The bivariate Rank-based Concordance Index for ordinal and tied data

Emanuela Raffinetti
Pier Alda Ferrari

Available at: https://works.bepress.com/emanuela_raffinetti/1/
The bivariate Rank-based Concordance Index for ordinal and tied data

Emanuela Raffinetti and Pier Alda Ferrari

Department of Economics, Management and Quantitative Methods
Università degli Studi di Milano
Via Conservatorio 7, 20122 Milano (Italy)
e-mail: emanuela.raffinetti@unimi.it, pieralda.ferrari@unimi.it

Abstract

The focus of this paper is to illustrate a particular statistical tool for quantitative rational science research involving ordinal data. Due to the non-metric ordinal variable nature, the study of dependence relationship results as an open problem. A recent concordance measure, hereafter called “Bivariate Rank-based Concordance Index” (and denoted by $RCI$), is here revisited in order to provide a solution. Due to its features, $RCI$ can take the role of measuring dependence relationships between any real-valued dependent variable and a quantitative (even tied) or ordinal independent one.

Keywords: monotonic dependence relationship, ordinal data, uncorrelation and dependence curves.

1. Introduction

The explosion in the development of methods for analyzing categorical data that began in the 1960s has continued apace in recent years. This phenomenon has been implied by the increasing availability of data-sets with ordinal responses in the social, behavioral and biomedical sciences, as well as in public health, human genetics, education, marketing and industrial quality control.

In literature, several methods have been proposed to the analysis of those variables. Agresti (2010) mentions a wide set of statistical contributions when dealing with categorical data. Examples are logit and probit models (see Aldrich and Nelson, 1984 and McCullagh and Nelder, 1989), where categorical data analysis focuses on methods for categorical response variables
while independent variables may be qualitative or quantitative, as in ordinary regression.

In our context of analysis, the categorical variable nature is referred to the explanatory variable and not the response one. As well known, there exist different types of categorical data. In general, a categorical variable has a measurement scale consisting of a set of categories. Typically, categorical variables have two primary types of scale. Variables having categories without a natural ordering are called nominal: since in this case the order of listing the categories is not relevant, the statistical analysis does not depend on that ordering. However, many categorical variables do have ordered categories and for this reason are called ordinal. Ordinal variables have ordered categories but distances between categories are unknown. Typically, researchers use the quantitative nature of ordinal variables by assigning numerical scores to categories. Obviously, the arbitrary scale choice implies difficulties especially with regard to the dependence analysis. In fact, an usually occurring problem is related to the application of the most wide used dependence measures, since based on a metric scale and thus more appropriate in a quantitative context.

An alternative measure of dependence suitable for ordinal data is here proposed. More in detail, the concordance measure originally presented by Muliere (1986) for the analysis of taxation problems, and further on extended by Giudici and Raffinetti (2011) and Raffinetti and Giudici (2012) to the statistical field, is here revisited with regard to the bivariate context. Our and their proposal is based on some specific statistical tools, such as the Lorenz and concordance curves.

In this paper the index presented in Raffinetti and Giudici (2012), hereafter called “Bivariate Rank-based Concordance Index” (and denoted by RCI), is reconsidered to make it suitable in dependence analysis when dependent variable takes any real-value and independent variable is ordinal or quantitative with tied values. Furthermore, we show that this index is able to capture exactly any monotonic dependence relationship between the involved variables.

The remainder of this paper is organized as follows. In Section 2 the main features of the original concordance measure based on Lorenz curves and developed in order to solve the “equity” problem of a taxation with regard to income data is briefly recalled. Section 3 gives account of our proposal consisting on one hand in building the classical Lorenz curve for a real-valued variable and on the other hand as presenting the RCI as a
dependence measure. Section 4 illustrates the RCI’s adequacy in case of ordinal and also quantitative tied data. Section 5 reports an example of possible application of RCI with the purpose of showing how the proposed method can lead to efficient results with respect to its main competitors also in presence of grouped data. Finally, Section 6 is addressed to summary conclusions.

