doi:10.20944/preprints202004.0276.v1

Peer-reviewed version available at *Symmetry* **2019**, *11*; doi:10.3390/sym11070872

Comparison of Correlation Measures for Nominal Data

Tanweer Ul Islam

National University of Sciences & Technology, Islamabad

Email: tanweer@s3h.nust.edu.pk

Mahvish Rizwan

National University of Sciences & Technology, Islamabad

Email: mahvishriz@yahoo.com

ABSTRACT

In social sciences, a plethora of studies utilize nominal data to establish the relationship between the variables. This, in turn, requires the correct use of correlation technique. The choice of correlation technique depends upon the underlying assumptions and power of the test of significance. The objective of the research is to explore the best measure of association for nominal data in terms of size, power and bias in estimation. Monte Carlo simulations reveal that the Phi and Pearson correlation statistics performs equally well in terms of size, power, and bias for naturally dichotomous variables. When both variables are artificially dichotomized, the Tetrachoric statistic has an edge in terms of bias to Pearson correlation statistic. If one variable is continuous and other is artificially dichotomized, the Biserial correlation measure turns out to be less biased as compared to Pearson statistic although both statistics exhibit similar power and size properties. If one variable is continuous and other is naturally dichotomized, it is hard to choose between the Point Biserial and Pearson correlation measures. Finally, if one variable is naturally dichotomous and other is artificially dichotomized, correlation coefficient *V* is compared with Pearson, Phi and Tetrachoric correlation techniques in terms of bias in estimate. The results indicate that the Tetrachoric statistic considerably overestimates the correlation value against non-normal distributions. Pearson and Phi correlation slightly underestimate the correlation value. In contrast, the correlation statistic *V* perform well.

Keywords: phi correlation, tetrachoric correlation, biserial correlation, point biserial correlation, correlation coefficient V, bias in estimate, size of test, power, Monte Carlo simulation

1. INTRODUCTION

Correlation analysis, though a highly used statistical technique, in general, is the one frequently misused in social and behavioural research, when analysing rating scales [1]. Onwuegbuzie & Daniel [2] provide a critical analysis on the use and misuse of Pearson Product Moment Correlation (PPMC). They highlight the shortcoming of statisticians and academicians when using the correlation coefficient. These include failure to properly consider the underlying assumptions of the coefficient used and power of the test. The literature is well established regarding the performance of PPMC under assumption of non-normality [3, 4] and comparison of PPMC with other non-parametric correlation statistics for ordinal data, such as Spearman, Kendall's tau, Goodman and Kruskall's Gamma [5-7].

Theoretical literature on correlation coefficient for nominal data is available, however, these statistics are not elaborated extensively in the mainstream textbooks. The scanty literature also lacks in terms of extensive power comparison of these statistics. Therefoer, the objective of the study is to explore the performance of the measures of association for nominal data. The list of selected correlation coefficients for the nominal scale includes phi correlation, tetrachoric correlation, biserial correlation, point biserial correlation and correlation coefficient V. These statistics are selected based on their extensive use in economics and social sciences [8-15].

2. REVIEW OF LITERATURE

Pearson product moment correlation is specified for continuous scale variable. For the naturally occurring correlated dichotomous variable, phi correlation is suggested [16,17]. Richardson & Stalnaker [18] proposes the point biserial correlation for correlation between continuous and naturally occurring dichotomous variable. Tetrachoric correlation is introduced by Pearson (1900) to measure correlation between artificially dichotomized variables. Pearson [19] gives the biserial coefficient to determine relation between artificially dichotomized and continuous variable. Ulrich & Wirtz [20] derive the coefficient V to determine the correlation between a naturally occurring and artificially occurring dichotomous variable.

For each measure of nominal correlation mentioned, the literature is sufficiently developed in terms of mathematical representation, individual significance testing, confidence interval measurement [21,22]. However, there is a dearth of literature regarding the power performance of correlation measures for nominal data. This section discusses the two available strands of literature: (i) statistical development of each correlation technique, (ii) efficiency analysis.

2.1. STATISTICAL PROPERTIES

Phi, point biserial and biserial correlation techniques give reliable estimates if there is equal split in the dichotomous variable [23]. However, Farrington [24] notes that the maximum value for phi correlation is one when the row and column total are equal, but it is less than one when the marginal distribution of rows and columns is dissimilar for a 2 x 2 contingency table. In order to overcome this drawback, he provides a corrected formula for phi correlation- phi max correlation. This new correlation is calculated by dividing the obtained phi correlation by the maximum possible phi correlation available. Later, Davenport & El - Sanhurry [25] analytically explore phi max and find it as an asymmetrical measure and shows that the corrected methodology for phi, phi max, is not robust for the very condition it was corrected for i.e. unequal total of rows and columns.

Literature has focused mostly on the impact of marginal distribution on the phi correlation. Liu, [26] provides the methodology to standardize the marginal of a 2 x 2 contingency table, so that the two tables are comparable for phi correlation. According to Warren [27], marginal frequency affects the phi correlation, though tetrachoric correlation does not depend on marginal frequency of the cell, unless the cell frequency is zero. Ekstrom [22] shows that there exists a continuous correspondence between phi correlation and tetrachoric correlation and it has no major implication if phi correlation is calculated using the assumption of tetrachoric correlation or vice versa. Demirtas [28] extends this correspondence in case the underlying distribution of variables is non-normal given marginal characteristic, degree of linear association, and marginal probabilities.

In his pioneer work, Lev [29] gives the mathematical formula for point biserial correlation along with test of significance and measure for confidence interval. The mathematical connection between biserial and point biserial correlation is also well established in the literature when the continuous variable follows normal distribution [30,31]. Demitras & Hedeker [32] further highlight this notion and propose a technique to calculate one coefficient from other given that the underlying distribution of the continuous variable is non-normal as opposed to normal distribution. Tate [33] builds on the framework given by Lev [29] and derives a mathematical model for point biserial correlation given equal marginal frequency for the dichotomous variable. Confidence limits are also calculated. Tate [34] observes that while point biserial has a range between ±1, the range for biseral correlation is unbounded for certain sample size and marginal probabilities. Furthermore, standard deviation is minimum for both point biserial and biserial correlation when dichotomization of binary variable is done at the mean, i.e. the marginal frequency is 0.5. Biserial is a good maximum likelihood estimate when population correlation is zero. Nevertheless, when compared to Pearson correlation, the standard deviation of biserial is more than the standard deviation of Pearson. Gupta [35] derives formula for large sample standard error of point biserial correlation.

Becker [36] provides correction to the formula of point biserial correlation because of unequal sample size. Furthermore, Gradstein [37] determines that the highest possible value obtained for point biserial correlation is 0.798 when marginal frequency is equal. Given the largest portion of dichotomous variable, Terrell [38,39] gives the table for values converted from point biserial correlation to biserial correlation and provides significance tables for point biserial and biserial correlation at 0.01 and 0.05 levels of significance. On balance, it is recommended that point biserial correlation should be used instead of biserial correlation because no assumption is made regarding the underlying distribution of the binary variable, however, if normality of distribution can be established, then biserial correlation is used.

Tetrachoric correlation was not a popular statistic to use in the decades after it was introduced because of computational difficulties. To overcome this problem, Guilford [40] gives tables of tetrachoric correlation for given sample size and marginal probabilities for 0.01 & 0.05 levels of significance. However, the later literature focuses on the development of easier to compute formula. Brown [41] gives a computational algorithm to calculate tetrachoric correlation and its standard error. A cell adjustment of 0.5 is proposed in case the cell frequency is zero. With this new adjustment, the bias in tetrachoric correlation is reduced considerably. Brown and Bennedetti [42] compare the estimates of mean and standard deviation of tetrachoric correlation with their expected values for various 2 x 2 contingency tables. From the results, it is shown that significant bias is present in the mean estimate of tetrachoric correlation when the cell frequency is less than five. To calculate the standard deviation of correlation, three different formulas are also given. Bonett and Price [21] propose a simpler form of formula to calculate the tetrachoric correlation coefficient.

Ulrich & Wirtz [20] introduce a correlation coefficient V in order to measure the correlation between artificially dichotomized (Y) and naturally dichotomous (X) variables. The article gives underlying assumptions for correlation V, and methodology to calculate this correlation coefficient and its standard error. For a 2x2 contingency table, if conditional mean of Y_L given X is equal to zero and conditional mean of Y_L given X is equal to one are equal, then correlation V yields the value of zero. It is also shown that correlation V is a maximum likelihood estimator and produces consistent and efficient results. The correlation V is compared to biserial and phi correlation for different level of marginal probabilities and cut off points for naturally dichotomous and underlying continuous variable respectively. Results conclude that the correlation V is invariant to point of cuts, biserial vary considerably but phi correlation varies more.

2.2. EFFICIENCY ANALYSIS

Tate [43] points out that biserial correlation statistic has minimum variance when variables are normally distributed, and mean is used for dichotomization of the variable. Biserial correlation has an efficiency of 1 when population correlation is zero, but when population correlation is large or close to ± 1, then its efficiency is zero. This is a major drawback, since mostly, larger correlation values are of interest in research. On other hand, point biserial assumes no distributional assumption for discrete variable and can be used for any sample size. Furthermore, it has efficiency of 1 for large sample of sizes, regardless of value of population correlation and minimum variance when point of dichotomy is at mean.

Martin [44] measures the effect of number of categories (scale points) on bias in estimates for Pearson, biserial and tertrachoric correlation. For the purpose of study, the data is simulated for continuous normally distributed variable with sample size of 1000. First, the effect of varying the number of categories (scale points) on Pearson correlation is determined. The results are then compared with tetrachoric and biserial correlation. It is found that the number of intervals has profound effect on the magnitude of the Pearson coefficient. Pearson correlation is not robust as the impact of information lost, when continuous variable is categorized into many scale points, is quite substantial. Compared to Pearson, biserial and tetrachoric correlation are less biased when continuous variable is categorized into scale points that are more than two. However, biserial correlation gives large standard error.

Fowler [45] measures theoretical power for Pearson product moment correlation, Spearman's point biserial correlation, and phi correlation coefficient for continuous variables X and Y. The aim of the study is to estimate the effect of non-normality on power of the correlation coefficients.

Power of each correlation test is computed as a percentage of power of Pearson correlation. The results show that Pearson correlation is robust for most non-normal distributions. If population median is used for dichotomization of data, point biserial correlation show relatively good power against some alternative distributions, but there is loss in power when the sample size is small. Point biserial is at advantage to Pearson correlation when the sample size is greater than 50 and the distribution considered is highly skewed or leptokurtic. Phi correlation performs poorly under all conditions. The results also indicate strong relative power of Spearman correlation under wide range of conditions. For leptokurtic distribution with small sample size, Pearson performs better than Spearman correlation.

Greer, Dunlap, & Beatty [46] conduct Monte Carlo simulations to investigate bias and standard error of tetrachoric correlation technique and compare it with phi correlation and Pearson correlation. For the purpose of analysis, they undertake 10,000 simulations for samples of size 50, 100 and 150 and population correlation of 0, 0.3 and 0.5 against the alternative distributions of normal and log normal with extremely skewed marginal probabilities. The data are dichotomized in two ways: specifying a threshold level above which the variable takes the value of one or forcing the specified proportion of sample to take the value of one. Following Brown and Benedetti [47] the 0.5 adjustment to cell frequency is used, when the cell has zero frequency in a 2x2 contingency table. The results from the study show that as the sample size increases, both the bias and standard error decrease. Overall, when population correlation is zero, phi and tetrachoric correlation measures give similar results for both distributions. When population correlation is not equal to zero, it is concluded that the tetrachoric is insensitive to the underlying distribution, and it is an especially useful estimate if the scores are dichotomized at the median. Also, tetrachoric can be used with extreme marginal probabilities, given that none of the cells has zero frequency, since it gives results that are less biased and have less standard error than Pearson correlation. Bonett and Price [21] provide a computationally simple method for computing tetrachoric correlation and derive its standard error and confidence interval. This new tetrachoric correlation measure shows better performance when applied to binary data. Phi correlation, in comparison, tends to underestimate the correlation between two artificially dichotomized variables.

Akkartal [48] conducts Monte Carlo simulation study to test the performance of point biserial correlation and compares it with two sample independent t-test against the alternative distributions like Normal, Beta and Chi squared. Both the tests show similar performance in terms of size and power. When effect size increases, both correlation techniques give high power regardless of sample size, distribution type or marginal frequency. The author, however, concludes that even though both tests give similar power and Type I error rate, it is better to use point biserial as a measure of correlation for naturally binary and continuous variable because of difference in interpretation; the t-test compares difference in mean.

Caution should also be taken while using tetrachoric and biserial correlation because these correlations are based on hypothetical underlying distribution which is strictly assumed to be normal for variables that are not directly observable observation [49,50].

3. CORRELATION MEASURES FOR DICHOTOMOUS DATA

The section will present mathematical description of the selected correlation statistics for dichotomous variables (Table 1).

Table 1: Correlation Measures for Nominal Data

		Variable Y		
		Naturally Dichotomous	Artificially Dichotomized	Interval/ Ratio Scale
Variable X	Naturally Dichotomous	Phi Correlation (ϕ)	Coefficient V (v)	Point Biserial Correlation (r_{pb})
	Artificially Dichotomized		Tetrachoric Correlation (r_{tet})	Biserial Correlation (r_b)
	Interval/ Ratio Scale			PPMC (ρ)

3.1. PHI CORRELATION

Phi correlation is a special case of Pearson Product Moment Correlation that gives linear association between variables that are both naturally dichotomous with underlying discreet distribution. It is nonparametric in nature and for a 2x2 contingency table (Table 2), it is calculated as:

$$\phi = \frac{bc - ad}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

Phi correlation is sensitive to marginal distribution of X & Y (table 3) as it assumes that the dichotomous proportions are equal. Phi correlation value is close to one if marginal probabilities are almost equal. Phi correlation gives symmetric percentage difference and is interpreted as percentage of variance in one variable explained by the other [30]. Phi correlation follows Chi-Square distribution as:

$$\chi^2 \ = n\,\Phi^2$$

with degree of freedom = (r – 1) (c – 1) where r & c represent the number of rows and columns respectively.

Table 2: Frequency distribution table used for ϕ and r_{tet}

		Variable Y		
		Y = 0	Y = 1	
Variable X	X = 1	a	b	(a + b)
	X = 0	c	d	(c + d)
		(a + c)	(b + d)	

Table 3: Probability distribution table used for ϕ and r_{tet}

		Variable Y		
		Y = 0	Y = 1	
Variable X	X = 1	π_{11}	π_{12}	π_{X1}
	X = 0	π_{21}	π_{22}	π_{X0}
		π_{Y0}	π_{Y1}	1

3.2. TETRACHORIC CORRELATION

Tetrachoric correlation is a parametric technique that represents degree of linear relationship. It is based on assumption that two variables, X and Y, are obtained by dichotomizing the two underlying interval or ratio scale latent variables, X_L and Y_L respectively, that follow bivariate normal distribution. It measures unobserved correlation between the latent variables and computed as follows:

$$r_{tet} = \frac{bc - ad}{\lambda_X \lambda_Y n^2}$$

where, λ_X is the ordinate (height) of standardized normal distribution for proportion of population obtaining X = 1 and λ_Y is the ordinate (height) of standardized normal distribution for proportion of population obtaining Y = 1. Alternately a computationally simpler formula to calculate tetrachoric correlation coefficient, that can be used, is expressed as [51]:

$$r_{tet} = Cos\left[\frac{180^0\,(\sqrt{ad})}{\sqrt{ad} + \sqrt{bc}}\right]$$

regarding testing the hypothesis, t-test is used (Sheskin, 2003). It is calculated as:

$$t = \frac{r_{tet}}{\sigma_r}$$

where, σ_r is the standard error of tetrachoric correlation.

3.3. BISERIAL CORRELATION

Biserial correlation is used to measure association between a continuous variable Y, and an artificially dichotomous variable X. The dichotomous variable is derived from a continuous variable X_L with normal distribution. It is an estimate of unobserved correlation between the latent X variable and Y. Mathematically it is calculated as [52]:

$$r_b = \frac{(\overline{Y_1} - \overline{Y_0})\,(p_X\ q_X)}{(\lambda) SD_Y}$$

where, p_X is proportion of sample scoring 1 on variable X, q_X is proportion of sample scoring 0 on variable X, $\overline{Y_1}$ is mean of Y for individuals scoring 1 on X, $\overline{Y_0}$ is mean of Y for individuals scoring 0 on X, λ is the ordinate of standardized normal distribution and Sd_Y is standard deviation of entire group. It is further assumed that the split between probabilities p_X and q_X is not large. To test the significance of biserial correlation, z score is calculated as:

$$z = \frac{r_b}{SE}$$

SE is the standard error of biserial correlation which is obtained using the formula:

$$SE = \frac{\sqrt{p_X \; q_X}}{\lambda \sqrt{N}}$$

The biserial correlation is generally used for large sample size and unlike Pearson product moment correlation and other correlation coefficients, the range of r_b is not limited to ± 1 (Garrett, 2005).

3.4. POINT BISERIAL CORRELATION

Point biserial correlation is a special case of Pearson product moment correlation. It gives correlation for two variables, X and Y, where Y is in interval or ratio scale with normal distribution while X is a naturally occurring dichotomous variable. Along with other assumptions that apply to Pearson product moment correlation, such as absence of outliers, it further assumes that Y has equal variance for each category of dichotomous variable. It is calculated as [53]:

$$r_{pb} = \frac{(\bar{Y}_1 - \bar{Y})}{Sd_Y} \sqrt{\frac{p_X}{q_X}}$$

where, Sd_Y is standard deviation of continuous variable Y, p_X is proportion of sample scoring 1 on variable X, q_X is proportion of sample scoring 0 on variable X, $\bar{Y}_1$is mean of Y for individuals scoring 1 on X, $\bar{Y}$is mean of Y. To check the hypothesis that correlation is significantly different from zero t-test is calculated using the formula:

$$t = \frac{r_{pb}\sqrt{N-2}}{\sqrt{1 - {r_{pb}}^2}}$$

with N-2 degrees of freedom and N is the total number of observations. Mathematically, the Pearson Product Moment Correlation can be modified to calculate point biserial correlation (Gradstein, 1986). The Pearson correlation is measured as follows:

$$r = \frac{E[\,(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

where, $\mu_X = P\,(X = 1) = p_X$ and $\sigma_X = \sqrt{p_X(1 - p_X}$. Substituting the mean and standard deviation for dichotomous variables in Pearson correlation formula, the Pearson correlation is transformed to obtain point biserial correlation as:

$$r_{pb} = \frac{E[\,(X - p_X)(Y - \mu_Y)]}{\sigma_Y\sqrt{p_X(1 - p_X)}} = \frac{E\,(XY) - p_X\mu_Y}{\sigma_Y\sqrt{p_X(1 - p_X)}}$$

The relationship between point biserial and biserial can be expressed as follows:

$$r_b = r_{pb}\frac{\sqrt{p_X\, q_X}}{(\lambda)}$$

Numerically, r_b obtained is always greater than r_{pb}.

3.5. CORRELATION COEFFICIENT V

Ulrich and Wirtz [20] propose this correlation coefficient to measures the correlation between a naturally occurring dichotomous and an artificially dichotomized variable, from an underlying latent variable with the normal distribution. The correlation coefficient V is conceptually like biserial correlation. It is assumed that the dichotomous variable X is a Bernoulli variable, that takes the value of 0 and 1 with the probabilities Pr (X= 0) and Pr (X =1) respectively. The variable Y is artificially dichotomized from an underlying latent continuous variable, Y_L, that follows the normal distribution. The marginal probabilities of X & Y used in the following formula is given in table 3:

$$V = \frac{\Delta}{\sqrt{\Delta^2 + \frac{1}{\pi_{1X}(1 - \pi_{1X})}}}$$

where, $\Delta = \varphi^{-1}\left[\frac{\pi_{11}}{\pi_{0X}}\right] - \varphi^{-1}\left[\frac{\pi_{21}}{\pi_{1X}}\right]$ and ϕ^{-1} denotes the inverse of cumulative standard normal distribution. If the conditional mean of $E[Y_L|\ X=1] = \mu_1$ and $E[Y_L|\ X=0] = \mu_0$ are equal, then correlation V is equal to zero. Furthermore, given that the distribution of Y_L is logistic instead of normal, the correlation coefficient can be simplified to the form:

$$V = \frac{\ln\left[\frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}\right]}{\sqrt{\ln\left[\frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}\right]^2 + \frac{2.89}{\pi_{1X}(1-\ \pi_{1X})}}}$$

4. DATA & METHODOLOGY

Monte Carlo procedures are conducted to probe the size, power and bias of selected correlation statistics for samples of size 10, 20, 40, 60, 80, 100, 150 and 200. The null hypothesis of zero correlation is tested against the alternative correlation values ranging from 0.2 to 0.8 with an interval of 0.2 at 1%, 5%, and 10% levels of significance.

The Bernoulli type distribution is considered for naturally occurring dichotomous variables. The alternative space for the artificially dichotomized variables consists of both normal and non-normal alternatives. Following Puth *et. al* [4], the non-normal alternatives are considered from the class of heavy tailed symmetric and asymmetric distributions. A normal tail and asymmetrical distribution cannot be taken since a skewed distribution is also leptokurtic [54]. Moreover, data sets like income, working hours, financial data etc. follow (a)symmetric heavy tailed distributions.

For the point biserial correlation when the continuous variable follows heavy tail distribution, the pre-specified correlation considered are 0.2, 0.4, 0.6 and 0.7 because the feasible correlation upper bound is 0.73 in this regard. Correlation V is undefined if any one of cell frequency is zero. Unlike tetrachoric correlation, zero cell adjustment for correlation V is not proposed by Ulrich and Wirtz [20] for 2x2 table. Correlation coefficient V is zero if [p_{00} / $p_{0.}$] = [p_{10} / $p_{1.}$]. It is also observed that for small sample size, the odds for obtaining zero correlation V are also high. Therefore, for comparison purpose, sample size of 10 and correlation greater than 0.6 are not included in simulations.

5. RESULTS & DISCUSSION

This section presents and discuss the results obtained when the correct correlation coefficient and the Pearson Product Moment Correlation is used for the specified data type. Monte Carlo procedures are called in to compute the size, power and bias of the selected correlation techniques for different sample sizes, correlation values, levels of significance and continuous variables are generated from normal and non-normal distributions.

5.1. NATURALLY DICHOTOMOUS VARIABLES

The Phi correlation, Pearson Product Moment Correlation and the tetrachoric correlation statistics are evaluated over the naturally dichotomous data in terms of their size, power and biases. The phi correlation is a special case of PPMC used for dichotomous data [25,30] and PPMC is also applicable on correlated naturally occurring dichotomous variables.

When evaluating the Type-I error rates of these statistics, tetrachoric correlation statistics exhibits size distortions for samples of size less than 100 (appendix, table 1) and rest of the statistics follow the nominal level of significance. In terms of power comparison, on balance, power increases with the increase in sample size and pre-specified correlation (Fig. 1) with Phi statistic having a slight edge for small samples and low correlation specifications. With low correlation, large samples are required to obtain a reason power for these statistics. Similar trends in power performance are found for 0.01 and 0.1 levels of significance[1].

Table 4 gives the bias in estimates for each test. The general trend is that there is negligible bias for phi and PPMC. However, the tetrachoric statistic shows some biases ranging from 10-21% for different correlation specifications. This may lead to incorrectly assuming a high degree of association among variables and causes researchers to draw incorrect inferences.

Overall, both Pearson and phi correlation perform well under various circumstance. Both give high power, good control over Type I error and insignificant bias in the estimates. No substantial distinction can be made regarding the consequence of using Pearson correlation on naturally dichotomous variables. Tetrachoric correlation, on the other hand, performs poorly.

Fig. 1: Power comparison of ϕ, ρ and r_{tet} ($\alpha = 0.05$) against normal distribution

[1] Power patterns of all selected statistics for other types of data generated from normal distribution are same. That's why results are not reported in the next sections.

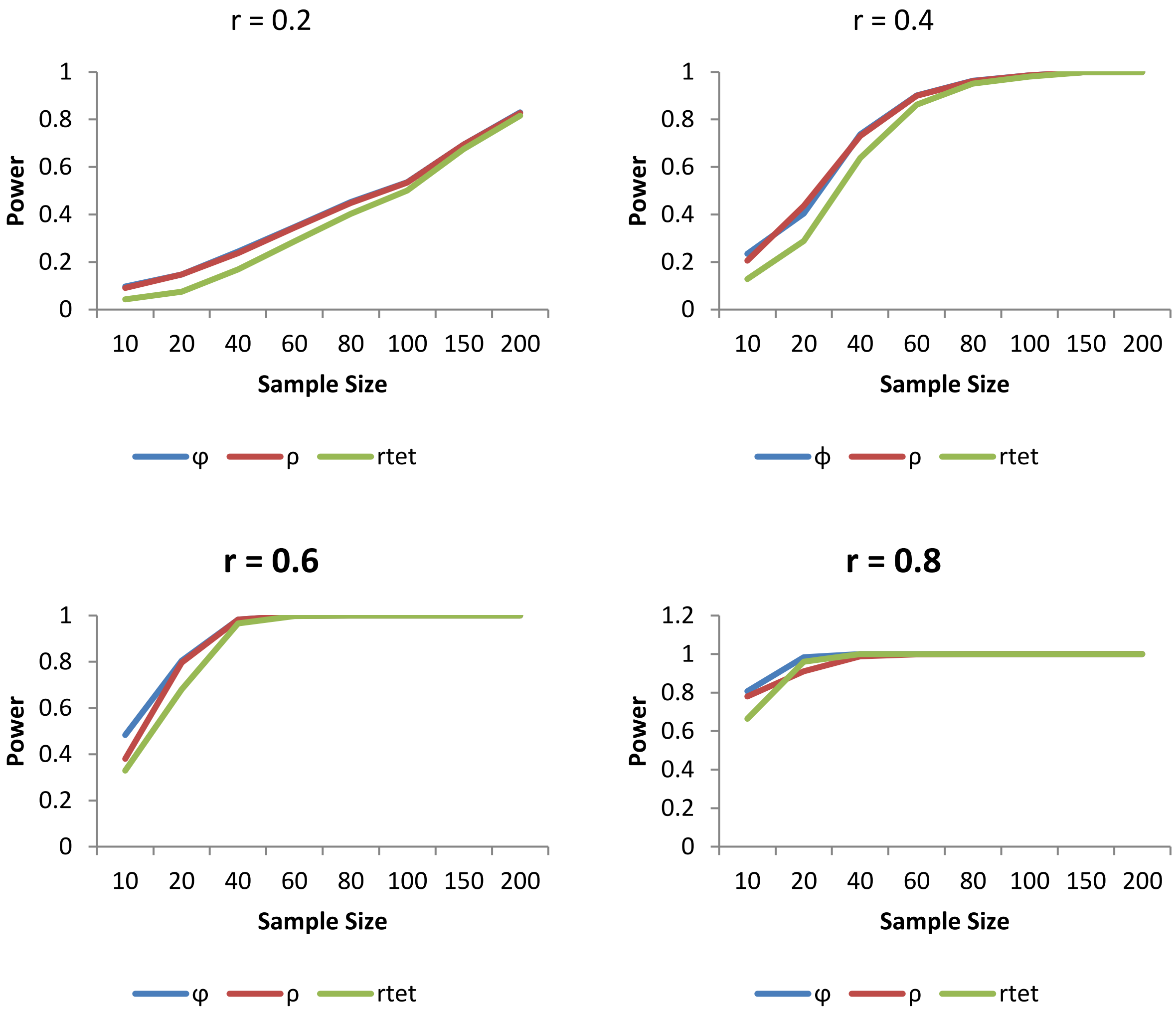


Table 4: Bias in estimate of ϕ, ρ and r_{tet}

Correlations	0.2	0.4	0.6	0.8
Φ	0.002	0.01	0.002	-0.001
ρ	0.001	0.003	0.002	-0.001
r_{tet}	0.108	0.187	0.206	0.149

Peer-reviewed version available at Symmetry 2019, 11; doi:10.3390/sym11070872

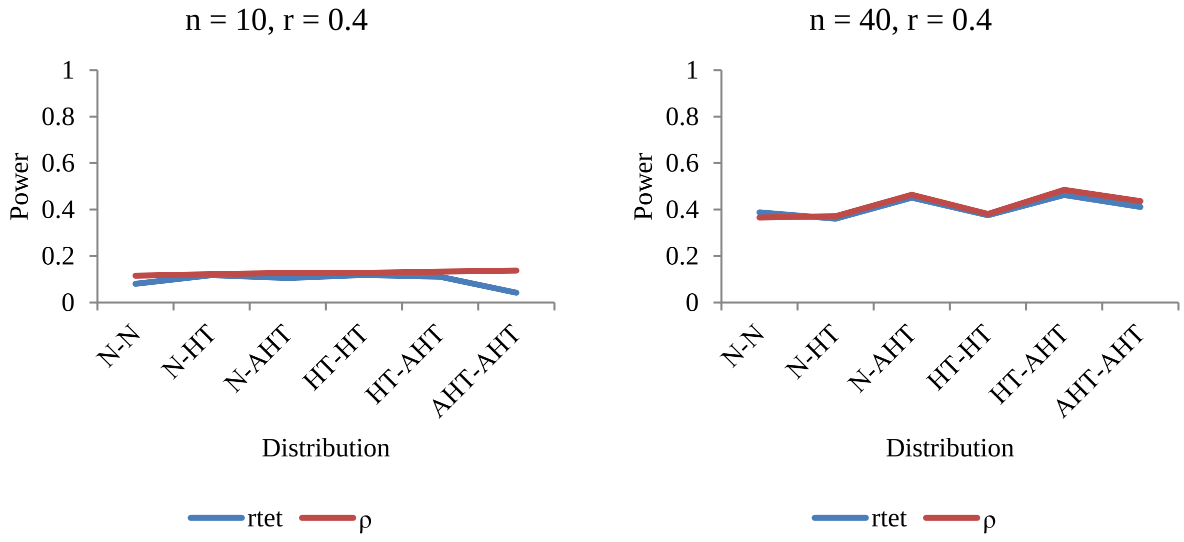
5.2. ARTIFICIALLY DICHOTOMOUS VARIABLES

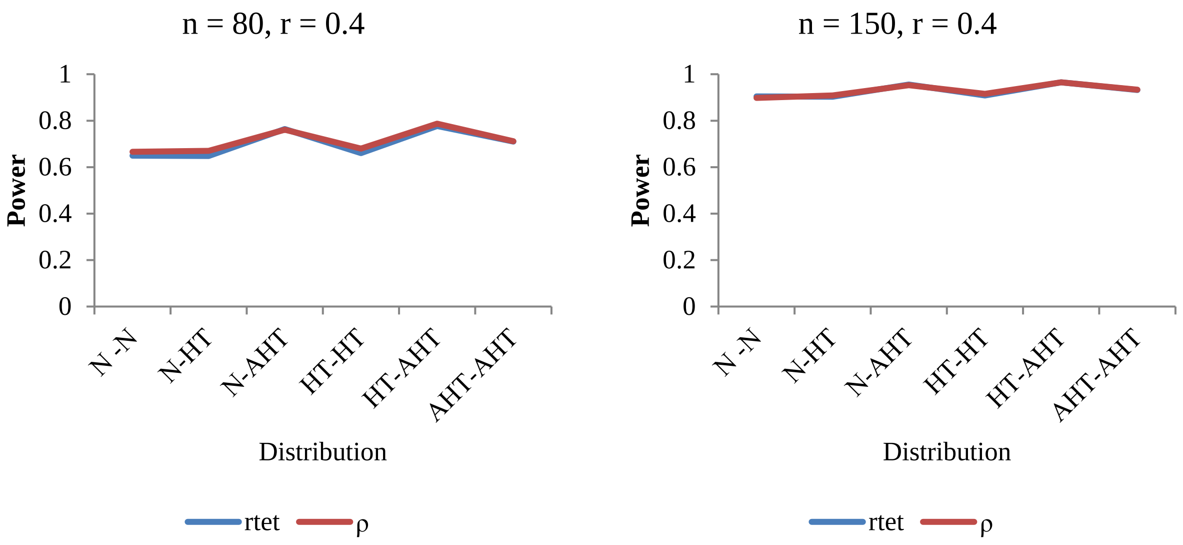
We selected different combinations of (a)symmetric, heavy-tailed (appendix, Table 2) distributions for the two latent variables which are artificially dichotomized. Tetrachoric and PPMC show no size distortions.

Table 5: Bias in estimates of r_{tet} and ρ

Distribution	Correlation							
	0.2		0.4		0.6		0.8	
	r_{tet}	ρ	r_{tet}	ρ	r_{tet}	ρ	r_{tet}	ρ
N – N	-0.007	-0.07	-0.013	-0.14	-0.02	-0.19	-0.024	-0.21
N – HT	-0.004	-0.071	-0.009	-0.135	-0.012	-0.183	-0.013	-0.198
N – AHT	0.015	-0.059	0.04	-0.104	0.077	-0.119	0.078	-0.111
HT – HT	-0.001	-0.068	-0.005	-0.131	-0.009	-0.182	-0.017	-0.202
HT - AHT	0.019	-0.057	0.05	-0.099	0.098	-0.096	0.077	-0.11
AHT – AHT	0.025	-0.057	0.025	-0.117	0.016	-0.166	-0.002	-0.187

Fig.2: Power comparison of r_{tet} and ρ against non-normal distributions





With respect to different distribution types, Fig. 2 provide power comparison between tetrachoric and Pearson correlation for given sample sizes and population correlation equal to 0.4. Both statistics exhibit low power against all type of distributions for small sample sizes. The most damaging combination is the HT-HT for all sample sizes. This might be because of outlying values in case of heavy-tailed distributions.

On balance, there is significant downwards bias in Pearson correlation (Table 5), as it tends to underestimate the true correlation value. The magnitude of bias increases as the population correlation increases. Tetrachoric correlation performs well when the distribution of latent variable is normal or skewed in one or both artificially dichotomized variables. If one of the variables has asymmetric heavy-tailed (AHT) distribution, the tetrachoric correlation shows considerable upward bias in estimate as the population correlation value increases. When both variables follow AHT distribution then the extent of bias is less, since both variables follow same distribution.

In terms of power comparison, it is hard to choose between the two statistics. Both test exhibit low power for small and moderate sample sizes and good power for large sample sizes. However, tetrachoric statistic is less biased as compared to Pearson's statistic.

5.3. ARTIFICIALLY DICHOTOMOUS & CONTINUOUS VARIABLES

When one variable is artificially dichotomized and other one is a continuous variable, we apply biserial correlations statistics. The performance of biserial statistic in terms of size, power and bias is contrasted with the Pearson correlation statistic. Both statistics don't suffer from size distortions for medium to large sample sizes. However, the biserial correlation statistic exhibits size distortions for small sample size (appendix, Table 3).

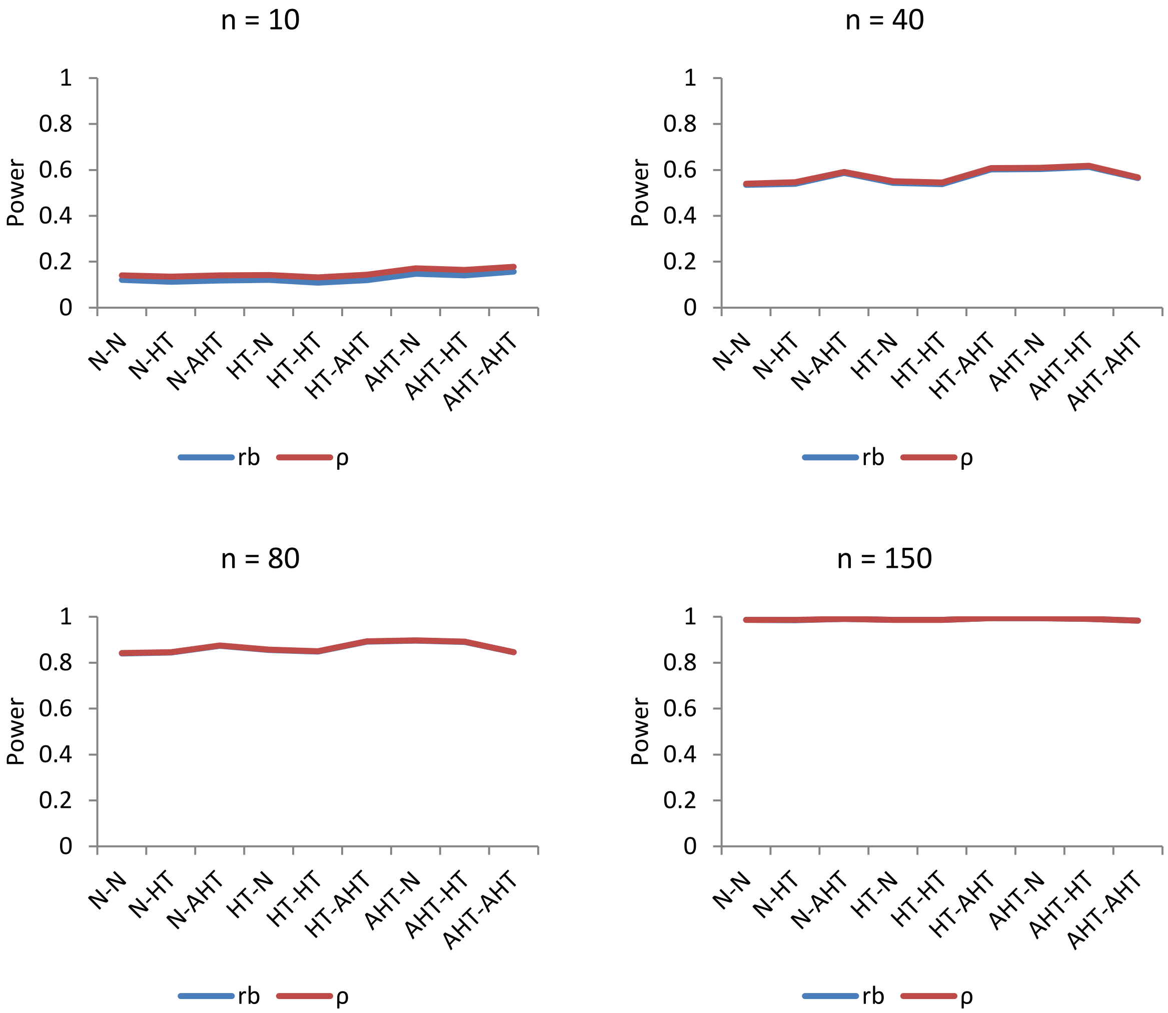
While comparing the biserial and Pearson correlation statistics in terms of power, it is hard to choose between the two statistics. However, both correlation measures have shown low power for small to medium sample sizes. Power goes to almost unity for sample sizes greater than 100. The underlying distributions have no significant impact on their power properties.

Table 6: Bias in estimates of r_b and ρ

Distribution	Correlation							
	0.2		0.4		0.6		0.8	
	r_b	ρ	r_b	ρ	r_b	ρ	r_b	ρ
N – N	-0.003	-0.041	-0.007	-0.082	-0.008	-0.123	-0.009	-0.161
N – HT	-0.005	-0.043	-0.009	-0.084	-0.018	-0.131	-0.035	-0.183
N – AHT	-0.001	-0.039	0.008	-0.071	0.031	-0.091	0.068	-0.099
HT – N	-0.002	-0.04	-0.001	-0.078	0.001	-0.116	0.003	-0.152
HT – HT	-0.002	-0.04	-0.006	-0.082	-0.017	-0.129	-0.039	-0.186
HT – AHT	0.002	-0.037	0.016	-0.064	0.054	-0.072	0.064	-0.103
AHT – N	0.016	-0.028	0.033	-0.054	0.056	-0.076	0.072	-0.103
AHT – HT	0.016	-0.027	0.031	-0.056	0.041	-0.088	-0.013	-0.17
AHT – AHT	0.008	-0.034	0.009	-0.073	0.009	-0.113	0.013	-0.161

On balance, in comparison to biserial correlation, Pearson correlation underestimates the true correlation and the bias in estimate increases as the population correlation increases. Biserial correlation overestimates the correlation slightly when one of the continuous variables follows AHT distribution and population correlation is large. Overall, biserial gives better performance, and is recommended, when purpose is to find correlation between artificially dichotomous variable and continuous variable.

Fig. 3: Power comparison of r_b and ρ



5.4. NATURALLY DICHOTOMOUS & CONTINUOUS VARIABLES

We compare the performance of point biserial correlation with Pearson correlation statistic for naturally dichotomous and continuous variables. On balance, size of these tests is close to the nominal level of significance, however, point biserial does not provide a good type-I error control in small samples (appendix, Table 4).

Fig. 4: Power comparison of r_{pb} and ρ

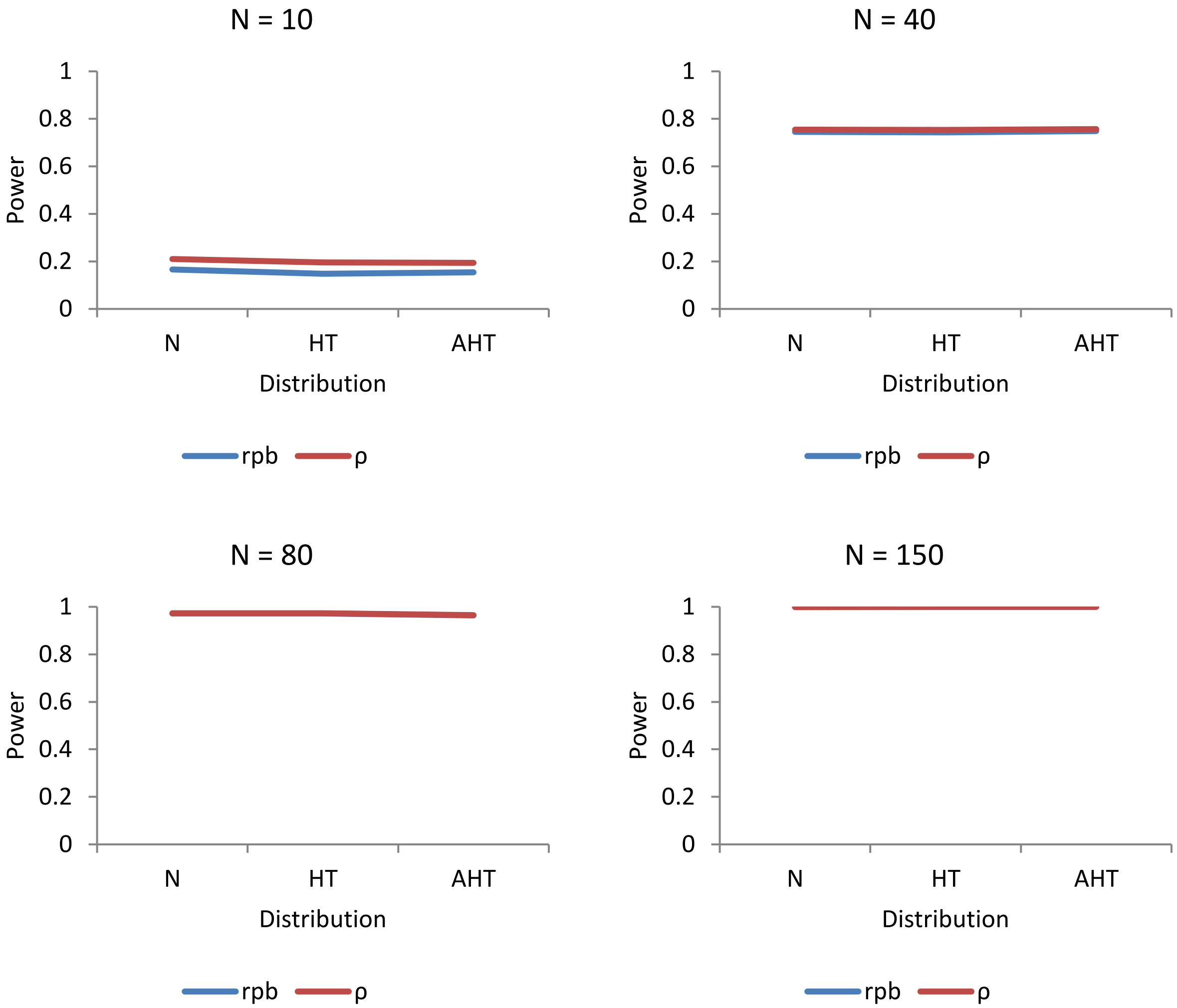


Figure 4 provides the power comparison of the correlation statistics when the underlying distribution is non-normal. Power is low for small samples and low population correlations and increases with the increase in both parameters of evaluations. our results corroborate with the findings in Akkartal [48].

At first, the value of the correlation matters for researchers. Table 7 shows bias in estimate under different normality conditions. Both point biserial and Pearson correlation show negligible bias in the estimate. It is hard to make any distinction between the two statistics in terms of size, power and bias estimates.

Table 7: Bias in estimates of r_{pb} and ρ

Distribution	Correlation									
	0.2		0.4		0.6		0.7		0.8	
	r_{pb}	ρ	r_{pb}	ρ	r_{pb}	ρ	r_{pb}	ρ	r_{pb}	ρ
N – N	-0.005	-0.002	-0.007	-0.001	-0.009	0	-	-	-0.011	0
N – HT	-0.003	0	-0.011	-0.006	-0.026	-0.017	-0.038	-0.028	-	-
N – AHT	-0.012	-0.009	-0.014	-0.008	0.009	0.019	-	-	-	-

5.5. NATURALLY DICHOTOMOUS & ARTIFICIAL DICHOTOMOUS VARIABLES

Regarding correlation measure between naturally dichotomous and artificially dichotomous variables, Ulrich and Wirtz (2004) propose a measure of correlation, *V*. However, literature does not provide any guidance regarding the computation of its power. Therefore, the comparison of this correlation statistic with Pearson, phi and tetrachoric correlation measures is based on bias in estimation only.

Table 8: Bias in estimate for *V*, P, r_{tet} & ϕ for given distribution of Y_L

Distribution of Y_L		Bias			
		V	P	r_{tet}	φ
N	0.2	-0.005	-0.039	0.048	-0.039
	0.4	-0.011	-0.066	0.096	-0.066
	0.6	-0.012	-0.058	0.146	-0.058
HT	0.2	-0.003	-0.037	0.051	-0.037
	0.4	-0.005	-0.06	0.103	-0.06
	0.6	-0.005	-0.048	0.156	-0.048
AHT	0.2	0.005	-0.033	0.062	-0.033
	0.4	0.024	-0.04	0.143	-0.04
	0.6	0.07	0.021	0.247	0.021

On balance, Pearson and Phi correlation measures underestimate population correlation. Tetrachoric, on the other hand overestimates the correlation considerably (Table 8). Tetrachoric correlation overestimates for higher values of pre-specified correlation and against non-normal distributions. In contrast to Pearson, Phi and Tetrachoric correlation measures, the *V* statistics performs well except for higher pre-specified correlation values which corroborates with the findings in Ulrich and Wirtz (2004).

6. CONCLUSION

This study provides the comparison of correlation techniques for nominal data in terms of size, power and bias. The correlation techniques considered for nominal data types are Phi correlation, tetrachoric correlation, point biserial correlation, biserial correlation and correlation coefficient *V*. Overall, power increases with the increase in sample size and pre-specified correlation values for all statistics under consideration. Followings are the key findings of the study.

1. For naturally dichotomous variables, the Phi and Pearson correlation statistics performs equally well in terms of size, power, and bias. However, the Tetrachoric correlation measure does not have good control over type-I error and tends to overestimate the correlation values.
2. When both variables are artificially dichotomized, the Tetrachoric and Pearson correlation measures exhibit similar power properties. The Tetrachoric statistic has an edge in terms of bias to Pearson correlation statistic.
3. If one variable is continuous and other is artificially dichotomized, the Biserial correlation measure turns out to be less biased as compared to Pearson statistic although both statistics exhibit similar power and size properties.
4. If one variable is continuous and other is naturally dichotomized, it is hard to distinguish between the Point Biserial and Pearson correlation measures as both statistics performs equally well in terms of size, power and bias.
5. Finally, if one variable is naturally dichotomous and other is artificially dichotomized, correlation coefficient *V* is compared with Pearson, Phi and Tetrachoric correlation techniques in terms of bias in estimate. The results indicate that the Tetrachoric statistic considerably overestimates the correlation value against non-normal distributions. Pearson and Phi correlation slightly underestimate the correlation value. In contrast, the correlation statistic *V* perform well.

REFERENCES

1. Choi, J., Peters, M., & Mueller, R. O. (2010). Correlational analysis of ordinal data: from Pearson's r to Bayesian polychoric correlation. *Asia Pacific education review*, *11*(4), 459-466.
2. Onwuegbuzie, A. J., & Daniel, L. G. (1999). Uses and misuses of the correlation coefficient.
3. Edgell, S. E., & Noon, S. M. (1984). Effect of violation of normality on the t test of the correlation coefficient. *Psychological bulletin*, *95*(3), 576.
4. Puth, M. T., Neuhäuser, M., & Ruxton, G. D. (2014). Effective use of Pearson's product–moment correlation coefficient. *Animal behaviour*, *93*, 183-189.
5. Puth, M. T., Neuhäuser, M., & Ruxton, G. D. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*, *102*, 77-84.
6. 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), 87-93.
7. Tuğran, E., Kocak, M., Mirtagioğlu, H., Yiğit, S., & Mendes, M. (2015). A simulation based comparison of correlation coefficients with regard to type I error rate and power. *Journal of Data Analysis and Information Processing*, *3*(03), 87.
8. Kim, S. Y. (2010). Do Asian values exist? Empirical tests of the four dimensions of Asian values. *Journal of East Asian Studies*, *10*(2), 315-344.
9. Fortanier, F., Kolk, A., & Pinkse, J. (2011). Harmonization in CSR reporting. *Management International Review*, *51*(5), 665.
10. Heikkila, T., & Schlager, E. C. (2012). Addressing the Issues: The Choice of Environmental Conflict-Resolution Venues in the United States. *American Journal of Political Science*, *56*(4), 774-786.
11. Gius, M. (2013). The effects of merit pay on teacher job satisfaction. *Applied Economics*, *45*(31), 4443-4451.
12. Bahmani-Oskooee, M., Hegerty, S. W., & Xu, J. (2013). Exchange-rate volatility and US–Hong Kong industry trade: is there evidence of a 'third country'effect?. *Applied Economics*, *45*(18), 2629-2651.
13. Barrett, D. E., Katsiyannis, A., Zhang, D., & Zhang, D. (2014). A structural equation modeling analysis of influences on juvenile delinquency. *Behavioral disorders*, *39*(3), 113-127.
14. Rai, R. K. (2015). Factors associated with nutritional status among adult women in urban India, 1998-2006. *Asia Pacific Journal of Public Health*, *27*(2), NP1241-NP1252.
15. Furtado, D. (2015). Can immigrants help women "have it all"? Immigrant labor and women's joint fertility and labor supply decisions. *IZA Journal of Migration*, *4*(1), 19.
16. Boas, F. (1909). Determination of the coefficient of correlation. *Science*, *29*(751), 823-824.
17. Pearson, K. (1900). I. Mathematical contributions to the theory of evolution. —VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *195*(262-273), 1-47.
18. Richardson, M. W., & Stalnaker, J. M. (1933). A note on the use of bi-serial r in test research. *The Journal of General Psychology*, *8*(2), 463-465.
19. Pearson, K. (1909). On the New Method for Determining the Correlation Between a Measure Character A, and a Character B. *Biometrika*, *7*, 96 - 105

20. Ulrich, R., & Wirtz, M. (2004). On the correlation of a naturally and an artificially dichotomized variable. *British Journal of Mathematical and Statistical Psychology*, *57*(2), 235-251.
21. Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. *Journal of Educational and Behavioral Statistics*, *30*(2), 213-225.
22. Ekström, J. (2011). The phi-coefficient, the tetrachoric correlation coefficient, and the Pearson-Yule Debate.
23. Vercruyssen, M., & Hendrick, H. W. (2011). *Behavioral research and analysis: an introduction to statistics within the context of experimental design*. CRC Press.
24. Farrington, D. P., & Loeber, R. (1989). Relative improvement over chance (RIOC) and phi as measures of predictive efficiency and strength of association in 2× 2 tables. *Journal of Quantitative Criminology*, *5*(3), 201-213.
25. Davenport Jr, E. C., & El-Sanhurry, N. A. (1991). Phi/phimax: review and synthesis. *Educational and psychological measurement*, *51*(4), 821-828.
26. Liu, R. (1980). A note on phi-coefficient comparison. *Research in Higher Education*, *13*(1), 3-8.
27. Warrens, M. J. (2008). On association coefficients for 2× 2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, *73*(4), 777.
28. Demirtas, H. (2016). A note on the relationship between the phi coefficient and the tetrachoric correlation under nonnormal underlying distributions. *The American Statistician*, *70*(2), 143-148.
29. Lev, J. (1949). The point biserial coefficient of correlation. *The Annals of Mathematical Statistics*, *20*(1), 125-126.
30. Chen, P. Y., Smithson, M., & Popovich, P. M. (2002). *Correlation: Parametric and nonparametric measures* (No. 139). Sage.
31. Glass, G., & Hopkins, K. (1996). Statistical methods in education and psychology. *Psyccritiques*, *41*(12).
32. Demirtas, H., & Hedeker, D. (2016). Computing the point-biserial correlation under any underlying continuous distribution. *Communications in Statistics-Simulation and Computation*, *45*(8), 2744-2751.
33. Tate, R. F. (1954). Correlation between a discrete and a continuous variable. Point-biserial correlation. *The Annals of mathematical statistics*, *25*(3), 603-607.
34. Tate, R. F. (1955). Applications of correlation models for biserial data. *Journal of the American Statistical Association*, *50*(272), 1078-1095.
35. Gupta, S. D. (1960). Point biserial correlation coefficient and its generalization. *Psychometrika*, *25*(4), 393-408.
36. Becker, G. (1986). Correcting the point-biserial correlation for attenuation owing to unequal sample size. *The Journal of Experimental Education*, *55*(1), 5-8.
37. Gradstein, M. (1986). Maximal Correlation Between Normal and Dichotomous Variables. *Journal of Educational and Behavioral Statistics, 11*(4), 259 - 261.
38. Terrell, C. D. (1982a). Significance Tables for the Biserial and the Point Biserial. *Educational and Psychological Measurement*, 42(4), 975 - 981.
39. Terrell, C. D. (1982b). Tables for Converting the Point Biserial to the Biserial. *Educational and Psychological Measurement*, 42(4), 983 - 986.
40. Guilford, J. P., & Lyons, T. C. (1942). On determining the reliability and significance of a tetrachoric coefficient of correlation. *Psychometrika*, *7*(4), 243-249.

