UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Social Sciences* 50

Contributions to the Theory of Measures of Association for Ordinal Variables

JOAKIM EKSTRÖM

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2009

ISSN 1652-9030
ISBN 978-91-554-7498-0
urn:nbn:se:uu:diva-100735

Dissertation presented at Uppsala University to be publicly examined in Sal IV, Universitetshuset, 753 12, Uppsala, Friday, May 15, 2009 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Ekström, J. 2009. Contributions to the Theory of Measures of Association for Ordinal Variables. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* 50. 32 pp. Uppsala. 978-91-554-7498-0.

In this thesis, we consider measures of association for ordinal variables from a theoretical perspective. In particular, we study the phi-coefficient, the tetrachoric correlation coefficient and the polychoric correlation coefficient. We also introduce a new measure of association for ordinal variables, the empirical polychoric correlation coefficient, which has better theoretical properties than the polychoric correlation coefficient, including greatly enhanced robustness.

In the first article, entitled ``On the relation between the phi-coefficient and the tetrachoric correlation coefficient", we show that under given marginal probabilities there exists a continuous bijection between the two measures of association. Furthermore, we show that the bijection has a fixed point at zero for all marginal probabilities. Consequently, the choice of which of these measures of association to use is for all practical purposes a matter of preference only.

In the second article, entitled ``A generalized definition of the tetrachoric correlation coefficient", we generalize the tetrachoric correlation coefficient so that a large class of parametric families of bivariate distributions can be assumed as underlying distributions. We also provide a necessary and sufficient condition for the generalized tetrachoric correlation coefficient to be well defined for a given parametric family of bivariate distributions. With examples, we illustrate the effects on the polychoric correlation coefficient of different distributional assumptions.

In the third article, entitled ``A generalized definition of the polychoric correlation coefficient", we generalize the polychoric correlation coefficient to a large class of parametric families of bivariate distributions, and show that the generalized and the conventional polychoric correlation coefficients agree on the family of bivariate normal distributions. With examples, we illustrate the effects of different distributional assumptions on the polychoric correlation coefficient. In combination with goodness-of-fit p-values, the association analysis can be enriched with a consideration of possible tail dependence.

In the fourth article, we propose a new measure of association for ordinal variables, named the empirical polychoric correlation coefficient. The empirical polychoric correlation coefficient relaxes the fundamental assumption of the polychoric correlation coefficient so that an underlying joint distribution is only assumed to exist, not to be of a particular parametric family. We also provide an asymptotical result, by which the empirical polychoric correlation coefficient converges almost surely to the true polychoric correlation under very general conditions. Thus, the proposed empirical polychoric correlation coefficient has better theoretical properties than the polychoric correlation coefficient.

Keywords: measure of association, ordinal variables, contingency table, phi-coefficient, tetrachoric correlation coefficient, polychoric correlation coefficient, empirical polychoric correlation coefficient, robustness

Joakim Ekström, Statistics, Box 513, Uppsala University, SE-75120 Uppsala, Sweden

ISSN 1652-9030
ISBN 978-91-554-7498-0
urn:nbn:se:uu:diva-100735 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-100735)

List of Articles

This thesis is based on the following articles, which are referred to in the text by their Roman numerals.

I Ekström, J. (2008) On the relation between the *phi*-coefficient and the tetrachoric correlation coefficient.

II Ekström, J. (2008) A generalized definition of the tetrachoric correlation coefficient.

III Ekström, J. (2008) A generalized definition of the polychoric correlation coefficient.

IV Ekström, J. (2009) An empirical polychoric correlation-coefficient.

Contents

1. Ordinal variables

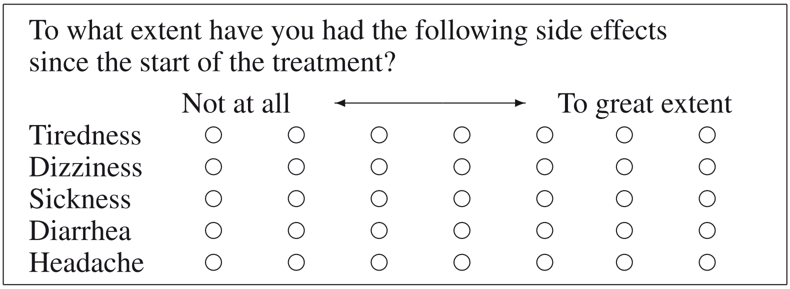
Some quantities can be measured with great accuracy. Time, for example, can be measured with mind-blowing precision; on a scale of 10^{-18} seconds. Other quantities for which there are amazing measurement instruments are weight, length, radiation, electric charge, just to name a few.

In the first half of the 20th century, one prominent application for statistical methods was the agricultural sciences. A common type of study was based on the following setup. A field is split into a number of subfields and the subfields are then randomly given different treatments, such as for instance different quantities of fertilizer. In this application, the amount of fertilizer used and the amount of agricultural production yielded can be measured with great accuracy. Generally speaking, in that time period statistical methods were developed for these kinds of easily measured quantities.

In one in many ways related scientific discipline, medicine, the situation is quite different. While it is often easy to measure the treatment, for example the amount of a pharmaceutical given to a patient, it is often more difficult to measure its effects. As an example, many pharmaceuticals are developed for the purpose of giving a medicinal treatment equivalent to existing pharmaceuticals, but with less severe side effects. However, the severity of a side effect is not easily measured. In practice, patients are often asked to describe the severity of the side effect on a scale of some kind, and that type of measurement has some inherent weaknesses, which will be discussed. Other variables that are difficult to measure are quality, design, user-friendliness, esthetics, emotions, opinions, utility, and many more.

Variables that can be measured only in terms of ranks are called ordinal variables. One can loosely say that if the measurement of a quantity is blurred to such an extent that it is only meaningful to compare different measurements in terms of their ranks, then the measured quantity is an ordinal variable. Formally, for every pair of values of an ordinal variable there exists an order relation, but no binary operations. Thus, values of an ordinal variable cannot be, for instance, added or multiplied. In this thesis, we will in particular study ordinal variables that have a range of finite cardinality, i.e. a range with a finite number of elements. Such ordinal variables are sometimes also called ordered categorical variables, and the cardinality of the range is called the number of categories.

Consider again the example with medical side effects. Patients are asked to describe the extent to which they have had some particular side effects



To what extent have you had the following side effects since the start of the treatment?