2. The RCI original formalization

In this section an overview of the original RCI formalization is provided starting from its employment in an a purely economical context (Muliere, 1986) up to its extension to a strictly statistical context where RCI was presented as a multivariate measure of concordance for a response variable Y and a set of k covariates (Raffinetti and Giudici, 2012).

In economical literature, Lorenz curves find wide application in all fields involving income data and the related taxation problem, in order to satisfy the “horizontal equity” condition according to which people who own the same income level have to be taxed for the same amount (see e.g. Musgrave, 1959). In such a context, an interesting contribution was given by Muliere (1986). Let X and Y be two random variables denoting the before and after taxation levels. The analysis was based on n pairs of ordered real values, (x_i, y_i), i = 1, 2, ..., n, whose components described respectively the i-th individual income levels before and after taxation. Being r(·) the rank of (·) and x_i ≠ x_j, with y_i ≠ y_j, i ≠ j, the situation of perfect horizontal equity is achieved when

\[ r(x_i) = r(y_i), \quad i = 1, 2, \ldots, n \]

whereas, the situation of perfect horizontal inequity occurs when

\[ r(y_i) = n + 1 - r(x_i), \quad i = 1, 2, \ldots, n. \]

Since the usual association indices, such as the Kendall’s \( \tau \), the Spearman’s \( \rho \) and the Gini index, remain unchanged also after the redistribution process in spite of each individual income extent substantially changed, a further measure suitable in quantifying the income extent was needed. For this reason, Muliere (1986) detected as possible solution to this problem a novel index, called “concordance index” (CI), built on specific statistical tools such as the Lorenz curves and the so called concordance curve.
Such measure assumes the expression provided below:

$$CI = \frac{2 \sum_{i=1}^{n} iy_i^* - n(n+1)M_Y}{2 \sum_{i=1}^{n} iy(i) - n(n+1)M_Y}.$$ (1)

In expression (1), $y_i^*$ is the after taxation income values $y$ ordered according to ranks assigned to the corresponding before taxation income values $x_i$ and $M_Y$ represents the after taxation income mean value.

Subsequently, by exploiting the same statistical tools, Raffinetti and Giudici (2012) reformulated this measure to interpret it as a multivariate concordance measure with regard to a multiple linear regression model. In this perspective a detailed description of Lorenz and concordance curves construction is provided. Let $Y$ and $X_1, \ldots, X_k$ be variables linked by a multiple linear regression model of $Y$ on $X_1, \ldots, X_k$. Let $L_Y$ be the response variable Lorenz curve characterized by the set of pairs $(i/n, (1/(nM_Y) \sum_{j=1}^{i} y(j)))$, where $y(i)$ are the observed values of $Y$ ordered in a non-decreasing sense, for $i = 1, \ldots, n$, and $M_Y$ the $Y$ mean value, and $L'_Y$ the corresponding dual Lorenz curve, characterized by the set of pairs ordered in reverse sense, i.e. $(i/n, (1/(nM_Y) \sum_{j=1}^{i} y(n+1-j)))$. The concordance curve is a similar curve obtained using the $Y$ values reordered according to the ranks of the $\hat{Y}$ values given by the least squares regression model $\hat{Y} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$. Such values ordered in this way are denoted by $y_i^*$. The set of pairs $(i/n, (1/(nM_Y) \sum_{j=1}^{i} y(j)))$ defines the $C$ concordance curve. An example of the $Y$ Lorenz curve, the $Y$ dual Lorenz curve and the $C$ curve are represented by the black line, the dashed line and the dotted line in Figure 1.

Through a direct comparison between the set of points that represent the Lorenz curve, $L_Y$, and the set of points that represent the concordance curve, one can see that a perfect “overlap” occurs if and only if

$$\sum_{j=1}^{i} y(j) = \sum_{j=1}^{i} y_j^* \text{ for every } i = 1, 2, \ldots, n,$$

that is if and only if $r(y_i) = r(\hat{y}_i)$.

