Correlation Coefficient for Continuous and Discrete Data Part 4 of 4

Chanoknath Sutanapong★ & Louangrath, P.I. ★★

About the author

★Chanoknath Sutanapong is an independent researcher. She may be reached by email at: Chanoknath.sutanapong@gmail.com

★★ Louangrath, P.I. is an Assistant Professor in Business Administration at Bangkok University, Bangkok, Thailand. For correspondence, he could be reached by email at: Lecturepedia@gmail.com

ABSTRACT

The purpose of this paper is to introduce researchers to correlation coefficient calculation. The Pearson Product Moment Correlation Coefficient is the most common type of correlation; however, the Pearson correlation coefficient may not be applicable in all cases. The Pearson r is used only when the data of the dependent and independent variables are quantitative. There are many types of correlation coefficient. The correct choice of correlation coefficient depends on the classification of the independent variable (X) and dependent variable (Y). Data are classified into one of three types: quantitative, nominal and ordinal. This writing explains various types of correlations on the basis of X-by-Y data type combination. Using the wrong type of correlation coefficient would lead to faulty inference; as the result, the researcher would commit Type 2 error.

Keywords: correlation coefficient, data type, Type 2 error

CITATION:

Sutanapong, C. and Louangrath, P.I. (2015). "Correlation Coefficient for Continuous and Discrete Data– Part 4 of 4." *Inter. J. Res. Methodol. Soc. Sci.*, Vol., 1, No. 4: pp. 25-50. (Oct. – Dec. 2015).

1.0 INTRODUCTION

Correlation coefficient is commonly used to measure the level of association between variables: independent (Y) and dependent (X) (Boddy and Smith, 2009). Incorrect type of correlation coefficient would lead to Type 2 error. Type 2 error is defined as a wrongful acceptance of the null hypothesis (Sherman, 2002). The objective of this paper is to provide a clear guidance on how to use correct type of correlation coefficient. We begin with a simple set up of linear equation in a form of $Y = a + bX + c$ where a is the Y-intercept, b is the slope of the linear regression line, and c is the forecast error. In basic statistics, the linear regression line equation is obtained through the following three statements:

$$I_{XY} = N\Sigma XY - (\Sigma X)(\Sigma Y)$$

$$II_X = N\Sigma X^2 - (\Sigma X)^2$$

$$III_Y = N\Sigma Y^2 - (\Sigma Y)^2$$

$$a = \overline{Y} - b\overline{X}$$

$$b = \frac{I_{XY}}{II_X}$$

$$c = \sqrt{\left[\frac{1}{N(N-2)}\right]\left[(III_Y) - \left(\frac{(I_{XY})^2}{II_X}\right)\right]}$$

The argument in $Y = a + bX + c$ asserts that there is a relationship between X and Y. This relationship is embodied in the slope *b*. The level of this association or relationship is measured by the ratio of the deviation in Y with respect to the change in X adjusted by *b*. This simple procedure may become erroneous if we use the Pearson correlation coefficient for all cases. The Pearson *r* is appropriate only when both X and Y are quantitative. However, in cases dealing with non-quantitative data, such as demographic information or preferential ranking, the Pearson *r* would not be applicable. This paper intends to address these non-Pearson *r* cases.

2.0 TYPES OF CORRELATION COEFFICIENT

Reliability test uses correlation coefficient as one of the means to test the degree of association. In reliability context, correlation coefficient is interpreted as the ability to replicate the data of a prior study as the current experiment represents one array and the prior study represents the second array. The function of correlation coefficient is to give an index of association between two data arrays. The researcher must be aware of various types of correlation coefficient calculations and which one to use in a given situation. The situation is defined by the type of data available. The variable may be defined as X and Y. The objective of correlation coefficient calculation is to determine the relationship between X and Y through the measurement of association between X and Y. The table below illustrates the type of data crossing according to data types.

Table 1. Types of correlation coefficient classified by data types

Variable (X,Y)	**Quantitative X**	**Ordinal X**	**Nominal X**	***Nota Bene***
Quantitative Y	Pearson *r*	Biserial r_b	Point Biserial r_{pb}	Determine the type of data for X and Y then select the appropriate correlation.
Ordinal Y	Biserial r_b	Spearman *rho* Tetrachoric & Polychoric	Rank Biserial r_{rb}	
Nominal Y	Point Biserial r_{pb}	Rank Biserial r_{rb}	Phi, L, C, Lambda	

This section of the writing includes ten types of correlation coefficients; each type of the correlation coefficient is used according to the characteristic of the data arrays: quantitative, ordinal, or nominal for the variables X and Y. There are nine common correlation coefficient types; one additional type is a variance of the tetrachoric correlation (2×2) made to accommodate $K \times L$ contingency data for ordinal-x-ordinal data. This extension of Pearson's 2×2 contingency is called *polychoric correlation*. Before examining each type of correlation, it is important to be familiar with data classification.

2.1 Pearson correlation coefficient

The Pearson correlation coefficient is the most commonly used form of correlation. The Pearson r is used when both X and Y are quantitative data. Quantitative data is the numerical measurement produced by the instrument without any intermediary interpretation, translation, or transformation. The raw data from the response itself may be read as a numerical data. This type of data may be accommodated by the Pearson correlation coefficient. The Pearson correlation coefficient is given by:

$$r = \frac{1}{n-1}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right) \tag{1}$$

where …

$Z = \frac{X_i - \bar{X}}{s_X}$ is the standard score measuring how far the individual data point is located away from the mean;

$\bar{X} = \frac{1}{n-1}\sum_{i=1}^{n} X_i$ is the mean of X_i, and

$\bar{Y} = \frac{1}{n-1}\sum_{i=1}^{n} Y_i$ is the mean of Y_i.

Another means to define r is to use the slope of the linear equation $Y = a + bX + c$ as the parameter and multiply the slope by the quotient of the standard deviation of X divided by the standard deviation of Y, thus:

$$r = b\left(\frac{s_X}{s_Y}\right) \tag{2}$$

where …

$$b = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2} \quad \text{where } X:(x_1, x_2, ..., x_n) \text{ and } Y:(y_1, y_2, ..., y_n) \tag{3}$$

and the standard deviation is generally given by:

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \quad \text{and } s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{4}$$

The value of the correlation coefficient ranges between -1 and +1. A value of zero means that there is no association between the two arrays. Negative coefficient means that there is an obverse association. If there is an increase in X, there is a corresponding decrease in Y and *vice* *versa*. A positive coefficient means that there is a perfect association. If there is an increase in X, there is also an increase in Y.

2.2 Biserial correlation coefficient

Biserial correlation is used when the X array is quantitative and the Y array is ordinal data. For example, X may represent the raw score of test performance of n number of students and Y represents the ranking placement of students according to their test scores, i.e. 100 = 1st place, 90 = 2nd, 80 = 3rd, and so on. The biserial correlation is given by:

$$r_b = \frac{(Y_1 - Y_0)\left(\frac{pq}{Y}\right)}{\sigma_Y} \qquad (5)$$

where …

Y = Y score means for data pairs with $x:(1,0)$:

$$Y_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i \quad \text{and} \quad Y_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} y_i$$

q = $1 - p$;

p = proportion of data pairs with scores $x:(1,0)$; and

σ_Y = population standard deviation for the y data and Y is the height of the standardized normal distribution at point z where $P(z' < z) = q$ and $P(z' > z) = p$.

