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

MDC_{go} takes up the association/correlation challenge for grouped-ordinal data

Emanuela Raffinetti · **Fabio Aimar**

Received: date / Accepted: date

Abstract The subjective assessment of quality of life, personal skills and the agreement with a certain opinion are common issues in clinical, social, behavioral and marketing research. A wide set of surveys providing ordinal data arises. Beside of such variables, other common surveys generate responses on a continuous scale, where the variable actual point value cannot be observed since data belong to certain groups. This paper introduces a re-formalization of the recent "Monotonic Dependence Coefficient" (*MDC*) suitable to all frameworks in which, given two variables, the independent variable is expressed in ordinal categories and the dependent variable is grouped. We denote this novel coefficient with MDC_{go}. The MDC_{go} behavior and the scenarios in which it presents better performance with respect to the alternative correlation/association measures, such as Spearman's r_S, Kendall's τ_b and Somers' Δ coefficients, are explored through a Monte Carlo simulation study. Finally, to shed light on the usefulness of the proposal in real surveys, an application to drug-expenditure data is considered.

Keywords grouped-ordinal data · dependence · correlation coefficients · association coefficients · Monte Carlo simulations

Mathematics Subject Classification (2000) 62-07 · 62H20 · 62P25

1 Introduction

The bivariate data analysis is strongly supported in literature by a wide set of classical dependence measures, including Pearson's r (e.g. [19]), Spearman's r_S (e.g. [25]) and Kendall's τ_b (e.g. [11]) correlation coefficients. Pearson's correlation provides a measure of how well two quantitative variables move together linearly (e.g. [20]). If the two variables are expressed in terms of ordinal categories, the most popular measures of the association degree are Spearman's r_S and Kendall's τ_b correlation coefficients. Beside Spearman's r_S and Kendall's τ_b correlation coefficients, several dependence/association measures

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

Fabio Aimar
School of Management and Economics, University of Turin, C.so Unione Sovietica 218 Bis, 10134 Turin, Italy
ASL CN1, Via C. Boggio 12, 12100 Cuneo, Italy
E-mail: fabio.aimar@unito.it, fabio.aimar@aslcn1.it

for the analysis of ordinal data have been introduced in the literature (e.g. [1]). Examples are Goodman's γ (e.g. [8]), Stuart's τ_c (e.g. [24]) and Somers' Δ coefficients (e.g. [23]). Such measures differ in the treatment of ties; if Goodman's γ coefficient ignores tied data, Stuart's τ_c and Somers' Δ coefficients use a correction for ties.

Currently, we are assisting to an explosion in the availability of ordinal data due to widespread attitudinal surveys, typically recorded through Likert-type scales (e.g. [12, 16]). In many cases, survey scales are also built on responses that are observed to belong to certain groups on a continuous scale (grouped variable). The main trouble arising in the presence of grouped data is a result of the measurement process, since the actual point value of the underlying variable is unobserved. Given h groups, the measurement problem may be addressed by encoding each group through a label (from 1 to h) and, subsequently, by assigning rank one to all the units included in the first ordered group, rank two to those included in the second ordered group and finally rank h to those included in the h-th ordered group. In such a way, the assessment of the direct or inverse dependence relationship may be carried out through Spearman's r_S or Kendall's τ_b correlation coefficients. While Spearman's coefficient is defined as the correlation between the ranks of two variables, Kendall's coefficient accounts for the pairs of concordant and discordant values of two variables. Similarly, Goodman's γ, Stuart's τ_c and Somers' Δ coefficients may be computed by using the number of concordant and discordant pairs of observations. This results in neglecting the original continuous nature of the grouped variable, since the information from the grouped variable has to be reduced to its ordinal information, too. A crucial issue is then related to dependence relationship studies when one variable is ordinal and the other variable is grouped. The "Monotonic Dependence Coefficient" (*MDC*), recently proposed by [7] as a dependence measure for quantitative and ordinal/tied data, is here re-formalized for the case of grouped and ordinal variables. By considering the mid-point of each grouped variable level, the re-formalized *MDC* coefficient (now called MDC_{go}, where *go* is the acronym of grouped-ordinal data) allows consideration of the original continuous variable nature providing a novel tool for dependence analysis. Moreover, through a Monte Carlo simulation study, some basic hints about the MDC_{go} performance in specific scenarios are given even in comparison with Spearman's, Kendall's and Somers' coefficients. The choice of addressing the study to the correlation/association indices cited above is motivated by two reasons. Indeed, the MDC_{go} coefficient, even if indirectly, resorts to ranks as do Spearman's r_S. Moreover, the *MDC* coefficient, as discussed in [7], can also be intended as a map that measures the concordance degree between the considered variables. This feature makes the comparison with the Kendall's τ_b and Somers' Δ coefficients appropriate, especially because, as with Somers' coefficient, even the *MDC* coefficient is an asymmetric measure.

As a case-study, we consider the analysis of the drug expenditure incurred by the Italian system for public health assistance (Azienda Sanitaria Locale (ASL)). The public drug expenditure appears as a relevant kind of assistance that involves a national spending of 9 billion Euros per year (e.g. [4]). The survey focuses on a topic of great interest for the Italian public health assistance systems, the role of age differences in the allocation of drug expenditure (incurred by the ASLs) both by considering overall patients and single sub-groups, differing in terms of gender and age.

The paper is structured as follows. In Section 2, an overview of the literature on bivariate dependence coefficients is presented. In Section 3, the extension of *MDC* and its properties to the case of ordinal and grouped data, jointly with Monte Carlo simulation results, is discussed. In Section 4, an application to real data concerning the public health drug expenditure analysis is taken into account. Finally, in Section 5, conclusive comments are provided.

2 An overview of the main correlation/association coefficients

Given two variables Y and X, we may be interested in evaluating how they jointly move, assessing if they are concordant or discordant. A pair of observations is said to be concordant if the observation with the

larger value of X has the larger value for Y, and discordant if the observation with the larger value of X has the smaller value of Y (e.g. [6]). In our perspective, Y and X are the dependent and independent variables, respectively.

The most commonly applied indices in social studies are Spearman's r_S, Kendall's τ_b, Goodman's γ, Stuart's τ_c and Somers' Δ coefficients. All of them are non-parametric statistics describing the strength and direction of a monotonic relationship between the two variables Y and X. Spearman's r_S and Kendall's τ_b correlation coefficients, as well as Goodman's γ, Stuart's τ_c and Somers' Δ coefficients, typically refer to the case of ordinal variables.

Spearman's r_S correlation coefficient is a rank-based version of Pearson's correlation coefficient. Its estimate can be written as follows

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

where $\bar{r}(x)$ and $\bar{r}(y)$ are the average ranks of X and Y. Tied values are ranked by assigning a rank equal to the average of all the tied positions. Spearman's r_S correlation coefficient varies from -1 to +1 and the absolute value describes the strength of the monotonic relationship. The closer the absolute value of r_S is to 0, the weaker the monotonic relationship between the variables is. Spearman's correlation coefficient can be zero for variables that are related in a non-monotonic way.

Besides Spearman's r_S, association between ordinal variables can be measured by Kendall's τ_b correlation coefficient. Before introducing Kendall's τ_b correlation coefficient, a premise is needed. Let us denote with U and W the ranks of variables X and Y, respectively (with ties replaced with average ranks). A pair of points (U_i, W_i) is said to be concordant if one of the following conditions is fulfilled: $(U_i < U_j)$ and $(W_i < W_j)$ or $(U_i > U_j)$ and $(W_i > W_j)$. On the contrary, a pair of points (U_i, W_i) is said to be discordant if one of the following conditions is fulfilled: $(U_i < U_j)$ and $(W_i > W_j)$ or $(U_i > U_j)$ and $(W_i < W_j)$. Pairs in which $(U_i = U_j)$ or $(W_i = W_j)$ are ignored, since they are not classified as concordant or discordant. Now, we can define the values T_X and T_Y as $T_X = \sum_{i=1}^{S_X}(t_{(X)i}^2 - t_{(X)i})$ and $T_Y = \sum_{i=1}^{S_Y}(t_{(Y)i}^2 - t_{(Y)i})$, where $t_{(X)i}$ is the number of ties in the i-th set of ties of the variable X, $t_{(Y)i}$ is the number of ties in the i-th set of ties of variable Y, S_X and S_Y are the sets of ties in variables X and Y, respectively. Kendall's τ_b correlation coefficient estimate is given by

$$\tau_b = \frac{2(n_C - n_D)}{\sqrt{n(n-1)-T_X}\sqrt{n(n-1)-T_Y}}, \tag{2}$$

where n_C is the total number of concordant pairs and n_D is the total number of discordant pairs. As well as Spearman's r_S, Kendalls τ_b takes values in a close range $[-1,+1]$ with the absolute value indicating the strength of the monotonic relationship between the two considered variables.

Similarly to Kendall's τ_b correlation coefficient, other measures of association based on the difference between the number of concordant and discordant pairs are Stuart's τ_c, Somers' Δ and Goodman's γ coefficients. All these indices vary in the range $[-1,+1]$ with the same meaning of Spearman's r_S and Kendall's τ_b correlation coefficients. Here, we restrict the study on Somers' Δ index, this measure accounting for ties as opposed to Goodman's γ which ignores tied pairs, and assuming the specification of both the dependent and independent variables in line with the *MDC* coefficient (as shown below) but opposed to Stuart's τ_c.

Somers' Δ coefficient is defined as

$$\Delta = \frac{n_C - n_D}{n_C + n_D + n_{tiedX}}, \text{ where } n_{tiedX} \text{ is the number of tied pairs on } X. \tag{3}$$

Somers' Δ in Equation (3) is Goodman's γ coefficient modified to penalize for pairs tied only on X, if X is the independent variable. This is why Somers' Δ appears as an asymmetric measure, being the

number of tied pairs computed only over the independent variable.