Similarly, the comparison between the set of points that represent the $Y$ dual Lorenz curve, $L'_Y$, and the set of points that represent the concordance curve, allows us to conclude that a perfect “overlap” occurs if and only if
\[ \sum_{j=1}^{i} y_{(n+1-j)} = \sum_{j=1}^{i} y_{j}^* \text{ for every } i = 1, 2, \ldots, n, \]

that is if and only if \( r(y_{n+1-i}) = r(y_{i}).\)

Since it is possible to prove that (see Marshall and Olkin, 1979)

\[
\begin{cases} 
\sum_{j=1}^{i} y_{j}^* \geq \sum_{j=1}^{i} y_{(j)} \\
\sum_{j=1}^{i} y_{j}^* \leq \sum_{j=1}^{i} y_{(n+1-j)}
\end{cases}
\]

it results that \( L_Y \leq C \leq L_Y' \) and the concordance index, called “Multivariate Ranks-based Concordance Index” (and denoted here by RCI), can be defined as follows:

\[
RCI = \frac{\sum_{i=1}^{n-1} \left\{ i/n - (1/(nM_Y)) \sum_{j=1}^{i} y_{j}^* \right\}}{\sum_{i=1}^{n} \left\{ i/n - (1/(nM_Y)) \sum_{j=1}^{i} y_{(j)} \right\}} = \frac{2 \sum_{i=1}^{n} iy_{i}^* - n(n+1)M_Y}{2 \sum_{i=1}^{n} iy_{(i)} - n(n+1)M_Y}.
\]

Also satisfies the following property: \(-1 \leq RCI \leq +1\) (see Raffinetti and Giudici, 2012).
3. Proposal

By exploiting the original CI and RCI properties and features, our purpose is to provide a similar index able to assess the strength of a monotonic dependence relationship between a quantitative response variable and a quantitative (even tied) or ordinal covariate. In order to achieve such goal, two problems have to be solved. The first one is related to the need of considering also negative values for the response variable and then to provide an appropriate construction of the classical Lorenz curve in such a case. On the second hand an investigation about the role of dependence measure is provided.

3.1. Extension to any real-valued variable

CI and RCI suffer from a relevant restriction to non-negative response variable, since based on the classical Lorenz curve definition (see e.g. Lorenz, 1905). Nevertheless, in many cases the response variable is a continuous variable taking any real value. An extension of the Lorenz curve and the corresponding MRCI construction to any real-valued response variable and with regard to the bivariate case are advised.

More specifically, to reach this goal we resort to an appropriate translational procedure. Given a real-valued $Y$ variable, we add the minimum assumed negative value, denoted by $y^-$, to all values of the variable $Y$, so that the values of the new $Y^t$ variable are obtained by:

$$Y^t = (y_1 - y^-, y_2 - y^-, \ldots, y_n - y^-)$$

and the new variable $Y^t$ will be considered in the analysis. Thus, if for example a simple linear regression model of $Y^t$ on $X$ is applied, the ordinary least squared $\hat{Y}^t$ estimated values are computed and the $Y^t$ values are then ordered according to ranks of $\hat{Y}^t$ and pointed out with $\hat{y}^t_i$. The set of pairs $(i/n, (1/(nM_{Y^t})) \sum_{j=1}^{i} y^*_j))$ detects the $C$ concordance curve.

It is then needed to verify that all the properties beforehand described are satisfied by the new proposed index. In the translational context, the left terms of both the inequalities in (2) become $\sum_{j=1}^{i} y^*_j = \sum_{j=1}^{i} [y_j^* - y^-] = \sum_{j=1}^{i} y_j^* - iy^-$, while the right terms in both the inequalities in (2) respectively become $\sum_{j=1}^{i} y_{(j)} = \sum_{j=1}^{i} [y(j) - y^-] = \sum_{j=1}^{i} y_{(j)} - iy^-$ and $\sum_{j=1}^{i} y_{(n+1-j)} = \sum_{j=1}^{i} [y(n+1-j) - y^-] = \sum_{j=1}^{i} y_{(n+1-j)} - iy^-$. It is immediate to check that
both the inequalities in (2) still hold and thus the \( C \) curve position, i.e. \( L_{y^*} \leq C \leq L'_{y^*} \), is preserved.

