INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0118

COPULA-TYPE REPRESENTATION FOR RANDOM COUPLES WITH BERNOULLI MARGINS

A. TAJAR, M. DENUIT and P. LAMBERT

http://www.stat.ucl.ac.be

Copula-type representation for random couples with Bernoulli margins

Abdelouahid Tajar *, Michel Denuit and Philippe Lambert

Institut de statistique
Université catholique de Louvain.
Voie du Roman Pays, 20
B-1348 Louvain-la-Neuve, Belgium

June 14, 2001

Abstract

We propose a copula-type representation for random couples with Bernoulli margins. Some dependence measures for binary data are re-examined. It is stressed that satisfactory dependence measure should only depend on the discrete copula, and not on the margins.

Key words and phrases: Binary Data, Copulas, Fréchet class, Gamma Measure, Kendall's tau, Odds ratio, Somers'd, Spearman's rho.

*Corresponding author.
E-mail address: tajar@stat.ucl.ac.be (Abdelouahid Tajar)

1 Introduction

The modelling of dependence in binary data has recently received considerable attention in the literature. Such data arise in studies where some dichotomous outcome is measured. For example the occurrence or non-occurrence of an insured peril in actuarial science, a binary trait as the presence of a disease in medical science, the results of educational tests in psychology (success/failure) or data on voting behaviour in political science (voting for a party or not for members of the same family). We refer the interested reader e.g. to Ekholm et al. (1995, 2000), Le Cessie and Van Houwelingen (1994), Trégouet et al. (1999), and Frostig (2000) for applications. For a detailed study of binary data we refer the interested reader e.g. to Collett (1991) and Cox and Snell (1989).

One of the most useful tool for handling multivariate distributions with given univariate marginals is the copula function. This concept originated in the late 50's in the context of probabilistic metric spaces. The idea behind is as follows: for multivariate distributions, the univariate marginals and the dependence structure can be separated and the latter may be represented by a copula. For a nice presentation of the copula construction, we refer the interested reader e.g. to Nelsen (1999).

When the marginals are continuous, any question about the dependence structure can be answered with the knowledge of the copula alone. However, as noted by Marshall (1996), things are not so simple with discrete marginals. Problems stem from the fact that copulas are no more unique in this case. As pointed out by Marshall (1996, Example 1.1), there exist copulas which yield the Fréchet upper bound for some marginal pairs and the Fréchet lower bound for other marginal pairs. Because of this, the range of the marginals drastically affects the properties of the joint distribution. Marshall (1996, Proposition 2.3) also proved that there is no nonconstant measure of dependence which depends only on the copula when the marginals are not bounded to be continuous. His reasoning precisely uses Bernoulli marginals.

A list of desirable properties for a measure of association between random variables is proposed by Scarsini (1984). In the continuous case, it is well known that such dependence measures only depend on the underlying copula (and not on the marginals). For instance, it is well known from Shih and Huang (1992) that unless the marginal distributions of two random variables differ only in location and/or scale parameters, the range of Pearson's correlation coefficient is narrower than [-1,1] and depends on the marginals. This can be regarded as a consequence of the fact that Pearson's correlation coefficient cannot be expressed in function of the copula alone. Contrarily to Pearson's coefficient, Kendall's τ and Spearman's ρ are coherent association measures for bivariate distributions with continuous marginals because they only depend on the copula and are not influenced by the marginals (see e.g. Schweizer and Wolff (1981) and Nelsen (1991)).

This paper aims to further discuss dependence between binary random variables. More precisely, we propose a new representation of the dependence structure, inspired from the classical copula construction. This new approach enjoys several nice properties: the representation is unique and carries all the information about the correlation structure. Whereas the copula approach fails to account for all the information about dependence in the discrete case, our approach permits to dissociate the Bernoulli marginals from the dependence structure. Moreover, it will be seen that all the (commonly used) appropriate measures of

$X\backslash Y$	0	1		$U\backslash V$	0	1	
0	p_{00}	p_{01}	$1-p_1$	0	γ_{00}	γ_{01}	$1/2$
1	p_{10}	p_{11}	p_1	1	γ_{10}	γ_{11}	$1/2$
	$1-p_2$	p_2	1		$1/2$	$1/2$	1

Table 2.1: Contingency Tables describing the joint distributions of (X, Y) and of the associated (U, V).

dependence only depend on that copula-type representation.

