Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.4, No.11, 2014

Robust Method for Testing the Significance of Bivariate Correlation of Ordinal Data

Odimientimi D. Agbedeyi* Amechi H Igweze
Department of Mathematics and Statistics, Delta State Polytechnic Ogwashi-uku, Delta State, Nigeria
* E-mail of the corresponding author: desmondagbedeyi@gmail.com

Abstract
The Tau's statistics were introduced to solve the problems of tied data but its effect on shape of table cannot be ascertained. This study compares the non parametric approaches for testing the significance of bivariate correlation for ordinal data based on table shape (square and rectangular table). The bootstrap method was used to compare the magnitude of the correlation values and the associated standard error of the values of tau's -b, c and Gamma over a square and rectangular table. The Gamma was found to be the most robust statistic for computing the correlation of ordinal data

Keywords: Tau statistics, Gamma, ordinal correlation, Bivariate correlation

1. Introduction

Bivariate data are data in which two variables are measured on an individual. Often time interest could be in describing the relationship between them. One measure used to describe the strength of linear relation between two variables is the linear correlation coefficient. Two variables are positively correlated if, whenever the value of one variable increases, the value of the other variable also increases. A negative correlation occurs when the value of one variable increases and the value of the other variable decreases. ρ is the Greek letter used in representing the parameter. The correlation coefficient has values ranging from -1 to 1, where -1 shows a perfect negative correlation; and 1 indicates a perfect positive linear association. Values of 0 indicate little or no linear association between the two variables.

The Kendall's rank correlation coefficient, frequently referred to as Kendall's tau (τ), measures the relationship between two measured quantities. The tau test is classified as a non-parametric tool which was specifically designed to test for independence.

The Tau's family of statistics is non parametric measures of correlation between two qualitative variables. It is most appropriate for square tables. Kendall's Tau is equivalent to Spearman's Rho, with regard to the underlying assumptions, but Spearman's Rho and Kendall's Tau are not identical in magnitude, since their underlying logic and computational formulae are quite different. In most cases, these values are very similar, and when discrepancies occur, it is probably safer to interpret the lower value. More importantly, Kendall's Tau and Spearman's Rho imply different interpretations. Spearman's Rho is considered as the regular Pearson's correlation coefficient in terms of the proportion of variability accounted for, whereas Kendall's Tau represents a probability, i.e., the difference between the probabilities that the observed data are in the same order versus the probability that the observed data are not in the same order.

There is no well-defined intuitive meaning for Tau -b, which is the surplus of concordant over discordant pairs as a percentage of concordant, discordant, and approximately one-half of tied pairs. The rationale for this is that if the direction of causation is unknown, then the surplus of concordant over discordant pairs should be compared with the total of all relevant pairs, where those relevant are the concordant pairs, the discordant pairs, plus either the X-ties or Y-ties but not both, and since direction is not known, the geometric mean is used as an estimate of relevant tied pairs.

Tau-b requires binary or ordinal data. It reaches 1.0 (or -1.0 for negative relationships) only for square tables when all entries are on one diagonal. Tau-b equals 0 under statistical independence for both square and non-square tables. Tau-c is used for non-square tables.

Kendall's Tau-c, also called Kendall-Stuart Tau-c, is a variant of Tau-b for larger tables. It equals the excess of concordant over discordant pairs, multiplied by a term representing an adjustment for the size of the table.

Thus, Gamma is the surplus of concordant pairs over discordant pairs, as a percentage of all pairs, ignoring ties. Gamma defines perfect association as weak monotonicity. Under statistical independence, Gamma will be 0, but it can be 0 at other times as well (whenever concordant minus discordant pairs are 0).

Gamma is a symmetric measure and computes the same coefficient value, regardless of which is the independent (column) variable. Its value ranges between +1 to –1.

In terms of the underlying assumptions, Gamma is equivalent to Spearman's Rho or Kendall's Tau; but in terms of its interpretation and computation, it is more similar to Kendall's Tau than Spearman's Rho. Gamma statistic is, however, preferable to Spearman's Rho and Kandall's Tau, when the data contain many tied observations.

Another useful way of looking at the relationship between two nominal (or categorical) variables is to cross-classify the data and get a count of the number of cases sharing a given combination of levels (i.e., categories), and then create a contingency table (cross-tabulation) showing the levels and the counts.