An alternative expression of \( RCI \), based on the classical Lorenz curve of a real-valued response variable \( Y \), is defined as:

\[
RCI = \frac{\sum^{n-1}_{i=1} \left\{ \frac{i}{n} - \frac{1}{(nM_Y^t)} \sum^i_{j=1} y^*_j \right\}}{\sum^{n-1}_{i=1} \left\{ \frac{i}{n} - \frac{1}{(nM_Y^t)} \sum^i_{j=1} [y(j) - y^-] \right\} 
= \frac{\sum^{n-1}_{i=1} \left\{ \frac{i}{n} - \frac{1}{(n[M_Y - y^-])} \sum^i_{j=1} y^*_j \right\}}{\sum^{n-1}_{i=1} \left\{ \frac{i}{n} - \frac{1}{(n[M_Y - y^-])} \sum^i_{j=1} [y(j) - iy^-] \right\}},
\]

being \( M_Y^t = \frac{1}{n} \sum^n_{i=1} (y_i - y^-) = \frac{1}{n} (\sum^n_{i=1} y_i - ny^-) = \frac{1}{n} \sum^n_{i=1} y_i - y^- = M_Y - y^- \).

A more general expression of (4) can be obtained by setting \( y_0 = \min(0, y^-) \) and by multiplying both numerator and denominator of (4) for \( n[M_Y - y^-] \), i.e.

\[
RCI = \frac{(M_Y - y_0) \sum^n_{i=1} i - \sum^n_{i=1} \sum^i_{j=1} y^*_j}{(M_Y - y_0) \sum^n_{i=1} i - \sum^n_{i=1} \sum^i_{j=1} (y(j) - y_0)}.
\]

Since \( \sum^n_{i=1} i = \frac{n(n+1)}{2} \), (5) can be written as

\[
RCI = \frac{n(n+1)(M_Y - y_0) - 2 \sum^n_{i=1} \sum^i_{j=1} y^*_j}{n(n+1)(M_Y - y_0) - 2 \sum^n_{i=1} \sum^i_{j=1} (y(j) - y_0)}
\]

and by substituting \( \sum^n_{i=1} \sum^i_{j=1} y^*_j = n(n+1)(M_Y - y_0) - \sum^n_{i=1} iy^*_i \) and \( \sum^n_{i=1} \sum^i_{j=1} (y(j) - y_0) = n(n+1)(M_Y - y_0) - \sum^n_{i=1} i(y(i) - y_0) \) in (3.1), we have

\[
RCI = \frac{2 \sum^n_{i=1} iy^*_i - n(n+1)(M_Y - y_0)}{2 \sum^n_{i=1} i(y(i) - y_0) - n(n+1)(M_Y - y_0)}.
\]

Note that if the response variable \( Y \) is characterized by non-negative values, \( y^*_i = y_i \), \( y_0 = 0 \) and (6) corresponds to (3).

The original \( RCI \) properties converted into the translational context are reported below.
Property 1. The RCI assumes values in the $[-1, +1]$ close range.

Proof 1. To show that $RCI \leq +1$, it is sufficient to prove that $\sum_{i=1}^{n} iy_{(i)}^t \geq \sum_{i=1}^{n} iy_{(i)}^t$. Since from the first inequality in (2) it results $\sum_{j=1}^{i} y_{j}^* \geq \sum_{j=1}^{i} y_{(j)}$, by applying the translation procedure one obtains

$$\sum_{j=1}^{i} [y_{j}^* - y^-] \geq \sum_{j=1}^{i} [y_{(j)} - y^-] \Rightarrow \sum_{j=1}^{i} y_{j}^* \geq \sum_{j=1}^{i} y_{(j)}.$$  \hspace{2cm} (7)

Since (7) is intuitively true for all $i$, then one also has that

$$\sum_{i=1}^{n} \sum_{j=1}^{i} y_{j}^* \geq \sum_{i=1}^{n} \sum_{j=1}^{i} y_{(j)}.$$

Being $M_{Y}^t = M_{Y} - y_0$ and verified that

$$\sum_{i=1}^{n} \sum_{j=1}^{i} y_{j}^* = \sum_{i=1}^{n} (n+1-i)y_{i}^* = n(n+1)M_{Y}^t - \sum_{i=1}^{n} iy_{i}^*$$

and

$$\sum_{i=1}^{n} \sum_{j=1}^{i} y_{j}^* = \sum_{i=1}^{n} (n+1-i)y_{(i)}^* = n(n+1)M_{Y}^t - \sum_{i=1}^{n} iy_{(i)}^*$$

are jointly true, one can show that

$$n(n+1)M_{Y}^t - \sum_{i=1}^{n} iy_{i}^* \geq n(n+1)M_{Y}^t - \sum_{i=1}^{n} iy_{(i)}^*,$$

so that

$$\sum_{i=1}^{n} iy_{(i)}^* \geq \sum_{i=1}^{n} iy_{i}^*, \forall i = 1, \ldots, n,$$

where $\sum_{i=1}^{n} iy_{(i)}^* = \sum_{i=1}^{n} i(y_{(i)} - y_0)$.

To prove $RCI \geq -1$ one can note that

$$\sum_{i=1}^{n} (n+1-i)y_{(i)}^* = \sum_{i=1}^{n} y_{(n+1-i)}^i.$$
thus
\[ \sum_{i=1}^{n} iy_{(i)}^t = n(n + 1) M_Y - \sum_{i=1}^{n} y_{(n+1-i)}^t. \]

By applying this equivalence in the denominator of (6) one can provide an alternative but equivalent formulation of the RCI index based on the dual Lorenz curve \( L_Y' \):

\[ RCI = \frac{2 \sum_{i=1}^{n} iy_{(i)}^t - n(n + 1)(M_Y - y_0)}{n(n + 1)M_Y - 2 \sum_{i=1}^{n} y_{(n+1-i)}^t}, \]

or
\[ RCI = \frac{-2 \sum_{i=1}^{n} iy_{(i)}^t - n(n + 1)(M_Y - y_0)}{2 \sum_{i=1}^{n} y_{(n+1-i)}^t - n(n + 1)M_Y}. \]

According to the second inequality in (3), transformed according to the translational procedure, one gets that
\[ \sum_{j=1}^{i} y_{j}^t \leq \sum_{j=1}^{i} y_{(n+1-i)}^t \forall i, \]
and the result follows similarly to proof of Property 1. \( \square \)

**Property 2.** \( RCI = +1 \) if and only if the \( C \) curve perfectly overlaps with the response variable Lorenz curve.

**Proof 2.** In the translational scenario, \( RCI = +1 \) if and only if
\[ \sum_{j=1}^{i} y_{j}^t = \sum_{j=1}^{i} y_{(j)}^t, \]
that is if and only if \( r(y_{j}^t) = r(y_{j}^i), \forall i = 1, \ldots, n. \) \( \square \)

**Property 3.** \( RCI = -1 \) if and only if the \( C \) curve perfectly overlaps with the response variable dual Lorenz curve.

**Proof 3.** In the translational scenario, \( RCI = -1 \) if and only if
\[ \sum_{j=1}^{i} y_{j}^t = \sum_{j=1}^{i} y_{(n+1-j)}^t, \]
that is if and only if \( r(y_i^*t) = r(y_{n+1-i}^t), \forall i = 1, \ldots, n. \)

3.2. RCI as a dependence measure

RCI given in (6) is here interpreted as a dependence measure of the \( Y \) variable from the other one on the basis of the \( C \) curve position within the area that defines the Gini measure which is the result of a combination between the area within the egalitarian and the \( Y \) Lorenz curves and the area within the egalitarian and the \( Y \) dual Lorenz curve. In order to better clarify the \( C \) curve position meaning, we refer to three extreme scenarios, by referring to the case of a simple linear regression model with both quantitative variables.

Let then \( Y \) and \( X \) be respectively the dependent and the independent variables and \( C \) the concordance curve defined in Section 2. If we suppose the existence of a perfect positive linear relationship between \( Y \) and \( X \), the \( C \) curve, represented by the dashed curve (Figure 2), overlaps with the response variable Lorenz curve (also defined in Section 2) leading through (6) to \( RCI = +1 \). If, on the contrary, a reverse relation between \( Y \) and \( \hat{Y} \) occurs, a perfect overlapping between the \( C \) curve and the dual Lorenz curve is obtained (Figure 3) and \( RCI = -1 \).

Suppose now that in the linear regression model \( E(Y|X) = E(Y) \). Such situation represents the case of uncorrelation between \( Y \) and \( X \), then the \( \hat{Y} \) values are all equal and thus a problem related to the original values reordering arises. Conventionally and consistently with the described construction
and proposals already presented in literature (i.e. Spearman), we suggest to consider for each observed value their mean, making the ordering trivial. The concordance curve graphically corresponds to the set of pairs \((i/n, i/n)\) for \(i = 1, \ldots, n\) (Figure 4). We name it uncorrelation curve, since no providing any information about the existence of linear dependence relationship between the two involved variables.

Due to its meaning, the concordance or \(C\) curve can be called dependence curve for underlying the fact that it is built on the original response variable values reordered according to the existing dependence relation with the explanatory variable.

Until now we have illustrated our proposal with regard to linear relation. However, one can prove that it holds for any monotonic relation. This result is supported by the \(RCI\) construction. In fact, \(RCI\) is built by ordering the original response variable \(Y\) values with respect to ranks assigned to their corresponding estimated values \(\hat{Y}\).

Hereafter we present some examples highlighting the \(RCI\) attitude in catching any monotonic dependence relationship.

Let \(Y\) and \(X\) be linked by the polynomial relationship \(Y = X^6 - X^3 + X\). Let us suppose to choose the first ten integer values for covariate \(X\) and set the corresponding \(Y\) values. Through the least squares linear regression model, the expected \(\hat{Y}\) values are also calculated. Since \(\hat{Y}\) values maintain the same order of the corresponding \(Y\) values, \(RCI\) achieves value +1, while the corresponding Pearson’s \(r\)-correlation coefficient is 0.802. This result shows the real attitude of our index to catch in the exact way the existing perfect dependence relation between \(Y\) and \(X\), while the Pearson’s \(r\)-correlation coefficient catches it only in approximative way.

Analogous findings are achieved when a logarithmic \(Y = \log(X)\) or an exponential \(Y = e^X\) relationship is considered for \(Y\) and \(X\). In case of the logarithmic relationship the Pearson’s \(r\)-correlation coefficient is 0.952, whereas when the exponential relationship holds, the Pearson’s \(r\)-correlation coefficient is 0.717, versus \(RCI\) taking always its maximum value +1.

4. \(RCI\) adequacy for ordinal and tied data

In many fields, several interesting phenomena are described by ordinal variables. This happens, for instance, in many surveys concerning customer satisfaction, personal attitude and self-assessment programs. In such situations, the study of dependence relationship among variables represents an
attractive issue, since ordinal variables are not specified according to a metric scale and for this reason the corresponding arbitrary assigned values can affect the obtained results. In fact, results changing with respect to the different ordered scales imply the need of defining novel measures preserving the same conclusions in terms of existing dependence relationships, meaning that such measures have to be invariant with respect to quantification of categories of the variable. Our proposed RCI features allow us to satisfy this condition since RCI is based on the dependent variable $Y$ original values ordered according to ranks of the corresponding estimated values. We call this scale invariance property. The RCI invariance property can be described by the first example introduced in Subsection 3.2. According to such example, the covariate $X$ is characterized by the first ten integer values $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and the relationship between $Y$ and $X$ is expressed as $Y = X^6 - X^3 + X$. Let us now modify the previous covariate $X$ by considering the squared values, i.e. $X = \{1, 4, 9, 16, 25, 36, 49, 64, 81, 100\}$. Even if the corresponding Pearson’s $r$-correlation coefficient decreases taking respectively values 0.776 (polynomial relationship), 0.864 (logarithmic relationship) and 0.632 (exponential relationship) the RCI achieves one more time value $+1$. This conclusion confirms that the RCI’s invariance property holds.

