Noname manuscript No.
(will be inserted by the editor)

A dependence measure flow tree through Monte Carlo simulations

Emanuela Raffinetti · Pier Alda Ferrari

Received: date / Accepted: date

Abstract In applied psychological, behavioral and sociological research the majority of data are typically mixed (continuous and discrete) or, if continuous, they violate the normality condition. Given a dependent and an independent variables: a) both the variables may appear with distinct values (continuous variables); b) the dependent variable may present distinct values (continuous variable) and the independent variable tied values (discrete variable); c) the dependent variable may present tied values (discrete variable) and the independent variable distinct values (continuous variable). The dependence relationship between the variables could be assessed through the common correlation coefficients, i.e., the Pearson's, Spearman's and Kendall's coefficients, jointly with a recently revisited monotonic dependence coefficient, called "Monotonic Dependence Coefficient". But, the choice of the most suitable dependence measure in different scenarios may become problematic.
The aim of the paper is to show which dependence measure to use to discover dependence relationships. A flow tree displaying how to find the best dependence measures is proposed by means of a Monte Carlo simulation study. Both Normal and non-Normal distributions producing continuous and discrete data, together with the possibility of transforming discrete data into continuous ones, are considered. Finally, validation of simulation findings on real data is also introduced.

Keywords normal and non-normal distributed data · mixed data · dependence coefficient · "continuous-ation" approach

Emanuela Raffinetti
Department of Economics, Management and Quantitative Methods, University of Milan, Via Conservatorio 7, 20122 Milan, Italy
Tel.: +39 02 503 21531
E-mail: emanuela.raffinetti@unimi.it

Pier Alda Ferrari
Department of Economics, Management and Quantitative Methods, University of Milan, Via Conservatorio 7, 20122 Milan, Italy

1 Introduction and literature review

Dependency relationships take place in various scenarios of our daily life, such as for instance, in natural, sociological, educational, medical and several further contexts. In social science research, applied scholars and methodologists became familiar with coefficients for studying dependence relationships (see e.g., Mari and Kotz, 2001). Given two variables, X and Y whose underlying joint distribution is Normal, the Pearson's correlation coefficient (Pearson 1907) appears as the most suitable index to measure the dependence relationship between two variables, as it corresponds to a parameter of a bivariate Normal distribution.

In psychological research, the available data may make the use of the Pearson's correlation coefficient problematic. The main drawbacks may occur in two situations: 1) at least one of the two distributions has a discrete nature or 2) one or both of the two distributions are continuous but non-Normal (see e.g. Blanca et al. 2013 and more recently Bishara et al. 2018). In the former case, categorizing one measurement variable tends to reduce the magnitude of the Pearson's coefficient (see e.g. Bedrickt 1995). In the latter case, being the Pearson's coefficient a measure of linear dependence, it might be not the optimal choice. In any case, the awareness of the partial inadequacy of the Pearson's coefficient led researchers to introduce new measures or specific adjustments when dealing with mixed and non-normally distributed data.

Regarding the first issue, some authors propose the use of the Spearman's (Spearman 1904) or Kendall's (Kendall 1938) correlation coefficients. More recently, Denuit and Lambert (2005) suggest to transform discrete variables into continuous variables, by simply adding a uniform random component to the discrete variable categories before the calculation of the indices.

In presence of non-normally distributed data, Bishara and Hittner (2015) use the Pearson's coefficient to assess the dependence relationship after resorting to some ad-hoc transformations able to convert the continuous sample distribution into distributions whose shape is approximately Normal. Some other researchers exploit bootstrap procedures to reach an unbiased estimation of the linear dependence relationship (for a review, see e.g. Rasmussen 1987; Sideridis and Simos 2010).

This paper presents a further extension in the field of dependence analysis when mixed and non-normal data are involved. We start from the previous concerns on the Pearson's correlation coefficient and wonder how to assess a general monotonic dependence relationship between two observed variables, Y and X, looking for an index which can work also when the Pearson's coefficient partially fails. Recently, a novel non-parametric index, called "Monotonic Dependence Coefficient" (henceforth denoted with MDC), suitable in catching any monotonic dependence relationship between two variables, was introduced by Ferrari and Raffinetti (2015). It fulfills some interesting properties: it detects any monotonic dependence relationship and can be also used for discrete dependent or independent variables, appearing as an attractive alternative to the Pearson's, Spearman's and Kendall's correlation coefficients.

Here, we aim at drawing up a flow tree representing the procedures for detecting the most suitable dependence measures, i.e. measures able to catch the actual dependence relationship between the variables, even when some pieces of information in data are lost or latent. This is obtained through a Monte Carlo simulation study, by generating normally and non-normally distributed variables and accounting for the multiple scenarios which may arise in the data collection process (that is, when both variables are continuous or when only one of the two variables is continuous and the other is recorded on the discrete scale).
The paper is organized as follows. We first recall the main features of the Pearson's, Spearman's, Kendall's and MDC coefficients. Due to the concerns about these coefficients when at least one of the two variables is expressed on discrete scale and the data are tied, we resort also to the approach of Denuit and Lambert (2005), based on the transformation of the discrete variable into continuous variable. After, we also introduce a statistical test to evaluate whether the difference in the Monte Carlo values of the considered coefficients is significant. Consequently, a flow tree for the choice of the most appropriate index in relation to the different scenarios is provided. The paper ends with an application to real data and with some brief conclusions.

2 Proposal

2.1 Dependence measures

In applied research, a care analysis of the data to be examined is required in order to select the most suitable measure for the assessment of dependence relationships. Given a dependent (Y) and an independent (X) variables, the usual scenarios which may arise are:

- scenario a): both the variables have distinct values. We refer to this as scenario with continuous variables;
- scenario b): only the dependent variable has distinct values, while for the independent variable tied data are observed. We call this situation continuous/discrete scenario;
- scenario c): only the dependent variable has tied data, while the independent variable has distinct values. This scenario is defined as discrete/continuous scenario.

If data are continuous, as in scenario a), the Pearson's (r), Spearman's (r_S), Kendall's (τ) coefficients are computed by resorting to the following formulas:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{for } i = 1, \ldots, n, \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the sample means,

$$r_S = 1 - \frac{6\sum_{i=1}^{n}(r(x_i)-r(y_i))^2}{n(n^2-1)}, \qquad \text{for } i=1,\dots,n, \tag{2}$$

where $r(x)$ and $r(y)$ are the ranks of the values of X and Y

$$\tau = \frac{2(C-D)}{n(n-1)}, \tag{3}$$

where C and D are the number of concordant and discordant[1] pairs of Y and X.

The MDC coefficient, proposed by Ferrari and Raffinetti (2015), is formulated by comparing the distance between the so-called concordance curve and the Lorenz curve (see, e.g. Lorenz 1905). Specifically, given a dependent variable Y and an independent variable X, the Lorenz curve is obtained by ordering the Y values according to the corresponding non-decreasing ranks, while the concordance curve is built by ordering the Y values according to the non-decreasing ranks of X. If Y is a real-valued variable, in order to build a Lorenz curve lying within the unit square, the transformation $Y^t = Y - y_0$, where $y_0 = min(0, y_{min})$ and y_{min} is the lowest negative value of Y, is taken into account. The L_Y Lorenz curve corresponds to the set of points $(i/n, \sum_{j=1}^{i} y_{(J)}^t / \sum_{i=1}^{n} y_{(J)}^t)$ and the C concordance curve is given by the set of points $(i/n, \sum_{j=1}^{i} y_{(i)}^{t*} / \sum_{i=1}^{n} y_{(i)}^t)$, where $i = 1,\dots,n$, $j = 1,\dots,i$, y^{t*}'s are the y^t's values ordered according to the ranks of variable X and $y_{(\cdot)}^t$'s are the y^t's values ordered in non-decreasing sense[2]. A graphical representation of the two curves, together with the $L_Y^{'}$ dual Lorenz curve, obtained by reordering the Y variable values in non-increasing sense, is provided in Figure 1.

The MDC coefficient is defined as the ratio between the sum of the distances between the points lying on the bisector curve and the those lying on the concordance curve and the sum of the distances between the points lying on the bisector curve and those lying on the Lorenz curve. Through some mathematical manipulations (see, Ferrari and Raffinetti 2015), the MDC coefficient can be expressed as

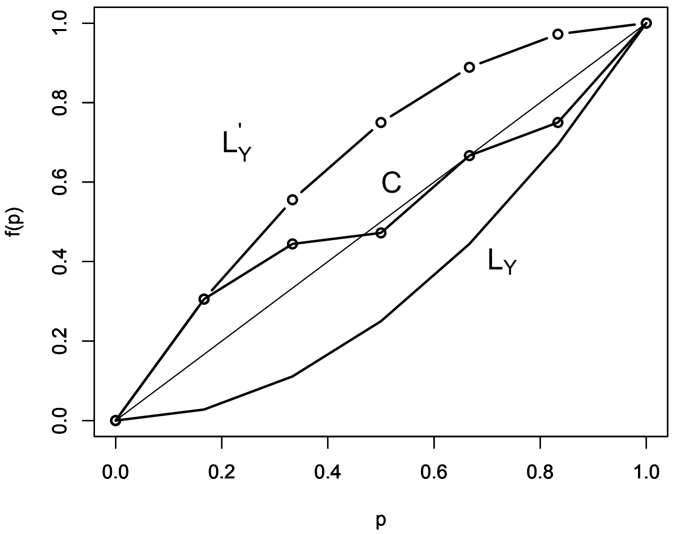
$$MDC = \frac{2\sum_{i=1}^{n} i y_i^{t*} - n(n+1)(M_Y - y_0)}{2\sum_{i=1}^{n} i y_{(i)}^{t} - n(n+1)(M_Y - y_0)}. \tag{4}$$

The MDC coefficient appears as a map that measures the co-variation (association) between the variables. Thus, it presents some similarities with

[1] Note that a pair of observations is said to be concordant if, given two variables Y and X, the observation with larger value of Y corresponds to the observation with larger value of X. Analogously, if the observation with larger value of Y corresponds to the observation with smaller value of X, the pair of observations is said to be discordant (see e.g. Denuit and Lambert 2005).

[2] For the sake of simplicity, the Y variable values are ordered directly with respect to the X variable values, as originally suggested by Schezhtman and Yitzhaki (1987), while in Ferrari and Raffinetti (2015) they were reordered according to the $\hat{Y}$ values coming from the least-squares linear regression model $\hat{Y} = \hat{\alpha} + \hat{\beta}X$.

Fig. 1: The L_Y Lorenz curve, $L_Y^{'}$ dual Lorenz curve and the C concordance curve.



the Spearman's and Kendall's correlation coefficients. On the one hand, the MDC measure exploits the ranks to re-order the Y variable values. On the other hand, it focuses on the concordance/discordance between the dependent variable Y values, arranged in non-decreasing sense, and the same Y values arranged according to the non-decreasing ranks of the independent variable X.

All the above coefficients take values -1 and +1 as lower and upper bounds corresponding to a perfect inverse and direct dependence relationship (specifically, linear relationship for Pearson's coefficient), respectively, crossing the value zero in the case of independence between the variables.

While r, r_S and τ are interdependence indices, since they evaluate how the two variables relate to each other, MDC is an asymmetric coefficient and for this reason it is sensitive to the choice of the dependent and independent variables. Clearly, if the role of the variables is inverted, the results about the dependence relationship between the variables might change, making the index a proper dependence measure. Moreover, the MDC coefficient catches any monotonic dependence relationship as well as r_S and τ.

If data are specified through a continuous dependent variable and a discrete independent variable, as in scenario b), the Pearson's correlation coefficient preserves its expression in (1), while the computation of the Spearman's and Kendall's coefficients has to be based on the equations reported below:

$$r_S = \frac{\sum_{i=1}^{n}(r(x_i) - \bar{r}(x))(r(y_i) - \bar{r}(y))}{\sqrt{\sum_{i=1}^{n}(r(x_i) - \bar{r}(x))^2}\sqrt{\sum_{i=1}^{n}(r(y_i) - \bar{r}(y))^2}}, \qquad \text{for } i = 1, \ldots, n, \tag{5}$$

where $\bar{r}(x)$ and $\bar{r}(y)$ are the average ranks of X and Y and tied values are assigned a rank equal to the average of their positions in the ascending order of the values,

$$\tau = \frac{2(C - D)}{\sqrt{n^2 - n - \sum_{i=1}^{n} s_i(s_i - 1)}\sqrt{n^2 - n - \sum_{i=1}^{n} t_i(t_i - 1)}}, \qquad \text{for } i = 1, \ldots, n, \tag{6}$$

where s_i and t_i are the number of tied x_i and y_i values in the i-th tied group, respectively, while MDC is defined as

$$MDC = \frac{2\sum_{g=1}^{k}\sum_{i=n_{g-1}^*+1}^{n_g^*} i\bar{y}_g^t - n(n+1)(M_Y - y_0)}{2\sum_{i=1}^{n} iy_{(i)}^t - n(n+1)(M_Y - y_0)}, \tag{7}$$

where, given k ordered categories $X_1, \ldots, X_g, \ldots, X_k$, with $g = 1, \ldots, k$, n_g^* is the cumulative frequency of the first g categories and $\bar{y}_g^t$ is the average of the Y values corresponding to X_g. Trivially, if $g = k$ then $n_k^* = n$ (Ferrari and Raffinetti 2015).

