*J. Ind. Soc. Agril. Statist.*
*61(2), 2007 : 179-191*

# Copula Functions for Modelling Dependence Structure with Applications in the Analysis of Clinical Data

Pranesh Kumar[1] and Mohamed M. Shoukri[2]
*King Faisal Specialist Hospital and Research Center, MBC #03,*
*P. O. Box 3354, Riyadh 11211, Saudi Arabia*

**SUMMARY**

Since the Pearson's linear correlation coefficient is not a complete and accurate description of dependence structure between variables even when there exists a straight-line relationship between them, copula as an alternative dependence measure is described. Copulas allow modelling linear and non-linear dependence using any choice of marginal distributions. Since many families of copulas are known, copula based approach provides flexibility in modelling and simulating the data. We have illustrated how to compute copula functions and use them to simulate data by considering a clinical trial of epileptic patients suffering from simple or complex partial seizures. A comparison with the correlation based analysis has indicated that the suggested copula based methodology is more appropriate and is capable of modelling the skewed behavior of measurements which correlation model fails to do.

*Key words* : Copula functions, Dependence measures, Archimedean copulas.

## 1. INTRODUCTION

Clinical data may arise primarily from a prospective, retrospective, case-control, clinical or longitudinal study. A particular study may result from a sequence of experiments, each one leading to the next. Possible studies may range from small laboratory experiments to the large and expensive experiments involving humans, to observational studies. Statistical methodologies are helpful in placing interpretations and inferences in their proper context. Appreciation of statistical methodology often leads to the design of study with increased precision and consequently a smaller sample size. Most biomedical or clinical data are multivariate (multifactor). In the multivariate situation, in addition to describing the frequency with which each value of each variable occurs, it is also of interest to study the association and relationship among the risk factors. Pearson's correlation coefficient, non parametric correlations like Kendall's and Spearman's rank correlations and multiple regressions are often applied to study the association and relationship. Embrechts *et al.* (1999) have critically examined that the Pearson's correlation and hence the methodologies based on this measure do not possess the desired properties of a good dependence measure. In particular, this measure fails to describe the tail-end (skewed) behavior of data or the extreme endpoints. In most survival and clinical studies, data distributions are fat-tailed and non-elliptical and thus the usual analyses based on the Pearson's correlation are not appropriate. There are number of ways to discuss and to measure dependence between random variables. Jogdeo (1982) states: "Dependence relation between random variables is one of the most widely studied subject in probability and statistics. The nature of the dependence can take a variety of forms and unless some specific assumptions are made about the dependence, no meaningful statistical model can be contemplated."

To study the dependence structure, Sklar (1959) used copula to describe functions which join together one-dimensional distribution functions to form multivariate distribution functions. Copulas, however, are a recent

[1] *University of Northern British Columbia, Prince George, BC, Canada*
[2] *Schulich School of Medicine, University of Western Ontario, London, Ontario, Canada*

phenomenon. There is no entry in the nine volumes of the encyclopedia of statistical sciences or in the supplement volume. The first update volume published in 1997 does show an entry (Fisher 1997). Only eleven papers mention copulas in the first eighteen volumes (1975-92) of the current index to statistics (CIS) and twenty references in the four recent volumes (1993-96). During 1995-2000, the search of copula in CIS has resulted in 61 hits, 17 being in year 2000 only. It may, however, be noted that the search in CIS does not extend to the working/discussion papers and unpublished research material which are often found in the homepages of researchers. The earliest paper explicitly relating copulas to the study of dependence among random variables is due to Schweizer and Wolff (1981). Copulas appear implicitly in earlier works on dependence, foremost being Hoeffding (1940, 1941). Deheuvels (1979, 1981 a, b, c) used empirical dependence functions (empirical copulas) to estimate the population copula and to construct various nonparametric tests of independence. A large and growing statistical literature on copulas have developed over the past few years [Nelsen (1995, 1999), Fisher (1997), Genest (1987), Schweizer and Sklar (1961), Schweizer (1991), Schweizer and Wolff (1981), Whitt (1976)]. Copulas are of interest to statisticians for two main reasons: Firstly, as a way of studying scale-free measures of dependence; and secondly as a starting point for constructing families of bivariate distributions with a view to simulation (Fisher 1997). Zheng and Klein (1995) proposed a copula-graphic estimator where the dependence between lifetime and censoring variable is described by copula. Rivest and Wells (2001) derived an explicit form for this estimator of the copula as Archimedean. Recently, Kumar and Shoukri (2007 a, b) have shown the advantages of copula based methodology in analyzing the correlated data and Herath and Kumar (2007) discussed the new research directions in engineering economics based on modeling dependence with copulas.

In this paper, we describe copula functions as a means of modelling the dependence measure and illustrate their applications in the analysis of clinical studies by simulating multivariate data. Section 2 describes shortcomings of correlation coefficient as dependence measure, desirable properties of the dependence measures and some parametric and non-parametric dependence measures. Copula functions and the Archimedean copulas are discussed in Section 3. In Section 4, we present an epilepsy trial and analyze patients' seizures data using the copula based methodology. Concluding remarks follow in Section 5.

## 2. DEPENDENCE MEASURES AND PROPERTIES

Let for two real-valued, non-degenerate random variables X and Y with finite variances $\sigma_x^2$ and $\sigma_y^2$ and marginal distributions F(x) and G(y) their joint behavior is described by the joint distribution $H(x, y) = P(X \leq x, Y \leq y)$. Linear correlation or simply correlation (r) between X and Y is only one particular measure of stochastic dependence among many dependence measures. It is the canonical measure in the world of multivariate normal distributions, and for spherical and elliptical distributions. Pitfalls and fallacies associated with correlation arise from the naive assumption that dependence properties of the elliptical world also hold in the non-elliptical world. An excellent paper by Embrechts *et al.* (1999) highlights problems of correlation and discusses alternative dependence measures and simulation algorithms avoiding correlation shortcomings. Correlation is favored by practitioners since for many bivariate distributions it is simple to calculate variances and covariance and hence the correlation coefficient. The generalization of correlation to more than two random variables is straightforward. Correlation and covariance are easy to manipulate under linear operations. However, the shortcomings of correlation are: Variances of X and Y must be finite else the linear correlation is not defined. Independence of two random variables implies they are uncorrelated (linear correlation equal to zero) but zero correlation does not in general imply independence. It is not invariant under nonlinear strictly increasing transformations. Non-parametric correlations often used are the Spearman's rank correlation $\rho$ and Kendall's rank correlation $\tau$. The rank correlation $\rho$ is defined as r(F(x), G(y)) where r is the Pearson's correlation. Let $(X_1, Y_1)$ and $(X_2, Y_2)$ be two independent pairs of random variables from joint distribution of X and Y, then the Kendall's rank correlation

$$\tau = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]$$

Both $\rho$ and $\tau$ measure the degree of monotonic dependence of X and Y, whereas linear correlation r measures the degree of linear dependence only.