The previous illustrated case neglects a relevant issue that concerns ordinal data. When considering ordinal covariates, the problem of tied categories can occur. The term tied data is used to point out data described by a frequency distribution. It can happen that some $X$ values are characterized by a frequency greater than one implying that the related regression estimates $\hat{Y}$ take the same value. It is worth noting that such aspect can appear also when the covariate is quantitative with repeated values. This condition generates some difficulties in defining the corresponding original $Y$ values ordering. In such a context, the rule introduced in Subsection 3.2 for the uncorrelation case is still once adopted: the original $Y$ values associated to the equal $\hat{Y}$ values are substituted by their mean. To better appreciate the RCI’s adequacy in dealing with ordinal or tied data, here we introduce an alternative expression. Let $r(\hat{y}_1) = \ldots = r(\hat{y}_{n_1}) < r(\hat{y}_{n_1+1}) = \ldots = r(\hat{y}_{n_2}) < \ldots < r(\hat{y}_{n_{j+1}}) = \ldots = r(\hat{y}_{n_k})$ be the sequence of the $Y$ linear estimated values ranks. In such a case, $k$ groups of equal linear estimated values are recognizable. Each $Y$ value characterized by equal rank belongs to the same group, that is the components are exactly $y_1, \ldots, y_{n_1}$ for group $j = 1$, $y_{n_1+1}, \ldots, y_{n_2}$ for group $j = 2$ until $y_{n_{j+1}}, \ldots, y_{n_k}$ for group $j = k$. Due to the re-ordering problem
related to $\hat{Y}$ values with the same ranks, within each group we compute the
$Y$ values mean value, such that $\bar{y}_1 = \frac{\sum_{i=1}^{n_1} y_i}{n_1}$ for group $j = 1$, $\bar{y}_2 = \frac{\sum_{i=n_1+1}^{n_2} y_i}{n_2-n_1}$
for group $j = 2$ until $\bar{y}_k = \frac{\sum_{i=n_j+1}^{n_k} y_i}{n_k-n_j}$ for group $j = k$. Then, all the $Y$ values
belonging to the same group are substituted by their group average value.

The $RCI$ index extended to the ordinal context, where $n = n_k$, can be then translated in the following expression:

$$RCI = \frac{2 \sum_{j=1}^{k} \sum_{i=n_{j-1}+1}^{n_j} i\bar{y}_i - n(n+1)(M_Y - y_0)}{2 \sum_{i=1}^{n} iy_i' - n(n+1)(M_Y - y_0)}.$$ (8)

In the following Section, the $RCI$ performance when ordinal or tied data are involved will be compared with that of the Pearson’s $r$-correlation coefficient.

5. Further investigations in application contexts

In this section we introduce an application of $RCI$ in order to make a comparison between its behavior and that associated to the Pearson’s $r$-correlation coefficient. We remind that the aim of such example is not addressed to provide a detailed analysis of all data involved and described in the considered dataset. Here, we simply show how the $RCI$ index can lead to more robust results with respect to the Pearson’s $r$-correlation coefficient in different scenarios.

Some information about the used dataset are due. The employed dataset is available as an SPSS Data file and it is called “Employee Data.sav”. The file contains data extracted from a bank’s employee records in an investigation into discrimination in 1987 and it is built on 473 statistical units.

As discussed above, $RCI$ can be usefully applied both in a quantitative context, when studying the dependence relationship between two quantitative variables, and in a context when the dependent variable is quantitative and the independent one is ordinal or discrete. With regard to this issue, here the focus is based on dependence of a variable $Y$, representing the beginning salary (in dollars), on another variable $X$, representing the individual education years. This variable takes values according to the frequency distribution represented in Table 1.