Finally, if data are structured as in scenario c), once again the Pearson's and MDC coefficients are computed as in (1) and (4), while the Spearman's and Kendall's coefficients refer to formulas (5) and (6). In this last case, we do not resort to the MDC coefficient formula in (7), since this formula accounts for tied values only on the independent variable as happens in scenario b).

It is worth remarking that when one of the two variables is expressed through a discrete scale, the Pearson's, Spearman's and Kendall's coefficients never reach the extreme bounds -1 and +1. On the contrary, the MDC coefficient may reach the extreme bounds only in scenario c). For the sake of clarity, three simple toy examples of positive dependence are introduced with the aim of specifying a situation in which all the four indices take value +1 and two situations in which they take multiple values.

- First toy example: let Y and X be two variables such that $Y = \{1, 2, 3, 4, 5\}$ and $X = \{2, 3, 4, 5, 6\}$. No tied data are involved. By denoting with $MDC_{Y|X}$ the MDC coefficient of Y given X, the results are $MDC_{Y|X} = 1$, $r = 1$, $r_s = 1$ and $\tau = 1$. All the indices reach the upper bound due that the relationship between Y and X is perfectly linear and corresponding to $Y = X + 1$. By taking into account the MDC coefficient feature of being an asymmetric index, let us reverse the conditioning between the variables by computing the value of $MDC_{X|Y}$. In such a case $MDC_{X|Y} = MDC_{Y|X} = 1$ since no tied data are present.
- Second toy example: let Y and X be two variables such that $Y = \{1, 2, 3, 4, 5\}$ and $X = \{2, 3, 5, 6, 8\}$. As in the first example, no tied values appear but

the relationship between Y and X is monotonic but not perfectly linear. Thus, $MDC_{Y|X} = MDC_{X|Y} = r_S = \tau = 1$, while $r = 0.9934 < 1$.

- Third toy example: let Y and X be two variables such that $Y = \{1, 2, 3, 4, 5\}$ and $X = \{2, 2, 4, 5, 6\}$. Tied vales are present in X. $MDC_{Y|X}$, r, r_S and τ reach values smaller than $+1$, and more exactly $MDC_{Y|X} = 0.95$, $r = 0.9723$, $r_S = 0.9747$ and $\tau = 0.9487$. In such a case $MDC_{Y|X}$ takes values which are intermediate with respect to those of the Pearson's and Spearman's coefficients and that of the Kendall's coefficient. Nevertheless, if we reverse the conditioning by taking into account the X and Y variables as the dependent and the independent variables, respectively, it derives that $MDC_{X|Y} = 1$, being X perfectly depending on Y.

***Remark* 1** *The MDC coefficient appears as a predictability measure of a variable given the other, since it provides the increase of predictability of Y due to the knowledge of X. The more the Y values ordered according to the non-decreasing ranks of X approaches to the Y values arranged in non-decreasing sense, the more variable X predicts variable Y. Contrary to the Goodman and Kruskal's asymmetric predictability measure (see, e.g. Goodman and Kruskal 1954), which finds application in the case of categorical variables, the MDC coefficient is employed in the case of continuous and mixed variables.*

In order to account for the effects on the dependence coefficient behavior if specific adjustments are applied to discrete data, the proposal of Denuit and Labert (2005) (henceforth denoted with D&L) is considered. The D&L approach addresses to transform discrete variables into continuous variables. Specifically, let X be a discrete variable taking non-negative integers as values belonging to a subset Ψ of the set of natural numbers $\mathbb{N}$ and such that $f_X(x) = P(X = x)$, for $x \in \Psi$. In order to define the corresponding continuous variable X^*, D&L propose the following continuous-ation procedure:

$$X^* = X + (U - 1), \tag{8}$$

where U is a continuous random variable taking values in the range $(0, 1)$, independent of X and with a strictly increasing cumulative density function $F_U(u)$ on $(0, 1)$ sharing no parameters with the distribution of X. It results that X is "continued" by U producing the continuous variable X^*. Trivially, $X^* \leq X$ almost surely and its distribution function (cumulative density function) for $x^* \in \mathbb{R}$ is

$$F_{X^*}(x^*) = Pr(X^* \leq x^*) = F_X([x^*]) + F_U(x^* - [x^*]) f_X([x^*] + 1), \tag{9}$$

where $[x^*]$ is the integer part of $x^* \in \mathbb{R}$. As argued by D&L, the most natural choice for U is the uniform distribution on $(0, 1)$, such that $F_U(x^* - [x^*]) = x^* - [x^*]$. Thus, expression in (9) becomes

$$F_{X^*}(x^*) = F_X([x^*]) + (x^* - [x^*]) f_X([x^*] + 1). \tag{10}$$

***Remark* 2** *Several surveys gives rise to ordinal data, typically expressed through the Likert-type scales (see, e.g. Likert, 1932). Trivially, the dependence relationship in presence of an ordinal independent variable and a continuous dependent variable may be measured by resorting to the MDC formula in (7). In the case of an ordinal dependent variable and a continuous independent variable, a novel formalization of the MDC coefficient was introduced by Raffinetti (2019), who replaced the ordered categories of the dependent variable with the average of the continuous independent variable values corresponding to a specific ordered category of the dependent variable.*

2.2 Simulation design

In this Section, an extensive Monte Carlo simulation study, considering both different distributions generating the data and various measurement scales adopted to collect the data, is built. More precisely, we assume that the two observed variables are derived by either Normal or non-Normal distributions. In both situations, we first consider the data just as they are. This is the case of continuous-continuous variables (scenario a), labeled as "CC").
Then we assume that the dependent variable is continuous but the independent variable is collected as a discrete variable (scenario b), labeled as "CD"). Finally, the case where the dependent variable is recorded as a discrete variable and the independent variable is maintained continuous is taken into account (scenario c), labeled as "DC").
Since the involved discrete variable is generated from a continuous variable, recalling the proposal of D&L to transform the discrete variables into continuous variables, scenarios b) and c) are replicated in order to include the "continuous-sation" of the discrete variable as possible sub-scenarios. We denote with labels "CD-D&L" (scenario b)) and "DC-D&L" (scenario c)) these possible sub-scenarios. It is worth noting that the first and second letters of the labels always refer to the nature of the dependent and independent variables, respectively.
We then proceed to data generation. Data are first generated from bivariate Normal distributions with multiple values of the pairwise correlation coefficient ρ. The pairwise correlation coefficient takes the role of benchmark for the relationship between the variables, which in such a case is linear. In order to account for the several degrees of linear dependence between the variables, the pairwise correlation coefficient ρ is set equal to $\rho = \{0.1, 0.3, 0.5, 0.7\}$, to include the cases of weak, low, medium and high dependence between the variables. Thus, samples of size equal to 500 are drawn and the process is iterated 10,000 times.
If on the one hand, scenario a) is easier to build, on the other hand scenarios b) and c) deserve care not only in terms of the data generation process, but also in terms of the computation of the considered dependence coefficients. Since one of the two variable has discrete nature, we follow the proposal by Ferrari and Barbiero (2012) for generating correlated discrete variables coming

from Normal distributions, which was further translated into the "GenOrd" R package (see, e.g. Barbiero and Ferrari 2015). The crucial point here is the presence of only one discrete variable. In order to deal with this issue, an *"ad hoc"* adjustment of the GenOrd package is implemented.

The discretization process is extended both to asymmetrical and uniform discrete variable distributions. Moreover, the case of non-normally distributed data is also considered in our simulation study. The first contribution in generating non-Normal variables is due to Fleishman (1978) who defined, in the univariate case, a non-Normal variable Y as a linear combination of the first three powers of a standard Normal variable X, that is

$$Y = a + bX + cX^2 + dX^3, \tag{11}$$

where the constants a, b, c and d are determined by expanding (11) to express the first four moments of the non-Normal variable Y in terms of the fourteen moments of X which are known constants. Through some manipulations, the values of the constants a, b, c and d can be obtained as the solutions of a system of non-linear equations. Thus, the univariate non-Normal variable can be generated transforming the Normal variable by resorting both to the constants a, b, c, d and equation (11). Vale and Maurelli (1983) developed the same method for generating multivariate non-Normal distributions with specified inter-correlations and marginal means, variances, skewness, and kurtosis. The two parameters affecting the normality condition are skewness and kurtosis, where the kurtosis is measured by the parameter "excess of kurtosis". The more the excess of kurtosis moves from value 0, the more the normality condition becomes weaker. This occurs also for the skewness parameter. A value different from zero of skewness leads to an asymmetrical distribution. The conditions to obtain data from non-Normal distributions include, in addition to those already specified for drawing normally distributed data, the determination of the skewness γ and kurtosis κ parameters which are set equal to $\gamma = (1, 1)$ and $\kappa = (5.5, 5.5)$. The R code written by Zopluoglu (2011) and referring to the Vale and Maurelli's procedure is then employed.

In order to provide a summary of the main simulation study settings, an outline is provided in Table 1.

Table 1: Simulation study outline

Scenario	Simulation design
a) Continuous dependent variable and continuous independent variable	1) Sampling 10,000 samples of size 500 units from Normal distributions ($\rho = \{0.1, 0.3, 0.5, 0.7\}$) → Compute MDC, r, r_S and τ 2) Sampling 10,000 samples of size 500 units from non-Normal distributions ($\rho = \{0.1, 0.3, 0.5, 0.7\}$, $\gamma = (1, 1)$, $\kappa = (5.5, 5.5)$) → Compute MDC, r, r_S and τ
b) Continuous dependent variable and discrete independent variable	1) Sampling 10,000 samples of size 500 units from Normal distributions ($\rho = \{0.1, 0.3, 0.5, 0.7\}$) using GenOrd adjusted for discretizing the independent variable - Compute directly MDC, r, r_S and τ - Transform the discrete variable into a continuous variable by resorting to the Denuit and Lambert's procedure → Compute MDC, r, r_S and τ 2) Sampling 10,000 samples of size 500 units from non-Normal distributions ($\rho = \{0.1, 0.3, 0.5, 0.7\}$, $\gamma = (1, 1)$, $\kappa = (5.5, 5.5)$) using GenOrd adjusted for sampling from non-Normal distributions and for discretizing the independent variable - Compute directly MDC, r, r_S and τ - Transform the discrete variable into continuous variable by resorting to the Denuit and Lambert's procedure → Compute MDC, r, r_S and τ
c) Discrete dependent variable and continuous independent variable	1) Sampling 10,000 samples of size 500 units from Normal distributions ($\rho = \{0.1, 0.3, 0.5, 0.7\}$) using GenOrd adjusted for discretizing the dependent variable - Compute directly MDC, r, r_S and τ - Transform the discrete variable into continuous variable by resorting to the Denuit and Lambert's procedure → Compute MDC, r, r_S and τ 2) Sampling 10,000 samples of size 500 units from non-Normal distributions ($\rho = \{0.1, 0.3, 0.5, 0.7\}$, $\gamma = (1, 1)$, $\kappa = (5.5, 5.5)$) using GenOrd adjusted for sampling from non-Normal distributions and for discretizing the dependent variable - Compute directly MDC, r, r_S and τ - Transform the discrete variable into continuous variable by resorting to the Denuit and Lambert's procedure → Compute MDC, τ, r_S and r

3 Simulation results

3.1 Simulation results under normality

The simulation findings, when data are generated from Normal distributions, are displayed in Figures 2, 3, 4 and 5, which report the boxplots of the Monte Carlo values taken by the four indices in each drawn sample. Each figure, corresponding to the different values of the pairwise correlation coefficient $\rho = \{0.1, 0.3, 0.5, 0.7\}$, graphically denoted with the red dashed line, is split into three main boxes. The first main box shows scenario a) ("CC"). The second main box is devoted to scenario b) and includes two sub-boxes where the two different ways of proceeding when handling mixed data are taken into account. Specifically, in the first and second sub-boxes we consider the cases where the discrete variable nature is preserved ("CD") and it is transformed into a continuous variable through the approach of D&L ("CD-D&L"). The third main box, referred to scenario c) ("DC", if the discrete variable nature is preserved and "DC-D&L", if it is transformed into a continuous variable through the approach of D&L), is structured as scenario b). Without loss of generality, we discretize one of the two variables into four asymmetrical and uniform categories. The probability distribution associated with the four asymmetrical (p_a) and uniform (p_u) categories are defined as $p_a = \{0.1, 0.2, 0.3, 0.4\}$ and $p_u = \{0.25, 0.25, 0.25, 0.25\}$. For the sake of brevity, we report only the boxplots corresponding to the case of four asymmetrical categories. This because, even if the values of the indices are different, their behavior in presence of uniform categories is similar to that of asymmetrical categories. By looking at boxplots in Figures 2, 3, 4 and 5, the Monte Carlo distributions of the considered coefficients are symmetrical, as confirmed by the Monte Carlo median and mean values included in Table 2.Table 2 highlights also a further interesting issue concerning the variability of the indices, measured by the standard deviation (sd). In general, as the pairwise correlation coefficient increases, the variability reduces in magnitude.