The paper is organized as follows. Section 2 describes the copula-type construction. The similarities of this construction with the classical Sklar's representation are outlined. In Section 3, we show that the appropriate (in the sense of Scarsini, 1984) measures of dependence for binary couples used in the literature are based on the representation of Section 2. Section 4 is devoted to a discussion and to possible extensions of the present work.

2 Copula-type representation for bivariate binary data

In this section, we propose a new representation of the dependence between binary random variables.

Let X and Y be Bernoulli random variables with respective means p_1 and p_2 and joint distribution given in the left panel of Table 2.1.
We propose to associate to (X, Y) a random couple (U, V) with discrete uniform marginals, i.e.

$$\Pr[U = 0] = \Pr[U = 1] = \Pr[V = 0] = \Pr[V = 1] = \frac{1}{2},$$

and joint distribution given in the right panel of Table 2.1. The link between the original table and its standardized version is given by the formula

$$p_{ij} = \alpha_i \beta_j \gamma_{ij} \quad i, j = 0, 1. \tag{2.1}$$

The role of the factors α_i and β_j involved in (2.1) is to capture the influence of the marginals, while the γ_{ij}'s summarize the type of dependence existing between X and Y, regardless of the values of p_1 and p_2. Note that γ_{11} is the second order dependence ratio introduced by Ekholm *et al.* (2000).

Because of the very simple structure of the marginals, we obviously have

$$\gamma_{01} = 1/2 - \gamma_{00},\ \gamma_{10} = 1/2 - \gamma_{00},\ \gamma_{11} = \gamma_{00} \tag{2.2}$$

so that the standardized contingency table only depends on one parameter, namely γ_{00}. The copula-like representation is now the contingency table displayed in the right panel of table 2.1.

It is easily seen that U and V are independent if, and only if, X and Y are independent. Indeed, if U and V are independent, $\gamma_{ij} = \frac{1}{4}$ for all i and j, and therefore $p_{ij} = \frac{1}{4}\alpha_i\beta_j$,

$i, j = 0, 1$, which yields

$$\frac{p_{00}}{p_{10}} = \frac{p_{01}}{p_{11}} = \frac{\alpha_0}{\alpha_1}$$

so that X and Y are independent.

In order to link the γ_{ij}'s to the p_{ij}'s, the easiest way is to notice that both contingency tables possess the same odd's ratio, that is

$$\psi = \frac{p_{00}p_{11}}{p_{01}p_{10}} = \frac{\gamma_{00}\gamma_{11}}{\gamma_{01}\gamma_{10}}. \tag{2.3}$$

Now, if in (2.3) we replace $\gamma_{11}, \gamma_{01}, \gamma_{10}$ with their corresponding values according to (2.2), we obtain (provided that $\psi \neq 1$) the following second degree polynomial equation in γ_{00}:

$$\gamma_{00}^2(\psi - 1) - \psi\gamma_{00} + \frac{\psi}{4} = 0. \tag{2.4}$$

Equation (2.4) has two roots γ_{00}^+ and γ_{00}^- given by

$$\gamma_{00}^{\pm} = \frac{\psi \pm \sqrt{\psi}}{2(\psi - 1)}$$

Because of the constraints $0 \leq \gamma_{00}^+ < 1/2$ and $0 \leq \gamma_{00}^- < 1/2$, the root γ_{00}^+ is not admissible. On the contrary, the root γ_{00}^- defines a bona fide bivariate distribution with discrete uniform margins. Henceforth, we drop the superscript "-" and later refer to γ_{00}^- by γ_{00}. When $\psi = 1$ (independence), we return to (2.4) and find $\gamma_{00} = 1/4$.

In the continuous case, Sklar's representation associates to each random couple a bivariate distribution with unit uniform marginals called the copula. The construction described above proceeds similarly for bivariate binary data: to each random couple with Bernoulli margins is associated a unique random couple (U, V) with discrete uniform marginals. We will see in the next section that (U, V) is a useful representation of the dependence structure existing between the original X and Y.