We now describe the desired properties of dependence measures. A measure of dependence, like linear correlation, summarizes the dependence structure of two random variables in a single number. Let d(.,.) be a dependence measure which assigns a real number to any pair of real-valued random variables X and Y. Then ideally, we desire a dependence measure to fulfill the following properties :

P1. Symmetry: $d(X, Y) = d(Y, X)$

P2. Normalization: $-1 \le d(X,Y) \le +1$

P3. (i) $d(X,Y) = +1 \Leftrightarrow X, Y$ comonotonic
(ii) $d(X,Y) = -1 \Leftrightarrow X, Y$ countermonotonic

P4. For a transformation $T:\Re \to \Re$ strictly monotonic on the range of X
(i) d(T(X),Y) = d(X, Y), if T increasing
(ii) d(T(X),Y) = –d(X, Y), if T decreasing

These properties could be altered or extended in various ways (Hutchinson and Lai 1990, Chapter 11). Another property we might desire is:

P5. $d(X,Y) = 0 \Leftrightarrow X, Y$ are independent

Unfortunately, this contradicts property P4. There is no dependence measure satisfying P4 and P5. If we require P5, then we can consider dependence measures with the amended properties:

P2b. $0 \le d(X,Y) \le 1$

P3b. $d(X,Y) = 1 \Leftrightarrow X, Y$ comonotonic

P4b. For a $T:\Re \to \Re$ strictly monotonic
d(T(X),Y) = d(X, Y)

Linear correlation (r) fulfills properties P1 and P2 only. Both rank correlations $\rho$ and $\tau$ have the properties P1, P2, P3 and P4. As far as P5 is concerned, the spherical distributions provide examples where pairwise rank correlations are zero, despite the presence of dependence (Embrechts *et al.* 1999).

A measure which satisfies all of P1, P2b, P3b, P4b and P5 (with the exception of the implication d(X, Y) = 1 $\Leftrightarrow$ X, Y comonotonic) is monotone correlation, $d(X,Y) = \sup_{f,g} r[f(X), g(Y)]$, where r represents linear correlation and the supreme is taken over all monotonic functions f and g such that $0 \le \sigma_x^2, \sigma_y^2 < \infty$ (Kimeldorf and Sampson 1989). The disadvantage of all these measures is that they are constrained to give nonnegative values and as such cannot differentiate between positive and negative dependence. It is often not clear how to estimate them. An overview of dependence measures and their statistical estimation is given by Tjostheim (1996). Schweizer and Wolff (1981) used distance criterion for measuring dependence and proposed

$$d_1(X,Y) = 12 \int_0^1 \int_0^1 \left| C(u,v) - uv \right| du\, dv \qquad (2.1)$$

$$d_2(X,Y) = (90 \int_0^1 \int_0^1 |C((u, v) - uv)|^2 du\, dv)^{1/2} \qquad (2.2)$$

$$d_3(X, Y) = 4 \sup_{u,v \in [0,1]} |C(u,v) - uv| \qquad (2.3)$$

where C(u, v) is the joint distribution function of F(x) and G(y) called the copula of random variables X and Y or the bivariate distribution H(x, y) . Next section is devoted to further discussions of copula functions. Copulas are the measures that satisfy amended set of properties including P5 but are constrained to give non-negative measurements and as such cannot differentiate between positive and negative dependence. A further disadvantage of these measures is statistical. Whereas statistical estimation of $\rho$ and $\tau$ from data is straightforward [Gibbons (1988) for the estimators and Tjostheim (1996) for asymptotic estimation theory] it is much less clear how we estimate measures like $d_1(X,Y)$, $d_2(X,Y)$, $d_3(X,Y)$. The following theorem summarizes the properties of $\rho$ and $\tau$ (Embrecht *et al.* 1999).

**Theorem 2.1.** Let X and Y be random variables with continuous distributions F(x) and G(y), joint distribution H(x, y) and copula C(u,v) which is the joint distribution function of F(x) and G(y) where U, V~Uniform (0, 1) then the following hold

(i) $\rho(X, Y) = \rho(Y, X), \ \tau(X, Y) = \tau(Y, X)$

(ii) If X and Y are independent
$\rho(X, Y) = \tau(X, Y) = 0$

(iii) $-1 \le \rho(X, Y), \tau(X, Y) \le +1$

(iv) $\rho = 12 \int_0^1 \int_0^1 [C(u,v) - u,v]\, du\, dv$

(v) $\tau = 4 \int_0^1 \int_0^1 C(u,v)\, dC(u,v) - 1$

(vi) For $T:\Re \to \Re$ strictly monotonic on the range of X, both $\rho$ and $\tau$ satisfy P4.

(vii) $\rho(X,Y) = \tau(X,Y) = 1$

$\Leftrightarrow C_L(x,y) = P(U = x,\ 1 - U = y)$

$\Leftrightarrow Y = T(X)$ a.s. with T increasing

(viii) $\rho(X,Y) = \tau(X,Y) = -1$

$\Leftrightarrow C_U(x,y) = P(U = x, U = y)$

$\Leftrightarrow Y = T(X)$ a.s. with T decreasing

The lower bound $C_L(.,.)$ and upper bound $C_U(.,.)$ are bivariate distribution functions of the vectors (U, 1 – U) and (U, U), where U~Uniform (0,1). $C_L(.,.)$ and $C_U(.,.)$ describe perfect positive and perfect negative dependence respectively. These bounds are well known as the Fréchet (1951)-Hoeffding (1940, 1941) copula boundaries are as follows.

**Minimum copula:** The lower bound for all copulas. In the bivariate case, it represents perfect negative dependence and is

$$C_L(u,v) = \max(0, u + v - 1) \tag{2.4}$$

**Maximum copula:** This is the upper bound for all copulas. It represents perfect positive dependence and is

$$C_U(u,v) = \min(u,v) \tag{2.5}$$

For all copulas C(u, v)

$$\max(0, u + v - 1) \le C(u,v) \le \min(u,v) \tag{2.6}$$

## 3. COPULA FUNCTIONS

Copulas are functions that join or couple multivariate distribution functions to their one-dimensional marginal distribution functions. Alternatively, copulas are multivariate distributions whose one-dimensional margins are uniform on the interval [0, 1]. For two random variables X and Y with respective marginal distributions F(x) and G(y), their joint behavior is described by joint distribution H(x,y). Then, the joint distribution for every $(u, v) \in [0,1]^2$ can be expressed by copula function as

$$\begin{aligned} C(u,v) &= P[F(x) \le u, G(y) \le v] \\ &= P\left(X \le F^{-1}(u), Y \le G^{-1}(v)\right) \\ &= H\left[F^{-1}(u), G^{-1}(v)\right] \end{aligned} \tag{3.1}$$

where $F^{-1}(u)$ and $G^{-1}(v)$ are the quantile functions. Some copula properties:

(i) $C(u,0) = C(0,v) = 0$

(ii) $C(u,1) = C(1,v) = v$

If F(x) and G(y) are continuous then C(u, v) is unique. An important feature of copulas is that any choice of marginal distributions can be used. Hence, copulas are constructed based on the assumption that marginal distribution functions are known. The two standard non-parametric rank correlations, Kendall's $\tau$ and Spearman's $\rho$ are expressed in copula form as

$$\tau = 4 \iint_{I^2} C(u,v)\, dC(u,v) - 1 \tag{3.2}$$

$$\rho = 12 \iint_{I^2} C(u,v)\, du\, dv - 3 \tag{3.3}$$

The explicit expressions for $\tau$ and $\rho$ for Archimedean copulas considered in this paper are presented in Subsection 3.3.

### 3.1 How to Construct Copula Functions?

There are several methods of constructing copulas or specifying families of copulas. We consider the following:

#### 3.1.1 Inversion method

This method for constructing bivariate copula is based on the Sklar (1959) theorem to construct copulas directly from the joint distribution functions. To illustrate the construction, we adapt an example from Nelsen (1999). Let X and Y be random variables with joint distribution function

$$H(x,y)=\begin{cases}\dfrac{(x+1)(e^{y}-1)}{x+2e^{y}-1} & (x,y)\in[-1,\ 1]\times[0,\infty] \\ 1-e^{-y} & (x,y)\in(1,\ \infty]\times[0,\infty] \\ 0 & \text{elsewhere}\end{cases} \tag{3.4}$$

The marginal distribution functions F(x) and G(y) are given by

$$F(x)=\begin{cases}0 & x<-1 \\ \dfrac{x+1}{2} & x\in[-1,\ 1] \\ 1 & x>1\end{cases} \tag{3.5}$$

$$G(y)=\begin{cases}0 & y<0 \\ 1-e^{-y} & y\geq 0\end{cases} \tag{3.6}$$

Let $u=\dfrac{x+1}{2}$ and $v=1-e^{-y}$, then the inverse of F and G are $F^{-1}(u)=2u-1$ and $G^{-1}(v)=-\ln(1-v)$ for $u, v \in [0, 1]$. The copula function C(u, v) is

$$C(u,v)=H(F^{-1}(u),G^{-1}(v))=\frac{uv}{u+v-uv} \tag{3.7}$$

In the inversion method one has to know the joint density function. See Nelsen (1999) and Frees and Valdez (1998) for other limitations.

### 3.1.2 Archimedean approach

Archimedean approach is a general method of constructing both bivariate and multivariate copulas. These copulas are called Archimedean copulas and arise from mathematical theory of associativity. Archimedean copulas are an important class of copulas which are easier to construct. They possess nice properties and many known copula families belong to this class (Nelson 1999).

Let $\varphi$ be a continuous decreasing function from $I=[0,1]$ to $[0,\infty]$ such that $\varphi(1)=0$ and $\varphi^{-1}$ its inverse given by

$$\varphi^{-1}(t)=\begin{cases}\varphi^{-1}(t) & 0\leq t\leq\varphi(0) \\ 0 & \varphi(0)\leq t\leq\infty\end{cases} \tag{3.8}$$

Then the function

$$C\ (u,v)=\varphi^{-1}(\varphi(u)+\varphi(v)) \tag{3.9}$$

for $u, v\in[0,1]$ is a copula. A copula of the form (3.9) is called an Archimedean copula. The function $\varphi$ is called a generator of the copula C(u, v). In order to construct Archimedean copulas using Equation (3.9) we need to find functions which serve as generators, i.e., continuous decreasing functions from 1 to $[0,\infty]$ such that $\varphi(1)=0$. Different choices of generator function give different families of copulas. To illustrate this method, we use the same example as considered above. Consider the generator function as $\varphi(t)=\dfrac{1-t}{t}$. Then the inverse function is $\varphi^{-1}(t)=\dfrac{1}{t+1}$. Substituting in equation (3.9), we have Archimedean copula

$$C\ (u,v)\ =\varphi^{-1}(\varphi(u)+\varphi(v))$$

$$=\varphi^{-1}\left(\frac{1-u}{u}+\frac{1-v}{v}\right)\ =\ \frac{uv}{u-uv+v} \tag{3.10}$$

### 3.1.3 Compound method

The Marshall and Olkin (1988) suggested the method for construction of copulas which involves the Laplace transform and its inverse function. Let $\psi(t)$ denote the Laplace transform of a positive random variable $\gamma$ then

$$\psi(t)=E_{\gamma}(e^{-t\gamma})=\int_{0}^{\infty}e^{-t\theta}\ dF_{\gamma}(\theta) \tag{3.11}$$

where $F_{\gamma}(\theta)$ is the distribution function of $\gamma$. Notice that $\psi(-t)$ is the moment generating function of $F_{\gamma}$. Let X and Y be random variables whose conditional distributions given positive latent variables $\gamma_X$ and $\gamma_y$ are specified by $H_x(x\,|\,\gamma_x)=H_x(x)^{\gamma_x}$ and $H_y(y\,|\,\gamma_y)=H_y(y)^{\gamma_y}$ where $H_x$ and $H_y$ are baseline distribution functions. Consider a bivariate distribution function of the form

$$H(x,y)=E[K(H_x(x)^{\gamma_x},H_y(y)^{\gamma_y})] \tag{3.12}$$

where K(.,.) is a distribution function with uniform marginals and the expectation is over $\gamma_x$ and $\gamma_y$. As a special case consider both the latent variables equal i.e.

