ statistics has standard normal distribution. Besides making adjustments for ties it is most suitable for rectangular tables. Value of −1 is 100% negative association or perfect inversion whereas value of +1 is 100% positive association, or perfect agreement. A value of zero indicates no association (Brown and Benedetti, 1977a; SAS, 2010).

Kendall's tau-c, also called Stuart's tau-c or Kendall-Stuart tau-c, is a special case of tau-b for larger tables. It also makes adjustments for the size of the cross table (Lohninger, 1999).

2.6 Somers' d

Somers' d(*C*|*R*) and Somers' *d*(*R*|*C*) are asymmetric modifications of tau-b. *C*|*R* represents that the row variable X is treated as an independent variable, whereas the column variable Y is treated as dependent. Similarly, *R*|*C* represents the reverse interpretation. Somers'd differ from tau-b in that it only makes a correction for tied pairs on the independent variable. Somers' d can be calculated only when both variables are ordered. It varies in the range $-1 \leq d \leq 1$. Formulas for Somers' d is obtained according to the position of independent variable. For instance, if the row variable X is treated to be independent then Somers' d can be calculated as

$$\mathbf{d_{Y/X}} = \frac{\mathbf{P - Q}}{\mathbf{D_r}} \tag{2.25}$$

and its general standard error is defined as below:

$$\mathbf{ASE_1} = \frac{2}{\mathbf{D_r^2}} \sqrt{\sum_{i,j} \mathbf{f_{ij}} \left\{ \mathbf{D_r (C_{ij} - D_{ij}) - (P - Q)(W - R_i)} \right\}^2} \tag{2.26}$$

or, under the null hypothesis of independence its standard error may be written as:

$$\mathbf{ASE_0} = \frac{2}{\mathbf{D_r}} \sqrt{\sum_{i,j} \mathbf{f_{ij}} (\mathbf{C_{ij} - D_{ij}})^2 - \frac{1}{\mathbf{W}} (\mathbf{P - Q})^2} \tag{2.27}$$

by interchanging the roles of X and Y, the formulas for Somers' d with X as the dependent variable can be obtained with only a minor change in the denominator by replacing $\mathbf{D_r}$ with $\mathbf{D_c}$.

If both variables are ignored to be either independent or dependent, symmetric version of Somers' d is appropriate and it is calculated as follows:

$$\mathbf{d_{symetric}} = \frac{\mathbf{(P-Q)}}{\frac{\mathbf{1}}{\mathbf{2}}\mathbf{(D_c + D_r)}} \tag{2.28}$$

and its standard error is simplified as follows:

$$\mathbf{ASE_1} = \frac{\mathbf{2\sigma^2_{\tau_b}}}{\mathbf{(D_r + D_c)}}\sqrt{\mathbf{D_r D_c}} \tag{2.29}$$

where $\sigma^2_{\tau_b}$ is the variance of Kendall's τ_b. Under the null hypothesis of no association its standard error may be obtained as follows:

$$\mathbf{ASE_0} = \frac{\mathbf{4}}{\mathbf{(D_c + D_r)}}\sqrt{\sum_{i,j}\mathbf{f_{ij}(C_{ij} - D_{ij})^2} - \frac{\mathbf{1}}{\mathbf{W}}\mathbf{(P-Q)^2}} \tag{2.30}$$

Somers' d value of −1 is 100% negative association or perfect inversion whereas value of +1 is 100% positive association, or perfect agreement (Somers, 1962; Goodman and Kruskal, 1963; Liebetrau, 1983; SAS, 2010).

A value of zero indicates no association. Under the null hypothesis of independence, the following statistics asymptotically has standard normal distribution

$$\mathbf{Z_{calculated}} = \frac{\mathbf{d_{symetric}}}{\mathbf{ASE_0}} \tag{2.31}$$

3 Generation of doubly ordered contingency table

In order to generate a doubly ordered contingency table, there are lots of techniques in the literature of Statistical simulation. For instance, a doubly ordered contingency table may be generated from the uniform association model (Agresti, 2010). In our study we present a new way of generating a doubly ordered contingency table using bivariate standard normal distribution. In the first step we generate two identically independently distributed random variables, as $\mathbf{X_1} \sim \mathbf{N(0,1)}$ and $\mathbf{X_2} \sim \mathbf{N(0,1)}$. To generate two random variables (X and Y) from the bivariate normal distribution with certain correlation (ρ) for a specific sample size, we apply the followings:

$$\mathbf{X} = \mathbf{aX_1} + \mathbf{bX_2} \tag{3.1}$$

$$\mathbf{Y} = \mathbf{bX_1} + \mathbf{aX_2} \tag{3.2}$$

where $\mathbf{a^2 + b^2 = 1}$ and $\mathbf{2ab = \rho}$, and hence a and b are obtained as $\mathbf{a = \frac{\sqrt{1+\rho} + \sqrt{1-\rho}}{2}}$ and $\mathbf{b = \frac{\sqrt{1+\rho} - \sqrt{1-\rho}}{2}}$.

To generate two random variables for certain correlation from the bivariate normal distribution, a and b are calculated and presented in Table 1.

Table 1: For specific correlations the values of a and b.