	Not at all		←		→		To great extent
Tiredness	○	○	○	○	○	○	○
Dizziness	○	○	○	○	○	○	○
Sickness	○	○	○	○	○	○	○
Diarrhea	○	○	○	○	○	○	○
Headache	○	○	○	○	○	○	○

Figure 1.1: Example of a questionnaire used for measuring the severity of side effects of a pharmaceutical.

on a scale of one to seven, where one is "Not at all" and seven is "To great extent". An example of such a questionnaire is pictured in Figure 1.1. The stated severity of a side effect is an ordinal variable, since it is only meaningful to compare different observations in terms of their ranks. In this application, it is not meaningful to define a binary operation, such as addition, on the set of values of the variable. Furthermore, since the variable can only attain seven separate values, the variable is a category variable with seven categories.

While ordinal variables are in practice very common, the statistical methodology is not as well developed for ordinal variables as for ordinary quantitative variables. There are many reasons for this unfortunate fact. Noticeably, the methodological difficulties of the statistical analysis of ordinal variables start at the very definition of a random variable. By Kolmogorov's construction, a random variable is a measurable function from a probability space to an algebraical field, most commonly the real field **R**. The range of an ordinal variable, on the other hand, is algebraically only a totally ordered set. Consequently, for (random) ordinal variables it is not meaningful, nor algebraically possible, to apply such a fundamental thing as the expectation operator, as an example.

Because of the severity of these difficulties, a common approach to the statistical analysis of ordinal variables is to assume that the variables (in fact) are not ordinal. Technically, this is done by assuming a metric. Consider the example with medical side effects, say that a medical doctor finds an ingenious argument why category one in fact represents the value zero, and that the distance between each pair of subsequent categories equals unity. Under this convenient assumption, the medical doctor circumvents every statistical problem associated with ordinal variables, and can then continue using any ordinary statistical method he, or she, feels like, such as e.g. regression analysis, time series analysis, analysis of variance, factor analysis, and so on. At the time of writing of this thesis, metric-assumptions based on the inverse normal distribution function are common. The problem with assuming a metric, however, is that the choice of metric is completely arbitrary. The arbitrariness

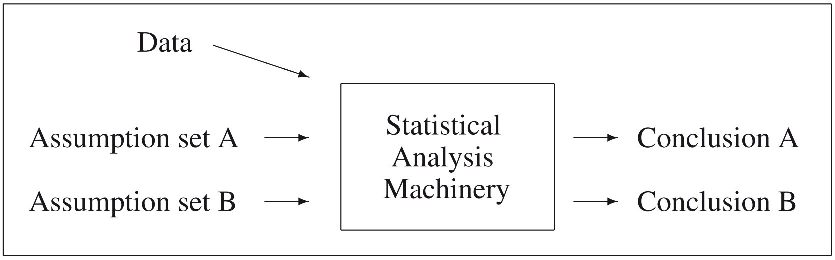


Figure 1.2: Flowchart of a statistical method that takes data and an assumption set and yields a conclusion. In the chart, two conclusions are produced from a fixed data set under two (both reasonable) assumption sets. If the conclusions (appreciably) differ, then the statistical method is non-robust.

comes from the very nature of the problem. This is easily realized after considering the simple fact that if the metric was indeed known, then it would not make sense to analyze the variable as an ordinal variable in the first place.

Fundamentally, statistical methods take empirical data in combination with assumptions and yield statistical conclusions. If for a method, the statistical analysis of a given data set yields conclusions that are more or less consistent under different (reasonable) assumptions, then the statistical method is said to be robust. If, on the other hand, the method yields conclusions that are appreciably different under different, again reasonable, assumptions, then the statistical method is non-robust. A flowchart of this process is pictured in Figure 1.2.

Ideally, a statistical conclusion follows from empirical evidence only. For non-robust statistical methods, however, the conclusion may be a direct, or indirect, consequence of the assumptions. This is a serious problem, because the purpose of statistical conclusions is that they should be interpretable as existence of empirical evidence. For example, if the conclusion of the statistical analysis of the medical side effects, see Figure 1.1, is that the new pharmaceutical has less severe side effects than existing pharmaceuticals, then one wants to be able interpret the conclusion as existence of empirical evidence that the new pharmaceutical really has less severe side effects than existing ones. One does not want the statistical conclusion to follow as a consequence of anything else.

An equally serious problem is that non-robust statistical methods open up the possibility for less experienced scientists, or perhaps scientists with conflicts of interest, to reverse-engineer the set of assumptions needed to reach a desired conclusion. Once a set of assumptions suitable for the desired conclusion is found, a scientist can most often find some arguments why the reverse-engineered set of assumptions is natural for the particular application. For the scientific community, this situation can be quite difficult to deal with. Statisticians have a special responsibility to raise awareness of this unfortunate reverse-engineering possibility. In the situation with an assumed metric for an ordinal variable, described in a preceding paragraph, there is in general no reason to believe that the statistical analysis of the ordinal variable will be robust to changes of the metric-assumption.

In this thesis, we study measures of association for ordinal variables, with a particular interest for robustness properties.

2. Measures of association

Statistical independence between random variables is one of the most fundamental concepts of multivariate statistics. Random variables that are not independent are called dependent. Dependence relations between random variables is indeed one of the most studied subjects in statistics.

In almost all scientific disciplines, it is of great interest to study how variables are related to each other. If two random variables are dependent, then one intuitively can say that they contain information on each other. As a consequence, one can make predictions of a random variable, X, using information provided by observations of other random variables that are interdependent with X. Furthermore, in applications one is often interested in controlling the variable X by influencing other variables that are interdependent with X. For example in medicine, one is interested in minimizing the severity of side effects by adjusting other variables, such as, e.g., environmental factors and levels of pharmaceutical substances, that are interdependent with the side effects. For another example, European governments are at the time of writing of this thesis trying to control tobacco consumption by legislatively adjusting variables that are believed to be interdependent with tobacco usage.

If two random variables, X and Y, are each completely determined by the other, then X and Y are said to be perfectly dependent. It is easily realized that perfect dependence corresponds to existence of a bijection between the random variables, i.e. a function that is one-one and onto. A function f of a random variable X is said to be a bijection almost surely if it is a bijection everywhere except possibly on a set with zero probability, i.e. $P(\{\omega : X(\omega) \mapsto f(X(\omega)) \,\mathrm{not\,a\,bijection}\}) = 0$. If $X = f(Y)$ is a bijection almost surely, then the conditional variance $\mathrm{Var}(X|Y)$ equals zero. On the other hand, if the random variables X and Y are independent, then $\mathrm{Var}(X|Y) = \mathrm{Var}(X)$. This suggests the existence of a continuum of dependence, with perfect dependence and independence as the two extremes.