$\gamma_x = \gamma_y = \gamma$ and use distribution functions corresponding to independent marginals. Marshall and Olkin (1988) showed that

$$H(x, y) = E[(H_x(x)^{\gamma_x} . H_y(y)^{\gamma_y})]$$

$$= E[(H_x(x)^{\gamma} . H_y(y)^{\gamma})]$$

$$= \psi[(\psi^{-1}(F(x)) + \psi^{-1}(G(y))] \qquad (3.13)$$

where F(.) and G(.) are marginal distribution functions and $\psi(.)$ is the Laplace transform of $\gamma$. We consider an example (Marshall and Olkin 1988 and Joe 1993) in which convex sums lead to copulas constructed from Laplace transforms of distribution functions. Let H(u, v) be a convex sum (or mixture) of powers of distribution function. Set

$$H(u,v) = \int_0^\infty F^\theta(u)\, G^\theta(v)\, d\Lambda(\theta) \qquad (3.14)$$

and assume that $\Lambda(\theta) = 0$. Further let $F(u) = \exp[-\varphi^{-1}(u)]$ and $G(v) = \exp[-\varphi^{-1}(v)]$. Then from (3.13)

$$H(u, v) = \int_0^\infty \exp[-\theta(\varphi^{-1}(u) + \varphi^{-1}(v)]d\Lambda(\theta)$$

$$= \varphi[\varphi^{-1}(u) + \varphi^{-1}(v)] \qquad (3.15)$$

Thus H(u,v) is a copula C(u,v). Notice that the right hand side represents the broader class of Archimedean copulas. Laplace transform of a distribution function $\psi$ have well defined inverse functions. As seen from (3.9) and (3.14) the inverse function $\varphi^{-1}(u)$ serves as a generator for Archimedean copulas.

Once the copulas are known then as a consequence of Sklar (1959) theorem, we obtain bivariate or multivariate distributions with whatever marginal distributions we want. This is an elegant property that can be exploited. There are several ways to generate observations (x, y) of a pair of random variables (X,Y) whose joint distribution function is H(x, y). Copulas, however, express the non-parametric nature of dependence between two random variables and as such are a powerful tool for modelling dependencies among random variables.

### 3.2 Archimedean Copulas

We consider three one parameter ($\theta$) Archimedean copulas namely Frank copula (1979), Clayton copula (1978) and Gumbel copula (1960). Nelsen (1999, p. 94-97) tabulates one parameter families of the Archimedean copulas. Archimedean copulas are easy to apply and have nice properties. The parameter $\theta$ in each case measures the degree of dependence and controls the association between the two variables. When $\theta \to 0$, there is no dependence and if $\theta \to \infty$, there is perfect dependence. Schweizer and Wolff (1981) showed that the dependence parameter $\theta$ which characterizes each family of Archimedean copulas can be related to Kendall's $\tau$. This property can be used to empirically determine the applicable copula form.

#### 3.2.1 Frank copula

(a) Generator

$$\varphi(t) = -\ln \frac{e^{-\theta t} - 1}{e^{-\theta} - 1}$$

(b) Bivariate Copula

$$C(u,v) = -\frac{1}{\theta} \ln\left(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right)$$

(c) Laplace Transform

$$\varphi(t) = \varphi^{-1}(t) = \theta^{-1} \ln\left[1 + e^t(e^\theta - 1)\right]$$

(d) Kendall's $\tau$

$$\tau = 1 - \frac{4}{\theta}[1 - D_1(\theta)]$$

where $D_k(x)$ is the Debye function for any positive integer k, given by $D_k(x) = \int_0^x \frac{t^k}{e^t - 1}\, dt$

#### 3.2.2 Clayton copula

(a) Generator

$$\varphi(t) = (t^{-\theta} - 1)$$

(b) Bivariate Copula

$$C(u,v) = \left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{1}{\theta}}$$

(c) Laplace Transform

$$\varphi(t) = \varphi^{-1}(t) = (1-t)^{-\frac{1}{\theta}}$$

(d) Kendall's $\tau$

$$\tau = \frac{\theta}{\theta+2}$$

### 3.2.3 Gumbel copula

(a) Generator

$$\varphi(t) = (-\ln(t))^{\theta}$$

(b) Bivariate Copula

$$C(u,v) = \exp\left\{-\left[(-\ln u)^{\theta} + (-\ln v)^{\theta}\right]^{\frac{1}{\theta}}\right\}$$

(c) Laplace Transform

$$\varphi(t) = \varphi^{-1}(t) = \exp(-t^{1/\theta})$$

(d) Kendall's $\tau$

$$\tau = \frac{\theta-1}{\theta}$$

### 3.3 Normal Copula

Normal (Gaussian) copula is constructed from the bivariate normal distribution using the Sklar's theorem. For the random variables X and Y which are distributed as standard bivariate normal with correlation r, the normal copula function is

$$C(u,v) = \Phi\left(\Phi^{-1}(u), \Phi^{-1}(v)\right) \qquad (3.16)$$

where the marginals U and V are N(0,1) distributions and $\Phi$ denotes the cumulative normal probability distribution.

### 3.4 Choosing Right Copula

The first step in modelling and simulation is identifying the appropriate copula form. The procedure (Genest and Rivest 1993) involves verifying how close different copulas fit the data by comparing the closeness of the copula with the empirical copula. The steps follow

**Step 1**: Estimate the Kendall's $\tau$ from the data by

$$\tau = \binom{n}{2}^{-1} \sum_{i<j}^{n} \text{Sign}\left[(x_i - x_j)(y_i - y_j)\right]$$

**Step 2**: Construct an empirical copula function as follows

(i) Determine the pseudo observations

$T_i = \{\text{Number of } (x_j < x_i) \text{ such that } x_j \le x_i$ and $y_j \le y_i\} / (n-1)$

(ii) The empirical copula

$K_E(t) = \text{proportion of } T_i\text{'s} \le t, \quad 0 \le t \le 1$

**Step 3:** Construct the Archimedean copula function

$$K_C(t) = t - \frac{\varphi(t)}{\varphi'(t)},$$ where $\varphi'(t)$ is the first derivative of $\varphi(t)$.

In order to select the Archimedean copula which best fits the application data, we choose that copula which minimizes the non-parametric distance measure

$$DM := \int \left[K_C(t) - K_E(t)\right]^2 dK_E(t) \qquad (3.17)$$

We derive the expressions of for the above described three Archimedean copulas in Table 1.

**Table 1.** Archimedean copula functions

| Copula | $K_C(t)$ |
|---|---|
| Frank | $\dfrac{\theta t - [1-\exp(\theta t)]\ln\left[\dfrac{\exp(-\theta t)-1}{\exp(-\theta)}\right]}{\theta}$ |
| Clayton | $\dfrac{t\,(1+\theta-t^{\theta})}{\theta}$ |
| Gumbel | $\dfrac{t\,(\theta-\ln t)}{\theta}$ |

### 3.4 Tail-Dependence and Copula Functions

Tail-Dependence refers to the degree of dependence in the corner of the lower-left quadrant or upper-right quadrant of a bivariate distribution. It describes the limiting proportion that one margin exceeds a certain threshold given that the other margin has already exceeded that threshold. For two random variables X and Y with marginal distributions F(x) and F(y), the upper tail-dependence is defined as

$$\lambda_{upper} = \lim_{u \to 1} \Pr[Y \geq F_Y^{-1}(u) \mid X \geq F_X^{-1}(u)] \qquad (3.18)$$

and the lower tail dependence

$$\lambda_{lower} = \lim_{u \to 0} \Pr[Y \leq F_Y^{-1}(u) \mid X \leq F_X^{-1}(u)] \qquad (3.19)$$

provided limit exists where $F^{-1}(.)$ is the inverse distribution function and U is a uniform random variable defined over (0,1). A distribution is upper tail dependent if $\lambda_{upper} > 0$ and upper tail independent if $\lambda_{upper} = 0$. Similarly we interpret $\lambda_{lower}$.

The following representation shows that tail dependence is a copula property. An equivalent definition (for continuous random variables) of tail dependence in terms of a bivariate copula function C(u, v) is

$$\lambda_{upper} = \lim_{u \to 1} \frac{1 - 2u + C(u,u)}{1-u} \qquad (3.20)$$

and

$$\lambda_{lower} = \lim_{u \to 0} \frac{C(u,u)}{u} \qquad (3.21)$$

## 4. ILLUSTRATION: EPILEPSY TRIAL

For illustration we consider data from a clinical trial of 59 epileptics (Leppik *et al.* 1985). Patients suffering from simple or complex partial seizures were randomized to receive either the anti-epileptic drug progabide or a placebo as an adjuvant to standard chemotherapy. At each of the four successive post randomization clinic visits, the number of seizures occurring over the previous two weeks was recorded. Each patient subsequently was crossed over to the other treatment. For the purpose of illustration, we analyze data on 30 patients from the treatment group only. Let the random variable X denote the number of seizures during base period and Y the number of seizures end of the treatment. We apply the copula simulation to generate data sets and obtain the confidence intervals for difference in mean number of pre- and post- treatment seizures to conclude that the treatment was effective in reducing the number of seizures. The data and summary statistics are given in

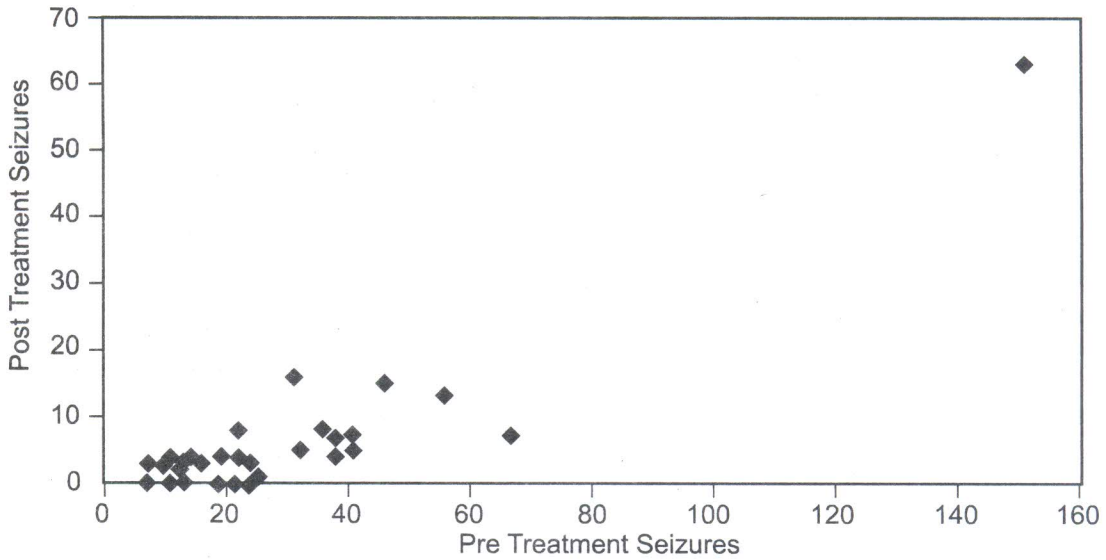


**Fig 1.** Scatter plot of pre-and post-treatment seizures from 30 patients

**Table 2.** Data on the number of pre (X)- and post (Y)-treatment seizures and summary statistics

| Patient | Pre-treatment (X) | Post-Treatment (Y) |
|---|---|---|
| 1 | 76 | 8 |
| 2 | 38 | 4 |
| 3 | 19 | 0 |
| 4 | 10 | 3 |
| 5 | 19 | 4 |
| 6 | 24 | 3 |
| 7 | 31 | 16 |
| 8 | 14 | 4 |
| 9 | 11 | 4 |
| 10 | 67 | 7 |
| 11 | 41 | 5 |
| 12 | 7 | 0 |
| 13 | 22 | 0 |
| 14 | 13 | 3 |
| 15 | 46 | 15 |
| 16 | 36 | 8 |
| 17 | 38 | 7 |
| 18 | 7 | 3 |
| 19 | 36 | 8 |
| 20 | 11 | 0 |
| 21 | 22 | 4 |
| 22 | 41 | 7 |
| 23 | 32 | 5 |
| 24 | 56 | 13 |
| 25 | 24 | 0 |
| 26 | 16 | 3 |
| 27 | 22 | 8 |
| 28 | 25 | 1 |
| 29 | 13 | 0 |
| 30 | 12 | 2 |
| Mean | 27.63 | 4.83 |
| Standard Error | 3.17 | 0.78 |
| Skewness | 1.15 | 1.15 |
| Standard Error | 0.43 | 0.43 |
| Kurtosis | 1.11 | 1.00 |
| Standard Error | 0.83 | 0.83 |
| Pearson Correlation | 0.616 | |
| p-value | 0.0001 | |
| Kendall Rank Correlation | 0.495 | |
| p-value | 0.0001 | |

Table 2. A scatter plot of the pre- and post-treatment seizure counts is shown in Fig. 1. The descriptive analysis in Table 2 indicates that the distributions of both the pre- and post-treatment seizures are not symmetrical (skewness coefficients being 1.147 and 1.107, respectively), cluster more and have longer tails (kurtosis coefficients being 1.108 and 1.402, respectively). From

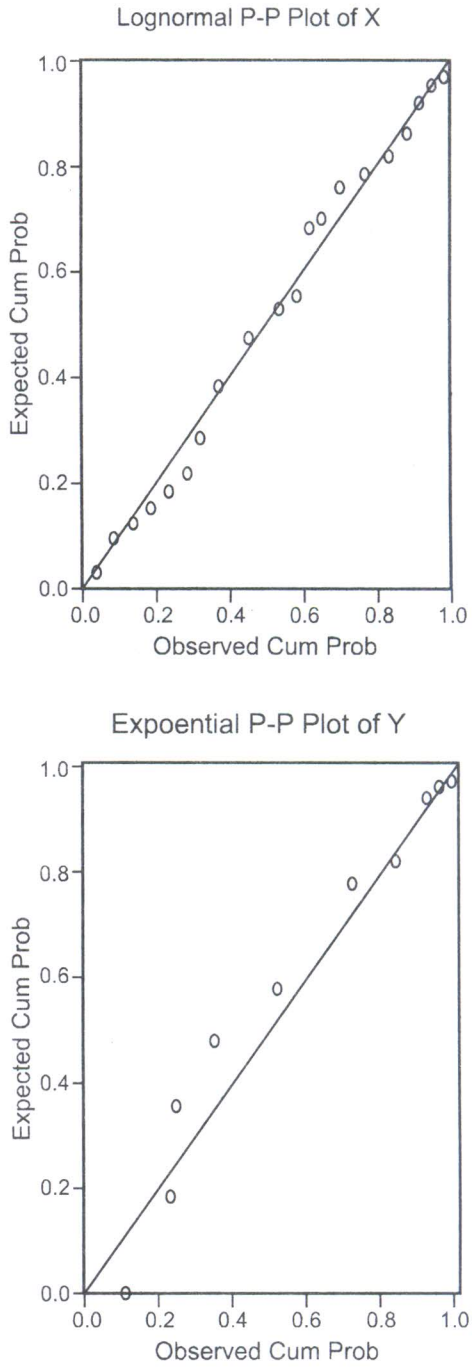


**Fig. 2.** Probability plots of pre (X)-and post (Y)-treatment seizures

the probability plots of X and Y in Fig. 2, we estimate that X ~ lognormal (22.88, 0.637) and Y ~ exponential (0.207). The conventional measure of dependence between X and Y, the correlation coefficient r = 0.616 ($p < 0.0001$) and non-parametric Kendall's rank correlation $\tau = 0.495$ ($p < 0.0001$). For making comparisons, we assume that the distributions X and Y are normal. The estimates (M.) and 95% confidence intervals (CI) of mean of X and Y and their difference are

$M_X = 27.63$ ($p < 0.0001$; 95%CI = 21.14, 37.12)

$M_Y = 4.83$ ($p < 0.0001$; 95%CI = 3.23, 6.43)

$D := M_{X-Y} = 22.8$ ($p < 0.0001$; 95%CI = 16.14, 29.45)

### 4.1 Choosing the Best Copula

Three copulas of the Archimedean family, Gumbel, Clayton and Frank copulas, and empirical copula are estimated from the data. The estimated copula parameters are given in Table 3 and copulas are plotted in Fig. 3. The non parametric distance measure $DM := \int [K_C(t) - K_E(t)]^2 dK_E(t)$ for the Gumbel, Clayton and Frank copulas are respectively 0.143, 0.233 and 0.150. According to the minimum distance criterion for choosing the copula, we find that the minimum distance 0.143 is for the Gumbel copula implying that the Gumbel copula is the best fit to the given data.

**Table 3.** Estimated copula parameters and distance measure

| | Gumbel | Clayton | Frank |
|---|---|---|---|
| $\tau$ | 0.495 | 0.4950 | 0.495 |
| $\theta$ | 1.9957 | 1.6924 | 5.0276 |
| $DM := \int [K_C(t) - K_E(t)]^2 dK_E(t)$ | 0.143 | 0.233 | 0.150 |

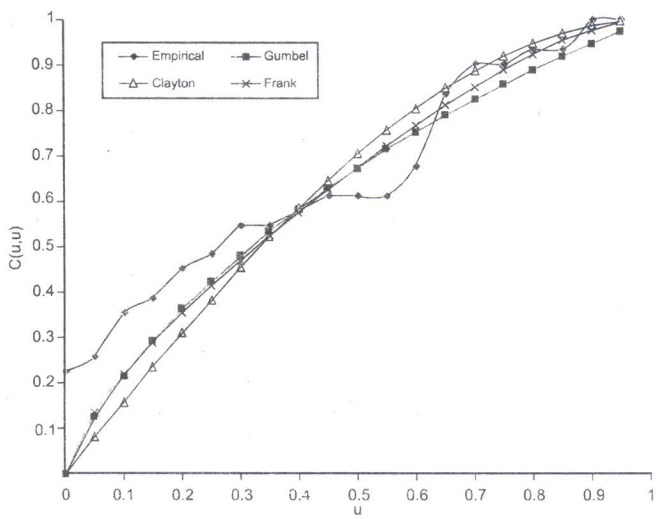


**Fig. 3.** Which copula is the right one?

### 4.2 Tail Dependence

We use the Gumbel and normal copulas to study tail dependence. It is seen from Fig. 4a that when $u \to 0$, the lower tail dependence tends to zero for both – normal and Gumbel copulas. When $u \to 1$, Fig. 4b shows that the upper tail dependence for the normal copula tends to zero, however, it remains almost constant for the Gumbel copula.

The tail dependence analysis thus indicates that the Gumbel copula have upper tail dependence but does not have the lower tail dependence whereas the normal copula have neither. Therefore, the conventional statistical

