KU LEUVEN

FACULTY OF SCIENCE

Blomqvist's Beta: An Unconventional Measure of Association

General properties, Asymptotic Distributions and Connection with Copulas

Louis CLAEYS
Amos NICODEMUS
Samuel VERGAUWEN

Supervisor: Prof. Clément Cerovecki

Bachelor thesis submitted in fulfilment of the requirements for the degree in Bachelor of Science in Mathematics

Academic year 2022-2023

Contents

Introduction

In many applied sciences, the first step in establishing a causal relation between variables is a study of their correlation. In a mathematical context, *correlation* usually refers to Pearson's ρ, which is a specific way of measuring linear dependence. However, more general measures of dependence have been introduced, which are capable of capturing nonlinear relationships or have desirable properties like robustness and known asymptotic distributions (for an overview, see [4]). The goal of this paper is to study one particular example of such an *association measure*, namely Blomqvist's beta (denoted β), introduced by Nils Blomqvist in 1950.

Chapter 1 briefly recalls the concepts related to association measures by summarizing the most important properties of the Pearson correlation coefficient.

In Chapter 2, Blomqvist's beta is introduced, based on his original paper [1]. We study the sample and population versions and derive the distribution of an estimator for β under independence. This yields a nonparametric independence test, which we apply to recent financial data. We perform a simulation test to verify the general asymptotic distribution of the sample version of β proposed by Blomqvist, and find that the empirical asymptotic variance does not agree with Blomqvist's theory.

In pursuit of a multivariate generalization of Blomqvist's beta, the notion of *copulas* is introduced in Chapter 3. We limit ourselves to the main results, so the chapter may be skipped without loss of continuity by readers who are familiar with the topic.

Finally, in Chapter 4 we discuss the multivariate version of Blomqvist's beta proposed by Schmid and Schmidt in their 2007 paper [8], which also leads to an expression for the asymptotic variance that agrees with experiment.

All code that was used to perform the simulations and generate the figures is available on github: https://github.com/AmosNico/Blomqvists-Beta.

Chapter 1

Pearson's Correlation Coefficient

Before investigating Blomqvist's beta, we discuss the most commonly used correlation coefficient in statistics, the Pearson correlation coefficient. The aim of this chapter is to familiarize with Pearson's correlation and its main characteristics. We will limit ourselves to properties that will be investigated again in Chapter 2 for the Blomqviqt's beta. This can be helpful for the understanding of subsequent chapters. The material in this chapter is based on chapter two of Kruskal's book *Ordinal Measures of Association* [4].

1.1 Definition

Definition 1.1 (Pearson)**.** Consider (X, Y), a pair of random variables, and the corresponding bivariate distribution. The correlation coefficient ρ_{XY} is defined as

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}\, X \,\mathrm{Var}\, Y}}. \tag{1.1}$$

To give an interpretation of this definition, take α and β so that $\alpha + \beta X$ is the best linear estimate for Y given X. This means that $a = \alpha$ and $b = \beta$ minimize $E[Y - a - bX]^2$. The minimizing values are found to be $\beta = \mathrm{Cov}(X, Y)/\,\mathrm{Var}\, X$ and $\alpha = E[Y] - \beta E[X]$. By inserting this into (1.1) we obtain

$$\rho^2_{XY} = \frac{\mathrm{Var}\, Y - E[Y - \alpha - \beta X]^2}{\mathrm{Var}\, Y}, \tag{1.2}$$

and that

$$\rho^2_{XY} = \frac{\mathrm{Var}(\alpha + \beta X)}{\mathrm{Var}\, Y} = \frac{E[\alpha + \beta X - E[Y]]^2}{\mathrm{Var}\, Y}. \tag{1.3}$$

It is not difficult to interpret these formulas. For example, equation (1.3) says that ρ^2_{XY} is the variance of the best linear estimate of Y, relative to the variance of Y. With these equations it becomes clear that ρ^2_{XY} is a tool to measure the strength of the linear relation between X and Y.

Note that $\rho_{XY} \in [-1, 1]$. This can be seen by taking the square root on both sides of the Cauchy-Schwarz inequality,

$$(\mathrm{Cov}(X, Y))^2 \leq \mathrm{Var}\, X \;\; \mathrm{Var}\, Y,$$

we get

$$-\sqrt{\operatorname{Var} X \operatorname{Var} Y} \leq \operatorname{Cov}(X,Y) \leq \sqrt{\operatorname{Var} X \operatorname{Var} Y}$$
$$\Leftrightarrow -1 \leq \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var} X \operatorname{Var} Y}} \leq 1.$$

The conditions to reach $-1, 0, 1$ are given by:

- $\rho_{XY} = \pm 1 \Leftrightarrow E[Y - \alpha - \beta X] = 0 \Leftrightarrow$ the joint distribution of X and Y lies entirely on the straight line $Y = \alpha + \beta X$
- $\rho_{XY} = 0 \Leftrightarrow \operatorname{Cov}(X,Y) = E[XY] - E[X]E[Y] = 0 \Leftrightarrow E[XY] = E[X]E[Y] \Leftarrow X \text{and } Y \text{independent}$

1.2 Sample version

When given a sample from (X, Y), it is possible to calculate the *sample correlation coefficient* by substituting sample variance and covariances into (1.1).

Definition 1.2 (Sample correlation)**.** Suppose we observe a sample $(x_1, y_1), \ldots, (x_n, y_n)$ from the bivariate distribution (X, Y). Denote $\bar{x}$ and $\bar{y}$ the respective sample means. Then the sample correlation is given by

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

In general, the asymptotic distribution of r_{xy} can be quite complex. For a more detailed discussion, the interested reader can consult [5]. We mention here only a specific case where the finite sample distribution can be calculated exactly:

Lemma 1.3. *Suppose (X, Y) follow an uncorrelated ($\rho = 0$) bivariate normal distribution. Denote r_n the sample correlation for a sample of size n. Then the following holds:*

$$r_n \sqrt{\frac{n-2}{1-r_n^2}} \sim t(n-2),$$

where $t(n-2)$ denotes a Student t distribution with $n-2$ degrees of freedom.

Further discussion and proof of the above result are given in [7].

Chapter 2

Classical Discussion of Blomqvist's Beta

In this section we discuss some classical results on Blomqvist's measure of dependence, as introduced in his 1950 paper [1]. The results of this section are all based on [1], but proofs are original unless stated otherwise.

2.1 Definition

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a realisation of a random vector (X, Y) with a continuous cumulative distribution function (cdf) $F(x, y)$. The idea of Blomqvist's beta is to describe the way in which the sample points are distributed around the marginal sample medians M_X and M_Y. By drawing the lines $x = M_X$ and $y = M_Y$ the x, y-plane is divided in four regions, see Figure 2.1. If X and Y are positively correlated, we expect many points to lie in the first and third quadrants (where $(X - M_X)(Y - M_Y) > 0$), whereas if X and Y are independent we do not expect a significant difference with the number of points in other quadrants (where $(X - M_X)(Y - M_Y) < 0$). This motivates the definition of a measure of dependence based on these quadrants.

Assume for now that $n = 2k$ is even. Let a, b, c and d denote the number of sample points in respectively the first, second, third and fourth quadrant. By definition of the median, we have $a + b = k$, $b + c = k$, $c + d = k$ and $d + a = k$. This implies that $c = a$ and $b = d = k - a$, so all four values are determined by the number of sample points in the first quadrant.

If $n = 2k + 1$ is odd then the lines $x = M_X$ and $y = M_Y$ both contain one, possibly the same, sample point. One has to make a choice whether or not to count these points. The conventions we use ensure that the equations $a = c$ and $b = d = k - a$ remain valid. If one point lays on the intersection of the lines, then we will not count it at all. If both lines contain a different point, then there is one quadrant that touches both points. We will say that one of the points belongs to this quadrant and discard the other. If one would allow two sample points to have the same x- or y-coordinate by dropping the condition that F is continuous, then even in a simple case as in Figure 2.2 there is no obvious way to determine which points to count. Hence we will restrict ourselves to the case of

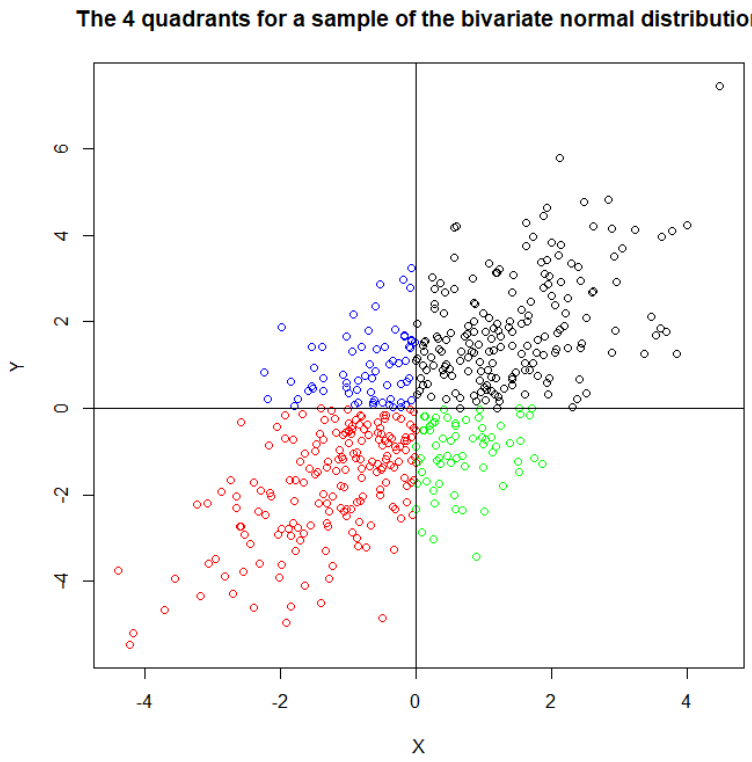

Figure 2.1: The four quadrants for a random sample of 500 points out of a bivariate normal distribution. In this case a and c are clearly larger than b and d.

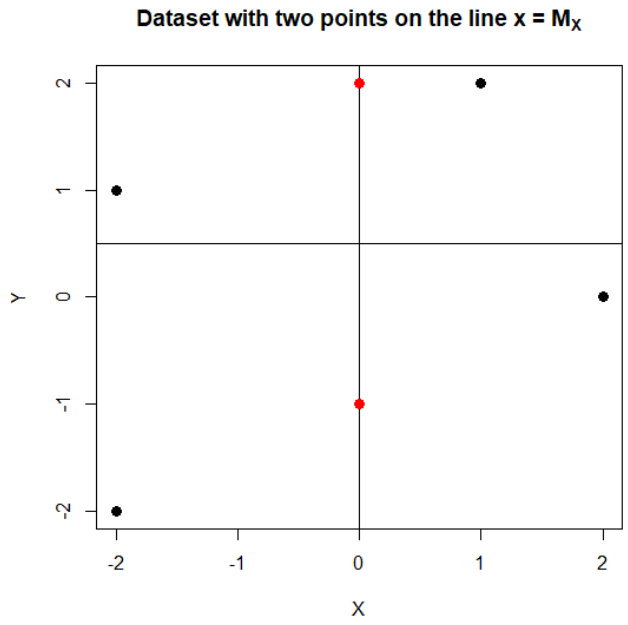

Figure 2.2: An example for $n = 6$ where there is no satisfying way to assign the two red points to quadrants. If one either removes both points or counts them in each adjacent quadrant, then the formula $b = d = k - a = k - c$ (with $k = 3$) no longer holds. A solution could be to assign the points to a random adjacent quadrant.

continuous distributions.

From now on we will refer to the number of sample points in the first and third quadrant as $n_1 := a + c$ and to that of the second and fourth as $n_2 := b + d$. Both n_1 and n_2 are clearly even. We will quantify the association between X and Y by the amount of points in the first and third quadrant compared to the others. Typically one expects a measure of association to lie between -1 and 1 and to be zero if X and Y are independent. In case of independence we expect n_1 and n_2 to be equal, which leads to the following definition of Blomqvist's coefficient:

Definition 2.1 (Blomqvist's beta, sample version)**.** Blomqvist's beta (sometimes called the quadrant measure or medial correlation coefficient) is defined to be

$$\hat{\beta} := \frac{n_1 - n_2}{n_1 + n_2} = \frac{2n_1}{n_1 + n_2} - 1, \qquad \text{with } -1 \leq \hat{\beta} \leq 1. \tag{2.1}$$

2.2 Sample Distribution of $\hat{\beta}$

Denote (X, Y) the bivariate random vector we are interested in, with cdf $F(x, y)$. We wish to find the distribution and asymptotic behaviour of the quadrant measure (2.1). Under fairly general conditions, the sample medians are consistent estimates for the population values:

1. The population medians are uniquely defined (and assumed 0).
2. The marginal distributions of $F(x, y)$ admit density functions $f_1(x), f_2(y)$ which are continuous in some neighbourhood of the origin.
3. $f_1(0)$ and $f_2(0)$ are nonzero.

Define now, for any x, y:

$$\begin{aligned} a(x, y) &= P\left(X > x, Y > y\right), \\ b(x, y) &= P\left(X \leq x, Y > y\right), \\ c(x, y) &= P\left(X \leq x, Y \leq y\right), \\ d(x, y) &= P\left(X > x, Y \leq y\right). \end{aligned}$$

Clearly we have $a + b + c + d = 1$. We wish to determine the distribution of n_1 when observing a sample of size $n = 2k + 1$. This calculation was performed by Blomqvist [1]:

Lemma 2.2. *Let $x, y \in \mathbb{R}$ and $r \in \{0, 1, \ldots, k\}$. The probability of the combined event $n_1 = 2r, M_X \in (x, x + dx), M_Y \in (y, y + dy)$ is given by*

$$p_k(2r; x, y) = \frac{(2k + 1)!}{(r!)^2 (k - r)!^2} (ac)^r (bd)^{k-r} S, \tag{2.2}$$

where

$$S = \frac{r}{a} d_x a \; d_y a - \frac{k - r}{b} d_x b \; d_y b + \frac{r}{c} d_x c \; d_y c - \frac{k - r}{d} d_x d \; d_y d + dF.$$

Proof. We wish to calculate

$$p_k(2r; x, y) := P(n_1 = 2r, M_X \in (x, x + dx), M_Y \in (y, y + dy)), \tag{2.3}$$

where M_X, M_Y denote the sample median of X, Y. These medians may be determined by one or two sample points.

One sample point: The probability that this point is in $(x, x+dx) \times (y, y+dy)$ is simply $dF(x,y)$. The probability that the remaining $2k$ points are distributed in such a way that this point is the median and $n_1 = 2r$ is given by the chance that r of them are in the first and third quadrant and $k-r$ of them in the second and fourth. We thus obtain a factor

$$\frac{(2k)!}{(r!)^2(k-r)!^2}(ac)^r(bd)^{k-r}$$

upon using the multinomial distribution.

Two sample points: In this case one of the points has first component in $(x, x+dx)$ and the other has second component in $(y, y+dy)$. By convention, these points are counted together as one point in the quadrant they are both touching. For example, one can see that the points touch the first quadrant with probability

$$P(\text{first in } (x, x+dx) \times (y+dy, \infty),\ \text{second in } (x+dx, \infty) \times (y, y+dy)) = \frac{\partial a}{\partial x}\frac{\partial a}{\partial y}dxdy.$$

In order for these points to define the median and n_1 to be $2r$, the other $2k-1$ points need to be distributed so that $r-1$ lie in the first quadrant, $k-r$ lie in the second and fourth quadrant and r lie in the third. This gives a final result of

$$\frac{(2k-1)!}{r!(r-1)!(k-r)!^2}a^{r-1}c^r(bd)^{k-r}\ d_x a\ d_y a,$$

and we find similar expressions for the other quadrants.

The choice of the sample point(s) that define the median is arbitrary, this introduces a factor of $(2k+1)$ in the first case and $(2k+1)(2k)$ in the second. It is easy to verify this yields the required result. □

In order to obtain the probability of observing $n_1 = 2r$, this result is then integrated over all x, y:

$$P(n_1 = 2r) = \int p_k(2r; x, y). \tag{2.4}$$

In particular, writing Ψ_k for the combined cdf of the sample medians, from the definition of p_k given in (2.3) we can write Ψ_k as

$$d\Psi_k(x,y) = \sum_{r=0}^{k} p_k(2r; x, y).$$

Under the assumptions made earlier, the sample medians are consistent estimates for the population medians, so that

$$\lim_{k\to\infty} \Psi_k(x,y) = \mathbb{1}_{[0,\infty]\times[0,\infty]}(x,y). \tag{2.5}$$

To obtain more insight into (2.4), we wish to consider the special cases of asymptotically large sample sizes and independence.

Large sample: The goal is to study the limit of $P(n_1 \leq 2R)$ as $k, R \to \infty$ and $R/k \to$ constant (with $0 \leq R \leq k$). Blomqvist claims in [1] that we can use (2.5) to write

$$\lim P(n_1 \leq 2R) = \lim \int \frac{\sum_{r=0}^{R} p_k(2r; x, y)}{\sum_{r=0}^{k} p_k(2r; x, y)} d\Psi_k(x, y) = \lim \frac{\sum_{r=0}^{R} p_k(2r; 0, 0)}{\sum_{r=0}^{k} p_k(2r; 0, 0)}. \tag{2.6}$$

We omit the details here, but Blomqvist concludes from this that n_1 is has an asymptotically normal distribution with mean and variance given by

$$4ka_0, \qquad 8ka_0\left(\frac{1}{2} - a_0\right),$$

where $a_0 = a(0, 0) = \int_0^\infty \int_0^\infty dF$. It follows that $\hat{\beta}$ defined as

$$\hat{\beta} = \frac{2n_1}{2k} - 1 = \frac{n_1}{k} - 1$$

should be asymptotically normal with mean $4a_0 - 1$ and standard deviation $2\sqrt{a_0(1 - 2a_0)/k}$. We will later show that this property holds under the additional assumption that the marginals of F are known, but it was shown by Schmid and Schmidt in [8] that the asymptotic variance given above is not correct in general, and the exchange of limits performed in (2.6) is unjustified.

Independence: The expression given by (2.4) and (2.2) is not very tractable, but can be simplified considerably when X and Y are independent. In that case, denoting F_1 and F_2 for the marginals of F, we have

$$\begin{aligned}
a &= (1 - F_1(x))(1 - F_2(y)) & d_x a \; d_y a &= a dF \\
b &= F_1(x)(1 - F_2(y)) & d_x b \; d_y b &= -b dF \\
c &= F_1(x) F_2(y) & d_x c \; d_y c &= c dF \\
d &= 1 - F_1(x) F_2(y) & d_x d \; d_y d &= -d dF,
\end{aligned}$$

and therefore $S = (2k + 1)dF$. The integral in (2.4) then evaluates to be

$$\begin{aligned}
P(n_1 = 2r) &= \int \frac{(2k+1)!}{(r!)^2 (k-r)!^2} (ac)^r (bd)^{k-r} S \\
&= \frac{(2k+1)!(2k+1)}{(r!)^2 (k-r)!^2} \int F_1(x)^k (1 - F_1(x))^k F_2(y)^k (1 - F_2(y))^k dF_1 dF_2 \\
&= \frac{(2k+1)!(2k+1)}{(r!)^2 (k-r)!^2} B(k+1, k+1)^2,
\end{aligned}$$

where B denotes the Beta function. For natural numbers $n, m \in \mathbb{N}_0$ we have

$$B(n, m) = \frac{(m-1)!(n-1)!}{(m+n-1)!},$$

and therefore the previous expression is reduced to

$$P(n_1 = 2r) = \frac{(k!)^2}{(r!)^2 (k-r)!^2} \frac{(k!)^2}{(2k)!} = \frac{\binom{k}{r}^2}{\binom{2k}{k}},$$

which coincides with the value Blomqvist obtained by using a combinatorial argument. The distribution of $n_1/2$ is seen to be a hypergeometric distribution with parameters

$$N = 2k,\ n = K = k.$$

Since this distribution is symmetric about $r = k$, we obtain $E[n_1] = k$. To compute the variance of n_1, we will use Vandermonde's identity:

$$\binom{m+n}{k} = \sum_{r=0}^{k} \binom{m}{r}\binom{n}{k-r}.$$

We find that

$$\begin{aligned} E[n_1^2] = \sum_{r=0}^{k} 4r^2 P(n_1 = 2r) = \binom{2k}{k}^{-1} 4k^2 \sum_{r=0}^{k-1} \binom{k-1}{r}^2 \\ = \frac{4k^2\binom{2k-2}{k-1}}{\binom{2k}{k}} = \frac{2k^3}{2k-1}, \end{aligned}$$

and therefore the variance is given by

$$\mathrm{Var}(n_1) = E[n_1^2] - E[n_1]^2 = \frac{k^2}{2k-1}.$$

It is a known property of the hypergeometric distribution that if we let $N \to \infty$ while keeping the proportion of successes K/N constant, we obtain a binomial distribution with parameters $(n, K/N)$. In our case, this means

$$\frac{n_1}{2} \sim \mathrm{Binom}(k, 1/2)$$

for large k. By invoking the Central Limit Theorem, we see that for large sample sizes under independence,

$$\frac{n_1 - k}{k}\sqrt{2k-1} = \hat{\beta}\sqrt{2k-1}$$

is normally distributed with mean 0 and variance 1, which corresponds with Blomqvist's claim (as can be seen by inserting $a_0 = 1/4$).

Note that under the assumption of independence, the distribution of n_1 does not depend on the marginal distributions of X and Y. In a later section, we will use this result to construct a test of independence based on $\hat{\beta}$.