Note that the probability for p and q may be given by the Laplace Rule of Succession (Laplace, 1814):

$$p = \frac{s+1}{n+2} \quad \text{where } s = \text{number of success and } n = \text{total observations.} \qquad (6)$$

$$q = 1 - p \qquad (7)$$

The population standard deviation (σ) may not be known; however, it may be determined indirectly through two-steps process: (i) t-equation and (ii) Z-equation.

$$t = \frac{\overline{x} - \mu}{S/\sqrt{n}} \qquad (8)$$

From the t-equation, determine the population mean (μ), thus:

$$\mu = t\left(\frac{S}{\sqrt{n}}\right) - \overline{x} \qquad (9)$$

With known μ, the population standard deviation may be determined through the Z-equation. The Z-equation is given by:

$$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \qquad (10)$$

Now solve for the population standard deviation (σ), thus:

$$\sigma = \left(\frac{\bar{x} - \mu}{Z} \right) \sqrt{n} \qquad (11)$$

Note that p and q are used when discrete probability is involved. In point biserial correlation, discrete probability is used because Y (response variable) exists as a ranked or ordinal variable. The response $Y : (y_1, y_2, ... y_n)$ either falls with a rank placement: $[1^{st}, 2^{nd}, ..., i^{th}]$ or it does not. This "either or" argument dichotomizes the ordinal variable into {Yes | No} identifier which could be score as Yes = 1 and No = 0. Therefore, p and q of the discrete binomial probability is used. The test statistic for the binomial probability is given by:

$$Z_{bin} = \frac{\frac{X}{n} - p}{\sqrt{\frac{pq}{n}}} \qquad \textit{See infra.}$$

2.3 Point biserial correlation coefficient

There are two cases where point-biserial correlation is used: (i) X is nominal and Y is quantitative data, and (ii) X is quantitative data and Y is nominal. In addition, if one variable, such as Y in the series of X and Y is dichotomous, point-biserial correlation is also used. Dichotomous data are categorical data that gives a binomial distribution. This type of distribution is produced by {Yes | No} answer category. Although the point biserial correlation is equivalent to the Pearson correlation, the formula is different from the Pearson product moment correlation. The mathematical equivalence is: $r_{XY} = r_{pb}$. The point-biserial correlation (r_{pb}) is given by:

$$r_{pb} = \left(\frac{M_1 - M_0}{s_n} \right) \sqrt{\frac{n_1 n_0}{n^2}} \qquad (12)$$

where s_n is the standard deviation of the combined population or pooled standard deviation, thus:

$$s_n = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2} \qquad (13)$$

In order to obtain S_n, both arrays must be combined: $n_1 + n_2 = n$, and the standard deviation of the combined string n is calculated to obtain s_n. This pooled standard deviation is presented as S_n.

The term M_1 is the mean value for the continuous $X : x_1, x_2, ..., x_n$; therefore: $M_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$ for all data points in group 1 with size n_1 and $M_0 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_i$. The combined sample size is given by: $n = n_1 + n_2$. If a data comes from only one *sample* of the population, s_{n-1} is used for the standard deviation. Thus r_{pb} is written as:

$$r_{pb} = \left(\frac{M_1 - M_0}{s_{n-1}}\right)\sqrt{\frac{n_1 n_0}{n(n-1)}} \tag{14}$$

The standard deviation for the "sample only" data set is given by:

$$s_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \tag{15}$$

The two equations using s_n and s_{n-1} are equivalent, thus;

$$r_{pb} = \left(\frac{M_1 - M_0}{s_n}\right)\sqrt{\frac{n_1 n_0}{n^2}} = \left(\frac{M_1 - M_0}{s_{n-1}}\right)\sqrt{\frac{n_1 n_0}{n(n-1)}}$$

The test statistic is the t-test which is given by:

$$t_{pb} = r_{pb}\sqrt{\frac{n_1 + n_0 - 2}{1 - r_{pb}^2}} \tag{16}$$

The degrees of freedom is $v = n_1 + n_0 - 2$. If the data array of X is normally distributed, a more accurate biserial coefficient is given by:

$$r_b = \left(\frac{M_1 - M_0}{s_n}\right)\left(\frac{n_1 n_0}{n^2 u}\right) \tag{17}$$

where u is the *abscissa* or Y of the normal distribution $N(0,1)$. Normal distribution may be verified by the Anderson-Darling test (Stephen, 1974, 1986).

There are three types of point-biserial correlation, namely (i) the Pearson correlation between item scores and total test scores including the item scores, (ii) the Pearson correlation between item scores and total test score excluding the item scores, and (iii) correlation adjusted for the bias resulted from the inclusion of the items scores. The correlation adjusted for the bias resulted from the inclusion of the items scores is given by:

$$r_{upb} = \frac{M_1 - M_0 - 1}{\sqrt{\left(\frac{n^2 s_n^2}{n_1 n_0}\right) - 2\left(M_1 - M_0\right) + 1}} \tag{18}$$

Note that for point-wise or specific probability of X value, the binomial distribution for the categorical data is given by:

$$P(X) = \frac{n!}{(n-X)!n!}p^X q^{X-n} \tag{19}$$

where n is the total number of observations, X is the specified value to be predicted, p is the probability of success of the observed value over the total number of events, and q is $1-p$ or the probability of failure. The test statistic for the binomial distribution is:

$$Z_{bin} = \frac{\frac{X}{n} - p}{\sqrt{\frac{pq}{n}}} \tag{20}$$

Recall that the term a in the 2×2 contingency table is the frequency of for perfect match of {Yes: observed} and {Yes: forecast}. The frequency a is equal to X in $P(X)$ as illustrated in the table below.

Table 2. Contingency Table 2 x 2

	Y			
X		YES	NO	***Forecast***
	YES	a	b	$P(F) = a + b$
	NO	c	d	$1 - P(F)$
	Observation	$P(O) = a + c$	$1 - P(O)$	$a + b + c + d$

The term $a+b+c+d$ is the combined joint probability of all events in the set.

2.4 Spearman rho

The Spearman correlation coefficient is used when both the independent variable (X) and dependent variable (Y) are ordinal. Ordinal data is defined as a ranked order type of a well order set (Dauben, 1990; Moore, 1982; Suppes, 1972): {first, second, third, …, n^{th}}. However, there is a claim made by Lehman that the Spearman coefficient can be used for both continuous and discrete variable (Lehman, 2005). This section focuses on the ordinal data of both dependent and independent variables.

Assume that there are two arrays of data called independent variable: X_i and dependent variable: Y_i. The ordinal data of these variable are x_i and y_i respectively. The correlation coefficient of x_i and y_i is given by:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{21}$$

There is an alternative calculation of *rho* through the use of the difference of two ranked arrays (Myers and Well, 2003; Maritz, 1981) x_i and y_i, thus:

$$\rho = \frac{1 - 6\sum d_i}{n\left(n^2 - 1\right)} \tag{22}$$

where $d_i = x_i - y_i$ and n is the number of elements in the paired set: i in d. This method is not used if the researcher is looking for top X. Generally, equation (21) is used.

The test statistic used for the Spearman rank correlation is given by the Z-test or t-test. In order to use the Z-score test, it is necessary to find the Fisher's transformation of r. The Fisher's transformation for the correlation is given by:

$$F(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{23}$$

where $r = b\left(\frac{S_x}{S_y}\right)$ and $\ln$ is the natural log of base $e = 2.718...$ The test statistic under the Z-equation is given by:

$$z = \left(\sqrt{\frac{n-3}{1.06}}\right)F(r) \tag{24}$$

The null hypothesis is that $r = 0$ which means that there is a statistical independence (Choi, 1977; Feiller *et al.*, 2003) or no dependent association. In the alternative, the test statistic may also be determined by the t-test thus:

$$t_r = r\sqrt{\frac{n-2}{1-r^2}} \tag{25}$$

The degree of freedom (Press et al., 1992) is given by $df = n-2$. The argument in support of t_r approach rests on the idea of permutation (Kendall and Stuart, 1973). The equivalence of the above determination is the Kendall's *tau* (Kowalczyk, *et al.*, 2004). Kendall's *tau* is beyond the scope of the present topic.