In the case that only one of the two variables is ordinal and the other variable is continuous, the above dependence measures may appear as inappropriate for catching the monotonic dependence relationship. This because the Spearman's coefficient is built on the variable ranks and for this reason it does not take into account the actual values of the variable expressed on continuous scale. Similarly, the Kendall's and Somers' coefficients, which are determined by counting the pairs of observations that have concordant or discordant ranks, do not resort to the values of the continuous variable. Consequently, the use of these indices may produce a shrinkage of the existing dependence relationship strength. A solution to this problem was proposed by [7], who introduced the *MDC* coefficient. A similar coefficient, the Gini correlation, was originally formalized in terms of correlation measure by [5] and further developed by [22] especially when applied to the study of income inequality. Given two quantitative variables Y and X, in [7] the *MDC* index was formulated by resorting to the classical Lorenz and concordance curves (e.g. [13,15]). Since our aim does not extend to the interpretation of the *MDC* coefficient in terms of Lorenz curves, we re-express the *MDC* index by taking into account the original non-translated real valued Y variable. Thus, the *MDC* expression is defined as

$$MDC = \frac{2\sum_{i=1}^{n} iy_i^* - n(n+1)\mu_Y}{2\sum_{i=1}^{n} iy_i^{ord} - n(n+1)\mu_Y}, \tag{4}$$

where μ_Y is the Y variable mean value, y_i^{ord} represents the Y variable values ordered in non-decreasing sense and y_i^* represents the same Y variable values re-ordered according to ranks taken by the corresponding $\hat{Y}$ linear estimated values, obtained through to the least squares linear regression model $\hat{Y} = \hat{\alpha} + \hat{\beta}X$. Thus, contrary to the Spearman's, Kendall's and Somers' coefficients, *MDC* is built by considering the actual values taken by the dependent variable Y. Moreover, like the Somers' coefficient, the *MDC* index appears as an asymmetric measure since it assumes that the independent variable is well-specified.The *MDC* formula in Equation (4) can be made more familiar by providing an alternative expression based on the ratio of two correlation coefficients, $cor(i, y_i^*)$ and $cor(i, y_i^{ord})$. The proof is straightforward. By denoting with σ_i, $\sigma_{y^{ord}}$ and σ_{y^*} the standard deviations of i, y^{ord} and y^*, respectively, the *MDC* formula in terms of the correlation ratio becomes

$$MDC = \frac{\frac{cov(i,y^*)}{\sigma_i \sigma_{y^*}}}{\frac{cov(i,y^{ord})}{\sigma_i \sigma_{y^{ord}}}} = \frac{\frac{1}{n}\left[\sum_{i=1}^{n} iy_i^* - \frac{1}{n}\sum_{i=1}^{n} i \sum_{i=1}^{n} y_i^*\right]}{\frac{1}{n}\left[\sum_{i=1}^{n} iy_i^{ord} - \frac{1}{n}\sum_{i=1}^{n} i \sum_{i=1}^{n} y_i^{ord}\right]} = \frac{cor(i,y^*)}{cor(i,y^{ord})}. \tag{5}$$

By noting $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ and $\frac{1}{n}\sum_{i=1}^{n} y_i^{ord} = \frac{1}{n}\sum_{i=1}^{n} y_i^* = \mu_Y$, through some trivial re-arrangements, Equation (5) can be translated into Equation (4).

The measure originally proposed by [5] and further developed in [22] is similar to Equation (4), but it differs because the Y values are directly re-ordered according to the ranks (ordering) of X. In addition, [7] introduced an extension of such measure to the case of ordinal/tied data, in which a re-ordering problem may arise if some Y values are associated with equal $\hat{Y}$ values. Specifically, in such a case they replace the Y values, characterized by the same $\hat{Y}$ values, with their mean values. For the sake of clarity, a simple example is proposed. Let Y and X be two variables such that $Y = \{15, 10, 26, 21, 32, 45\}$ and $X = \{1, 2, 1, 3, 3, 2\}$. The procedure is as follows:

1. computing linear estimated values based on the least squares linear regression model $\hat{Y} = 18.833 + 3X$, i.e., $\hat{Y} = \{21.833, 24.833, 21.833, 27.833, 27.833, 24.833\}$;
2. since $\hat{Y}$ presents three pairs of equal values, the corresponding Y values are substituted according to their mean values, that is $Y = \{20.5, 27.5, 20.5, 26.5, 26.5, 27.5\}$, with $20.5 = \frac{15+26}{2}$, $27.5 = \frac{10+45}{2}$ and $26.5 = \frac{21+32}{2}$;

3. finally, the variable Y values, defined above, have to be ordered according to the ranks of the corresponding $\hat{Y}$ values to compute the numerator of the *MDC* index, while for the computation of the *MDC* index denominator it is sufficient to order the original Y variable values ($Y = \{15, 10, 26, 21, 32, 45\}$) non-decreasingly.

Similarly to Spearman's r_S, Kendall's τ_b and Somers' Δ, *MDC* appears as an oriented relative index, taking values in the close range $[-1, +1]$, whose extremes represent the two situations of perfect inverse or direct monotonic dependence, while the intermediate values, crossing zero in case of independence, are interpretable as previously described. More in detail,

- $MDC = +1$ if the ranks associated with the dependent variable values are the same as the ranks associated with their corresponding linear estimated values;
- $MDC = -1$ if the ranks associated with the dependent variable values are reverse with respect to the ranks associated with their corresponding linear estimated values;
- $MDC = 0$ if the dependent variable values increase (or decrease) and the independent variable values do not increase or decrease (but they stay the same). Suppose, as discussed by [7], that in the linear regression model the equality $\hat{Y} = E(Y|X) = E(Y) = \mu_Y$ holds for every value of X. Since the $\hat{Y}$ values are all equal, they are characterized by the same ranks and consequently it derives that $y_i^* = \mu_Y$ $\forall i = 1, \ldots, n$. Thus, the numerator of (4) becomes $2\mu_Y \sum_{i=1}^{n} i - n(n+1)\mu_Y$. Since $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, then $2\mu_Y \sum_{i=1}^{n} i = n(n+1)\mu_Y$, providing that $MDC = 0$.

3 Methodology

In this section, an extension of the formalization of the *MDC* measure to the context of grouped-ordinal data is proposed. The presence of ordinal data allows for the use of Spearman's r_S, Kendall's τ_b, Somers' Δ and *MDC* coefficients. A Monte Carlo simulation study was led, proving the attitude of the re-formalized *MDC* coefficient in dealing with both grouped and ordinal variables.

3.1 The *MDC* re-formalization for grouped-ordinal data

Let Y be a grouped variable, defined according to h classes (or groups), and let X be an ordinal variable expressed by k ordinal categories. Typically, for the grouped variable, the tendency measure is represented by the central value of the class. Given h classes, we denote the Y central value of each class by $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_h$, where $\bar{y}_1 < \bar{y}_2 < \ldots < \bar{y}_h$. Based on these assumptions, the aggregate phenomenon described by the Y grouped variable can be re-formulated in terms of single units. Let us associate with each unit belonging to a specific class of the grouped variable, the central value of the class. Thus,

$$\bar{Y} = \left\{ \underbrace{\bar{y}_1, \ldots, \bar{y}_1}_{n_1}, \underbrace{\bar{y}_2, \ldots, \bar{y}_2}_{n_2 - n_1}, \ldots, \underbrace{\bar{y}_h, \ldots, \bar{y}_h}_{n_h - n_{h-1}} \right\} \tag{6}$$

is the vector of the central values of the grouped variable expressed in terms of individual values; n_1, $n_2 - n_1$, …, $n_h - n_{h-1}$ correspond to the absolute frequencies related to the first group and to the second group until the last group such that n_h equals the total amount of involved units, that is $n_h = n$. Specifically, if we refer to the grouped variable expressed in terms of individual values, n_1 represents the position of the last unit in the first group, n_2 is the position of the last unit in the second group, until n_h, which is the position of the last unit in the last group.

In order to evaluate the existence and strength of the monotonic dependence relationship between Y and X, both the approaches of [5, 22] and [7] are considered. Indeed, the re-formalized *MDC* is built by

taking into account the issue of tied values, as suggested by [7], and by selecting the re-ordering criterion proposed by [5,22]: ordering the $\bar{Y}$ values according to the ranks of the corresponding X values. Thus, let us define $r(x_i)$ as the rank of the i-th X value and let $\bar{Y}$ be re-arranged according to the ranks taken by the variable X. We denote by $\bar{y}_i^*$, the $\bar{y}_i$ values associated with the ordered i-th X value. The ordinal nature of X yields some of the X categories to take the same values. Therefore, the original $\bar{Y}$ values associated with the same ranks of X are replaced by their mean values. More precisely, given a variable X with k ordered categories,

$$\bar{Y}^* = \left\{ \underbrace{\bar{y}_1^*, \ldots, \bar{y}_1^*}_{m_1}, \underbrace{\bar{y}_2^*, \ldots, \bar{y}_2^*}_{m_2 - m_1}, \ldots, \underbrace{\bar{y}_k^*, \ldots, \bar{y}_k^*}_{m_k - m_{k-1}} \right\} \tag{7}$$

is then the vector of the means of all the $\bar{Y}$ values belonging to the same ordered category, $m_1, m_2 - m_1, \ldots, m_k - m_{k-1}$ correspond to the absolute frequencies related to the $\bar{Y}^*$ values associated with the first, second until the last ordered category and such that m_k equals to the total amount of the involved units (i.e., $n = m_k$). Finally, m_1 represents the position of the last unit in the class related to the $\bar{Y}^*$ values associated with the first ordered category, m_2 is the position of the last unit in the class related to the $\bar{Y}^*$ values associated with the second ordered category until m_k which defines the position of the last unit in the class related to the the $\bar{Y}^*$ values associated with the k-th ordered category.

The expression of the novel *MDC* index (MDC_{go}) becomes

$$MDC_{go} = \frac{2\sum_{z=1}^{k} \sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* - n(n+1)\bar{\mu}}{2\sum_{j=1}^{h} \sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j - n(n+1)\bar{\mu}}, \tag{8}$$

where $\bar{\mu}$ is the mean of $\bar{Y}$, $n_{j-1} = n_0 = 0$ for $j = 1$, $m_{z-1} = m_0 = 0$ for $z = 1$, $\bar{y}_j$ is such that $\bar{y}_j < \bar{y}_{j+1}$ $\forall j = 1, \ldots, h$ and $\bar{y}_z^*$ is the mean of all $\bar{Y}$ values whose corresponding X values have the same ranks (i.e., the mean of the $\bar{Y}$ values that belong to the same ordered category z, with $z = 1, \ldots, k$).