The purpose of this application is threefold. More precisely, we aim at comparing the $RCI$ performance with respect to that of $r$ under three spe-
Specific scenarios. Firstly, we take into account both the variables $Y$ and $X$ as directly provided by the dataset and representing the real situation of analysis. This in order to assess the existence of a dependence relationship between the beginning salary (continuous variable) and education years (discrete variable). Secondly, the comparison between $RCI$ and $r$ is carried out as if the original data, concerning the explanatory variable $X$, are classified into five groups whose frequency distribution is provided in Table 2. More in detail, the education years variable is re-expressed according to the previous groups and a rank scale such that employees included in group 1, 2, 3, 4 and 5 are characterized by an education level encoded respectively by 1, 2, 3, 4 and 5. Finally, the study is focused on considering the average value of education years within each group as available data. To be more thorough the education years ranges within each group are also reported.

<table>
<thead>
<tr>
<th>Groups</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute frequencies</td>
<td>243</td>
<td>122</td>
<td>69</td>
<td>36</td>
<td>3</td>
</tr>
<tr>
<td>Education years ranges</td>
<td>[8, 14)</td>
<td>[14, 16)</td>
<td>[16, 18)</td>
<td>[18, 20)</td>
<td>over 20</td>
</tr>
<tr>
<td>Average of education years</td>
<td>11.128</td>
<td>14.951</td>
<td>16.159</td>
<td>18.750</td>
<td>20.333</td>
</tr>
</tbody>
</table>

Table 2: Frequency distribution of groups and average of education years within each group

By serving in this way, the effect associated to transformation of the
X variable scale on both the indices capability in catching the dependence relationship is evaluated.

Results regarding the two analysis are presented in Table 3.

<table>
<thead>
<tr>
<th>X-Education Years expressed as:</th>
<th>Original data (a)</th>
<th>Five ordered categories (b)</th>
<th>Average at group (c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( RCI )</td>
<td>0.787</td>
<td>0.781</td>
<td>0.781</td>
</tr>
<tr>
<td>( r )</td>
<td>0.633</td>
<td>0.746</td>
<td>0.688</td>
</tr>
</tbody>
</table>

Table 3: \( RCI \) and \( r \) values to evaluate dependence of beginning salary on education years

With regard to original data (a), the value of \( RCI \) (0.787) is higher than \( r \) (0.633) and such difference is well marked. This because of the original discrete dependent variable nature. As well known, the Pearson’s \( r \)-correlation coefficient is sensitive to variables nature: typically, it shrinks with variables which are not continuous. \( RCI \) is not affected by the variables nature since it is based only on re-ordering the response variable values according to the corresponding linear estimates.

In the second case (b), when data about the explanatory variable are encoded into five ordered categories, if on one hand \( RCI \) reaches almost the same value (0.781), on the other hand \( r \) raises considerably (0.746), highlighting its sensitivity to scale transformation.

The third case (c) reports the results based on data available only in terms of average years of education within each group. If on one hand \( r \) gets worse (0.688) with respect to case (b), our proposed \( RCI \) remains unchanged in its value (0.781), highlighting once again the unreliability of the Pearson’s \( r \)-correlation coefficient.

The described real example confirms the \( RCI \) robustness in catching the dependence relationships even when one of the variable is expressed according to different measurement scales. Such issue supports the \( RCI \)’s adequacy in dependence relationship investigation with respect to the Pearson’s \( r \)-correlation coefficient which can lead, as previously shown, to misleading results.
6. Conclusions

In this contribution we revisited a concordance measure able to catch any monotonic relationship between a real-valued response variable and a numerical or ordinal independent variable (even tied). Such measure is invariant with respect to the quantification of categories for the ordinal variable.

The behavior of $RCI$ was also investigated through an application to real data. Results highlighted the $RCI$’s capability to catch the dependence relationship between the involved variables, preserving it also when pieces of information are lost due to the discretization process involving the independent variable.
References