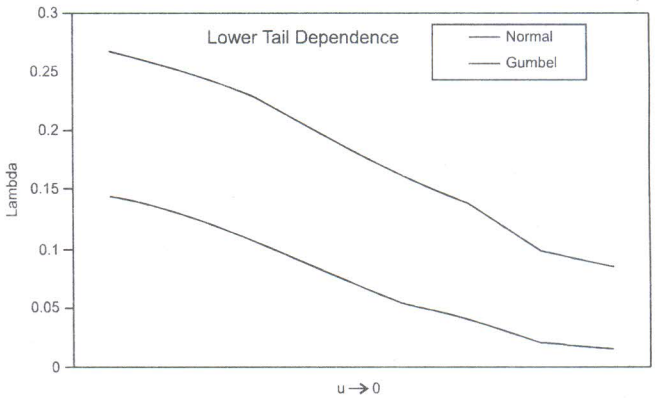


**Fig. 4a.** Lower Tail Dependence $\lambda_{lower}$

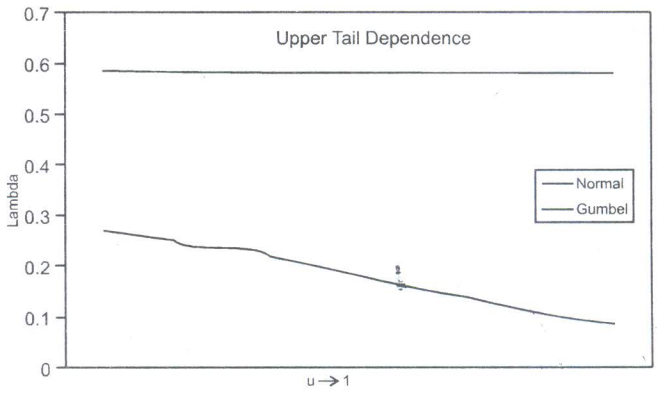


**Fig. 4b.** Upper Tail Dependence $\lambda_{upper}$

analysis based on the normality assumption and correlation as the measure of dependence is not appropriate. A Gumbel copula based analysis which models the tail dependence as well, is a right choice for the analysis.

Further, a question may occur whether one needs to carry out exercise, as done in Subsections 4.1 and 4.2, every time one has a data set and then decide on which copula to choose? Or, there is a better way? As an answer, it may be noted that one doesn't need to carry out calculations as done in both Subsections 4.1 and 4.2 to decide on which copula is appropriate for a given data set. What is required is to calculate distance measure DM for copulas and then to select one which has the minimum value of DM, as shown in Subsection 4.1. Calculations of measures of tail dependence in Subsection 4.2 additionally give an insight into the direction of dependence whether dependence is in right or left tail. There is no other better quantitative way right now to choose the best copula. This is a point worth further investigation.

### 4.3 Simulation

We carry out 50 and 100 Monte Carlo simulations of size 30, 50 and 75 using the Gumbel copula with the estimated marginal distributions X ~ lognormal (22.88,0.637) and Y ~ exponential (0.207). The VBA codes (Melchiori 2003) are used for executing the algorithm. The point and interval estimates of mean and of difference in mean numbers of the pre-and post-treatment seizures are presented in Table 4. The following indicative conclusions from the results may be noted:

(i) The null hypothesis of no difference, i.e. $H_0 : D = M_{X-Y} = 0$, is rejected on the basis of all the estimated confidence intervals.

(ii) For fixed sample size, increase in the number of simulations affects no significant changes in the point and interval estimates and consequently the width of the confidence intervals. Therefore, we report results from 50 and 100 simulations only.

(iii) For fixed number of simulations, increase in sample size (n = 30, 50, 75) like (ii) above brings no significant changes in the estimates.

(iv) Thus, there is strong evidence that the anti-epileptic drug progabide has been effective in reducing the number of seizures in the epileptic patients.

It is interesting to note that the copula based analysis performed better than the correlation based analysis since the width of the confidence intervals from copula simulation was smaller than the width of the confidence interval from the conventional method. Another noteworthy observation is that it is not necessary to carry out a large number of simulations when using the copula based methodology. Since the estimates are consistent and not affected by the sample size, copula based simulation provides a robust alternative for the analysis.

## 5. CONCLUDING REMARKS

We have emphasized that in modelling dependency, the Pearson's linear correlation coefficient is not a complete and accurate description of dependence structure between variables even when there exists a straight-line relationship between them. An alternative is to model the dependence structure using copulas that overcome the limitations of correlation. Copulas allow modelling linear and non-linear dependence. Using copulas any choice of marginal distribution functions can be used and extreme endpoints can be modeled too. We have illustrated how to compute copula functions and simulate data using a clinical trial of epileptic patients suffering from simple or complex partial seizures. A comparison with the conventional correlation based analysis has indicated that the suggested copula based methodology is more appropriate and is capable to capture and model the skewed behavior of the measurements which correlation model fails to do. Since many families of copulas are known to exist in literature, copula based approach provides flexibility in modelling various categories of clinical studies and simulating the data. In clinical trials and experiments, sample size is often an important consideration and is relatively small for scarce experimental units. Copula based analysis overcomes this limitation as well, because as described above, simulation algorithm can be applied to replicate any number of patients data. The copula simulation methodology discussed in this paper is simple and easy to implement in any computing environment.

**Table 4.** Point estimates and the 95% confidence intervals of mean number of pre (X)- and post(Y)- treatment seizures and of mean difference

| Sample Size | 30 | 50 | 75 | 30 | 50 | 75 |
|---|---|---|---|---|---|---|
| Simulations | 50 | 50 | 50 | 100 | 100 | 100 |
| $M_X$ | | | | | | |
| Mean | 29.49 | 29.47 | 29.59 | 29.52 | 29.62 | 29.63 |
| Standard Deviation | 6.74 | 6.68 | 6.84 | 6.78 | 6.86 | 7.00 |
| 95% Lower Confidence Limit | 29.22 | 29.21 | 29.32 | 29.38 | 29.48 | 29.49 |
| 95% Upper Confidence Limit | 29.76 | 29.74 | 29.86 | 29.65 | 29.76 | 29.77 |
| Width of Confidence Interval | 0.54 | 0.53 | 0.54 | 0.27 | 0.28 | 0.28 |
| $M_Y$ | | | | | | |
| Mean | 5.28 | 5.21 | 5.27 | 5.24 | 5.18 | 5.25 |
| Standard Deviation | 2.19 | 2.05 | 2.22 | 2.15 | 2.05 | 2.18 |
| 95% Lower Confidence Limit | 5.19 | 5.12 | 5.18 | 5.20 | 5.14 | 5.21 |
| 95% Upper Confidence Limit | 5.37 | 5.29 | 5.36 | 5.28 | 5.22 | 5.30 |
| Width of Confidence Interval | 0.18 | 0.17 | 0.18 | 0.08 | 0.08 | 0.09 |
| $M_{X-Y}$ | | | | | | |
| Mean | 24.18 | 22.73 | 24.32 | 24.26 | 23.67 | 24.38 |
| Standard Deviation | 8.24 | 9.15 | 8.32 | 8.22 | 8.78 | 8.44 |
| 95% Lower Confidence Limit | 21.18 | 20.23 | 22.42 | 21.31 | 21.25 | 22.45 |
| 95% Upper Confidence Limit | 27.19 | 25.23 | 26.22 | 27.21 | 26.08 | 26.31 |
| Width of Confidence Interval | 6.01 | 5.00 | 3.80 | 5.90 | 4.83 | 3.86 |

## ACKNOWLEDGEMENTS

Authors thank the referee for his valuable comments. They also acknowledge the Research Centre, KFSH&RC for sponsoring the research project # 2060 022. First author thanks Prof. Sudhir Gupta, Northern Illinois University, USA and Dr. Rajender Prasad, IASRI, New Delhi for inviting to present paper under the theme: Emerging Issues in Areas of Basic Statistical Research in the International Conference on Statistics and Informatics in Agricultural Research at IASRI, New Delhi, 27-30 December, 2006.