The MDC_{go} coefficient in (8) is well-defined, being the denominator different from zero. The only case in which the denominator takes zero value arises if there exists only a group, so that $h = 1$ and $\bar{Y}$ takes n_1 values all equal to $\bar{y}_1$. Thus, denominator in (8) becomes $2\sum_{i=1}^{n_1} i\bar{y}_1 - n(n+1)\bar{\mu}$ with $\bar{y}_1 = \bar{\mu}$. Since the number of the involved observations corresponds to n_1, $n_1 = n$ so that we get $2\sum_{i=1}^{n} i\bar{\mu} - n(n+1)\bar{\mu}$. Being $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, through some simple computations, it follows that $n(n+1)\bar{\mu} - n(n+1)\bar{\mu} = 0$. Since the presence of only a group does not make sense, we can conclude that the MDC_{go} is always well-defined.

Due to its construction, MDC_{go} appears as a more appropriate dependence index when dealing with grouped-ordinal data. This because it is computed by both considering the actual central values of each class (dependent variable) and then by re-ordering them according to the ranks of the ordinal independent variable X. Thus, as well as for the *MDC* coefficient, ranks are not directly involved in the MDC_{go} computation, but they are only used as a tool for the re-ordering process of the actual dependent variable values. Furthermore, the MDC_{go} coefficient preserves the feature of being an asymmetric measure, as does the original *MDC* coefficient, and fulfills the following properties.

Property 1 $-1 \leq MDC_{go} \leq +1$, meaning that in all intermediate situations the MDC_{go} index takes values always smaller than +1 or greater than -1, according to the existence of an increasing or decreasing dependence relationship between the variables.

Proof Following [7], inequalities

i) $\frac{2\sum_{z=1}^{k} \sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* - n(n+1)\bar{\mu}}{2\sum_{j=1}^{h} \sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j - n(n+1)\bar{\mu}} \leq +1$, and

ii) $\frac{2\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* - n(n+1)\bar{\mu}}{2\sum_{j=1}^{h}\sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j - n(n+1)\bar{\mu}} \geq -1$

have to be proven. Since the MDC_{go} denominator is always positive (i.e., $2\sum_{j=1}^{h}\sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j - n(n+1)\bar{\mu} > 0$), it trivially derives that inequality i) is equivalent to

$$\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* \leq \sum_{j=1}^{h}\sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j. \tag{9}$$

The term on the left side of inequality (9), i.e., $\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^*$, can be expressed as

$$\begin{aligned}\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* &= \sum_{i=1}^{m_1} i\bar{y}_1^* + \sum_{i=m_1+1}^{m_2} i\bar{y}_2^* + \ldots + \sum_{i=m_{k-1}+1}^{m_k} i\bar{y}_k^* \\ &= \underbrace{1\cdot\bar{y}_1^* + \ldots + m_1\bar{y}_1^*}_{m_1} + \underbrace{(m_1+1)\bar{y}_2^* + \ldots + m_2\bar{y}_2^*}_{m_2-m_1} + \ldots + \underbrace{(m_{k-1}+1)\bar{y}_k^* + \ldots + m_k\bar{y}_k^*}_{m_k-m_{k-1}} \\ &= \sum_{i=1}^{n} i\bar{u}_i^*,\end{aligned}$$

where $\bar{u}_i^* = \left\{\underbrace{\bar{y}_1^*,\ldots,\bar{y}_1^*}_{m_1},\underbrace{\bar{y}_2^*,\ldots,\bar{y}_2^*}_{m_2-m_1},\ldots,\underbrace{\bar{y}_k^*,\ldots,\bar{y}_k^*}_{m_k-m_{k-1}}\right\}$ and $n = \sum_{z=1}^{k}(m_z - m_{z-1})$, with $m_{z-1} = m_0 = 0$ if $z = 1$. Analogously, the term on the right side of inequality (9), $\sum_{j=1}^{h}\sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j$ can be written as $\sum_{i=1}^{n} i\bar{u}_i$, with $\bar{u}_i = \left\{\underbrace{\bar{y}_1,\ldots,\bar{y}_1}_{n_1},\underbrace{\bar{y}_2,\ldots,\bar{y}_2}_{n_2-n_1},\ldots,\underbrace{\bar{y}_h,\ldots,\bar{y}_h}_{n_h-n_{h-1}}\right\}$ and $n = \sum_{j=1}^{h}(n_j - n_{j-1})$, with $n_{j-1} = n_0 = 0$ if $j = 1$.

It is then sufficient to prove that

$$\sum_{i=1}^{n} i\bar{u}_i^* \leq \sum_{i=1}^{n} i\bar{u}_i. \tag{10}$$

From inequality $\sum_{w=1}^{i}\bar{u}_w^* \geq \sum_{w=1}^{i}\bar{u}_w$, it derives that $\sum_{i=1}^{n}\sum_{w=1}^{i}\bar{u}_w^* \geq \sum_{i=1}^{n}\sum_{w=1}^{i}\bar{u}_w$. Moreover, since $\sum_{i=1}^{n}\sum_{w=1}^{i}\bar{u}_w^* = n(n+1)\bar{\mu} - \sum_{i=1}^{n} i\bar{u}_i^*$ and $\sum_{i=1}^{n}\sum_{w=1}^{i}\bar{u}_w = n(n+1)\bar{\mu} - \sum_{i=1}^{n} i\bar{u}_i$, it follows that

$$\sum_{i=1}^{n} i\bar{u}_i^* \leq \sum_{i=1}^{n} i\bar{u}_i$$

and the result in inequality (10) is achieved.

Let us now consider inequality ii), from which it follows that

$$2\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* - n(n+1)\bar{\mu} \geq -2\sum_{j=1}^{h}\sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j + n(n+1)\bar{\mu} \tag{11}$$

Since, as previously shown, $\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* = \sum_{i=1}^{n} i\bar{u}_i^*$ and $\sum_{j=1}^{h}\sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j = \sum_{i=1}^{n} i\bar{u}_i$, inequality (11) becomes

$$2\sum_{i=1}^{n} i\bar{u}_i^* - n(n+1)\bar{\mu} \geq -2\sum_{i=1}^{n} i\bar{u}_i + n(n+1)\bar{\mu} \tag{12}$$

By noting that $\sum_{i=1}^{n}(n+1-i)\bar{u}_i = n(n+1)\bar{\mu} - \sum_{i=1}^{n} i\bar{u}_i$, it derives that $\sum_{i=1}^{n} i\bar{u}_i = n(n+1)\bar{\mu} - \sum_{i=1}^{n}(n+1-i)\bar{u}_i$. Since $\sum_{i=1}^{n}(n+1-i)\bar{u}_i = \sum_{i=1}^{n} i\bar{u}_{(n+1-i)}$, inequality (12) reduces to

$$\sum_{i=1}^{n} i\bar{u}_i^* \geq \sum_{i=1}^{n} i\bar{u}_{(n+1-i)}. \tag{13}$$

Finally, as the following inequality $\sum_{w=1}^{i} \bar{u}_w^* \leq \sum_{w=1}^{i} \bar{u}_{(n+1-w)}$ holds, then

$$\sum_{i=1}^{n}\sum_{w=1}^{i} \bar{u}_w^* \leq \sum_{i=1}^{n}\sum_{w=1}^{i} \bar{u}_{(n+1-w)}. \tag{14}$$

Due that $\sum_{i=1}^{n}\sum_{w=1}^{i} \bar{u}_{(n+1-w)} = n(n+1)\bar{\mu} - \sum_{i=1}^{n} i\bar{u}_{(n+1-i)}$ and $\sum_{i=1}^{n}\sum_{w=1}^{i} \bar{u}_w^* = n(n+1)\bar{\mu} - \sum_{i=1}^{n} i\bar{u}_i^*$, through some computations expression in (14) can be written as

$$\sum_{i=1}^{n} i\bar{u}_i^* \geq \sum_{i=1}^{n} i\bar{u}_{(n+1-i)}, \tag{15}$$

allowing the result in (13) to be obtained.

Specifically, MDC_{go} reaches the upper and lower bounds in the cases illustrated by Properties 2 and 3.

Property 2 $MDC_{go} = +1$ (perfect direct monotonic dependence) when

$$\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* = \sum_{j=1}^{h}\sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_j \Rightarrow \sum_{i=1}^{n} i\bar{u}_i^* = \sum_{i=1}^{n} i\bar{u}_i$$

that is if $k = h$, $n_j = m_j$, $\forall j = 1,\ldots,h$ and $r(\bar{u}_i^*) = r(\bar{u}_i)$, $\forall i = 1,\ldots,n$.

According to Property 2, the situation of perfect direct monotonic dependence is reached if the number h of groups equals the number k of considered ordered categories, and the units belonging to each j-th group are the same units belonging to the corresponding z-th ordered category. In this case, the position of the units within each group is preserved with respect to the re-ordering process based on the ranks of the respective ordered categories.

Property 3 $MDC_{go} = -1$ (perfect inverse monotonic dependence), when

$$\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* = \sum_{j=1}^{h}\sum_{i=n_{j-1}+1}^{n_j} (n+1-i)\bar{y}_j \Rightarrow \sum_{i=1}^{n} i\bar{u}_i^* = \sum_{i=1}^{n} i\bar{u}_{(n+1-i)}$$

that is if $k = h$, $n_j = m_{k+1-j}$, $\forall j = 1,\ldots,h$ and $r(\bar{u}_i^*) = r(\bar{u}_{(n+1-i)})$, $\forall i = 1,\ldots,n$.

Property 3 concerns the definition of the perfect inverse monotonic dependence relationship. Such a situation arises if the number h of groups equals the number k of considered ordered categories, and the units belonging to each j-th group are the same units belonging to the corresponding $(k+1-j)$-th ordered category. This is the scenario where the position of the units within each group overturns with respect to the re-ordering process based on the ranks of the respective ordered categories.
MDC_{go} crosses the value zero if no monotonic dependence relationship occurs between the considered variables, as stated by Property 4.

Property 4 $MDC_{go} = 0$ (lack of monotonic dependence), when

$$2\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* = n(n+1)\bar{\mu}$$

that is if $k = 1$ and all the $\bar{Y}$ variable values are characterized by the same rank of variable X. Indeed, $2\sum_{z=1}^{k}\sum_{i=m_{z-1}+1}^{m_z} i\bar{y}_z^* = 2\sum_{z=1}^{1}\sum_{i=1}^{m_1} i\bar{y}_1^* = 2\sum_{i=1}^{n} i\bar{\mu}$. Since $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, then $n(n+1)\bar{\mu} = 2\sum_{i=1}^{n} i\bar{\mu}$ and the result follows.

The finding included in Property 4 represents a theoretical scenario that rarely happens since it implies that all the units belong to a unique ordered category. However, the more we move away from this situation, the more the monotonic dependence relationship between the variables arises.

3.2 The MDC_{go} vs its challengers

In order to validate the features of the MDC_{go} index, with respect to Spearman's, Kendall's and Somers' coefficients, we used a Monte Carlo simulation study by considering specific experimental scenarios in the case the continuous nature of the variables is not preserved. We started by first considering samples from the bivariate Normal distribution, specifying the value to be taken by the pairwise correlation coefficient ρ. Subsequently, we proceeded to discretize one of the two variables and group the other variable. Finally, to take into account the effects associated with the presence of extreme values that typically characterize several real world scenarios, such as financial or health data, we replicated the same design of experiment by sampling from a bivariate t-distribution.

3.2.1 Bivariate Normal distribution vs t-distribution

Let $(Y,X) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where Y and X represent the dependent and independent variables. The bivariate Normal density function is defined as

$$f_{X,Y}(x,y) = \frac{1}{(2\pi)\sqrt{\det \Sigma}} \exp\Big(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(y-\mu)\Big), \quad x,y \in \mathbb{R}^2, \tag{16}$$

where $\boldsymbol{\mu} = [0,0]$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

are the mean vector and the correlation matrix defined by specific values of the pairwise correlation coefficient ρ and unit variances.

Although the bivariate Normal distribution appears as the most popular distribution among continuous distributions, many phenomena (especially in quantitative risk and insurance fields) are modeled through t-distributions (e.g. [21]). The t-distribution is a leptokurtic distribution, presenting higher kurtosis than the Normal distribution. Given a t-distribution and a Normal distribution with the same mean and variance, data coming from the t-distribution appear closer to the mean value or farther from the mean value than those typically normally distributed.

Let $(Y,X) \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with ν corresponding to the number of degrees of freedom. The bivariate t-density function is given by

$$f_{X,Y}(x,y) = \frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi\nu\sqrt{\det \Sigma}} \left(1 + \frac{(x-\mu)^T \Sigma^{-1}(y-\mu)}{\nu}\right)^{-\frac{\nu+2}{2}}, \quad x,y \in \mathbb{R}^2. \tag{17}$$

A clear overview of multivariate t-distributions can be found in [9], especially with regard to the data generation process. When generating data from bivariate t-distributions, two basic issues have to be taken into account: 1) even if the covariance matrices of the bivariate Normal distribution and t-distribution are differently defined, the correlation matrices Σ in Equation (16) and (17) have the same meaning, thus they define the strength of the existing relationship between the variables; 2) the parameter $\boldsymbol{\mu}$ in Equation (16) differs from the parameter $\boldsymbol{\mu}$ specified in (17). In order to take care of the correct mean when generating data from a bivariate t distribution, the adjustments illustrated by [9] have to be considered.

3.2.2 Simulation settings

Our Monte Carlo simulation study starts by first generating data from bivariate standard Normal distributions and then from t-distributions in order to assess the effect of kurtosis on the behavior of the most popular indices. The ν parameter for the number of degrees of freedom was fixed equal to $\nu = 4$ into the t-distribution data generation process. This choice is coherent with [9], who states that in many financial applications ν takes values between 3 and 5. The simulation study also considers three different cases of ρ: $\rho = \{0.2, 0.5, 0.8\}$, in order to include situations of low, mid and high degrees of dependence. For each of the three chosen ρ values, the following X variable discretization scenarios were taken into account:

(i) discretization with equal-width intervals (EW), in which all categories present the same width but different weights;
(ii) discretization with uniform probability (U), in which all categories present the same weights but different width;
(iii) discretization with asymmetrical probability (A), in which all categories present different width and weights, with greater weights on one side of the distribution.

Furthermore, the number k of categories was let vary in each discretization scenario by setting $k = \{3, 5, 7\}$. Consequently, the variable X discretized into three, five and seven equal-width categories, is denoted by EW3, EW5 and EW7, the variable X discretized into three, five and seven uniform categories is denoted by U3, U5 and U7, and finally, the variable X discretized into three, five and seven asymmetrical categories is denoted by A3, A5 and A7. The variable Y was transformed into a grouped variable characterized by h equal-width classes. In this case, the number h of groups was set by considering $h = \{3, 5, 7\}$. Finally, the sample size was chosen equal to $n = \{25, 50, 100, 500\}$ and the process was replicated 10,000 times to achieve a total number of 81 ($3 \times 3 \times 3 \times 3$) scenarios. Since a single scenario was accounted for the dependent variable Y (split into intervals of equal-width), more details are needed about the selected parameters concerning the way the independent variable X was discretized. In Table 1 both the number k of categories and the probabilities values p, defining the different distributions of the discretized variable X, are reported.

Table 1: Number k of categories for discretization with uniform probability p_u and asymmetrical probability p_a

k	p_u							p_a						
3	1/3	1/3	1/3					0.10	0.30	0.60				
5	1/5	1/5	1/5	1/5	1/5			0.05	0.10	0.15	0.20	0.50		
7	1/7	1/7	1/7	1/7	1/7	1/7	1/7	0.05	0.05	0.10	0.10	0.20	0.25	0.25

3.2.3 Simulation results

The results of the MDC_{go}, r_S, τ_b and Δ Monte Carlo estimates, related to the three different values of ρ (0.2,0.5,0.8), are displayed through the boxplots in Figures 1 and 2, in the case of the dependent variable

Y split into 3 equal-width classes, in Figures 3 and 4, in the case of the dependent variable Y split into 5 equal-width classes, and in Figures 5 and 6, in the case of the dependent variable Y split into 7 equal-width classes. In more detail, Figures 1, 3 and 5 refer to normally distributed data, while Figures 2, 4 and 6 are associated with data following a t-distribution. For the sake of brevity, we report only the findings for $n = 500$, since they are similar to those obtained with the other selected sample sizes ($n = \{25, 50, 100\}$)[1].

The performance of the correlation/association coefficients is assessed in terms of their median values resulting from the Monte Carlo simulation study. Specifically, the closer the median value is to the target correlation coefficient ρ, the better the performance of the measure is. The target coefficient ρ is graphically represented by a horizontal red and dashed line in Figures 1, 2, 3, 4, 5 and 6.

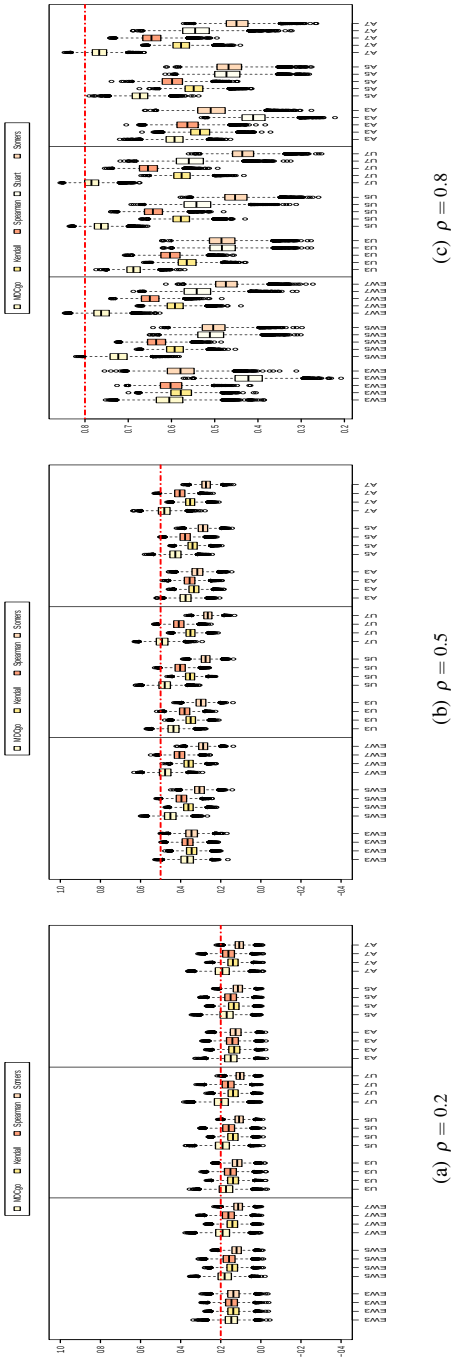
When data are normally distributed, the MDC_{go} index performs better than Spearman's, Kendall's and Somers' coefficients in most scenarios. Such behavior is also confirmed in terms of variability, here measured by the coefficient of variation (cv). The cv values of the three indices are graphically illustrated by the dot chart plots in Figure 7. In the case of a pairwise correlation coefficient $\rho = 0.2$, the four indices present a larger variability than in the cases in which the pairwise correlation coefficients are equal to 0.5 and 0.8. With respect to such a scenario, the MDC_{go} coefficient shows variability lower or at most equal to that of r_S, τ_b and Δ, except for the cases of Y split into three groups and X discretized into three equal-width categories, where the variability of MDC_{go} is greater with respect to that of r_S and τ_b. For $\rho = 0.5$, the MDC_{go} index is better or at most equal in terms of variability with respect to r_S and τ_b, except for the cases $Y = \{3,5,7\}$ and X discretized into three equal-width categories. Moreover, in the scenarios where $Y = \{5,7\}$ and X is discretized into three equal-width categories, MDC_{go} presents higher variability than Δ. Similar conclusions can be finally drawn when sampling with a pairwise correlation coefficient $\rho = 0.8$. In the case of t-distributed data, MDC_{go} seems to perform better than the challenger indices, but in some scenarios its variability is larger than that associated with results coming from Normal distributions, as displayed in Figure 8. Nevertheless the MDC_{go} coefficient has similar variability to that of the other indices in some scenarios, the cv values associated with the MDC_{go} index are smaller or higher than that of the other indices. If X is discretized into three, five and seven equal-width, uniform and asymmetrical categories, when $\rho = 0.5$ and $Y = \{3,5\}$ groups and when $\rho = 0.8$ and $Y = \{3,5,7\}$ groups, the MDC_{go} coefficient has the lowest variability. The variability of MDC_{go} increases with respect to the Spearman's and Kendall's correlation coefficients in most scenarios in which X is discretized into equal-width categories. This happens also with respect to the variability of the Somers index when $\rho = 0.2$, $Y = 3$ groups and X is discretized into 5 and 7 equal-width categories. Similar considerations arise when $\rho = \{0.5, 08\}$, $Y = \{5,7\}$ groups and X is defined in terms of seven equal-width categories.

Following [7] and in order to further validate the simulation results, an inferential analysis was led by resorting to Jonckheere's Trend Test (JT test) (e.g. [26, 10]) with the purpose of assessing if the differences in values of the indices can be considered significant. Given g independent populations with absolutely continuous underlying distributions, the JT test compares the location parameters μ_i ($i = 1, \ldots, g$) of g populations. There is no difference among g populations under the null hypothesis, while the distributions under the monotone ordering alternative differ by their location parameters μ_i, with $i = 1, \ldots, g$. The hypotheses are $H_0 : \mu_1 = \ldots = \mu_g$ against $H_1 : \mu_1 < \ldots < \mu_g$ or $H_1 : \mu_1 > \ldots > \mu_g$.

In our perspective, with the aim of avoiding that the often worse results with Kendall's and Somers' coefficients may strongly influence the test, single comparisons were performed by testing the alternative hypotheses $H_1 : r_S < MDC_{go}$, $H_1 : \tau_b < MDC_{go}$ and $H_1 : \Delta < MDC_{go}$[2]. For the sake of brevity, we report only the inferential findings about the r_S and the Δ coefficients, being that the hypothesis $H_1 : \tau_b < MDC_{go}$ is fulfilled, except for X discretized into three equal-width categories, $Y = 5$ and $\rho = \{0.2, 0.5, 0.8\}$, $Y = 7$ and $\rho = \{0.5, 0.8\}$ (with Normal distribution), and $Y = \{5,7\}$ and $\rho = \{0.2, 0.5, 0.8\}$ (with t-distribution). The inferential findings from $H_1 : r_s < MDC_{go}$ and $H_1 : \Delta < MDC_{go}$, for Y split into 3, 5, and 7 groups, are shown in Tables 2 and 3, i.e. when X is discretized into 3, 5 and 7 equal-width categories, Tables 4

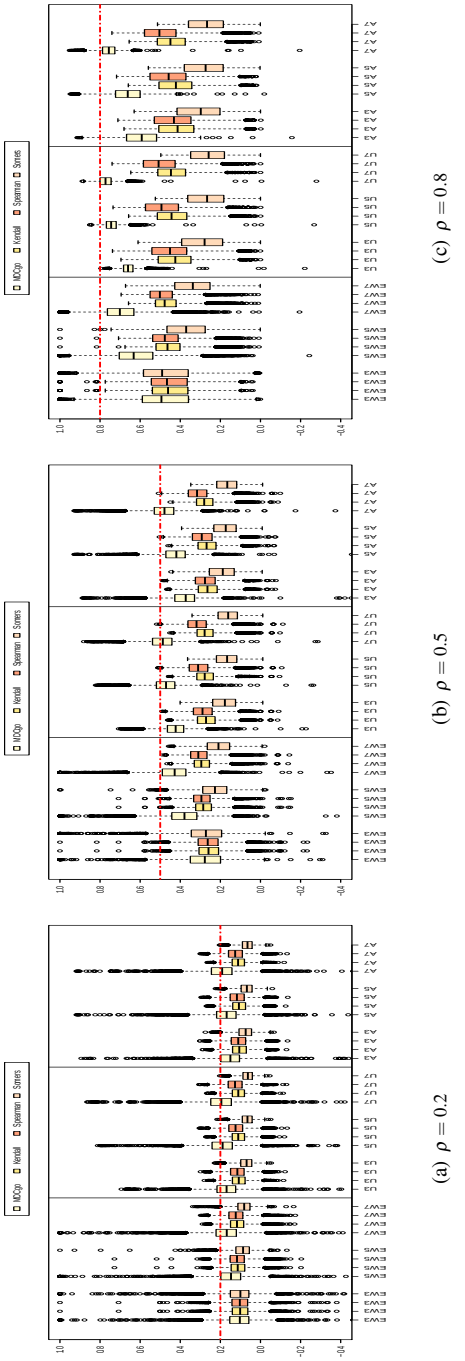
[1] Boxplots referring to $n = 50$ may be provided upon request.

[2] Even if the JT test is typically used when three or more populations are considered, it can be used for just two populations.



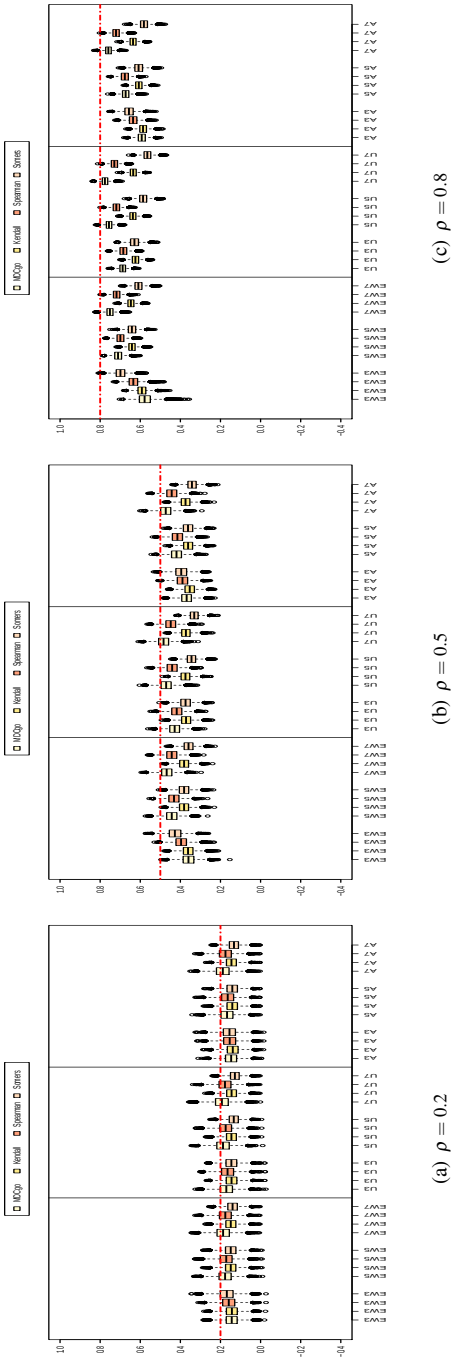
(a) $\rho = 0.2$ (b) $\rho = 0.5$ (c) $\rho = 0.8$

Fig. 1: Y split into 3 groups and X discretized in EW, U and A categories - Normal distribution



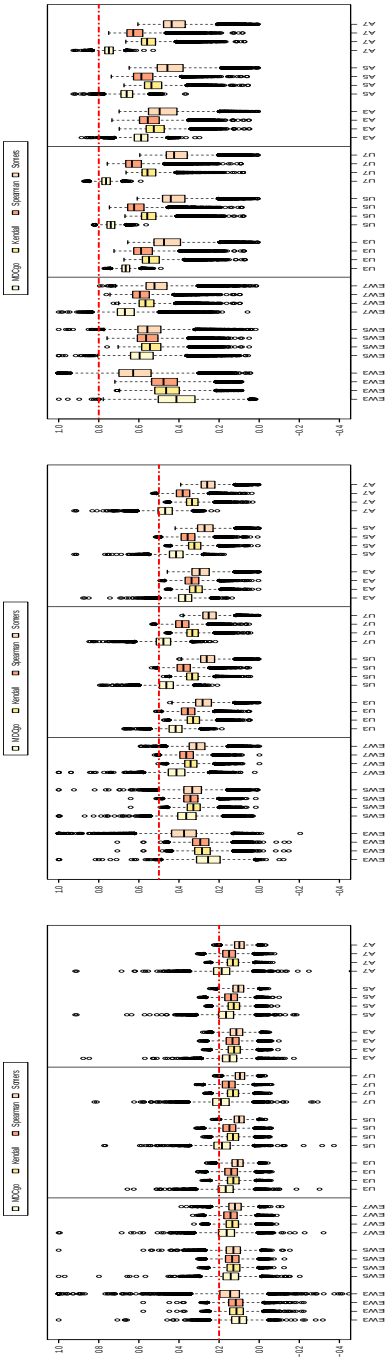
(a) $\rho = 0.2$ (b) $\rho = 0.5$ (c) $\rho = 0.8$

Fig. 2: Y split into 3 groups and X discretized in EW, U and A categories - t-distribution



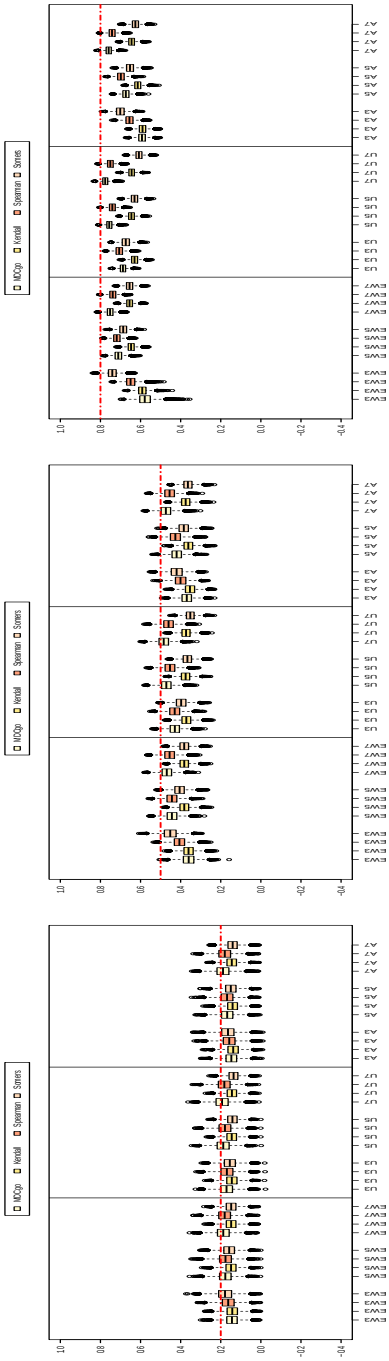
(a) $\rho = 0.2$ (b) $\rho = 0.5$ (c) $\rho = 0.8$

Fig. 3: Y split into 5 groups and X discretized in EW, U and A categories - Normal distribution



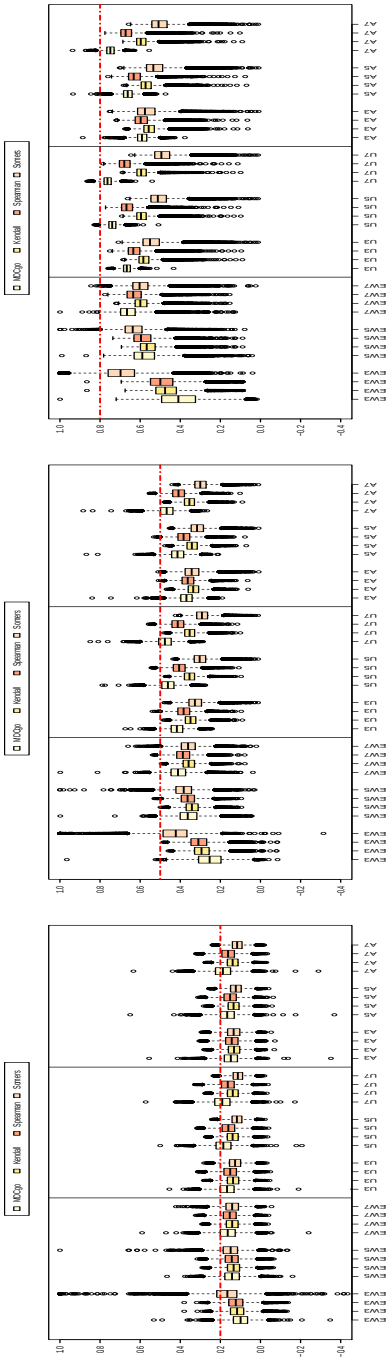
(a) $\rho = 0.2$ (b) $\rho = 0.5$ (c) $\rho = 0.8$

Fig. 4: Y split into 5 groups and X discretized in EW, U and A categories - t-distribution



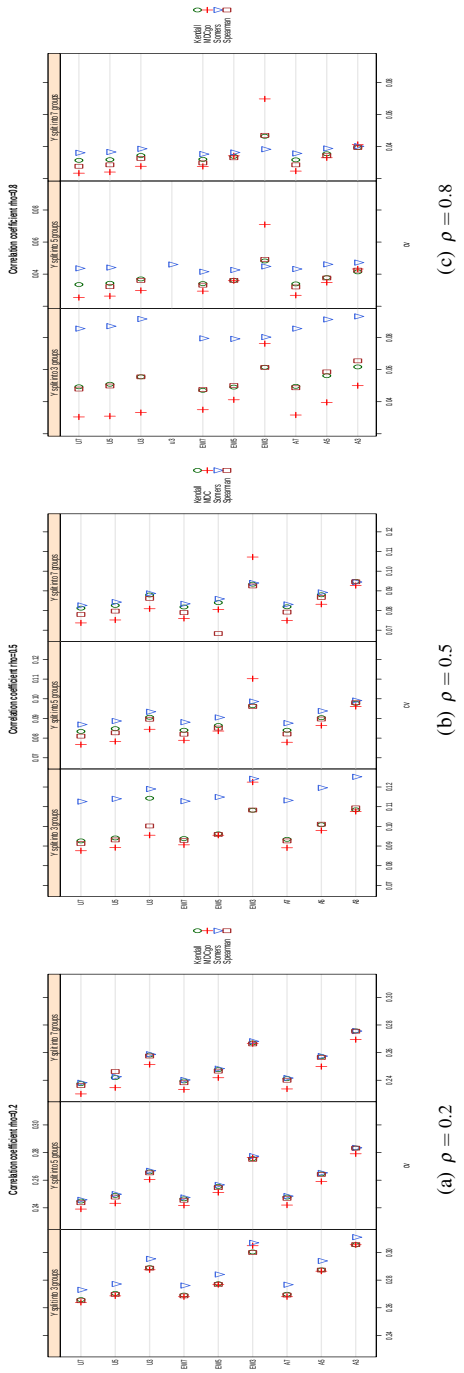
(a) $\rho = 0.2$ (b) $\rho = 0.5$ (c) $\rho = 0.8$

Fig. 5: Y split into 7 groups and X discretized in EW, U and A categories - Normal distribution



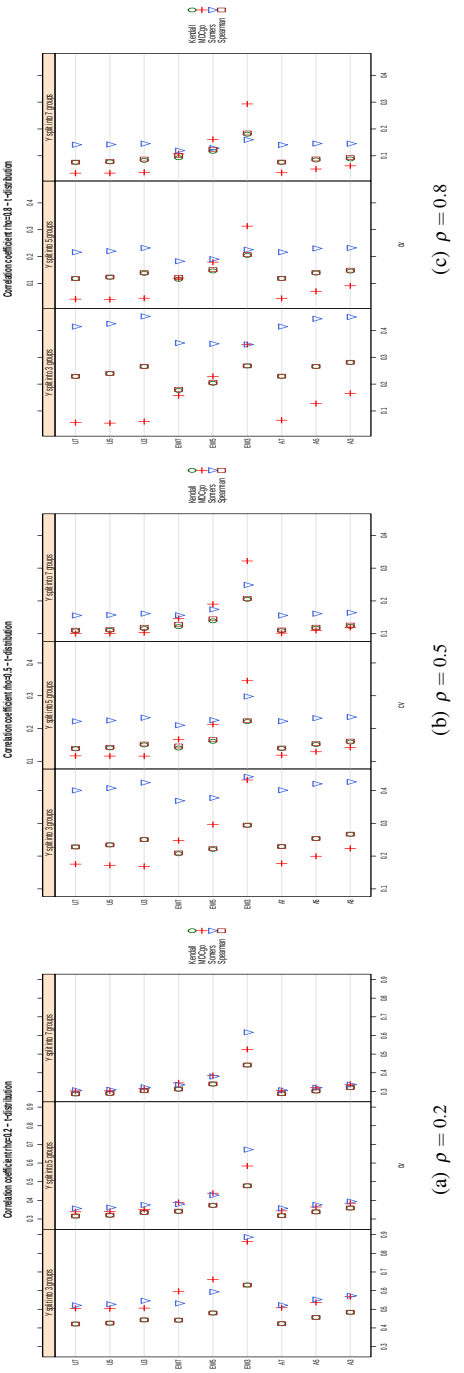
(a) $\rho = 0.2$ (b) $\rho = 0.5$ (c) $\rho = 0.8$

Fig. 6: Y split into 7 groups and X discretized in EW, U and A categories t-distribution



(a) $\rho = 0.2$ (b) $\rho = 0.5$ (c) $\rho = 0.8$

Fig. 7: cv for Y split into $\{3,5,7\}$ groups and X discretized into EW, U and A categories - Normal distribution



(a) $\rho = 0.2$ (b) $\rho = 0.5$ (c) $\rho = 0.8$

Fig. 8: cv for Y split into $\{3,5,7\}$ groups and X discretized into EW, U and A categories - t-distribution

and 5, i.e. when X is discretized into 3, 5 and 7 uniform categories, and Tables 6 and 7, i.e. when X is discretized into 3, 5 and 7 asymmetrical categories[3]. Note that, in all cases where the p-values associated with the different pairwise correlation coefficients $\rho = \{0.2, 0.5, 0.8\}$ are smaller than the selected significance level $\alpha = 0.05$, the corresponding table cells are colored in green. On the contrary, if at least a p-value takes a value greater than 0.05, table cells are colored yellow. Finally, if all p-values are higher than 0.05, the corresponding table cells are red. The inferential results organized according to the rule specified above work like a kind of traffic light for directing researchers in the use of the MDC_{go} index: by means of colored lights, typically red for stop, green for go, and yellow for proceed with caution.

Generally, the null hypothesis is rejected both in terms of Spearman's and Somers' coefficients, highlighting the better attitude of the MDC_{go} in preserving the original dependence relationship between the variables Y and X both when the data are normally and t-distributed. In the case of data whose underlying distribution is Normal, the scenarios where the null hypotheses $H_0 : r_S = MDC_{go}$ and $H_0 : \Delta = MDC_{go}$ are accepted arise when $\rho = \{0.2, 0.5, 0.8\}$, $Y = \{5, 7\}$ groups and X is discretized according to three equal-width categories; $\rho = \{0.2, 0.5, 0.8\}$, $Y = \{5, 7\}$ and X is discretized according to three asymmetrical categories. In addition, MDC_{go} performs as well as r_S if $\rho = 0.2$, Y is split into three groups and X is discretized into three equal-width categories; $\rho = 0.8$, Y is split into five groups and X is discretized according to five asymmetrical categories; $\rho = \{0.2, 0.5, 0.8\}$, Y is split into seven groups and X is discretized into five equal-width categories; $\rho = \{0.2, 0.5, 0.8\}$, Y is split into seven groups and X is discretized into five asymmetrical categories; $\rho = \{0.5, 0.8\}$, Y is split into seven groups and X is discretized into three uniform categories. Based on the above comments, it is interesting to compare the inferential results regarding the case of normally distributed data with those related to t-distributed data. When variable X is discretized into equal-width categories, the null hypothesis $H_0 : r_S = MDC_{go}$ is accepted in the same scenarios discussed for normally distributed data. This condition does not hold if $\rho = 0.2$, $Y = 3$ groups and X is discretized into three equal-width categories. Indeed, in this case MDC_{go} reaches values greater than Spearman's correlation coefficient. The null hypothesis $H_0 : \Delta = MDC_{go}$ is accepted when X is expressed into three equal-width categories, $\rho = 0.8$ and $Y = 3$ and if $\rho = \{0.2, 0.5, 0.8\}$ and $Y = \{5, 7\}$ groups, as in the case of Normal distributions, but also when X is defined through 5 equal-width categories. Thus, with respect to the case of Normal distributions, MDC_{go} reduces its performance in comparison with the Δ coefficient. In the cases of t-distributed data and X discretized into uniform categories, the null hypotheses $H_0 : r_S = MDC_{go}$ and $H_0 : \Delta = MDC_{go}$ are always rejected. Finally, if X appears as discretized through asymmetrical categories and data have t-distributions, the hypothesis $H_0 : \Delta = MDC_{go}$ is rejected in all the scenarios. This also happens for the hypothesis $H_0 : r_S = MDC_{go}$ which is always rejected except in the case of $\rho = 0.8$, $Y = 7$ groups and X discretized into three asymmetrical categories. Thus, when dealing with asymmetrical categories, the MDC_{go} index greatly improves its performance with respect to the case of normally distributed data. Such a conclusion generally also holds for the other scenarios, showing that the MDC_{go} is non-sensitive to the presence of leptokurtic data (with high kurtosis).