We define $R(X,Y)$ to be a measure of dependence if, for random variables X and Y, it satisfies the following properties.

(D1) R is defined for every pair of random variables.

(D2) $R(X,Y) = R(Y,X)$.

(D3) $0 \leq R(X,Y) \leq 1$.

(D4) $R(X,Y) = 0$ if and only if X and Y are independent.

(D5) $R(X,Y) = 1$ if and only if each of X and Y is a bijection almost surely of the other.

(D6) If f and g are bijections almost surely, then $R(f(X),g(Y)) = R(X,Y)$.

(D7) If (X,Y) and $\{(X_n,Y_n)\}_{n=1}^{\infty}$ are pairs of random variables with joint distribution functions H and H_n respectively, and if the sequence $\{H_n\}$ converges to H, then $\lim_{n\to\infty} R(X_n,Y_n) = R(X,Y)$.

These axioms were introduced in Rényi (1959) in a slightly modified form. Property D7, which is a continuity property, was not included in Rényi (1959), but in Schweizer & Wolff (1981).

If, in the setup above, bijections are restricted to strictly monotonic functions, then it becomes possible to define positive and negative dependence, or more commonly termed positive and negative association. A measure of association $S(X,Y)$, where X and Y are random variables, satisfies the following properties (Nelsen, 2006).

(E1) S is defined for every pair of random variables.

(E2) $S(X,Y) = S(Y,X)$.

(E3) $-1 \leq S(X,Y) \leq 1$, $S(X,X) = 1$, and $S(X,-X) = -1$.

(E4) If X and Y are independent, then $S(X,Y) = 0$.

(E5) $S(X,-Y) = S(-X,Y) = -S(X,Y)$.

(E6) If f and g are strictly increasing functions almost surely, then $S(f(X),g(Y)) = S(X,Y)$.

(E7) If (X,Y) and $\{(X_n,Y_n)\}_{n=1}^{\infty}$ are pairs of random variables with joint distribution functions H and H_n respectively, and if the sequence $\{H_n\}$ converges to H, then $\lim_{n\to\infty} S(X_n,Y_n) = S(X,Y)$.

Properties E3 and E6, above, imply that whenever there exists a strictly increasing function f such that $f(X) = Y$ almost surely, then $S(X,Y) = 1$. Furthermore, this in combination with Property E5 yields that whenever there exists a strictly decreasing function g such that $g(X) = Y$ almost surely, then $S(X,Y) = -1$. As a consequence, one can loosely say that a measure of association between X and Y contains information on the extent to which X and Y can be represented as strictly monotonic functions of each other.

One often thinks of the linear correlation coefficient, defined by $\rho(X,Y) = \mathrm{Cov}(X,Y)/\left(\mathrm{Var}(X)\mathrm{Var}(Y)\right)^{1/2}$, as a first choice of measure of association, but in fact ρ does not satisfy Property E6 above, something which is easily seen. However, the linear correlation coefficient satisfies a variant of Property E6, where functions are restricted to first-degree polynomials. Therefore, the linear correlation coefficient, ρ, is often referred to as a measure of linear association.

Let X and Y be continuous random variables, and let ρ_S be defined as $\rho_S(X,Y) = \rho(F(X),G(Y))$, where F and G are the distribution functions of X and Y, respectively. Then ρ_S is a measure of association, satisfying properties E1-E7, called the Spearman grade correlation. The Spearman grade correlation is the population analog of the well known Spearman rank correlation.

Because measures of association are invariant under strictly increasing transformations, it makes sense to standardize the distribution functions of the random variables X and Y. Since distribution functions are strictly increasing functions almost surely, it is easy to standardize the distribution functions of X and Y to be standard-uniformly distributed, i.e. uniformly distributed on the unit interval, $I = [0,1]$. This is easy, because if F is the distribution function of a random variable X, then $F(X)$ has a (discrete or continuous) standard uniform distribution, as is well known.

The joint distribution functions of pairs of uniformly distributed continuous random variables constitute a special class of functions called copulas. Formally, a copula is a function $C : I^2 \to I$ such that for every $u, v \in I$, $C(u,0) = C(0,v) = 0$, and $C(u,1) = u$, $C(1,v) = v$, and moreover, for all $u_1 \leq u_2$, $v_1 \leq v_2$, where u_1, u_2, v_1 and v_2 are elements of I,

$$C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0\,.$$

By Sklar's theorem, for every joint distribution function H with marginal distribution functions F and G, there exists a copula C such that $H(x,y) = C(F(x),G(y))$, for all points (x,y) in the domain of H. If F and G are continuous, then C is unique, otherwise C is uniquely determined on the cartesian product of the ranges of X and Y (Nelsen, 2006).

Copulas have many properties that are useful for the study of association. For every copula C, the inequality $\max(u+v-1,0) \leq C(u,v) \leq \min(u,v)$ holds. The functions $\max(u+v-1,0)$ and $\min(u,v)$ are copulas, which is easily verified from the definition, and are commonly denoted W and M, respectively. Furthermore, the function $\Pi(u,v) = uv$ is also a copula. The copulas W, Π and M are of particular interest for the study of association because they correspond to perfect negative dependence, independence and perfect positive dependence, respectively. Moreover, copulas are useful because, as a result of Property E6, measures of association for the random variables X and Y can be expressed as functionals of the copula C of X and Y. For example, the Spearman grade correlation of random variables X and Y, with copula C, can be written as

$$\rho_S(X,Y) = \rho_S(C) = 12 \int_{I^2} C d\lambda - 3\,,$$

where λ is the Lebesgue measure (Nelsen, 2006). This result will be used extensively in this thesis.

Because the conditions in the definition of a copula are so easily checked, the development of the theory of copulas has led to the discovery of many new bivariate distributions. In a number of examples in this thesis, we use bivariate distributions defined by their copulas, such as for example the Clayton, Frank and Genest-Ghoudi families.