2.5 Tetrachoric correlation coefficient

The tetrachoric correlation coefficient is used when both the independent (X) and dependent (Y) variables are dichotomous or binary data and both are ordinal. Generally, there are two correlation tests used for binary data, namely phi-coefficient and tetrachoric correlation coefficient. The data is commonly presented in 2×2 contingency table. Below is an example of the 2×2 contingency table and its scoring.

Table 3. Contingency table for frequency counts

		Y		
		Yes	*No*	
X	*Yes*	*a*	*b*	p_F
	No	*c*	*d*	$1-p_F$
		p_o	$1-p_o$	

A series of definition for the terms used in tetrachoric correlation coefficient must be provided in order to gain a clearer understanding. The definitions are based on the 2×2 contingency table below:

p_o and p_F = marginal frequencies;

p_o = probability of the observed and

p_F = probability of the forecast;

a = joint frequency of the contingency table

O = observations which is comprised of $O \to X_O : (x_{O1} + x_{O2} +, ..., x_{On})$

F = forecast which is comprised of $F \to X_F : (x_{F1} + x_{F2} +, ..., x_{Fn})$

The three frequencies: a, p_o and p_F determine the values in the table. The bias of this determination is given by:

$$Bias = \frac{P_F}{P_o} \quad \text{or } Bias = \frac{a+b}{a+c} \qquad (26)$$

Juras and Pasaric (2006) formally explained tetrachoric correlation coefficient as:

> "Let $z_O = \Phi^{-1}(P_O)$ and $z_F = \Phi^{-1}(P_F)$ be the standard normal deviates (SND) corresponding to marginal probabilities P_O and P_F, respectively. The tetrachoric correlation coefficient (TCC), introduced by Pearson (1900), is the correlation coefficient r that satisfies
>
> $$a = \int_{z_O}^{\infty} \int_{z_F}^{\infty} \phi(x_1, x_2, r)\, dx_1 / dx_2 \,, \qquad (2)$$
>
> Where $\phi(x_1, x_2, r)$ is the bivariate normal p.d.f.
>
> $$\phi(x_1, x_2, r) = \frac{1}{2\pi\sqrt{1-r^2}} \exp\left[-\frac{1}{2(1-r^2)} \left(x_1^2 - 2rx_1x_2 + x_2^2 \right) \right] .(3)$$
>
> The line $x_1 = z_O$ and $x_2 = z_F$ divide the bivariate normal into four quadrants whose probabilities correspond to relative frequencies in the 2×2 table.
>
> Clearly, the SDN_{-S} z_O and z_F are uniquely determine by P_O and P_F, respectively. The double integral in (2) can be expressed as (National Bureau of Standards, 1959):
>
> $$a = \frac{1}{2\pi} \int_{\arccos r}^{\pi} \exp\left[-\frac{1}{2} \left(z_O^2 + z_F^2 - 2z_O z_F \cos\omega \right) \operatorname{cosec}^2 \omega \right] d\omega \ (4)$$
>
> Showing that the joint frequency a is a monotone function of r is well defined by (2)."
> *See* Josip Juras and Zuran Pasaric (2006). "Application of tetrachoric and polychoric correlation coefficients to forecast verification." GEOFIZIKA, Vol. 23, No. 1, p. 64.

Another version, the tetrachoric correlation coefficient is defined as the solution given by r_{tc} to the integral equation:

$$p_a = \int_{\Phi=1(1-p_X)}^{\infty} \int_{\Phi=1(1-p_Y)}^{\infty} \phi_2(x_1, x_2, r_{tc}) dx_2 dx_1 \qquad (27)$$

where $\Phi(x)$ is the standard normal distribution and $\phi_2(x_1, x_2, \rho)$ is the bivariate standard normal density function. The term p_a may be written as:

$$p_a = \bar{\Phi}\left(\Phi^{-1}\left(1-p_X\right), \Phi^{-1}\left(1-p_Y\right), r_{tc}\right) \tag{28}$$

These formal definitions are not helpful for the actual calculation of the tetrachoric correlation coefficient. For practical purpose, assume that the 2×2 contingency table below as the basis for further discussion of the tetrachoric correlation coefficient: r_{tc}.

Table 4. Contingency table for joint probabilities

		Y		
		Pos.	*Neg.*	
X	*Pos.*	p_a	p_b	p_X
	Neg.	p_c	p_d	$1-p_X$
		p_Y	$1-p_Y$	

Juras and Pasaric gave an extensive treatment of the tetrachoric correlation coefficient when they provided the Peirce measure (s_P), Heidke measure (s_H) and the Doolitlle measure (s_D) as the estimate of r_{tc}. All these measures are comparable calculation for the tetrachoric correlation coefficient. These measures are provided as:

$$s_P = \frac{a - P_O P_F}{P_O(1-P_O)} \tag{29}$$

$$s_H = \frac{2\left(a - P_O P_F\right)}{P_O + P_F - 2P_O P_F} \tag{30}$$

$$s_D = \frac{a - P_O P_F}{\sqrt{P_O\left(1-P_O\right)P_F\left(1-P_F\right)}} \tag{31}$$

In addition, the Yule's odd ratio skill score is said to also give the approximation of the tetrachoric correlation coefficient:

$$S_Y = \frac{a - P_O P_F}{a\left[1-2\left(P_O + P_F\right)+2a\right]+P_O P_F} \tag{32}$$

Finally, the actually tetrachoric correlation coefficient is given by
(Juras, 1998; Johnson and Kots, 1972):

$$S_r = \sin\left(\frac{\pi}{2}(4a-1)\right) \tag{33}$$

Note that S with subscripts is equivalent to r_{tc}. One alternative to calculating tetrachoric correlation coefficient is given by the alpha ratio:

$$r_{tc} = \frac{\alpha - 1}{\alpha + 1} \tag{34}$$

where $\alpha = \left(\frac{AD}{BC}\right)^{\pi/4}$ and the equivalence of ABCD are $A = P_a$; $B = P_b$; $C = P_c$ and $D = P_d$. This short-hand version is less complicated than the Pearson's original version (Pearson, 1904). Yet another shorthand formula for tetrachoric correlation coefficient is given by:

$$r_{tc} = \cos\left(\frac{180}{1+\sqrt{(BC/AD)}}\right) \qquad (35)$$

where the contingency table is given by:

		Y		
		0	1	
X	1	A	B	$A+B$
	0	C	D	$C+D$
		$A+C$	$B+D$	

Assume that Table 8.4.0 has the following data set:

		Y		
		0	1	
X	1	$A=10$	$B=5$	$A+B=15$
	0	$C=5$	$D=10$	$C+D=15$
		$A+C=15$	$B+D=15$	30

The calculation for r_{tc} follows:

$$r_{tc} = \cos\left(\frac{180}{1+\sqrt{(BC/AD)}}\right)$$

$$r_{tc} = \cos\left(\frac{180}{1+\sqrt{(5)(5)/(10)(10)}}\right) = \cos\left(\frac{180}{1+\sqrt{25/100}}\right)$$

$$r_{tc} = \cos\left(\frac{180}{1+\sqrt{25/100}}\right) = \cos\left(\frac{180}{1+\sqrt{0.25}}\right) = \cos\left(\frac{180}{1+0.50}\right)$$

$$r_{tc} = \cos\left(\frac{180}{1.50}\right) = \cos(120) = 0.81$$

Using equation (34), the calculation for alpha follows:

$$\alpha = \left(\frac{AD}{BC}\right)^{\pi/4} = \left(\frac{10(10)}{5(5)}\right)^{3.14/4} = \left(\frac{100}{25}\right)^{3.14/4} = 4^{3.14/4}$$, then …

$$\alpha = 4^{0.79} = 2.97$$

$$r_{tc} = \frac{\alpha-1}{\alpha+1} = \frac{2.97-1}{2.97+1} = \frac{1.97}{3.97} = 0.50$$

Another means of determining the tetrachoric correlation coefficient is given by:

$$r_{tc}=\sin\left(\frac{\pi}{2}\left(\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)\right) \qquad (36)$$

According to this method, the calculation follows:

$$r_{tc}=\sin\left(\frac{\pi}{2}\left(\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)\right)=\sin\left(\frac{\pi}{2}\left(\frac{\sqrt{10(10)}-\sqrt{5(5)}}{\sqrt{10(10)}+\sqrt{5(5)}}\right)\right)$$

$$r_{tc}=\sin\left(\frac{\pi}{2}\left(\frac{\sqrt{100}-\sqrt{25}}{\sqrt{100}+\sqrt{25}}\right)\right)=\sin\left(\frac{\pi}{2}\left(\frac{10-5}{10+5}\right)\right)$$

$$r_{tc}=\sin\left(\frac{\pi}{2}\left(\frac{5}{15}\right)\right)=\sin\left(1.57(0.33)\right)=\sin(0.52)$$

$$r_{tc}=0.50$$

The result of the calculation shows that equations (10.34) and (10.36):

$r_{tc}=\dfrac{\alpha-1}{\alpha+1}$ and $r_{tc}=\sin\left(\dfrac{\pi}{2}\left(\dfrac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)\right)$ produce the same result and equation $r_{tc}=\cos\left(\dfrac{180}{1+\sqrt{\left(BC/AD\right)}}\right)$ produces a higher coefficient. In comparison, equation (8.33) yields the following computation:

$$S_r=\sin\left(\frac{\pi}{2}(4a-1)\right)$$

$$S_r=\sin\left(\frac{\pi}{2}(4(10)-1)\right)=\sin\left(\frac{\pi}{2}(40-1)\right)=\sin\left(\frac{\pi}{2}(39)\right)$$

$$S_r=\sin\left(\frac{3.14}{2}(39)\right)=\sin\left(1.57(39)\right)=\sin(61.23)$$

$$S_r=-1$$

The results of the computation are from various claims of method are not in agreement. Below are the computation of the Peirce's measure, Heike, Doolittle and Yule. For convenience the following definition and value are given:

$$P_O=a+c=10+5=15$$

$$P_F=a+b=10+5=15$$

The calculation for the Peirce measure follows:

$$s_P = \frac{a - P_O P_F}{P_O(1-P_O)} = \frac{10-(15)(15)}{15(1-15)} = \frac{10-225}{15(-14)} = \frac{-215}{-210}$$
$$s_P = 1.02$$

The result does not seem to be accurate because the range of a correlation coefficient is between -1 and +1. The calculation under the Heidke measure follows:

$$s_H = \frac{2\left(a - P_O P_F\right)}{P_O + P_F - 2P_O P_F} = \frac{2\left(10-15(15)\right)}{15+15-2(15(15))} = \frac{2\left(10-225\right)}{15+15-2(225)}$$
$$s_H = \frac{2\left(10-225\right)}{30-450} = \frac{2(-215)}{-420} = \frac{-430}{-420}$$
$$s_H = 1.02$$

Again a different number is obtained. The calculation under the Doolittle measure follows:

$$s_D = \frac{a - P_O P_F}{\sqrt{P_O\left(1-P_O\right)P_F\left(1-P_F\right)}} = \frac{10-15(15)}{\sqrt{15(1-15)15(1-15)}} = \frac{215}{\sqrt{15(14)15(14)}}$$
$$s_D = \frac{215}{\sqrt{15(14)15(14)}} = \frac{215}{\sqrt{15(14)210}} = \frac{215}{\sqrt{210(210)}} = \frac{215}{\sqrt{4410}} = \frac{215}{210}$$
$$s_D = 1.02$$

This result coincides with the Peirce measure. Lastly, under the Yule's method the calculation follows:

$$S_Y = \frac{a - P_O P_F}{a\left[1-2\left(P_O + P_F\right)+2a\right] + P_O P_F}$$
$$S_Y = \frac{10-(15)(15)}{10\left[1-2(15+15)+2(10)\right]+15(15)} = \frac{10-225}{10\left[1-2(30)+20\right]+225}$$
$$S_Y = \frac{-215}{10\left[1-60+20\right]+225} = \frac{-215}{10\left(-39\right)+225} = \frac{-215}{-390+225} = \frac{-215}{-165}$$
$$S_Y = 1.30$$

The result under the Yule's measure also exceeds 1.00. Bias in each case is determined by: $Bias = P_F\ /\ P_O = 15/15 = 1.00$. The results of the various measures may be summarized thus:

$$s_P = 1.02$$
$$s_H = 1.02$$
$$s_D = 1.02$$
$$s_Y = 1.30$$
$$S_r = -1.00$$

The values appear to be consistent, except the sign for S_r. With the exception of the negative sign of S_r, the interpretation of the correlation coefficient would otherwise be consistent. However, with the negative S_r, the S_r value would point to an opposite meaning in interpretation.

Table 5. Estimating level of significance using standard score: Z

X_i	$\bar{X}$	$\left(X_i - \bar{X}\right)$	S	$Z_i = \left(X_i - \bar{X}\right)/S$	$\hat{Z}_i$
1.02	0.672	0.348	0.9425	0.369	1.65
1.02	0.672	0.348	0.9425	0.369	1.65
1.02	0.672	0.348	0.9425	0.369	1.65
1.30	0.672	0.628	0.9425	0.666	1.65
-1.00	0.672	-1.672	0.9425	-1.774	1.65

Table 5 shows the determining of the level of significance of the for the variance estimates for the tetrachoric correlation. All estimates are consistent except S_r showing significant difference: $\left|-1.774\right| > 1.65$. The above analysis deals with correlation coefficient called tetrachoric for 2×2 contingency table. There is also a case where the categories of the contingency table is expanded to $K \times L$. The corresponding correlation coefficient of multiple categorical data is polychoric correlation coefficient.

2.6 Polychoric correlation coefficient

When both X and Y are ordinal data, i.e. X = 1st, 2nd, …, nth and Y = 1st, 2nd, …, nth , and the categories of the data exceeds two, the use of polychoric correlation is suggested (Holgado-Tello *et al.*, 2010). Polychoric correlation is an estimate of the correlation between two unobserved variables X and Y where both X and Y are continuous by using the observed variables X^* and Y^* as the basis. The variables X^* and Y^* are ordinal variables that are assumed to follow bivariate normal distribution.

First, collect the observations of X^* and Y^*. These are ordinal data. The values of X^* and Y^* are known as the *underlying* or *latent variables*. These variables cannot be measured by direct observations. For instance, IQ test is a score obtained from a test battery intended to measure the level of intelligence. IQ test scores are the observed data X^* and Y^*; intelligence is the unobservable X and Y. In this case, IQ is the score from an indirect test used to determine intelligence; however, the purpose of the illustration here is to give an example of a latent variable. These observations may be given as:

$$X^* = \begin{bmatrix} \overset{*}{x_1} \\ \overset{*}{x_2} \\ \vdots \\ \overset{*}{x_n} \end{bmatrix} \quad \text{and} \quad Y^* = \begin{bmatrix} \overset{*}{y_1} \\ \overset{*}{y_2} \\ \vdots \\ \overset{*}{y_n} \end{bmatrix}$$

Thus, $X^* = \{\overset{*}{x_1}, \overset{*}{x_2}, ..., \overset{*}{x_n}\}$ and $Y^* = \{\overset{*}{y_1}, \overset{*}{y_2}, ..., \overset{*}{y_n}\}$. These are observed data. Assume that there are discrete and random variables X and Y (note that there is no asterisk marking this “unobserved” set) which relates to X^* and Y^* where

$$x_{i=k} \quad \text{if} \quad \xi_{k-1} < \overset{*}{x_i} \le \xi_k \quad \text{and} \tag{37}$$

$$y_{i=l} \quad \text{if} \qquad \eta_{l-1} < y_i^* \le \eta_l \tag{38}$$

For x_i the threshold ξ_k comes from $\xi_1, \xi_2, ..., \xi_k$ and $\xi_0 = -\infty$ and $\xi_k = +\infty$. Similarly, the threshold for y_i is given by η_l which comes from $\eta_1 + \eta_2 + ... + \eta_l$ and $\eta_0 = -\infty$ and $\eta_l = +\infty$. The corresponding values of the observed $X*$ and $Y*$ and the unobserved X and Y may be represented as:

$$X* \to X = \begin{bmatrix} x_1^* \to x_1 \\ x_2^* \to x_2 \\ \vdots \\ x_n^* \to x_k \end{bmatrix} \qquad \text{and} \qquad Y* \to Y = \begin{bmatrix} y_1^* \to y_1 \\ y_2^* \to y_2 \\ \vdots \\ y_l^* \to y_\eta \end{bmatrix}$$

The joint distribution of the unobserved X and Y is given by:

$$P[x = k] = p_k \qquad \text{for } X \text{ and} \tag{39}$$

$$P[y = l] = q_l \qquad \text{for } Y \tag{40}$$

The cloud of the unobserved variables $X*$ and $Y*$ as defined by x_i^* and y_i^* may be projected onto a space, thus:

$$\begin{bmatrix} \\ \pi_{11} & \pi_{12} & \cdots & \pi_{1L} \\ \pi_{12} & \pi_{22} & \cdots & \pi_{2L} \\ \vdots & \vdots & \vdots & \vdots \\ \pi_{K1} & \pi_{K2} & \cdots & \pi_{KL} \end{bmatrix} \tag{41}$$

The term π_{KL} is called the *discrete cell proportion*. The objective of polychoric correlation is to find the value for π_{KL}. Recall that π is the joint probability of X and Y pair that is not observed.

With the above set up, polychoric correlation can now be discussed. Polychoric correlation is developed as the result of the inadequacy of the tetrachoric correlation to handle a $K \times L$ contingency table;[1] recall that the tetrachoric correlation is confined to a 2×2 contingency table. In

[1] Ritchie-Scott used the notation as $r \times s$ in labeling the contingency matrix. Zoran Pasoric and Josip Juras uses $K \times L$. Common statistics designated the multivariable contingency table as $K \times K$. In general, the $K \times K$ or $K \times L$ contingency table is provided as:

OBSERVATIONS			
C_1	C_2	$\cdots$	C_K

1900, Pearson (Pearson, 1900) introduced tetrachoric correlation calculation as an attempt to obtain a quantitative measurement of a continuous variable. However, Pearson's earlier attempt was confined to 2×2 scenario. The work was further expanded by Ritchie-Scott (Ritchie, 1918) to cover a $K\times L$ scenario which became known as polychoric correlation today. Where as Pearson's 2×2 tetrachoric handles dichotomous data, Ritchie-Scott's $K\times L$ handles polytomous tests.

The observed array $X*$ and $Y*$ are assumed to be bivariate normal, and the correlation between $X*$ and $Y*$ is given by ρ (*rho*) given series of unobserved data set of x_i^* and y_i^*. The objective of polychoric correlation is to obtained the correlation of the unobserved arrays X and Y from the product-moment or correlation of the observed arrays $X*$ and $Y*$ where x_i^* and y_i^* is jointly normally distributed. The polychoric correlation is estimated from the discrete cell proportion π_{KL}; thus, this estimated value is designated as $\hat{\pi}_{KL}$ from the $K\times L$ contingency table (8.5.0).

The *probability density function* (PDF) of the bivariate normal $X*$ and $Y*$ is given by:

$$\phi\left(x^*,y^*;\rho\right)=\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{x^{*2}-2\rho x^*y^*+y^*}{2\left(1-\rho^2\right)}\right] \tag{42}$$

The *cumulative distribution function* (CDF) for the bivariate normal $X*$ and $Y*$ is given by:

$$\Phi_2\left(\xi_i,\eta_i;\rho\right)=\int_{-\infty}^{\xi_1}\int_{-\infty}^{\xi_1}\phi\left(x^*,y^*;\rho\right)dy^*dx^* \tag{43}$$

for $x_i=1$ and $y_i=1$, the probability is:

FORECAST	C_1	P_{11}	P_{12}	$\cdots$	P_{2K}	P_{F1}
	C_2	P_{21}	P_{22}	$\cdots$	P_{2K}	P_{F2}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	P_{KK}	$\cdots$
	C_K	P_{K1}	P_{K2}	$\cdots$	P_{KK}	P_{FK}
		$P_{obs,1}$	$P_{obs,2}$	$\cdots$	$P_{obs,K}$	Polychoric Joint Probability

Bias = $\left(P_{F,1}/P_{1\square},\ldots,P_{F,K-1}/P_{0,K-1}\right)$ and $P_{obs}=\left(P_{obs,1},\ldots,P_{obs,k-1}\right)$. The notation above uses: $P_{row,column}$. Note that there is switching of observation and forecast above. This alternation does not change the number of interpretation of the result.

$$\pi_{11} = \phi(\xi_1, \eta_1; \rho) \tag{44}$$

The probability of $x_i = 1$ and $y_i = 1$ is a function of ρ. Recall that ρ is the correlation of the observed $X*$ and $Y*$ where the threshold is (ξ_1, η_1). From the cumulative distribution function Φ_2, the following generalization may be made:

$$h_{kl}(\theta) = \Phi_2(\xi_k, \eta_l) - \Phi_2(\xi_{k-1}, \eta_l) - \Phi_2(\xi_k, \eta_{l-1}) + \Phi_2(\xi_{k-1}, \eta_{l-1}) \tag{45}$$

The term h_{kl} stands for the *likelihood* of event *k* and *l* occurring and this likelihood is a function of theta θ and θ is given by:

$$\theta = [\rho, \theta_1] = [\rho, \xi_1, ..., \xi_{K-1}, \eta_1, ..., \eta'_{L-1}] \tag{46}$$

The *likelihood for the discrete cell proportion* is written as:

$$\pi_{kl} = h_{kl}(\theta) \tag{47}$$

A general statement may now be made about π; now let $\pi = [\pi_{11}, ..., \pi_{KL}]'$ and the likelihood of θ may be generally stated as $h[\theta] = [h_{11}(\theta), ..., h_{KL}(\theta)]'$. Now, the polychoric equation may be written as:

$$\pi = h(\theta) \tag{48}$$

Olsson (Olsson, 1979) provides a close estimate of π as the maximum log likelihood which is equivalent to the estimated theta $\hat{\theta}$. Olsson's maximum log likelihood is given by:

$$\ln L = \sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{kl} \log h_{kl}(\theta) \tag{49}$$

Recall that theta θ is the likelihood function. There are three kinds of maximum likelihood functions used according to the type of data distribution: (i) Bernoulli distribution, (ii) normal distribution, and (iii) Poisson distribution. Polychoric correlation deals with ordinal data. Ordinal data is ranked data set. The appropriate form of maximum likelihood function type is one that is used for normal distribution. The maximum likelihood function for normal distribution is provided thus:

$$f(x_1, x_2, ..., x_n \mid \mu, \sigma) = \Pi \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \tag{50}$$

which may be make into a general statement as:

$$f(x_1, x_2, ..., x_n \mid \mu, \sigma) = \frac{(2\pi)^{-\pi/2}}{\sigma^n} \exp\left[-\frac{\Sigma(x_i - \mu)^2}{2\sigma^2}\right] \tag{51}$$

The maximum likelihood is express as the natural log of the function; therefore,

$$\ln f = -\frac{1}{2} n \ln(2\pi) - n \ln \sigma - \frac{\Sigma\left(x_i - \mu\right)^2}{2\sigma^2} \tag{52}$$

To find the expected mean of the function, the derivative of the maximum likelihood function is taken by:

$$\frac{\partial(\ln f)}{\partial \mu} = \frac{\Sigma\left(x_i - \mu\right)}{\sigma^2} = 0 \quad \text{which gives the expected mean as:} \tag{53}$$

$$\hat{\mu} = \frac{\Sigma x_i}{n} \tag{54}$$

Using the same rationale, the expected standard deviation follows:

$$\frac{\partial\left(\ln f\right)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\Sigma\left(x_i - \mu\right)^2}{\sigma^3} = 0 \tag{55}$$

$$\hat{\sigma} = \sqrt{\frac{\Sigma\left(x_i - \hat{\mu}\right)^2}{n}} \tag{56}$$

The above steps obtained the maximum likelihood of mean and standard deviation as the mean and standard deviation of the sample. This may be a biased estimate; nevertheless, for purposes of demonstrating how the maximum likelihood is calculated in the context of the maximum likelihood of $\hat{\pi}_{KL}$, it is an adequate explanation.

2.7 Rank biserial correlation coefficient

In a case where Y is dichotomous and X is rank data, rank biserial correlation coefficient is used. The formula for the rank biserial correlation is given by (Glass and Hopkins, 1995):

$$r_{rb} = 2\left(\frac{M_1 - M_0}{n_1 + n_0}\right) \tag{57}$$

where the subscripts 1 and 0 refers to the score of 1 and 0 in the 2×2 contingency table; M is the mean of the frequency of the scores, and n is the sample size. The null hypothesis is that $r_{rb} = 0$, meaning there is no correlation. If the null hypothesis is true, the data is distributed as Mann-Whitney U.

The objective of the Mann-Whitney U Test is to verify the claim that the standard deviation of population A is the same as the standard deviation of population B; if so, then the two populations are identical, except for their locations, i.e. the populations in two cities have the same income. The case involves two population located at a different place; this is a case that could be termed *parallel group*. The claim by the alternative hypothesis (H_A) is that the two populations are the same and have the same population standard deviation. The logic follows that "if the two standard deviations are the same, there difference, i.e. $\sigma_1 - \sigma_2 = 0$, must equal to zero." The null hypothesis (H_0) states the obverse: "the two populations are different; their means are different. Therefore, $\sigma_1 - \sigma_2 \neq 0$."

The procedure for conducting the Mann-Whitney U test involves five steps. Each step is explained below thus:

1. Collect a sample from each population. The sample size of the two samples may be equal or unequal. Mark one sample as n_1 and the second sample n_2. It does not mater which one is designated as the *first* or the *second* sample. However, conventional practice dictates that treat the largest sample as n_1 and the smaller sample as n_2.
2. Combine the two samples into one array, i.e. one set as shown below:

$$n = n_1 + n_2 \tag{58}$$

3. Rank the combined sample (n) in an ascending order, i.e. from low to high so that the elements of the set is arranged as: $n1 < n2, \ldots < nN$.

4. calculate the test statistic for the Mann-Whitney U test according to the formula below:

$$Z = W_1 - \left(\frac{\frac{n_1(n_1 + n_2 + 1)}{2}}{S_w} \right) + C \tag{59}$$

where …

$$W_1 = \sum_{k=1}^{n_1} Rank(X_{lk}) \tag{60}$$

This (W_1) is called the rank sum. The standard deviation of the ranked set is given by:

$$S_w = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} - \left(\frac{n_1 n_2 \left(\sum_{i=1} t_i^3 - t_i \right)}{12(n_1 + n_2)(n_1 + n_2 - 1)} \right) \tag{61}$$

where t_1 = number of observations tied at value one;

t_2 = number of observations tied at value two, and so on.

C = correction factor. This number is fixed at 0.50 if the numerator if Z is *negative* and -0.50 if the numerator of Z is *positive*.

5. Use the following decision rule to determine whether to accept or reject the null hypothesis:

$H_A : \sigma = 0$, the decision rule is governed by $Z < Z_{\alpha/2}$ or $Z > Z_{1-_{\alpha/2}}$.
$H_A : \sigma > 0$, the decision rule is governed by $Z > Z_{1-_{\alpha/2}}$; and
$H_A : \sigma < 0$, the decision rule is governed by $Z < Z_{\alpha}$.

The Mann-Whitney U test is used in the following cases: (i) the test involves the comparison of two populations; (ii) the values of the data is non-parametric; therefore, it is an alternative test to the conventional t-test; (iii) the data is classified as *ordinal* and NOT interval scale. "Ordinal" means that the data score, i.e. answer choice, has the spacing between each score unit is unequal or non-constant. For example, a scale of 1 (lowest) to 5 (highest) would not be able to use the Mann-Whitney U test. Whereas, 1st place, 2nd place, and 3rd place type of answer choice, where the distance between the first, second, and third are not equal, may be appropriate for this test; and (iv) it is said that the Mann-Whitney test is more *robust* and *efficient* than the conventional t-test.

"Robustness" (Portnoy and He, 2000) means that the final result is not unduly affected by the outliers. *Outliers* are extreme value. If the system is robust, it will not be affected by extreme value. Generally, extreme value tends to create bias estimate by the estimator because outliers or extreme values creates greater variance and thus larger standard deviation. This problem is eliminated through the use of *ranking* the data by arranging the combined sets of $n = n_1 + n_2$ into one set ranking from lowest value to highest value.

"Efficiency" is the measure of the desirability of the estimator. The estimator is desirable if it yields an optimal result. It yields the optimal result it the observed data meets or comes closest to the expected value
(Everitt, 2002).

Recall that the conventional t-test is given by:

$$t = \frac{\overline{x} - \mu}{S / \sqrt{n}} \qquad (62)$$

The Mann-Whitney U Test requires the comparison of the population standard deviations $\sigma_1 - \sigma_2 = 0$. Recall further that in order to determine the population standard deviation one must use the Z-equation. The Z equation is given by:

$$Z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}} \qquad (63)$$

From equation (43), the population standard deviation may be written as:

$$\sigma = \left(\frac{\overline{x} - \mu}{Z} \right) \sqrt{n} \qquad (64)$$

Note that the population standard deviation in equation (64) may not be determined unless the conventional t-equation (62) is used to determine the population mean (μ). The value of μ is derived from equation (62) as:

$$\mu = t \left(\frac{S}{\sqrt{n}} \right) - \overline{x} \qquad (65)$$

Therefore, even the Mann-Whitney U test statistic, equation (59), shows no use of the t-equation and Z-equation, the researcher must understand the underlying functions and steps to illustrate the logic of the Mann-Whitney U Test.