A contingency table lists the frequency of the joint occurrence of two levels (or possible outcomes), one level for each of the two categorical variables. The levels for one of the categorical variables correspond to the columns of the table, and the levels for the other categorical variable correspond to the rows of the table. The primary interest in constructing contingency tables is usually to determine whether there is any association (in terms of statistical dependence) between the two categorical variables, whose counts are displayed in the table. A measure of the global association between the two categorical variables is the Chi-square statistic, which is computed as follows:

This statistic is distributed according to Pearson's Chi-square law with (k-1) ´(h-1) degrees of freedom. Thus, the statistical significance of the relationship between two categorical variables is tested by using the $\chi 2-$ test which essentially finds out whether the observed frequencies in a distribution differ significantly from the frequencies, which might be expected according to a certain hypothesis (say the hypothesis of independence between the two variables).

The reason for this assumption is that the Chi-square inherently tests the underlying probabilities in each cell; and when the expected cell frequencies fall, these probabilities cannot be estimated with sufficient precision. Hence, it is essential that the sample size should be large enough to guarantee the similarity between the theoretical and the sampling distribution of the X^2 – statistic. In the formula for computation of X^2, the expected value of the cell frequency is in the denominator. If this value is too small, the X^2 value would be overestimated and would result in the rejection of the null hypothesis.

To avoid making incorrect inferences from the c2–test, the general rule is that an expected frequency less than 5 in a cell is too small to use. When the contingency table contains more than one cell with an expected frequency < 5, one can combine them to get an expected frequency ³5. However, in doing so, the number of categories would be reduced and one would get less information.

However, Phi-square loses this nice property, when both dimensions of the table are greater than 2. By a simple manipulation of Phi–square, we get a measure (Cramer's V), which ranges from 0 to 1 for any size of the contingency table. Cramer's V is computed as follows:

The coefficient of contingency is a Chi-square -based measure of the relation between two categorical variables (proposed by Pearson, the originator of the Chi-square test). It is computed by the following formula:

Its advantage over the ordinary Chi-square is that it is more easily interpreted, since its range is always limited to 0 through 1 (where 0 means complete independence). The disadvantage of this statistic is that its specific upper limit is 'limited' by the size of the table; Contingency coefficient can reach the limit of 1, only if the number of categories is unlimited.

Tau-b is recommended for ordinal correlations while Cramer's V or phi for nominal correlations. Furthermore Phi is appropriate if one variable has only two categories. The advantage of using Cramer's V and tau-b is that when the numbers of categories of the row and column variables are roughly equal, they are measured more or less on the same scale.

According to Beversdorf and Sa (2011), many estimators exist for ρ and two of them are frequently used: the Pearson Product Moment Correlation, which is applied to ratio and interval data; and the Spearman Rank Order Correlation, which is used for ordinal data. When the data is not bivariate normal and the sample size is small, the nonparametric Spearman rank correlation is useful.

Outliers, unequal variances, non-normality, and nonlinearity unduly influence the Pearson correlation. Furthermore, the spearman correlation which is a non parametric version of the Pearson does not consider cases

of tied observation. The Tau's statistics were introduced to solve the problems of tied data but the effect of table shape cannot be ascertained.

2. Purpose of the Study

This study compares the non parametric approaches for testing the significance of bivariate correlation for ordinal data based on table shape:-square and rectangular table. The specific objective therefore is to compare the magnitude and standard errors of Tau-b, Tau-c and the Gamma over squared and rectangular tables

3. Literature Review

Different indices can be used to measure the correlations between variables. Some of these indices include: Pearson Product Moment Correlation coefficient (r), Spearman rank correlation coefficient (r_s), and Kendall's tau coefficient (τ). Some latest studies on non parametric correlations are reviewd.

According to Ebuh & Oyeka (2012), in the Kendall approach the populations of interest may be data on ordinal scale and need not be normally distributed or continuous. In other to estimate a tau correlation coefficient, samples of equal sizes are drawn from any two observations; each ranked either from the largest to the smallest or from the smallest to the largest. In each sample tied observations are as usual assigned to their mean ranks. The assigned ranks in each sample are then arranged in their natural order from 1 through n. The ranks assigned to the observations in the sample drawn from the second population, are then juxtaposed against the naturally arranged ranks for the correspondent subjects in the sample drawn from the first population.

Hauke & Kossowski (2011) argues that Kendall's tau can be used as an alternative to Spearman's rho for ranked data. The authors noted that Kendall tau measures the average between pairs of data. Hence it has been evidently recommended as a measure of the concordance between two variables. Other advantages of Kendall are that its distribution has to some extent better statistical properties; again, the direct interpretation of probabilities of observing concordant and discordant pairs is a worthwhile advantage.

In the study of Murray (2013), the Pearson, Spearman rho and Kendall tau_b analyses was conducted on scale data in other to determine if statistical tests on Likert scale data affect the conclusions. The study revealed that there was a positive relationship between all the permuted pairs of the variables at the $p<.05$ level. Though the relationship revealed a weak relationship for almost all the tests, the author concluded that the Pearson and Spearman rho were similar and that Kendall tau_b is different.