In figures 2.1, 2.2 and 2.3, the graphs of copulas W, Π and M, respectively, are pictured. Also, in Figure 2.4, the graph of the copula corresponding to

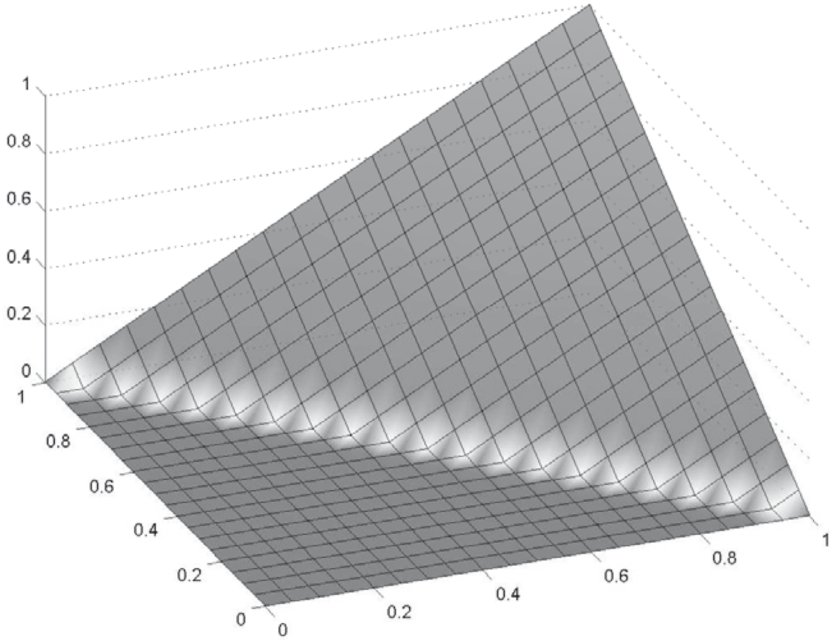


Figure 2.1: The copula W, representing perfect negative dependence between the two random variables.

the bivariate normal distribution with correlation parameter $\rho = 0.309$ is pictured. This particular copula will be of interest for an example in the next section. The family of copulas corresponding to bivariate normal distributions are commonly called Gaussian copulas.

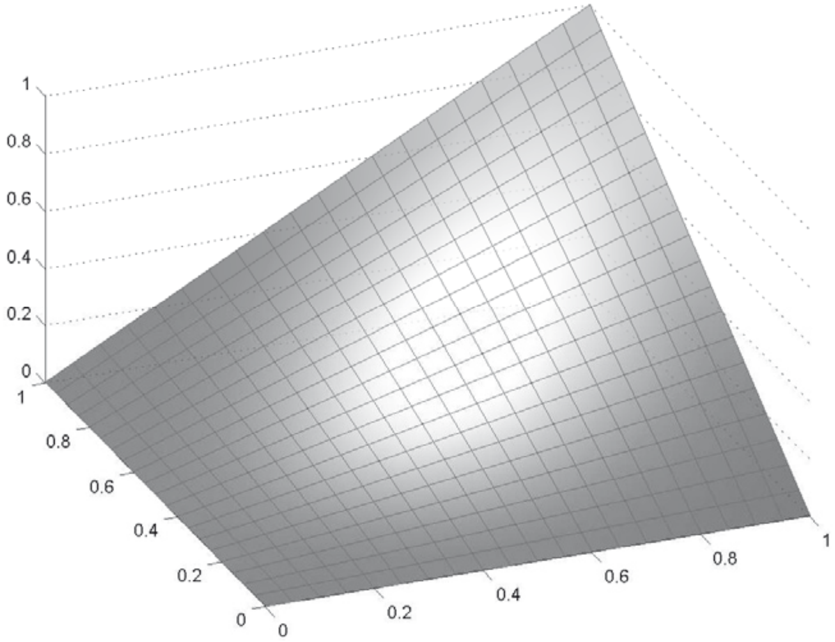


Figure 2.2: The copula Π, representing independence between the two random variables.

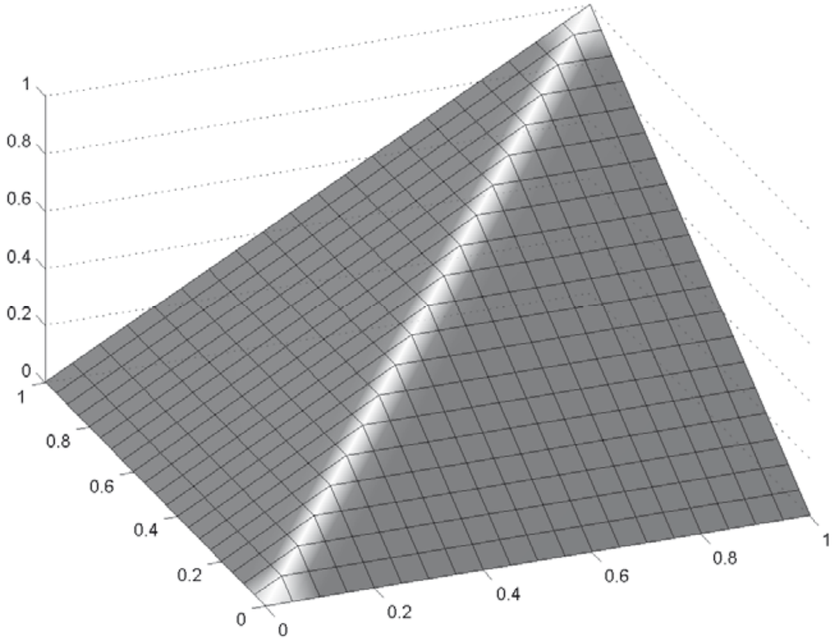


Figure 2.3: The copula *M*, representing perfect positive dependence between the two random variables.

