Title
The Phi-coefficient, the Tetrachoric Correlation Coefficient, and the Pearson-Yule Debate

Author
Ekström, Joakim

Publication Date
2011-10-25

Peer reviewed

THE PHI-COEFFICIENT, THE TETRACHORIC CORRELATION COEFFICIENT, AND THE PEARSON-YULE DEBATE

JOAKIM EKSTRÖM

Abstract. Two measures of association for dichotomous variables, the *phi*-coefficient and the tetrachoric correlation coefficient, are reviewed and differences between the two are discussed in the context of the famous so-called Pearson-Yule debate, that took place in the early 20th century. The two measures of association are given mathematically rigorous definitions, their underlying assumptions are formalized, and some key properties are derived. Furthermore, existence of a continuous bijection between the *phi*-coefficient and the tetrachoric correlation coefficient under given marginal probabilities is shown. As a consequence, the tetrachoric correlation coefficient can be computed using the assumptions of the *phi*-coefficient construction, and the *phi*-coefficient can be computed using the assumptions of the tetrachoric correlation construction. The efforts lead to an attempt to reconcile the Pearson-Yule debate, showing that the two measures of association are in fact more similar than different and that between the two, the choice of measure of association does not carry a substantial impact on the conclusions of the association analysis.

Key words and phrases. Phi-coefficient, Tetrachoric Correlation Coefficient, 2×2 Contingency Tables, Measures of Association, Dichotomous Variables.

Financial support from the Jan Wallander and Tom Hedelius Research Foundation, project P2008-0102:1, is gratefully acknowledged.

(a) Karl Pearson (1857-1936) (b) George Udny Yule (1871-1951)

FIGURE 1. Pearson portrait is from Pearson (1938), and is in the public domain. Yule portrait is from Yule et al. (1971), reproduced with the kind permission of Hodder & Stoughton.

1. Introduction

The *phi*-coefficient and the tetrachoric correlation coefficient are two measures of association for dichotomous variables. The association between variables is of fundamental interest in most scientific disciplines, and dichotomous variables occur in a wide range of applications. Consequently, measures of association for dichotomous variables are useful in many situations. For example in medicine, many phenomena can only be reliably measured in terms of dichotomous variables. Another example is psychology, where many conditions only can be reliably measured in terms of, for instance, *diagnosed* or *not diagnosed*. Data is often presented in the form of 2×2 contingency tables. A historically prominent example is Pearson's smallpox recovery data, see Table 1, studying possible association between vaccination against, and recovery from, smallpox infection. Another interesting data set is Pearson's diphtheria recovery data, Table 2, studying possible association between antitoxin serum treatment and recovery from diphtheria.

Measures of association for dichotomous variables is an area that has been studied from the very infancy of modern statistics. One of the first scholars to treat the subject was Karl Pearson, one of the fathers of modern statistics. In the 7th article in the seminal series *Mathematical contributions to the theory of evolution*, Pearson (1900) proposed what later became known as the tetrachoric correlation coefficient, as well as, Pearson would later argue, the *phi*-coefficient. The fundamental idea of the tetrachoric correlation coefficient is to consider the 2×2 contingency table as a double dichotomization of a bivariate standard normal distribution, and then to solve for the parameter such that the volumes of the dichotomized bivariate standard normal distribution equal the joint

FIGURE 2. Care at the Hampstead fever hospital, London 1872. One of many hospitals opened for the sick poor by the Metropolitan Asylums Board in the late 19th century. With the kind permission of workhouses.org.uk.

probabilities of the contingency table. The tetrachoric correlation coefficient is then defined as that parameter, which, of course, corresponds to the linear correlation of the bivariate normal distribution.

According to Pearson's colleague Burton H. Camp (1933), Pearson considered the tetrachoric correlation coefficient as being one of his most important contributions to the theory of statistics, right besides his system of continuous curves, the chi-square test and his contributions to small sample statistics. However, the tetrachoric correlation coefficient suffered in popularity because of the difficulty in its computation. Throughout his career, Pearson published statistical tables aimed at reducing that difficulty (Camp,

TABLE 1. Karl Pearson's smallpox recovery data.

	Recovery	Death	
Vaccinated	1562	42	1604
Unvaccinated	383	94	477
	1945	136	2081
Pearson's chi-square test for independence			
$\chi^2_{obs} = 176$ p-value < 0.0001			
Measures of association			
$r_{phi} = 0.3$ $r_{tc} = 0.6$			

Source: Metropolitan Asylums Board: Small-pox epidemic 1893. (Pearson, 1900)

TABLE 2. Karl Pearson's diphtheria recovery data.

	Recovery	Death	
With antitoxin	319	143	462
Without antitoxin	177	289	466
	496	432	928

Pearson's chi-square test for independence

$\chi^2_{obs} = 90$ p-value < 0.0001

Measures of association

$r_{phi} = 0.3$ $r_{tc} = 0.5$

Source: Metropolitan Asylums Board: Use of Antitoxin Serum 1896. (Pearson, 1900)

1933), reflecting an interest in promoting a wider adoption of the tetrachoric correlation coefficient among practitioners.

While the tetrachoric correlation coefficient is the linear correlation of a so-called *underlying* bivariate normal distribution, the *phi*-coefficient is the linear correlation of an underlying bivariate discrete distribution. This measure of association was independently proposed by Boas (1909), Pearson (1900), Yule (1912), and possibly others.

The question of whether the underlying bivariate distribution should be considered continuous or discrete is at the core of the so-called Pearson-Yule debate. In the historical context of the Pearson-Yule debate, though, it is important to understand that no one at the time looked upon these two measures of association as the linear correlations of different underlying distributions, the framework in which both were presented in the preceding paragraph. On the contrary, according to Yule (1912) the tetrachoric correlation coefficient is founded upon ideas entirely different from those of which the *phi*-coefficient is founded upon. The sentiment is echoed by Pearson & Heron (1913), which even claims that the *phi*-coefficient is not based on a reasoned theory, while at the same time arguing for the soundness of the tetrachoric correlation coefficient. In fact, the point of view that both measures of association are the linear correlations of underlying distributions is one of the contributions of the present article.

1.1. **The Pearson-Yule debate.** George Udny Yule, a former student of Pearson, favored the approach of an inherently discrete underlying distribution. Yule (1912) is a comprehensive review of the area of measures of association for dichotomous variables, as well as a response to Heron (1911), and contains blunt criticism of Pearson's tetrachoric correlation coefficient. Regarding the tetrachoric correlation coefficient's assumptions of underlying continuous variables, Yule (1912) reads:

> Here, I am concerned rather with the assumptions and their applicability. [...] Those who are unvaccinated are all equally non-vaccinated, and similarly, all those who have died of small-pox are all equally dead. [...] From

> this standpoint Professor Pearson's assumptions are quite inapplicable, and do not lead to the true correlation between the attributes. But this is not, apparently, the standpoint taken by Professor Pearson himself.

The example that Yule (1912) referes to is the smallpox recovery data which was prominently featured in Pearson (1900), see Table 1.

Yule (1912) also contains a bibliographical discussion which could be interpreted as a questioning of whether Pearson really is the originator of some of the ideas that Pearson claimed credit for. In all, Pearson quite evidently felt offended by some of Yule's wordings and was upset by his former student's publicly expressed, and in Pearson's opinion uninformed, misgivings about the tetrachoric correlation coefficient. And from there on, it is by most accounts fair to say that the debate lost all proportions.

Pearson & Heron (1913) is a scathing, almost 200 pages long reply. The introduction reads:

> The recent paper by Mr Yule calls for an early reply on two grounds, first because of its singularly acrimonious tone [...], and secondly because we believe that if Mr Yule's views are accepted, irreparable damage will be done to the growth of modern statistical theory. Mr Yule has invented a series of methods which are in no case based on a reasoned theory, but which possess the dangerous fascination of easy application [...], and therefore are seized upon by those who are without adequate training in statistical theory.

With regards to the smallpox recovery example, Pearson & Heron (1913) replies:

> Recovery and death in cases of small-pox were used to measure a continuous variable - the severity of the attack. [Moreover,] vaccination regarded as conferring immunity is an essentially continuous variable.

With respect to Yule's contrasting view of the dichotomous variables as inherently discrete, while still unidimensional, Pearson & Heron (1913) rhetorically counter-asks:

> Does Mr Yule look upon death as the addition of one unit to recovery?

Pearson may also have taken offense at the fact that Yule wrote a review on one of the regarded Professor's favorite topics. Pearson & Heron (1913) mentions Yule's statistical textbook on several occasions.

> It may be said that a vigorous protest against Mr Yule's coefficient is unnecessary. We believe on the contrary that, if not made now and made strongly, there will be great set-back to both modern statistical theory and practice. The publication of Mr Yule's text-book has resuscitated the use of his coefficient of association; it is now being used in all sorts of quarters on all sorts of unsuitable data. The coefficient of association is in our opinion wholly fallacious, it represents no true properties of the actual distribution, and it has no adequate physical interpretation.

The exchange became known as the Pearson-Yule debate. The tone was indeed caustic, many readers likely felt intimidated by the gravity of the accusations, and Camp (1933) acknowledges that it may have contributed to Pearson's reputation of being unkind. Though in the end, it is important to point out, Yule wrote Pearson's obituary for the Royal Society (Yule & Filon, 1936) and according to Kendall (1952), Yule was deeply affected by Pearson's death.

The unresolved nature of the debate must also have had the negative effect that practitioners and fellow statisticians alike were left in doubt about what measure of association to use in different situations. The tone of the debate leaves the reader with the impression that the choice of measure of association almost is a matter of life and death. And that is, of course, not quite the case. In fact, one of the conclusions of the present article is that between the two, the choice does not carry a substantial impact on the conclusions of the association analysis. So quite on the contrary, as it will be seen, practitioners have no reason to be anxious. And neither Pearson nor Yule, as will also be seen, had really any reason to fear for the future of modern statistics.

1.2. **Outline of the present article.** The core of the Pearson-Yule debate is about the assumptions implied by the two measures of association. In this article, a close look at the two measures of association will be taken and the implied assumptions will be pinpointed and formalized. Pearson & Heron (1913) argued that dichotomous variables should be considered dichotomizations of continuous underlying variables, while Yule (1912) argued that they should be considered inherently discrete. In this article, however, it is shown that under given marginal probabilities there exists a continuous bijection between the two, which moreover has a fixed point at zero for all marginal probabilities. Consequently, both measures of association can be computed equally well no matter whether the variables are considered dichotomizations of continuous variables or not. As long as one of the assumptions is deemed appropriate, it does not make a difference which one it is. As a consequence, it turns out, whether to use the tetrachoric correlation coefficient or the *phi*-coefficient is in principle a matter of preference only.

The main result of this article, that there exists a continuous bijection between the *phi*-coefficient and the tetrachoric correlation coefficient under given marginal probabilities, has not been found in the literature. Guilford & Perry (1951) and Perry & Michael (1952) use series expansion of the integral equation of the tetrachoric correlation coefficient to find an approximate formula of the tetrachoric correlation coefficient as a function of the *phi*-coefficient whose errors, according to Perry & Michael, "are negligible for values of [the approximate tetrachoric correlation coefficient] less than |0.35| and probably relatively small for values of [the approximate tetrachoric correlation coefficient] between |0.35| and |0.6|". Though Guilford & Perry and Perry & Michael consider the relationship *phi*-coefficient - tetrachoric correlation coefficient, their result does, however, not imply a continuous bijection.

In Section 2, the *phi*-coefficient and the tetrachoric correlation coefficient are introduced, necessary assumptions formalized, and a proof that the tetrachoric correlation

coefficient is well defined is given. In Section 3, the main theorem of this article is stated and proved, and its implications are briefly discussed. Thereafter, in Section 4, some numerical examples and graphs of the relation *phi*-coefficient - tetrachoric correlation coefficient are considered. And finally, the article is concluded with Section 5.

2. The two measures of association

2.1. **Dichotomous variables.** Let X and Y be two dichotomous variables. In the most general setting, the values of a dichotomous variable cannot be added, multiplied, ordered, or otherwise acted on by any binary operator, save projection. The algebraically most stringent way to model a dichotomous variable is to define it as a random element $X : \Omega \to \mathcal{C}$, where the sample space $\mathcal{C}$ is an abstract set $\{c_1, c_2\}$ with no binary operations defined. Label the values of the two dichotomous variables *positive* and *negative*, respectively, and let p_X and p_Y denote the probabilities of positive values of X and Y, respectively.

One basic question in multivariate statistics is whether the random variables are statistically independent. For this purpose a new random variable $Z : \Omega \to \mathcal{C}^2$ is defined by $Z = (X, Y)$. Let p_a, p_b, p_c, and p_d denote the probabilities of Z taking values (*pos.*, *pos.*), (*pos.*, *neg.*), (*neg.*, *pos.*), and (*neg.*, *neg.*), respectively. Hence, p_a is the joint probability of positive values of X and Y. The random variable Z is often illustrated with a 2×2 contingency table, see Table 3.

As always, Kolmogorov's axioms imply that the joint probabilities are elements of the unit interval, $I = [0, 1]$, and that they sum to one. For 2×2 contingency tables, this, together with the identities $p_X = p_a + p_b$ and $p_Y = p_a + p_c$, implies the inequalities

$$\max(p_X + p_Y - 1, 0) \leq p_a \leq \min(p_X, p_Y).$$

Moreover, the contingency table is fully determined by the triple $(p_X, p_Y, p_a) \in I^2 \times [\max(p_X + p_Y - 1, 0), \min(p_X, p_Y)]$. It is often necessary to separate the cases where the marginal probabilities, (p_X, p_Y), are elements of the boundary and the interior of the unit square, I^2, respectively. Because it is closed, the unit square is the disjoint union of its boundary and its interior, $I^2 = \partial I^2 \cup \mathrm{Int}(I^2)$.

If X and Y are statistically independent, then the joint probabilities are the products of the marginal probabilities. Given a sample, independence can be tested with, e.g., the Pearson chi-square test. If X and Y are found to be statistically dependent, then it may be of interest to estimate some measure of association. The linear correlation is often a first choice of measure of association, but because the sample space of Z has no additive notion, expected values cannot be computed. The *phi*-coefficient and the tetrachoric correlation coefficient, on the other hand, are two measures of association defined especially for dichotomous variables.

2.2. **The *phi*-coefficient.** The *phi*-coefficient is the linear correlation between postulated underlying discrete univariate distributions of X and Y. Formally, let T_X and

TABLE 3. Elements of the 2×2 contingency table.

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

T_Y be two mappings such that $T_X X(\Omega)$ and $T_Y Y(\Omega)$ are both subsets of $\mathbb{R}$. For example, T_X could be $a + b\mathbb{1}_{\{pos.\}}$, for some real constants a and b, $\mathbb{1}$ being the indicator function. Furthermore, denote $T_X(pos.) = \alpha$, $T_X(neg.) = \beta$, $T_Y(pos.) = \gamma$ and $T_Y(neg.) = \delta$. The technical conditions $\alpha \neq \beta$, $\gamma \neq \delta$ and $\text{sign}(\alpha - \beta) = \text{sign}(\gamma - \delta)$ are also needed. The *phi*-coefficient is then defined as the linear correlation of $T_X X$ and $T_Y Y$, i.e. $r_{phi} = \text{Corr}(T_X X, T_Y Y)$.

Proposition 1. *The* phi-*coefficient is given by*

$$r_{phi} = \frac{p_a - p_X p_Y}{(p_X p_Y (1 - p_X)(1 - p_Y))^{1/2}} \tag{1}$$

if $0 < p_X, p_Y < 1$*, and* $r_{phi} = 0$ *otherwise.*

Proof. For $(p_X, p_Y) \in \text{Int}(I^2)$, the result is an easy exercise in algebraic manipulation.

For $(p_X, p_Y) \in \partial I^2$, it follows that $p_a = p_X$ or $p_a = p_Y$. Suppose $p_X = 1$, then $p_a = p_Y$ and $p_c + p_d = 0$, and calculations yield $\text{Cov}(T_X X, T_Y Y) = 0$. If $p_X = 0$, then $p_a = p_X$ and a similar calculation yields $\text{Cov}(T_X X, T_Y Y) = 0$. The same holds if $p_Y = 1$ or $p_Y = 0$, by symmetry. And so, because random variables with covariance zero are uncorrelated, it is concluded that $r_{phi} = 0$. □

Note that without the requirement $\text{sign}(\alpha - \beta) = \text{sign}(\gamma - \delta)$ in the definition, a factor $\text{sign}((\alpha - \beta)(\gamma - \delta))$ would appear in the right hand side of Equation (1). However, if the sign equality does not hold, then the two values of one dichotomous variable might as well have their labels switched, since the labeling of the values of the dichotomous variable was arbitrary to begin with, and then the sign equality holds. So the condition is really no restriction. The next result follows.

Corollary 2. *Except for its sign, the* phi-*coefficient does not depend on the choice of mappings* T_X *and* T_Y*.*

An implication of Corollary 2 is that proliferation of *phi*-coefficients is avoided, and the amount of subjectivism inserted into the *phi*-coefficient construction is limited. Also, the *phi*-coefficient can be considered scale and origin free. The only assumption needed is that it is possible to, without loss (or distortion) of information, map the values of X and Y into $\mathbb{R}$. The assumption is formalized as follows.

Assumption A1. *The values of the dichotomous variables can without loss of information be mapped into the real ordered field,* $\mathbb{R}$*.*

One possible interpretation of Assumption A1 is that the dichotomous variables X and Y are inherently discrete and that both values of X and Y, respectively, represent the same thing but of different magnitudes. Note that Assumption A1 implies an order relation between the values of X and Y, respectively, i.e. that the variables are ordinal. Assumption A1 is not appropriate if a dichotomous variable is strictly nominal, e.g. if the values are *Male* and *Female*, *Cross* and *Self-Fertilization*, or *Treatment A* and *Treatment B*. Pearson & Heron (1913) call these cases *Mendelian*.

Another consequence of Corollary 2 is the well known result that for dichotomous variables, the rank correlation equals the *phi*-coefficient. This follows since observations from a dichotomous variable can have at most two distinct (average) ranks, and the rank correlation is, of course, the linear correlation of the ranks.

It seems as if Yule (1912) thought of the dichotomous variables as Bernoulli distributed, and that the linear correlation therefore could be computed without any assumptions whatsoever. In the present framework, however, such a random variable would be given by $\mathbb{1}_{\{pos.\}} \circ X = \mathbb{1}_{\{pos.\}}(X)$, which demonstrates that Assumption A1 is indeed implicit in Yule's construction. A general word of caution is that the labeling of the values of a categorical variable by numbers can lead to a range of unnecessary complications and errors.

For given marginal probabilities $0 < p_X, p_Y < 1$, r_{phi} is a polynomial of degree one, so r_{phi} is a continuous and monotonic function of p_a. By the intermediate value theorem, the range of the *phi*-coefficient r_{phi} is the closed interval with endpoints given by the two inequalities

$$\begin{aligned} r_{phi} &\geq \max\left(-\left(\tfrac{p_X p_Y}{(1-p_X)(1-p_Y)}\right)^{1/2}, -\left(\tfrac{(1-p_X)(1-p_Y)}{p_X p_Y}\right)^{1/2}\right) \\ r_{phi} &\leq \min\left(\left(\tfrac{p_X(1-p_Y)}{p_Y(1-p_X)}\right)^{1/2}, \left(\tfrac{p_Y(1-p_X)}{p_X(1-p_Y)}\right)^{1/2}\right). \end{aligned} \tag{2}$$

A special case is if $p_X = p_Y = 0.5$, then $-1 \leq r_{phi} \leq 1$. Other examples are if $p_X = p_Y = 0.2$ then $-0.25 \leq r_{phi} \leq 1$, or if $p_X = 0.2$ and $p_Y = 0.8$ then $-1 \leq r_{phi} \leq 0.25$. See Figures 3, 4 and 5.

2.3. **The tetrachoric correlation coefficient.** The tetrachoric correlation coefficient is the linear correlation between postulated underlying normal distributions of X and Y. The 2×2 contingency table is thought of as a double dichotomy of a bivariate normal distribution. One can visualize the bell-shaped bivariate normal density function standing atop the contingency table. Since the dichotomous variables are both scale and origin free, and the family of normal distributions is closed under linear transformations, the normal distribution can without loss of generality be set to standard normal. Changing the parameter value of the bivariate standard normal distribution will change the shape of the bell-shaped bivariate normal density function, and hence the probability masses over the four rectangles that results from the double dichotomization. The tetrachoric correlation coefficient is the parameter value for which the volumes of the double dichotomized bivariate standard normal distribution equal the joint probabilities of the

contingency table. The parameter value of the bivariate standard normal distribution equals, of course, the linear correlation of the postulated joint normal distributuion.

Since the contingency table is fully determined by the marginal probabilities and one joint probability, it suffices to choose parameter value such that, under given marginal probabilities, one joint probability equals the corresponding volume. Conventionally, that joint probability is chosen to be the probability p_a, corresponding to positive values of both dichotomous variables. Since a volume of a normal distribution cannot be expressed in closed form, computing the tetrachoric correlation coefficient amounts to solving an integral equation.

For marginal probabilities that satisfy $0 < p_X, p_Y < 1$, and for joint probabilities p_a such that $\max(p_X + p_Y - 1, 0) < p_a < \min(p_X, p_Y)$, the tetrachoric correlation coefficient is defined as the solution r_{tc} to the integral equation

$$p_a = \int_{\Phi^{-1}(1-p_X)}^{\infty} \int_{\Phi^{-1}(1-p_Y)}^{\infty} \phi_2(x_1, x_2, r_{tc}) dx_2 dx_1 \,, \tag{3}$$

where $\Phi(x)$ is the standard normal distribution function and $\phi_2(x_1, x_2, \rho)$ is the bivariate standard normal density function. If $p_a = \max(p_X + p_Y - 1, 0)$ or $p_a = \min(p_X, p_Y)$, then the tetrachoric correlation coefficient, r_{tc}, is defined to be -1 or 1 respectively. If the inequality $0 < p_X, p_Y < 1$ does not hold, then any value of r_{tc} will satisfy Equation (3). However, from the perspective of presuming statistical independence until evidence of dependence is found, the tetrachoric correlation coefficient is here defined to be zero.

The integral in Equation (3) is sometimes called the bivariate standard normal survival function, and denoted $\bar{\Phi}_2(k, l, \rho)$. Therefore, Equation (3) can be written $p_a = \bar{\Phi}_2(\Phi^{-1}(1 - p_X), \Phi^{-1}(1 - p_Y), r_{tc})$.

There are several theories why Pearson (1900) for the purpose of the above definition chose the parametric family of bivariate normal distributions. At the time, the normal distribution was prevalent, and according to Pearson & Heron (1913) there were no other bivariate distribution that up until the time had been discussed effectively. Furthermore, Pearson (1900) was primarily interested in applications in the fields of evolution and natural selection, which is evident from the article's title, and such variables were generally assumed to be normally distributed. Pearson's friend and mentor Francis Galton even had a philosophical argument why all variables in nature ought to be normally distributed. Also, the parameter of the parametric family of bivariate normal distributions happens to be a measure of association, and this in combination with other nice properties makes the choice of the bivariate normal distribution most convenient. Ekström (2009) has generalized the definition so that a large class of parametric families of bivariate distributions can be assumed.

The assumption on which the tetrachoric correlation coefficient rests is formalized as follows.

Assumption A2. *The dichotomous variables are dichotomized jointly normally distributed random variables.*

In particular, Assumption A2 implies that the values of the dichotomous variables have an ordering, i.e. that the variables are ordinal. In many of the examples of Pearson (1900), the dichotomous variables are actual dichotomizations of continuous variables such as the stature of fathers and sons. In practice, however, such variables are most often measured with greater precision than two categories, and for dichotomous variables which cannot be measured with greater precision it is in general not easy to make an assertion about the distribution of a postulated underlying continuous variable.

If a tetrachoric correlation coefficient exists and is unique for every contingency table, then the tetrachoric correlation coefficient is said to be well defined. The following theorem is of theoretical and practical importance, and has not been found in the literature.

Theorem 3. *The tetrachoric correlation coefficient is well defined.*

For the proof of the Theorem 3 the following lemma is needed.

Lemma 4. *Let* $\bar{\Phi}_2(k,l,\rho)$ *be the bivariate standard normal survival function*

$$\bar{\Phi}_2(k,l,\rho) = \int_k^\infty \int_l^\infty \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left\{-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}\right\} dydx$$

and let $\Phi(x)$ *be the univariate standard normal distribution function. Then it holds that*

$$\lim_{\rho \searrow -1} \bar{\Phi}_2(k,l,\rho) = \max(1-\Phi(k)-\Phi(l),0) \tag{4}$$

$$\lim_{\rho \nearrow 1} \bar{\Phi}_2(k,l,\rho) = 1-\max(\Phi(k),\Phi(l))\,. \tag{5}$$

Proof. With the change of variable $w = (y-\rho x)/(1-\rho^2)^{1/2}$ and writing the limits of integration using indicator functions, it follows that

$$\bar{\Phi}_2(k,l,\rho) = \int_{-\infty}^\infty \int_{-\infty}^\infty \mathbb{1}_{\{x>k\}} \mathbb{1}_{\left\{w>\frac{l-\rho x}{(1-\rho^2)^{1/2}}\right\}} \phi(x)\phi(w)dwdx\,,$$

where ϕ is the standard normal density function.

First, it is shown that $\lim_{\rho\searrow -1} \mathbb{1}_{\left\{w>(l-\rho x)/(1-\rho^2)^{1/2}\right\}} = \mathbb{1}_{\{x<-l\}}$ almost everywhere. This is because $(l-\rho x)/((1-\rho^2)^{1/2}) \to \operatorname{sign}(l+x)\cdot\infty$ as $\rho \searrow -1$, except for the case $l+x=0$. So the equality holds almost everywhere.

Now, because the bivariate standard normal density function is dominated (by 1), it follows by dominated convergence

$$\lim_{\rho \searrow -1} \bar{\Phi}_2(k,l,\rho) = \int_{-\infty}^\infty \int_{-\infty}^\infty \mathbb{1}_{\{x>k\}} \mathbb{1}_{\{x<-l\}} \phi(x)\phi(w)dwdx\,.$$

Since the integral over w equals one, (4) follows. Equation (5) is shown similarly, with the difference that $\lim_{\rho\nearrow 1} \mathbb{1}_{\left\{w>(l-\rho x)/(1-\rho^2)^{1/2}\right\}} = \mathbb{1}_{\{x>l\}}$ almost everywhere. □

Now comes the proof of Theorem 3.

Proof of Theorem 3. If $p_X = 1$, then $p_a = p_Y$, so the right hand side of Equation (3), as a function of r_{tc}, has a one-point range, $p_a = p_Y$, and by the definition a one-point domain. For $p_X = 1$, $p_a = p_Y$ and $r_{tc} = 0$, (3) is clearly satisfied. The same holds if $p_Y = 1$, by symmetry. If $p_X = 0$ then $p_a = 0$, and thus the right hand side of Equation (3), as a function of r_{tc}, has a one-point range and a one-point domain. And the same holds if $p_Y = 0$, by symmetry. So the tetrachoric correlation coefficient is well defined when $(p_X, p_Y) \in \partial I^2$.

Suppose now that $(p_X, p_Y) \in \mathrm{Int}(I^2)$. The result that $\frac{\partial}{\partial \rho}\bar{\Phi}_2(k, l, \rho) = \phi_2(k, l)$, see, e.g., Tallis (1962), is used. Thus, the right hand side of Equation (3) is continuous and strictly increasing in ρ for all $(k, l, \rho) \in \mathbb{R}^2 \times (-1, 1)$. By Lemma 4, $\bar{\Phi}_2(k, l, \rho)$ is also continuous at the limit points $\rho = \pm 1$.

By the mean value theorem, for fixed values of k and l, $\bar{\Phi}_2(k, l, \rho)$ is a continuous bijection from $[-1, 1]$ to the closed interval with endpoints given by inserting the limits of integration of the right hand side of Equation (3) into expressions (4) and (5), i.e. the closed interval $[\max(p_X + p_Y - 1, 0), \min(p_X, p_Y)]$. Since the left hand side of Equation (3) is an element of the same interval, a unique solution r_{tc} of Equation (3) is guaranteed. □

3. The relation between the two

The following result is the main theorem of this article, by which the function defined $f : r_{phi} \mapsto r_{tc}$ exists and is a continuous bijection under given marginal probabilities.

3.1. **Main theorem.**

Theorem 5. *Under given marginal probabilities, there exists a continuous bijection $f : U \to [-1, 1]$ between the* phi-*coefficient and the tetrachoric correlation coefficient. U is the closed interval with endpoints given by (2).*

Proof. If $(p_X, p_Y) \in \partial I^2$, the definitions are such that $r_{phi} = r_{tc} = 0$. So in this case, f has domain and range equal to $\{0\}$, and $f(0) = 0$. Thus, by the definitions of continuity, injectivity and surjectivity, f is a continuous bijection.

Suppose now that $(p_X, p_Y) \in \mathrm{Int}(I^2)$. For convenience, it will be shown that the inverse, $f^{-1} : r_{tc} \mapsto r_{phi}$ is a continuous bijection. The continuous bijection is treated as a composition $f^{-1} = g \circ h$, where h is the function $h : r_{tc} \mapsto p_a$ and g is the function $g : p_a \mapsto r_{phi}$. The function h is given by Equation (3) and g is given by Equation (1).

First, continuity is shown. The function h is continuous on the domain $[-1, 1]$ by the proof of Theorem 3. Under given marginal probabilities p_X and p_Y, Equation (1) shows that r_{phi} is a polynomial of degree one, so g is also continuous. Thus f^{-1} is continuous. Moreover, the function h is strictly monotonic on the domain $[-1, 1]$ by the proof of Theorem 3. Since g is a polynomial of degree one, g is also strictly monotonic. Thus, f^{-1} is strictly monotonic and thus injective.

To show surjectivity, note first that because f^{-1} is continuous and the domain is an interval, by the intermediate value theorem the range $f^{-1}[-1, 1]$ is an interval. The range is closed since $\mathbb{R}$ is Hausdorff and the continuous image of a compact space is compact.

And because f^{-1} is monotonic, the endpoints of the interval $f^{-1}[-1,1]$ are $f^{-1}(-1)$ and $f^{-1}(1)$, respectively. By the definition of the tetrachoric correlation coefficient, $h(-1) = \max(p_X + p_Y - 1, 0)$ and $h(1) = \min(p_X, p_Y)$. So the endpoints of the range are $f^{-1}(-1) = g \circ h(-1) = g(\max(p_X + p_Y - 1, 0))$ and $f^{-1}(1) = g \circ h(1) = g(\min(p_X, p_Y))$ which equal the endpoints of (2). Thus, it is established that the range $f^{-1}[-1,1]$ is a closed interval with endpoints given by (2) which proves surjectiveness and this completes the proof. □

An expression for the continuous bijection can be found by solving Equation (1) for the joint probability, p_a, and equating with Equation (3), yielding the *phi*-coefficient, r_{phi}, as an integral-expression of the marginal probabilities, p_X and p_Y, and the tetrachoric correlation coefficient, r_{tc}. Furthermore, by a result from Kendall (1941), the integral of Equation (3) can be written as a so-called tetrachoric series. This expression was also discussed in Guilford & Perry (1951). Thus, inserting p_X, p_Y and r_{tc} in the so obtained integral-expression yields r_{phi}.

The use of tetrachoric series for approximating the integral of Equation (3) is, in all likelihood, the origin of the tetrachoric correlation coefficient's name. However, the name is probably of later date since neither Pearson & Heron (1913) nor Yule (1912) used it. Nowadays the integral of Equation (3) is quite effortlessly computed with computer assisted numerical optimization, but the name lingers on nevertheless.

The continuous bijection f has an interesting property. For all marginal probabilities, it has a fixed point at zero, which moreover corresponds to statistical independence of the dichotomous variables.

Proposition 6. *The following statements are equivalent.*

(a) *The dichotomous variables are statistically independent.*
(b) *The* phi-*coefficient is zero.*
(c) *The tetrachoric correlation coefficient is zero.*

Proof. (a) $\implies$ (b). Assume that the dichotomous variables X and Y are statistically independent. Since a 2×2 contingency table is fully determined by the triple (p_X, p_Y, p_a), the statement is equivalent to $p_a = p_X p_Y$. By Proposition 1, it follows immediately that $p_a = p_X p_Y$ implies that the *phi*-coefficient is zero.

(b) $\implies$ (c). Assume $r_{phi} = 0$. If $(p_X, p_Y) \in \partial I^2$, then $r_{tc} = 0$ by definition. Otherwise, if $(p_X, p_Y) \in \mathrm{Int}(I^2)$, then it follows from Equation (1) that $p_a = p_X p_Y$. Since $r_{tc} = 0$ is a solution of the integral equation (3) with $p_a = p_X p_Y$, which moreover is unique by Theorem 3, the tetrachoric correlation coefficient is zero.

(c) $\implies$ (a). Assume $r_{tc} = 0$. If $(p_X, p_Y) \in \partial I^2$, then $p_a = p_X p_Y$ follows from the inequalities $\max(p_X + p_Y - 1, 0) \leq p_a \leq \min(p_X, p_Y)$. Otherwise, if $(p_X, p_Y) \in \mathrm{Int}(I^2)$, then $p_a = p_X p_Y$ follows from Equation (3). And since a 2×2 contingency table is fully determined by the triple (p_X, p_Y, p_a), $p_a = p_X p_Y$ is equivalent to the dichotomous variables being statistically independent. □

A further consequence of Theorem 5 and Proposition 6 is that the *phi*-coefficient is positive if and only if the tetrachoric correlation coefficient is positive, and that the *phi*-coefficient is negative if and only if the tetrachoric correlation coefficient is negative.

Proposition 7. *The* phi-*coefficient is positive (negative) if and only if the tetrachoric correlation coefficient is positive (negative).*

Proof. Assume that the *phi*-coefficient is positive (negative). Then, by Proposition 6 the tetrachoric correlation coefficient is non-zero and by Proposition 1, $(p_X, p_Y) \in \text{Int}(I^2)$. Consequently, the continuous bijection defined $f : r_{phi} \mapsto r_{tc}$, whose existence is guaranteed by Theorem 5, can be decomposed into $f^{-1} = g \circ h$, where h is the function $h : r_{tc} \mapsto p_a$, given by Equation (1), and g is the function $g : p_a \mapsto r_{phi}$, given by Equation (3). The function h is increasing in r_{tc} by the proof of Theorem 3 and the function g is clearly increasing in p_a. Thus, the continuous bijection $f : r_{phi} \mapsto r_{tc}$ is increasing. And since the bijection has a fixed point at zero for all marginal probabilities, by Proposition 6, it follows that the tetrachoric correlation coefficient is positive (negative). Because $f : r_{phi} \mapsto r_{tc}$ is a bijection, the converse implication holds as well. □

3.2. **Implications.** An implication of Theorem 5 is that under given marginal probabilities, whenever the *phi*-coefficient is known, the tetrachoric correlation coefficient is also known, and vice versa. Thus, the *phi*-coefficient can be computed under Assumption A2 and the tetrachoric correlation coefficient can be computed under Assumption A1. In that sense neither assumption is more restrictive than the other, at least not in this particular setup.

In fact, one can argue that Assumptions A1 and A2 are quite similar. Both assumptions imply order relations on the values of the two dichotomous variables. And both assumptions imply underlying joint probability distributions on the real plane, $\mathbb{R}^2$. Assumption A2 implies the Gaussian probability measure on $\mathbb{R}^2$, while Assumption A1 implies a probability measure on $\mathbb{R}^2$ which is zero everywhere except on the four point set $\{(\alpha, \gamma), (\alpha, \delta), (\beta, \gamma), (\beta, \delta)\}$. And moreover, both the *phi*-coefficient and the tetrachoric correlation coefficient are the linear correlations of the postulated underlying bivariate distributions, and as such both are scale and origin free. The only difference between the *phi*-coefficient and the tetrachoric correlation coefficient is the joint probability distribution postulated.

Furthermore, in the case one of the two assumptions is correct but the other is erroneously made nonetheless, the conclusions of the association analysis will not change appreciably. Both measures of association equal zero if and only if the dichotomous variables are statistically independent, and the *phi*-coefficient is positive if and only if the tetrachoric correlation coefficient is positive. Since there exists a continuous bijection $f : r_{phi} \mapsto r_{tc}$, the only difference between the two measures of association is that individual values can be somewhat different, see Figures 3, 4 and 5. But since individual values of measures of association on the interior of the unit interval, $(0, 1)$, do not carry

any particular meaning other than that they are in between zero and one, the conclusions of the association analysis will not appreciably change.

