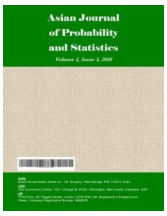

Asian Journal of Probability and Statistics

2(3): 1-12, 2018; Article no.AJPAS.45938

Methods for Determining the Tetrachoric Correlation Coefficient for Binary Variables

E. F. El-Hashash[1*] and K. M. El-Absy[2]

[1]Department of Agronomy, Faculty of Agriculture, Al-Azhar University, Cairo, Egypt.
[2]Department of Biology, Faculty of Science, Tabuk University, Tayma Branch, Tabuk, Saudi Arabia.

Authors' contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2018/v2i328782
Editor(s):
(1) Dr. Jiteng Jia, School of Mathematics and Statistics, Xidian University, China.
Reviewers:
(1) Alexandre Ripamonti, University of Sao Paulo, Brazil.
(2) Eric S. Hall, USA.
Complete Peer review History: http://www.sciencedomain.org/review-history/28019

Original Research Article

Received: 23 September 2018
Accepted: 07 December 2018
Published: 31 December 2018

Abstract

The tetrachoric correlation coefficient (r_t) is a special case of the statistical covariation between two variables measured on a dichotomous scale, but assuming an underlying bivariate normal distribution. Our goal was to provide an analysis of seven different methods used to calculate r_t. The r_t approximation was then used to derive its standard error and its associated confidence interval. Computation of r_t is not straightforward and is usually not available in standard statistical packages. This paper introduces seven methods for computing the r_t value and three methods used to provide the standard error estimation $\{SE(r_t)\}$. These methods were illustrated using data from questionnaires that were used to evaluate public awareness regarding Electronic Waste hazards. The different algorithmic/mathematical methods used to estimate r_t and $SE(r_t)$ yielded values that were equal to (or very close to) each other and the estimates obtained from SAS statistical analysis software. Method 6 and Method 1 used to estimate r_t and $SE(r_t)$ work very well, the equations are easy to understand, are computationally simple and are ideally suited for use. Additionally, the width of the confidence intervals for these methods are equal to (or closely approximates) the widths calculated by the SAS statistical analysis computer program.

Keywords: Methods; tetrachoric correlation coefficient; standard error; binary variables.

**Corresponding author: E-mail: dressamelhashash@yahoo.com;*

1 Introduction

Measures of association for dichotomous variables are an area that has been studied from the very infancy of modern statistics [1]. One of the first scholars to treat the subject was Karl Pearson, one of the fathers of modern statistics [1]. In the 7th article in the seminal series Mathematical contributions to the theory of evolution, Pearson [2] proposed what later became known as the tetrachoric correlation coefficient (r_t).

The r_t is one of the oldest and most interesting bivariate measures of correlation [3]. The r_t is a product–moment correlation between two unobserved quantitative (continuous) variables that have each been measured on a dichotomous scale [2]. The r_t assumes that the two variables under study are essentially continuous and would be normally distributed if it were possible to obtain scores or exact measures and thus be able to classify both variables into frequency distributions [4]. The r_t is, therefore, an estimate of the product–moment correlation that would have been obtained with the underlying continuous variables if its joint distribution were bivariate normal [5].

The fundamental idea of the r_t is to consider the 2 × 2 contingency table as a double dichotomization of a bivariate standard normal distribution, and then solve for the parameter such that the volumes of the dichotomized bivariate standard normal distribution equal the joint probabilities of the contingency table [1]. The r_t is then defined as that parameter, which, of course, corresponds to the linear correlation of the bivariate normal distribution [1].

While the r_t is typically used to measure the correlation between two independent dichotomous variables, it also can be used to assess the reliability of a single rater when two raters independently rate n objects on a dichotomous scale [6]. In addition, the r_t is often used to measure rater agreement and is preferred by some researchers over Cohen's kappa for this purpose [7]. When assessing agreement between experts, it is important to distinguish between disagreements that can and cannot be explained by different placing of the boundaries between categories. Cohen's kappa statistic is affected by both types of disagreement, and r_t only by the second type [8].

An asymptotic approximation to the standard error $SE(r_t)$ of the r_t was given by Pearson [9]. However, the accuracy and, therefore, usefulness of Pearson's standard error has repeatedly been called into question [7]. For example, Kendall and Stuart [10] noted that the sampling distribution and the standard error of the r_t are not known with any precision, and they noted further that it is not known for what sample size the standard error may safely be used.