3 Dependence measures

Our purpose in this section is to show that any question about the dependence structure of the pair (X, Y) can be answered with the help of (U, V). Let us now review different measures of association for binary data. It will be seen that only those based on γ_{00} are coherent with the Rényi axioms, Rényi (1959) (suitably modified by Scarsini (1984)). There are seven axioms which concern the domain, the symmetry, the coherence, the range, the independence, the change of sign and the continuity. An appropriate measure of concordance should satisfy all these axioms.

3.1 Kendall's tau

Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed random vectors, each with distribution function H. Then the population version of Kendall's τ is defined as the probability of concordance minus the probability of discordance, that is,

$$\tau = \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0] \tag{3.1}$$

In the case of discontinuous joint distribution functions, the definition of concordance is no longer acceptable as stated, because of the presence of ties. In this case, τ is replaced with τ_b (Kendall (1945)) defined as

$$\tau_b = \frac{\Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0]}{\sqrt{\Pr[X_1 \neq X_2]\Pr[Y_1 \neq Y_2]}}.$$

Let us consider two independent random vectors (X_1, Y_1) and (X_2, Y_2) distributed as (X, Y) in Table 2.1. Let us now give the expression of τ and τ_b in this case. We obtain by replacing the probabilities of concordance and discordance by their respective values:

$$\tau = 2p_{00} - 2(1 - p_1)(1 - p_2). \tag{3.2}$$

Hence, Kendall's τ_b is easily obtained from (3.2) and is given by

$$\tau_b = \frac{p_{00} - (1 - p_1)(1 - p_2)}{\sqrt{p_1 p_2 (1 - p_1)(1 - p_2)}}.$$

We note that τ (and hence τ_b) depends on γ_{00} but also on the marginals p_1 and p_2. Consequently, τ is not an appropriate measure of dependence for couples of Bernoulli's.

By virtue of Fréchet inequalities

$$\max\{p_1 + p_2 - 1, 0\} \leq p_{00} \leq \min\{p_1, p_2\}, \tag{3.3}$$

so that we get the inequalities $\tau^{\min} \leq \tau \leq \tau^{\max}$ where

$$\begin{aligned} \tau^{\max} &= 2\min\{1 - p_1, 1 - p_2\} - 2(1 - p_1)(1 - p_2) \\ &= \begin{cases} 2p_2(1 - p_1) & \text{if } \; p_1 \geq p_2 \\ 2p_1(1 - p_2) & \text{if } \; p_1 \leq p_2 \end{cases} \end{aligned}$$

and

$$\begin{aligned} \tau^{\min} &= 2\max\{0, 1 - p_1 - p_2\} - 2(1 - p_1)(1 - p_2) \\ &= \begin{cases} -2p_1 p_2 & \text{if } \; 2 - p_1 - p_2 > 1 \\ -2(1 - p_1)(1 - p_2) & \text{if } \; 2 - p_1 - p_2 \leq 1 \end{cases} \end{aligned}$$

We remark that the extremal values $\tau^{\min}$ and $\tau^{\max}$ depend on marginal probabilities p_1 and p_2. The maximal (minimal) Kendall's τ does not in general attain 1 (-1).

Taking the correction for ties into account yields

$$\tau_b^{\max} = \begin{cases} \sqrt{\frac{p_2(1-p_1)}{p_1(1-p_2)}} & \text{if} \quad p_1 \geq p_2 \\ \sqrt{\frac{p_1(1-p_2)}{p_2(1-p_1)}} & \text{if} \quad p_1 \leq p_2 \end{cases} \tag{3.4}$$

and

$$\tau_b^{\min} = \begin{cases} -\sqrt{\frac{p_1 p_2}{(1-p_1)(1-p_2)}} & \text{if} \quad p_1 + p_2 < 1 \\ -\sqrt{\frac{(1-p_1)(1-p_2)}{p_1 p_2}} & \text{if} \quad p_1 + p_2 \geq 1 \end{cases} \tag{3.5}$$

The maximal value of 1 for Kendall's τ_b is reached if, and only if, $p_1 = p_2 = 0.5$ and the minimal value -1 is attained if, and only if, $p_1 = 1 - p_2$ or $p_1 = p_2 = 0.5$. Bounds on Kendall's τ and Kendall's τ_b are shown on Figures 3.1 and 3.2.

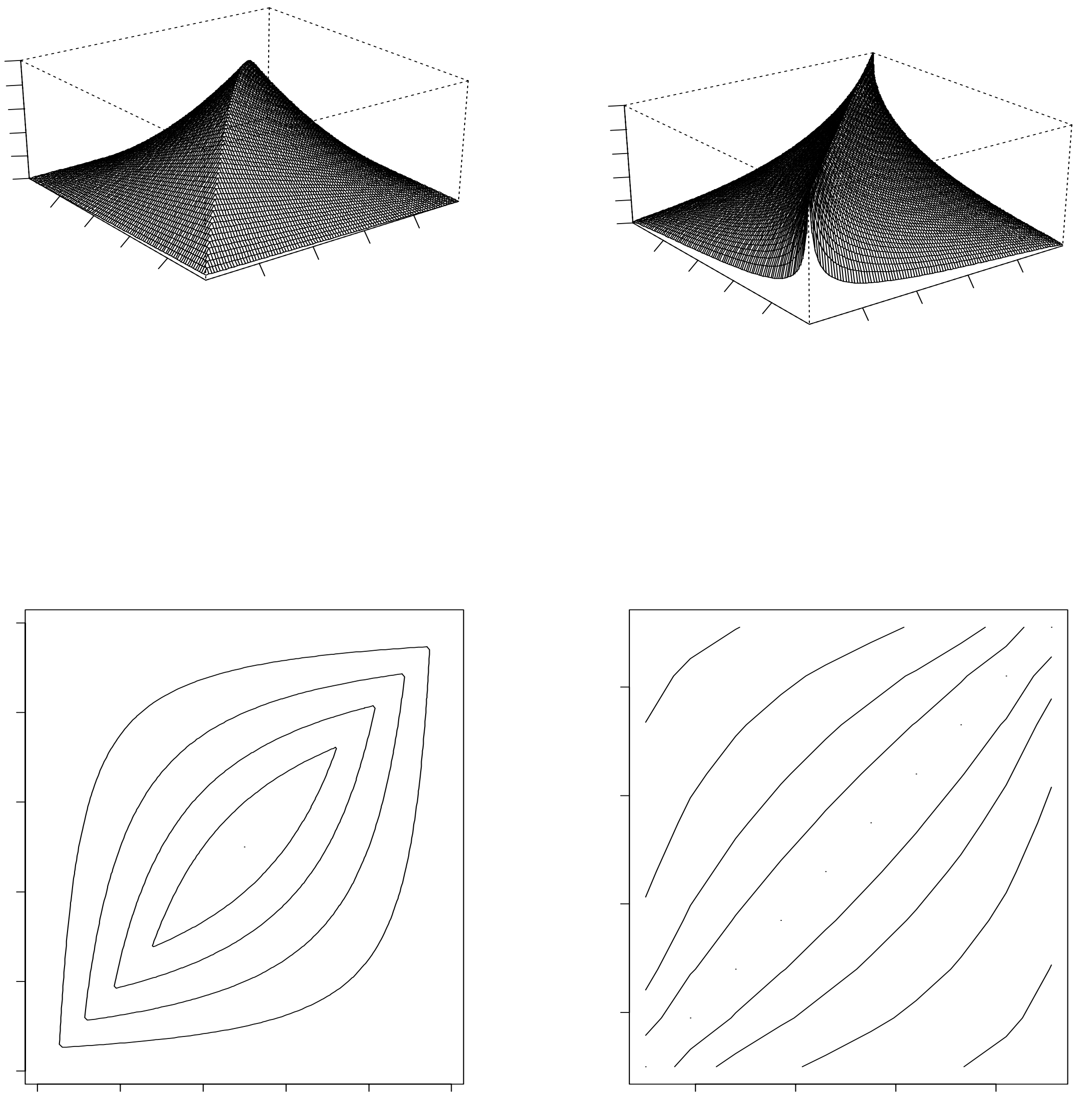

Figure 3.1: *Maximal Kendall's* τ *and Kendall's* τ_b *(top) and their corresponding contour graphs (bottom) for Bernoulli marginals with means* p_1 *and* p_2*.*