4 The case-study of the drug-expenditure for the ASL CN1 of Cuneo in Italy

In this section we further stress the effectiveness of our proposed measure through an illustrative example concerning the drug-expenditure and the so-called 'daily defined dose' (DDD) data provided by the ASL CN1 of Cuneo in Italy. The data, characterized by single drug prescriptions during the first quarter of 2014, cover 133,723 consumers of public drug prescriptions on a total amount of 420,000 patients, living over the ASL CN1 area. The purpose is investigating the phenomenon of the ASL CN1 pharmaceutical

[3] The JT test was led by resorting to the `R` package "`clinfun`" which uses the statistic $JT = \sum_{k<l} \sum_{ij} I(X_{ik} < X_{jl}) + 0.5I(X_{ik} = X_{jl})$, where i, j are observations in groups k and l respectively, and $I(\psi)$ equals one if ψ is true and zero otherwise. Since the JT test refers to a large sample size (i.e. the values obtained by the indices in 10,000 iterations), the p-values provided here are based on normal approximation of the standardized test statistic $Z = (JT - E(JT))/\sqrt{var(JT)}$.

expenditure via an exploratory analysis based only on MDC_{go} and Spearman's r_S coefficients, since from the led simulations, the remaining challengers are less powerful in catching monotonic dependence relationships.

The study was carried out by considering variables of specific demographic features of the patients (i.e., age and gender) and the associated amount of both the drug expenditure, expressed in Euros and supported by the Public Health Service (PHS), and the defined daily dose (DDD), intended as the drug consumption level. As reported by [2], 'The DDD represents the assumed average maintenance dose per day for a drug used for its main indication in adults'.

We aim at assessing and measuring the dependence relationship between the drug expenditure or DDD and the age of assisted patients. In addition to a gender-blind analysis, even the analysis identifying male and female patients was considered to account for gender differences in health care expenditures (e.g. [18, 14,3]). The available dataset is characterized by 59,022 males and 74,701 females, and the patient ages range from 0-102. We referred to the traditional OECD (Organisation for Economic Co-operation and Development) criteria to group the patients' ages as follows: $[0,15)$, $[15,45)$, $[45,65)$, $[65,75)$, $[75,85)$, $[85,95)$ and $[95,103)$ (e.g. [17]). To provide a better picture of the drug expenditure problem, our data were re-expressed in terms of the per-capita public health drug expenditure (PHDE) and the per-capita defined daily dose (DDD).

In our perspective, the dependent variable is per-capita PHDE or per-capita DDD, while the independent variable is patient age. In addition, the independent variable, split into classes according to traditional OECD criteria, was encoded according to ordinal labels. Consequently, seven ordered categories were defined by assigning numerical codes to the age categories, starting with code 1 for the age level $[0,15)$ and ending with code 7 for the final age level $[95,103)$. Following the methodology described in Subsection 3.1, the dependent variable PHDE, and subsequently DDD, were in turn split into three intervals of equal-width with the purpose of showing low, medium and high levels of PHDE or DDD associated with each age category. The thresholds of PHDE and DDD, used to delimit the groups (low, medium and high levels), are reported in Table 8. The comparison between MDC_{go} and r_S performance was carried out by computing the two indices on the original data (i.e., without splitting PHDE and DDD into intervals and discretizing the patients' ages) and on grouped-discretized data. The results related to PHDE are displayed in Table 9. MDC_{go} and r_S performances on the original data are similar and denote a strong monotone dependence relationship between PHDE and patient age. Nevertheless, MDC_{go} presents higher values than r_S in the case of grouped-ordinal data. Such findings suggest the better attitude of the MDC_{go} index in preserving information about the existence of the monotone dependence relationship if data are grouped and expressed according to ordered categories. Similar findings were obtained when the dependent variable was represented by DDD (see Table 10).

Table 2: p-values of the test for the null hypothesis $H_0 : r_S = MDC_{go}$ for $\rho = \{0.2, 0.5, 0.8\}$ and X discretized in equal-width categories

Y split into:	$H_1 : r_S < MDC_{go}$			$H_1 : r_S < MDC_{go}$ - ***t*-distribution**		
3 groups	$\rho = 0.2$ (p-value=0.141) $\rho = 0.5$ (p-value=0.002) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value=0.002) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)
5 groups	$\rho = 0.2$ (p-value>0.999) $\rho = 0.5$ (p-value>0.999) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value>0.999) $\rho = 0.5$ (p-value>0.999) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)
7 groups	$\rho = 0.2$ (p-value>0.999) $\rho = 0.5$ (p-value>0.999) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value=0.068) $\rho = 0.5$ (p-value=0.582) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value>0.999) $\rho = 0.5$ (p-value>0.999) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value=0.9719) $\rho = 0.5$ (p-value=0.9894) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)
X number of categories:	**3 equal-width categories**	**5 equal-width categories**	**7 equal-width categories**	**3 equal-width categories**	**5 equal-width categories**	**7 equal-width categories**

Table 3: p-values of the test for the null hypothesis $H_0 : \Delta = MDC_{go}$ for $\rho = \{0.2, 0.5, 0.8\}$ and X discretized in equal-width categories

Y split into:	$H_1 : \Delta < MDC_{go}$ - ***Normal distribution***			$H_1 : \Delta < MDC_{go}$ - ***t*-distribution**		
3 groups	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value=0.006) $\rho = 0.5$ (p-value=0.006) $\rho = 0.8$ (p-value=0.073)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)
5 groups	$\rho = 0.2$ (p-value>0.999) $\rho = 0.5$ (p-value>0.999) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value>0.999) $\rho = 0.5$ (p-value>0.999) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value>0.999) $\rho = 0.8$ (p-value<0.001)
7 groups	$\rho = 0.2$ (p-value>0.999) $\rho = 0.5$ (p-value>0.999) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value>0.999) $\rho = 0.5$ (p-value>0.999) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value>0.999) $\rho = 0.5$ (p-value>0.999) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)
X number of categories:	**3 equal-width categories**	**5 equal-width categories**	**7 equal-width categories**	**3 equal-width categories**	**5 equal-width categories**	**7 equal-width categories**

Table 4: p-values of the test for the null hypothesis $H_0 : r_S = MDC_{go}$ for $\rho = \{0.2, 0.5, 0.8\}$ and X discretized in uniform categories

Y split into:	$H_1 : r_S < MDC_{go}$ - ***Normal distribution***			$H_1 : r_S < MDC_{go}$ - ***t*-distribution**		
3 groups	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)
5 groups	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)
7 groups	$\rho = 0.2$ (p-value=0.004) $\rho = 0.5$ (p-value=0.941) $\rho = 0.8$ (p-value>0.999)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)	$\rho = 0.2$ (p-value<0.001) $\rho = 0.5$ (p-value<0.001) $\rho = 0.8$ (p-value<0.001)
X number of categories:	**3 uniform categories**	**5 uniform categories**	**7 uniform categories**	**3 uniform categories**	**5 uniform categories**	**7 uniform categories**

Table 5: p-values of the test for the null hypothesis $H_0 : \Delta = MDC_{go}$ for $\rho = \{0.2, 0.5, 0.8\}$ and X discretized in uniform categories

Y split into:	$H_1 : \Delta < MDC_{go}$ - *Normal distribution*			$H_1 : \Delta < MDC_{go}$ - *t-distribution*		
3 groups	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)
5 groups	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)
7 groups	$\rho = 0.2$ (p-value;0.001) $\rho = 0.5$ (p-value;0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)
X number of categories:	3 uniform categories	5 uniform categories	7 uniform categories	3 uniform categories	5 uniform categories	7 uniform categories

Table 6: p-values of the test for the null hypothesis $H_0 : r_S = MDC_{go}$ for $\rho = \{0.2, 0.5, 0.8\}$ and X discretized in asymmetrical categories

Y split into:	$H_1 : r_S < MDC_{go}$ - *Normal distribution*			$H_1 : r_S < MDC_{go}$ - *t-distribution*		
3 groups	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)
5 groups	$\rho = 0.2$ (p-value$>$0.999) $\rho = 0.5$ (p-value$>$0.999) $\rho = 0.8$ (p-value$>$0.999)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$>$0.999)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)
7 groups	$\rho = 0.2$ (p-value$>$0.999) $\rho = 0.5$ (p-value$>$0.999) $\rho = 0.8$ (p-value$>$0.999)	$\rho = 0.2$ (p-value=0.889) $\rho = 0.5$ (p-value$>$0.999) $\rho = 0.8$ (p-value$>$0.999)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$>$0.999)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)
X number of categories:	3 asymmetrical categories	5 asymmetrical categories	7 asymmetrical categories	3 asymmetrical categories	5 asymmetrical categories	7 asymmetrical categories

Table 7: p-values of the test for the null hypothesis $H_0 : \Delta = MDC_{go}$ for $\rho = \{0.2, 0.5, 0.8\}$ and X discretized in asymmetrical categories