Because of the extensive calculations necessary to compute the r_t, it has not been a popular statistic, despite its usefulness [7]. With the advent of high speed computing, the r_t has seen a resurrection in fields such as psychology, psychopathology, radiology, and genetics [11]. If you have the original data, don't bother dichotomizing them, because r_t has an efficiency of 0.40 compared with the efficient Pearson correlation estimate [12].

The r_t value may be a more appropriate measure of reliability for 2 × 2 contingency tables [3]. It may be preferable to factor analyze r_t rather than phi coefficients for a set of dichotomously scored items because differences in item difficulties will introduce additional factors when phi coefficients are factor analyzed [13]. The r_t also may be preferred to the popular kappa coefficient in applications where a consistency measure of reliability is preferred to an agreement measure of reliability [3].

The purpose of this paper is to provide methods used to estimate r_t from 2 × 2 contingency tables. The r_t approximation is then used to derive simple methods of determining $SE(r_t)$ and its confidence interval. To illustrate the methodology, questionnaire data for E-waste management practices, was used.

2 Materials and Methods

2.1 Participants and questionnaires

In this study, the questionnaires were used to evaluate public awareness regarding Electronic Waste hazards, to identify the status of E-waste management practices, determine the effect of E-waste on humans and environment, and the disposal of E-waste. The questionnaires contained a short description of this study and the intended use of collected data, which was also distributed to homeowners by hand. This created room for 'one-on-one' interaction with the respondents. The data was collected through the distribution of 200 well-structured questionnaires. The questionnaire is shown in Table 1.

Table 1. Questionnaire on assessment of E-waste management during 2018

No.	Questions
Q1	Do you know what electronic wastes are? Yes ( ) – No ( )
Q2	Do you use these household electronics frequently? Yes ( ) – No ( )
Q3	Do you recover any of the electronic equipment/components from waste? Yes ( ) – No ( )
Q4	Are you aware of any health risks associated with electronic wastes? Yes ( ) – No ( )
Q5	Do you know that some components of electronic devices contain toxic/hazardous materials? Yes ( ) – No ( )

2.2 Tetrachoric correlation coefficient (r_t)

The 2x2 contingency table of frequencies below represents schematically the arrangement of a four-element table. Both variables, X and Y are classified into two categories (Table 2), here f_{ij} *(i, j=0,1)* is cell frequency. P_X and P_Y are marginal probabilities corresponding to $\frac{f_{00}+f_{10}}{N}$ and $\frac{f_{00}+f_{01}}{N}$, respectively [14].

Table 2. Elements of the 2x2 contingency table

		X		
		X_0	X_1	
Y	Y_0	f_{00}	f_{01}	P_Y
	Y_1	f_{10}	f_{11}	P_Y
		P_X	P_X	N

The table entries f_{00} and f_{11} are the cluster of frequencies into the cells displayed in Table 2, and values in these two cells indicate close agreement and positive correlation. Clusters of entries in f_{01} and f_{10} indicate disagreement and negative correlation. Equal numbers of frequencies in each of the 4 cells means that there is no relationship and the correlation value is zero [4].

For marginal probabilities that satisfy $0 < P_X, P_Y < 1$, and for joint probabilities f_{00} such that $\max(P_X + P_Y - 1,0 < f_{00} < \min(P_X, P_Y))$, the tetrachoric correlation coefficient, introduced by Pearson [2], is defined as the solution r_t by to the integral equation (1):

$$f_{00} = \int_{\Phi^{-1}(1-p_X)}^{\infty} \int_{\Phi^{-1}(1-p_Y)}^{\infty} \phi_2(X, Y, \rho_{tet}) dX dY$$

where $\Phi(x)$ is the standard normal distribution function and $\phi_2(X, Y, \rho_{tet})$ is the bivariate standard normal density function [15]. If $f_{00} = \max(P_X + P_Y - 1,0)$ or $f_{00} = \min(P_X, P_Y)$, then the r_t, is defined to be -1 or 1 respectively [1]. If the inequality $0 < P_X, P_Y < 1$ does not hold, then any value of r_t will satisfy above

equation. However, from the perspective of presuming statistical independence until evidence of dependence is found, r_t is here defined to be zero. The full equation above for r_t is algebraically complex and requires the solution of a quadratic equation to compute r_t [4]. Fortunately, there are several useful approximations to r_t which •are sufficiently accurate for most purposes [4]; hence, the long formula will not be reproduced here.

2.3 Methods used to determine r_t

Approximation methods for determining r_t are provided, with their equations, in Table 3.