2.3 Population Variables

The statistic $\hat{\beta}$ given above arises from the idea that variables are positively/negatively correlated when deviations from the median have a tendency to be in the same/different direction. This idea can be formulated as a property of the distribution as follows:

Definition 2.3 (Blomqvist's beta, population version)**.** For two random variables X and Y we define the population version of Blomqvist's beta to be

$$\beta = 2\left[P\left(X < \tilde{X}, Y < \tilde{Y}\right) + P\left(X > \tilde{X}, Y > \tilde{Y}\right)\right] - 1 = 4P\left(X < \tilde{X}, Y < \tilde{Y}\right) - 1,$$

where $\tilde{X}$ and $\tilde{Y}$ are the population medians of X and Y.

Under the assumption that the medians of X and Y are 0 this corresponds to $\beta = 4a_0 - 1$. The estimator $\hat{\beta}$ is consistent for the above quantity. If the term a_0 can be calculated, we can obtain an explicit expression for β. For the case of a bivariate normal distribution, this allows us to obtain a relation between β and the standard correlation ρ. Denote (Z_1, Z_2) a bivariate standard normal with correlation coefficient ρ. Then

$$a_0 = P(Z_1 > 0, Z_2 > 0) = \frac{1}{4} + \frac{\sin^{-1}\rho}{2\pi}.$$

The derivation of this expression is attributed to Cramer [2], and leads to the following relation between ρ and β:

$$\beta = \frac{2}{\pi}\sin^{-1}\rho.$$

Incidentally, for the particular case of bivariate normality, this is the same quantity estimated by Kendall's tau, another common measure of association (see for example [3], page 167).

2.4 Simulation Study

We wish to check the asymptotic normality of the sample version of Blomqvist's beta by means of a simulation study. We will simulate bivariate data from a random vector (X, Y), whose cdf is given by that of the so-called *Clayton copula*

$$F(x, y) = \left(\max\left\{x^{-\theta} + y^{-\theta} - 1, 0\right\}\right)^{-1/\theta},$$

where $x, y \in [0, 1]$ and $\theta \in [-1, \infty)\setminus\{0\}$ is a parameter. The marginals of this distribution are uniform, so that the population medians of X and Y are given by $1/2$. The constant a_0 for this cdf is then

$$a_0 = F(1/2, 1/2) = \left(2^{\theta+1} - 1\right)^{-1/\theta},$$

which we will denote as h_θ. Suppose we have a random sample of (X, Y) of size n. Denote $\hat{\beta}$ the sample version of Blomqvist's beta as defined in the previous section. Blomqvist claims that, for n large enough, we have

$$\sqrt{n}(\hat{\beta} - \beta) \sim \mathcal{N}\left(0, 8h_\theta(1 - 2h_\theta)\right), \tag{2.7}$$

where $\beta = 4h_\theta - 1$.

First, we will discuss the asymptotic normality condition and impact of the sample size n by investigating the empirical distribution of $\hat{\beta}$ for different values of n and a fixed θ. Later, we sample n elements from the Clayton copula, calculate Blomqvist's beta for this sample, and repeat the process N times for several values of θ. The goal is to find out

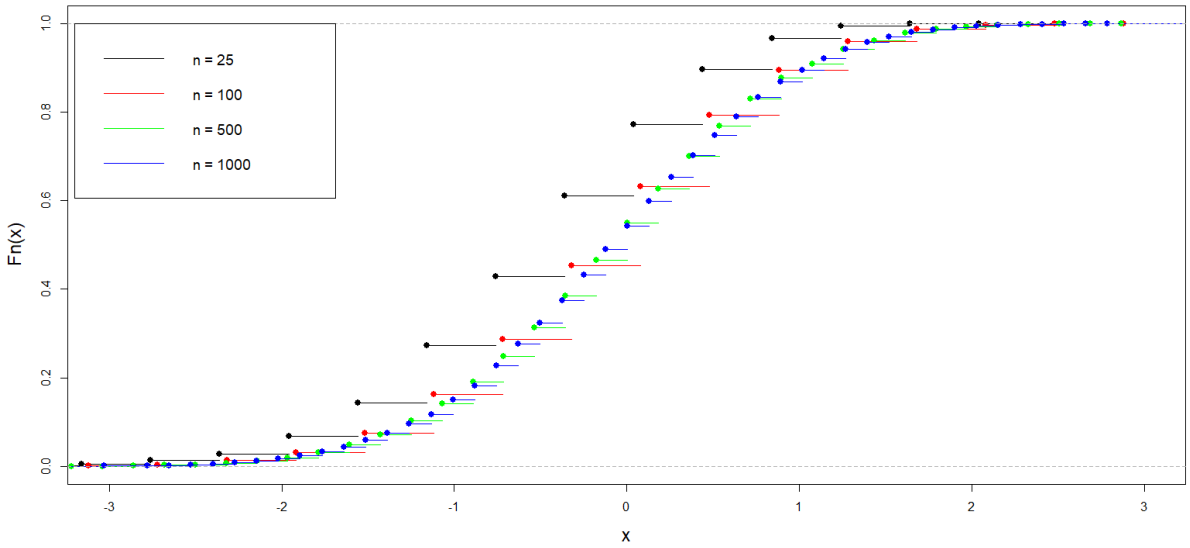

Figure 2.3: ECDF of $\sqrt{n}(\hat{\beta} - \beta)$ for several sample sizes, showing there is indeed convergence as n grows.

whether the mean and variance of (2.7) are plausible.

Asymptotic Normality and Sample Size n: In Figure 2.3, the empirical cumulative distribution function (ECDF) of $\sqrt{n}(\hat{\beta} - \beta)$ is shown for $N = 5000$ samples of size $n = 25, 100, 500, 1000$ from the Clayton copula with parameter $\theta = 2$. The convergence of the distribution is clear from the graph, and the limiting distribution appears to be normal. This is verified by considering normal quantile plots for each of the sample sizes, as is done in Figure 2.4. The quantile plots each show a linear dependence, indicating that the normality of $\hat{\beta}$ is valid even for relatively small sample sizes n. However, Figure 2.5 indicates that for finite sample sizes, the estimator $\hat{\beta}$ could be biased. The empirical mean of $\hat{\beta}$ is consistently lower than the population variable β, and the deviation decays as n gets larger.

Asymptotic Mean and Variance: Denote $\bar{\beta}$ the sample mean of N realisations of $\hat{\beta}$, and s^2 the empirical variance. As a result of asymptotic normality, we can construct an approximate $(1 - \alpha)\%$ confidence interval for $E(\hat{\beta})$ as

$$\left[\bar{\beta} - \frac{t_{N-1,1-\alpha/2}}{\sqrt{N}} s, \bar{\beta} + \frac{t_{N-1,1-\alpha/2}}{\sqrt{N}} s\right],$$

where $t_{N-1,1-\alpha/2}$ denotes the $(1 - \alpha/2)$-quantile of the Student t distribution with $N - 1$ degrees of freedom.

For the asymptotic variance, we use the relation between the sample variance of a normal distribution and the χ^2 distribution to obtain an approximate $(1-\alpha)\%$ confidence interval for the variance of $\sqrt{n}(\hat{\beta} - \beta)$ as

$$\left[\frac{ns^2(N-1)}{\chi^2_{N-1,1-\alpha/2}}, \frac{ns^2(N-1)}{\chi^2_{N-1,\alpha/2}}\right],$$

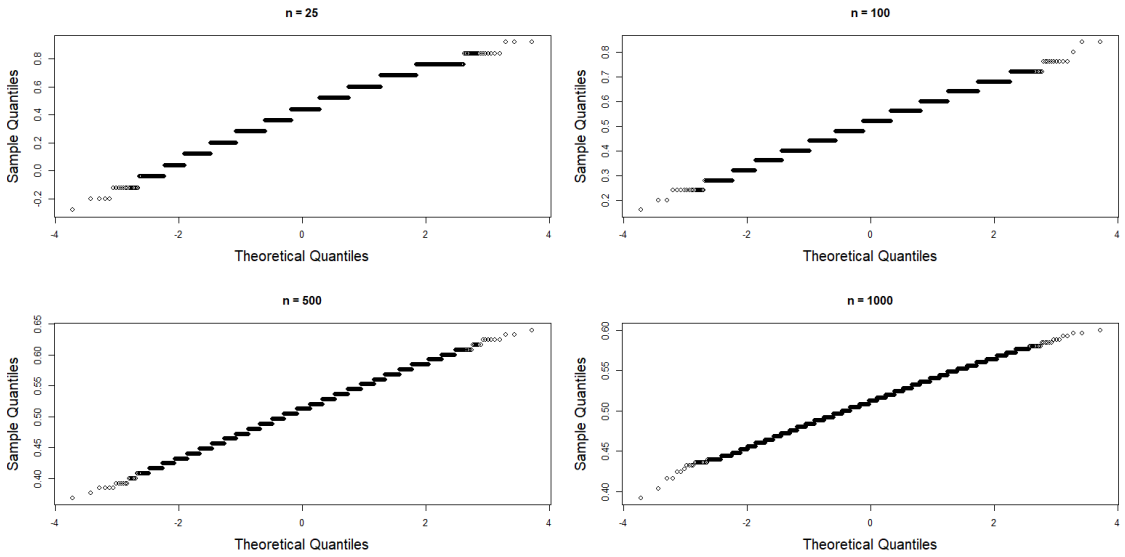

Figure 2.4: Normal QQ Plot of $\hat{\beta}$ for several sample sizes, illustrating the asymptotic normality result.

where the denominators denote quantiles of the χ^2 distribution with $N-1$ degrees of freedom.

The empirical means obtained from the simulation are given in Figure 2.5, along with their 95% confidence intervals. For $|\theta| < 1$, β lies within the confidence interval for $E(\hat{\beta})$, but for larger values of θ we see that the population variable β lies a bit higher than the mean of the estimator $\hat{\beta}$. We believe this discrepancy is an artifact of a finite sample bias, as supported by the second graph in Figure 2.5.

The empirical variances obtained from the simulation are given in Figure 2.6, along with their 95% confidence intervals. Around $\theta = 0$, where the distribution is close to an independent bivariate uniform distribution [1], the predicted variance lies within the confidence intervals. Indeed, in the case of independence, we have explicitly shown that Blomqvist's formula holds. However, it seems that when taking θ further from zero, Blomqvist's formula is no longer true. For nearly every θ with $|\theta| \geq 0.5$, we reject the hypothesis that the variance is given by Blomqvist's formula with 95% confidence. In Chapter 4, we will show that the alternative formula given by Schmid and Schmidt aligns much more closely with the variances observed through the simulation experiment.

2.5 Testing Independence

As a concrete application of testing for independence using Blomqvist's beta, suppose we are interested in whether or not the stock prices of Tesla and Twitter are independent (see Figure 2.7). A scatterplot of the prices is given in Figure 2.8, showing how the Blomqvist coefficient is calculated in this case. There are 693 data points, so that we obtain $k = 346$.

[1] see Chapter 3

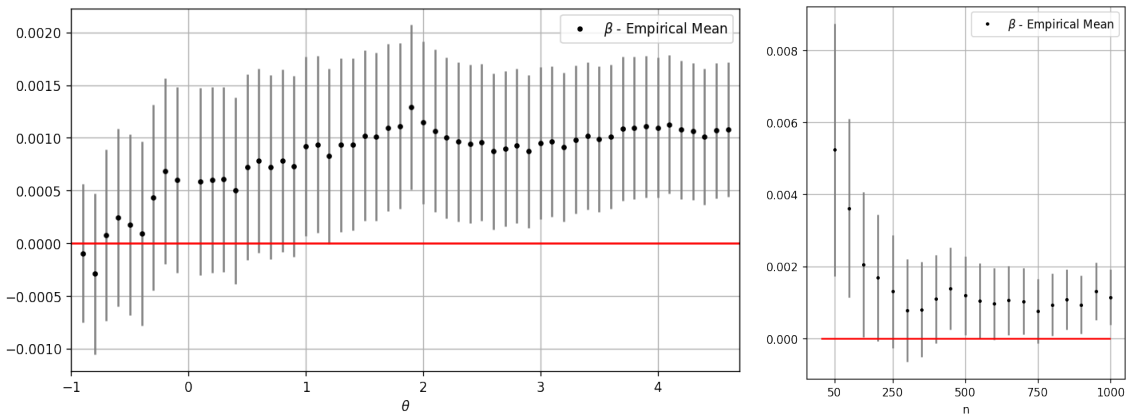

Figure 2.5: Left: Empirical mean compared with population β for $N = 5000$ calculations of Blomqvist's beta for $n = 1000$ samples each, for different values of the distribution parameter θ. The bars denote the 95% confidence intervals. Right: Empirical mean compared with population β for $N = 5000$ calculations of Blomqvist's beta for various sample sizes n, where $\theta = 2$.

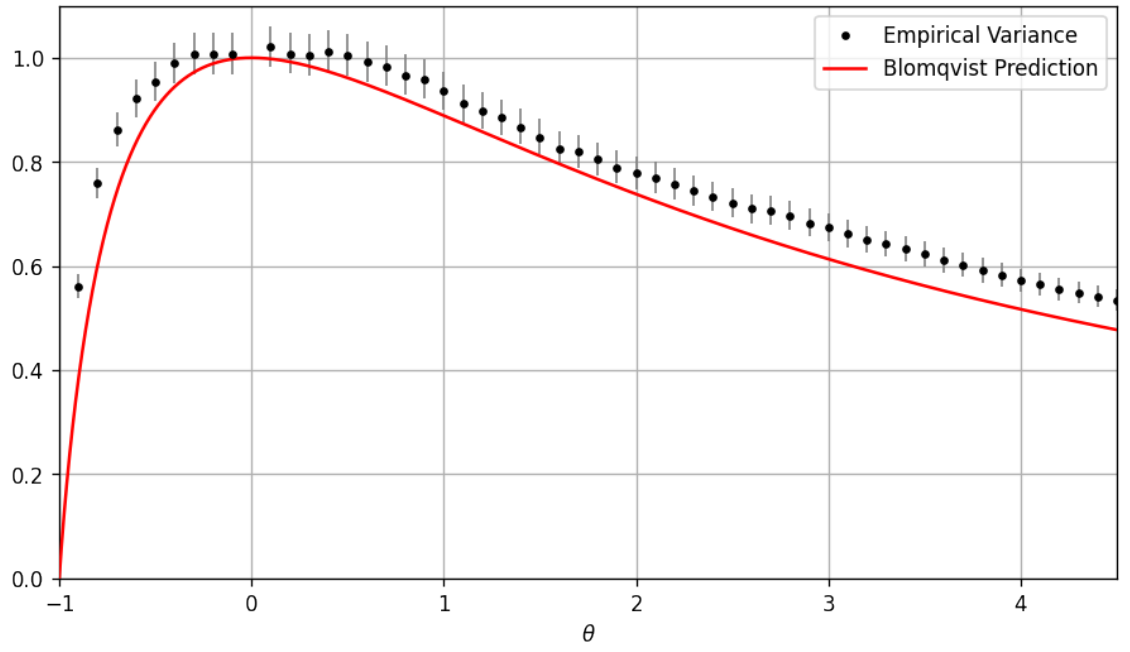

Figure 2.6: Empirical variance for $N = 5000$ calculations of Blomqvist's beta for $n = 1000$ samples each, for different values of the distribution parameter θ. The bars denote the 95% confidence intervals. The red line indicates the value predicted by Blomqvist (2.7)

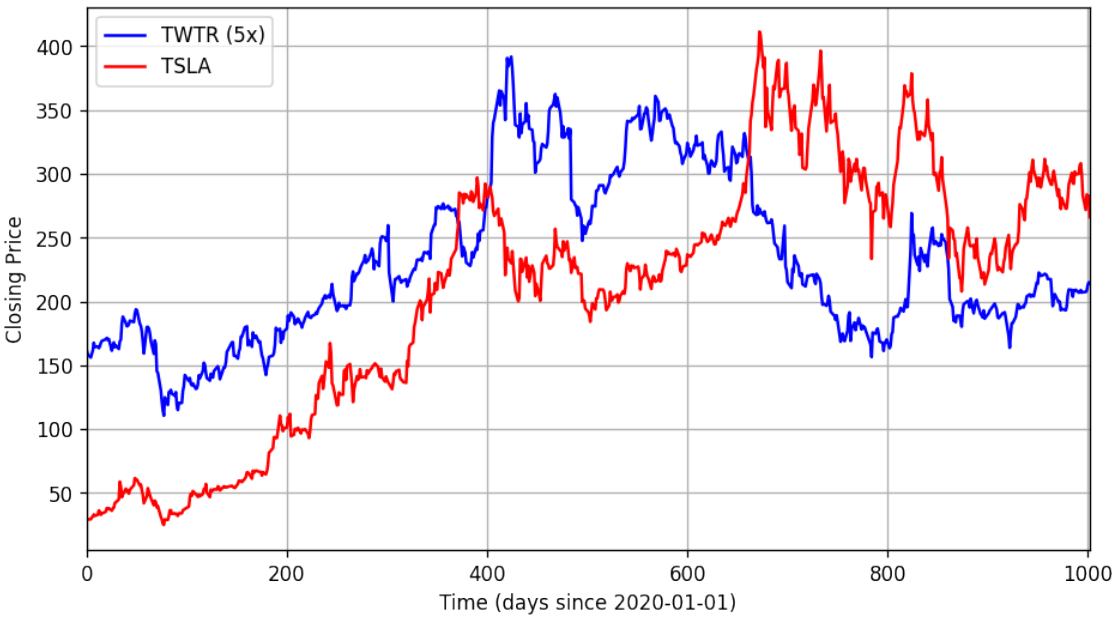

Figure 2.7: Daily closing price of TWTR and TSLA from 2020-01-01 to 2022-10-1.

Under the null hypothesis of independence, we expect (approximately)

$$n_1 \sim \mathcal{N}\left(k, \frac{k^2}{2k-1}\right).$$

So we can perform an approximate test using the statistic

$$Z = \left(\frac{n_1}{k} - 1\right)\sqrt{2k-1} \sim \mathcal{N}(0,1).$$

In the present case, we have $n_1 = 408$, and therefore $z = 4.71$. The associated p-value is

$$p = P(|Z| > 4.71) = 2.5e-6,$$

so that we can reject the null hypothesis with high confidence and conclude that the stock prices are not independent.

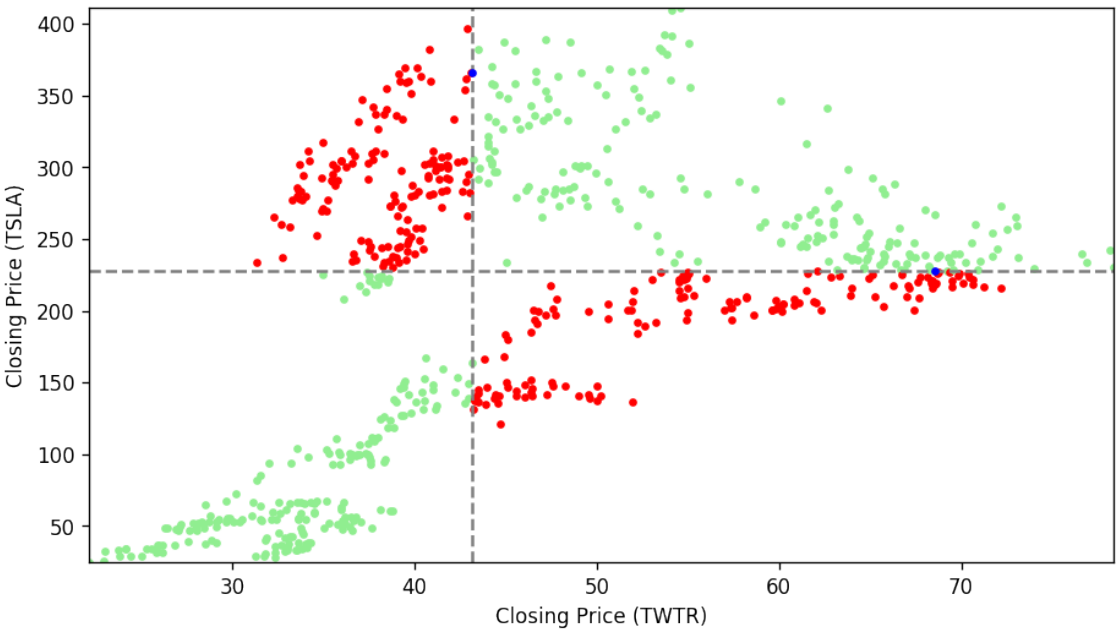

Figure 2.8: Closing prices for TWTR and TSLA from 2020-01-01 to 2022-10-1. The dotted lines indicate the sample medians. The green and red points show how n_1 resp. n_2 is calculated, and the blue points are the ones determining the medians.

Chapter 3

A Brief Introduction to Copulas

Before continuing our study of Blomqvist's beta, we will first introduce Copula Theory, a piece of statistical machinery developed in 1959 by mathematician Abe Sklar to better describe dependence between random variables. This chapter is based on Chapter 2 of Nelson's book *An Introduction to Copulas* [6]. We will only introduce the basic concepts related to copulas that are needed in the next chapters, which implies that a lot of topics are not touched or only very briefly mentioned. For a more in-depth treatment, it is recommended to read Nelson's introduction. The first part of the chapter discusses 2-dimensional copulas and afterwards the concepts are generalized to n dimensions. For simplicity reasons all stochastic variables are assumed to be continuous, but it is possible to define copulas for arbitrary distributions. Only for a few results of the chapter a proof is given, all these proofs are original.

3.1 Copulas in 2 Dimensions

In order to define copulas, we need the concept of a 2-increasing function.

