ON THE RELATION BETWEEN THE POLYCHORIC CORRELATION COEFFICIENT AND SPEARMAN'S RANK CORRELATION COEFFICIENT

JOAKIM EKSTRÖM

Abstract. Spearman's rank correlation coefficient is shown to be a deterministic transformation of the empirical polychoric correlation coefficient. The transformation is a homeomorphism under given marginal probabilities, and has a fixed point at zero. Moreover, the two measures of association for ordinal variables are asymptotically equivalent, in a certain sense. If the ordinal variables arise from discretizations, such as groupings of values into categories, Spearman's rank correlation coefficient has some undesirable properties, and the empirical polychoric correlation coefficient is better suited for statistical inference about the association of the underlying, non-discretized variables.

Key words and phrases. Contingency table, Measure of association, Ordinal variable, Polychoric correlation coefficient, Spearman's rank correlation coefficient.

1. Introduction

The polychoric correlation coefficient and Spearman's rank correlation coefficient are two measures of association for ordinal variables. Ordinal variables are variables whose values can only be compared in terms of their ordering. Sometimes referred to as ordered categorical variables, ordinal variables are common in many scientific fields such as the health and social sciences. Data for a pair of ordinal variables is often presented in the form of a contingency table.

A measure of association is, loosely, a function that maps a pair of random variables to a subset of the real line, and its value is meant to be interpreted as the degree to which the two random variables can be represented as monotonic functions of each other. The first, and likely most well-known, measure of association is the linear correlation, introduced by Francis Galton (1888).

The idea of the polychoric correlation coefficient was proposed by Galton's protégé Karl Pearson (1900) as a correlation coefficient for ordinal variables. The measure of association rests on an assumption of an underlying joint bivariate normal distribution, meaning that the contingency table of the two ordinal variables is assumed to be the result of a discretization of a bivariate normal distribution, cf. Figure 1. Given a contingency table, a bivariate normal distribution is fitted to the table, and the polychoric correlation coefficient then corresponds to the linear correlation of the fitted bivariate normal distribution. Implicit in the construction is that the ordinal variables, while only observed in terms of ordered categories, are considered as fundamentally continuous in nature.

Charles Spearman was an English psychologist who made large contributions to the theory of multivariate statistics, notably the idea of factor analysis. Spearman's rank correlation coefficient is quite simply the linear correlation of the ranks of the observations, and as such it is a measure of association for ordinal variables.

Karl Pearson, who considered the polychoric correlation coefficient as one of his most important contributions to the theory of statistics (see Camp, 1933), was never convinced of the appropriateness of Spearman's rank correlation coefficient. On the contrary, Pearson (1907) contains a comprehensive discussion on correlation of ranks and some rather blunt criticism. For example, the article reads:

> Dr Spearman has proposed that rank in a population for any variate should be considered as in itself the quantitative measure of the character, and he proceeds to correlate ranks as if they were quantitative measures of character, without any reference to the true value of the variate. This seems to me a retrograde step; hitherto we have dealt with grade or rank as an index to the variate, and to make rank into a unit of itself cannot fail, I believe, to lead to grave misconception.

Part of the explanation for Pearson's dedication to contingency tables and measures of association for ordinal variables can be found in his book *The Grammar of Science*

(Pearson, 1911). In the chapter *Contingency and Correlation*, the contingency table is described as a universal tool for utilizing empirical evidence for the advancement of science. The chapter summary reads: "Whether phenomena are qualitative or quantitative a classification leads to a contingency table, and from such a table we can measure the degree of dependence between any two phenomena."

In several articles, Pearson referred to the linear correlation as the *true correlation*, implying that the linear correlation is more true, in some sense, than other measures of association. However, following the axiomatic definition of Rényi (1959), measures of association are nowadays considered merely an abstract continuous mapping of a pair of random variables to a subset of the real line, subject to certain conditions. Consequently, the modern point of view is that no measure of association is more true or otherwise more valuable than any other.

Pearson's polychoric correlation coefficient and Spearman's rank correlation coefficient are based upon seemingly very different constructions. As an indication of this perception, Pearson's former student George Udny Yule (1912) claim that the polychoric correlation coefficient is founded upon ideas entirely different from those of which Spearman's rank correlation coefficient is founded upon. The sentiment is echoed by Pearson & Heron (1913), who even claim that Spearman's rank correlation coefficient is not based on a reasoned theory, while arguing for the soundness of the polychoric correlation coefficient. In the present article, though, it shall be seen that the polychoric correlation coefficient and Spearman's rank correlation coefficient, in spite of the perceived dissimilarity and Karl Pearson's opposition to the use of latter, are quite similar theoretical constructs.

2. The two measures of association

2.1. **Ordinal variables.** Ordinal variables are variables whose values are ordered but cannot in general be added, multiplied, or otherwise acted on by any binary operator save projection. In spite of the common occurrence of ordinal variables, both in practice and in scientific fields such as the health and social sciences, no algebraic framework with the level of formalism sought for in the present article has been found in the literature. Therefore such a formalistic framework is given as follows.