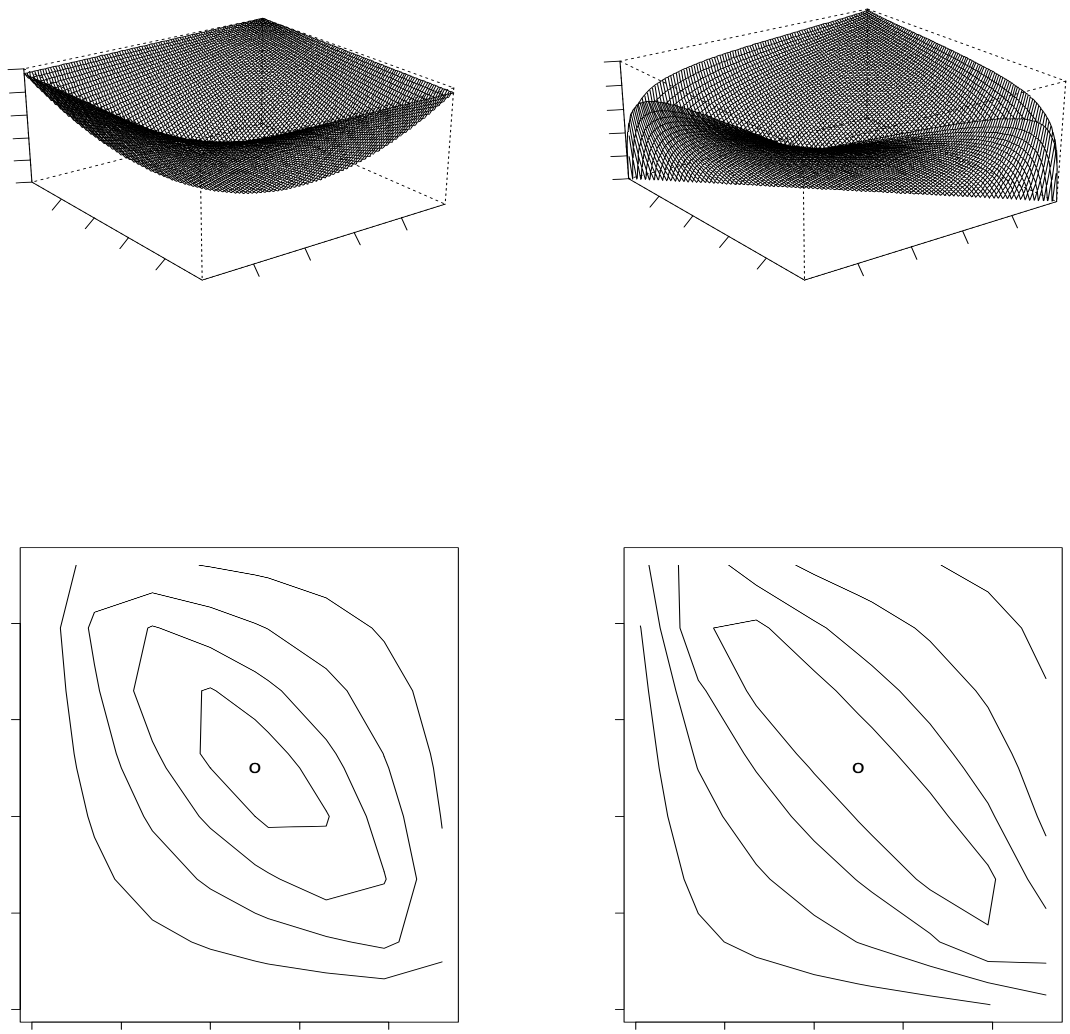

Figure 3.2: *Minimal Kendall's τ and Kendall's τ_b (top) and their corresponding contours graphs (bottom) for Bernoulli marginals with means p_1 and p_2.*

3.2 Spearman's rho

Another measure of association based on concordance and discordance is Spearman's ρ. To obtain the population version of this measure (Kruskal (1958)), we consider $(X_1, Y_1), (X_2, Y_2)$ and (X_3, Y_3), three independent random vectors with common joint distribution function H. The population version of Spearman's ρ is defined to be proportional to the probability of concordance minus the probability of discordance for the two vectors (X_1, Y_1) and (X_2, Y_3), that is,

$$\rho = 3\{\Pr[(X_1 - X_2)(Y_1 - Y_3) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_3) < 0]\} \tag{3.6}$$

Let us now assume that the (X_i, Y_i)'s are distributed as in Table 2.1. The corresponding form for Spearman's ρ is

$$\rho = 3p_{00} - 3(1 - p_1)(1 - p_2) \tag{3.7}$$

By combining Equations (3.2) and (3.7), we conclude that

$$\rho = \frac{3}{2}\tau. \tag{3.8}$$

Like Kendall's τ, Spearman's ρ depends on the marginal probabilities p_1 and p_2. So does its range. Therefore, it is not an appropriate measure of association for binary random couples.