41. Brown, M. B. (1977). Algorithm AS 116: The Tetrachoric Correlation and its Asymptotic Standard Error. *Journal of the Royal Statistical Society*, 26(3), 343 - 351.
42. Brown, M. B., & Benedetti, J. K. (1977). On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, *42*(3), 347-355.
43. Tate, R. F. (1950). *The biserial and point correlation coefficients*. North Carolina State University. Dept. of Statistics.

44. Martin, W. S. (1978). Effects of Scaling on the Correlation Coefficient: Additional Considerations. *Journal of Marketing Research*, *15*(2), 304 - 308.

45. Fowler, R. L. (1987). Power and robustness in product-moment correlation. *Applied Psychological Measurement*, *11*(4), 419-428.
46. Greer, T., Dunlap, W. P., & Beatty, G. O. (2003). A Monte Carlo evaluation of the tetrachoric correlation coefficient. *Educational and Psychological Measurement*, *63*(6), 931-950.
47. Brown, M. B., & Benedetti, J. K. (1977). On the Mean and Variance of the Tetrachoric Correlation Coefficient. *Psychometrika*, *43*(3), 347 - 355.
48. Akkartal, E. (2016). A Monte Carlo Simulation Study to Investigate the Performance of Point Biserial Correlation. *Sylwan*, 160(12), 44 - 53.
49. Cohen, J. 8: Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences. *Hillsdale, NJ: Lawrence Erlbaum*.
50. Myers, J. L., Well, A. D., & Lorch Jr., R. F. (2013). *Research design and Statistical Analysis* (3 ed.). Routledge.
51. Sheskin, D. J. (2003). *Handbook of Parametric and Nonparametric Procedures* (3 ed.). CRS Press.
52. Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using R. *Sage publications*.
53. Anderson, J. A. (1994). Point biserial correlation. *Stata Technical Bulletin*, *3*(17).
54. Hopkins, K. D., & Weeks, D. L. (1990). Tests for normality and measures of skewness and kurtosis: Their place in research reporting. *Educational and Psychological Measurement*, *50*(4), 717-729.

Appendix

Table 1: Estimated Type I error for ϕ, ρ and r_{tet}

				$\alpha = 0.05$				
n	10	20	40	60	80	100	150	200
Φ	0.052	0.05	0.05	0.05	0.052	0.059	0.05	0.054
ρ	0.049	0.05	0.05	0.05	0.051	0.058	0.049	0.053
r_{tet}	0.021	0.02	0.02	0.031	0.039	0.047	0.043	0.049
				$\alpha = 0.01$				
Φ	0.013	0.012	0.009	0.01	0.009	0.012	0.011	0.009
ρ	0.018	0.012	0.01	0.01	0.009	0.013	0.011	0.009
r_{tet}	0.003	0.005	0.006	0.006	0.006	0.009	0.009	0.007
				$\alpha = 0.1$				
Φ	0.135	0.118	0.097	0.096	0.102	0.097	0.101	0.101
ρ	0.091	0.107	0.096	0.094	0.099	0.097	0.097	0.101
r_{tet}	0.033	0.056	0.056	0.063	0.07	0.069	0.077	0.085

Table 2: Estimated Type I error for rtet and ρ at $\alpha = 0.05$

Distribution	Correlation Statistic	Sample Size							
		10	20	40	60	80	100	150	200
N – N	r_{tet}	0.046	0.053	0.051	0.050	0.048	0.044	0.046	0.051
	ρ	0.050	0.061	0.047	0.056	0.051	0.054	0.051	0.053
N – HT	r_{tet}	0.046	0.047	0.049	0.055	0.048	0.050	0.047	0.049
	ρ	0.049	0.054	0.049	0.053	0.053	0.050	0.052	0.053
N – AHT	r_{tet}	0.040	0.051	0.048	0.049	0.049	0.053	0.051	0.049
	ρ	0.044	0.053	0.049	0.033	0.054	0.050	0.053	0.048
HT - HT	r_{tet}	0.046	0.049	0.046	0.056	0.048	0.049	0.048	0.048
	ρ	0.049	0.054	0.049	0.053	0.053	0.059	0.053	0.052
HT - AHT	r_{tet}	0.040	0.051	0.046	0.048	0.048	0.051	0.053	0.049
	ρ	0.049	0.052	0.053	0.050	0.051	0.050	0.054	0.052
AHT – AHT	r_{tet}	0.032	0.053	0.046	0.047	0.050	0.044	0.050	0.044
	ρ	0.045	0.051	0.048	0.051	0.538	0.053	0.051	0.048

Table 3: Estimated Type I error for rb and ρ at α = 0.05

Distribution		Sample Size							
		10	20	40	60	80	100	150	200
N - N	r_b	0.039	0.045	0.049	0.048	0.052	0.049	0.051	0.051
	ρ	0.047	0.049	0.051	0.048	0.052	0.049	0.051	0.051
N - HT	r_b	0.039	0.045	0.049	0.048	0.052	0.049	0.051	0.051
	ρ	0.047	0.049	0.051	0.048	0.052	0.049	0.051	0.051
N - AHT	r_b	0.032	0.041	0.045	0.048	0.049	0.049	0.050	0.048
	ρ	0.039	0.043	0.046	0.049	0.050	0.049	0.050	0.048
HT - N	r_b	0.037	0.050	0.048	0.050	0.054	0.047	0.047	0.049
	ρ	0.045	0.053	0.049	0.050	0.054	0.048	0.048	0.049
HT - HT	r_b	0.035	0.042	0.047	0.047	0.044	0.049	0.046	0.051
	ρ	0.044	0.044	0.048	0.047	0.045	0.049	0.046	0.051
HT - AHT	r_b	0.039	0.042	0.046	0.047	0.049	0.046	0.047	0.050
	ρ	0.047	0.047	0.047	0.047	0.050	0.047	0.048	0.050
AHT – N	r_b	0.039	0.047	0.050	0.053	0.051	0.050	0.049	0.051
	ρ	0.049	0.050	0.051	0.054	0.051	0.050	0.049	0.051
AHT - HT	r_b	0.037	0.048	0.045	0.048	0.047	0.049	0.047	0.049
	ρ	0.047	0.05	0.047	0.048	0.047	0.049	0.048	0.049
AHT - AHT	r_b	0.039	0.042	0.042	0.050	0.051	0.050	0.046	0.052
	ρ	0.047	0.044	0.043	0.051	0.052	0.051	0.046	0.052

Table 4: Estimated Type I error rates for r_{pb} and ρ at α = 0.05

Distribution	Correlation	Sample Size							
		10	20	40	60	80	100	150	200
N – N	r_{pb}	0.037	0.045	0.047	0.048	0.049	0.051	0.052	0.048
	ρ	0.053	0.051	0.051	0.049	0.05	0.053	0.052	0.049
N – HT	r_{pb}	0.034	0.043	0.042	0.049	0.046	0.048	0.047	0.049
	ρ	0.049	0.049	0.046	0.051	0.047	0.05	0.047	0.05
N - AHT	r_{pb}	0.032	0.04	0.044	0.047	0.05	0.048	0.048	0.05
	ρ	0.045	0.047	0.046	0.048	0.052	0.049	0.049	0.05