Table 3. Summary of equations for the seven methods used to determine r_t.

Methods	Governing equation	Authors/Users/Developers
1	$r_t = cos\left(\frac{180^{\circ} x\sqrt{f_{01}f_{10}}}{\sqrt{f_{00}f_{11}}+\sqrt{f_{01}f_{10}}}\right)$	Garrett and Woodworth [4]
2	$r_t = \frac{f_{00}f_{11}}{f_{01}f_{10}}$	Garrett and Woodworth [4]
3	$\rho = \frac{\alpha-1}{\alpha+1}$	Edwards and Edwards [16]
4	$\rho = cos\left(\frac{\pi}{1+\omega^c}\right)$	Bonett and Price [3]
5	$\hat{\rho} = cos\left(\frac{\pi}{1+\hat{\omega}^{\hat{c}}}\right)$	Bonett and Price [3]
6	$\rho = \cos(\pi/\delta)$	Unknown
7	$r_{tet} = \frac{f_{01}f_{10}-f_{00}f_{11}}{\lambda_X\lambda_Y N^2}$	Chen and Popovich [17]

where:

r_t, ρ, r_{tet} and $\hat{\rho}$: the tetrachoric correlation coefficient; f_{00}, f_{01}, f_{10} and f_{11}: are cell frequencies; $\alpha = \left(\frac{f_{00}f_{11}}{f_{01}f_{10}}\right)^{\pi/4}$; $\pi = 3.14$; $N = f_{00}+f_{01}+f_{10}+f_{11}$; $\omega = \frac{f_{00}f_{11}}{f_{01}f_{10}}$; $c = \left[1-\frac{|p_{1+}-p_{+1}|}{5}-(0.5-p_{min})^2\right]/2$; $p_{1+} = (f_{00}+f_{01})/N$; $p_{+1} = (f_{00}+f_{10})/N$; p_{min}: the smallest marginal proportion; where $\widehat{\omega}$ and $\hat{c}$ are computed as above (ω and c) from sample proportions after 0.5 has been added to each cell frequency; $\delta = 1+\sqrt{\frac{f_{00}f_{11}}{f_{01}f_{10}}}$; λ_X: the ordinate of the standardized normal distribution at $\frac{f_{00}+f_{01}}{N}$, proportion of subjects obtaining X=1; λ_Y: the ordinate of the standardized normal distribution at $\frac{f_{01}+f_{11}}{N}$, proportion of subjects obtaining Y=1

In Method 1, the value of r_t is read directly from the table which gives the cosines of angles from 0° to 90° from Garrett and Woodworth [4]. While, entering the estimated values table of r_t corresponding to values of the ratio, $f_{00}f_{11}/f_{01}f_{10}$ in Method 2, we read r_t directly from Davidoff and Goheen [18].

2.4 Standard error of r_t:

The accuracy of estimators of association for parameters is characterized by the standard errors (SE) of their sampling distributions [19]. The SE of r_t is mathematically complex and is too long to be useful practically [4]. The simple approximate value of standard error for r_t will be found by the following methods:

Method 1 [16]:

$$SE = \sqrt{\left(\frac{\pi\alpha}{2(1+\alpha)^2}\right)^2\left(\frac{1}{f_{00}}+\frac{1}{f_{01}}+\frac{1}{f_{10}}+\frac{1}{f_{11}}\right)}$$

Method 2 [3]:

$$\bar{s}e(\hat{\rho}) = k\sqrt{\left(\frac{1}{f_{00}+0.5}+\frac{1}{f_{01}+0.5}+\frac{1}{f_{10}+0.5}+\frac{1}{f_{11}+0.5}\right)}$$

where:

$$k = \frac{\left(\pi\hat{c}\hat{\omega}^{\hat{c}}\right)sin\left[\pi/(1+\hat{\omega}^{\hat{c}})\right]}{(1+\hat{\omega}^{\hat{c}})^2}$$

Method 3 (Unknown):

$$SE = \sqrt{\frac{p_0(1-p_0)p_1(1-p_1)}{N(h_0h_1)^2}}$$

where:

$$p_0 = (f_{00}+f_{01})/N;\ p_1 = (f_{00}+f_{10})/N \text{ and}$$

h_0 and h_1: the normal distribution at p_0 and p_1, respectively.