Analogously to Kolmogorov's definition of random variables, an ordinal variable is defined as a measurable function from a probability space Ω to a sample space, $\mathcal{C}$. The sample space $\mathcal{C} = \{c_1, c_2, \dots\}$ is totally ordered, i.e. for any c_i and c_j it holds that either $c_i \preceq c_j$, $c_i \succeq c_j$, or both. But characteristically, the sample space is not by definition equipped with any binary operation. The equality notation $c_i = c_j$ is shorthand for $c_i \preceq c_j$ and $c_i \succeq c_j$, and the strict notation $c_i \prec c_j$ is shorthand for $c_i \preceq c_j$ and $c_i \not\succeq c_j$.

In the present context, the only characteristic of the elements of the sample space that is of relevance is their ordering, and therefore all elements that have the same order are considered equal. Let $[c]_{\mathcal{C}}$ denote the equivalence class $\{x \in \mathcal{C} : x = c\}$, and let $\lfloor c \rfloor_{\mathcal{C}}$ denote the lower half-space $\{x \in \mathcal{C} : x \preceq c\}$. The index $\mathcal{C}$ is sometimes omitted when the

ordered set is clear from the context. Let the sets $\{x \in \mathcal{C} : c_i \preceq x \preceq c_j\}$ be closed, and as such be a basis for a topology on $\mathcal{C}$. The set $\mathcal{B}$ of Borel sets is the smallest σ-algebra containing the topology, and hence the so-constructed pair $(\mathcal{C}, \mathcal{B})$ is a measurable space.

For an ordinal variable $X : \Omega \to \mathcal{C}$ it is assumed without loss of generality that the strict inequalities $c_1 \prec c_2 \prec c_3 \prec \cdots$ hold. The fact that this can be assumed is easily realized when considering that it is always possible to map each equivalence class to any element of the class, relabel them if necessary, and then get a totally ordered set for which the strict inequalities hold. The values of an ordinal variable are sometimes referred to as *categories*, the ordinal variable as an *ordered categorical variable*, and the cardinality of the sample space as the *number of categories*.

As a note of caution to the reader, other definitions of ordinal variables than the one given above exist. For example, some authors impose the additional condition that the sample space must not be a real ordered field. In the present definition, though, there are no conditions on group structure, or the absence thereof. In particular, real-valued random variables satisfy the conditions of the present definition.

Let X and Y be two ordinal variables, each with a finite number of categories, whose association is to be studied and denote their numbers of categories r and s, respectively. Let the cumulative marginal probabilities be denoted $u_0, \ldots, u_r$ for X, i.e. $u_0 = 0$, $u_r = 1$ and $u_i = P(X \preceq c_i)$, and $v_0, \ldots, v_s$ for Y. The marginal probabilities are denoted ∇u_i and ∇v_j, respectively, where the symbol ∇ can be interpreted as a difference operator, yielding $\nabla u_i = u_i - u_{i-1} = P(X = c_i)$.

The joint probabilities of X and Y are sometimes denoted with double index, each referring to a value of one of the ordinal variables. In the present article, though, the joint probabilities is denoted with single index, $p_1, \ldots, p_{rs}$, each index referring to a specific cell of the contingency table. The way in which the cells of the contingency table is enumerated is not of importance. For example, the cells could be enumerated column-wise, row-wise, or via Cantor's diagonal method.

2.2. **The polychoric correlation coefficient.** The fundamental idea of the polychoric correlation coefficient construction is to assume that the two ordinal variables are, into r and s ordered categories respectively, discretized random variables with a continuous joint distribution belonging to some family of bivariate distributions. The discretization cuts the domain of the bivariate density function into rectangles corresponding to the cells of the contingency table, see Figure 1 for an illustration. For later reference, the fundamental assumption is formalized as follows.

Assumption A1. *The two ordinal variables are, into r and s ordered categories respectively, discretized random variables with a continuous joint distribution belonging to the family of bivariate distributions $\{H_\theta\}_{\theta \in \Theta}$.*

Pearson (1900) studied the case assuming a bivariate standard normal distribution. The definition given in this section is the generalized definition (see Ekström, 2008),

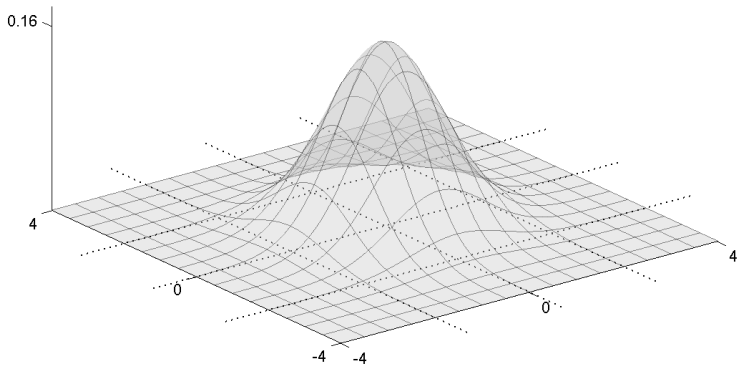


FIGURE 1. Illustration of the domain of the standard normal density function being discretized by the dotted lines into a 4×4 contingency table.

which agrees with Pearson's original definition under a joint normal distribution assumption.