Definition 3.1 (2-increasing function)**.** A function $H : A \times B \subseteq \mathbb{R}^2 \to \mathbb{R}$ is called 2-increasing if for all rectangles $B = [x_1, x_2] \times [y_1, y_2]$ such that $x_1, x_2 \in A$ and $y_1, y_2 \in B$ it holds that

$$H(x_2, y_2) - H(x_1, y_2) - H(x_2, y_1) + H(x_1, y_1) \geq 0. \tag{3.1}$$

Example 3.2. The joint cumulative distribution $F_{X,Y}$ of two stochastic variables X and Y is 2-increasing, since

$$F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_2) - F_{X,Y}(x_2, y_1) + F_{X,Y}(x_1, y_1) = P(x_1 < X \leq x_2, y_1 < Y \leq y_2) \geq 0$$

for all $x_1 < x_2$ and $y_1 < y_2$.

Definition 3.3 (Copula)**.** A two-dimensional copula is a 2-increasing function $C : I^2 \to I$ (where $I = [0, 1]$) that satisfies the following equalities for all $u, v \in I$:

$$\begin{aligned} C(u, 0) &= C(0, v) = 0 \\ C(u, 1) &= u \\ C(1, v) &= v. \end{aligned}$$

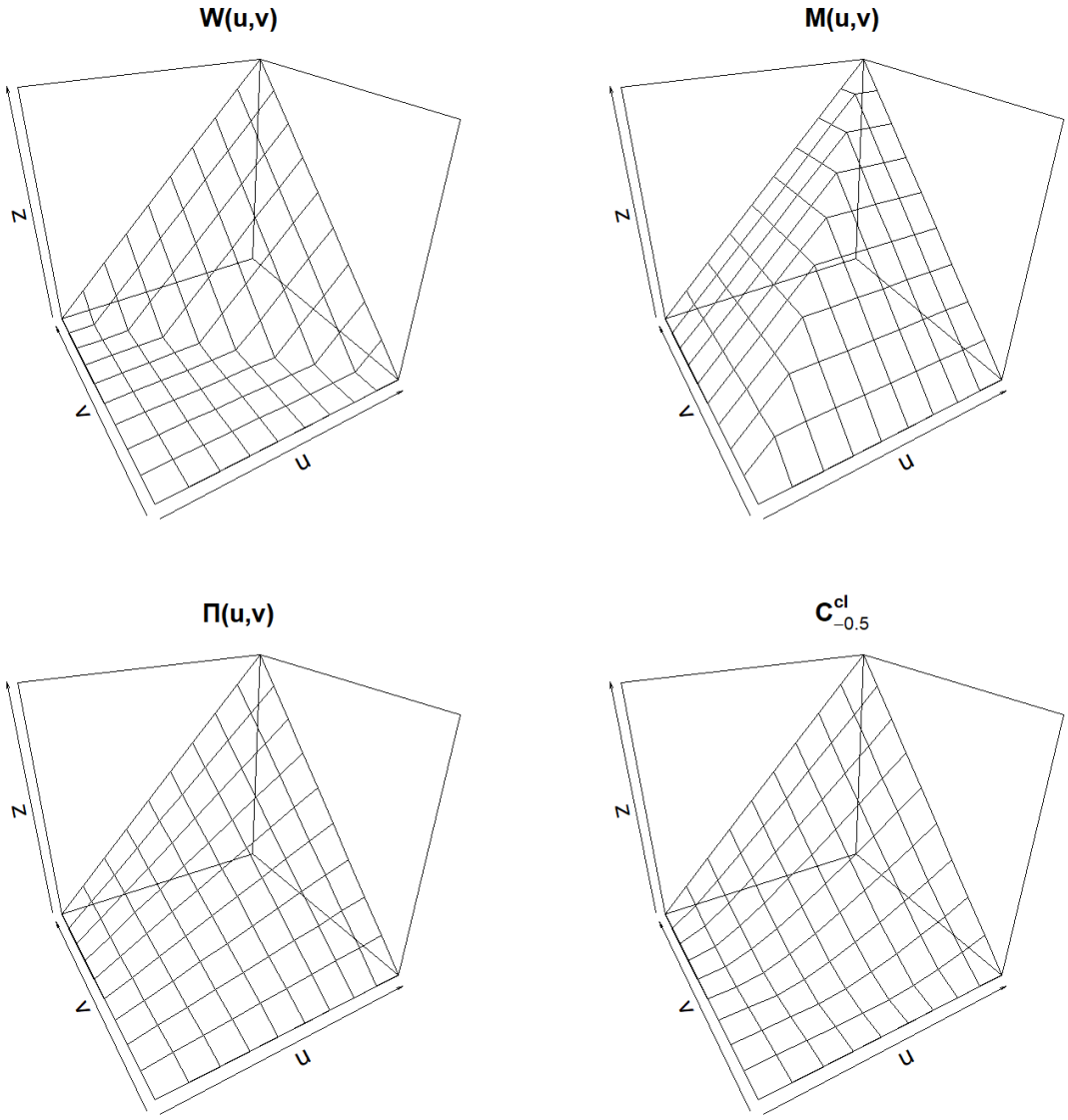

Figure 3.1: The Fréchet-Hoeffding lower- and upper bound copulas, the product copula and the Clayton copula for $\theta = -1/2$

The copulas in the examples below can be seen in Figure 3.1.

Example 3.4. Some important copulas are:

- the product copula $\Pi(u, v) = uv$;
- the Fréchet-Hoeffding lower bound $W(u, v) = \max\{u + v - 1, 0\}$;
- and the Fréchet-Hoeffding upper bound $M(u, v) = \min\{v, u\}$.

Example 3.5. The Clayton family of copulas is given by

$$C_\theta^{cl}(u, v) = \left(\max\left\{u^{-\theta} + v^{-\theta} - 1, 0\right\}\right)^{-1/\theta},$$

where $\theta \in [-1, +\infty) \setminus \{0\}$. For $\theta = -1$ this is the Fréchet-Hoeffding lower bound. For $u > 0, v > 0$ and θ large enough (for example $\theta > \max\{\log_u 2, \log_v 2\}$) we have $u^{-\theta} + v^{-\theta} - 1 > 0$ and hence

$$\lim_{\theta\to 0} C_\theta^{cl}(u, v) = \lim_{\theta\to 0} \left(u^{-\theta} + v^{-\theta} - 1\right)^{-1/\theta} = \exp\left(-\lim_{\theta\to 0} \frac{\ln\left(u^{-\theta} + v^{-\theta} - 1\right)}{\theta}\right).$$

By L'Hôpital's rule we get

$$\lim_{\theta\to 0} C_\theta^{cl}(u, v) = \exp\left(\lim_{\theta\to 0} \frac{(\ln(u)u^{-\theta} + \ln(v)v^{-\theta}}{u^{-\theta} + v^{-\theta} - 1}\right) = uv. \tag{3.2}$$

Hence C_θ^{cl} converges to the product copula for $\theta \to 0$. Similarly to (3.2) it holds that

$$\lim_{\theta\to\infty} C_\theta^{cl}(u, v) = \exp\left(\lim_{\theta\to\infty} \frac{(\ln(u)u^{-\theta} + \ln(v)v^{-\theta}}{u^{-\theta} + v^{-\theta} - 1}\right),$$

and by doing a case analysis for $u > v$, $u < v$ and $u = v$ one can show that this expression equals $\min\{v, u\}$. Thus for $\theta \to \infty$ the Clayton copula converges to the Fréchet-Hoeffding upper bound.

Lemma 3.6. *A copula C is non-decreasing in each of its arguments.*

Proof. Let $x_1, x_2, y \in I$ with $x_1 \leq x_2$, and take $y_1 = 0$ and $y_2 = y$ in (3.1), then

$$C(x_2, y) - C(x_1, y) = C(x_2, y) - C(x_1, y) - C(x_2, 0) + C(x_1, 0) \geq 0.$$

The proof for the second argument is completely analogous. □

Theorem 3.7. *If C is a copula, then for every $u, v \in I$ it holds that $W(u, v) \leq C(u, v) \leq M(u, v)$.*

Proof. Lemma 3.6 gives $C(u, v) \leq C(u, 1) = u$ and $C(u, v) \leq C(1, v) = v$, which proves the upper bound. For the lower-bound we have $C(u, v) \geq C(0, v) = 0$ and by the 2-increasing property $C(u, v) \geq C(u, 1) + C(1, v) - C(1, 1) = u + v - 1$. □

The following theorem lays at the heart of copulas, it was first proven by Sklar in 1959 [9]. It shows that each joint distribution function can be uniquely described by the marginal distributions together with a copula. The copula in essence "couples" the joint distribution function with its marginal distributions, which explains the name "copula". For the proof of the theorem we refer to Nelson [6].

Theorem 3.8 (Sklar's theorem)**.** *Let X and Y be continuous random variables with cumulative distribution functions F_X and F_Y and let H be their joint distribution. Then there exists a unique copula $C_{X,Y}$ such that*

$$H(x,y) = C_{X,Y}\left(F_X(x), F_Y(y)\right), \tag{3.3}$$

for all $x, y \in \mathbb{R}$. The copula $C_{X,Y}$ can be found by inverting F_X and F_Y:

$$C_{X,Y}(u,v) = H\left(F_X^{-1}(u), F_Y^{-1}(v)\right), \tag{3.4}$$

for all $u, v \in (0,1)$. Conversely, if C is a copula and F_X and F_Y are univariate distributions, then $C(F_X(x), F_Y(y))$ is a bivariate distribution with marginals F_X and F_Y.

By extending its domain, a copula C can be interpreted as being the cdf of a bivariate distribution. If we define

$$F_C(u,v) = \begin{cases} C(u,v) & \text{if } u,v \in I \\ 0 & \text{if } u < 0 \text{ or } v < 0 \\ u & \text{if } v > 1 \text{ and } u \in I \\ v & \text{if } u > 1 \text{ and } v \in I \\ 1 & \text{if } u > 1 \text{ and } v > 1 \end{cases},$$

then F_C is a bivariate distribution where the marginals are uniform distributions on I. For two stochastic variables X, Y this distribution $F_{C_{X,Y}}$ captures all information about the dependence between X and Y where all information of the marginal distributions is eliminated.

A direct consequence of Sklar's theorem is the next characterisation of independence of two random variables.

Theorem 3.9. *Two continuous random variables X and Y are independent if and only if $C_{X,Y} = \Pi$.*

3.2 Multivariate Copulas

In order to generalize equation (3.1) in the definition of a 2-increasing function, we define the notion of an H-volume. The notation $[\mathbf{a}, \mathbf{b}]$ is used for the *n-dimensional box* $B = \prod_{i=1}^{n}[a_i, b_i] = [a_1, b_1] \times \ldots \times [a_n, b_n]$.

Definition 3.10 (H-volume)**.** Let $H : \prod_{i=1}^{n} S_i \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function and let $B = [\mathbf{a}, \mathbf{b}]$ be an *n-dimensional box* such that $a_i, b_i \in S_i$ for all i. The *H-volume of B* is defined to be

$$V_H(B) = \Delta_{1\,a_1}^{b_1} \Delta_{2\,a_2}^{b_2} \cdots \Delta_{n\,a_n}^{b_n} H(x_1, \ldots, x_n),$$

where $\Delta_{i\,a_i}^{b_i}$ is the first order difference in the i-component

$$\Delta_{i\,a_i}^{b_i} H(x_1, \ldots, x_i, \ldots, x_n) = H(x_1, \ldots, b_i, \ldots, x_n) - H(x_1, \ldots, a_i, \ldots, x_n).$$

Example 3.11. Let $(X_1, \ldots, X_n)$ be a stochastic vector with cumulative density function F and let $B = [\mathbf{a}, \mathbf{b}]$, then

$$V_F(B) = P(a_1 < X_1 \leq b_1, \ldots, a_n < X_n \leq b_n).$$

Definition 3.12 (n-increasing function)**.** Let $H : \prod_{i=1}^{n} S_i \subseteq \mathbb{R}^n \to \mathbb{R}$ be a function, then H is said to be n-increasing if $V_H([\mathbf{a}, \mathbf{b}]) \geq 0$ for each $\mathbf{a}, \mathbf{b} \in \prod_{i=1}^{n} S_i$ with $\mathbf{a} \leq \mathbf{b}$.

In the 2-dimensional case this corresponds to the definition of a 2-increasing function, since

$$V_H([x_1, y_1] \times [x_2, y_2]) = H(x_2, y_2) - H(x_1, y_2) - H(x_2, y_1) + H(x_1, y_1).$$

Definition 3.13 (n-copula)**.** An n-dimensional copula (or n-copula) is an n-increasing function $C : I^n \to I$ that for all $\mathbf{u} \in I^n$ satisfies

$$\begin{aligned} C(\mathbf{u}) &= 0 && \text{if at least one component of } \mathbf{u} \text{ is zero;} \\ C(\mathbf{u}) &= u_k && \text{if all components of } \mathbf{u} \text{ except } u_k \text{ are 1.} \end{aligned}$$

Example 3.14. As in 2 dimensions the product copula $\Pi(\mathbf{u}) = \prod_{i=1}^{n} u_i$ and the Fréchet-Hoeffding upper bound $M(\mathbf{u}) = \min\{u_1, \ldots, u_n\}$ are copulas for arbitrary n. However, the Fréchet-Hoeffding lower bound $W(\mathbf{u}) = \max\{u_1 + \cdots + u_n - n + 1, 0\}$ is only a copula if $n = 2$.

As the names of M and W suggest, Theorem 3.7 remains valid in arbitrary dimensions.

Theorem 3.15. *If C is a n-dimensional copula, then for every $\mathbf{u} \in I^n$ it holds that $W(\mathbf{u}) \leq C(\mathbf{u}) \leq M(\mathbf{u})$.*

Again Sklar's theorem plays a central role in the theory of copulas.

Theorem 3.16 (Sklar's theorem in n dimensions)**.** *Let $F_1, F_2, \ldots, F_n$ be the cumulative distributions of continuous stochastic variables $X_1, X_2, \ldots, X_n$, and let H be their joint distribution, then there exists a unique copula C such that*

$$H(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)),$$

for all $(x_1, \ldots x_n) \in \mathbb{R}^n$.

As in the 2-dimensional case, the product-copula forms a characterization for independence.

Theorem 3.17. *Let $X_1, X_2, \ldots, X_n$ be continuous stochastic variables, then $X_1, X_2, \ldots, X_n$ are independent if and only if their n-copula is the product copula Π.*

A similar characterization exists for the Fréchet-Hoeffding upper-bound M.

Theorem 3.18. *Let $X_1, X_2, \ldots, X_n$ be continuous stochastic variables, then their n-copula is M if and only if each of the variables is almost surely a strictly increasing function of the others.*

An n-copula C assigns to every n-dimensional box $[\mathbf{a}, \mathbf{b}]$ a non-negative value via $V_C(B)$. In general this is a quite complicated expression in terms of C, but for $\mathbf{a} = \mathbf{0}$ it reduces to $V_C([\mathbf{0}, \mathbf{b}]) = C(\mathbf{b})$. In the next chapter we will see that it is useful the to have a similar expression if $\mathbf{b} = \mathbf{1}$, hence we define the *survival function* $\bar{C}$ to be $\bar{C}(\mathbf{a}) = V_C([\mathbf{a}, \mathbf{1}])$. In two dimensions one has

$$\bar{C}(u, v) = C(1, 1) - C(u, 1) - C(1, v) + C(u, v) = 1 - v - u + C(u, v).$$

In particular $\bar{C}(1/2, 1/2) = C(1/2, 1/2)$.

Chapter 4

Blomqvist's Beta Revisited

The aim of this chapter is to express Blomqvist's beta in terms of copulas. Aside from providing more insight in the behaviour in 2 dimensions, it will be helpful to generalize Blomqvist's beta to $n > 2$ dimensions. It will also lead us to an expression for the asymptotic variance of $\hat{\beta}$ that more closely agrees with the simulation performed in Chapter 2. This chapter is based on [8].

4.1 Bivariate Population

Remember that for a pair (X, Y) of continuous random variables, the population version of the Blomqvist's beta is defined to be

$$\beta = 2\left[P\left(X < \tilde{X}, Y < \tilde{Y}\right) + P\left(X > \tilde{X}, Y > \tilde{Y}\right)\right] - 1 = 4P\left(X < \tilde{X}, Y < \tilde{Y}\right) - 1,$$

where $\tilde{X}$ and $\tilde{Y}$ are the medians of X and Y. Using Sklar's Theorem we have

$$\begin{aligned} &P\left(X < \tilde{X}, Y < \tilde{Y}\right) = F_{X,Y}(F_X^{-1}(1/2), F_Y^{-1}(1/2)) = C_{X,Y}(1/2, 1/2) \text{ and} \\ &P\left(X > \tilde{X}, Y > \tilde{Y}\right) = \bar{C}_{X,Y}(1/2, 1/2), \end{aligned}$$

where $C_{X,Y}$ denotes the copula associated with (X, Y). We obtain

$$\beta = 2\left(C_{X,Y}(1/2, 1/2) + \bar{C}_{X,Y}(1/2, 1/2)\right) - 1 = 4C_{X,Y}(1/2, 1/2) - 1. \tag{4.1}$$

If X and Y are independent, then $C_{X,Y} = \Pi$, hence $\beta = 4\Pi(1/2, 1/2) - 1 = 0$. Conversely, if $\beta = 0$ then X and Y are not necessarily independent, as the following example shows:

Example 4.1. Let $C(u, v) = (W(u, v) + M(u, v))/2$ be the average of the Fréchet-Hoeffding bounds (this is again a copula), then $C(1/2, 1/2) = 1/4 = \Pi(1/2, 1/2)$. Hence any pair of stochastic variables X and Y with this copula has $\beta = 0$, but since $\Pi \neq C$, they are not independent.

4.2 Multivariate Population

We now wish to extend the definition of Blomqvist's beta to the multivariate case. Suppose we have a d-dimensional random variable $\mathbf{X} = (X_1, \ldots, X_d)$. We wish to express the correlation based on the probability that all components are larger or smaller than their respective medians, which is

$$P(\forall i : X_i \leq \tilde{X}_i) + P(\forall i : X_i \geq \tilde{X}_i) = C(\mathbf{1/2}) + \bar{C}(\mathbf{1/2}),$$

where C denotes the copula of $\mathbf{X}$ and $\mathbf{1/2} = (1/2, \ldots, 1/2)$. However, this quantity is not properly normalized, since we need $\beta = 0$ in the case of independence. We therefore alter the formula to

$$C(\mathbf{1/2}) - \Pi(\mathbf{1/2}) + \bar{C}(\mathbf{1/2}) - \bar{\Pi}(\mathbf{1/2}),$$

with Π the independence copula. Finally, we demand $|\beta| \leq 1$. This can be done by using the Fréchet-Hoeffding bounds, leading to the following definition:

Definition 4.2 (multivariate Blomqvist's beta)**.** Let C be the copula of a random vector $\mathbf{X} = (X_1, ..., X_d)$. The multivariate version of Blomqvist's beta is defined by

$$\beta := \frac{C(\mathbf{1/2}) - \Pi(\mathbf{1/2}) + \bar{C}(\mathbf{1/2}) - \bar{\Pi}(\mathbf{1/2})}{M(\mathbf{1/2}) - \Pi(\mathbf{1/2}) + \bar{M}(\mathbf{1/2}) - \bar{\Pi}(\mathbf{1/2})} = h_d\{C(\mathbf{1/2}) + \bar{C}(\mathbf{1/2}) - 2^{1-d}\},$$

where $h_d = 2^{d-1}/(2^{d-1} - 1)$. The second equality follows from the fact that $M(\mathbf{1/2}) = \bar{M}(\mathbf{1/2}) = 1/2$ and $\Pi(\mathbf{1/2}) = \bar{\Pi}(\mathbf{1/2}) = 2^{-d}$. For $d = 2$ this expression can be simplified to (4.1) by using $\bar{C}(\mathbf{1/2}) = C(\mathbf{1/2})$. For $d \geq 3$ this is not possible, because then $\bar{C}(\mathbf{1/2}) = C(\mathbf{1/2})$ does not necessarily hold.

From the above definition, it is clear that $\beta = 0$ is in no way sufficient for independence. The coefficient β is only dependent on the copula and survival function in a single point, so that $\beta = 0$ is satisfied by any copula C with $C(\mathbf{1/2}) + \bar{C}(\mathbf{1/2}) = 2^{1-d}$.

4.3 Sample Version (known margins)

Suppose we are interested in estimating β for a d-dimensional random vector $\mathbf{X}$ with cdf F, copula C and marginal distributions F_j, and we have a random sample $(\mathbf{X}_i)_{i=1,\ldots,n}$ at our disposal. Furthermore, we treat the marginal distributions F_j as given, so that $U_{ij} = F_j(X_{ij})$ is uniform for each i, j. The natural estimators for the copula terms are

$$\widehat{C(\mathbf{1/2})} = \frac{1}{n}\#\left\{i \mid \forall j : U_{ij} \leq 1/2\right\} \qquad \text{and} \qquad \widehat{\bar{C}(\mathbf{1/2})} = \frac{1}{n}\#\left\{i \mid \forall j : U_{ij} \geq 1/2\right\},$$

which can be rewritten to yield the following estimator for β:

$$\hat{\beta}_n = h_d\left(\frac{1}{n}\sum_{i=1}^{n}\left(\prod_{j=1}^{d}\mathbb{1}_{\{U_{ij}\leq 1/2\}} + \prod_{j=1}^{d}\mathbb{1}_{\{U_{ij}\geq 1/2\}}\right) - 2^{1-d}\right). \tag{4.2}$$

It is clear that this estimator is consistent for β. For each i, the random variable

$$\prod_{j=1}^{d}\mathbb{1}_{\{U_{ij}\leq 1/2\}} + \prod_{j=1}^{d}\mathbb{1}_{\{U_{ij}\geq 1/2\}}$$

has a Bernouilli distribution with probability of success given by $C(\mathbf{1/2}) + \bar{C}(\mathbf{1/2})$. By the CLT, the average in expression (4.2) is asymptotically normal with variance

$$\frac{1}{n}\left(C(\mathbf{1/2}) + \bar{C}(\mathbf{1/2}) - \left(C(\mathbf{1/2}) + \bar{C}(\mathbf{1/2})\right)^2\right),$$

and therefore we have the following asymptotic normality result:

Lemma 4.3. *The estimator $\hat{\beta}_n$ satisfies*

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \to \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = 2^{2d-2}\left(C(\mathbf{1/2}) + \bar{C}(\mathbf{1/2}) - \left(C(\mathbf{1/2}) + \bar{C}(\mathbf{1/2})\right)^2\right) / \left(2^{d-1} - 1\right)^2$.