2.5 Confidence interval estimation

A Wald confidence interval for r_t is

$$\rho \pm z_{\alpha/2}SE(\rho)$$

where ρ is an estimate of the tetrachoric correlation, SE(ρ) is an estimate of the tetrachoric standard error estimate, and $z_{\alpha/2}$ is the $(1-\alpha/2)$ quartile of the standard normal distribution [3]. The $100(1-\alpha)$ % lower and upper interval estimates for the seven equations of the tetrachoric correlation are as follows:

$$Lower = \rho - Z_{\alpha/2}SE$$
$$Upper = \rho + Z_{\alpha/2}SE$$

3 Results and Discussion

The results of the methods applied to r_t using our generated data set are shown in Table 4. The relationships between all questions were highly significant (P<0.01), except the relationship between the Q2 and Q3 was found to be non-significant. The estimates of r_t by the seven methods recorded the highest values between Q1 and Q4 followed by Q4 and Q5, and Q1 and Q2. While, the lowest values of r_t were found between Q1 and Q3, and Q2 and Q3. On the other hand, the values of r_t between the other (questions) relationships were moderate.

All relationships between the questions showed positive values of r_t using Methods 1, 2, 3, 4, 5 and 6, while, the Method 7 gave negative values of r_t, due to, the values of $f_{00}f_{11}$ being greater than the values of $f_{01}f_{10}$. The numerator of the Method 7 is $f_{01}f_{10} - f_{00}f_{11}$. The sign of r_t is positive when the $f_{00}f_{11}$ entries (agreements) exceeds the $f_{01}f_{10}$ entries (disagreements), and negative when the $f_{01}f_{10}$ entries exceed the $f_{00}f_{11}$. When $f_{01}f_{10}$ is greater than $f_{00}f_{11}$, put $f_{00}f_{11}$ in the numerator instead of $f_{01}f_{10}$ values. When the correlation is negative, the minus sign is affixed by the experimenter [4].

The values of r_t using the first six methods were equal or close for all relationships between the questions, while, the range (maximum value – minimum value) between these methods was 0.01 or 002. Excluding negative 'signal', the highest values of r_t were obtained by the seventh Method; therefore, this Method differed little from the other methods. Tetrachoric *r's* read from the computing charts are usually accurate to at least 0.01 when compared to r_t*'s* calculating using the full formula [4].

All the methods for calculating r_t were consistently associated (P<0.01) or non-significant with each other, indicating that they were identical in calculating r_t.

And all methods seem to work quite well for calculating of r_t. Fortunately, there are several useful approximations to r_t which ·are sufficiently accurate for most purposes [4].

Table 4. Estimates of r_t using the seven methods

Questions	Methods	Questions							
		Q2		Q3		Q4		Q5	
Q1	1	0.809	**0.81****	0.848	**0.85****	0.875	**0.88****	0.588	**0.59****
	2	0.810	**0.81****	0.850	**0.85****	0.880	**0.88****	0.590	**0.59****
	3	0.800	**0.80****	0.832	**0.83****	0.862	**0.86****	0.588	**0.59****
	4	0.805	**0.81****	0.834	**0.83****	0.863	**0.86****	0.579	**0.58****
	5	0.799	**0.80****	0.828	**0.83****	0.856	**0.86****	0.574	**0.57****
	6	0.813	**0.81****	0.846	**0.85****	0.876	**0.88****	0.594	**0.59****
	7	-0.948	**-0.95****	-0.973	**-0.97****	-0.995	**-1.00****	-0.623	**-0.62****
Q2	1			0.087	**0.09**	0.616	**0.62****	0.588	**0.59****
	2			0.090	**0.09**	0.610	**0.61****	0.580	**0.58****
	3			0.089	**0.09**	0.605	**0.61****	0.574	**0.57****
	4			0.086	**0.09**	0.601	**0.60****	0.575	**0.58****
	5			0.085	**0.09**	0.596	**0.60****	0.570	**0.57****
	6			0.090	**0.09**	0.612	**0.61****	0.580	**0.58****
	7			-0.087	**-0.09**	-0.666	**-0.67****	-0.622	**-0.62****
Q3	1					0.588	**0.59****	0.777	**0.78****
	2					0.580	**0.58****	0.770	**0.77****
	3					0.576	**0.58****	0.760	**0.76****
	4					0.573	**0.57****	0.764	**0.76****
	5					0.569	**0.57****	0.758	**0.76****
	6					0.582	**0.58****	0.772	**0.77****
	7					-0.625	**-0.63****	-0.866	**-0.87****
Q4	1							0.819	**0.82****
	2							0.820	**0.82****
	3							0.807	**0.81****
	4							0.818	**0.82****
	5							0.812	**0.81****
	6							0.820	**0.82****
	7							-0.960	**-0.96****

***Note**: The values from the approximations to the methods for calculating the values of r_t are printed in bold.*

Findings on the comparison of the standard error $SE(r_t)$ methods are given in Table 5. The values for standard error of r_t, obtained from the third Method are highest, followed by the second and first methods. The $SE(r_t)$ calculating using the first and second methods are equal to (or reasonable close) to each other, while, the difference between the two methods was very low (0.01). Garrett and Woodworth [4] reported the tetrachoric *r* is most stable when (1) N is large and the 'cuts' in X and in Y close are the median of each variable, and least stable when; (2) N is small and the splits in X and Y depart sharply from 0.50. They added, the $SE(r_t)$ is from 50% to 100% larger than the SE of a product-moment r of the same size and based

upon the same N size. If r is computed from 100 cases, for example, for r_t to be equally stable, the values should be computed using at least 150 to 200 cases [4].

Table 5. Estimates of standard error for the tetrachoric correlation coefficient using the three methods

Questions	Methods	Questions							
		Q2		Q3		Q4		Q5	
Q1	1	0.051	**0.05**	0.048	**0.05**	0.045	**0.05**	0.083	**0.08**
	2	0.054	**0.05**	0.052	**0.05**	0.049	**0.05**	0.084	**0.08**
	3	0.113	**0.11**	0.112	**0.11**	0.112	**0.11**	0.112	**0.11**
Q2	1			0.118	**0.12**	0.081	**0.08**	0.083	**0.08**
	2			0.113	**0.11**	0.083	**0.08**	0.085	**0.09**
	3			0.115	**0.12**	0.116	**0.12**	0.114	**0.11**
Q3	1					0.083	**0.08**	0.059	**0.06**
	2					0.085	**0.09**	0.062	**0.06**
	3					0.115	**0.12**	0.112	**0.11**
Q4	1							0.049	**0.05**
	2							0.052	**0.05**
	3							0.111	**0.11**

***Note**: The values from the approximations to the methods for calculating the values of standard error are printed in bold.*

Generally, the values of r_t using the seven methods were positive and equal to (or close to) each other, except Chen and Popovich Method had negative values and differed little from each other. On the other hand, we suggest that, subject to the data structure, the seven methods can be used to approximate value of r_t. The three methods of $SE(r_t)$ generated values that were equal to (or very close to) each other. Using this information and the associated findings, it can be noted that r_t and $SE(r_t)$ are estimated quite well by the sixth and first methods, respectively.

3.1 Numerical illustration of methods

Guilford and Fruchter [20] reported on a 2 × 2 table for two questions in a personality inventory, for which 930 respondents answered either Yes or No to each question. This is a very common use of the r_t [3]. This example was also analyzed by Bonnet and Price [3], which serves to illustrate the methods of r_t and check the accuracy of our results. The sample of cell and marginal frequencies is presented in Table 6.

Table 6. The sample of cell and marginal frequencies

Variables		X		Total
		X_0	X_1	
Y	Y_0	$f_{00}(203)$	$f_{01}(186)$	389
	Y_1	$f_{10}(167)$	$f_{11}(374)$	541
Total		370	560	930

The Tetrachoric Correlation Coefficient:

1- Calculation of Bonnet and Price [3]:

The tetrachoric correlation is $(\hat{\rho}) = 0.333$, the standard error $(\bar{s}e(\hat{\rho})) = 0.048$ and the 95% confidence interval is between 0.237 and 0.424. After adding 0.5 to each cell, SAS gives an estimated tetrachoric correlation of 0.335, an estimated standard error of 0.048, and a 95% confidence interval of (0.240, 0.429).

2- Calculation of the seven methods in this study:

The results from the seven methods applied to our generated data set are shown in Table 7.

Method 1 [4]:

$$r_t = cos\left(\frac{180^{\circ}x\sqrt{f_{01}f_{10}}}{\sqrt{f_{00}f_{11}}+\sqrt{f_{01}f_{10}}}\right) = cos\left(\frac{180^{\circ}x\sqrt{186x167}}{\sqrt{203x374}+\sqrt{186x167}}\right) = cos\left(\frac{180^{\circ}x\sqrt{31062}}{\sqrt{75922}+\sqrt{31062}}\right)$$
$$= cos\left(\frac{180^{\circ}x176.24}{275.54+176.24}\right) = cos\left(\frac{31723.95}{451.78}\right) = cos70.22^{\circ} \approx 70^{\circ}$$

From cosine table of an angle of 70º is found to be 0.342 and, accordingly, r_t = 0.342.