For a bivariate probability distribution H and a rectangle $A = [a, b] \times [c, d]$, the volume of the rectangle equals $H(A) = H(b, d) - H(b, c) - H(a, d) + H(a, c)$. If the distribution function is absolutely continuous, i.e. has a density function, then the volume $H(A)$ equals the integral of the density function over the rectangle A. This special case illustrates the more general fact that if Z is a bivariate random variable with distribution function H, then $P(Z \in A) = H(A)$.

For all $i = 1, \ldots, r$ and $j = 1, \ldots, s$, create rectangles $[u_{i-1}, u_i] \times [v_{j-1}, v_j]$, enumerate them in the same way as the joint probabilities, $p_1, \ldots, p_{rs}$, and denote them $A_1, \ldots, A_{rs}$. The rectangles $A_1, \ldots, A_{rs}$ are interpreted as the result of the discretization of the domain of the bivariate copula distribution function, cf. Figure 1. Under Assumption A1, it should ideally hold that the volumes of the rectangles equal the joint probabilities of the two ordinal variables. Hence it should hold that

$$(H_\theta(A_1), \ldots, H_\theta(A_{rs})) = (p_1, \ldots, p_{rs})\,. \tag{1}$$

The equation above is often referred to as the defining relation of the polychoric correlation coefficient.

For the solution θ to Equation (1), the polychoric correlation coefficient is defined as

$$r_{pc} = 2\sin(\rho_S(H_\theta)\pi/6),$$

where ρ_S denotes the Spearman grade correlation, which is the population analogue of Spearman's rank correlation coefficient (see, e.g., Nelsen, 2006). If all points (u_i, v_j) are elements of the boundary of the unit square, ∂I^2, then any parameter θ satisfies Equation (1). However, in this case the polychoric correlation coefficient is defined to

be zero, in part because of a reasoning of presuming independence until evidence of association is found.

By Sklar's theorem, every continuous joint distribution function has a unique corresponding copula (see, e.g., Nelsen, 2006). In this setting, it is both mathematically and practically convenient to use the copula corresponding to a bivariate distribution instead of the bivariate distribution function itself. Assuming that the continuous joint distribution H is a copula, the Spearman grade correlation of the such jointly distributed random variables can be expressed as

$$\rho_S = 12 \int_{I^2} H d\lambda \, - \, 3 \, , \tag{2}$$

where I is the unit interval, $[0, 1]$, and λ is the Lebesgue measure (see, e.g., Nelsen, 2006).

Given data, the joint probabilities on the right-hand side of Equation (1) are estimated by their corresponding relative frequencies. Under Assumption A1, the relative frequencies will in general differ from their corresponding joint probabilities due to, for instance, fixed sample sizes and noisy observations. If the numbers of categories, r and s, both equal 2 then a unique solution to the sample analogue of Equation (1) always exists under some general conditions of the family of bivariate distributions. If one of r and s is greater than 2 and the other is greater than or equal to 2, on the other hand, then a solution to the sample analogue of Equation (1) does in general not exist. In that case it is standard statistical procedure to look for a best fit of the parameter θ with respect to some loss function.

In many situations, it is clear from the context whether the name polychoric correlation coefficient refers to the theoretical population construct or the sample analogue. When the sample and population variants are discussed in relation to each other, such as in Section 3, the theoretical population polychoric correlation is denoted ρ_{pc} and the sample polychoric correlation coefficient is denoted r_{pc}.

Examples in Ekström (2008) indicate that the polychoric correlation coefficient is not statistically robust to changes of distributional assumption, nor changes of loss function. For example, the polychoric correlation coefficient can change from positive to negative only because of a change of distributional assumption, and the same thing can occur because of a change of loss function. The lack of robustness is a problem for the polychoric correlation coefficient whenever there is uncertainty about which specific family of distributions that satisfies the statement of Assumption A1. In those commonly occurring cases, the empirical polychoric correlation coefficient is an in many ways attractive non-parametric alternative.

2.3. **The empirical polychoric correlation coefficient.** The empirical polychoric correlation coefficient is a relaxed version of the polychoric correlation coefficient which rests only on the assumption that the two ordinal variables are discretized random variables with a joint continuous distribution. In other words, for the empirical polychoric

correlation coefficient an underlying continuous joint distribution is only assumed to exist, not to be of any particular distributional family. Proposed by Ekström (2009), the non-parametric empirical polychoric correlation coefficient approximates the joint distribution by means of the empirical copula.