Since scenario a) reflects the case of normally distributed data, the sample Pearson's-r correlation coefficient entirely catches the linear dependence relationship and this translates into an overlapping between the Monte Carlo r median (and mean) value and the value of the pairwise correlation coefficient ρ. Analogously, the MDC coefficient also appears as an unbiased estimator of ρ. A comparison between the two indices in terms of efficiency is then interesting. The MDC coefficient presents a Monte Carlo standard deviation a little bit higher that that associated with the Pearson's-r correlation coefficient, but this difference is close to zero. Thus, both the r and MDC coefficients may be equally used to measure the existing linear dependence relationship.

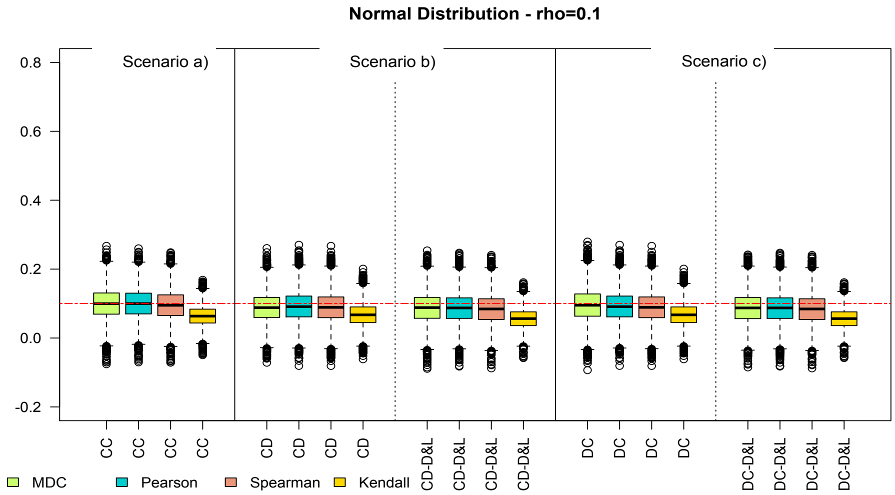


Fig. 2: Boxplots of MDC, r, r_S, and τ distributions in scenarios a), b) and c) for four categories, asymmetrical discrete variable - Normal distribution, $\rho = 0.1$.

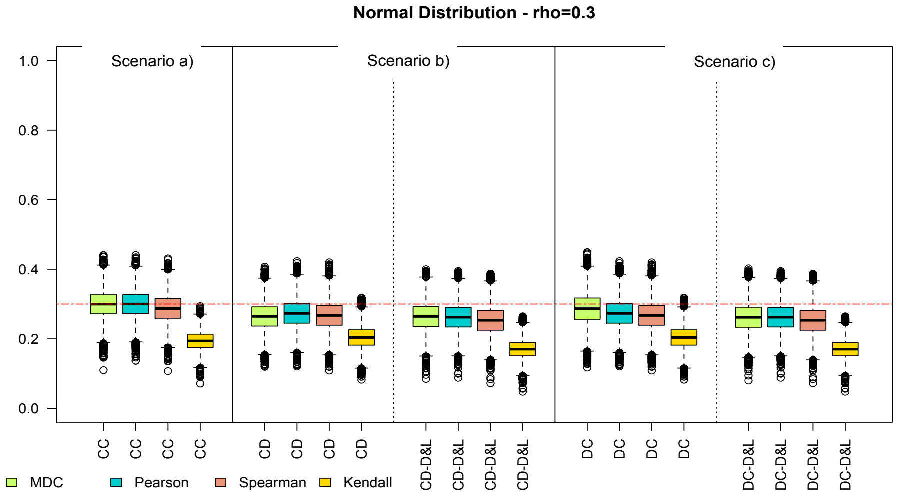


Fig. 3: Boxplots of MDC, r, r_S, and τ distributions in scenarios a), b) and c) for four categories, asymmetrical discrete variable - Normal distribution, $\rho = 0.3$.

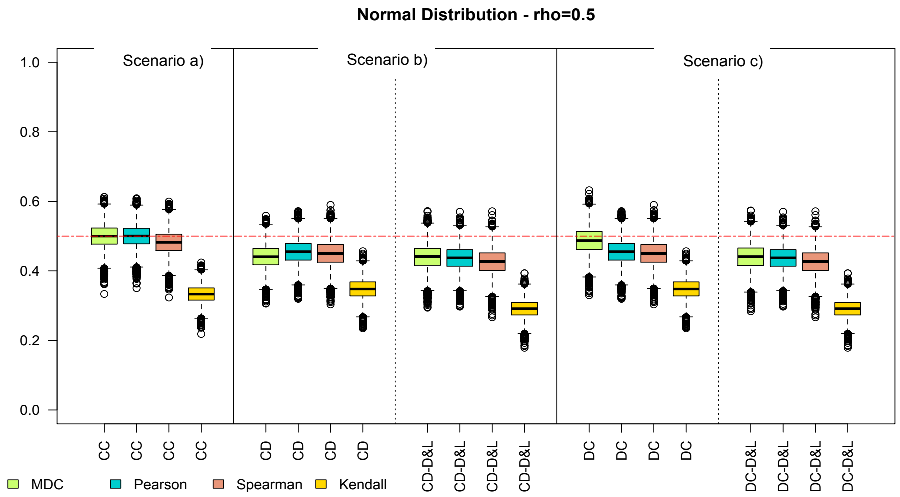


Fig. 4: Boxplots of MDC, r, r_S, and τ distributions in scenarios a), b) and c) for four categories, asymmetrical discrete variable - Normal distribution, $\rho = 0.5$.

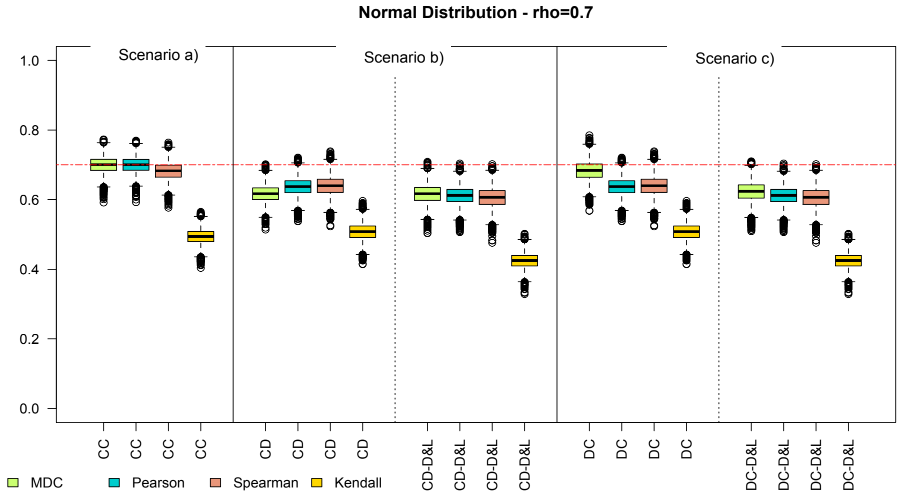


Fig. 5: Boxplots of MDC, r, r_S, and τ distributions in scenarios a), b) and c) for four categories, asymmetrical discrete variable - Normal distribution, $\rho = 0.7$.

Table 2: Monte Carlo median, mean and standard deviation (sd) of MDC, r, r_S and τ - Normal distributions, $\rho = \{0.1, 0.3, 0.5, 0.7\}$

	$\rho = 0.1$			$\rho = 0.3$			$\rho = 0.5$			$\rho = 0.7$		
scenario a) - CC	*median*	*mean*	*sd*	*median*	*mean*	*sd*	*median*	*mean*	*sd*	*median*	*mean*	*sd*
MDC	0.0997	0.0998	0.0450	0.2998	0.2997	0.0416	0.5000	0.4996	0.0345	0.7004	0.6996	0.0237
r	0.0999	0.1000	0.0440	0.3001	0.2997	0.0405	0.5003	0.4996	0.0334	0.7003	0.6996	0.0228
r_S	0.0950	0.0953	0.0440	0.2870	0.2870	0.0412	0.4820	0.4817	0.0353	0.6827	0.6818	0.0256
τ	0.0636	0.0637	0.0295	0.1939	0.1939	0.0284	0.3332	0.3333	0.0260	0.4941	0.4936	0.0217
scenario b) - CD	*median*	*mean*	*sd*	*median*	*mean*	*sd*	*median*	*mean*	*sd*			
MDC	0.0881	0.0883	0.0433	0.2645	0.2642	0.0404	0.4407	0.4403	0.0342	0.6170	0.6165	0.0248
r	0.0910	0.0912	0.0446	0.2732	0.2729	0.0416	0.4552	0.4547	0.0352	0.6374	0.6367	0.0253
r_S	0.0889	0.0890	0.0447	0.2673	0.2674	0.0424	0.4501	0.4498	0.0371	0.6399	0.6394	0.0281
τ	0.0673	0.0674	0.0339	0.2038	0.2038	0.0326	0.3479	0.3477	0.0296	0.5080	0.5077	0.0242
scenario b) - CD-D&L	*median*	*mean*	*sd*	*median*	*mean*	*sd*	*median*	*mean*	*sd*			
MDC	0.0883	0.0876	0.0452	0.2647	0.2640	0.0424	0.4413	0.4400	0.0362	0.6170	0.6161	0.0270
r	0.0873	0.0869	0.0441	0.2628	0.2621	0.0414	0.4374	0.4367	0.0352	0.6122	0.6114	0.0261
r_S	0.0843	0.0837	0.0441	0.2541	0.2534	0.0421	0.4269	0.4262	0.0371	0.6068	0.6058	0.0289
τ	0.0563	0.0559	0.0295	0.1709	0.1705	0.0287	0.2913	0.2908	0.0264	0.4250	0.4245	0.0226
scenario c) - DC	*median*	*mean*	*sd*	*median*	*mean*	*sd*	*median*	*mean*	*sd*			
MDC	0.0952	0.0955	0.0479	0.2867	0.2866	0.0451	0.4817	0.4809	0.0390	0.6819	0.6811	0.0286
r	0.0910	0.0912	0.0446	0.2732	0.2729	0.0416	0.4552	0.4547	0.0352	0.6374	0.6367	0.0253
r_S	0.0889	0.0890	0.0447	0.2673	0.2674	0.0424	0.4501	0.4498	0.0371	0.6399	0.6394	0.0281
τ	0.0673	0.0674	0.0339	0.2038	0.2038	0.0326	0.3479	0.3477	0.0296	0.5080	0.5077	0.0242
scenario c) - DC-D&L	*median*	*mean*	*sd*	*median*	*mean*	*sd*	*median*	*mean*	*sd*			
MDC	0.0874	0.0867	0.0449	0.2626	0.2622	0.0427	0.4410	0.4400	0.0371	0.6239	0.6231	0.0280
r	0.0873	0.0869	0.0441	0.2628	0.2621	0.0414	0.4374	0.4367	0.0352	0.6122	0.6114	0.0261
r_S	0.0843	0.0837	0.0441	0.2541	0.2534	0.0421	0.4269	0.4262	0.0371	0.6068	0.6058	0.0289
τ	0.0563	0.0559	0.0295	0.1709	0.1705	0.0287	0.2913	0.2908	0.0264	0.4250	0.4245	0.0226

As expected, the Spearman's correlation coefficient takes a Monte Carlo median (and mean) value which is slightly lower than the actual value of the population parameter ρ. The bias is more evident for the Kendall's correlation coefficient. In addition, the Kendall's correlation coefficient has the lowest variability, taking values concentrated around the Monte Carlo median (mean) value, which is very far from the fixed parameter ρ.
The Monte Carlo coefficient performance reflects what exactly happens in theory. The relation $\rho_\tau < \rho_S < \rho$, where ρ, ρ_τ and ρ_S correspond to the population Pearson's, Kendall's and Spearman's correlation coefficients (see e.g., McNeil, 2005), also holds on average for the Monte Carlo sample indices, encouraging us to state that the sample MDC unknown behavior is similar to that of the sample Pearson's correlation coefficient, resulting

$$\tau < r_S < r \cong MDC. \tag{12}$$

Scenario b) deserves more space for discussion. As one variable takes discrete nature, we expect that both the Pearson's and MDC Monte Carlo median (and mean) value decreases in magnitude with respect to the fixed pairwise correlation coefficient. Actually, as highlighted by Table 2 and Figures 3 and 4, for $\rho = \{0.3, 0.5\}$, the best estimator is r, but r_S also seems working good. On the contrary, the behavior of the MDC coefficient worsens with respect to scenario a), since tied data in the independent variable make the coefficient unable to completely account for the multiple values of the dependent variable.
If the procedure of D&L is implemented (Figures 2, 3, 4, 5 and Table 2), a worsening occurs only for r and r_S. Indeed, the MDC coefficient performance seems not to be affected by the continuous-ation process. In this case, the MDC coefficient is more robust to the different ways of dealing with discrete variables.
Last comments address scenario c). If the two variables are treated as they present, the MDC coefficient reaches a Monte Carlo median value which is closer than its competitors to the value of ρ, presenting as the best estimator of ρ, even if the MDC standard deviation is a little bit higher than that of r and r_S. These considerations lead us to conclude that generally the relationship

$$\tau < r_S < r < MDC \tag{13}$$

in average holds. Only when $\rho = 0.7$ (Figure 5), the median value of r_S seems to exceed the median value of r, showing a reverse in the previous relation, that is

$$\tau < r < r_S < MDC. \tag{14}$$

Finally, note that also in this scenario, the D&L continuous-ation process tends to negatively affect the behavior of all the coefficients.

3.2 Simulation results under non-normality

A parallel study is carried out on non-normally distributed data. Figures 6, 7, 8 and 9 show the simulation results related to scenarios a), b) and c).

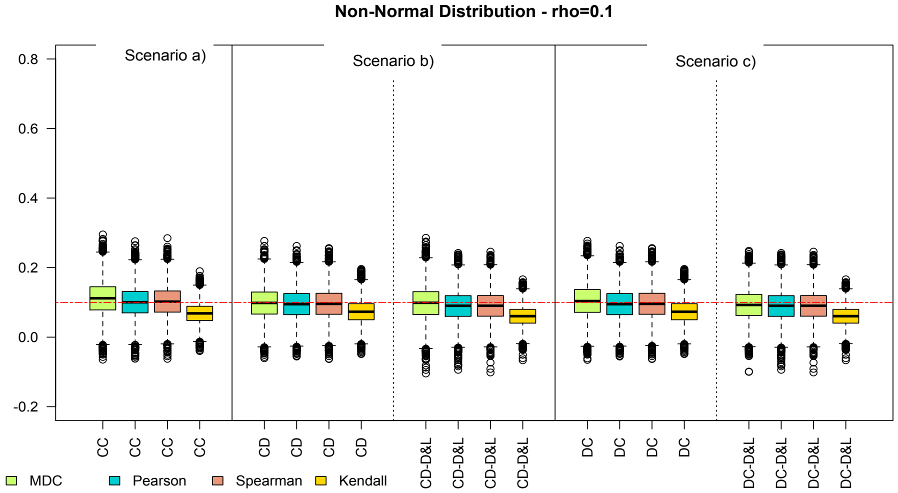


Fig. 6: Boxplots of MDC, r, r_S, and τ distributions in scenarios a), b) and c) for four categories, asymmetrical discrete variable - Non-Normal distribution, $\gamma = (1, 1)$, $\kappa = (5.5, 5.5)$ and $\rho = 0.1$.

As well as for the case of normally distributed data, the Monte Carlo distributions associated with the indices are symmetrical and the variability reduces as ρ increases (see Table 3).
We remark that in such a case ρ does not represent the parameter to be estimated, being the underlying data distribution non-Normal. For this reason, it makes no sense to recall the notion of unbiased and efficient estimator.
Findings referred to scenario a) highlight that in general the MDC and r_S indices reach greater values than those taken by the other competitor coefficients. In particular, their Monte Carlo median (mean) values are higher than the values fixed for ρ. This result is expected. Indeed, the MDC and r_S coefficients cover a wider concept of dependence, non-limited to the linear one. Nevertheless, when computed on continuous variables, the Spearman's coefficient may not compete with the MDC coefficient, which appears as the most informative dependence index. This because it is computed on the observed values rather than on ranks.

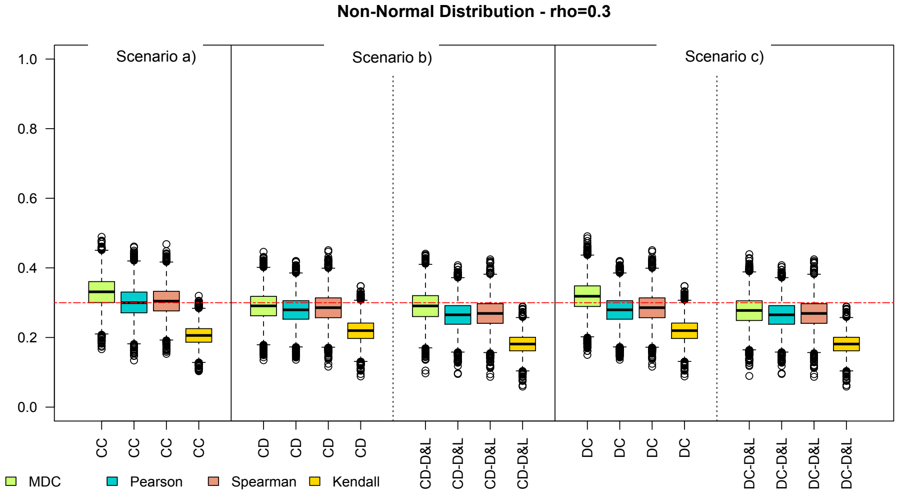


Fig. 7: Boxplots of MDC, r, r_S, and τ distributions in scenarios a), b) and c) for four categories, asymmetrical discrete variable - Non-Normal distribution, $\gamma = (1, 1)$, $\kappa = (5.5, 5.5)$ and $\rho = 0.3$.

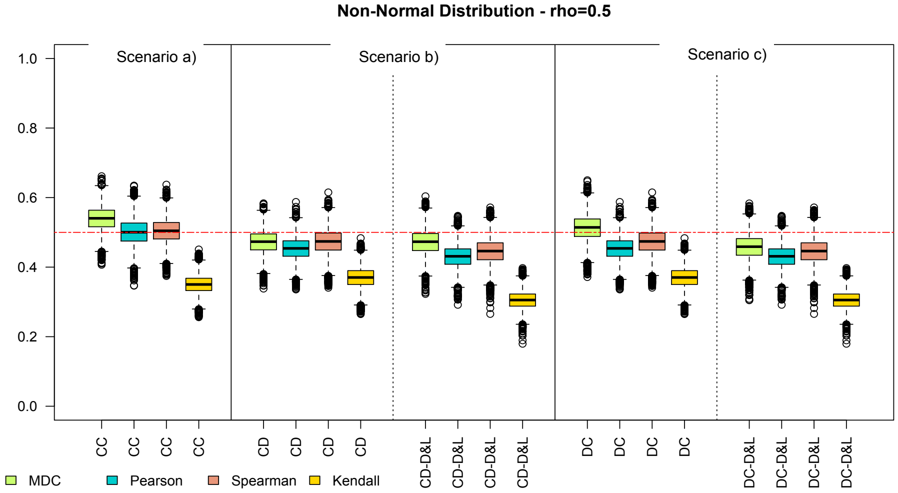


Fig. 8: Boxplots of MDC, r, r_S, and τ distributions in scenarios a), b) and c) for four categories, asymmetrical discrete variable - Non-Normal distribution, $\gamma = (1, 1)$, $\kappa = (5.5, 5.5)$ and $\rho = 0.5$.

As for normally distributed data, also in the case of non-normally distributed data, a novel relation among the indices may be defined as follows:

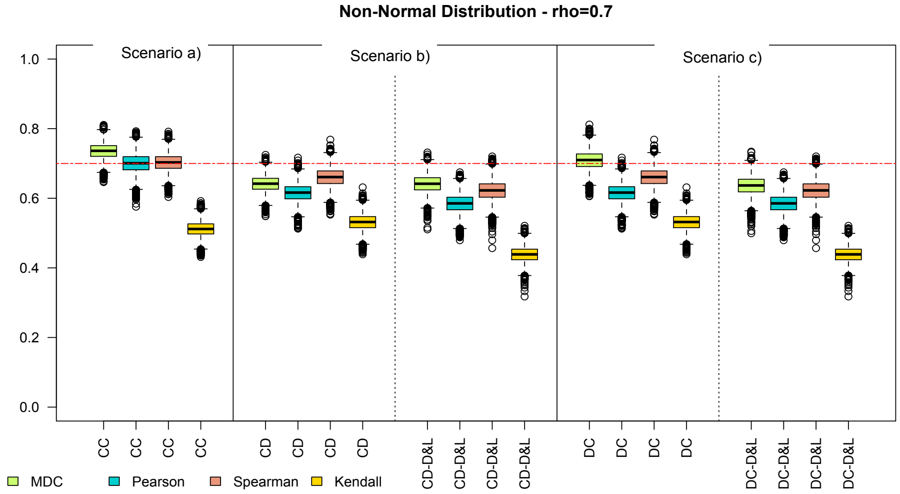


Fig. 9: Boxplots of MDC, r, r_S, and τ distributions in scenarios a), b) and c) for four categories, asymmetrical discrete variable - Non-Normal distribution, $\gamma = (1,1)$, $\kappa = (5.5, 5.5)$ and $\rho = 0.7$.

$$\tau < r < r_S < MDC. \tag{15}$$

Conclusions on the coefficient behavior are not so direct and evident in scenario b). The MDC coefficient reaches a Monte Carlo median value very close to those of the Spearman's coefficient, for $\rho = \{0.3, 0.5\}$. Consequently, MDC and r_S may be equivalently used. Nevertheless, if $\rho = 0.7$, the Monte Carlo median value of MDC is smaller than that of r_S, but if $\rho = 0.1$, the Monte Carlo median value of MDC is greater than that of r_S. As expected, the Pearson's correlation coefficient takes lower values than those of MDC and r_S, due to both the linearity condition violation outside Normal distributions and the discretization process of one variable.

Finally, in scenario c) the MDC coefficient shows as the best measure of the existing monotonic dependence relationship.

Based on the above considerations, the relation

$$MDC > r_S > r > \tau \tag{16}$$

is in general fulfilled. Moreover, it also holds when the D&L approach is implemented.

The previous simulation conditions are set by considering marginal distributions with the same values of skewness and kurtosis parameters ($\gamma = (1,1)$, $\kappa = (5.5, 5.5)$). The presence of marginal distributions characterized by equal values for the skewness and kurtosis parameters makes the specification of the variable role useless, yielding that each of the two variables may be considered as the independent or dependent one. In this case, the asymmetric MDC index is not sensitive to the choice of the dependent and independent variables. Nevertheless, if the marginal distribution skewness and kurtosis parameters vary, the specification of the dependent and independent variables is required for the MDC computation. Let us defined the mixed-case with skewness and kurtosis parameters set equal to $\gamma = (2, 1)$ and $\kappa = (12, 5.5)$, respectively. If we replicate the simulation conditions (sample size equal to 500, number of iterations corresponding to 10,000 and $\rho = 0.5$), the MDC coefficient takes different values depending on the variable which is chosen as the dependent variable. This result is displayed by the boxplots reported in Figure 10. In each scenario, the MDC coefficient is computed by first conditioning variable Y on variable X ($MDC_{Y|X}$) and then variable X on variable Y ($MDC_{X|Y}$). Even if in scenario a), where the variables are preserved with their original continuous nature, the difference between the two MDC indices is negligible, this difference becomes more evident if one of the two continuous variables is discretized (scenario b) or scenario c)).

Table 3: Monte Carlo median, mean and standard deviation (sd) of MDC, r, r_S and τ - non-Normal distributions, $\rho = \{0.1, 0.3, 0.5, 0.7\}$

	$\rho = 0.1$			$\rho = 0.3$			$\rho = 0.5$			$\rho = 0.7$		
scenario a) - CC	*median*	*mean*	*sd*	*median*	*mean*	*sd*	*median*	*mean*	*sd*			
MDC	0.1119	0.1118	0.0494	0.3312	0.3305	0.0445	0.5404	0.5395	0.0353	0.7363	0.7355	0.0227
r	0.1004	0.1006	0.0455	0.3001	0.3006	0.0439	0.5004	0.5005	0.0385	0.7006	0.7002	0.0281
r_S	0.1022	0.1023	0.0448	0.3045	0.3042	0.0416	0.5045	0.5041	0.0351	0.7031	0.7023	0.0248
τ	0.0683	0.0684	0.0300	0.2059	0.2059	0.0288	0.3502	0.3501	0.0261	0.5118	0.5118	0.0215
scenario b) - CD	*median*	*mean*	*sd*	*median*	*mean*	*sd*	*median*	*mean*	*sd*			
MDC	0.0985	0.0984	0.0464	0.2909	0.2902	0.0418	0.4732	0.4722	0.0338	0.6420	0.6411	0.0231
r	0.0950	0.0949	0.0442	0.2795	0.2788	0.0394	0.4539	0.4531	0.0324	0.6163	0.6155	0.0256
r_S	0.0958	0.0958	0.0449	0.2857	0.2852	0.0419	0.4740	0.4731	0.0362	0.6609	0.6602	0.0269
τ	0.0730	0.0731	0.0343	0.2198	0.2192	0.0326	0.3703	0.3698	0.0293	0.5319	0.5313	0.0236
scenario b) - CD-D&L	*median*	*mean*	*sd*	*median*	*mean*	*sd*	*median*	*mean*	*sd*			
MDC	0.0988	0.0980	0.0486	0.2908	0.2900	0.0446	0.4731	0.4719	0.0365	0.6417	0.6410	0.0261
r	0.0903	0.0896	0.0441	0.2652	0.2647	0.0397	0.4312	0.4303	0.0332	0.5854	0.5847	0.0270
r_S	0.0903	0.0897	0.0441	0.2691	0.2685	0.0417	0.4462	0.4454	0.0364	0.6227	0.6217	0.0284
τ	0.0604	0.0600	0.0296	0.1812	0.1809	0.0285	0.3054	0.3051	0.0262	0.4388	0.4385	0.0225
scenario c) - DC	*median*	*mean*	*sd*	*median*	*mean*	*sd*	*median*	*mean*	*sd*			
MDC	0.1037	0.1038	0.0483	0.3089	0.3082	0.0447	0.5110	0.5097	0.0377	0.7091	0.7081	0.0268
r	0.0950	0.0949	0.0442	0.2795	0.2788	0.0394	0.4539	0.4531	0.0324	0.6163	0.6155	0.0256
r_S	0.0958	0.0958	0.0449	0.2857	0.2852	0.0419	0.4740	0.4731	0.0362	0.6609	0.6602	0.0269
τ	0.0730	0.0731	0.0343	0.2198	0.2192	0.0326	0.3703	0.3698	0.0293	0.5319	0.5313	0.0236
scenario c) - DC-D&L	*median*	*mean*	*sd*	*median*	*mean*	*sd*	*median*	*mean*	*sd*			
MDC	0.0927	0.0926	0.0449	0.2777	0.2769	0.0420	0.4588	0.4578	0.0360	0.6369	0.6361	0.0271
r	0.0903	0.0896	0.0436	0.2652	0.2647	0.0397	0.4312	0.4303	0.0332	0.5854	0.5847	0.0270
r_S	0.0903	0.0897	0.0441	0.2691	0.2685	0.0417	0.4462	0.4454	0.0364	0.6227	0.6217	0.0284
τ	0.0604	0.0600	0.0296	0.1812	0.1809	0.0285	0.3054	0.3051	0.0262	0.4388	0.4385	0.0225

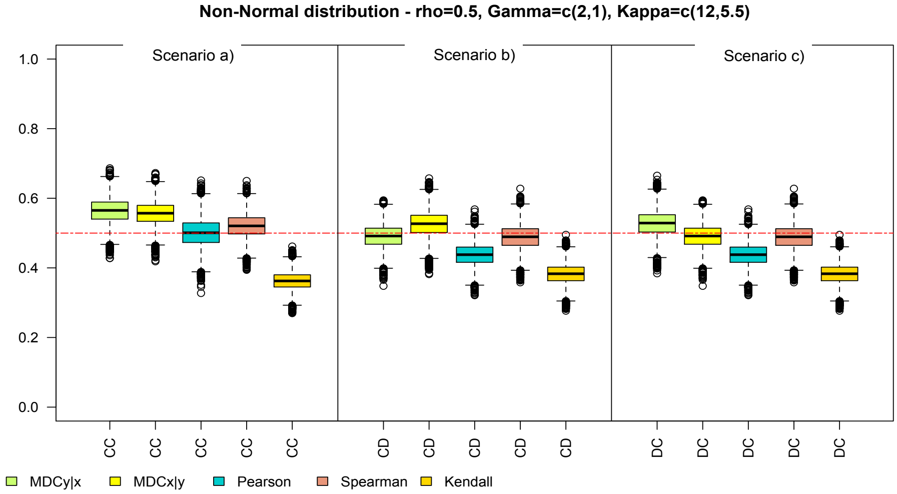
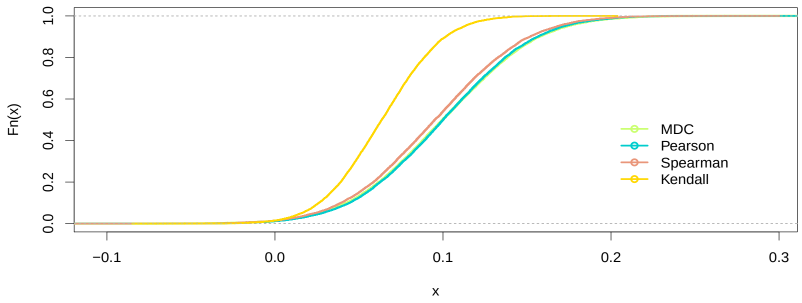


Fig. 10: Boxplots of $MDC_{Y|X}$, $MDC_{X|Y}$, r, r_S and τ distributions in scenarios a), b) and c) for four categories asymmetrical discrete variable sampling from Non-Normal distributions, $\gamma = (2, 1)$, $\kappa = (12, 5.5)$ and $\rho = 0.5$.

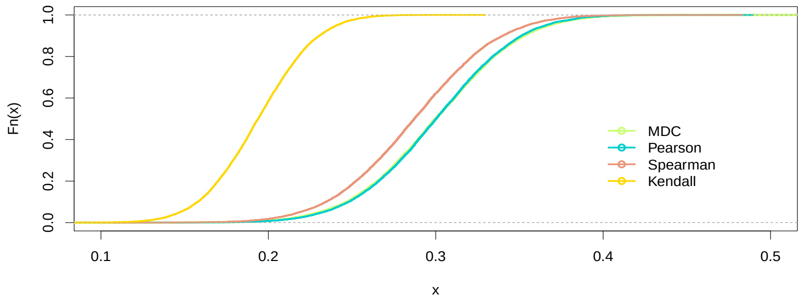
3.3 Hypotheses testing

To gain more insights into the behavior of the considered indices, we now test if the differences between their empirical cumulative distribution functions are significant. The empirical cumulative distribution functions related to scenarios a), b), and c) and different values of ρ are displayed in Figures 11, 12 and 13, if data are normally distributed, and in Figures 14, 15 and 16, if data are non-normally distributed. Since the approach of D&L does not provide any improvement in the performance of the considered indices, it is not included in this additional study.

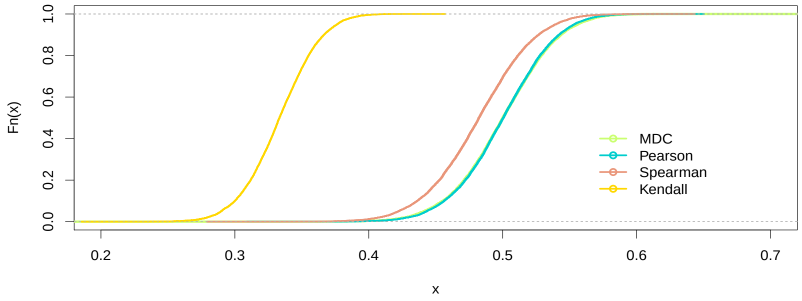
Fig. 11: Empirical cumulative density functions - Scenario a) (Normal distribution)



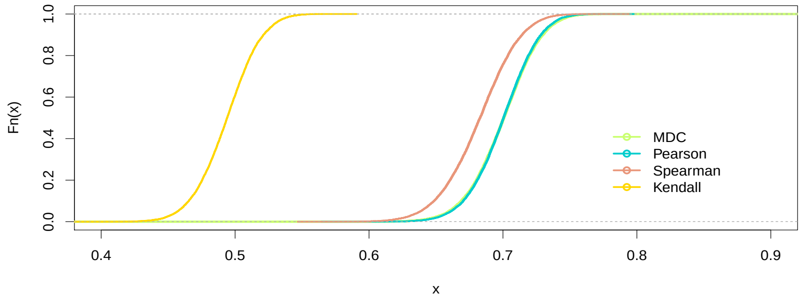
(a) $\rho = 0.1$



(b) $\rho = 0.3$

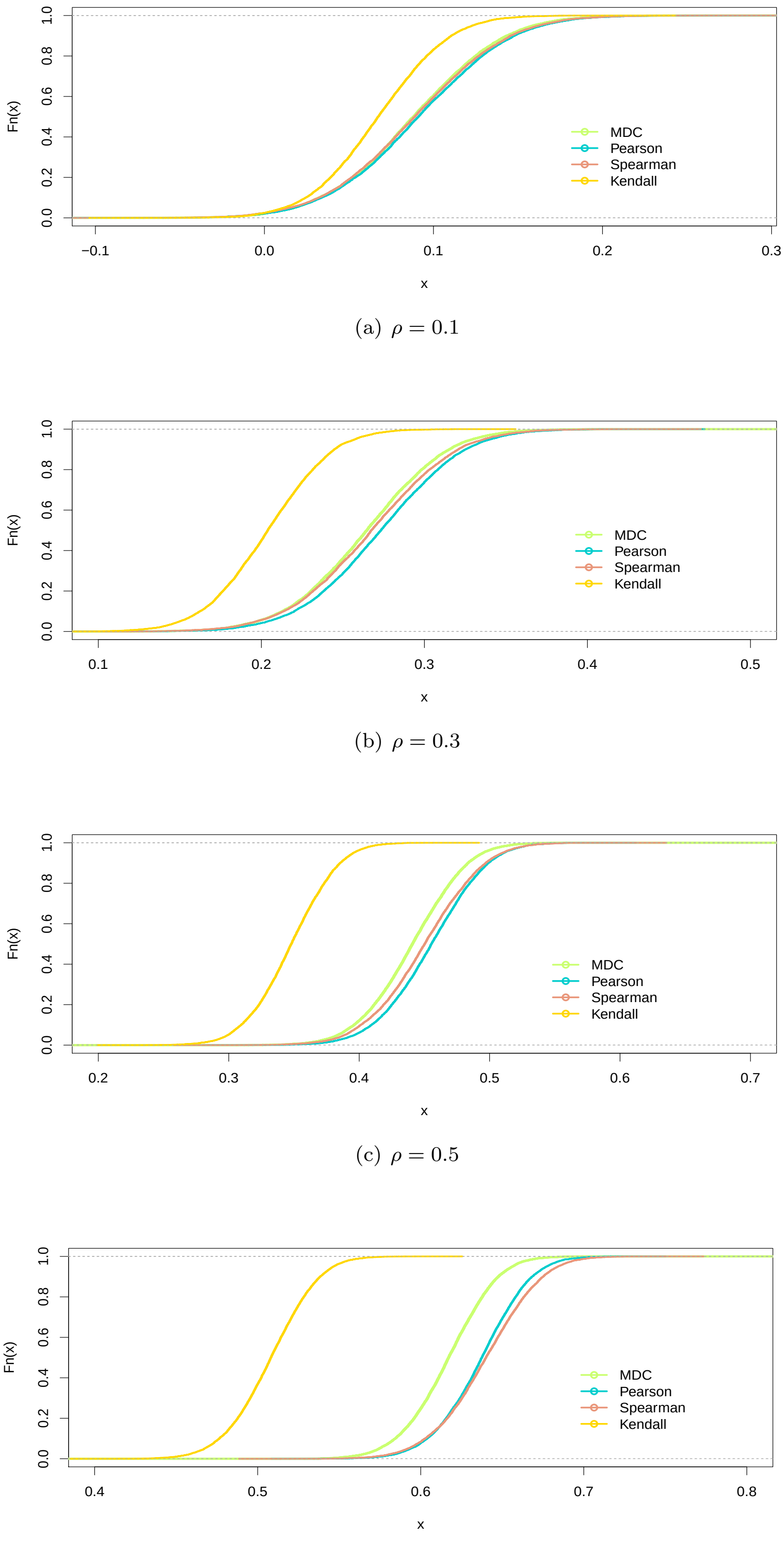


(c) $\rho = 0.5$



(d) $\rho = 0.7$

Fig. 12: Empirical cumulative density functions - Scenario b) (Normal distribution)



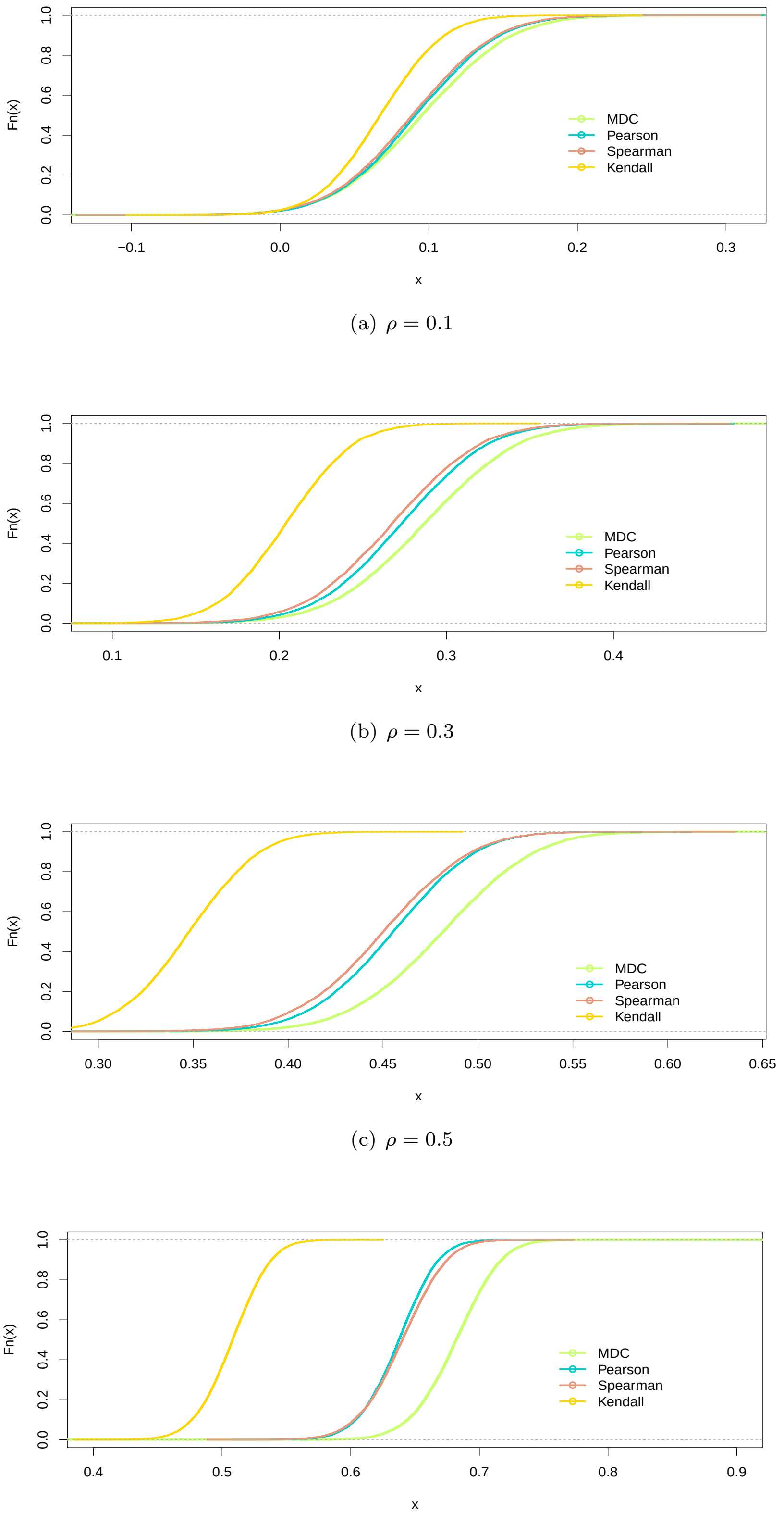
(a) $\rho = 0.1$

(b) $\rho = 0.3$

(c) $\rho = 0.5$

(d) $\rho = 0.7$

Fig. 13: Empirical cumulative density functions - Scenario c) (Normal distribution)



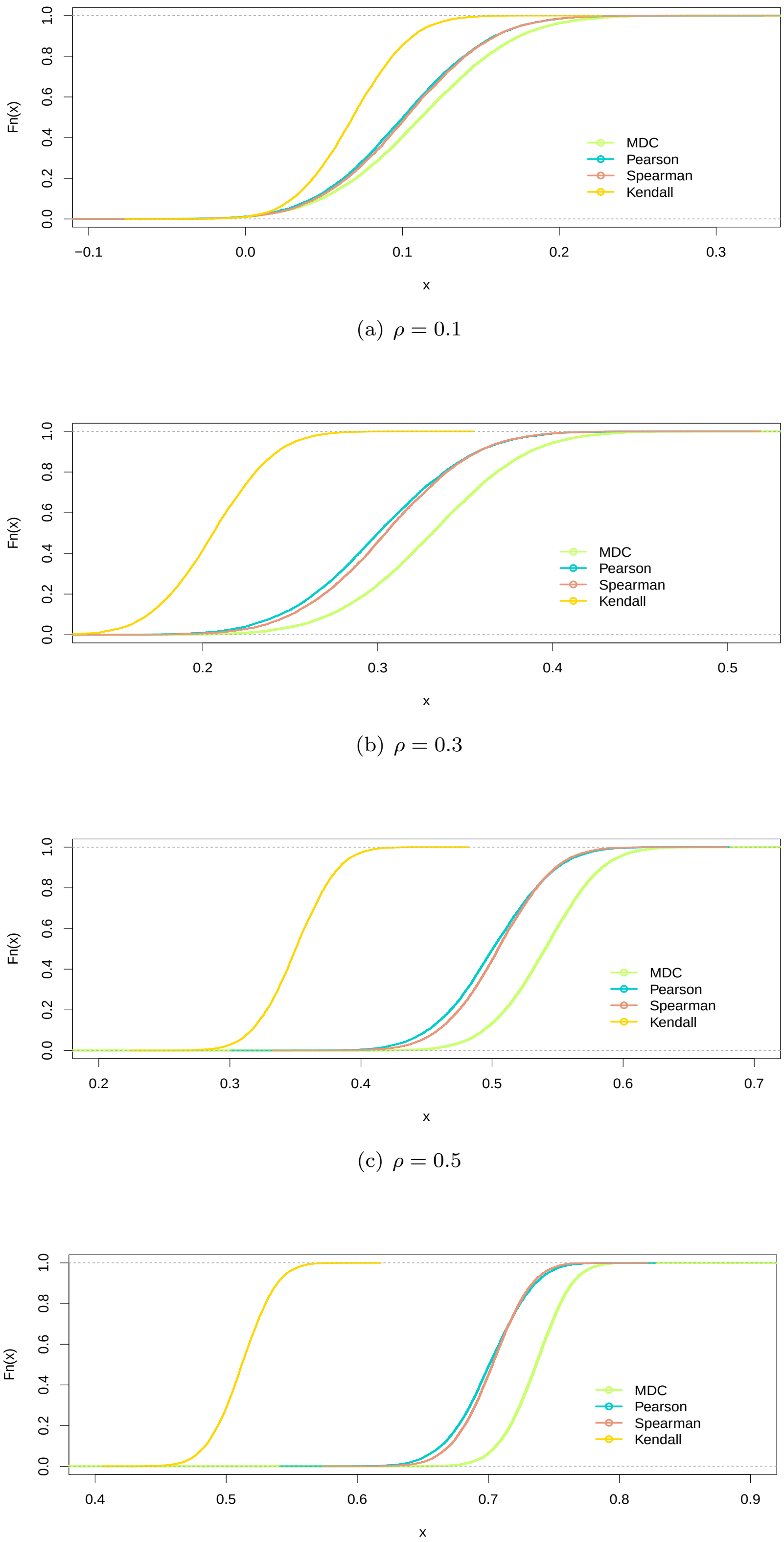
(a) $\rho = 0.1$

(b) $\rho = 0.3$

(c) $\rho = 0.5$

(d) $\rho = 0.7$

Fig. 14: Empirical cumulative density functions - Scenario a) (Non-Normal distribution)



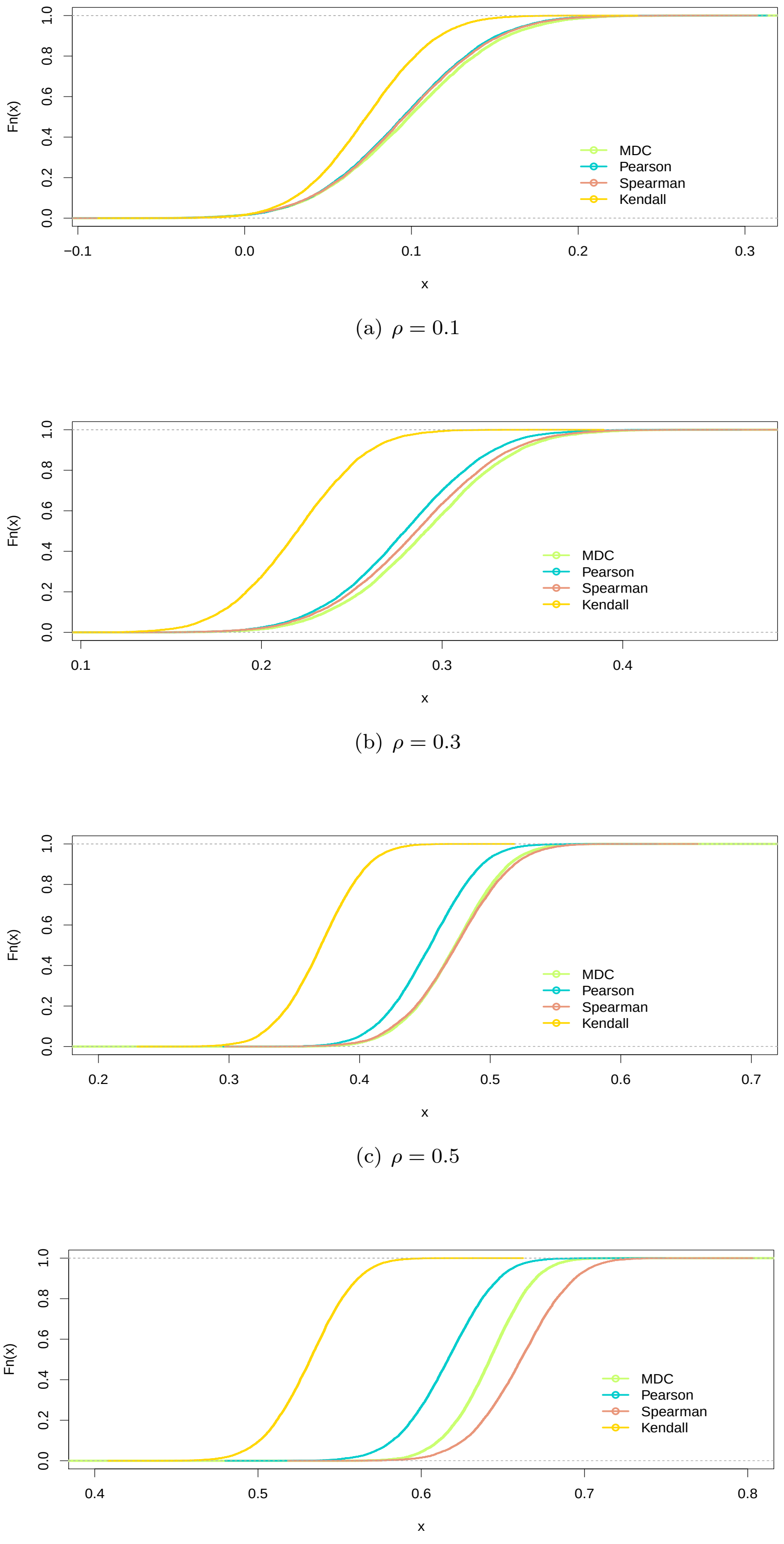
(a) $\rho = 0.1$

(b) $\rho = 0.3$

(c) $\rho = 0.5$

(d) $\rho = 0.7$

Fig. 15: Empirical cumulative density functions - Scenario b) (Non-Normal distribution)



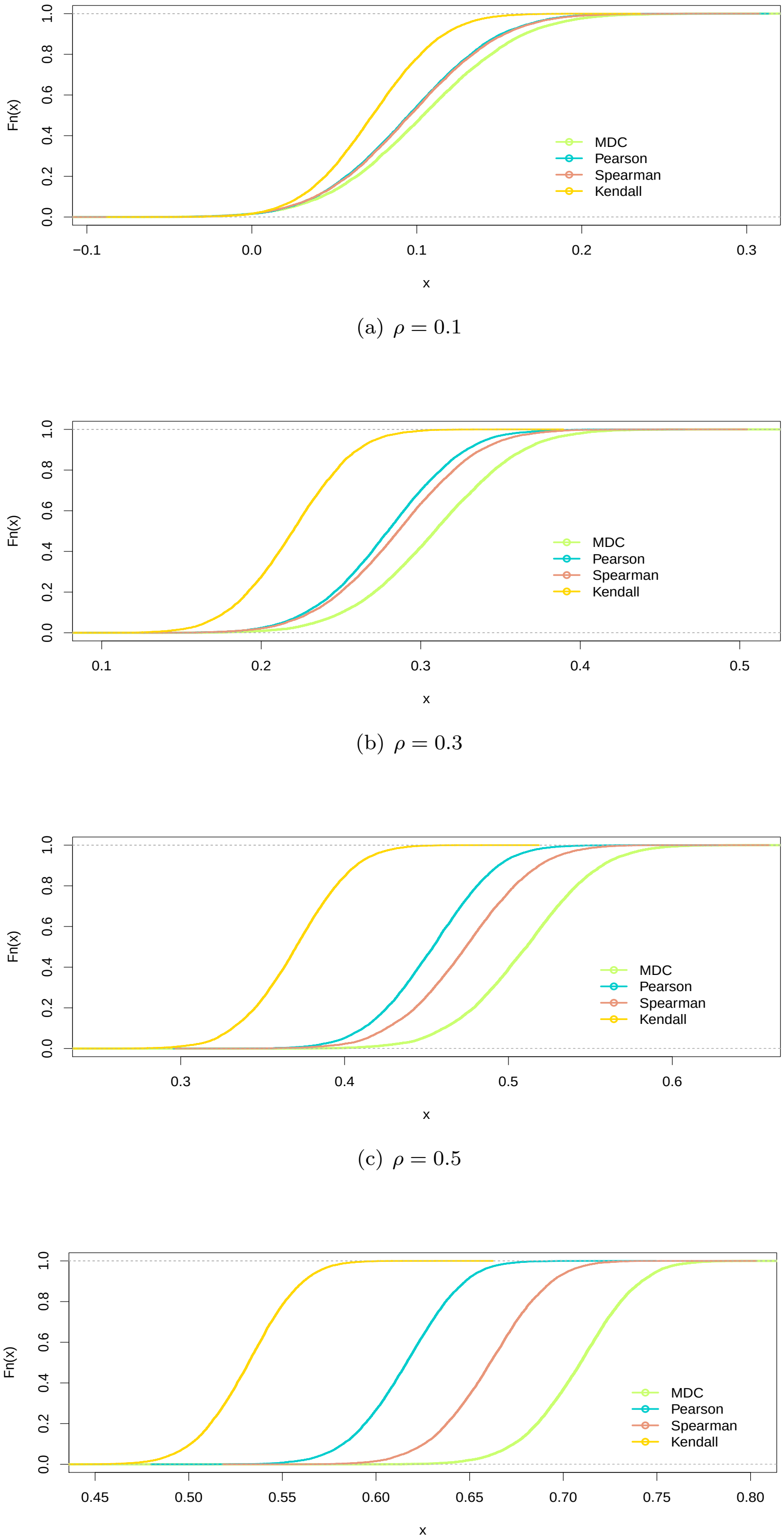
(a) $\rho = 0.1$

(b) $\rho = 0.3$

(c) $\rho = 0.5$

(d) $\rho = 0.7$

Fig. 16: Empirical cumulative density functions - Scenario c) (Non-Normal distribution)



(a) $\rho = 0.1$

(b) $\rho = 0.3$

(c) $\rho = 0.5$

(d) $\rho = 0.7$

Through a direct comparison of the empirical cumulative distribution functions of the indices, statements in terms of stochastic dominance can be introduced. The first order stochastic dominance defines an order relationship between cumulative distribution functions (see e.g. Heathcote et al. 2010). Given two variables Y and X, let $F(x)$ and $F(y)$ be the associated continuous cumulative distribution functions, such that $F(x), F(y) : \mathbb{R} \to [0, 1]$. $F(x)$ dominates $F(y)$ if and only if $F(x) < F(y)$, $\forall x, y \in \mathbb{R}$. From Figures 11, 12, 13, 14, 15 and 16 one can realise that the Kendall's coefficient empirical cumulative distribution function is always dominated by the empirical cumulative distribution functions of each of the other indices. For the remaining indices, these differences are not so evident. Inferential analyses are then required to evaluate if these differences may be considered as significant. We resort to the Page's test (Page 1963), which is based on punctual comparison of the coefficient empirical cumulative distribution functions. Following Solaro et al. (2017), let $F_{(MDC)}$, $F_{(r)}$ and $F_{(r_S)}$ be the cumulative distribution functions of the MDC, r and r_s values in each iteration. We test the null hypothesis

$$H_0 : F_{(MDC)} = F_{(r)} = F_{(r_S)}, \tag{17}$$

stating that there is no difference between the expected ranks for the three coefficients in each iteration, against each of the following six ordered alternative hypotheses:

$$\begin{aligned} 1 &= F_{(MDC)} > F_{(r)} > F_{(r_S)} \\ 2 &= F_{(r)} > F_{(MDC)} > F_{(r_S)} \\ 3 &= F_{(r_S)} > F_{(r)} > F_{(MDC)} \\ 4 &= F_{(MDC)} > F_{(r_S)} > F_{(r)} \\ 5 &= F_{(r)} > F_{(r_S)} > F_{(MDC)} \\ 6 &= F_{(r_S)} > F_{(MDC)} > F_{(r)}. \end{aligned} \tag{18}$$

The significance level α is set equal to 0.05. For each scenario, we assume that the choice of the coefficient is detected as the one with the smallest p-value. In the case of equivalent p-values, the index with the highest test statistics value is chosen. Results are reported in Table 4.

Table 4: Page's test results: the number of the alternative hypothesis chosen varying scenario, ρ value and distribution

	Normal distributions				**Non-Normal distributions**			
ρ	**0.1**	**0.3**	**0.5**	**0.7**	**0.1**	**0.3**	**0.5**	**0.7**
Scenario a)	1-2	1-2	1-2	1-2	4	4	4	4
Scenario b)	5	5	5	3	4	4	4-6	6
Scenario c)	1	1	1	4	1	4	4	4

Each cell in Table 4 displays the number assigned to the alternative hypothesis characterized by the smallest p-value among those in (18). Situations, in which two numbers associated with two alternative hypotheses appear, occur wether the test results present the same p-values and similar values of the test statistics. The alternative hypotheses based on the ordering with $F_{(MDC)}$, $F_{(r)}$ and $F_{(r_S)}$ as the first element are those labeled as 1 and 4, 2 and 5, 3 and 6, respectively.

Beside the above findings, it is crucial to assess how these measures compare with type-1 (false positive rate) and type-2 (false negative rate) errors. To do so, we add to the simulation results for the case of normally distributed data characterized by a pairwise correlation coefficient $\rho = 0.7$, a new simulation corresponding to the case of a pairwise correlation coefficient $\rho = 0.6$. When data follow Normal distributions, the pairwise correlation coefficient essentially coincides with the empirical observed correlation between the variables. We aim at detecting a linear dependence relationship close to 0.7 and we fix 0.68 as a threshold value. The hypotheses to be tested are:

$$\begin{aligned} &H_0 : \rho_{MDC} \geq 0.68 \quad \text{vs} \quad H_1 : \rho_{MDC} < 0.68 \\ &H_0 : \rho \geq 0.68 \quad \text{vs} \quad H_1 : \rho < 0.68 \\ &H_0 : \rho_S \geq 0.68 \quad \text{vs} \quad H_1 : \rho_S < 0.68, \end{aligned}$$

where ρ_{MDC}, ρ and ρ_S correspond to the population MDC, Pearson's and Spearman's coefficients. Specifically, the false positive rate (type-1 error rate) is computed on data from Normal distribution with $\rho = 0.7$, while the false negative rate (type-2 error rate) is determined on data from Normal distribution with $\rho = 0.6$. Results are reported in Table 5.

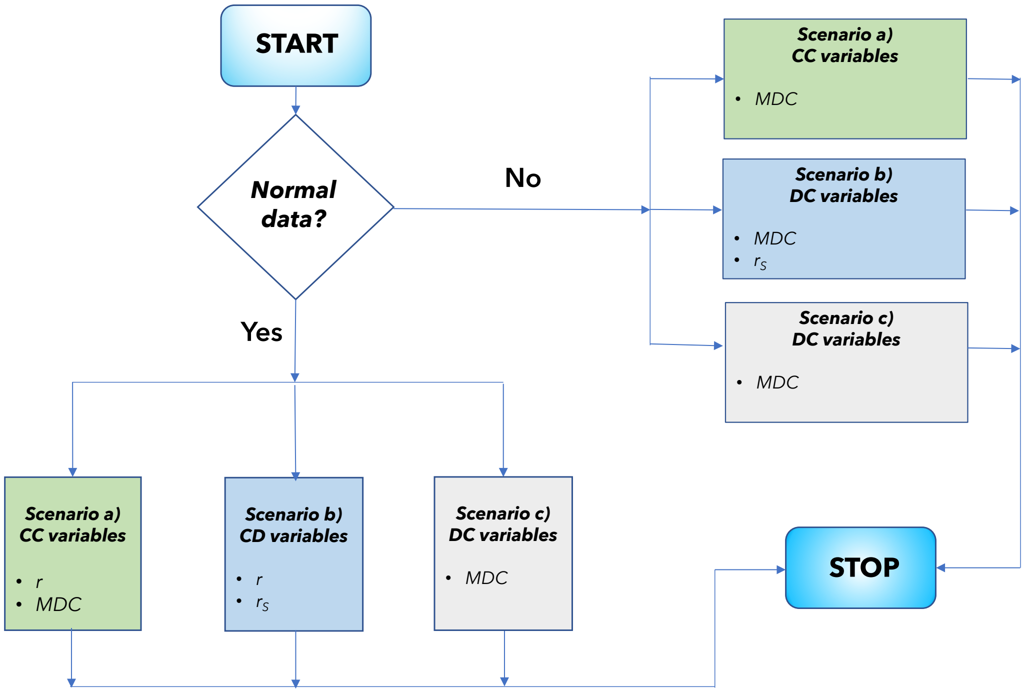
Table 5: Type-1 and type-2 error rates

	Type-1 error rate			**Type-2 error rate**		
	Scenario a)	**Scenario b)**	**Scenario c)**	**Scenario a)**	**Scenario b)**	**Scenario c)**
MDC	0.1993	0.9961	0.4477	0.0013	0	0.0010
r	0.1935	0.9620	0.9620	0.0012	0	0
r_S	0.4594	0.9319	0.9319	0.0002	0	0

From Table 5 it results that the type-1 error related to the MDC coefficient is: similar to that associated with the Pearson's correlation coefficient in scenario a); the greatest one in scenario b); the smallest one in scenario c). The type-2 errors are instead close to zero.

The elements for tracing all the possible procedures for the dependence relationship discovery are now available. Specifically, a flow tree addressing to the dependence coefficient correct choice is displayed in Figure 17 for the cases of normal and non-normal (continuous and discrete) data.