ρ	**a**	**b**
0	1	0
0.5	0.9659258263	0.25881190451
0.9	0.8473163206	0.5310885546

If we would like to generate a doubly ordered contingency table for a certain number of rows R and certain number of column C, say RxC table, we split the range of generated data for X variable into R sub equal intervals and for Y variable into C sub equal intervals. And then we recode the variables into new variables according to the sub equal intervals. How we do that is quite simple. The recoding is performed for instance if a datum falls into the first interval then its recode value is 1, for general if it falls into *i*-th interval then its recode value is *i* and so on. An example of generating 4x4 doubly ordered contingency table for 100 sample size when there is no correlation has been given step by step below.

Table 2 presents both the generated data from the uncorrelated bivariate standard normal distribution for the sample size 100 and the recoded new variables according to the subintervals presented in Table 3.

Table 3 presents the subintervals of each variable and their code values. For instance the range of the generated data from X is -2.69058 for lower bound and 2.97257 for upper bound. This range has been split into four equal subintervals as (-2.69058;-1.27479) for the first subinterval and (-1.27479;0.14100) for the second subinterval and so on.

Table 4 presents the generated doubly ordered contingency table that is obtained by cross tabulating the X coded and Y coded variable presented in Table 3.

Table 2: Uncorrelated X and Y from the bivariate normal distribution with their codes.

NO	X	X coded	Y	Y coded	NO	X	X coded	Y	Y coded
1	-0.2536	2	-0.7163	2	**51**	1.0623	3	0.0464	3
2	-1.0397	2	-0.2695	2	**52**	1.5019	3	-0.3173	2
3	-1.0620	2	-0.7119	2	**53**	1.0088	3	0.4799	3
4	-0.1434	2	-0.2043	2	**54**	-0.6655	2	-1.5755	1
5	0.5554	3	-0.6441	2	**55**	0.4410	3	-0.5301	2
6	-0.9337	2	0.5972	3	**56**	-0.2411	2	-0.2260	2
7	-0.6532	2	-1.3105	1	**57**	-0.6315	2	-1.6726	1
8	-0.0312	2	-0.5532	2	**58**	1.2348	3	-0.0832	2
9	0.1456	3	-0.7766	2	**59**	2.9726	4	-1.1554	2
10	-0.1454	2	-1.7343	1	**60**	-0.2394	2	-0.0542	2
11	-0.5058	2	0.6856	3	**61**	-0.1062	2	-0.1400	2
12	-1.2511	2	1.1864	3	**62**	0.9309	3	0.0107	3
13	-0.5854	2	-0.4971	2	**63**	0.2709	3	-0.2638	2
14	0.9921	3	0.1856	3	**64**	-0.5009	2	0.5375	3
15	0.4573	3	1.0416	3	**65**	0.7381	3	-1.2461	1
16	-0.8605	2	-1.3636	1	**66**	0.0861	2	-0.1343	2
17	0.3545	3	1.0972	3	**67**	-0.2575	2	-1.3048	1
18	0.4592	3	-0.7049	2	**68**	-0.1921	2	-0.0969	2
19	0.0779	2	0.1284	3	**69**	-0.9413	2	1.6775	4
20	-1.2302	2	0.1972	3	**70**	0.8649	3	1.5616	4
21	-1.6566	1	-0.6006	2	**71**	-0.3182	2	0.1286	3
22	-0.3763	2	2.3653	4	**72**	2.1642	4	-1.5743	1
23	0.1518	3	0.4474	3	**73**	1.4203	3	-1.3141	1
24	2.8291	4	0.1628	3	**74**	-0.8289	2	-2.4796	1
25	-0.2974	2	0.3574	3	**75**	-1.7606	1	0.8185	3
26	-1.7357	1	0.2520	3	**76**	0.7911	3	-0.6351	2
27	-1.3409	1	-1.2586	1	**77**	-0.9899	2	0.7008	3
28	0.0192	2	1.2798	4	**78**	-0.3139	2	-0.7316	2
29	-2.6906	1	0.5039	3	**79**	1.5227	3	0.1013	3
30	1.0547	3	1.3173	4	**80**	-1.5821	1	1.2279	3
31	-0.5241	2	-1.1634	2	**81**	-1.8928	1	0.2019	3
32	-0.1484	2	0.0275	3	**82**	-0.9889	2	0.4336	3
33	-0.3196	2	0.1960	3	**83**	0.5171	3	-0.3009	2
34	1.7460	4	0.1461	3	**84**	1.3806	3	0.4284	3
35	1.2623	3	0.5115	3	**85**	0.7285	3	-0.6359	2
36	0.0753	2	-1.8280	1	**86**	0.4379	3	-1.4755	1
37	1.0775	3	-0.9509	2	**87**	0.2492	3	-0.6649	2
38	-0.5648	2	0.5449	3	**88**	0.1890	3	-0.7737	2
39	-0.8080	2	1.0685	3	**89**	1.7847	4	0.4966	3
40	-1.4783	1	0.1946	3	**90**	-0.3535	2	-0.6291	2
41	-0.0221	2	0.4106	3	**91**	0.1400	2	-1.4845	1
42	-1.2005	2	-0.5804	2	**92**	-1.8057	1	2.4985	3
43	-0.1499	2	-0.0788	2	**93**	-1.0551	2	-0.4196	2
44	0.0909	2	-0.2291	2	**94**	-1.6230	1	0.0321	3
45	0.8470	3	0.4373	3	**95**	0.4015	3	-1.4419	1
46	-1.3933	1	0.2019	3	**96**	0.0687	2	-0.8363	2
47	-0.7364	2	-1.4237	1	**97**	-0.3731	2	-1.6131	1
48	0.2612	3	1.2662	4	**98**	0.2875	3	-2.0094	1
49	-1.2693	2	-0.4225	2	**99**	-0.3149	2	0.7019	3
50	0.7355	3	0.1085	3	**100**	0.6373	3	-0.8100	2

Table 3: The sub intervals of each variable and their code values.