Ferguson, Genest & Hallin (2000) showed that the Kendall's tau can be modified to test for a serial dependence in a univariate time series. They gave formulae for both the mean and variance of circular and non-circular versions of the statistic and proved its asymptotic normality under the hypothesis of independence. They further presented a Monte Carlo study comparing the power and size of a test based on Kendall's tau to that of competing procedures based on alternative parametric and nonparametric measures of serial dependence. Their simulations revealed that Kendall's tau outperforms Spearman's rho in detecting first-order autoregressive dependence, despite the fact that these two statistics are asymptotically equivalent under the null hypothesis.

Although Galla (1987) however argued that tau was presented for a special case in which no ties exist, Hawkes (1971) showed it to also be appropriate for the case in which ties do exist. The major drawback to this procedure is that the sampling distribution for tau was not yet determined. There are, therefore, no tests of significance for the partial rank correlation coefficient at that time.

Hauke & Kossowski (2011) compared the coefficients and statistical significance of Pearson's product-moment correlation and Spearman's rank correlation over some data sets. They observed that the significance of Spearman's correlation could lead to the significance or non-significance of Pearson's correlation coefficient even for large data set. However, this does not hold in the case of the significance of Pearson's coefficient translating into the significance of Spearman's coefficient. They also noted that it is possible to meet a situation where the coefficient of Pearson correlation is negative while Spearman's coefficient is positive. In conclusion they advised that the Spearman's rank correlation should not be over interpreted as a significant measure of the strength of the associations between two variables

Norman (2010) suggested in His study that parametric test can be conducted on Likert data without fear of arriving at the wrong conclusion.

Newson (2001) reviewed the uses of three non parametric statistics namely Somers' D, Kendall's Tau and the Hodges-Lehmann median difference. The author found that the confidence limits for these parameters, and their differences, were more informative than the traditional practice of reporting only *P*-values

According to Fredricks & Nelsen (2007), the sample value of Spearman's rho is about 50% larger than the sample value of Kendall's tau. They explained this behavior by proving that, the ratio of rho to tau tends to 3/2 as the joint distribution approaches that of two independent random variables. They also found adequate conditions for determining the inequality direction between three times tau and twice rho when the underlying joint distribution is absolutely continuous.

Maturi & Elsayigh (2010) compared ten correlation coefficients using a three-step bootstrap approach (TSB) to determine the optimal repetitions, *B* and to estimate the standard error of the statistic with some degree of accuracy. The coefficients include: Pearson product moment, Spearman's rho, Spearman's Footrule, Symmetric Footrule, Kendall's tau, the Top - Down, the greatest deviation, Weighted Kendall's tau, Blest, and Symmetric Blest's coefficient. The authors considered a standard error criterion for their comparisons. However, since the rank correlation coefficients suffer from the tied problem resulting from the bootstrap technique, they employed the use existing modified formulae for some rank correlation coefficients; otherwise, the randomization tied-treatment was applied. They concluded that the Pearson correlation coefficient should be used if the data meets the assumption of normality; otherwise, the greatest deviation performed well especially when the data has outliers. However, when we want emphasis on the initial (top) data, the Symmetric Blest's coefficient has lowest standard error amongst other weighted correlation coefficients.

4. Methods

This study compares the bootstrap values for tau's b, c and Gamma over a square and rectangular table. A nominal data on grades of students who did their entry registration over a varied period of time was used for the study.
The various statistics employed are:

TAU-B: *Kendall's Tau-b* measures the association of a two ordinal variable, often used for a 2-by-2 tables. It is computed as the excess of concordant over discordant pairs (C - D), divided by a the geometric mean between the number of pairs not tied on *X* (X_0) and the number not tied on *Y* (Y_0

$$\text{Tau-b} = \frac{(C-D)}{\sqrt{[(C+D+X_0)(C+D+Y_0)}}$$

TAU-C: is a variant of *Tau-b* for larger tables. It the excess of concordant over discordant pairs, multiplied by an adjustment for the size of the table.

$$\text{Tau-c} = (C-D) * \left(\frac{2m}{n^2(m-1)}\right)$$

Where:

m = the number of rows or columns, whichever is smaller

n = the sample size.