## REFERENCES

Clayton, D.G. (1978). A model for association in Bivariate Life Tables and its applications in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141-151.

Deheuvels, P. (1979). La fonction de dépendence empirique et ses propriétés. Un test non paramétrique d'indépendence, *Acad. Roy. Belg. Bull. Cl. Sci.*, **65(5)**, 274-292 .

Deheuvels, P. (1981a). A Kolmogorov-Smirnov type test for independence and multivariate samples. *Rev. Roumaine Math. Pures Appl.*, **26**, 213-226 .

Deheuvels, P. (1981a). A non parametric test for independence. *Publ. Inst. Statist.*, University of Paris, **26**, 29-50.

Deheuvels, P. (1981c). Multivariate tests of independence. *Analytical Methods in Probability* (Oberwolfach, 1980), *Lecture Notes in Mathematics*, **861**, Spring-Verlag, Berlin, 42-50 .

Embrechts, P., Mcneil, A.J. and Straumann, D. (1999). Correlation and dependence in risk management: Properties and pitfalls. In: *Risk Management: Value at Risk and Beyond,* (ed.) M. Dempster and H.K. Moffatt, Cambridge University Press.

Fisher, N.I. (1997). Copulas. In : *Encyclopedia of Statistical Sciences*, Updated Vol. 1, S. Kotz, C.B. Read, D.L. Banks, (eds.), John Wiley and Sons, New York, 159-163 .

Frank, M.J. (1979). On the simultaneous associativity of F(x, y) and x + y – F(x, y). *Aequationes Math.* , **19**, 194-226.

Fréchet, M. (1951). Sur les tableaux de corrélation don't les marges sont données. *Ann. Univ. Lyon,* **A9**, 53-77.

Frees, E.W. and Valdez, E. (1998) Understanding relationships using copulas. *North American Actuarial Journal*, **2(1)**, 1-25.

Genest, C. (1987). Frank's family of bivariate distributions. *Biometrika*, **74**, 549-555.

Genest, C. and Rivest, L. (1993). Statistical inference procedures for bivariate archimedean copulas. *J. Amer. Statis. Assoc.*, **88**, 1034-1043.

Gibbons, J.D. (1988). *Nonparametric Statistical Inference.* Dekker, New York.

Gumbel, E.J. (1960). Distributions des valeurs extremes en plusiers dimensions. *Publ. Inst. Statist.,* University of Paris, **9**, 171-173.

Herath, H. and Kumar, Pranesh (2007) . New research directions in engineering economics – modeling dependencies with copulas. *The Engineering Economist*, **52(4)**, 305-331.

Hoeffding, W. (1940). (reprinted as) Scale-invariant correlation theory. In : *The Collected Works of Wassily Hoeffding,* N.I. Fisher and P. Sen (eds.), Springler-Verlag, New York, 57-107, 1994.

Hoeffding, W. (1941). (reprinted as) Scale-invariant correlation measures for discontinuous distributions. In : *The Collected Works of Wassily Hoeffding*, N.I. Fisher and P. Sen, (eds.). Springler-Verlag, New York, 109-133, 1994 .

Hutchinson, T.P. and Lai, C.D. (1990). *Continuous Bivariate Distributions, Emphasising Applications*. Rumsby Scientific Publishing, Adelaide.

Joe, H. (1993). Parametric families of multivariate distributions with given marginals. *J. Mult. Anal.*, **46**, 262-282 .

Jogdeo, K. (1982). Concepts of dependence. In : *Encyclopedia of Statistical Sciences*, Vol. 1, S. Kotz and N.L. Johnson, (eds.), John Wiley & Sons, New York, 324-334 .

Kimberling, G. and Sampson, A. (1989). A framework for positive dependence. *Ann. Inst. Statist. Math.*, **41,** 31-45.

Kumar, Pranesh and Shoukri, Mohammed M. (2007a). Copula based prediction models: An application to an aortic regurgitation study *BMC Medical Research Methodology*, **7(21).**

Kumar, Pranesh and Shoukri, Mohammed M. (In press, 2007b) Evaluating aortic stenosis using the Archimedean copula methodology. *J. Data Sci.*

Leppik, I.E. *et al.* (1985). A double-blind crossover evaluation of progabide in partial seizures. *Neurology*, **35**, 285.

Marshall, A.W. and Olkin, I. (1988). Families of multivariate distributions. *J. Amer. Statist. Assoc.*, **83**, 834-841.

Melchiori, M.R. (2003). Which Archimedean copula is the right one? *Yield Curve*, **37**, 1-20.

Nelsen, R.B. (1999). *An Introduction to Copulas*. Springer-Verlag, New York, Inc.

Nelsen, R.B. (1995). Copulas, characterization, correlation and counter examples. *Mathematics Magazine*, **68(3)**, 193-198.

Rivest, L. and Wells, M.T. (2001). A martingale approach to the copula-graphic estimator for the survival analysis function under dependant censoring. *J. Multivariate Anal.*, **79**, 138-155 .

Schweizer, B. (1991). Thirty years of copulas. In : *Advances in Probability Distributions with Given Marginals*, G. Dall'Aglio, S. Kotz, G. Salinetti (eds.), Kluwer Academic Publishers, Dordrecht, 13-50 .

Schweizer, B. and Sklar, A. (1961). Topology and Tchebycheff. *The American Mathematical Monthly,* **68(8)**, 760-762 .

Schweizer, B. and Wolff, E.F. (1981). On nonparametric Measures of dependence for random variables. *Ann. Statist.*, **9(4)** , 879-885.

Sklar, A. (1959). Functions de repartition a n dimensions et leurs merges. *Publ. Inst. Statist.*, University of Paris, **8**, 229-231.

Tjøstheim, D. (1996). Measures of dependence and tests of independence. *Statistics*, **28**, 249–284 .

Whitt, W. (1976). Bivariate distributions with given marginals. *Ann. Statist.*, **4(6),** 1280-1289 .

Zheng, M. and Klein, J.P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika,* **82**, 127-138.