	X				Y		
	Lover Bound	**Upper Bound**	**Code**		**Lover Bound**	**Upper Bound**	**Code**
Interval 1	-2.69058	-1.27479	**1**	**Interval 1**	-2.47961	-1.23508	**1**
Interval 2	-1.27479	0.14100	**2**	**Interval 2**	-1.23508	0.00946	**2**
Interval 3	0.14100	1.55679	**3**	**Interval 3**	0.00946	1.25399	**3**
Interval 4	1.55679	2.97257	**4**	**Interval 4**	1.25399	2.49853	**4**

Table 4: The generated 4x4 doubly ordered contingency table.

		Y				
		1	**2**	**3**	**4**	**Total**
X	**1**	1	1	9	0	**11**
	2	11	20	16	3	**50**
	3	5	14	12	3	**34**
	4	1	1	3	0	**5**
	Total	**18**	**36**	**40**	**6**	**100**

4 Simulation study

The simulation work has been designed in terms of sample size, table dimension and degree of association. For a fair degree of ordinal association, the correlation between variables generated has been declared as 0.5 and for a strong degree of ordinal association it is declared to be 0.9. Seven different square table dimensions which are 3x3, 4x4, 5x5, 6x6, 7x7, 8x8, 9x9 have been generated for each correlation. Also for each correlation and table dimension, eight different sample sizes which are 50, 100, 150, 200, 250, 500, 750, 1000 are studied. For each correlation, table dimension and sample size, the process has been repeated 10000 times. The comparisons have been made according to the mean of 10000 replications of the degree of ordinal measure of associations.

Since it is not easy to judge the results recorded in tables, it is decided to perform line plots. Therefore the results obtained are presented in line plots to clarify the effect of both the sample size and the table dimension. Actually the results are presented in two types of line plots. The first type has been performed to investigate the effect of table dimension, whereas the second has been performed to investigate the effect of sample size. For instance the line plots are presented in Figure 1 (ρ=0.5) and Figure 2 (ρ=0.9) give an idea of how the table dimension affects the result of each of the expected mean of ordinal measure of association. Figure 3 (ρ=0.5) and Figure 4 (ρ=0.9) give an idea of how the sample size affects the result of each of the expected mean of ordinal measure of association.

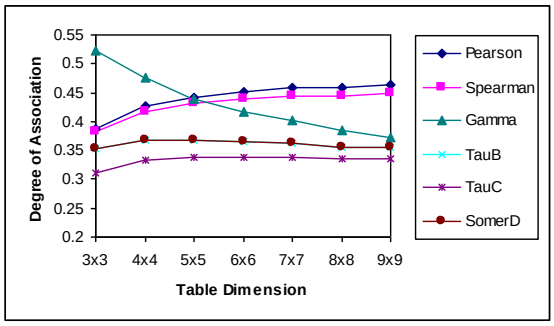

Figure 1a: Table dimension against degree of the ordinal measure of associations for ρ =0.5 and n=50.

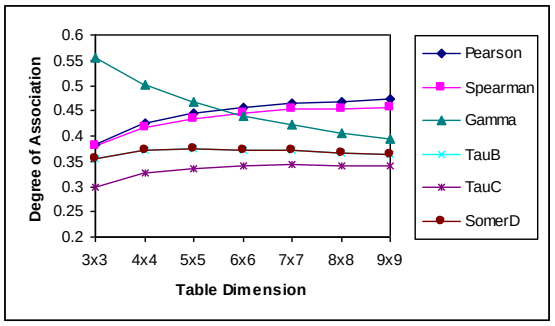

Figure 1b: Table dimension against degree of the ordinal measure of associations for ρ =0.5 and n=100.

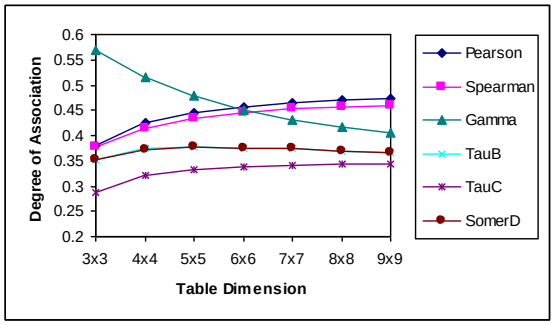

Figure 1c: Table dimension against degree of the ordinal measure of associations for ρ =0.5 and n=150.

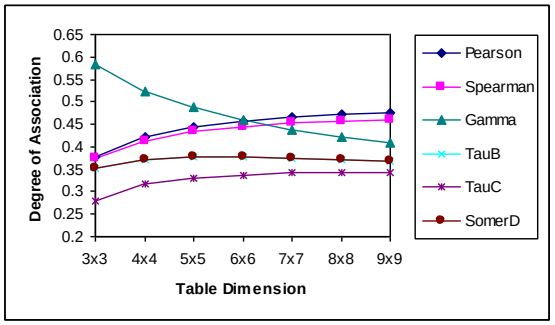

Figure 1d: Table dimension against degree of the ordinal measure of associations for ρ =0.5 and n=200.

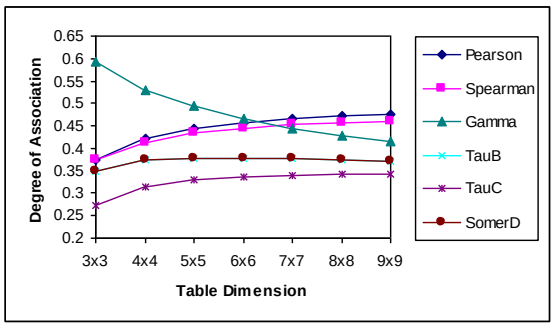

Figure 1e: Table dimension against degree of the ordinal measure of associations for ρ =0.5 and n=250.

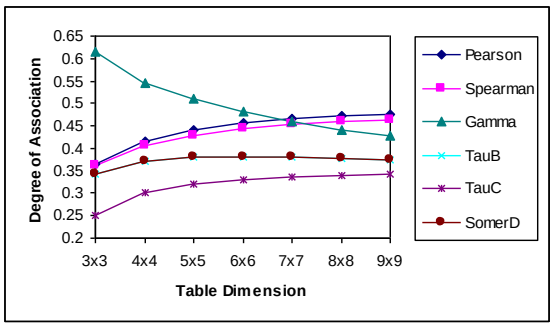

Figure 1f: Table dimension against degree of the ordinal measure of associations for ρ =0.5 and n=500.

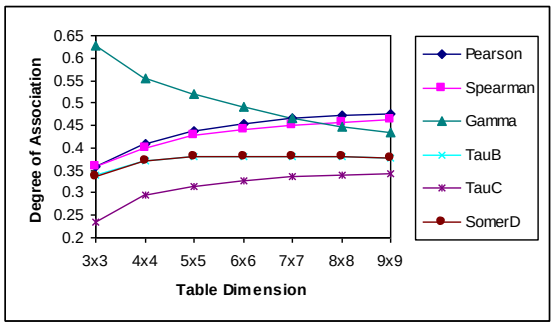

Figure 1g: Table dimension against degree of the ordinal measure of associations for ρ =0.5 and n=750.

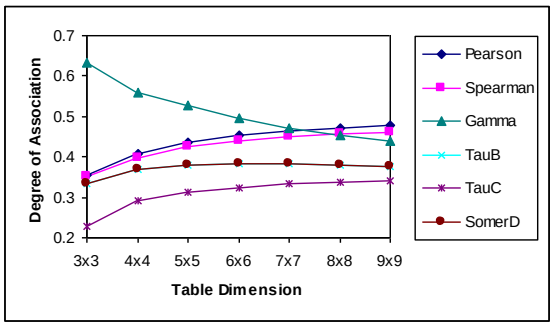

Figure 1h: Table dimension against degree of the ordinal measure of associations for ρ =0.5 and n=1000.

Figure 1: Table dimension against the mean of the ordinal measure of associations ρ =0.5 and sample size n=50, 100, 150, 200, 250, 500, 750, 1000.

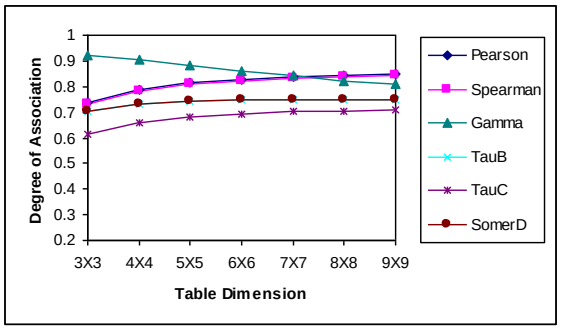

Figure 2a: Table dimension against degree of the ordinal measure of associations for ρ =0.9 and n=50.

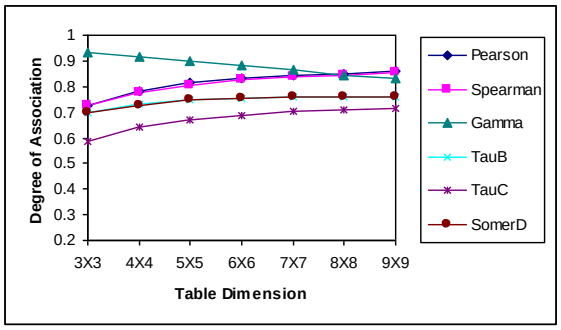

Figure 2b: Table dimension against degree of the ordinal measure of associations for ρ =0.9 and n=100.