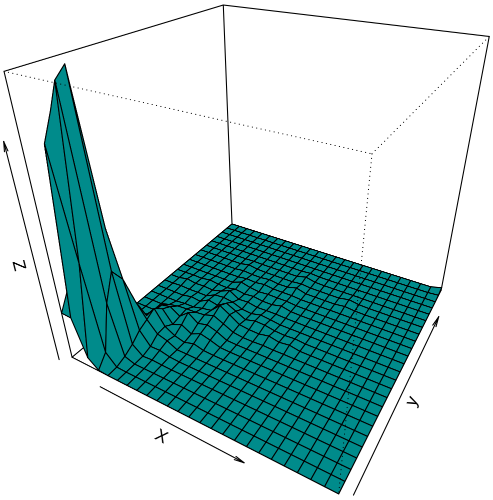
Fig. 17: A flow tree for discovering dependence relationships with Normal and non-Normal data - scenario a), scenario b), scenario c).



4 Application to real data

The dataset “Employee Data.sav” from `https://www.spss-tutorials.com/spss- opening-data-with-syntax/` is used to illustrate the coherence between the simulation results and those obtained on actual data.
The file contains data extracted from a bank’s employee records in an investigation into discrimination in 1987 and focuses on employees’ gender, educational degree, employment category, months since hire, previous experience (months), minority classification (i.e., whether ethnic minority), initial and current salaries.
In order to accomplish scenario a), we select as variables of interest, the current salary (in dollars), which is the dependent variable, and the initial salary (in dollars), which is the independent variable. The joint distribution of the initial and current salary is displayed in Figure 18. The skewness and kurtosis indices are equal to $\gamma = (2.11, 2.86)$ and $\kappa = (12.17, 5.26)$.

Fig. 18: Joint distribution of current and initial salary.



Since the joint distribution of the two considered variables is non-Normal, the MDC coefficient seems the most appropriate measure to be used. Computed on data, it detects the existence of a strong monotonic dependence relationship. Also the Pearson's-r, Spearman's-r_S and Kendall's-τ coefficients are computed (see Table 8) showing the coherence with the ordering arisen from the simulation study.
Let us now suppose that, for reasons related to the data collection process, the initial salary is recorded in terms of discrete categories. Thus, the independent variable is not expressed by single values, but by tied data even if its underlying distribution is continuous. This case corresponds to scenario b). This scenario is built by first splitting the initial salary into three classes and subsequently by encoding each class by three ordered categories, as illustrated in Table 6.

Table 6: Distribution of the initial salary variable discretized into three categories

Categories of initial salary	1	2	3
Distribution	0.20	0.63	0.17

Due to the non-normality condition of the original variable joint distribution, we may resort to the use of both the Spearman's and MDC coefficients (see Table 8).
The loss of information when one of the two variables is categorized is also determined. The indices with the smallest loss of information are r_S and τ. Even if the Kendall's coefficient reports the minimum loss of information, it has not to be considered as a good measure, since its performance on the original data is the worst one. The highest loss of information is associated to r, while MDC leads to an intermediate level, smaller than that of the Pearson's coefficient but substantially bigger than that of the Spearman's coefficient (Table 8).
Finally, in order to determine scenario c), let us suppose that the dependent variable (current salary) is recorded through discrete categories as displayed in Table 7.

Table 7: Distribution of the current salary variable discretized into three categories

Categories of initial salary	1	2	3
Distribution	0.55	0.41	0.04

MDC takes value equal to 0.8021, while the remaining coefficients take values which are far lower. Moreover, the relative loss of information yielded by the MDC coefficient is the lowest one with respect to the that provided by the other coefficients (Table 8).

Table 8: Coefficient value in scenario a), coefficient values and loss of information in scenarios b) and c)

	Scenario a)	**Scenario b)**		**Scenario c)**	
Coefficients	**Value ($v_{a)}$)**	**Value ($v_{b)}$)**	$\frac{v_{a)}-v_{b)}}{v_{a)}}$	**Value ($v_{c)}$)**	$\frac{v_{a)}-v_{c)}}{v_{a)}}$
MDC	0.8903	0.7590	0.1475	0.8021	0.0991
r	0.8802	0.7343	0.1658	0.6477	0.2641
r_s	0.8253	0.7526	0.0881	0.6975	0.1548
τ	0.6555	0.6325	0.0351	0.5856	0.1066

These findings allow us to conclude that the guidelines suggested by our proposed tree flow are also fulfilled on real data.
To further investigate the coherence between the results on simulated and real data and supposing that the distribution generating sample data is unknown, an additional simulation study based on bootstrap resampling techniques is led on the available dataset. Then, 10,000 randomly samples for variables Y and X are drawn, with replacement, from the whole dataset and the value of each coefficient computed in each selected sample. Table 9 reports the associated bootstrap estimates with their standard errors (se) and coefficients of variation (cv).

Table 9: MDC, r, r_S, and τ in scenarios a), b) and c): original value, bootstrap estimate, se, and cv.

	Scenario a)			
Coefficients	***Original Value***	***Bootstrap estimate***	***Bootstrap*** se	***Bootstrap*** cv
MDC	0.8903	0.8897	0.0118	0.0133
r	0.8802	0.8805	0.0167	0.0190
r_s	0.8253	0.8243	0.0186	0.0226
τ	0.6555	0.6554	0.0196	0.0299
	Scenario b)			
Coefficients	***Original Value***	***Bootstrap estimate***	***Bootstrap*** se	***Bootstrap*** cv
MDC	0.7590	0.7583	0.0195	0.0257
r	0.7343	0.7353	0.0171	0.0233
r_S	0.7526	0.7519	0.0222	0.0295
τ	0.6325	0.6322	0.0197	0.0312
	Scenario c)			
Coefficients	***Original Value***	***Bootstrap estimate***	***Bootstrap*** se	***Bootstrap*** cv
MDC	0.8021	0.8023	0.0255	0.0318
r	0.6477	0.6489	0.0245	0.0378
r_S	0.6975	0.6968	0.0240	0.0344
τ	0.5856	0.5855	0.0207	0.0354

In general, from the comparison between the original coefficient values computed on sample data and their bootstrap values, it appears that typically the bootstrap estimates are unbiased. Focusing on variability, the MDC index provides the lowest relative variability (measured by the coefficient of variation cv) in scenarios a) and c). In scenario b), the MDC coefficient of variation is slightly greater than that of the Pearson's coefficient, but its value is very close to that of the Pearson's coefficient. Thus, the MDC coefficient results as the measure that better catches the dependence relationship between the variables with the smallest relative variability. By combining these considerations with the simulation study findings, we get a further confirmation of the general coherence with the guidelines provided by the flow tree.

5 Conclusions

In the paper we provide a flow tree procedure for the discovery of bivariate dependence relationships. Beside the commonly used correlation coefficients, i.e., the Pearson's, Spearman's and Kendall's coefficients, the assessment is extended also to a recently revisited monotonic dependence measure, called MDC coefficient. The procedure is obtained through the construction of a Monte Carlo simulation study, involving both the case of data generated from Normal and non-Normal distributions and observed according to three main scenarios. Scenario a) addresses the continuous variables, while scenarios b) and c) deal with the presence of a continuous and a discrete variables. Scenarios b) and c) are further extended by considering the case where the discrete variable is transformed into continuous variable through a specific "continuousation" approach.

The absence of knowledge on the distribution generating the data leads to prefer the MDC coefficient, similar or better than the Pearson's correlation coefficient and always better than the Spearman's and Kendall's coefficients, especially if data are continuous and normally distributed or mixed with an underlying non-Normal distribution.
Our proposal is useful not only from a theoretical view point but also from its professional implications. Through our flow tree we develop clear guidelines for the dependence measure optimal choice reducing the risk of biased results. We believe that the proposed guidelines may find outstanding interest and applicability in many research fields, such as for instance in Psychology, Education and Sociology, where the collection and the analysis of large amounts of mixed and non-normally distributed data are involved.

Acknowledgements Authors are grateful to the two anonymous reviewers for their helpful comments and suggestions.

References

1. Barbiero, A., Ferrari, P.A.: R package GenOrd. https://cran.r-project.org/web/packages/GenOrd/GenOrd.pdf) (2015)
2. Bedrickt, E.J. (1995). A note on the attenuation of correlation, *Educational & Psychological Measurement*, 48, 271-280 (1995). https://doi.org/10.1111/j.2044-8317.1995.tb01064.x
3. Bishara, A.J., Hittner, J.B.: Reducing Bias and Error in the Correlation Coefficient Due to Nonnormality. Educational & Psychological Measurement 75(5), 785-804 (2015). https://doi.org/10.1177/0013164414557639
4. Bishara, A.J., Li , J., Nash, T.: Asymptotic confidence intervals for the Pearson correlation via skewness and kurtosis. British Journal of Mathematical and Statistical Psychology 71(1), 167-185 (2018). https://doi.org/10.1111/bmsp.12113
5. Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., Bendayan, R.: Skewness and kurtosis in real data samples. Methodology. European Journal of Research Methods for the Behavioral and Social Sciences 9, 78-84 (2013). https://doi.org/10.1027/1614-2241/a000057
6. Denuit, M., Lambert, P.: Constraints on concordance measures in bivariate discrete data. Journal of Multivariate Analysis 93, 40-57 (2005). https://doi.org/10.1016/j.jmva.2004.01.004
7. Ferrari, P.A., Barbiero, A.: Simulating ordinal data. Multivariate Behavioral Research 47(4), 566-589 (2012). https://doi.org/10.1080/00273171.2012.692630
8. Ferrari, P.A., Raffinetti, E.: A different approach to Dependence Analysis. Multivariate Behavioral Research 50(2), 248-264 (2015). https://doi.org/10.1080/00273171.2014.973099
9. Fleishman, A. I.: A method for simulating non-normal distributions. Psychometrika 43, 521-532 (1978)
10. Goodman, L.A, Kruskal, W.H.: Measures of Association for Cross Classifications. Journal of the American Statistical Association 49(268), 732-764 (1954). https://doi.org/10.2307/2281536
11. Heathcote, A, Brown, S., Wagenmakers E.J., Eidels, A.: Distribution-free tests of stochastic dominance for small samples. Journal of Mathematical Psychology 54, 454-463 (2010). https://doi.org/10.1016/j.jmp.2010.06.005
12. Kendall, M.: A New Measure of Rank Correlation. Biometrika 30(1-2), 81-89 (1938). https://doi.org/10.1093/biomet/30.1-2.81
13. Likert, R.: A technique for the measurement of attitudes. Archives of Psychology 22(140), 5-55 (1932)

14. Lorenz, M.O.: Methods of measuring the concentration of wealth. Publications of the American Statistical Association 9(70), 209-219 (1905)
15. Mari, D.D., Kotz, S.: Correlation & Dependence. World Scientific (2001)
16. Page, E.B.: Ordered hypotheses for multiple treatments: A significance test for linear ranks.Journal of the American Statistical Association 58(301), 216-230 (1963).https://doi.org/10.2307/2282965
17. Pearson, K.: Mathematical Contributions to the Theory of Evolution. XVI. On Further Methods of Determining Correlation. Draper's Research Memoirs, Biometric Series, IV. Cambridge University Press (1907)
18. Raffinetti, E.: A Note on the Dependence Measurement for Ordinal-Continuous Data. Biostatistics and Biometrics Open Access Journal, 9(5), 129-134 (2019). https://doi.org/10.19080/BBOAJ.2019.09.555775.
19. Rasmussen, J.L.: Estimating correlation coefficients: Bootstrap and parametric approaches. Psychological Bulletin 101, 136-139 (1987). https://doi.org/10.1037/0033-2909.101.1.136
20. Schezhtman, E., Yitzhaki, S.: A Measure of Association Based On Gini's Mean Difference. Communications in Statistics-Theory and Methods 16(1), 207-231 (1987). https://doi.org/10.1080/03610928708829359
21. Sideridis, G.D., Simos, P.: What is the Actual Correlation Between Expressive and Receptive Measures of Vocabulary? Approximating the Sampling Distribution of the Correlation Coefficient Using the Bootstrapping Method. The International Journal of Educational & Psychological Assessment 5, 117-133 (2010)
22. Solaro, N., Barbiero, A., Manzi, G., Ferrari, P.A.: A sequential distance-based approach for imputing missing data: Forward Imputation. Advances in Data Analysis & Classification 11, 395-414 (2017). https://doi.org/10.1007/s11634-016-0243-0
23. Spearman, C.: The proof and measurement of correlation between two things. American Journal of Psychology, 15, 72-101 (1904). http://doi.org/10.2307/1412159
24. Vale, C.D., Maurelli, V.A.: Simulating multivariate nonnormal distributions. Psychometrika 48(3), 465-471 (1983). https://doi.org/10.1007/BF02293687
25. Zopluoglu, C.: Application in R: Generating Multivariate Non-normal Variables. https://www.dropbox.com/s/ldcu4f3mnf89fby/gennonnormal.r (2011)