Let the two ordinal variables X and Y have sample spaces $\mathcal{C} = \{c_1, \ldots, c_r\}$ and $\mathcal{D} = \{d_1, \ldots, d_s\}$, respectively, and let $(x_k, y_k)_{k=1}^n$ be a sample of (X, Y) of size n. The empirical copula $\hat{C}_n$ of the sample $(x_k, y_k)_{k=1}^n$ is the function given by

$$\hat{C}_n(u_i, v_j) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\lfloor c_i \rfloor \times \lfloor d_j \rfloor}(x_k, y_k),$$

where $\mathbb{1}_A$ is the indicator function of the set A, and u_i and v_j are the cumulative marginal probabilities corresponding to values c_i and d_j, respectively, of the two ordinal variables. The empirical copula is only defined on the set of cumulative marginal probabilities (u_i, v_j) for $i = 0, \ldots, r$ and $j = 0, \ldots, s$. In practice, the cumulative marginal probabilities, $u_0, \ldots, u_r$ and $v_0, \ldots, v_s$, are estimated by their sample analogues, i.e. $n^{-1} \sum_{k=1}^{n} \mathbb{1}_{\lfloor c_i \rfloor}(x_k)$.

For the purpose of the empirical polychoric correlation coefficient, the postulated underlying joint distribution function H is approximated by means of the simple function $\hat{E}_n = \sum_{k=1}^{rs} a_k \mathbb{1}_{A_k}$, where, if A_k is the rectangle $[u_{i-1}, u_i) \times [v_{j-1}, v_j)$, a_k is given by the mean of the empirical copula values of the vertices of A_k, i.e.

$$a_k = \frac{1}{4} \left(\hat{C}_n(u_i, v_j) + \hat{C}_n(u_i, v_{j-1}) + \hat{C}_n(u_{i-1}, v_j) + \hat{C}_n(u_{i-1}, v_{j-1}) \right).$$

The empirical polychoric correlation coefficient is then defined analogously to the conventional polychoric correlation coefficient with the copula approximated by $\hat{E}_n$, hence $r_{epc} = 2\sin(\rho_S(\hat{E}_n)\pi/6)$, where the functional $\rho_S(\hat{E}_n)$ is given by Expression (2), i.e. $\rho_S(\hat{E}_n) = 12 \int \hat{E}_n d\lambda - 3$. Since $\hat{E}_n$ is simple, the integral reduces to $\sum_{k=1}^{rs} a_k \lambda(A_k)$.

The empirical polychoric correlation coefficient is well defined, takes values on the interval $[-1, 1]$, and converges almost surely to the theoretical population analogue $\rho_{pc} = 2\sin(\rho_S(H)\pi/6)$ as the numbers of categories and the sample size go to infinity, cf. Theorem 5. For fixed sample sizes, a simulation study in Ekström (2009) indicates that the empirical polychoric correlation coefficient is robust in terms of the underlying distribution, and is in terms of standard deviation more stable than conventional polychoric correlation coefficients. The simulation study also indicates that the empirical polychoric correlation coefficient, while unbiased at zero, have 4 to 20 percent too small absolute values, depending on the number of categories. In the same simulation study, the bias is constant under different sample sizes. In conclusion, therefore, the empirical polychoric correlation coefficient can be considered a conservative estimate of the theoretical population polychoric correlation, which is statistically robust and, in terms of standard deviation, more stable than conventional polychoric correlation coefficients.

2.4. **Spearman's rank correlation coefficient.** Spearman's rank correlation coefficient, proposed by Charles Spearman (1904), is quite simply the linear correlation of the

sample ranks. The function $rank$, which maps each observation to its average rank, is defined as

$$rank(x_i) = \frac{1}{2} + \sum_{k=1}^{n} \mathbb{1}_{\lfloor x_i \rfloor}(x_k) - \frac{1}{2} \sum_{k=1}^{n} \mathbb{1}_{[x_i]}(x_k).$$

Denoted r_S, Spearman's rank correlation coefficient is the sample correlation coefficient of the ranks of $(x_k, y_k)_{k=1}^n$, or more explicitly the sample correlation coefficient of $(rank(x_k), rank(y_k))_{k=1}^n$.

Notice that if the ordinal variables X and Y both have sample spaces with infinite cardinality and the maximal probability of an individual value is zero, then with probability one, no ties will be present in the sample. An example of such a case is if the ordinal variables are real-valued continuous random variables. If no ties are present the function $rank/n$ reduces to the empirical distribution function, and hence the sample correlation of $(rank(x_k), rank(y_k))_{k=1}^n$ equals the sample correlation of the empirical distribution function values of $(x_k, y_k)_{k=1}^n$. Thus, it is clear, by the Glivenko-Cantelli lemma and the Slutsky theorem, that Spearman's rank correlation coefficient converges to the Spearman grade correlation with probability one as the sample size goes to infinity.

For two non-continuous random variables, Nešlehová (2007) has defined the Spearman grade correlation as $\tilde{\rho}_S = \rho_S(C^S) / \left((1 - ||\nabla \vec{u}||_3^3)(1 - ||\nabla \vec{v}||_3^3)\right)^{1/2}$, where ρ_S is given by Expression (2) and $||x||_3$ denotes the L^3 norm of the vector x. The function C^S denotes the Schweizer-Sklar standard extension copula defined $C^S = \sum_{k=1}^{rs} h_k \mathbb{1}_{A_k}$, where $h_k : A_k \to [0,1]$ is the linear interpolant of the copula values of the vertices of the rectangle A_k. Clearly, if the maximal marginal probability is zero, i.e. the random variables are continuous, the non-continuous definition agrees with the conventional definition. Nešlehová (2007) also shows that the non-continuous Spearman grade correlation of the empirical copula equals Spearman's rank correlation coefficient, a result which is used extensively in the next section. The Schweizer-Sklar standard extension of the empirical copula is denoted $\hat{C}_n^S$. The graph of a Scheweizer-Sklar standard extension of the empirical copula for a particular data set is pictured in Figure 2.

3. Relations between the two

The aim of the present section is to find an expression for the empirical polychoric correlation coefficient as a function of Spearman's rank correlation coefficient, and thereby establishing a relation between Spearman's rank correlation coefficient and the polychoric correlation coefficient.

Proposition 1. *In the notation of Section 2, for any rectangle A_k it holds that $\int h_k \mathbb{1}_{A_k} d\lambda = a_k \lambda(A_k)$.*

Proof. Let $A_k = [u_{i-1}, u_i) \times [v_{j-1}, v_j)$. The only non-constants of h_k are the interpolation coefficients, and these are all of the form $(x - u_{i-1})/(u_i - u_{i-1})$ and $(u_i - x)/(u_i - u_{i-1})$. Integration over the interval $[u_{i-1}, u_i)$ with respect to x yields $(u_i - u_{i-1})/2$. Thus, a

factor $(u_i - u_{i-1})(v_j - v_{j-1})/4 = \lambda(A_k)/4$ breaks out, and the statement follows after collecting terms and substituting for a_k. □

Corollary 2. *In the notation of Section 2, it holds that* $\int \hat{C}_n^S d\lambda = \int \hat{E}_n d\lambda$.

A consequence of Corollary 2 is that for the purpose of computing the Spearman grade correlation it is equivalent whether the postulated underlying continuous joint distribution function is approximated by the simple function $\hat{E}_n$ or by the Schweizer-Sklar standard extension $\hat{C}_n^S$. Therefore, the empirical polychoric correlation coefficient can equivalently be defined as $r_{epc} = 2\sin(\rho_S(\hat{C}_n^S)\pi/6)$, i.e. as a function of the Schweizer-Sklar standard extension of the empirical copula.

The following theorem is the main result of this section.

Theorem 3. *In the notation of Section 2, for any contingency table it holds that*

$$r_{epc} = 2\sin\Big(r_S\left((1 - ||\nabla\vec{u}||_3^3)(1 - ||\nabla\vec{v}||_3^3)\right)^{1/2}\pi/6\Big). \tag{3}$$

Proof. By definition, $r_{epc} = 2\sin(\rho_S(\hat{E}_n)\pi/6)$, and by Nešlehová (2007) it holds that $r_S = \rho_S(\hat{C}_n^S)/\left((1 - ||\nabla\vec{u}||_3^3)(1 - ||\nabla\vec{v}||_3^3)\right)^{1/2}$. The statement then follows by Corollary 2 and substitution. □

Some numerical examples of Equation (3) are the following. If the maximal marginal probability of the contingency table is one third, the relative difference between the empirical polychoric correlation coefficient and Spearman's rank correlation coefficient, i.e. $|(r_{epc} - r_S)/r_S|$, is less than 11%. If the maximal marginal probability of the contingency table is one quarter, the relative difference is less than 6%, and if the maximal marginal probability is one fifth, the relative difference is less than 4%. For most applications, a relative difference of 6% or less can be considered negligible, in the sense that it carries no appreciable impact on the conclusions of the association analysis.

The next theorem states some properties of the relationship.

Theorem 4. *The relation between Spearman's rank correlation coefficient,* r_S*, and the empirical polychoric correlation coefficient,* r_{epc}*, has the following properties:*

(a) the function $f : r_S \mapsto r_{epc}$ *is a homeomorphism under given marginal probabilities,*
(b) Spearman's rank correlation coefficient is zero if and only if the empirical polychoric correlation coefficient is zero,
(c) Spearman's rank correlation coefficient is positive (negative) if and only if the empirical polychoric correlation coefficient is positive (negative).

Proof. (a). If either $||\nabla\vec{u}||_3^3$ or $||\nabla\vec{v}||_3^3$ equals one, then some value have probability one. Then r_S is zero because $(rank(x_k), rank(y_k))_{k=1}^n$ have sample covariance zero, and r_{epc} is zero as well, and the function is clearly a homeomorphism. Otherwise, note that the function $g(x) = \sin(x\pi/6)$ is a homeomorphism on the domain $[-1, 1]$, so for all $c \in (0, 1]$ the function $\tilde{g}(x) = \sin(cx\pi/6)$ is also a homeomorphism on the same domain. Thus, the function given by Equation (3) is a homeomorphism.

(b). If either $||\nabla\vec{u}||_3^3$ or $||\nabla\vec{v}||_3^3$ equals one, then some value have probability one, so both r_S and r_{epc} are zero. Otherwise, the function $f : r_S \mapsto r_{epc}$, given by Equation (3), clearly has a fixed point at zero, and since it is a bijection it has no other zeroes.

(c). Because the factor $\big((1-||\nabla\vec{u}||_3^3)(1-||\nabla\vec{v}||_3^3)\big)^{1/2}$ is non-negative, the function $f : r_S \mapsto r_{epc}$, given by Equation (3), is non-decreasing. The statement then follows by (a) and (b). □

As a consequence of Theorem 4, the conclusions of association analyzes conducted with the empirical polychoric correlation coefficient and Spearman's rank correlation coefficient, respectively, can differ only in terms of the strength of the association. From Equation (3) it is clear that the absolute value of Spearman's rank correlation coefficient in general is greater than that of the empirical polychoric correlation coefficient, resulting in the conclusion of a stronger association when using Spearman's rank correlation coefficient versus the empirical polychoric correlation coefficient. The relative difference is, however, generally small. The statistical reasons for the difference in values between the empirical polychoric correlation coefficient and Spearman's rank correlation coefficient is discussed in Section 4.

Let $\tilde{r}_S$ be the variant of r_S given by $\tilde{r}_S = 2\sin(r_S\pi/6)$. The term variant is used because the maximum absolute difference, $\sup|\tilde{r}_S - r_S|$, is less than 0.02, and the maximum relative difference, $\sup|(\tilde{r}_S - r_S)/r_S|$, is less than 0.05, i.e. five percent. Therefore, the difference between r_S and the variant $\tilde{r}_S$ can for most purposes be considered negligible. By the next result, the variant of Spearman's rank correlation coefficient and the empirical polychoric correlation coefficient are asymptotically equivalent, in a certain sense.

Theorem 5. *For a given underlying joint distribution, if the numbers of categories, r and s, increase such that the maximal difference of cumulative marginal probabilities goes to zero as $r, s \to \infty$, then:*

(a)

$$\lim_{n\to\infty}\lim_{r,s\to\infty} \tilde{r}_S - r_{epc} = 0 \quad \textit{almost surely},$$

(b)

$$\lim_{n\to\infty}\lim_{r,s\to\infty} \tilde{r}_S = \rho_{pc} \quad \textit{almost surely}.$$

Proof. (a). By hypothesis, $||\nabla\vec{u}||_3^3$ and $||\nabla\vec{v}||_3^3$ go to zero as $r, s \to \infty$. By the strong law of large numbers and the Slutsky theorem, the sample analogues go to zero almost surely. The statement then follows by Equation (3) and continuity.

(b). By Theorem 2 of Ekström (2009), $\lim_n \lim_{r,s} r_{epc} = \rho_{pc}$ almost surely under the present hypothesis. Therefore, the statement then follows by (a) and the Slutsky theorem. □

Corollary 6. *If the statement of Assumption A1 is true, then the polychoric correlation coefficient and the variant of Spearman's rank correlation coefficient are asymptotically equivalent in the sense of Theorem 5.*

The interpretation of Theorem 5 and Corollary 6 is that the empirical polychoric correlation coefficient, r_{epc}, and the variant of Spearman's rank correlation coefficient, $\tilde{r}_S$, approximate each other and the theoretical population polychoric correlation arbitrarily well, if only the sample size and the numbers of categories are large enough. General rules-of-thumb are, of course, difficult to establish. One of the main techniques for studying the rate of convergence in situations such as the present is so-called simulation studies, see, e.g., Ekström (2009).

4. Differences between the two

Between the empirical polychoric correlation coefficient and Spearman's rank correlation coefficient there are two distinctive differences. In terms of construction, all polychoric correlation coefficients rest on an assumption of a continuous underlying joint distribution. When applied to contingency tables, Spearman's rank correlation coefficient, being the linear correlation of the ranks, implicitly assumes an underlying joint distribution which is discrete. In terms of value, the absolute value of Spearman's rank correlation coefficient is in general greater than that of the empirical polychoric correlation coefficient, and Spearman's rank correlation coefficient attains the boundary values ± 1 in certain cases when the empirical polychoric correlation coefficient does not. Of these two distinctive differences, the latter is a consequence of the former.

If the two ordinal variables have equal numbers of categories and all joint probabilities off one of the main diagonals are zero, then Spearman's rank correlation coefficient have absolute value one, while the empirical polychoric correlation coefficient have absolute value less than one. Following the axiomatic definition of Rényi (1959), and later Schweizer & Wolff (1981), a measure of association should attain the boundary values ± 1 only if the two variables are strictly monotonic functions of each other. Whether the two variables are strictly monotonic functions of each other or not, in this case, depends on the underlying joint distribution.

In precise mathematical terms, when applied to contingency tables Spearman's rank correlation coefficient being equal to one implies that the variables are strictly increasing functions of each other if and only if the support of the underlying joint distribution is $\mathcal{C} \times \mathcal{D}$, the Cartesian product of the sample spaces of the two ordinal variables, respectively. However, if at least one of the two ordinal variables is the result of some form of discretization, such as grouping of values into categories or even numerical rounding, the implication does not hold. For instance if the two ordinal variables have an underlying continuous joint distribution, cf. Assumption A1, then Spearman's rank correlation coefficient does not satisfy the axiomatic definitions of Rényi (1959) and Schweizer & Wolff (1981).

In practice, many variables are ordinal as a result of difficulties of measurement; examples include quality, design, user-friendliness, esthetics, emotions, opinions, utility, and many more. However, subject experts often regard these mentioned variables as

fundamentally continuous in nature; therefore the underlying continuous joint distribution assumption. The use of Spearman's rank correlation coefficient on ordinal variables of this kind can potentially lead the analyst into concluding that the underlying joint distribution exhibits perfect dependence, even though empirical data cannot possibly imply the conclusion. The empirical polychoric correlation coefficient, on the other hand, attains the bounds ± 1 only if the underlying distribution is perfectly dependent, which is the desired property.

5. Illustrational example and visualizations

In this section, the use of the two measures of association, the empirical polychoric correlation coefficient and Spearman's rank correlation coefficient, is discussed in the light of a data set. Moreover, the data set serves to illustrate visualization techniques for the association of ordinal variables, under the assumption of a postulated continuous joint distribution.

Table 1 shows an excerpt from the World Health Organization (WHO) report *Alcohol, Gender and Drinking Problems* (Obot & Room, 2005). Brazilian men and women, age 17 and older, were surveyed about their alcohol consumption. Table 1 shows how 595 respondents reported their education, and their alcohol consumption during the past twelve months.

When questioned about alcohol consumption, respondents were asked to convert numbers of beers, glasses of wine, et cetera, into grammes of alcohol and then choose one of seven categories. Consequently, alcohol consumption is an inherently continuous variable that has been grouped into ordered discrete categories. Since education is measured in terms of time, education is also considered as being a grouped, or discretized, continuous variable. Therefore, both education and alcohol consumption are considered to have continuous underlying distributions.

For this data set, though, it is difficult to make an assertion about the distributional family of the postulated continuous joint distribution, and therefore the conventional polychoric correlation coefficient is not particularly suitable to use as a measure of association. The empirical polychoric correlation coefficient, on the other hand, is deemed suitable since it rests only on the assumption of existence of an underlying continuous joint distribution. Also of interest for the association analysis, Pearson's chi-square test for independence is rejected on all conventional significance levels.

For the data set of Table 1, the empirical polychoric correlation coefficient is 0.24. A 95% confidence interval, constructed using the percentile method under non-parametric bootstrap, is estimated to $(0.16, 0.31)$. Spearman's rank correlation coefficient, for comparison, is 0.26 with its 95% confidence interval estimated to $(0.18, 0.33)$. The difference between the two, which also is given by Equation (3), is negligible in the sense that is does not impact the conclusions of the association analysis.

TABLE 1. Alcohol consumption versus education, survey data from Brazil 2005.

Education	Alcohol consumption						
	None	←				→	Heavy
≤ 7 years	147	36	42	10	5	4	8
8 to 11 years	80	50	48	26	12	11	10
≥ 12 years	25	32	25	9	4	6	5

Note: Figures represent numbers of respondents.
Source: Obot & Room (2005)