Resuming notation from earlier, in the two-dimensional case one has $d = 2$ and $C(\mathbf{1/2}) = \bar{C}(\mathbf{1/2}) = a_0$, so that

$$\sigma^2 = 8a_0(1 - 2a_0).$$

By comparing with the corresponding result in Chapter 2, this is seen to be the asymptotic variance Blomqvist claims for $\hat{\beta}$. However, in this derivation we have assumed the population medians to be known and used in the estimator $\hat{\beta}$, whereas Blomqvist's original estimator uses the sample medians of the observation. In the next section, we will see that this introduces an increase in variance that was not accounted for in Blomqvist's calculation.

4.4 Sample Version (unknown margins)

When $\hat{\beta}$ was introduced in Chapter 2, we did not assume the population medians to be known. In general, we cannot assume F_i to be known and these need to be estimated by using their sample counterparts

$$\hat{F}_i(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{\{X_{ij} \le x\}},$$

and we will work with $\hat{U}_{ij} = \hat{F}_i(X_j)$ to construct $\hat{\beta}$ using the *empirical copulas*

$$C_n(\mathbf{1/2}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} 1_{\{\hat{U}_{ij} \le 1/2\}} \qquad \text{and} \qquad \bar{C}_n(\mathbf{1/2}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} 1_{\{\hat{U}_{ij} \ge 1/2\}},$$

where we then define

$$\hat{\beta}_n = h_d\left(C_n(\mathbf{1/2}) + \bar{C}_n(\mathbf{1/2}) - 2^{1-d}\right).$$

The asymptotic behaviour of this estimator is derived in [8]. The details of this result are beyond the scope of this paper, but we will state the result here:

Lemma 4.4. *If the i-th partial derivatives D_iC and $D_i\bar{C}$ exist and are continuous at the point $\mathbf{1/2}$, then $\hat{\beta}_n$ satisfies*

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \tag{4.3}$$

where $\sigma^2 = E\{\mathbb{G}(\mathbf{1/2}, \mathbf{1/2})^2\}$ *and*

$$\mathbb{G}(\mathbf{u}, \mathbf{v}) = h_d(\mathbf{u}, \mathbf{v}) \left[\mathbb{B}_C(\mathbf{u}) + \mathbb{B}_{\bar{C}}(\mathbf{v}) - \sum_{i=1}^{d} \left\{ D_i C(\mathbf{u}) \mathbb{B}_C \left(\mathbf{u}^{(i)} \right) + D_i \bar{C}(\mathbf{v}) \mathbb{B}_C \left(\mathbf{v}^{(i)} \right) \right\} \right],$$

with $\mathbf{u}^{(i)} := (1, ..., 1, u_i, 1, ..., 1)$.

Here, $\mathbb{B}_C$ *and* $\mathbb{B}_{\bar{C}}$ *are centered tight Gaussian processes on* $[0,1]^d$ *with covariance functions* $E[\mathbb{B}_C(\mathbf{u_1})\mathbb{B}_C(\mathbf{u_2})] = C(\mathbf{u_1} \wedge \mathbf{u_2}) - C(\mathbf{u_1})C(\mathbf{u_2})$ *and* $E[\mathbb{B}_{\bar{C}}(\mathbf{u_1})\mathbb{B}_{\bar{C}}(\mathbf{u_2})] = \bar{C}(\mathbf{u_1} \vee \mathbf{u_2}) - \bar{C}(\mathbf{u_1})\bar{C}(\mathbf{u_2})$, *where* $\mathbf{u_1} \wedge \mathbf{u_2}$ *and* $\mathbf{u_1} \vee \mathbf{u_2}$ *are respectively the components wise minimum and maximum of* $\mathbf{u_1}$ *and* $\mathbf{u_2}$. *Furthermore,* $\mathbb{B}_C(\mathbf{u})$ *and* $\mathbb{B}_{\bar{C}}(\mathbf{v})$ *are jointly normally distributed with covariance*

$$E[\mathbb{B}_C(\mathbf{u})\mathbb{B}_{\bar{C}}(\mathbf{v})] = -C(\mathbf{u})\bar{C}(\mathbf{v}).$$

In Appendix A we use the expressions in the theorem to explicitly calculate σ^2. In the two-dimensional case ($d = 2$) we get

$$\begin{aligned} \sigma^2 &= 4 \left(D_1 C(\mathbf{1/2}) - D_2 C(\mathbf{1/2}) \right)^2 \\ &\quad + 16 C(\mathbf{1/2}) \left[2 D_1 C(\mathbf{1/2}) D_2 C(\mathbf{1/2}) - D_1 C(\mathbf{1/2}) - D_2 C(\mathbf{1/2}) + 1 - C(\mathbf{1/2}) \right]. \end{aligned}$$

We can now look back at our example of the Clayton copula and use the above expression to calculate the variance of this specific copula.

Lemma 4.5. *For the Clayton Copula, the asymptotic variance of* $\hat{\beta}$ (4.3) *is given by*

$$\sigma^2 = 8 h_\theta \left(1 - 2 h_\theta + \left(4 h_\theta^{\theta+1} 2^\theta - 1 \right)^2 \right), \tag{4.4}$$

where $h_\theta = \left(2^{\theta+1} - 1 \right)^{-1/\theta}$.

Proof. For the Clayton copula we have:

$$\begin{aligned} C(\mathbf{1/2}) &= (2^{\theta+1} - 1)^{-1/\theta} = h_\theta \\ D_1 C(\mathbf{1/2}) &= D_2 C(\mathbf{1/2}) = h_\theta^{\theta+1} 2^{\theta+1}. \end{aligned}$$

Substituting this into our expression for σ^2 gives:

$$\begin{aligned} \sigma^2 &= 16 h_\theta \left[2 (h_\theta^{\theta+1} 2^{\theta+1})^2 - 2 (h_\theta^{\theta+1} 2^{\theta+1}) + 1 - h_\theta \right] \\ &= 8 h_\theta \left[2 - 2 h_\theta + (4 h_\theta^{\theta+1} 2^\theta)^2 - 8 (h_\theta^{\theta+1} 2^\theta) \right] \\ &= 8 h_\theta \left(1 - 2 h_\theta + \left(4 h_\theta^{\theta+1} 2^\theta - 1 \right)^2 \right), \end{aligned}$$

which is precisely (4.4). □

Figure 4.1 shows that the value obtained by this formula agrees with the results of the simulation study performed in Chapter 2. Indeed, for every value of θ tested, the predicted variance (4.4) lies within the 95% confidence region obtained by the simulation.

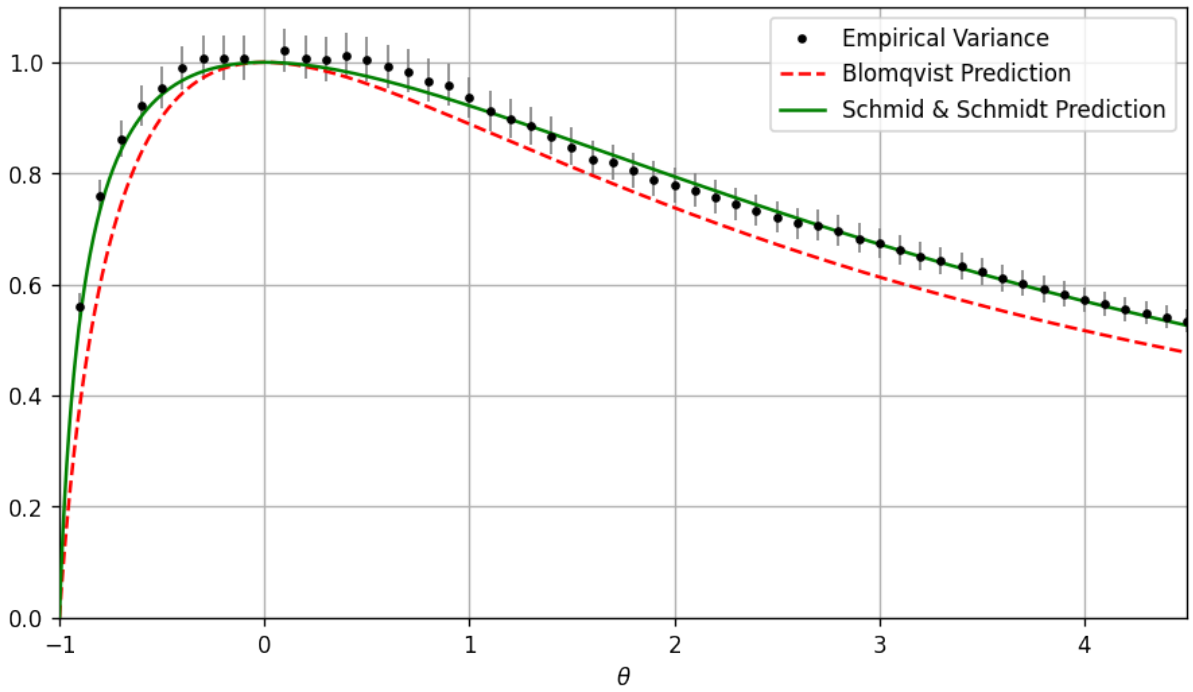

Figure 4.1: Variances obtained from the simulation (see Figure 2.6) along with the value predicted by Schmid and Schmidt (4.4). It is clear that the Schmid and Schmidt prediction lies within the 95% confidence interval for every θ, and that the prediction of Blomqvist only agrees when $\theta \to 0$ or $\theta = -1$.

Conclusion

Although Pearson's correlation coefficient ρ is ubiquitous in introductory statistics texts, there is a large variety of different correlation measures to be studied. In dissecting Blomqvist's beta, it became clear that even a simple definition based on the sample medians can lead to interesting properties and asymptotic behaviour.

Indeed, the asymptotic variance of Blomqvist's original paper [1] turned out to be incorrect, a fact that was only discovered recently by Schmid and Schmidt [8]. By means of a simulation study, we have given numerical proof for the proposed adjustment to the formula.

The study of correlation has advanced much since Blomqvist's 1950 paper, and the theory of copulas certainly deserves recognition in this regard. It allows us to fully capture relations between variables while disregarding the marginal distributions, and we have shown that it forms a natural environment to think about what Blomqvist's coefficient could mean in a more generalized context.