Method 2 [4]:

$$r_t = \frac{f_{00}f_{11}}{f_{01}f_{10}} = \frac{203x374}{186x167} = \frac{75922}{31062} = 2.44$$

From estimated values table of r_t corresponding to values of the ratio, $f_{00}f_{11}/f_{01}f_{10}$, the corresponding r_t is 0.34.

Method 3 [16]:

$$\alpha = \left(\frac{f_{00}f_{11}}{f_{01}f_{10}}\right)^{\pi/4} = \left(\frac{203x374}{186x167}\right)^{3.14/4} = \left(\frac{75922}{31062}\right)^{0.785} = (2.44)^{0.785} = 2.017$$
$$\rho = \frac{\alpha-1}{\alpha+1} = \frac{2.017-1}{2.017+1} = \frac{1.017}{3.017} = 0.337$$

Method 4 [3]:

$$\omega = \frac{f_{00}f_{11}}{f_{01}f_{10}} = \frac{203x374}{186x167} = \frac{75922}{31062} = 2.444$$
$$p_{1+} = \frac{f_{00}+f_{01}}{N} = \frac{203+186}{930} = \frac{389}{930} = 0.42 \; p_{+1} = \frac{f_{00}+f_{10}}{N} = \frac{203+167}{930} = \frac{370}{930} = 0.40$$

$$c = \frac{1-\frac{|p_{1+}-p_{+1}|}{5}-(0.5-p_{min})^2}{2} = \frac{1-\frac{|0.42-0.40|}{5}-(0.5-0.40)^2}{2} = \frac{1-0.004-0.01}{2} = \frac{0.986}{2}$$
$$= 0.493$$
$$\rho = cos\left(\frac{\pi}{1+\omega^c}\right) = cos\left(\frac{3.14}{1+2.444^{0.493}}\right) = cos\left(\frac{3.14}{1+1.553}\right) = cos\left(\frac{3.14}{2.553}\right) = \cos(1.230)$$
$$= 0.334$$

Method 5 [3]:

$$\widehat{\omega} = \frac{(f_{00}+0.5)(f_{11}+0.5)}{(f_{01}+0.5)(f_{10}+0.5)} = \frac{(203+0.5)x(374+0.5)}{(186+0.5)x(167+0.5)} = \frac{203.5x374.5}{186.5x167.5} = \frac{76210.75}{31238.75} = 2.440$$
$$p_{1+} = \frac{(f_{00}+0.5)+(f_{01}+0.5)}{N+2} = \frac{(203+0.5)+(186+0.5)}{932} = \frac{390}{932} = 0.42$$
$$p_{+1} = \frac{(f_{00}+0.5)+(f_{10}+0.5)}{N+2} = \frac{(203+0.5)+(167+0.5)}{932} = \frac{371}{932} = 0.40$$

$$\hat{c} = \frac{1 - \frac{|p_{1+} - p_{+1}|}{5} - (0.5 - p_{min})^2}{2} = \frac{1 - \frac{|0.42 - 0.40|}{5} - (0.5 - 0.40)^2}{2} = \frac{1 - 0.004 - 0.01}{2} = \frac{0.986}{2} = 0.493$$

$$\hat{\rho} = cos\left(\frac{\pi}{1 + \hat{\omega}^{\hat{c}}}\right) = cos\left(\frac{3.14}{1 + 2.44^{0.493}}\right) = cos\left(\frac{3.14}{1 + 1.552}\right) = cos\left(\frac{3.14}{2.552}\right) = \cos(1.231) = 0.333$$

Method 6 (Unknown):

$$\delta = 1 + \sqrt{\frac{f_{00}f_{11}}{f_{01}f_{10}}} = 1 + \sqrt{\frac{203x374}{186x167}} = 1 + \sqrt{\frac{75922}{31062}} = 1 + \sqrt{2.44} = 1 + 1.563 = 2.563$$

$$\rho = \cos\left(\frac{\pi}{\delta}\right) = \cos\left(\frac{3.14}{1.563}\right) = \cos(1.225) = 0.339$$

Method 7 [17]:

$$\lambda_X = \frac{f_{00} + f_{01}}{N} = \frac{203 + 186}{930} = \frac{389}{930} = 0.4183$$