2.8 Phi correlation coefficient

In case where the data of X and Y are both nominal, the *phi* correlation coefficient is used. The *phi* equation is given by:

$$r_{phi} = \phi = \frac{p_a - p_X p_Y}{\sqrt{\left(p_X p_Y (1-p_X)(1-p_Y)\right)}} \tag{66}$$

There is an equivalence of equation (66) by contingency coding method of blocks ABCD in the table:

$$r_{phi} = \frac{(BC - AD)}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \tag{67}$$

The calculation according to equation (10.67) follows:

$$r_{phi} = \frac{(BC - AD)}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

$$r_{phi} = \frac{(25-100)}{\sqrt{(15)(15)(15)(15)}} = \frac{-75}{\sqrt{50,625}} = \frac{-75}{225}$$

$$r_{phi} = -0.33$$

Note that equations (66) and (67) is equivalent to:

$$\phi = \sqrt{\frac{\chi^2}{n}} \tag{68}$$

where $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ or $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$.

2.9 Pearson contingency C

Another case where both X and Y are nominal data, the Pearson contingency coefficient is used. This is known as Pearson's C which is given by:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \tag{69}$$

where χ^2 is chi square and N is the grand total of observations. Generally, the range for correlation coefficient is -1 and +1; however, C does not reach this range 9Pearson, 1904; p. 16). For a 2×2 table, it can reach 0.707 and 0.870 for 4×4 table. In order to reach the interval maximum, more categories has to be added (Smith and Albaum, 2004).

2.10 Goodman and Kruskal lambda

The Goodman and Kruskal's lambda is a measurement of reduction in error ratio. This type of correlation measurement is used for the measurement of association. To the extent that it is applicable to "reliability," GK's lambda is usable only if reliability is defined in terms of association of polytomies, i.e. the answer to the survey question contains more than 2 choices. The lambda equation is given by:

$$\lambda = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1} \qquad (70)$$

where ε_1 is the overall non-modal frequency; and ε_2 is the sum of the non-modal frequencies for each value of independent variable. The range of lambda is $0 \le \lambda \le 1$. Zero means there is no association between the independent and dependent variables, and one means there is a perfect association between the two variables.

Goodman and Kruskal deals with optimal prediction of two polytomies (multiples) which are asymmetrical where there is no underlying continua and no ordering of interest (Goodman and Kruskal, 1954). The Goodman and Kruskal (GK) polytomy is described by $A \times B$ crossing in the table below:

Table 6. Measure of association under Goodman-Kruskal method

A	B				
	B_1	B_2	$\cdots$	B_β	Total
A_1	ρ_{11}	ρ_{12}	$\cdots$	$\rho_{1\beta}$	$\rho_{1\bullet}$
A_2	ρ_{21}	ρ_{22}	$\cdots$	$\rho_{2\beta}$	$\rho_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_α	$\rho_{\alpha 1}$	$\rho_{\alpha 2}$	$\cdots$	$\rho_{\alpha\beta}$	$\rho_{\alpha\bullet}$
Total	$\rho_{\bullet 1}$	$\rho_{\bullet 2}$	$\cdots$	$\rho_{\bullet\beta}$	1

In Table 6, A divides the population into *alpha* (α) classes where $\alpha : (A_1, A_2, ..., A_\alpha)$. Similarly, B divides the population into *beta* (β) classes where $\beta : (B_1, B_2, ..., B_\beta)$. The proportion that classified as both A_α and B_β is $\rho_{\alpha\beta}$. The marginal proportion $\rho_{\alpha\bullet}$ is the proportion of the population classified as A_α and $\rho_{\bullet\beta}$ is the proportion of the population classified as B_β. *See* Goodman & Kruskal (1954), p. 734.

Goodman and Kruskall originally proposed the measure of association as:

$$\lambda_b = \frac{P(e_1) - P(e_2)}{P(e_1)} \qquad (71)$$

which can be written as:

$$\lambda_b = \frac{\sum_a \rho_{am} - \rho_{\bullet m}}{1 - \rho_{\bullet m}} \qquad (72)$$

The expression above is the relative decrease in probability of error from B_b as between A_a unknown and A_a known. The value λ_b gives the error proportion which can be eliminated when A is known. Goodman and Kruskal defined λ_α as:

$$\lambda_a = \frac{\sum_b \rho_{mb} - \rho_{m\bullet}}{1 - \rho_{m\bullet}} \qquad (73)$$

where: $\rho_m = \underset{a}{Max}\, \rho_{a\bullet}$ and $\rho_{mb} = \underset{a}{Max}\, \rho_{ab}$

The interpretation of the meaning of λ_a is opposite of λ_b. The meaning of λ_a is "the relative decrease in probability of error in guessing A_a as between B_b unknown and known" (Goodman and Kruskal, p. 742). Goodman and Kruskal stated that the value of λ_a and λ_b were given by Guttman (Guttman, 1941) from which they derived the following lambda:

$$\lambda = \frac{0.5\left[\sum_a \rho_{am} + \sum_b \rho_{mb} - \rho_{\bullet m} - \rho_{m\bullet}\right]}{1 - 0.5\left(\rho_{\bullet m} + \rho_{m\bullet}\right)} \qquad (74)$$

The lambda proposed by Goodman and Kruskal lies between λ_a and λ_b described by Guttman. The range of GK's lambda is $0 \le \lambda \le 1$. Goodman and Kruskal alternatively the terms in lambda by "[l]et v be the total number of individuals in the population, $v_{ab} = v\rho_{ab}, v_{am} = v\rho_{am}, v_{mb} = v\rho_{mb}$, and so on" 9Goodman and Kruskal, p. 743). Under this general definition, the Guttment λ_a and λ_b becomes:

$$\lambda_b = \frac{\sum_a v_{am} - v_{\bullet m}}{v - v_{\bullet m}} \qquad \text{and} \qquad (75)$$

$$\lambda_a = \frac{\sum_b v_{mb} - v_{m\bullet}}{v - v_{m\bullet}} \qquad (76)$$

The general GK's lambda then is given by:

$$\lambda = \frac{\sum_a v_{am} + \sum_b v_{mb} - v_{\bullet m} - v_{m\bullet}}{2v - \left(v_{\bullet m} + v_{m\bullet}\right)} \qquad (77)$$

The following table demonstrates GK's new definition of the population and its components.

Table 7. Example of Goodman-Kruskal measure of association

A	B				
	B_1	B_2	B_3	B_4	$v_{a\square}$
A_1	1768	807	189	47	2811
A_2	946	1387	746	53	3132
A_3	115	438	288	16	857
$v_{\square b}$	2829	2632	1223	116	$v = 6800$

The numerical examples in Table 7 are taken from Kendall's work and also was reproduced in Goodman and Kruskal's article (p. 744). *See* Kendall, Maurice G. (1948). *The Advanced Theory of Statistics*, London, Charles Griffin and Co., Ltd. 1948; p. 300.