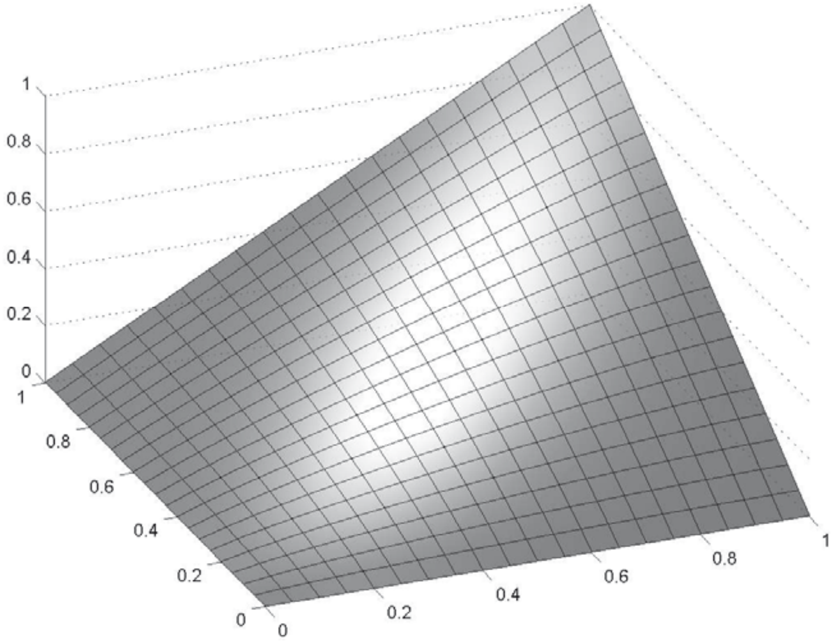


Figure 2.4: The copula corresponding to the bivariate normal distribution with correlation parameter $\rho = 0.309$. This copula is also called the Gaussian copula with parameter $\rho = 0.309$. Random variables with this copula are weakly positively associated. Note that for all points of the domain, the graph is sandwiched by the graphs of the copulas Π and M. This implies that the pair of random variables has a dependence relation that is somewhere in between independence and perfect positive dependence.

3. The tetrachoric and polychoric correlation coefficients

The tetrachoric and polychoric correlation coefficients are measures of association for ordinal variables. In this thesis, the tetrachoric and polychoric correlation coefficients play prominent roles.

One interesting historical note on the tetrachoric and polychoric correlation coefficients is that they were proposed by Karl Pearson. Karl Pearson (1857-1936) was originally a physician who became interested in the theory of natural selection and evolution. In a series of 18 articles entitled “Mathematical contributions to the theory of evolution”, Pearson suggested some methods that could be useful for the study of evolution. In the first article in the series, for example, he used a concept borrowed from elasticity called bending moments. This was the embryo of what today is known as moments of random variables, and the Pearsonian method of moments. In the second article, Pearson suggested a system of seven continuous distributions to complement the normal distribution, which at the time was prevalent. These distributions are today known as, e.g., the beta distribution, the gamma distribution, the chi-square distribution, the exponential distribution and the F distribution. Later in the series, Pearson explored the relation between linear regression and correlation, and introduced the chi-square goodness-of-fit test. Especially noteworthy is that the chi-square goodness-of-fit suggestion contained the embryo of what is now known as statistical decision theory. In 1901, Pearson co-founded the statistical journal Biometrika. Karl Pearson is by many considered one of the fathers of modern statistics.

In the 7th of the 18 articles entitled “Mathematical contributions to the theory of evolution”, with subtitle “On the correlation of characters not quantitatively measurable”, Pearson introduced the fundamental idea of what has later become known as the tetrachoric and polychoric correlation coefficients. The idea is to consider a 2×2 contingency table as a dichotomization of a bivariate standard normal distribution.

A 2×2 contingency table has four elements, but since the probabilities sum up to one, the table is completely determined by the triple (p_X, p_Y, p_a), where p_X and p_Y are the marginal probabilities of “positive” values of the two dichotomous variables X and Y, respectively, and p_a is the joint probability of “positive” values of both variables. Pearson suggested finding the parameter of the bivariate standard normal distribution such that the volumes of the distribution equal the joint probabilities of the contingency table. Of course, this parameter represents the linear correlation coefficient of the normally distributed random variables postulated. The equation can be written

$$p_a = \int_{\Phi^{-1}(1-p_X)}^{\infty} \int_{\Phi^{-1}(1-p_Y)}^{\infty} \phi(x,y,r_{tc})dxdy, \tag{3.1}$$

where Φ is the univariate standard normal distribution function and ϕ is the bivariate standard normal density function. By Pearson's suggestion, the parameter r_{tc} that solves the integral equation (3.1) can be considered the correlation of the contingency table. The fact that Equation (3.1) always has a unique solution, r_{tc}, for every contingency table is proved in the first article of this thesis.

As an example, consider the contingency table $(p_X, p_Y, p_a) = (0.5, 0.5, 0.3)$. For this contingency table, the lower limits of integration in the right hand side of Equation (3.1) are both $\Phi^{-1}(0.5) = 0$. Thus, the integral in the right hand side of Equation (3.1) is over the first quadrant of the real plane, i.e. the upper-right quadrant in the standard coordinate axis system.

In figures 3.1, 3.2, 3.3 and 3.4, the graphs of density functions of bivariate normal distributions with parameter values -0.8, 0, 0.309 and 0.8, respectively, are pictured. The figures illustrate how Pearson imagined that one could control the probability mass over the quadrant by adjusting the correlation parameter, ρ. In Figure 3.1, the correlation parameter is $\rho = -0.8$, and the probability mass over the first quadrant, which in this example corresponds to the integral of Equation (3.1), equals 0.1024. Since this is less than the corresponding joint probability of the contingency table, $p_a = 0.3$, Equation (3.1) is not satisfied. In Figure 3.2, the correlation parameter is $\rho = 0$ and the probability mass over the first quadrant is 0.25, which is also less than the corresponding joint probability of the contingency table, $p_a = 0.3$, and thus Equation (3.1) is not satisfied with this parameter value either. However in Figure 3.3, the correlation parameter is $\rho = 0.309$, and here the probability mass over the first quadrant equals 0.3 and since the corresponding joint probability of the contingency table, p_a, also equals 0.3, Equation (3.1) is satisfied. Thus, the correlation of the contingency table $(0.5, 0.5, 0.3)$, as of Pearson's reasoning, is $r_{tc} = 0.309$. The graph of the copula corresponding to this bivariate normal distribution, with correlation parameter $\rho = 0.309$, is pictured in Figure 2.4. Finally, to conclude this example, in Figure 3.4, the correlation parameter is $\rho = 0.8$ and the probability mass over the first quadrant is 0.3976, which is greater than the corresponding joint probability of the contingency table, $p_a = 0.3$.

The article (Pearson, 1900) is 48 pages long, but the fundamental idea, for which the article has become known, is presented in less than half a page. The bulk of the article is about methods to approximate a solution to the integral equation (3.1), above. This was done by method of series expansion. Later, tetrachoric series were used, and it is most probably therefrom the name tetrachoric correlation comes. Pearson (1900), however, refers to the measure of association only as “the method of the present memoir”. Nowadays, the integral equation is easily solved using computer assisted numerical optimization, and consequently, the method of series expansion has become obsolete. In spite of this, the name lingers on.

The polychoric correlation coefficient was introduced by Ritchie-Scott (1918), and is an extension of the tetrachoric correlation coefficient to general ordinal variables, i.e. to general $r \times s$ contingency tables. The extension is not trivial because the integral equation analogous to (3.1) does in general not have a solution. Ritchie-Scott suggested to dichotomize the variables of the $r \times s$ contingency table in all possible ways and then to find a tetrachoric correlation coefficient for every dichotomization. The polychoric correlation coefficient then corresponds to a weighted average of those so obtained tetrachoric correlation coefficients.

Tallis (1962) suggested that a polychoric correlation coefficient can be fitted to the contingency table with respect to a multiplicative loss function referred to as a likelihood. Martinson & Hamdan (1971) merged the idea of Tallis (1962) with the works of Pearson and Ritchie-Scott, along with some computational simplifications. Martinson & Hamdan (1971) also provided some additional suggestions of loss functions. Moreover, Olsson (1979) suggested a slightly modified approach, allowing for reclassifications.

Since Pearson’s article was published at the turn of the 19th century, the research on these measures of association has largely been confined to numerical optimization algorithms and simulation studies. Theoretically, though, not much has been written on the subject. The measures of association have not even been given rigorous definitions, and a proof of existence and uniqueness of a solution to the integral equation (3.1) has never been published. An unfortunately flawed attempt was made by Juras & Pasaric (2006).

Many historians would say that Karl Pearson’s strength was not mathematical rigor. Instead, Pearson’s great contribution to the theory of modern statistics was to publish lots of ideas that kick-started scientific progress and would constitute a foundation for statistics for many years. Other more rigorous statisticians would later derive the properties of many of the methods Pearson suggested.

In this thesis, we take a close look at certain properties of some measures of association for ordinal variables, including the tetrachoric and polychoric correlation coefficients.

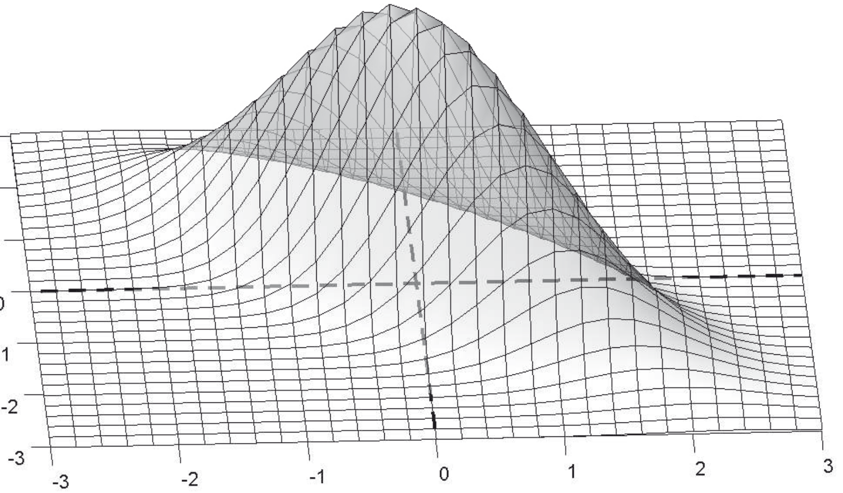


Figure 3.1: The density function of the bivariate standard normal distribution with correlation parameter $\rho = -0.8$. The probability mass over the upper-right quadrant equals 0.1024.

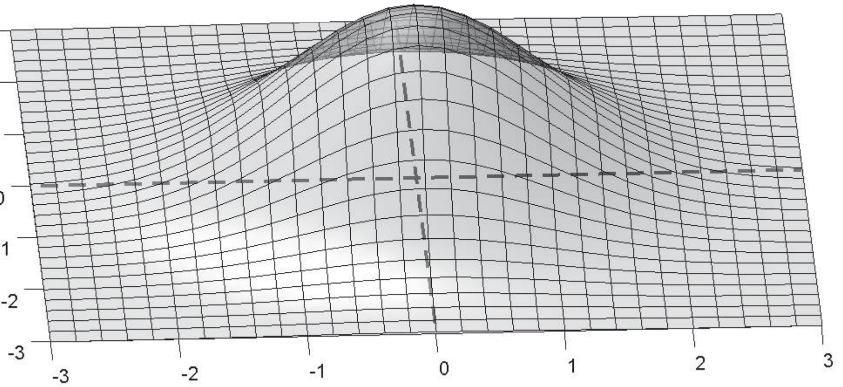


Figure 3.2: The density function of the bivariate standard normal distribution with correlation parameter $\rho = 0$. Random variables with this bivariate distribution are independent. The probability mass over the upper-right quadrant equals 0.25.

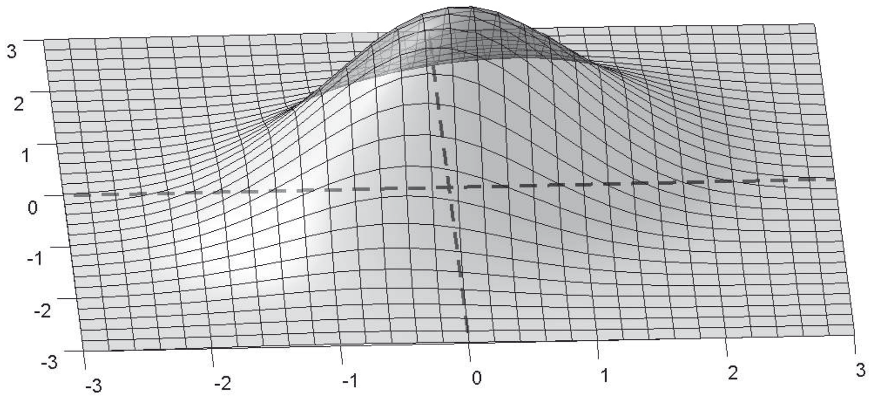


Figure 3.3: The density function of the bivariate standard normal distribution with correlation parameter $\rho = 0.309$. The probability mass over the upper-right quadrant equals 0.3.