Goodman-Kruskal Gamma: the Gamma is based on the difference between concordant pairs (C) and discordant pairs (D). Gamma is computed as follows:

G = (C-D)/(C+D)

Thus, Gamma is the surplus of concordant pairs over discordant pairs, as a percentage of all pairs, ignoring ties. Gamma defines perfect association as weak monotonicity. Under statistical independence, Gamma will be 0, but it can be 0 at other times as well (whenever concordant minus discordant pairs are 0)

Table 1. Time/Grade Cross Tabulation Data

TIME * GRADE Cross tabulation (5X5)							
Count							
		GRADE					
		A	B	C	D	F	Total
TIME	1.00	64	42	50	83	21	260
	2.00	20	25	43	102	31	221
	3.00	18	40	64	87	28	237
	4.00	12	33	36	77	19	177
	5.00	12	22	20	31	20	105
Total		126	162	213	380	119	1000

5. Discussion of Finding

Table 2. Bootstrap Result for Squared Table

BOOTSTRAP RESULT FOR SQUARED TABLE												
	5x5			4X4			3X3			2X2		
	Value	**Bias**	**Std. Error**	**Value**	**Bias**	**Std. Error**	**Value**	**Bias**	**Std. Error**	**Value**	**Bias**	**Std. Error**
Kendall's tau-b	.082	-.002	.028	.111	.000	.030	.218	.000	.045	.147	-.001	.081
Kendall's tau-c	.078	-.002	.027	.107	.000	.029	.213	.000	.044	.133	-.002	.073
Gamma	.105	-.002	.037	.154	.000	.042	.329	.000	.066	.311	-.007	.161

The comparison of the bootstrap results for the nonparametric methods of testing correlation of ordinal data in a squared table has revealed a number of findings: Firstly, the Gamma statistic consistently gave the highest correlation coefficient (value) followed by the Kendall's tau-b. The Kendall's tau-c gave the least correlation value in all.

The result also shows a low negative bias (of -0.007) for gamma for a 2x2 table and (-0.002 for a 5x5 table). In all the tau-c recorded the least bias in all square table while gamma had the highest bias in all square table.

Table 2. Bootstrap Result for Rectangular Table

BOOTSRTAP RESULT FOR RECTANGULAR TABLE									
	5X4			3X4			2X3		
	Value	Bias	Std. Error	Value	Bias	Std. Error	Value	Bias	Std. Error
Kendall's tau-b	.069	.001	.029	.120	.001	.035	.187	.000	.059
Kendall's tau-c	.068	.000	.029	.124	.000	.036	.206	-.001	.065
Gamma	.093	.001	.039	.175	.001	.050	.332	-.001	.100

The comparison of the bootstrap results for a rectangular table shows that in a 5x4 table, the Gamma again consistently gave the highest correlation value, the tau-b fared slightly better than tau-c with a value of 0.01 and the same standard error of 0.029. In the smaller tables of 3x4 and 2x3, Gamma consistently gave the highest correlation value followed by tau-c. The tau-b gave the least values as the size of the table reduced. The gap between the correlation value of gamma and tau-c widened as the size of the table reduces.

6. Conclusion

From the results, the Gamma statistic consistently fared better than the Tau-c and Tau-b. Furthermore the Tau-c did better than tau-b in a rectangular table. This leads to the conclusion that the Gamma is the most robust statistic for calculating the coefficient of correlation for ordinal data

Reference

Beversdorf, L. M and Sa, P. (2011) Tests for Correlation on Bivariate Non-Normal Data. Journal of Modern Applied Statistical Methods, Vol. 10, No. 2, 699-709

Ebuh, G.U. and Oyeka, I.C.A. (2012) A Non Parametric Method for Estimating Partial Correlation Coefficient. J. BiomBiostat 3(8)http://dx.doi.org/10.4172/2155-6180.1000156.

Ferguson, T. S., Genest, C., & Hallin, M. (2000) Kendall's tau for Serial Dependence. The Canadian Journal of StatisticsVol. 28,

Fredricks, G. A., & Nelsen R. B. (2007) On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. Journal of Statistical Planning and Inference 137 (2007) 2143 – 2150

Galla, J. P. (1987) Kendall's tau and Kendall's partial correlation: Two BASIC programs for microcomputers. Behavior Research Methods, Instruments, & Computersn1987, 19 (1), 55-56

Hauke, J. & Kossowski, T (2011) Comparison Of Values Of Pearson's And Spearman's Correlation Coefficients On The Same Sets Of Data. Quaestiones Geographicae 30(2)

Maturi, T, A. & Elsayigh, A. (2010) A Comparison of Correlation Coefficients via a Three-Step Bootstrap Approach. Journal of Mathematics Research Vol. 2, No. 2, May 2010

Murray, J (2013) Likert Data: What to Use, Parametric or Non-Parametric? International Journal of Business and Social Science Vol. 4 No. 11

Newson, R (2001) Parameters behind "non-parametric" statistics: Kendall's a, Somers' D and median differences. The Stata Journal vol 1(1), pp. 1–20

Norman, G. (2010). Likert scales, levels of measurement and the laws. Adv in Health Sci Educ, 15:625-632,DOI 10,1007/s10459-010-9222-y.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar