

Copula associated to order statistics

Ulisses U. dos Anjos, Nikolai Kolev and Nelson I. Tanaka

University of Sao Paulo

Abstract: We exhibit a copula representation of the (r, s) -th bivariate order statistics from an independent sample of size n . We give conditions when such a representation converges weakly to a bivariate Gaussian copula. A recurrence relationship between the density of the order statistics is presented and related Fréchet bounds are given. The usefulness of those results are stressed through examples.

Key words: Bivariate binomial, copula, Fréchet bounds, normal asymptotics, order statistics.

1 Introduction

Copulas are important in statistical modeling since they connect the marginal distributions to restore the joint distribution, or as typically explained “a copula *couples* a joint distribution function to its univariate margins”, see e.g. Nelsen (1999, p. 15). The copula theory has an incredible evolution during the last decade, motivated by its application in probability theory, statistics, finance, insurance, economics, see for example Cherubini *et al.* (2004) and references therein. One important dependence structure in Statistics is the order statistics. In this paper we bring some contribution on the bivariate order statistic distribution based on copula.

At first, let us outline several basic facts concerning bivariate copulas. A two-dimensional copula is a function $C : [0, 1]^2 \rightarrow [0, 1]$ such that

- (i) $C(t, 0) = C(0, t) = 0$ and $C(t, 1) = C(1, t) = t$ for all t in $[0, 1]$;
- (ii) C is 2-increasing, i.e.

$$V_C([u_1, u_2] \times [v_1, v_2]) := C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0,$$

for all $u_1, u_2, v_1, v_2 \in [0, 1]$ with $u_1 \leq u_2$ and $v_1 \leq v_2$. Alternatively, copulas can be defined as follows: Let X and Y be continuous random variables with distribution functions $F(x) = P(X \leq x)$ and $G(y) = P(Y \leq y)$, and joint distribution function $H(x, y) = P(X \leq x, Y \leq y)$. For every (x, y) in $[-\infty, \infty]^2$ consider the point in $[0, 1]^3$ with coordinates $(F(x), G(y), H(x, y))$. This mapping from $[0, 1]^2$ to $[0, 1]$ is a copula.

Both copula definitions are connected by the following basic theorem, e.g. Sklar (1959), which also partially explains the gist of copulas.

Sklar's Theorem *Let H be a two-dimensional distribution function with marginal distribution functions F and G . Then there exists a copula C such that $H(x, y) = C(F(x), G(y))$. Conversely, for any distribution functions F and G and any copula C , the function H defined above is a two-dimensional distribution function with marginals F and G . Furthermore, if F and G are continuous, C is unique.*

Given a joint distribution function H with continuous marginals F and G , as in Sklar's Theorem, it is easy to construct the corresponding copula: $C(u, v) = H(F^{(-1)}(u), G^{(-1)}(v))$, where $F^{(-1)}$ is the cadlag inverse of F , given by $F^{(-1)}(u) = \sup\{x | F(x) \leq u\}$ (and similarly for $G^{(-1)}$). Note as well that if X and Y are continuous random variables with distribution functions as above, then C is the joint distribution function for the random variables $U = F(X)$ and $V = G(Y)$ which are uniformly distributed on $[0, 1]$, to be denoted hereon by $U(0, 1)$.

It is easy to show that if H is a bivariate distribution function with marginals F and G , then $\max\{F(x) + G(y) - 1, 0\} \leq H(x, y) \leq \min\{F(x), G(y)\}$ or, since $H(x, y) = C(F(x), G(y))$, $\max\{u + v - 1, 0\} \leq C(u, v) \leq \min\{u, v\}$. Those inequalities are known as the Fréchet-Hoeffding bounds.

There are few results in literature relating the order statistics and associated copulas. The random variables $\max(X, Y)$ and $\min(X, Y)$ are the order statistics for X and Y . Then, e.g. Nelsen (1999, p. 25),

$$P(\max(X, Y) \leq t) = C(F(t), G(t))$$

and

$$P(\min(X, Y) \leq t) = F(t) + G(t) - C(F(t), G(t)).$$

The above relations are generalized by Georges *et al.* (2001) as follows: Let (X_1, \dots, X_n) be a set of continuous random variables with $F_i(x) = P(X_i \leq x)$, $i = 1, 2, \dots, n$. Denote by C_n the associated copula and let $X_{r:n}$ be r -th order statistic ($1 \leq r \leq n$). Then its distribution function $F_{r:n}(t) = P(X_{r:n} \leq t)$ is given by

$$F_{r:n}(t) = \sum_{k=r}^n \left[\sum_{l=r}^k (-1)^{k-l} \binom{k}{l} \sum C_n(v_1, \dots, v_n) \right], \quad (1.1)$$

where \sum denotes summation over the set

$$\left\{ (v_1, \dots, v_n) \in [0, 1]^n \mid v_i \in \{F_i(t), 1\}, \sum_{i=1}^n \delta_{\{1\}}(v_i) = n - k \right\}$$

with $\delta_{\{1\}}(v_i) = 1$ if $v_i = 1$, and 0 otherwise. It is not hard to see from (1.1), that

$$F_{1:n}(t) = 1 - \overline{C}_n(S_1(t), \dots, S_n(t)),$$

where \overline{C}_n is the survival copula and $S_i(t) = 1 - F_i(t)$. We also note that $X_{n:n} = \max(X_1, \dots, X_n)$ and its distribution function is the diagonal section of the multivariate distribution $F_{n:n}(t) = C_n(F_1(t), \dots, F_n(t))$.

We may also characterize other statistics which are relevant in reliability, life modeling or risk analysis. For example, one could be interested in the range $X_{n:n} - X_{1:n}$ or subranges $X_{r_1:n} - X_{r_2:n}$ for $r_1 > r_2$. However, in order to derive explicit formulas, we need the joint distribution of $X_{r_1:n}$ and $X_{r_2:n}$. In the case of independent and identically distributed random variables, Balakrishnan and Cohen (1991) give more friendly formulas for the density. Nelsen (2003) found the copula $C_{1,n}$ of $X_{1:n}$ and $X_{n:n}$:

$$C_{1,n}(u, v) = v - [\max\{(1 - u)^{\frac{1}{n}} + v^{\frac{1}{n}} - 1, 0\}]^n.$$

In the general case, the problem is open. One solution is then to use Monte Carlo methods, as suggested by Georges *et al.* (2001). A recent study on the degree of association of pairs of ordered random variables is provided Avérous *et al.* (2005).

The purpose of this paper is to shed light in terms of copula on the dependence structure between r -th and s -th order statistics corresponding to n independent observations from (X, Y) . In Section 2 we give a copula representation of the joint distribution function of r -th and s -th order statistics corresponding to X and Y given the associated copula C , as well as the related Fréchet bounds in the last section. We find in Section 3 the asymptotic distribution in the case when $r/n \rightarrow \lambda_1$ and $s/n \rightarrow \lambda_2$ as $n \rightarrow \infty$ such that $0 \leq \lambda_1, \lambda_2 < 1$ or $\lambda_1 = 0$ and $\lambda_2 = 1$, the increasing rank case in Barakat (2001). In Section 4 we show a recurrence relation. The usefulness of results is demonstrated with numerical examples.

2 Order statistics copula

Consider a bivariate distribution function with continuous margins and n independent observations from the population (X, Y) . Let $(X_1, Y_1), \dots, (X_n, Y_n)$, $n \geq 2$, be a sample from continuous distribution with copula C and marginals F and G respectively. Let $X_{r:n}$ and $Y_{s:n}$ be the order statistics of the sample, $1 \leq r, s \leq n$. In this section we find the copula $C_{X_{r:n}, Y_{s:n}}$ associated to the order statistics $X_{r:n}$ and $Y_{s:n}$ as a function of C .

Define, for each pair $(x, y) \in [-\infty, \infty]^2$

$$R_x = \sum_{j=1}^n I\{X_j \leq x\} \quad \text{and} \quad R_y = \sum_{j=1}^n I\{Y_j \leq y\},$$

where $I\{\cdot\}$ is the indicator function. Therefore, $R_x \sim \text{Bin}(n, p)$ and $R_y \sim \text{Bin}(n, q)$, with $p = P(X \leq x) = F(x)$ and $q = P(Y \leq y) = G(y)$.

Since $F(x)$ and $G(y)$ are continuous the pairs $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ can be transformed into $\{(U_1, V_1), \dots, (U_n, V_n)\}$ by $U_i = F(X_i) \sim U(0, 1)$ and $V_i = G(Y_i) \sim U(0, 1)$. Therefore, $p = F(x) = u$ and $q = G(y) = v$. Now, note that for all $(u, v) \in [0, 1]^2$ for which $u = F(x)$, $v = G(y)$, $R_u = \sum_{j=1}^n I\{U_j \leq u\}$ and $R_v = \sum_{j=1}^n I\{V_j \leq v\}$ we get

$$P(X_{r:n} \leq x, Y_{s:n} \leq y) = P(R_u \geq r, R_v \geq s) = P(U_{r:n} \leq u, V_{s:n} \leq v),$$

where $U_{r:n}$ and $V_{s:n}$ are r -th and s -th order statistics corresponding to n independent observations from (U, V) .

Since the joint distribution of (R_u, R_v) is bivariate Binomial, we have that

$$\begin{aligned} H_{U_{r:n}, V_{s:n}}(u, v) &= P(U_{r:n} \leq u, V_{s:n} \leq v) = P(R_u \geq r, R_v \geq s) \\ &= \sum_{j=r}^n \sum_{k=s}^n \sum_m \frac{n! \theta^m (u - \theta)^{j-m} (v - \theta)^{k-m} (1 - u - v + \theta)^{n-j-k+m}}{m! (j-m)! (k-m)! (n-j-k+m)!}, \end{aligned}$$

where $\theta = P(U \leq u, V \leq v) = C(u, v)$ and m is the number of pairs (U_j, V_j) such that $U_j \leq u$ and $V_j \leq v$, $j = 1, \dots, n$, i.e., $\max(0, j+k-n) \leq m \leq \min(j, k)$.

The marginal distributions of $P(U_{r:n} \leq u, V_{s:n} \leq v)$ are

$$P(U_{r:n} \leq u) = \sum_{j=r}^n \binom{n}{j} u^j (1-u)^{n-j} \quad \text{and} \quad P(V_{s:n} \leq v) = \sum_{k=s}^n \binom{n}{k} v^k (1-v)^{n-k},$$

which, in fact, are Beta distributed random variables, i.e. $U_{r:n} \sim \text{Beta}(r, n-r+1)$ and $V_{s:n} \sim \text{Beta}(s, n-s+1)$. Let $\beta_{r, n-r+1}^{-1}$ and $\beta_{s, n-s+1}^{-1}$ be the inverses of these Beta distributions.

The copula associated to order statistics of the pair $(X_{r:n}, Y_{s:n})$ is the same copula of the pair $(U_{r:n}, V_{s:n})$, see also Lemma 6 in Avérous *et al.* (2005), i.e.

$$C_{X_{r:n}, Y_{s:n}}(w, t) = C_{U_{r:n}, V_{s:n}}(w, t) = H_{U_{r:n}, V_{s:n}}(\beta_{r, n-r+1}^{-1}(w), \beta_{s, n-s+1}^{-1}(t)).$$

Thus, we obtain the following statement.

Theorem 1 *Under the above notations the copula $C_{U_{r:n}, V_{s:n}}$ is given by*

$$\begin{aligned} C_{U_{r:n}, V_{s:n}}(w, t) &= \sum_{j=r}^n \sum_{k=s}^n \sum_m \frac{n! C(\beta_{r, n-r+1}^{-1}(w), \beta_{s, n-s+1}^{-1}(t))^m}{m! (j-m)! (k-m)! (n-j-k+m)!} \\ &\quad \times [\beta_{r, n-r+1}^{-1}(w) - C(\beta_{r, n-r+1}^{-1}(w), \beta_{s, n-s+1}^{-1}(t))]^{j-m} \\ &\quad \times [\beta_{s, n-s+1}^{-1}(t) - C(\beta_{r, n-r+1}^{-1}(w), \beta_{s, n-s+1}^{-1}(t))]^{k-m} \\ &\quad \times [1 - \beta_{r, n-r+1}^{-1}(w) - \beta_{s, n-s+1}^{-1}(t) + C(\beta_{r, n-r+1}^{-1}(w), \beta_{s, n-s+1}^{-1}(t))]^{n-j-k+m}. \end{aligned}$$

The formula given by Theorem 1 presents a relation between the copula C associated to the random vector (X, Y) and the copula of order statistics $C_{U_{r:n}, V_{s:n}}$. Therefore, we can do inferences about $C_{U_{r:n}, V_{s:n}}$ knowing C .

In fact, the statement of Theorem 1 can be obtained as a consequence of Exercise 2.2.2 given in David (1981, p. 25) and relation $C(u, v) = H(F^{(-1)}(u), G^{(-1)}(v))$.

3 Asymptotic copula

In this section we derive from the limit distribution of the pair (R_u, R_v) the asymptotic copula in order to find an approximation to the joint distribution of $(X_{r:n}, Y_{s:n})$.

In Barakat (2001) properties of joint distribution (R_u, R_v) are investigated, and as consequences, limiting distribution results are obtained for the vector $(X_{r:n}, Y_{s:n})$ where $1 \leq r, s \leq n$. For fixed $r, s \geq 1$, as $n \rightarrow \infty$, the pair (r, s) is called fixed rank (or the case of extreme order statistics). When $r, s \rightarrow \infty$ as $n \rightarrow \infty$, (r, s) is called increasing rank. One particular rate of increase of special interest in this work is when $r/n \rightarrow \lambda_1$ and $s/n \rightarrow \lambda_2$ as $n \rightarrow \infty$ such that $0 \leq \lambda_1, \lambda_2 < 1$ or $\lambda_1 = 0, \lambda_2 = 1$. Additionally, Barakat (2001) presents nine other cases covering the possible asymptotic distributions of bivariate order statistics. We consider only the increasing rank case, but the method elaborated here can be applied similarly in the other cases.

We extract the asymptotic copula, denoted by C^a , from the limiting distribution of (R_u, R_v) and use the corresponding approximation to evaluate the joint distribution function of order statistics $(X_{r:n}, Y_{s:n})$, as follows.

The basic result in Barakat (2001) is Theorem 2.2, which gives the conditions for the following convergence

$$\left(\frac{R_u - nu}{\sqrt{nu(1-u)}}, \frac{R_v - nv}{\sqrt{nv(1-v)}} \right) \xrightarrow[n \rightarrow \infty]{\mathbf{d}} \mathcal{N}_\rho,$$

where $\xrightarrow[n \rightarrow \infty]{\mathbf{d}}$ means a convergence in distribution and \mathcal{N}_ρ denotes the bivariate Normal distribution with zero mean vector and correlation coefficient given by

$$\rho = \frac{C(u,v) - uv}{\sqrt{uv(1-u)(1-v)}}.$$

The above theorem in our notation has the following form,

Theorem 2 *Let $\min(n - r, r) \rightarrow \infty$ and $\min(n - s, s) \rightarrow \infty$ when $n \rightarrow \infty$. Furthermore, let $r/n \rightarrow \lambda_1$ and $s/n \rightarrow \lambda_2$ as $n \rightarrow \infty$ such that $0 \leq \lambda_1, \lambda_2 < 1$ or $\lambda_1 = 0$ and $\lambda_2 = 1$. If*

$$\rho_{R_u, R_v} = \text{Corr}(R_u, R_v) \xrightarrow[n \rightarrow \infty]{} \rho = \frac{C(u, v) - uv}{\sqrt{uv(1-u)(1-v)}}$$

for a fixed value of ρ such that $|\rho| \leq 1$,

$$\frac{r - nu}{\sqrt{nu(1-u)}} \rightarrow \tau_1 \quad \text{and} \quad \frac{s - nv}{\sqrt{nv(1-v)}} \rightarrow \tau_2$$

hold for fixed constants τ_1 and τ_2 , then

$$P\left(\frac{R_u - nu}{\sqrt{nu(1-u)}} \leq \frac{r - nu}{\sqrt{nu(1-u)}}, \frac{R_v - nv}{\sqrt{nv(1-v)}} \leq \frac{s - nv}{\sqrt{nv(1-v)}} \right) \xrightarrow[n \rightarrow \infty]{\mathbf{d}} \Phi_\rho(\tau_1, \tau_2),$$

where Φ_ρ is the accumulated joint distribution function of \mathcal{N}_ρ . □

Under the above conditions and from the fact that $C(w, t) = H(F^{-1}(w), G^{-1}(t))$, the asymptotic copula C^a of (R_u, R_v) is given by

$$C^a(w, t) = \Phi_\rho(\Phi^{-1}(w), \Phi^{-1}(t))$$

and the corresponding survival copula has the form

$$\overline{C}^a(w, t) = \overline{\Phi}_\rho(\overline{\Phi}^{-1}(w), \overline{\Phi}^{-1}(t)).$$

Using the asymptotic survival copula \overline{C}^a we can find an approximation for the join survival distribution of (R_u, R_v) . The marginals $R_u \sim \text{Bin}(n, u)$ and $R_v \sim \text{Bin}(n, v)$ and under conditions of Theorem 2 we have

$$\begin{aligned} \overline{H}_{R_u, R_v}(r, s) &\approx \overline{C}^a\left(\overline{B}(n, u, r), \overline{B}(n, v, s)\right) \\ &= \overline{\Phi}_\rho\left(\overline{\Phi}^{-1}\left(\overline{B}(n, u, r)\right), \overline{\Phi}^{-1}\left(\overline{B}(n, v, s)\right)\right). \end{aligned}$$

where $\overline{B}(n, u, r) = P(R_u \geq r)$ and $\overline{B}(n, v, s) = P(R_v \geq s)$. Since $P(R_u \geq r) = P(U_{r:n} \leq u)$ and $P(R_v \geq s) = P(V_{s:n} \leq v)$, we get

$$\begin{aligned} &\overline{\Phi}_\rho\left(\overline{\Phi}^{-1}\left(\overline{B}(n, u, r)\right), \overline{\Phi}^{-1}\left(\overline{B}(n, v, s)\right)\right) \\ &= \overline{\Phi}_\rho\left(\overline{\Phi}^{-1}\left(\beta_{r, n-r+1}(u)\right), \overline{\Phi}^{-1}\left(\beta_{s, n-s+1}(v)\right)\right). \end{aligned}$$

But $H_{U_{r:n}, V_{s:n}}(u, v) = \overline{H}_{R_u, R_v}(r, s)$ and under conditions of Theorem 2 $H_{U_{r:n}, V_{s:n}}(u, v)$ can be approximated by

$$H_{U_{r:n}, V_{s:n}}(u, v) \approx \overline{\Phi}_\rho\left(\overline{\Phi}^{-1}\left(\beta_{r, n-r+1}(u)\right), \overline{\Phi}^{-1}\left(\beta_{s, n-s+1}(v)\right)\right).$$

Now, using the well know relation $\overline{C}(w, t) = w + t - 1 + C(1 - w, 1 - t)$ where $\overline{C}(w, t)$ is the survival copula, we have

$$\begin{aligned} H_{U_{r:n}, V_{s:n}}(u, v) &\approx \beta_{r, n-r+1}(u) + \beta_{s, n-s+1}(v) - 1 + \\ &\quad \Phi_\rho\left(\Phi^{-1}(1 - \beta_{r, n-r+1}(u)), \Phi^{-1}(1 - \beta_{s, n-s+1}(v))\right). \end{aligned}$$

Finally, using $u = F(x)$ and $v = G(y)$ and the relationship $C_{X_{r:n}, Y_{s:n}} = C_{U_{r:n}, V_{s:n}}$, we obtain

$$\begin{aligned} H_{X_{r:n}, Y_{s:n}}(x, y) &\approx \beta_{r, n-r+1}(F(x)) + \beta_{s, n-s+1}(G(y)) - 1 + \\ &\quad \Phi_\rho\left(\Phi^{-1}(1 - \beta_{r, n-r+1}(F(x))), \Phi^{-1}(1 - \beta_{s, n-s+1}(G(y)))\right). \end{aligned} \tag{3.1}$$

Remark 1. It is worth to note that the copula of $(U_{r:n}, V_{s:n})$ is different than the copula of (R_u, R_v) . In this work, we extract the asymptotic copula of (R_u, R_v) in order to find an approximation for $P(R_u \geq r, R_v \geq s)$ and then by relationship $P(U_{r:n} \leq u, V_{s:n} \leq v) = P(R_u \geq r, R_v \geq s)$ we obtain the asymptotic result. \square

Next, we give an application example to (3.1).

Example 1 Consider the bivariate Normal distribution with zero mean, unit variances and correlation coefficient -0.5 . Let us calculate $P(X_{9:10} \leq x, Y_{10:10} \leq y)$. For $x = 1.5$ and $y = 1.8$, we have $F(x) = 0.93319$, $G(y) = 0.96406$ and $C(F(x), G(y)) = 0.897319$. The exact value of $P(X_{9:10} \leq x, Y_{10:10} \leq y)$ can be computed by the formula given in Theorem 1, i.e.

$$\begin{aligned} P(X_{9:10} \leq x, Y_{10:10} \leq y) &= P(U_{9:10} \leq u, V_{10:10} \leq v) \\ &= P(U_{9:10} \leq F(x), V_{10:10} \leq G(y)) \\ &= 10 \times C(F(x), G(y))^9 [G(y) - C(F(x), G(y))] + C(F(x), G(y))^{10} \\ &= 0.5902. \end{aligned}$$

For our data we calculate

$$\begin{aligned} \rho &= \frac{C(F(x), G(y)) - F(x)G(y)}{\sqrt{F(x)G(y)(1 - G(y))(1 - F(x))}} \\ &= \frac{0.897319 - 0.93319 \times 0.96406}{\sqrt{0.93319 \times 0.96406(1 - 0.93319)(1 - 0.96406)}} \\ &= -0.0504337 \end{aligned}$$

and using (3.1) we obtain

$$\begin{aligned} H_{X_{r:n}, Y_{s:n}}(x, y) &\approx \bar{\Phi}_\rho \left(\bar{\Phi}^{-1}(\beta_{9,10-9+1}(F(x))), \bar{\Phi}^{-1}(\beta_{10,10-10+1}(G(y))) \right) \\ &= \bar{\Phi}_\rho \left(\bar{\Phi}^{-1}(0.85942), \bar{\Phi}^{-1}(0.69356) \right) \\ &= 0.85942 + 0.69356 - 1 \\ &\quad + \Phi_\rho \left(\Phi^{-1}(1 - 0.85942), \Phi^{-1}(1 - 0.69356) \right) \\ &= 0.5922. \end{aligned}$$

As we can see, the asymptotic copula is easy to calculate and gives a good approximation even when a small sample size is used. \square

4 A recurrence relation

In the univariate case, a number of recurrence relations between the densities and the moments of order statistics are available, e.g. Arnold and Balakrishnan (1998) and David (1981). We shall derive a similar result for the distribution of bivariate order statistics.

Let us define the events

$$A = \{U_{r-1:n-1} \leq u < U_{r:n-1}, V_{s:n-1} \leq v\},$$

$$B = \{U_{r-1:n-1} \leq u < U_{r:n-1}, V_{s-1:n-1} \leq v < V_{s:n-1}\}$$

and

$$D = \{U_{r:n} \leq u < U_{r+1:n}, V_{s:n} \leq v\}$$

for $u, v \in [0, 1]$. Then

$$P(A) = P(U_{r-1:n-1} \leq u, V_{s:n-1} \leq v) - P(U_{r:n-1} \leq u, V_{s:n-1} \leq v),$$

$$\begin{aligned} P(B) &= P(U_{r-1:n-1} \leq u, V_{s-1:n-1} \leq v) - P(U_{r-1:n-1} \leq u, V_{s:n-1} \leq v) \\ &\quad - P(U_{r:n-1} \leq u, V_{s-1:n-1} \leq v) + P(U_{r:n-1} \leq u, V_{s:n-1} \leq v) \end{aligned}$$

and

$$P(D) = P(U_{r:n} \leq u, V_{s:n} \leq v) - P(U_{r+1:n} \leq u, V_{s:n} \leq v).$$

Theorem 3 *Under the above notations the following relationship holds*

$$nP(B)C(u, v) = rP(D) + nP(A)u \quad (4.1)$$

for $2 \leq r, s \leq n-1$, $n \geq 2$.

Proof. Consider a sample of size n for which the event D occurs. Partition this sample randomly into two subsamples, one of size $(n-1)$ and one with size 1. Consider the event

$$E = \{\text{the observation singled out has } U\text{-value} \leq u\}.$$

Then we have $P(D \cap E) = P(D)P(E|D) = \frac{rP(D)}{n}$. Denote by A_1 and B_1 the events

$$Q_U = \left\{ \begin{array}{l} \text{the event A occurs for the sample size } n-1 \\ \text{and the observation singled out has } U\text{-value} \leq u \end{array} \right\}$$

and

$$Q_V = \left\{ \begin{array}{l} \text{the event B occurs for the sample size } n-1 \\ \text{and the observation singled out has} \\ U\text{-value} \leq u \text{ and } V\text{-value} \leq v \end{array} \right\}.$$

Clearly, Q_U and Q_V are disjoint, and the event $D \cap E$ can occur if and only if either Q_U or Q_V occurs. Hence,

$$P(D \cap E) = P(Q_U) + P(Q_V) = \frac{rP(D)}{n} = uP(A) + P(U \leq u, V \leq v)P(B).$$

From Sklar's Theorem we have

$$P(U_{r:n} \leq u, V_{s:n} \leq v) = C_{U_{r:n}, V_{s:n}}(\beta_{r, n-r+1}(u), \beta_{s, n-s+1}(v)) = C_{U_{r:n}, V_{s:n}}(w, t),$$

where $u = \beta_{r,n-r+1}^{-1}(w)$ and $v = \beta_{s,n-s+1}^{-1}(t)$ are the inverses of $Beta(r, n - r + 1)$ and $Beta(s, n - s + 1)$ distributions, respectively.

Then,

$$\begin{aligned} P(A) &= P(U_{r-1:n-1} \leq u, V_{s:n-1} \leq v) - P(U_{r:n-1} \leq u, V_{s:n-1} \leq v) \\ &= C_{U_{r-1:n-1}, V_{s:n-1}}(w, t) - C_{U_{r:n-1}, V_{s:n-1}}(w, t), \end{aligned}$$

$$\begin{aligned} P(B) &= P(U_{r-1:n-1} \leq u, V_{s-1:n-1} \leq v) - P(U_{r-1:n-1} \leq u, V_{s:n-1} \leq v) \\ &\quad - P(U_{r:n-1} \leq u, V_{s-1:n-1} \leq v) + P(U_{r:n-1} \leq u, V_{s:n-1} \leq v) \\ &= C_{U_{r-1:n-1}, V_{s-1:n-1}}(w, t) - C_{U_{r:n-1}, V_{s-1:n-1}}(w, t) \\ &\quad - C_{U_{r-1:n-1}, V_{s:n-1}}(w, t) + C_{U_{r:n-1}, V_{s:n-1}}(w, t), \end{aligned}$$

$$\begin{aligned} P(D) &= P(U_{r:n} \leq u, V_{s:n} \leq v) - P(U_{r+1:n} \leq u, V_{s:n} \leq v) \\ &= C_{U_{r:n}, V_{s:n}}(w, t) - C_{U_{r+1:n}, V_{s:n}}(w, t), \end{aligned}$$

and

$$P(U \leq u, V \leq v) = C(u, v) = C(\beta_{r,n-r+1}^{-1}(w), \beta_{s,n-s+1}^{-1}(t)).$$

Combining the last five relations we obtain

$$C(\beta_{r,n-r+1}^{-1}(w), \beta_{s,n-s+1}^{-1}(t)) = \frac{r}{n} \frac{P(D)}{P(B)} - \beta_{r,n-r+1}^{-1}(w) \frac{P(A)}{P(B)},$$

i.e. (4.1). □

Remark 2 For convenience, let us substitute

$$C_{r,s;n}(w, t) = C_{U_{r:n}, V_{s:n}}(\beta_{r,n-r+1}(u), \beta_{s,n-s+1}(v)) = P(U_{r:n} \leq u, V_{s:n} \leq v)$$

and introduce the difference notations Δ_1 and Δ_2 defined as

$$\begin{aligned} \Delta_1 C_{r,s;n}(w, t) &= C_{r+1,s;n}(w, t) - C_{r,s;n}(w, t), \\ \Delta_2 C_{r,s;n}(w, t) &= C_{r,s+1;n}(w, t) - C_{r,s;n}(w, t) \end{aligned}$$

and

$$\Delta_{1,2} C_{r,s;n}(w, t) = \Delta_1[\Delta_2 C_{r,s;n}(w, t)].$$

Then, note that

$$P(A) = -\Delta_1 C_{r-1,s;n-1}(w, t), \quad P(D) = -\Delta_1 C_{r,s;n}(w, t)$$

and $P(B) = \Delta_{1,2} C_{r-1,s-1;n-1}(w, t)$. Thus, the relation (3) can be rewritten as

$$\frac{r}{n} \Delta_1 C_{r,s;n}(w, t) = u \Delta_1 C_{r-1,s;n-1}(w, t) - C(w, t) \Delta_{1,2} C_{r-1,s-1;n-1}(w, t).$$

The last formula can be used to compute recursively the value of the joint density of order statistics. □

5 Fréchet bounds

The inclusion-exclusion formula states that if A_1, \dots, A_n are n events, and if the probability of occurrence of at least r of them is denoted by $P(r; n)$, then

$$P(r; n) = \sum_{m=r}^n (-1)^{m-r} \binom{m-1}{r-1} W(m),$$

where

$$W(m) = \sum_{1 \leq i_1 < \dots < i_m \leq n} P(\cap_{j=1}^m A_{i_j}),$$

see e.g. Feller (1968). If one defines $A_i = \{X_i \leq x\}$, $i = 1, \dots, n$, then for the corresponding order statistics $X_{r:n}$ we have

$$P(X_{r:n} \leq x) = \sum_{m=r}^n (-1)^{m-r} \binom{m-1}{r-1} \sum_{1 \leq i_1 < \dots < i_m \leq n} P(\cap_{j=1}^m \{X_{i_j} \leq x\}). \quad (5.1)$$

To find the joint distribution function $P(X_{r:n} \leq x, Y_{s:n} \leq y)$ of order statistics from dependent random variables X and Y , one is led to consider an extension of (5.1) for the case of $K \geq 2$ (finite) classes of events.

Let A_1, \dots, A_n and B_1, \dots, B_n be two classes of events. For integers r and s , $1 \leq r, s \leq n$ define

$$P[r, s; n] = P\{\text{exactly } r \text{ } A_i\text{'s and exactly } s \text{ } B_i\text{'s occur}\}$$

and

$$P(r, s; n) = P\{\text{at least } r \text{ } A_i\text{'s and at least } s \text{ } B_i\text{'s occur}\}.$$

Let

$$W(r, s) = \sum P(\cap_{j=1}^r A_{i_j} \cap_{i=1}^s B_{j_i})$$

where \sum denotes summation over the indices $1 \leq i_1 < \dots < i_r \leq n$; $1 \leq j_1 < \dots < j_s \leq n$. It is known, see Fréchet (1943), that

$$P[r, s; n] = \sum_{t=r+s}^{2n} \sum_{i+j=t} (-1)^{t-r-s} \binom{i}{r} \binom{j}{s} W(i, j).$$

Then,

$$P(r, s; n) = \sum_{\alpha=r}^n \sum_{\beta=s}^n P[\alpha, \beta; n] = \sum_{i=r}^n \sum_{j=s}^n (-1)^{i+j-(r+s)} \binom{i-1}{r-1} \binom{j-1}{s-1} W(i, j).$$

Thus, the following bivariate forms of Bonferroni inequalities for any non-negative integer $k \geq 0$ are given by

$$\sum_{t=r+s}^{r+s+2k+1} \sum_{i+j=t} g(i, j, t) \leq P(r, s; n) \leq \sum_{t=r+s}^{r+s+2k} \sum_{i+j=t} g(i, j, t), \quad (5.2)$$

where $g(i, j, t) = (-1)^{t-(r+s)} \binom{i-1}{r-1} \binom{j-1}{s-1} W(i, j)$, see Meyer (1969). For $2k \geq n - r - s$ the two bounds coincide, hence yielding an equality.

Our aim is to evaluate $C_{U_{r:n}, V_{s:n}}(w, t)$ using (5.2) by making the function $W(i, j)$ more specific. Consider the events $A_i = \{U_i \leq \beta_{r, n-r+1}^{-1}(w)\}$ and $B_i = \{V_i \leq \beta_{s, n-s+1}^{-1}(t)\}$ for $U_i, V_i \sim U(0, 1)$, $i = 1, \dots, n$, arranged in an $2 \times n$ matrix. Then $P(r, s; n) = P(U_{r:n} \leq \beta_{r, n-r+1}^{-1}(w), V_{s:n} \leq \beta_{s, n-s+1}^{-1}(t)) = C_{U_{r:n}, V_{s:n}}(w, t)$. The elements taken from different rows are independent and each element in any given column has the same probability. Then, $W(i, j) = \sum_d T(i, j; d)$, where in the summation the contribution terms $T(i, j; d)$ consist of $\max(i + j - n) \leq d \leq \min(i, j)$ pairs $A_i \cap B_i$ and the remaining pairs being taken from different columns, i.e.

$$T(i, j; d) = \binom{n}{d} [C(\beta_{r, n-r+1}^{-1}(w), \beta_{s, n-s+1}^{-1}(t))]^d \times \binom{n-d}{i-d} [\beta_{r, n-r+1}^{-1}(w)]^{i-d} \times \binom{n-i}{j-d} [\beta_{s, n-s+1}^{-1}(t)]^{j-d}, \tag{5.3}$$

see e.g. Galambos (1975). While the formula for the bounds in (5.2) may seem complicated, its actual computation is quite simple and fast. We first build a table for binomial coefficients and for the exponents of $C(\beta_{r, n-r+1}^{-1}(w), \beta_{s, n-s+1}^{-1}(t))$, $\beta_{r, n-r+1}^{-1}(w)$ and $\beta_{s, n-s+1}^{-1}(t)$ occurring in (5.3) for given values of n, i, j and d . The method of calculation is the same whatever the sample size and the bivariate distribution are.

Let us substitute $k = 0$ in (5.2). Then, the simplest Fréchet bounds are given by

$$W(r, s) - rW(r + 1, s) - sW(r, s + 1) \leq C_{U_{r:n}, V_{s:n}}(w, t) \leq W(r, s).$$

Example 1 (continued) For the data of Example 1 we computed by taking $w = 0.8594$, $t = 0.6936$ the exact value $C_{U_{9:10}, V_{10:10}}(w, t) = 0.5902$. Following the above described procedure we obtain the following inequalities:

$$\begin{aligned} k = 0 : \quad & 0.5902 \leq C_{U_{9:10}, V_{10:10}}(w, t) \leq 3.6361; \\ k = 1 : \quad & 0.5902 \leq C_{U_{9:10}, V_{10:10}}(w, t) \leq 0.5902 \end{aligned}$$

and no further correction was obtained by increasing k in (5.2). Now, for $r = 2$ and $s = 3$ by choosing $w = 0.1405798$, $t = 0.9980501$ the exact value is

$$C_{U_{2:10}, V_{3:10}}(w, t) = 0.1398483.$$

Using the Fréchet bounds we obtain the following inequalities:

$$\begin{aligned} k = 3 : \quad & 0.1255057 \leq C_{U_{2:10}, V_{3:10}}(w, t) \leq 0.3034248; \\ k = 4 : \quad & 0.1398175 \leq C_{U_{2:10}, V_{3:10}}(w, t) \leq 0.1406579; \\ k = 5 : \quad & 0.1398482 \leq C_{U_{2:10}, V_{3:10}}(w, t) \leq 0.1398491; \\ k = 6 : \quad & 0.1398483 \leq C_{U_{2:10}, V_{3:10}}(w, t) \leq 0.1398483 \end{aligned}$$

and no further correction was obtained by increasing k in (5.2). \square

The general multivariate form of inequalities (5.2) is given by Meyer (1969). While the bivariate distribution for each pair of the components may be reasonable, the vector behaviour in higher dimensions may be either unknown, or may be such that the computation of these distribution poses difficulty.

Acknowledgements

The first author thanks for a financial support from CNPq Grant 141503/02-5. The second author is partially supported by FAPESP, Grant 03/10105-2 and PROBRAL (CAPES/DAAD), Grant 171-04. The third author was partially supported by Projeto Temático FAPESP, number 99/10611-8.

(Received December, 2004. Accepted April, 2005.)

References

- Arnold, B. and Balakrishnan, N. (1989). *Relations, Bounds and Approximations for Order Statistics*. New York: Springer Verlag.
- Avérous, J., Genest, C. and Kochar, S. (2005). On the dependence structure of order statistics. To appear in *Journal of Multivariate Analysis*, available at <http://www.mat.ulaval.ca/pages/genest/>.
- Balakrishnan, N. and Cohen, A. C. (1991). *Order Statistics and Inference*. San Diego: Academic Press.
- Barakat, H. (2001). The asymptotic distribution theory of bivariate order statistics. *Ann. Inst. Stat. Math.*, **53**, 487-497.
- Cherubini, U., Luciano, E. and Vecchiato, W. (2004). *Copula Methods in Finance*. Chichester: Wiley Finance.
- David, H. (1981). *Order Statistics*, 2.ed. edition. New York: John Wiley and Sons.
- Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Volume I, 3.ed. edition. New York: John Wiley and Sons.
- Fréchet, M. (1943). Les Probabilités Associées à um Systém d' Événements Compatibles et Dépendantes. *Exposés d' Analyse Geral*, **942**. Paris: Hermann.
- Galambos, J. (1975). Order statistics of sample from multivariate distributions. *Journal of the American Statistical Association*, **70**, 674-680.

- Georges, P., Lamy, A-G., Nicolas, G., Quibel, G. and Roncalli, T. (2001). Multivariate survival modeling: a unified approach with copulas. *Crédit Lyonnais* (Working paper).
Available at http://gro.creditlyonnais.fr/content/rd/home_copulas.htm.
- Meyer, R. (1969). A note on a “multivariate” form of Bonferroni’s inequalities. *The Annals of Mathematical Statistics*, **40**, 692-693.
- Nelsen, R. (1999). *An Introduction to Copulas*. New York: Springer.
- Nelsen, R. (2003). Properties and applications of copulas: a brief survey. In *Proceedings of the First Brazilian Conference on Statistical Modeling in Insurance and Finance*, Dhaene, J., Kolev, N. and Morettin, P. A. (eds), São Paulo: University Press USP, 10-28.
- Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, **8**, 229-231.

Ulisses U. dos Anjos, Nikolai Kolev and Nelson I. Tanaka

Department of Statistics

University of São Paulo

Cx. Postal 66.281, 05311-970, São Paulo, SP, Brazil

E-mails: anjos@ime.usp.br, nkolev@ime.usp.br and nitanaka@ime.usp.br