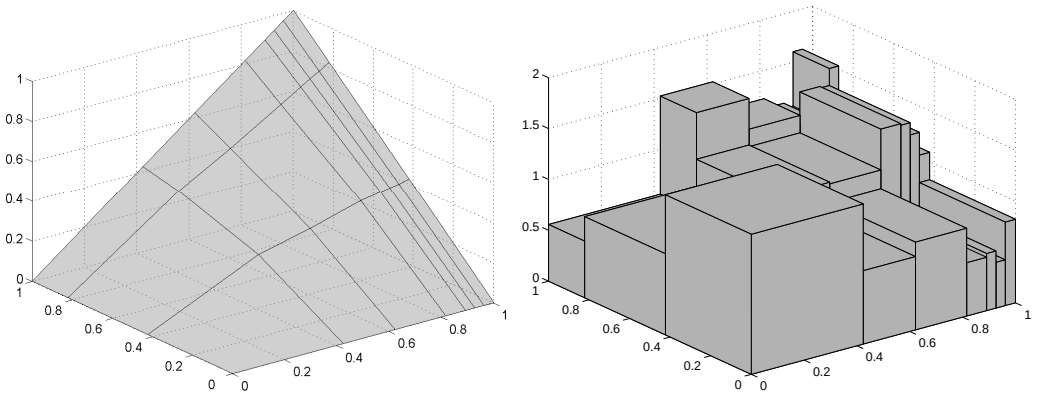
From a theoretical point of view, for this data set the empirical polychoric correlation coefficient is more suitable than Spearman's rank correlation coefficient, since the assumption of a continuous joint distribution is more appropriate than an assumption of an inherently discrete joint distribution. On the other hand, as is illustrated by this example and discussed in Sections 3 and 4, the choice between the two generally makes little difference for the conclusions of the association analysis. Though, if in doubt whether the postulated joint distribution should be considered continuous or discrete, i.e. whether to use the empirical polychoric correlation coefficient or Spearman's rank correlation coefficient, the former is the safer choice in the sense that it is a more conservative estimate of the theoretical population polychoric correlation.

The Schweizer-Sklar standard extension of the empirical copula of the data set of Table 1 is pictured in Figure 2. Each rectangle on the graph corresponds to a cell of Table 1, and the vertices on the graph correspond to the values of the empirical copula. By Corollary 2, the empirical polychoric correlation coefficient and Spearman's rank correlation coefficient are both functions of the integral of the Schweizer-Sklar standard extension of the empirical copula, $\hat{C}_n^S$. The fact that the graph is concave implies that there is a positive association between the two ordinal variables.

Also pictured in Figure 2 is the graph of the density function corresponding to the Schweizer-Sklar standard extension of the empirical copula. In the density graph it is seen that there is more probability mass on the positive diagonal than on average. Hence, the positive association is visualized. In the graph, other aspects of the association, such as for example possible tail dependence, can also be seen. All in all, the graph of the density function is an informative illustration of the association between the two ordinal variables.

6. CONCLUSIONS

By the main theorem of the present article, Spearman's rank correlation coefficient can be expressed as a deterministic transformation of the empirical polychoric correlation coefficient, and vice versa. The transformation is a homeomorphism under given marginal probabilities, has a fixed point at zero, and the two measures of association are



(a) Copula (b) Density

FIGURE 2. Graph of the Schweizer-Sklar standard extension of the empirical copula corresponding to Table 1, and its density function.

asymptotically equivalent in a certain sense. In general, the absolute value of Spearman's rank correlation coefficient is greater than that of the empirical polychoric correlation coefficient.

If one or both of the ordinal variables is the result of some form or discretization, such as grouping of values into categories, Spearman's rank correlation coefficient has the undesired property that the measure of association can equal ± 1 even though empirical data cannot possibly imply that the non-discretized variables are strictly monotonic functions of each other. In this respect, the empirical polychoric correlation coefficient is more conservative and better suited for statistical inference about the association of the underlying, non-discretized variables. Furthermore, from the perspective of Theorem 5 association studies should be designed so that ordinal variables have the largest numbers of categories feasible.

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 Bengt Muthén generously provided valuable comments. My thanks also to two anonymous reviewers and an associate editor who provided valuable comments. This work was supported by the Jan Wallander and Tom Hedelius Research Foundation, project P2008-0102:1.

REFERENCES

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

Ekström, J. (2008). A generalized definition of the polychoric correlation coefficient. In *Contributions to the Theory of Measures of Association for Ordinal Variables*. Ph.D. thesis, Uppsala: Acta Universitatis Upsaliensis.

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

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proc. Roy. Soc. London*, *45*, 135–145.

Nelsen, R. B. (2006). *An Introduction to Copulas, 2nd ed*. New York: Springer.

Nešlehová, J. (2007). On rank correlation measures for non-continuous random variables. *J. Multivariate Anal.*, *98*, 544–567.

Obot, I. S., & Room, R. (Eds.) (2005). *Alcohol, Gender and Drinking Problems*. Geneva: WHO.

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. (1907). *Mathematical contributions to the theory of evolution. XVI. On further methods of determining correlation*, vol. 4 of *Drapers' Company Research Memoirs, Biometric series*. London: Cambridge University Press.

Pearson, K. (1911). *The Grammar of Science, 3rd ed*. London: Adam and Charles Black.

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

Rényi, A. (1959). On non-parametric measures of dependence for random variables. *Acta. Math. Acad. Sci. Hungar.*, *10*, 441–451.

Schweizer, B., & Wolff, E. F. (1981). On measures of dependence. *Ann. Statist.*, *9*, 879–885.

Spearman, C. (1904). The proof and measurement of association between two things. *Amer. J. Psychol.*, *15*, 72–101.

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

UCLA Department of Statistics, 8125 Mathematical Sciences Building, Box 951554, Los Angeles CA, 90095-1554

E-mail address: joakim.ekstrom@stat.ucla.edu