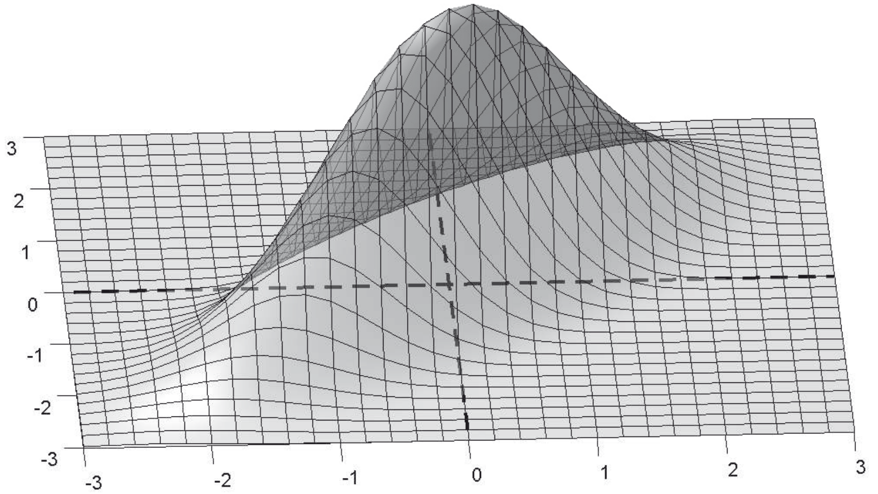


Figure 3.4: The density function of the bivariate standard normal distribution with correlation parameter $\rho = 0.8$. The probability mass over the upper-right quadrant equals 0.3976.

4. Contributions in this thesis

In this thesis, we consider measures of association for ordinal variables from a theoretical perspective. In particular, we study the *phi*-coefficient, the tetrachoric correlation coefficient and the polychoric correlation coefficient. We also introduce a new measure of association for ordinal variables, the empirical polychoric correlation coefficient, which has better theoretical properties than the polychoric correlation coefficient, including greatly enhanced robustness.

All measures of association studied are given mathematically rigorous definitions, and all necessary assumptions are formally stated. We provide important results on basic properties such as, e.g., necessary and sufficient conditions for existence and uniqueness of coefficients under various circumstances.

Furthermore, we provide some results on relations between different measures of association. Most notably, we show existence of a continuous bijection between the *phi*-coefficient and the tetrachoric correlation coefficient under given marginal probabilities. As a consequence, whether to use the *phi*-coefficient or the tetrachoric correlation coefficient is a matter of preference only. Moreover, this previously unknown result amounts to a new take on the so-called Pearson-Yule debate, see Yule (1912) and Pearson & Heron (1913).

One theoretically important result in this thesis is that the tetrachoric and polychoric correlation coefficients can be expressed as a functional of the joint distribution function. More precisely, if H is the joint distribution function of the ordinal variables, as given by the fundamental assumption, then the tetrachoric and the polychoric correlation coefficients, r_{pc}, are given by the identity

$$r_{pc} = 2\sin(\rho_S(H)\,\pi/6)\,, \tag{4.1}$$

where ρ_S is the Spearman grade correlation (see Section 2).

Using this identity, we generalize the tetrachoric and polychoric correlation coefficients so that a large class of parametric families of bivariate distributions can be assumed as underlying distributions. As a consequence of the generalization, however, it becomes evident that the tetrachoric and polychoric correlation coefficients are not robust to changes of the distributional assumption. Furthermore, examples illustrate that the polychoric correlation coefficient, which in general has to be fitted, is not robust to changes of the choice of loss function either. This severe lack of robustness amounts to a major methodological problem for these two measures of association.

However, this thesis also contains a suggestion that addresses the deficiencies discussed in the preceding paragraph. We suggest a new measure of association for ordinal variables that is based on Pearson's original idea, but relaxes the assumptions and greatly enhances robustness. The new measure of association for ordinal variables, named the empirical polychoric correlation coefficient, has better theoretical properties than the polychoric correlation coefficient and it also performs better in terms of robustness and standard deviation in a simulation study conducted. In a reference to a comment in Section 1, the empirical polychoric correlation coefficient does not require the implicit assumption of a metric, only the assumption of the existence of a metric.

Throughout the thesis, the theory is illustrated with several examples, both constructed examples and examples based on real-world data.

5. Summary of articles

In the first article of this thesis, entitled "On the relation between the *phi*-coefficient and the tetrachoric correlation coefficient", we rigorously define the two measures of association, show that they are well defined for all 2×2 contingency tables, and that under given marginal probabilities there exists a continuous bijection between the two. Furthermore, we show that the bijection has a fixed point at zero for all marginal probabilities. Consequently, the choice of which of these measures of association to use is for all practical purposes a matter of preference only. This result also gives new input to the so-called Pearson-Yule debate. Moreover, the result can be used to construct a numerical table of tetrachoric correlation coefficients, converted from the marginal probabilities and the *phi*-coefficient, which is easily calculated by hand.

In the second article, entitled "A generalized definition of the tetrachoric correlation coefficient", we generalize the tetrachoric correlation coefficient so that a large class of parametric families of bivariate distributions can be assumed as underlying distributions. We also provide a necessary and sufficient condition for the generalized tetrachoric correlation coefficient to be well defined for a given parametric family of bivariate distributions. Furthermore, we provide some sufficient criteria which can be useful for practical purposes. We also show that the generalized and conventional definitions agree on the parametric family of bivariate normal distributions. With examples, we illustrate the implications of different distributional assumptions, and discover that the tetrachoric correlation coefficient is not robust to changes of the distributional assumption. In fact, quite the opposite seems to hold true. There are even examples where the conclusion of the association analysis under one assumption is that the ordinal variables are perfectly dependent, while the conclusion under a different assumption is that the ordinal variables are independent. Consequently, the conclusion of the association analysis can vary from the variables being perfectly dependent, to the variables being independent, or anything in between, only as a consequence of a change of the distributional assumption. Furthermore, with S&P 100 stock data, as of year 2006, we exemplify the fact that a correct distributional assumption is vitally important for the conclusions of the tetrachoric correlation association analysis.