The ordinate (height) of the standardized normal distribution at 0.4183 is 0.3905

$$\lambda_Y = \frac{f_{01} + f_{11}}{N} = \frac{186 + 374}{739} = \frac{560}{930} = 0.6022$$

The ordinate (height) of the standardized normal distribution at 0.6022 is 0.3858

$$r_{tet} = \frac{f_{01}f_{10} - f_{00}f_{11}}{\lambda_X\lambda_Y N^2} = \frac{186x167 - 203x374}{0.3905x0.3858x930^2} = \frac{31062 - 75922}{0.1507x864900} = \frac{-44860}{130312.91} = -0.344$$

Table 7. Accuracy of the seven methods used for r_t approximation.

Methods	Tetrachoric *r* values	Approximation
1	0.342	0.34
2	0.340	0.34
3	0.337	0.34
4	0.334	0.33
5	0.333	0.33
6	0.339	0.34
7	-0.344	-0.34
Check		
Bonnet and Price (2005)	0.333	0.33
SAS	0.335	0.34

Standard error methods $SE(r_t)$:

The results of the three methods applied to our generated data set are shown in Table 8.

Method 1 [16]:

$$SE = \sqrt{\left(\frac{\pi\alpha}{2(1+\alpha)^2}\right)^2\left(\frac{1}{f_{00}} + \frac{1}{f_{01}} + \frac{1}{f_{10}} + \frac{1}{f_{11}}\right)} = \sqrt{\left(\frac{3.14x2.017}{2(1+2.017)^2}\right)^2\left(\frac{1}{203} + \frac{1}{186} + \frac{1}{167} + \frac{1}{374}\right)}$$

$$= \sqrt{\left(\frac{6.34}{18.20}\right)^2(0.005 + 0.005 + 0.006 + 0.003)} = \sqrt{0.121x0.019} = \sqrt{0.002} = 0.048$$

Method 2 [3]:

$$k = \frac{(\pi\hat{c}\hat{\omega}^{\hat{c}})sin[\pi/(1+\hat{\omega}^{\hat{c}})]}{(1+\hat{\omega}^{\hat{c}})^2} = \frac{(3.14x0.493x2.44^{0.493})sin[3.14/(1+2.44^{0.493})]}{(1+2.44^{0.493})^2}$$

$$= \frac{(3.14x0.493x1.552)sin[3.14/(1+1.552)]}{(1+1.552)^2} = \frac{(2.402)\sin(1.23)}{6.512} = \frac{2.402x0.943}{6.512}$$

$$= 0.348$$

$$\bar{s}e(\hat{\rho}) = k\sqrt{\left(\frac{1}{f_{00}+0.5}+\frac{1}{f_{01}+0.5}+\frac{1}{f_{10}+0.5}+\frac{1}{f_{11}+0.5}\right)} = 0.348\sqrt{\left(\frac{1}{203.5}+\frac{1}{186.5}+\frac{1}{167.5}+\frac{1}{374.5}\right)} =$$

$$= 0.348\sqrt{0.005+0.005+0.006+0.003} = 0.348\sqrt{0.019} = 0.048$$

Method 3 (Unknown):

$$p_0 = \frac{f_{00}+f_{01}}{N} = \frac{203+186}{930} = 0.4183 \qquad h_0 = 0.3905$$

$$p_1 = \frac{f_{00}+f_{10}}{N} = \frac{203+167}{930} = 0.3978 \qquad h_1 = 0.3858$$

$$SSE = \sqrt{\frac{p_0(1-p_0)p_1(1-p_1)}{N(h_0h_1)^2}} = \sqrt{\frac{0.4183(1-0.4183)x\,0.3978(1-0.3978)}{930(0.3905x0.3858)^2}} = \sqrt{\frac{0.2433x0.2396}{930x0.0227}}$$

$$= \sqrt{\frac{0.0583}{21.112}} = \sqrt{0.0028} = 0.053$$

Table 8. Accuracy of the methods for standard error $SE(r_t)$ approximations

Methods	$SE(r_t)$	Approximations
1	0.048	0.05
2	0.048	0.05
3	0.053	0.05
Check		
Bonnet and Price (2005)	0.048	0.05
SAS	0.048	0.05

Confidence Interval Estimation:

$$Lower = \rho - Z_{\alpha/2}SE(r_t) = 0.33 - (1.96x0.048) = 0.33 - 0.094 = 0.236$$
$$Upper = \rho + Z_{\alpha/2}SE(r_t) = 0.33 + (1.96x0.048) = 0.33 + 0.094 = 0.424$$

The results of the other methods applied to our generated data set are shown in Table 9.