Goodman and Kruskal provided the calculation as:

$$\begin{array}{ll}
v_{1m} = 1768 & v_{m1} = 1768 \\
v_{2m} = 1387 & v_{m2} = 1387 \\
v_{3m} = 438 & v_{m3} = 746 \\
 & v_{m4} = 53 \\
v_{\square m} = 2829 & v_{m\square} = 3132
\end{array}$$

The calculation for Guttman's λ_a and λ_b follows:

$$\lambda_a = \frac{3954-3132}{6800-3132} = \frac{822}{3668} = 0.2241$$

$$\lambda_b = \frac{3593-2829}{6800-2829} = \frac{764}{3971} = 0.1924$$

The calculation for GK's lambda follows:

$$\lambda = \frac{822+764}{3668+3971} = \frac{1586}{7639} = 0.2076$$

It is tempted to treat GK's lambda λ as the average of Guttman's λ_a and λ_b; however, the following calculation shows:

$$\lambda = \frac{\lambda_a + \lambda_b}{2} = \frac{0.2241+0.1924}{2} = \frac{0.4165}{2} = 0.2083$$

There is a minor difference of $0.2083 - 0.2076 = 0.0006$ or 0.065%. It is a good approximation. For that reason, GK's lambda may be preferential to Guttman's λ_a and λ_b which requires a two-step process.

In reliability test, the GK's lambda is a tool to measure the degree of reliability through the interpretation of the reduction of the error ratio. If the error ratio is reduced, it is said that the study is reliable. Is this reliability test relevant to the instrument itself or the entire score set produced by the survey? It is worth noting that the issue of reliability, when it relates to the survey, fulfills the requirement of replication. For instrument assessment, the issue of reliability attests to the efficacy of the instrument, i.e. does it give predictable result or scores? GK's lambda, as well as Guttman's λ_a and λ_b, measures association between two polytomies: A_a and B_b. On the issue of instrumentation, GK's lambda is not a tool for instrument calibration. On the issue of survey reliability, GK's is not a tool for testing the reliability of the survey. GK's lambda is a tool to measure association of two random variables. Only if reliability is measured as the degree of association would GK's lambda be a usable tool for reliability analysis.

3.0 CONCLUSION

This paper underscores the importance of the use of correct type of correlation coefficient. Correlation coefficient measures the level of association between variables. If the type of correlation is not correct, the inference made from it would also be faulty: a case of Type 2 error. The correct type of correlation coefficient depends on the type of data of the variables. Assume that the relationship is captured by the dependent (Y) and independent (X) variables, the correct correlation coefficient to use depends on two questions: what type of data is Y? and what type of data is X? Data are classified into three types: quantitative, ordinal and nominal. The crossing of X-by-Y determines the type of correlation coefficient. Although this is not a new knowledge, this paper has put all three possible pairs of QON into context and explained how each type of correlation coefficient is calculated. This attempt helps facilitate correct calculation of association among variables.

REFERENCES

Boddy, Richard; Smith, Gordon (2009). *Statistical methods in practice: for scientists and technologists*. Chichester, U.K.: Wiley. pp. 95–96. ISBN 978-0-470-74664-6.

Choi, S. C. (1977). "Tests of Equality of Dependent Correlation Coefficients." *Biometrika* **64**(3): 645–647.

Dauben, J. W. (1990). *Georg Cantor: His Mathematics and Philosophy of the Infinite*. Princeton, NJ: Princeton University Press; p. 199.

Everitt, Brian S. (2002). *The Cambridge Dictionary of Statistics*. Cambridge University Press. ISBN 0-521-81099-X. p. 128.

Fieller, E. C.; Hartley, H. O.; Pearson, E. S. (1957). "Tests for rank correlation coefficients. I." *Biometrika* **44**; 470–481.

Guttman, Luis (1941). "An outline of the statistical theory of prediction," Supplementary Study B-1, pp. 253-318 in Horst, Paul *et al.* (eds.), The Prediction of Personal Adjustment, Bulletin 48, Social Science Research Council, New York (1941).

Goodman, Leo A. & Kruskal, William H. (1954). "Measure of Association for Cross Classifications." *Journal of the American Statistical Association*, Vol 49, No. 268 (Dec 1954), pp. 732-764.

Holgado–Tello, F.P., Chacón–Moscoso, S, Barbero–García, I. and Vila–Abad E (2010). "Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables." *Quality and Quantity* **44**(1), 153-166.

Johnson, N.L. and Kotz, S. (1972). *Distributions in Statistics. 4. Continuous Multivariate Distributions*. Wiley, New York, p. 333.

Juras and Pasaric, p. 67 citing Sheppard (1998). "On the application of the theory of error to cases of normal distribution and normal correlations." *Philos. Tr. R. Soc. S.-A*, 192, 101-167.

Kendall, M. G.; Stuart, A. (1973). *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*. Griffin. ISBN 0-85264-215-6. (Sections 31.19, 31.21).

Kowalczyk, T.; Pleszczyńska E. , Ruland F. (eds.) (2004). "Grade Models and Methods for Data Analysis with Applications for the Analysis of Data Populations." *Studies in Fuzziness and Soft Computing* 151. Berlin Heidelberg New York: Springer Verlag. ISBN 978-3-540-21120-4.

Lehman, Ann (2005). *For Basic Univariate And Multivariate Statistics: A Step-by-step Guide*. Cary, NC: SAS Press. p. 123. ISBN 1-59047-576-3.

Laplace, Pierre-Simon (1814). Essai philosophique sur les probabilities. Paris: Courcier.

Olsson, U. (1979). "Maximum likelihood estimation of the polychoric correlation coefficient." *Psychometriko*, **44**: 443-460.

Pearson, Karl (1904). "Mathematical Contributions to the Theory of Evolution. XIII. On the Theory of Contingency and Its Relation to Association and Normal Correlation." *Biometric Series* I. Draper's Company Research Memoirs; pp.. 5-21.

Pearson, K (1900). "Mathematical contributions to the theory of evolution. Vii. On the correlation of characters not quantitatively measurable." Philosophical Transaction of the Royal Society of London. Series A, 195: 1-47.

Portnoy S. and He, X. (2000). "A Robust Journey in the New Millennium," Journal of the American Statistical Association Vol. 95, No. 452 (Dec., 2000), 1331–1335.Glass, G.V. and Hopkins, K.D. (1995). *Statistical Methods in Education and Psychology* (3rd ed.). Allyn & Bacon. ISBN 0-205-14212-5.

Press; Vettering; Teukolsky; Flannery (1992). *Numerical Recipes in C: The Art of Scientific Computing* (2nd ed.). p. 640.

Ritchie-Scott, A. (1918). "A new measure of rank correlation." *Biometrika*, **30**:81-93.

Maritz, J. S. (1981). *Distribution-Free Statistical Methods*. Chapman & Hall. p. 217. ISBN 0-412-15940-6.

Moore, G. H. (1982). *Zermelo's Axiom of Choice: Its Origin, Development, and Influence*. New York: Springer-Verlag; p. 52.

Myers, Jerome L.; Well, Arnold D. (2003). *Research Design and Statistical Analysis* (2nd ed.). Lawrence Erlbaum. p. 508. ISBN 0-8058-4037-0.

Smith, S. C., & Albaum, G. S. (2004). *Fundamentals of marketing research*. Sage: Thousand Oaks, CA. p. 631.

Shermer, Michael (2002). *The Skeptic Encyclopedia of Pseudoscience 2 volume set*. ABC-CLIO. p. 455. ISBN 1-57607-653-9. Retrieved 10 January 2011.

Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons." Journal of the American Statistical Association 69: 730–737.

Stephens, M.A. (1986). "Tests Based on EDF Statistics." In D'Agostino, R.B. and Stephens, M.A. Goodness-of-Fit Techniques. New York: Marcel Dekker. ISBN 0-8247-7487-6.

Suppes, P. (1972). *Axiomatic Set Theory*. New York: Dover; p. 129.