It is worth mentioning however, that we have limited ourselves to one specific correlation coefficient, and there exist other coefficients in many shapes and sizes. We refer the interested reader to the overview given in [4], where a general definition of dependence measures is also given. The section on correlation in [6] would also be of interest to readers wanting to know more about applications of copulas in this context.

Appendix

Appendix A: Derivation of the Asymptotic Variance of $\hat{\beta}$

Lemma 4.4 gives an expression for the asymptotic variance of $\hat{\beta}$ in terms of tight Gaussian processes. As an original result we will explicitly calculate the value of σ^2. We need the following expressions based on the lemma (see also Corollary 3 in [8]):

$$
\begin{aligned}
&E\left[\mathbb{B}_C(\mathbf{1/2})\mathbb{B}_C(\mathbf{1/2})\right] = C(\mathbf{1/2} \wedge \mathbf{1/2}) - C(\mathbf{1/2})C(\mathbf{1/2}) = C(\mathbf{1/2}) - C(\mathbf{1/2})^2, \\
&E\left[\mathbb{B}_C(\mathbf{1/2})\mathbb{B}_C\left(\mathbf{1/2}^{(i)}\right)\right] = C(\mathbf{1/2} \wedge \mathbf{1/2}^{(i)}) - C(\mathbf{1/2})C(\mathbf{1/2}^{(i)}) = \frac{1}{2}C(\mathbf{1/2}), \\
&E\left[\mathbb{B}_C\left(\mathbf{1/2}^{(i)}\right)\mathbb{B}_C\left(\mathbf{1/2}^{(j)}\right)\right] = C(\mathbf{1/2^{(i)}} \wedge \mathbf{1/2^{(j)}}) - C(\mathbf{1/2^{(i)}})C(\mathbf{1/2^{(j)}}) \\
&\qquad\qquad = \begin{cases} \frac{1}{2} - \left(\frac{1}{2}\right)^2 & i = j \\ C_{i,j}\left(\frac{1}{2}, \frac{1}{2}\right) - \left(\frac{1}{2}\right)^2 & i \neq j \end{cases}, \\
&E\left[\mathbb{B}_C(\mathbf{1/2})\mathbb{B}_{\bar{C}}(\mathbf{1/2})\right] = -C(\mathbf{1/2})\bar{C}(\mathbf{1/2}), \\
&E\left[\mathbb{B}_C\left(\mathbf{1/2}^{(i)}\right)\mathbb{B}_{\bar{C}}(\mathbf{1/2})\right] = -C\left(\mathbf{1/2^{(i)}}\right)\bar{C}(\mathbf{1/2}) = -\frac{1}{2}\bar{C}(\mathbf{1/2}).
\end{aligned}
$$

Here $C_{i,j}$ is the marginal copula for the i-th and j-th components.

Using these expressions, we can calculate σ^2:

$$
\begin{aligned}
\sigma^2 &= E\left[\mathbb{G}(\mathbf{1/2}, \mathbf{1/2})^2\right] \\
&= h_d(\mathbf{1/2}, \mathbf{1/2})^2 \Bigg[E\left[\mathbb{B}_C(\mathbf{1/2})^2\right] + E\left[\mathbb{B}_{\bar{C}}(\mathbf{1/2})^2\right] \\
&\quad + E\left[\left(\sum_{i=1}^{d} \mathbb{B}_C\left(\mathbf{1/2}^{(i)}\right)\left(D_i C(\mathbf{1/2}) + D_i \bar{C}(\mathbf{1/2})\right)\right)^2\right] \\
&\quad + 2E\left[\mathbb{B}_C(\mathbf{1/2})\mathbb{B}_{\bar{C}}(\mathbf{1/2})\right] - 2E\left[\mathbb{B}_C(\mathbf{1/2})\sum_{i=1}^{d} \mathbb{B}_C\left(\mathbf{1/2}^{(i)}\right)\left(D_i C(\mathbf{1/2}) + D_i \bar{C}(\mathbf{1/2})\right)\right] \\
&\quad - 2E\left[\mathbb{B}_{\bar{C}}(\mathbf{1/2})\sum_{i=1}^{d} \mathbb{B}_C\left(\mathbf{1/2}^{(i)}\right)\left(D_i C(\mathbf{1/2}) + D_i \bar{C}(\mathbf{1/2})\right)\right]\Bigg]
\end{aligned}
$$

$$
\begin{aligned}
&= h_d(\mathbf{1/2},\mathbf{1/2})^2 \Bigg[\left(C(\mathbf{1/2}) - C(\mathbf{1/2})^2\right) + \left(\bar{C}(\mathbf{1/2}) - \bar{C}(\mathbf{1/2})^2\right) + 2\left(-C(\mathbf{1/2})\bar{C}(\mathbf{1/2})\right) \\
&\quad + E\left[\sum_{i=1}^{d} \mathbb{B}_C\left(\mathbf{1/2}^{(i)}\right) \mathbb{B}_C\left(\mathbf{1/2}^{(i)}\right) \left(D_i C(\mathbf{1/2}) + D_i \bar{C}(\mathbf{1/2})\right)^2\right] \\
&\quad + 2E\left[\sum_{i<j} \mathbb{B}_C\left(\mathbf{1/2}^{(i)}\right) \mathbb{B}_C\left(\mathbf{1/2}^{(j)}\right) \left(D_i C(\mathbf{1/2}) + D_i \bar{C}(\mathbf{1/2})\right) \left(D_j C(\mathbf{1/2}) + D_j \bar{C}(\mathbf{1/2})\right)\right] \\
&\quad - 2\sum_{i=1}^{d} \frac{1}{2} C(\mathbf{1/2}) \left(D_i C(\mathbf{1/2}) + D_i \bar{C}(\mathbf{1/2})\right) + 2\sum_{i=1}^{d} \frac{1}{2} \bar{C}(\mathbf{1/2}) \left(D_i C(\mathbf{1/2}) + D_i \bar{C}(\mathbf{1/2})\right) \Bigg] \\
&= h_d(\mathbf{1/2},\mathbf{1/2})^2 \Bigg[\left(C(\mathbf{1/2}) + \bar{C}(\mathbf{1/2})\right) - \left(\bar{C}(\mathbf{1/2}) + C(\mathbf{1/2})\right)^2 + \sum_{i=1}^{d} \frac{1}{4} \left(D_i C(\mathbf{1/2}) + D_i \bar{C}(\mathbf{1/2})\right)^2 \\
&\quad + 2\sum_{i<j} \left(C_{i,j}\left(\frac{1}{2},\frac{1}{2}\right) - \frac{1}{4}\right) \left(D_i C(\mathbf{1/2}) + D_i \bar{C}(\mathbf{1/2})\right) \left(D_j C(\mathbf{1/2}) + D_j \bar{C}(\mathbf{1/2})\right) \\
&\quad + \left(\bar{C}(\mathbf{1/2}) - C(\mathbf{1/2})\right) \sum_{i=1}^{d} \left(D_i C(\mathbf{1/2}) + D_i \bar{C}(\mathbf{1/2})\right) \Bigg].
\end{aligned}
$$

In case of $d = 2$, we have $\bar{C}(u,v) = 1 - u - v + C(u,v)$ hence $D_1\bar{C}(\mathbf{1/2}) = D_1 C(\mathbf{1/2}) - 1$ and $D_2\bar{C}(\mathbf{1/2}) = D_2 C(\mathbf{1/2}) - 1$. Together with $C(\mathbf{1/2}) = \bar{C}(\mathbf{1/2}) = C_{1,2}(\frac{1}{2},\frac{1}{2})$ and $h_d = 2^{d-1}/(2^{d-1}-1) = 2$ the expression for σ^2 simplifies to:

$$
\begin{aligned}
\sigma^2 &= 4\Big[\left(2C(\mathbf{1/2}) - 4C(\mathbf{1/2})^2\right) + \frac{1}{4}\left[(2D_1 C(\mathbf{1/2}) - 1)^2 + (2D_2 C(\mathbf{1/2}) - 1)^2\right] \\
&\qquad + 2\left(C(\mathbf{1/2}) - \frac{1}{4}\right)(2D_1 C(\mathbf{1/2}) - 1)(2D_2 C(\mathbf{1/2}) - 1)\Big] \\
&= 8C(\mathbf{1/2}) - 16C(\mathbf{1/2})^2 + \left[(2D_1 C(\mathbf{1/2}) - 1) - (2D_2 C(\mathbf{1/2}) - 1)\right]^2 \\
&\qquad + 8C(\mathbf{1/2})(2D_1 C(\mathbf{1/2}) - 1)(2D_2 C(\mathbf{1/2}) - 1) \\
&= 4\left(D_1 C(\mathbf{1/2}) - D_2 C(\mathbf{1/2})\right)^2 \\
&\qquad + 16C(\mathbf{1/2})\left[2D_1 C(\mathbf{1/2}) D_2 C(\mathbf{1/2}) - D_1 C(\mathbf{1/2}) - D_2 C(\mathbf{1/2}) + 1 - C(\mathbf{1/2})\right].
\end{aligned}
$$

Bibliography

[1] Nils Blomqvist. "On a Measure of Dependence Between two Random Variables". English. In: *The Annals of mathematical statistics* 21.4 (1950), pp. 593–600. ISSN: 0003-4851.

[2] Harald Cramér. *Mathematical methods of statistics*. English. Vol. 9. Princeton Math. Ser. Princeton University Press, Princeton, NJ, 1946.

[3] Maurice Kendall and Jean Dickinson Gibbons. *Rank correlation methods*. English. 5th ed. London: Edward Arnold, a div. of Hodder &— Stoughton, 1990. ISBN: 0-85264-305-5.

[4] William H. Kruskal. "Ordinal Measures of Association". English. In: *Journal of the American Statistical Association* 53.284 (1958), pp. 814–861. ISSN: 0162-1459.

[5] I. S. Lunev and V. V. Neknitkin. "A Remark on Certain Classic Criteria of Mathematical Statistics". eng. In: *Vestnik, St. Petersburg University. Mathematics* 52.2 (2019), pp. 154–161. ISSN: 1063-4541.

[6] Roger B. Nelsen. *An Introduction to Copulas*. English. 2nd ed. 2006. Springer Series in Statistics. New York, NY: Springer New York, 2006. ISBN: 1-280-93833-1.

[7] N. A. Rahman. *A Course in Theoretical Statistics*. Charles Griffin and Company, 1968, 1968.

[8] Friedrich Schmid and Rafael Schmidt. "Nonparametric inference on multivariate versions of Blomqvist's beta and related measures of tail dependence". English. In: *Metrika* 66.3 (2007), pp. 323–354. ISSN: 0026-1335. DOI: 10.1007/s00184-006-0114-3.

[9] Abe Sklar. "Fonctions de répartition à n dimensions et leurs marges". French. In: *Publ. Inst. Statist. Univ. Paris* 8 (1959), pp. 229–231.

Department of Mathematics
Celestijnenlaan 200B
B-3001 Leuven (Heverlee)
tel. + 32 16 32 70 07
fax + 32 16 32 79 98
www.kuleuven.be