Recall that the polychoric correlation coefficient is an extension of the tetrachoric correlation coefficient for dichotomous variables to general ordinal variables, i.e. from 2×2 contingency tables to general $r \times s$ contingency tables. In the third article, entitled “A generalized definition of the polychoric correlation coefficient”, we give a rigorous definition of the polychoric correlation coefficient, and show that a solution of the equation, via which the measure of association is defined, does in general not exist if one of the numbers of categories, r and s, is greater than 2 and the other number is greater than or equal to 2. We present some loss functions with which a polychoric correlation coefficient can be fitted, and give a necessary and sufficient condition for existence of a fitted coefficient. We generalize the polychoric correlation coefficient to a large class of parametric families of bivariate distributions, and show that the generalized and the conventional polychoric correlation coefficients agree on the family of bivariate normal distributions. Because a coefficient in general needs to be fitted, it is possible to test different distributional assumptions based on goodness of fit. We provide general suggestions for goodness-of-fit tests. With examples, we illustrate the effects on the polychoric correlation coefficient of different distributional assumptions. In combination with the goodness-of-fit p-values, the association analysis can be enriched with a consideration of possible tail dependence. In general, however, the polychoric correlation coefficient is not robust to changes of the distributional assumption, nor to changes of the loss function.

In the fourth and final article, we propose a new measure of association for ordinal variables, named the empirical polychoric correlation coefficient. The simple idea is to plug in the empirical distribution into the functional (4.1), above. The approach demands a bit of care, however, since the empirical copula is only defined on certain points, and since we want the new measure of association to have as good properties as possible. The empirical polychoric correlation coefficient relaxes the fundamental assumption of the polychoric correlation coefficient so that an underlying joint distribution is only assumed to exist, not to be of a particular parametric family. Put differently, the empirical polychoric correlation coefficient does not need an implicit assumption of a metric, only an assumption of the existence of a metric. The latter assumption is substantially weaker than the first, and represents a new approach to statistical analysis of ordinal variables. We show that the empirical polychoric correlation coefficient is well defined for all contingency tables, that it has range $[-1,1]$ and that it is zero if the ordinal variables are independent. We also provide an asymptotical result, by which the empirical polychoric correlation coefficient converges almost surely to the true polychoric correlation under very general conditions. Thus, the proposed empirical polychoric correlation coefficient has better theoretical properties than the polychoric correlation coefficient.

In a simulation study, the empirical polychoric correlation coefficient performs considerably better in terms of robustness and generally better in terms of standard deviation, than the polychoric correlation coefficient. The empirical polychoric correlation coefficient is, however, biased so that its absolute value is too small. For 3×3 contingency tables, the empirical polychoric correlation coefficient was consistently 20% too small, for 5 × 5 contingency tables 8% too small and for 7 × 7 contingency tables 4% too small. The fact that the bias goes to zero follows by the asymptotic result previously mentioned. Moreover, the empirical polychoric correlation coefficient, by design, fits every contingency table perfectly, and demands neither fitting nor optimization. In stark contrast to the polychoric correlation coefficient, the empirical polychoric correlation coefficient is easily calculated by hand.

For the practitioner, the empirical polychoric correlation coefficient has the advantages that neither a parametric family of distributions nor a loss function needs to be chosen and argued for. If the underlying distribution is a mixture, then that does not pose a problem either. Furthermore, goodness-of-fit tests are unnecessary because the empirical coefficient fits every contingency table perfectly every time. All in all, the empirical polychoric correlation coefficient brings considerable advantages, both theoretical and practical, for the practitioner.

6. Acknowledgements

I am grateful to all people that have been supportive during my time as a PhD-student. In particular my supervisor Rolf Larsson, who has been always encouraging, available and readily helpful. During the last few months, you stood by me and supported me steadfastly. I am deeply appreciative of our pleasant working relationship and happy to have been your student. Adam Taube, who always lifts my spirits with enthusiasm and contagious cheerfulness. Thank you for introducing me to the field of applied biostatistics. Anders Ågren, who quietly but noticeably shows consideration and approval. I appreciate your constructive feedback on my manuscripts and our discussions on statistical history. A special acknowledgement to my assistant supervisor Lars Forsberg for help with the S&P financial data set. Thanks also to all other people at the department, including the administrative staff, that have been helpful with a range of matters, from proofreading to working the bureaucracy.

I have greatly enjoyed and appreciated the camaraderie of my PhD-student colleagues. Nicklas Korsell and Daniel Preve, you gave me an early, no-nonsense explanation of the life as a PhD-student, that I have valued enormously. Daniel, thank you for a wonderful stay in Singapore and great feedback on my manuscripts. Jim Blevins, Hao Luo and Petra Ornstein, thank you for friendship and support. Myrsini Katsikatsou, Ronnie Pingel and Martin Solberger, I wish you all the best, and good luck with everything. Jenny Eriksson Lundström, thank you for your caring advice.

Katrin Kraus, my dear friend and colleague, working with you has been a privilege. Our discussions have been most rewarding, and your manuscript feedback has always been the most insightful and valuable. In fact, I owe much of this thesis to you. I feel blessed having been able to share this experience with you.

I am also grateful to two great teachers, Tobias Ekholm and Svante Jansson, who have inspired with joy and love of science, and have set a standard of excellence that I will continue try reaching for throughout my career.

I am truly blessed to have a great many dear friends that I appreciate enormously. In terms of recreation, comfort and joy, nothing beats simply hanging out with you. Because of this thesis, I have lately not been able to spend as much time with you as I would like, but knowing I have friends that care and are ready to help out is always greatly reassuring. I value each and every one of you tremendously.

I owe this achievement to my family, including my extended family. You have given me values, work ethic and resilience, and always encouraged and supported my decisions, morally as well as financially. With such a network of unconditional care, love and support, how can you not make it?

Uppsala, April 2009
Joakim Ekström

Bibliography

Juras, J., & Pasaric, Z. (2006). Application of tetrachoric and polychoric correlation coefficients to forecast verification. *Geofizika*, *23*, 59–81.

Martinson, E. O., & Hamdan, M. A. (1971). Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables. *Journal of Statistical Computation and Simulation*, *1*, 45–54.

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

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A*, *195*, 1–47.

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

Ritchie-Scott, A. (1918). The correlation coefficient of a polychoric table. *Biometrika*, *12*, 93–133.

Rényi, A. (1959). On non-parametric measures of dependence for random variables. *Acta Mathematica Academiae Scientiarum Hungaricae*, *10*, 441–451.

Rudin, W. (1976). *Principles of Mathematical Analysis. 2nd ed.* New York: McGraw-Hill.

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

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. *Journal of the Royal Statistical Society*, *75*, 579–652.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 50

Editor: The Dean of the Faculty of Social Sciences

Distribution: publications.uu.se
urn:nbn:se:uu:diva-100735

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2009