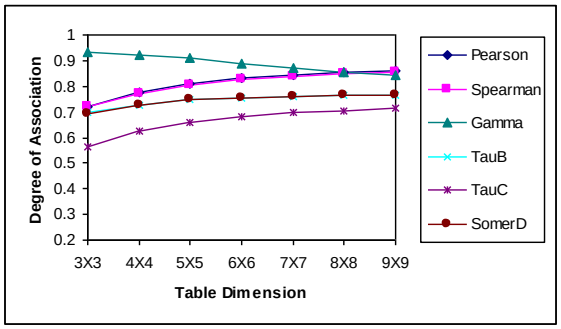

Figure 2c: Table dimension against degree of the ordinal measure of associations for ρ =0.9 and n=150.

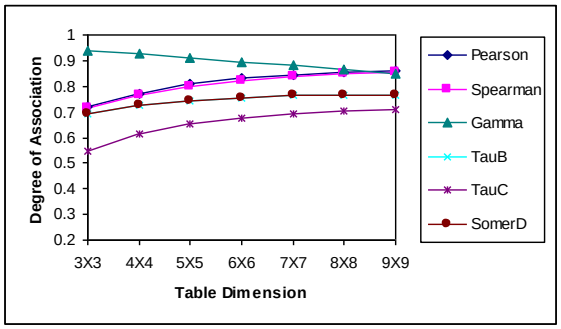

Figure 2d: Table dimension against degree of the ordinal measure of associations for ρ =0.9 and n=200.

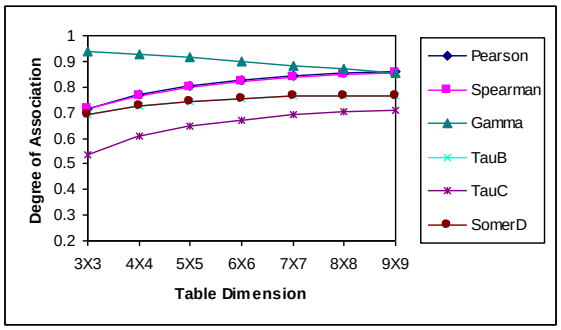

Figure 2e: Table dimension against degree of the ordinal measure of associations for ρ =0.9 and n=250.

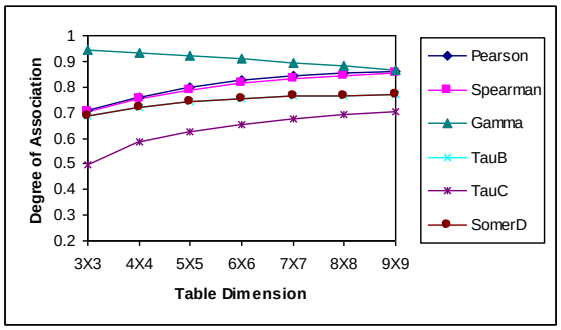

Figure 2f: Table dimension against degree of the ordinal measure of associations for ρ =0.9 and n=500.

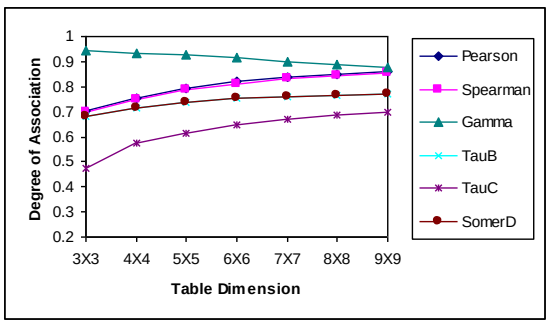

Figure 2g: Table dimension against degree of the ordinal measure of associations for ρ =0.9 and n=750.

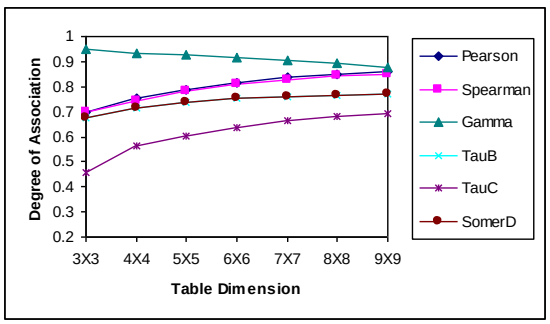

Figure 2h: Table dimension against degree of the ordinal measure of associations for ρ =0.9 and n=1000.

Figure 2: Table dimension against degree of the ordinal measure of associations for ρ =0.9 and n=50, 100, 150, 200, 250, 500, 750, 1000.

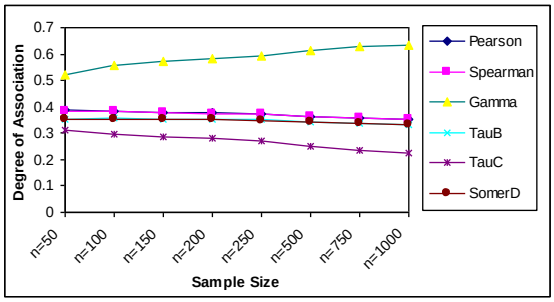

Figure 3a: Sample size against degree of the ordinal measure of associations for ρ =0.5 and 3x3 Figure.