3.3 Covariance, Pearson's linear correlation and Somers' d

Esary *et al.* (1967) consider, as measure of association between binary random variables X and Y, the covariance $\mathbb{C}\text{ov}[X, Y]$ or the linear correlation $r(X, Y)$. For (X, Y) distributed as described in Table 2.1,

$$\mathbb{C}\text{ov}[X, Y] = \frac{\tau}{2} \text{ and } r(X, Y) = \tau_b.$$

Somers (1962) proposed a measure for which pairs untied on X serve as the base rather than only those untied on both X and Y. The population version of Somers's d is

$$d = \frac{\tau}{\Pr[X_1 = X_2]} \tag{3.9}$$

Which becomes for Bernoulli marginals

$$d = \frac{2p_{00} - 2(1 - p_1)(1 - p_2)}{p_1^2 + (1 - p_1)^2}. \tag{3.10}$$

Equation (3.10) show the monotonocity of the measure d with respect to p_{00}. The maximal and minimal values for d are respectively

$$d^{max} = \frac{2\max\{p_1, p_2\}(1 - \min\{p_1, p_2\})}{2p_1(1 - p_1) + 1}$$

and

$$d^{min} = \frac{-2\min\{p_1, 1 - p_2\}\min\{p_2, 1 - p_1\}}{p_1^2 + (1 - p_1)^2}.$$

3.4 Odds ratio

The odds ratio ψ is a widely used measure of association between binary random variables. It is defined by

$$\psi = \frac{p_{00}p_{11}}{p_{10}p_{01}} \tag{3.11}$$

with independence corresponding to $\psi = 1$.

The odds ratio is increasing with p_{00} and with Kendall's τ. Indeed, one can check that ψ can be written as a function of p_{00} and of the marginal probabilities p_1 and p_2 as

$$\psi(p_{00}) = \frac{\{p_{00} + p_1 + p_2 - 1\}p_{00}}{\{(1-p_2) - p_{00}\}\{(1-p_1) - p_{00}\}} \tag{3.12}$$

and as function of τ, p_1 and p_2 as

$$\psi(\tau) = \frac{\{\tau + 2p_1p_2\}\{\tau + 2(1-p_1)(1-p_2)\}}{\{\tau - 2p_1(1-p_2)\}\{\tau - 2p_2(1-p_1)\}} \tag{3.13}$$

which are both non-decreasing. This can be deduced using (3.13) and the fact that τ is bounded between τ^{min} and τ^{max}.

The odds ratio ψ has range $\mathbb{R}^+$: it does not depend on the marginal probabilities. Moreover, from (2.4), we know that ψ only depends on γ_{00} (and thus not on the marginals). This is the only common measure of association considered so far with that property.

3.5 Gamma Measure

This measure of association proposed by Goodman and Kruskal (1954) is equal to the ratio of the difference between the probability of concordance P_c and discordance P_d to the sum of those probabilities. Using the same notation as in Section 3.1, we get

$$\begin{aligned}
\gamma &= \frac{\tau}{\Pr\{(X_1-X_2)(Y_1-Y_2)>0\} + \Pr\{(X_1-X_2)(Y_1-Y_2)<0\}} \\
&= \frac{p_{00}p_{11} - p_{10}p_{01}}{p_{00}p_{11} + p_{10}p_{01}} \\
&= \frac{2p_{00} - 2(1-p_1)(1-p_2)}{p_{00}\{4(p_1+p_2) - 6\} + 4p_{00}^2 + 2(1-p_1)(1-p_2)}
\end{aligned}$$

From the Fréchet bounds in (3.3), we get $\gamma^{\max} = 1$ and $\gamma^{\min} = -1$. One can also easily show the monotonic relationship between γ and the odds ratio ψ

$$\gamma = \frac{\psi - 1}{\psi + 1}.$$

By combining this with the results of Section 3.4, we conclude that γ is a monotonic function of p_{00}. Further, one can easily check that γ, satisfies the axioms proposed by Scarsini (1984) for a measure of concordance to be adequate.