Table 9. Accuracy of 95% confidence interval approximations

Methods	r_t value	$SE(r_t)$	95% confidence interval			
			Lower		Upper	
4 and 5	0.33	0.048	0.236	**0.24**	0.424	**0.42**
		0.053	0.226	**0.23**	0.434	**0.43**
Other methods	0.34	0.048	0.246	**0.25**	0.434	**0.43**
		0.053	0.236	**0.24**	0.444	**0.44**
Check						
Bonnet and Price	0.333	0.048	0.237	**0.24**	0.424	**0.42**
SAS	0.335	0.048	0.240	**0.24**	0.429	**0.43**

***Note**: The approximation values of standard error are printed in bold.*

4 Conclusion

The tetrachoric correlation coefficient (r_t) has important historical and practical importance, but our consulting experiences suggest that most researchers and graduate students have a very limited understanding of this interesting measure of association. Increased use of the r_t may occur in the future if students are exposed to a computationally simple and complete set of inferential methods instead of the brief and incomplete discussions found in modern texts [3]. Therefore, we propose that, the Method 6 and Method 1 for calculating the r_t and $SE(r_t)$ respectively works well, the equations are understandable, computationally simple, and ideally suited for users.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Ekström J. The Phi-coefficient, the Tetrachoric correlation coefficient, and the Pearson-Yule Debate. UCLA, Department of Statistics Papers; 2011.
Available:https://escholarship.org/uc/item/7qp4604r

[2] Pearson E. Mathematical contribution to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society of London. 1900;195A:1-47.

[3] Bonett DG, Price RM. Inferential methods for the Tetrachoric correlation coefficient. Journal of Educational and Behavioral Statistics. 2005;30(2):213–225.

[4] Garrett HE, Woodworth RS. Statistics in psychology and education. G. U. Mehta for Vakils, Feffer and Simons Ltd. Bombay 400 038, India; 1966.

[5] Lorenzo-Seva U, Ferrando PJ. TETRA-COM: A comprehensive SPSS program for estimating the tetrachoric correlation. Behav. Res. 2012;44:1191–1196.

[6] Fleiss JL. Statistical methods for rates and proportions (2nd ed.). New York: John Wiley; 1981.

[7] Long MA, Berry KJ, Mielke PW. Tetrachoric correlation a permutation alternative. Educational and Psychological Measurement. 2009;69(3):429-437.

[8] Hutchinson TP. Kappa muddles together two sources of disagreement: Tetrachoric correlation is preferable. Research in Nursing & Health. 1993;16:313-315.

[9] Pearson K. On the probable error of a coefficient of correlation as found from a fourfold table. Biometrika. 1913;9:22-27.

[10] Kendall MG, Stuart A. The advanced theory of statistics (Vol. 2). New York: Hafner; 1961.

[11] Greer T, Dunlap WP, Beatty GO. A Monte Carlo evaluation of the tetrachoric correlation coefficient. Educational and Psychological Measurement. 2003;63:931-950.

[12] SYSTAT. SYSTAT ® 8.0 Statistics. Chapter 6: Correlations, Similarities, and Distance Measures, p 122. SPSS Inc. Printed in the United States of America; 1998.

[13] Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison Wesley; 1968.

[14] Noyan F, Şimşek GG. Tetrachoric correlation as a measure of default correlation. Procedia - Social and Behavioral Sciences. 2012;62:1230–1234.

[15] Ogasawara H. Accurate distribution and its asymptotic expansion for the tetrachoric correlation coefficient. Journal of Multivariate Analysis. 2010;101:936948.

[16] Edwards JH, Edwards AWF. Approximating the tetrachoric correlation coefficient. Biometrics 1984;40:563.

[17] Chen PY, Popovich PM. Correlation: parametric and non-parametric measures. Sage University Papers Series on quantitative Applications in the Social Sciences, 07-139. Thousand Oaks, CA: Sage; 2002.

[18] Davidoff MD, Goheen HW. Psychometrika. 1953;18:115-121.

[19] Agresti A. Categorical data analysis (2nd ed.) John Wiley & Sons, Inc., Hoboken, New Jersey; 2002.

[20] Guilford JP, Fruchter B. Fundamental statistics in psychology and education (5th ed.). New York: McGraw-Hill; 1973.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sciencedomain.org/review-history/28019