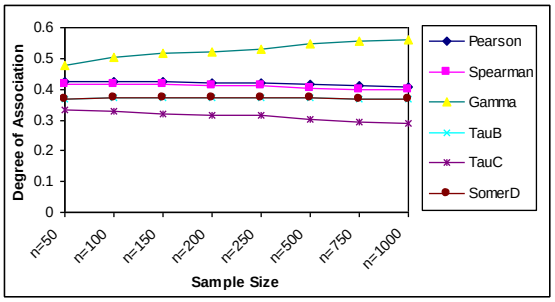

Figure 3b: Sample size against degree of the ordinal measure of associations for ρ =0.5 and 4x4 Figure.

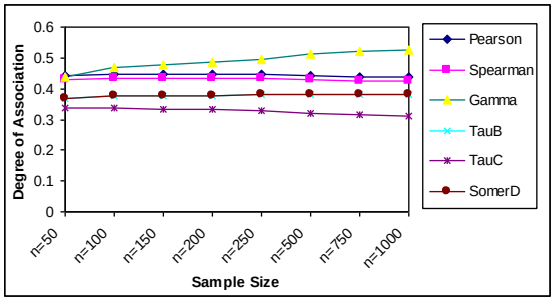

Figure 3c: Sample size against degree of the ordinal measure of associations for ρ =0.5 and 5x5 Figure.

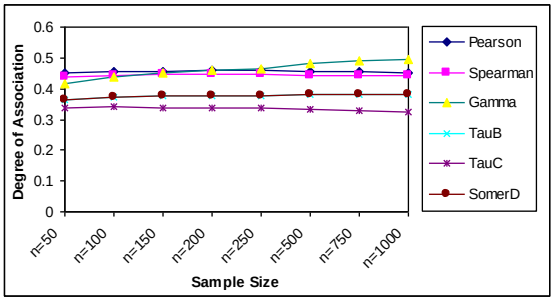

Figure 3d: Sample size against degree of the ordinal measure of associations for ρ =0.5 and 6x6 Figure.

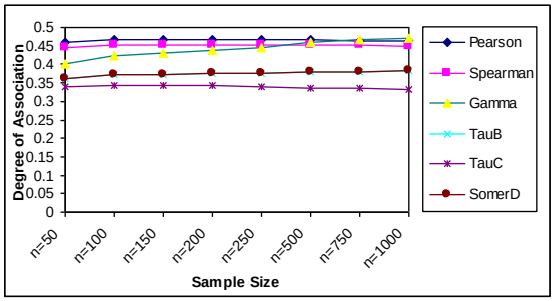

Figure 3e: Sample size against degree of the ordinal measure of associations for ρ =0.5 and 7x7 Figure.

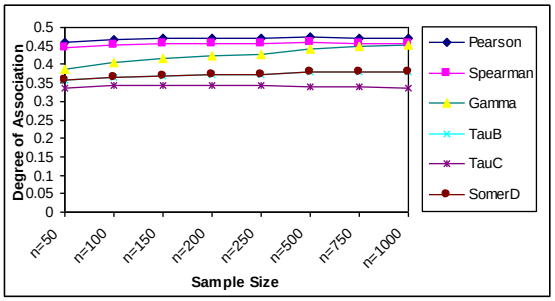

Figure 3f: Sample size against degree of the ordinal measure of associations for ρ =0.5 and 8x8 Figure.

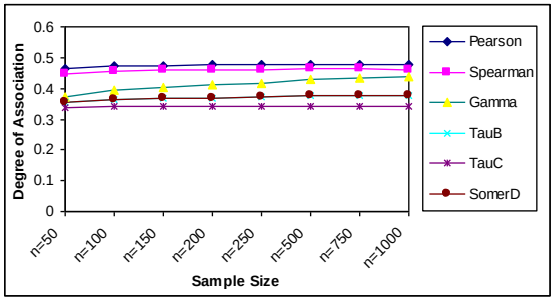

Figure 3g: Sample size against degree of the ordinal measure of associations for ρ =0.5 and 9x9 Figure.

Figure 3: Sample size against degree of the ordinal measure of associations for ρ =0.5 and 3x3, 4x4, 5x5, 6x6, 7x7, 8x8, 9x9.

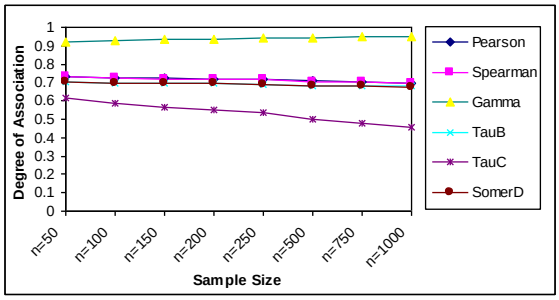

Figure 4a: Sample size against degree of the ordinal measure of associations for ρ =0.9 and 3x3 Figure.

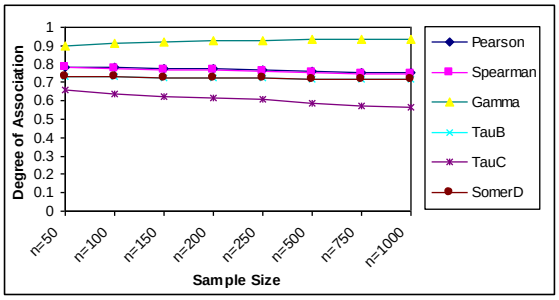

Figure 4b: Sample size against degree of the ordinal measure of associations for ρ =0.9 and 4x4 Figure.