Note also that γ only depends on γ_{00} (and not on the marginal probabilities) as

$$\gamma = \frac{4\gamma_{00} - 1}{8\gamma_{00}^2 - 4\gamma_{00} + 4} \tag{3.14}$$

4 Conclusion

This paper has been devoted to the study of dependence measures for binary data. In Section 2, we proposed a new representation of the dependence structure with γ_{00}, closely related to Sklar's (1959) copula representation: our approach can be regarded as an alternative for discrete margins. Like in the continuous case where the joint density can be expressed as a product of the marginal densities and a quantity measuring the dependence, the representation in (2.1) can be viewed as the equivalent discrete version.

Seven measures were considered, six of them based on concordance , the other one being the odds ratio. Among the commonly used measures, only the odds ratio and gamma were found satisfying. The others have a range which depends on the margins. This is true more generally with discrete margins (see e.g. Vandenhende and Lambert (2000) with ordinal responses). Among other things, the monotonocity of each measure with respect to p_{00} was also studied.

Acknowledgements

The support of the Belgian Government under contract "Project d'Action de Recherche Concertées" ARC 98/03-217 is gratefully acknowledged. The authors particularly thank Professor Jean-Marie Rolin for his valuable comments and suggestions .

References

Collett, D., 1991. Modeling Binary Data. Chapman and Hall, London.

Cox, D.R., Snell, E.J., 1989. Analysis of Binary Data, Chapman and Hall, London.

Ekholm, A., Smith, P.W.F., McDonald, J.W., 1995. Marginal regression analysis of a multivariate binary response, Biometrika, 82, 847-54.

Ekholm, A., McDonald, J.W., Smith, P.W.F., 2000. Association Models for a Multivariate Binary Response, Biometrics 56,712-718.

Esary, J.D., Proschan, F., Walkup, D.W., 1967. Association of Random Variables, with Applications, Ann.Math.Statis. 38, 1466-1474.

Frostig, E., 2000. Comparison of portfolios that depend on multivariate Bernoulli random variables with fixed marginals. Paper presented in Fourth International Congress on Insurance: Mathematics and Economics.Barcelona, Spain.

Goodman, L.A., Kruskal, W.H., 1954, Measures of Association for Cross Classification. J. Amer. Statist. Assoc. 49, 732-764.

Le Cessie S., Van Houwelingen, J.C., 1994. Logistic Regression for Correlated Binary Data, App.Statistics 43, 95-108.

Kruskal, W.H., 1958. Ordinal measures of association, J. Amer. Statist. Assoc. 53, 814-861.

Marshall, A.W., 1996. Copulas, marginals and joint distributions,in:Distributions with Fixed Marginals and Related Topics. L. Rüschendorf, B. Schweizer, and M.D. Taylor, Editors Institute of Mathematical Statistics, Hayward (CA), pp.213-222.

Nelsen, R.B., 1991. Copulas and Association, in: Advances in Probability Distributions with Given Marginals, G.Dall'Aglio,S.Kotz, andG. Salinetti, editors (Kluwer Academic Publishers, Dordrecht),pp.51-74.

Nelsen, R.B., 1999. An Introduction to Copulas, Springer. New York.

Rényi, A., 1959. On measures of concordance, Acta.Math.Acad.Sci.Hungar. 10, 441-451.

Scarsini, M., 1984. On measures of concordance, Stochastica 8, 201-218.

Schweizer, B., Wolff, E.F., 1981. On nonparametric measures of dependence for random variables, Ann. Statist. 9, 870-885.

Sklar, A., 1959. Fonctions de répartitions à n dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris 8, 229-231.

Somers, R.H., 1962. A New asymmetric measure of association for ordinal variables, Amer.Soc.Rev. 27, 799-811.

Trégouet, D.A., Ducimetiere, P., Bocquet, V., Visvikis, S., Soubrier, F., Tiret, L., 1999. A parametric copula model for Analysis of familial binary data, Amer.J.Hum.Genet. 64, 886-893.

Vandenhende, F., Lambert, P., 2000. Modeling repeated ordered categorical data using copulas, DP0025. Institute of Statistics. U.C.L. (Belgium).