The greatest difference in value between the *phi*-coefficient and the tetrachoric correlation coefficient occurs for contingency tables where the marginal probabilities are near zero or one. For example, consider a contingency table (p_X, p_Y, p_a) where the joint probability p_a is zero while the marginal probabilities, p_X and p_Y, are small but still non-zero. For these contingency tables, the tetrachoric correlation coefficient is identically minus one, while the *phi*-coefficient may be near zero. In fact, the limit of the *phi*-coefficient, as the vector of marginal probabilities (both positive) goes to the zero vector, is zero. For fixed positive marginal probabilities, however, the *phi*-coefficient is strictly negative, in accordance with Proposition 7. For the smallpox recovery data, Table 1, the marginal probabilities are approximately $p_X = 0.8$ and $p_Y = 0.9$, which in this setting can be considered relatively near one. The *phi*-coefficient is $r_{phi} = 0.3$, and hence smaller than the tetrachoric correlation coefficient which is $r_{tc} = 0.6$, see Figure 6. For the diphtheria recovery data, Table 2, the marginal probabilities are both approximately $p_X = p_Y = 0.5$ and therefore Figure 3 illustrates the relation between the two measures of association.

The limitation in range has sometimes been interpreted as a weakness of the *phi*-coefficient in comparison with the tetrachoric correlation coefficient, see, e.g., Guilford (1956). This point was also made by Pearson & Heron (1913). However, from some perspectives this interpretation is questionable. For example, consider a sample of fixed size from two independent dichotomous variables with marginal probabilities that are small but non-zero. Then the expected value of the number of observations in the cell (*pos.*,*pos.*) may be near zero, and if the number of observations indeed is zero, the tetrachoric correlation coefficient equals negative unity. The *phi*-coefficient, while also negative, is near zero and is therefore better reflecting the population association which is zero in this example. At the core of this matter is whether zero observations in a cell, given fixed sample sizes, should be interpreted as evidence of perfect association. The value of the tetrachoric correlation coefficient implies precisely that, while the value of the *phi*-coefficient does not.

4. Numerical examples

Using for example MATLAB, it is possible to compute *phi*-coefficients and tetrachoric correlation coefficients for a number of contingency tables. For fixed marginal probabilities, contingency tables can be generated by choosing some values of the joint probability p_a from $\max(p_X + p_Y - 1, 0)$ up to $\min(p_X, p_Y)$. In Figures 3, 4 and 5, tetrachoric correlation coefficients are plotted against *phi*-coefficients for marginal probabilities $p_X = p_Y = 0.5$, $p_X = p_Y = 0.2$ and $p_X = 0.2$, $p_Y = 0.8$ respectively. Using linear interpolation, an approximation of the continuous bijection f, whose existence is guaranteed by Theorem 5, is seen.

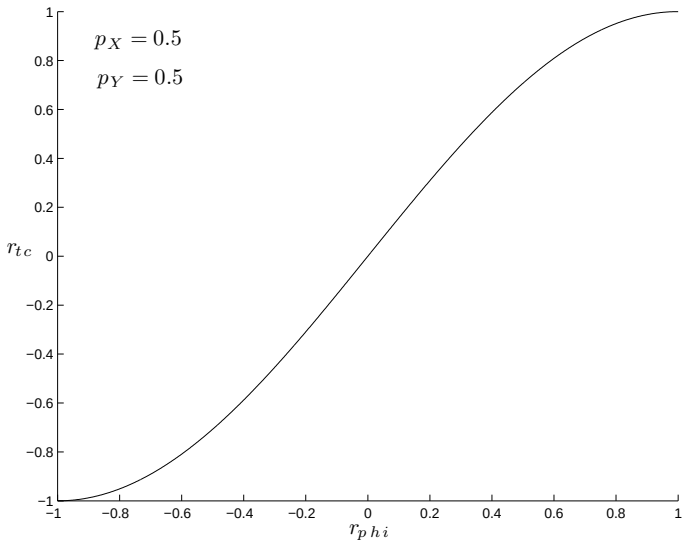

FIGURE 3. Numerical computation of the bijection graph given marginal probabilities $p_X = p_Y = 0.5$.

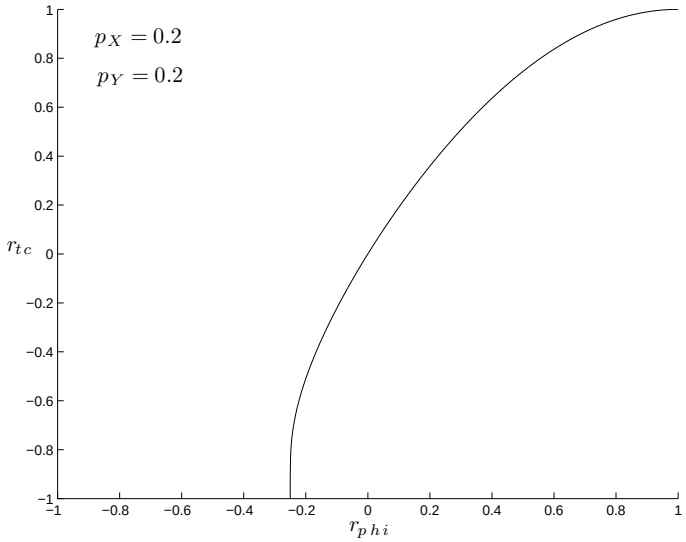

FIGURE 4. Numerical computation of the bijection graph given marginal probabilities $p_X = p_Y = 0.2$.

In Figure 6, the continuous bijection given the marginal probabilities of Pearson's smallpox recovery data, Table 1, is seen. Also, the two measures of association for the same data set are plotted in dotted lines. In the figure, it is clearly seen that the *phi*-coefficient is limited in range given marginal probabilities that are close to zero or one, and consequently, the *phi*-coefficient is smaller than the tetrachoric correlation coefficient for this data set.

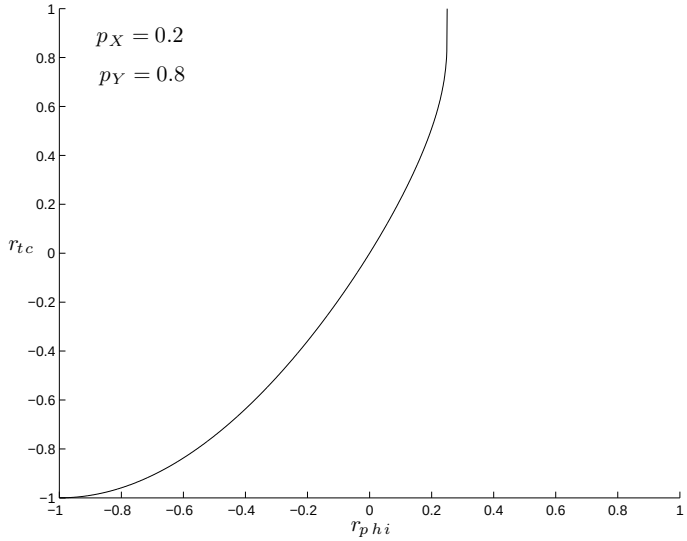

FIGURE 5. Numerical computation of the bijection graph given marginal probabilities $p_X = 0.2$ and $p_Y = 0.8$.