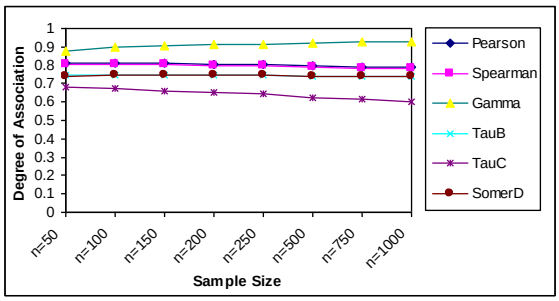

Figure 4c: Sample size against degree of the ordinal measure of associations for ρ =0.9 and 5x5 Figure.

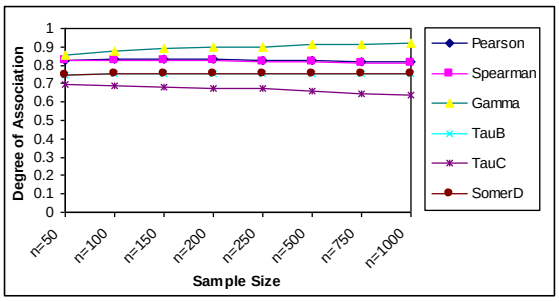

Figure 4d: Sample size against degree of the ordinal measure of associations for ρ =0.9 and 6x6 Figure.

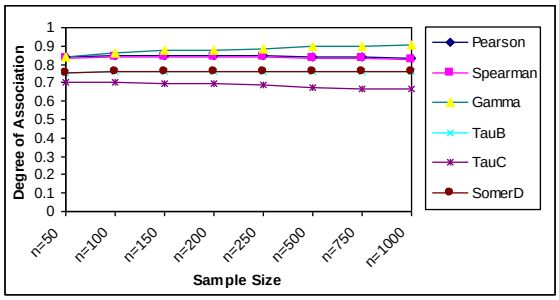

Figure 4e: Sample size against degree of the ordinal measure of associations for ρ =0.9 and 7x7 Figure.

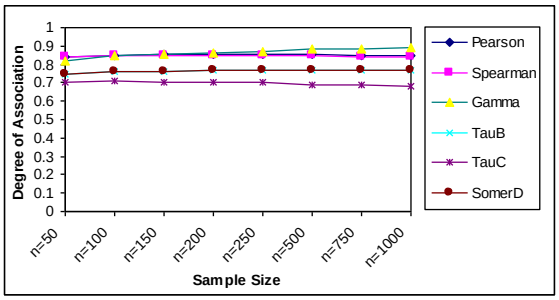

Figure 4f: Sample size against degree of the ordinal measure of associations for ρ =0.9 and 8x8 Figure.

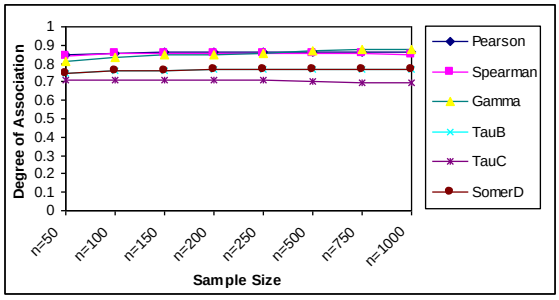

Figure 4g: f Sample size against degree of the ordinal measure of associations or ρ =0.9 and 9x9 Figure.

Figure 4: Sample size against degree of the ordinal measure of associations for ρ =0.9 and 3x3, 4x4, 5x5, 6x6, 7x7, 8x8, 9x9.

5 Results and remarks

As can be seen from Figure 1a for Table dimension 3x3 and 4x4 Gamma ordinal measure of association presents closer estimate of the degree of association which is expected to be 0.5 in comparison with the other ordinal measures. However it can be concluded from the same figure that Pearson's correlation and Spearman's rank correlation presents a good estimate of the actual degree of association as table dimension increases. In the meantime, Kendall's Tau-b, Kendall's Tau-c and Somers' d for such a small sample present a poor estimate of the actual degree of ordinal association in average for any table dimension. As table dimension increases Gamma coefficient decreases too.

When the sample size increases from 50 to 100 sizes, the results are presented in Figure 1b. A similar line plot has clearly been observed. There is not much difference at all. In fact, when the sample size increases to a fair size, for small table dimensions Gamma overestimate the actual the degree of ordinal association. Pearson's correlation and Spearman's rank correlation always tend to increase to a much better estimation as table dimension increases. Although as table dimension increases, Kendall's Tau-b, Kendall's Tau-c and Somers' d for such a fair and large sample present a slight different estimate of the actual degree of association. They are all underestimating the actual degree of ordinal measure of association. Meanwhile for large samples Kendall's Tau-c is the worst no matter how large the table dimension is.

Increasing the actual degree of association from 0.5 to 0.9 does not make any difference in the results obtained. The picture in Figure 2 shows that the lines representing the ordinal measure of association are shifted up by the amount of the increment of the actual degree of association.