Y split into:	$H_1 : \Delta < MDC_{go}$ - *Normal distribution*			$H_1 : \Delta < MDC_{go}$ - *t-distribution*		
3 groups	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)
5 groups	$\rho = 0.2$ (p-value$>$0.999) $\rho = 0.5$ (p-value$>$0.999) $\rho = 0.8$ (p-value$>$0.999)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)
7 groups	$\rho = 0.2$ (p-value$>$0.999) $\rho = 0.5$ (p-value$>$0.999) $\rho = 0.8$ (p-value$>$0.999)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)	$\rho = 0.2$ (p-value$<$0.001) $\rho = 0.5$ (p-value$<$0.001) $\rho = 0.8$ (p-value$<$0.001)
X number of categories:	3 asymmetrical categories	5 asymmetrical categories	7 asymmetrical categories	3 asymmetrical categories	5 asymmetrical categories	7 asymmetrical categories

Table 8: Thresholds of the PHDE and DDD delimiting the groups (low, medium and high levels)

PHDE (Euros)	**DDD (quantities)**
Low level	
$[0,50)$	$[0,120)$
Medium Level	
$[50,100)$	$[120,240)$
High level	
$[100,150)$	$[240,360)$

Table 9: MDC_{go} and r_S values to assess the dependence of PHDE on age and gender. A comparison between MDC_{go} and r_S performances on original and grouped-ordinal data

MDC_{go} (original data)	*r_S (original data)*	*MDC_{go} (grouped-ordinal data)*	*r_S (grouped-ordinal data)*
Public health drug-expenditure (PHDE) - Males and Females			
0.8942	0.9086	0.8643	0.8123
Public health drug-expenditure (PHDE) - Males			
0.8977	0.9072	0.8808	0.8375
Public health drug-expenditure (PHDE) - Females			
0.9004	0.9097	0.8830	0.7857

Table 10: MDC_{go} and r_S values to assess the dependence of DDD on age and gender. A comparison between the MDC_{go} and r_S performances on original and grouped-ordinal data

MDC_{go} (original data)	*r_S (original data)*	*MDC_{go} (grouped-ordinal data)*	*r_S (grouped-ordinal data)*
Defined daily dose (DDD) - Males and Females			
0.9318	0.9343	0.9147	0.8661
Defined daily dose (DDD) - Males			
0.9227	0.9307	0.9130	0.8583
Defined daily dose (DDD) - Females			
0.9428	0.9464	0.9063	0.8502

The relative loss of information with respect to original data was determined and reported in Table 11 (PHDE as the dependent variable) and Table 12 (DDD as the dependent variable).

Table 11: Relative loss of information with respect to original data for PHDE. A comparison between MDC_{go} and r_S values computed on grouped-ordinal data with respect to the corresponding values computed on original data

MDC_{go}	*r_S*
Public health drug-expenditure (PHDE) - Males and Females	
0.0335	0.1060
Public health drug-expenditure (PHDE) - Males	
0.0189	0.0768
Public health drug-expenditure (PHDE) - Females	
0.0194	0.1363

Typically, the grouping process of one of the two variables and the discretization process of the other variable translate into a shrinkage of the dependence relationship's strength detected on the original variables. This reduction is confirmed for both the MDC_{go} and r_S indices. Nevertheless, for both PHDE and DDD, the relative loss of information associated with the MDC_{go} index is always lower than 4% contrary

Table 12: Relative loss of information with respect to original data for DDD. A comparison between MDC_{go} and r_S values computed on grouped-ordinal data with respect to the corresponding values computed on original data

MDC_{go}	r_S
Defined daily dose (DDD) - Males and Females	
0.0184	0.0730
Defined daily dose (DDD) - Males	
0.0105	0.0814
Defined daily dose (DDD) - Females	
0.0387	0.1016

to the r_S index which achieves a minimum percentage of the relative loss of information at 7.30%. A second comparison between MDC_{go} and r_S behavior when considering grouped-ordinal data with respect to the same r_S value computed on the original data was introduced with the purpose of making the analysis more exhaustive. Results are shown in Tables 13 and 14 for PHDE and DDD, respectively. Even in this scenario, the maximum relative loss of information related to MDC_{go} is smaller than 5% against a minimum loss of information equal to 7.30% for Spearman's correlation coefficient.

Table 13: Relative loss of information with respect to original data for PHDE. A comparison between MDC_{go} and r_S values computed on grouped-ordinal data with respect to r_S values computed on the original data

MDC_{go}	r_S
Public health drug-expenditure (PHDE) - Males and Females	
0.0488	0.1060
Public health drug-expenditure (PHDE) - Males	
0.0291	0.0768
Public health drug-expenditure (PHDE) - Females	
0.0294	0.1363

Table 14: Relative loss of information with respect to original data for DDD. A comparison between MDC_{go} and r_S values computed on grouped-ordinal data with respect to r_S values computed on original data

MDC_{go}	r_S
Defined daily dose (DDD) - Males and Females	
0.0210	0.0730
Defined daily dose (DDD) - Males	
0.0229	0.0814
Defined daily dose (DDD) - Females	
0.0423	0.1016

5 Conclusions

In this paper, a re-formalization of the *MDC* index (now called MDC_{go}) for the case of grouped-ordinal data is proposed. It is based on the comparison between the central values of the grouped dependent variable and the same values properly re-ordered according to their relationship with the independent variable expressed in terms of ordered categories. MDC_{go} presents as a relative index, taking values in the close range $[-1,+1]$, with the same numerical features as the Spearman's, Kendall's and Somers' coefficients.

The behavior of MDC_{go} was analyzed through a Monte Carlo simulation study, and MDC_{go} estimates were compared with those of Spearman's, Kendall's, and Somers' coefficients. The performance of the

indices was evaluated by sampling from both bivariate Normal and t-distributions with different levels of correlation ($\rho = \{0.2, 0.5, 0.8\}$) and subsequently by grouping the dependent variable and discretizing the independent variable. The simulation results allow to detect the scenarios in which MDC_{go} is better than Spearman's, Kendall's and Somers' coefficients. As shown by the inferential analysis based on the JT test, on the 81 possible scenarios MDC_{go} performs better than Spearman's, Somers'and Kendall's coefficients in 42 cases if data have an underlying Normal distribution, and in 55 trials if data have an underlying t-distribution.

The MDC_{go} index is moreover validated on an application concerning the drug expenditure data provided by the ASL CN1 (Italy), whose purpose is to assess the linkage of PHDE and DDD with the patient age.

The obtained findings lead us to believe that the proposed methodology may find wide applicability in all research fields where data are not available in terms of point data but in terms of grouped-ordinal data, since appearing as a better tool for catching dependence relationships than the most popular correlation/association measures compared in the paper.

Acknowledgements The authors gratefully acknowledge the ASL CN1 of Cuneo (Italy) for making available the dataset representing the case-study illustrated and discussed in the paper. A special thanks goes to the reviewers for their helpful suggestions that allowed us to improve the quality of the paper.
Acknowledgements go to the Associate Editor and the two anonymous reviewers for their helpful comments and suggestions that allowed to improve the paper.

References

1. Agresti A, Analysis of ordinal categorical data, Second Edition, Wiley, New York (2010)
2. Ahrens W, Pigeot I, Handobook of Epidemiology. Springer-Verlag Berlin Heidelberg (2005)
3. Aimar F, Local reimbursed pharmaceutical expenditure monitoring through the use of statistical tools, Ph.D. Dissertation Thesis, University of Turin (2012)
4. Bernardini AC, Spandonaro F, Pharmaceutical Assistance: access to innovation, sustainability and selectivity, 10^{th} Health Report, Chapter 9, (in Italian) (2014)
5. Blitz, RC, Brittain, JA, An extension of the Lorenz diagram to the correlation of two variables, Metron, XXIII(14), 137143 (1964)
6. Denuit M, Lambert P, Constraints on concordance measures in bivariate discrete data, Journal of Multivariate Analysis, 93, 40-57 (2005)
7. Ferrari PA, Raffinetti E, A different approach to Dependence Analysis, Multivariate Behavioral Research, 50(2), 248-264 (2015)
8. Goodman LA, Kruskal WH, Measures of association for cross classifications, Journal of American Statistical Association, 49, 732764 (1954)
9. Hofert M, On Sampling from the Multivariate t Distribution, The R Journal 5(2), 129-136 (2013)
10. Jonckheere AR, A Distribution-Free k-Sample. Tests Against Ordered Alternatives, Biometrika, 41, 133-145 (1954)
11. Kendall K, New Measure of Rank Correlation, Biometrika, 30(12), 81-89 (1938)
12. Likert R, Technique for the measure of attitudes, Arch. Psycho., 22(140) (1932)
13. Lorenz MO, Methods of measuring the concentration of wealth, Publications of the American Statistical Association, 9(70), 209-219 (1905)
14. Martin J, Gonzales MPL, Garcia DC, Review of the literature of the determinants of healthcare expenditure, Applied Economics, 43, 19-46 (2011)
15. Marshall AW, Olkin I, Arnold CA, Inequalities: Theory of Majorization and Its Applications, Second Edition. Springer (2011)
16. Norman G, Likert scales, levels of measurement and the law of statistics, Advances in Health Science Education, 15, 625-632 (2010)
17. OECD, Estimating expenditure by desease, age and gender under the system of health accounts (SHA) framework, Final report, available at link `http://www.oecd.org/els/health-systems/EstimatingExpenditurebyDiseaseAgeandGender_FinalReport.pdf`, 2008.
18. Owens GM, Gender differences in health care expenditures, resource utilization, and quality of care, Journal of Managed Care Pharmacy, 14 (3 Suppl), 2-6 (2008)
19. 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, Cambridge (1907)
20. Rodgers JL, Nicewander WA, Thirteen Ways to Look at the Correlation Coefficient, The American Statistician, 42(1), 59-66 (1988)

21. Roth M, On the Multivariate t Distribution, Report no.: LiTH-ISY-R-3059 (2013)
22. 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)
23. Somers RH, A new asymmetric measure of association for ordinal variables, American Sociological Review, 27 799811 (1962)
24. Stuart A, The estimation and comparison of strengths of association in contingency tables, Biometrika, 40, 105110 (1953)
25. Spearman C, The proof and measurement of correlation between two things, American Journal of Psychology, 15, 72-101 (1904)
26. Terpstra TJ, The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking, Indigationes Mathematicae, 14, 327-333 (1952)