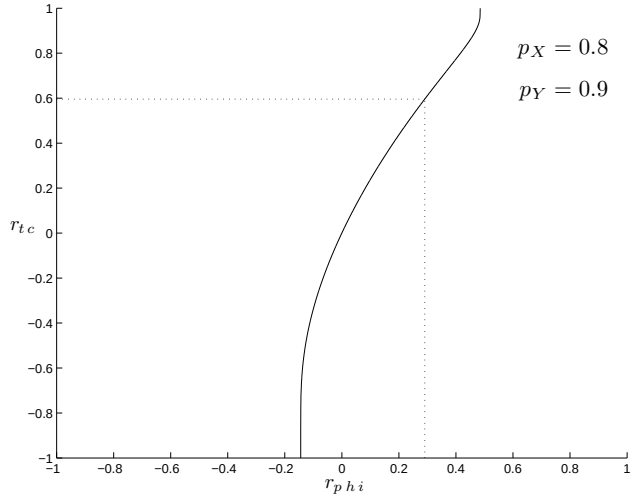

FIGURE 6. Numerical computation of the bijection graph given the marginal probabilities of the smallpox recovery data (solid line) and the two measures of association for the same data set (dotted lines), see Table 1. Numbers are rounded off to one decimal.

In all figures, the continuous bijection is increasing and has a fixed point at zero, as is stated in Propositions 6 and 7. By Theorem 5, the bijection domain is the interval with endpoints given by Equation (2) and the range is the interval $[-1, 1]$, something that is also clearly seen in the figures.

5. Conclusions

Two measures of association for dichotomous variables, the *phi*-coefficient and the tetrachoric correlation coefficient, have been rigorously defined, their assumptions formalized, and some key properties derived. The *phi*-coefficient is the linear correlation between postulated underlying discrete variables, while the tetrachoric correlation coefficient is the linear correlation between postulated underlying normally distributed variables. Both measures of association imply ordinality of values of the the two dichotomous variables.

The assumptions on which the two measures of association rest are quite similar in the sense that both imply underlying joint probability distributions on the real plane. The difference is that the tetrachoric correlation coefficient assumption implies the bivariate normal distribution, while the *phi*-coefficient implies an inherently discrete distribution with a four point support.

By the main theorem of this article, Theorem 5, there exists a continuous bijection between the *phi*-coefficient and the tetrachoric correlation coefficient under given marginal probabilities. Thus, the *phi*-coefficient can be computed using the assumptions of the tetrachoric correlation coefficient construction and vice versa. Hence, between the two the choice of measure of association is in principle a matter of preference only.

The reasoning carries over in an attempt to reconcile the famous Pearson-Yule debate. Because both measures of association can be computed under either assumption, and since differences in values resulting from making the erroneous assumption will not in general appreciably change the conclusions of the association analysis, the choice of measure of association is not crucial. Whether the underlying joint distribution is normal or discrete does not have a substantial impact on the conclusions of the association analysis.

Lastly, despite the caustic tone of the Pearson-Yule debate, there is no reason for practitioners to feel anxious about the choice between the two measures of association. The two measures of association are in principle similar theoretical constructions and whatever the choice is, it will not not carry a substantial impact on the conclusions of the association analysis. And despite Karl Pearson's fears, the use of any of the two measures of association will not put the future of modern statistics in peril.

Acknowledgements

This article was prepared during a visit to UCLA Department of Statistics, and the author is grateful for the generosity and hospitality of all department faculty and staff, and particularly Distinguished Professor Jan de Leeuw. In the manuscript preparation, Professor Thomas Ferguson generously provided valuable input and comments. This work was supported by the Jan Wallander and Tom Hedelius Research Foundation (project P2008-0102:1).

References

Boas, F. (1909). Determination of the coefficient of correlation. *Science*, *29*, 823–824.

Camp, B. H. (1933). Karl Pearson and Mathematical Statistics. *J. Amer. Statist. Assoc.*, *28*, 395–401.

Ekström, J. (2009). A generalized definition of the tetrachoric correlation coefficient. In *Contributions to the Theory of Measures of Association for Ordinal Variables*. Ph.D. thesis, Uppsala: Acta Universitatis Upsaliensis.

Guilford, J. P. (1956). *Fundamental statistics in psychology and education, 3rd ed*. New York: McGraw-Hill.

Guilford, J. P., & Perry, N. C. (1951). Estimation of other coefficients of correlation from the phi coefficient. *Psychometrika*, *16*, 335–346.

Heron, D. (1911). The danger of certain formulae suggested as substitutes for the correlation coefficient. *Biometrika*, *8*, 109–122.

Kendall, M. G. (1941). Proof of relations connected with the tetrachoric series and its generalization. *Biometrika*, *32*, 196–198.

Kendall, M. G. (1952). George Udny Yule C.B.E., F.R.S. *J. Roy. Statist. Soc. Ser. A*, *115*, 156–161.

Pearson, E. S. (1938). *Karl Pearson: An appreciation of some aspects of his life and work*. Cambridge: The University press.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, *195*, 1–47.

Pearson, K., & Heron, D. (1913). On theories of association. *Biometrika*, *9*, 159–315.

Perry, N. C., & Michael, W. B. (1952). The relationship of the tetrachoric correlation coefficient to the phi coefficient estimated from the extreme tails of a normal distribution of criterion scores. *Educ. Psychol. Meas.*, *12*, 778–786.

Tallis, G. M. (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, *18*, 342–353.

Yule, G. U. (1912). On the methods of measuring the association between two attributes. *J. Roy. Statist. Soc.*, *75*, 579–652.

Yule, G. U., & Filon, L. N. G. (1936). Karl Pearson. 1857-1936. *Obituary Notices of Fellows of the Royal Society*, *2*, 73–110.

Yule, G. U., Stuart, A., & Kendall, M. G. (1971). *Statistical papers of George Udny Yule, selected by Alan Stuart and Maurice G. Kendall*. London: Griffin.