Figure 3 has been prepared from the same results of Figure 1 except that the horizontal axis represents the sample size. The aim of drawing the picture in Figure 3 is to investigate the effect of sample size. It can be seen from Figure 3a that is drawn for the table dimension 3x3 as the sample size increases the Gamma ordinal measure of association not only increases but also overestimates the actual degree of association. However as the sample size increases Pearson's correlation, Spearman's rank correlation, Kendall's Tau-b, Kendall's Tau-c and Somers' d slightly decrease and underestimate the actual degree of association for the 3x3 contingency table in average. The picture in Figure 3b that is drawn for the table dimension 4x4 shows no significant difference in comparison with the 3x3 contingency table. When the results obtained for the rest of the contingency table dimensions are analyzed, it is found that there is no much difference at all except a little shift up. To this end, when the table dimension is large enough, Pearson's correlation and Spearman's rank correlation tend to present a better estimate of the actual degree of ordinal measure of association no matter what the sample size is, but the other ordinal measure of associations except Gamma always underestimate

the actual degree of association. Gamma tends to increase as the sample size increases. Figure 4 presents similar results of the results presented in Figure 3 apart from that the actual degree of association has changed from 0.5 to 0.9. There cannot be observed any different results in comparison with Figure 3.

6 Conclusions and discussion

As the number of table dimension increases, no matter what the sample size of the contingency table is Pearson's correlation and Spearman rank correlation coefficients of the generated tables from the bivariate standard normal distribution increase in average, but they still slightly underestimate the actual degree of ordinal measure of association.

Although Kendall's Tau-b, Tau-c and Somers' d increase as the table dimension increases, those measures of association always underestimate the actual degree of association of the generated tables.

Pearson's correlation is slightly larger than Spearman correlation. Those two measures are good at fairly large dimensional doubly ordered tables.

Gamma coefficient is good when the table dimension is small for relatively small sample sizes. It increases and overestimates as the sample size increases for any certain type of table dimension. In overall, for square tables Gamma presents the best estimation of the actual degree of the association in average.

There is another measure of association called polychoric correlation which has not been included in this study. Since that measure of association uses iterative methods which are hard to estimate the actual degree of association, it has been excluded from the study. However, a new study may be performed with this correlation to make a comparison with the ordinal measure of associations that are presented in our study.

A future study may be performed to develop a new measure of ordinal association that is free of both table dimension and sample size.

References

[1] Agresti, A. (2010): *Analysis of ordinal categorical data*. 2nd ed., New York: Wiley.

[2] Brown, M.B. and Benedetti, J.K. (1977a): On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, **41**, 347-355.

[3] Brown, M.B. and Benedetti, J.K. (1977b): Sampling behaviour of tests for correlation in two-way contingency tables. *Journal of the American Statistical Association*, **72**, *309-315***.**

[4] Cyrus, R.M. and Nitin, R.P. (1995): Exact Tests. http://support.spss.com/ProductsExt/SPSS/ESD/17/Download/User%20Manuals/English/SPSS%20Exact%20Tests.pdf.

[5] Garson, G.D. (2008): *Ordinal Association: Gamma, Kendall's tau-b and tau-c, Somers' d.*, http://faculty.chass.ncsu.edu/garson/PA765/assocordinal.htm.

[6] Goodman, L.A. and Kruskal, W.H. (1963): Measures of association for cross classifications. *Journal of the American Statistical Association*, **58**, 310–364.

[7] Goodman, L.A. and Kruskal, W.H. (1979): *Measures of Association for Cross Classifications*. New York: Springer.

[8] Hoeffding, W. (1948): A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, **19**, 293-325.

[9] Kendall, M.G. (1955): *Rank Correlation Methods*. New York: Hafner Publishing Co.

[10] Kendall, M.G. and Stuart, A. (1973): *The Advanced Theory of Statistics*. Volume 2: Inference and Relationship, London: Griffin.

[11] Kruskal, W.H. (1958): Ordinal measures of association. *Journal of the American Statistical Association*, **53**, 814-861.

[12] Lane, D.M. (1997): HyperStat Online Statistics Textbook. http://davidmlane.com/hyperstat/A34739.html.

[13] Liebetrau**,** A.M**.** (1983): *Measures of Association*. Beverly Hills: Sage. Publications, Inc.

[14] Lohninger, H. (1999): *Teach/Me Data Analysis*. Berlin-New York-Tokyo: Springer-Verlag.

[15] Rezanková, H. (2009): *Cluster Analysis and Categorical Data.* http://panda.hyperlink.cz/cestapdf/pdf09c3/rezankova.pdf.

[16] SAS (2010): *Measures of Association.* http://v8doc.sas.com/sashtml/stat/chap28/sect20.htm.

[17] Svensson, E. (2000): Concordance between ratings using different scales for the same variable. *Statistics in Medicine*, **19**, 3483-3496.

[18] Somers, R.H. (1962): A new asymmetric measure of association for ordinal variables. *American Sociological Review*, **27,** 799-811.

[19] UCLA (2007): Academic Technology Services. http://www.ats.ucla.edu/stat/mult_pkg/whatstat/nominal_ordinal_interval.htm

[20] Tu, X. (2007): Goodman-Kruskal Gamma, Kendall's tau-b, Stuart's tau-c and Somer's d. http://www.urmc.rochester.edu/smd/biostat/people/faculty/TuSite/bst466/documents/chap2.6.pdf.