Rapporto di ricerca n. 218

The Bipolar Mean Dependece

Autori
Luca Bagnato, Walter Maffenini and Mariangela Zenga

Novembre 2011

Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali
Università degli Studi di Milano Bicocca
Via Bicocca degli Arcimboldi, 8 - 20126 Milano - Italia
Tel +39/02/64483102/3 - Fax +39/02/64483105

THE BIPOLAR MEAN DEPENDENCE

Luca Bagnato*, Walter Maffenini* and Mariangela Zenga*

SUMMARY

In this paper an alternative formulation of the mean deviation about the bipolar mean is provided. In addition to this, the concept of bipolar mean-independence is defined in order to evaluate how conditional bipolar means change according to the value of another variable jointly considered. This concept arises from the well-known notion of mean dependence but it exploits the idea of the bipolar mean which can be used also for qualitative ordinal variables. A normalized index that measures the strength of the bipolar mean-dependence is also provided. An application is presented which highlights the simplicity and interpretability of the proposed methodology.

Keywords: *Bipolar Mean, Mean Deviation S_η, Mean Dependence*

1. INTRODUCTION

For studying the dependence between two variables, the sphere of the possible methodologies that can be used naturally depends on the nature of the variables at hand. Usually, the more informative are the variables, the more are the techniques one can adopt in order to investigate different type of dependences. If, for example, all the variables are quantitative, several type of dependence can be investigated, both linear and nonlinear . Suppose that the goal is to analyze the conditional dependence of the quantitative variable X given Y of any type. One can study the mean dependence through the well-known correlation ratio of Pearson. If X was only ordinal we should not be able to measure the mean dependence since by definition an ordinal variable does not posse mean. In the most of the cases the first natural numbers are given to each ordinal category of the variable X. The problem of which values assign to the different categories has been treated by several authors such as Coombs (1953) and Tufte (1970) whose underlined that differences among categories should be seriously analyzed instead of simply assign equal distances among them.

In this work we define a new concept of conditional dependence based on the bipolar mean introduced by Maffenini and Zenga (2005). The aim is to extend the notion of mean dependence to the case in which the conditional variable is qualitative ordinal. The paper also presents an alternative definition of the mean deviation about the bipolar mean proposed in Maffenini and Zenga (2006). This new formulation does not require the frequency distribution associated to the bipolar mean. because it uses only cumulative and retro-cumulative frequencies,.

*Department of Quantitative Methods for Business and Economic Sciences - University of Milan-Bicocca - Piazza dell'Ateneo Nuovo, 1, 20126 MILANO (e-mail: luca.bagnato@unimib.it, walter.maffenini@unimib.it, mariangela.zenga@unimib.it).

The article is organized as follows. Section 2 presents the new formulation of the mean deviation about the bipolar mean. Section 3 introduces the concept of bipolar mean-independence (Section 3.1) and provides a normalized index to evaluate the strength of the possible bipolar mean-dependence (Section 3.2). In Section 4 an application about the level of attendance in the Italian macro-area is provided. Conclusions and discussion are finally given in Section 5.

2. AN ALTERNATIVE DEFINITION OF S_η

Let X be at least an ordinal variable, taking values $(b_1, b_2, \ldots, b_k)$. Let the couples

$$\{(b_j, n_j)\,; j = 1, 2, \ldots, k\} \tag{1}$$

denote the frequency distribution (*fd*) of X observed over a finite population of n units and consider

$$\begin{aligned} \mu &= \frac{1}{n}\sum_{j=1}^{k} R_j \\ &= \frac{1}{n}\sum_{j=1}^{k} j \cdot n_j \end{aligned} \tag{2}$$

where R_j, $j = 1, \ldots, k$, are retro-cumulative frequencies of the observed X and $n = \sum_j^k n_j$. Then, the Bipolar Mean η of X is defined as follows:

I) if μ is the integer s, $(s = 1, 2, \ldots, k)$, the bipolar mean η is defined as the *fd* that concentrates n on the value $s = \mu$, so we have $n_s = n$ and $n_i = 0\ \forall i = 1, \ldots, k$, $i \neq s$.

II) if $s < \mu < s + 1$, $(s = 1, 2, \ldots, k - 1)$, the bipolar mean η is defined as the *fd* that concentrates n on the values s and $s + 1$, respectively with frequencies $n_s = n(s + 1 - \mu)$ and $n_{s+1} = n(\mu - s)$.

The bipolar mean, which can be seen as a particular frequencies distribution, is particularly useful since it can be applied, as well as quantitative variables, to qualitative ordinal variables.

In order to measure the variability of the bipolar mean, Maffenini and Zenga (2006) have introduced the mean deviation about the bipolar mean:

$$S_\eta = \frac{1}{n}\sum_{j=1}^{k} \left| N_j - \tilde{N}_j \right| , \tag{3}$$

where N_j and $\tilde{N}_j$, $j = 1, 2, \ldots, k$, are the cumulative frequencies respectively of the observed X and of the bipolar mean η. An important feature of S_η is that it can be interpreted as the total number of "unitary steps" that serve to transform the distribution of X into that of the bipolar mean. For a review regarding S_η and the maximum value which it can assume see Brentari et al. (2009). In the following theorem we will provide an alternative formula for calculate (3).

THEOREM 1 *Let X be at least an ordinal variable and let the couples*

$$\{(b_j, n_j) : j = 1, 2, \ldots, k\} \ ,$$

denote the fd of X observed over a finite population of n units. The mean deviation (3) of X about its bipolar mean η can be calculated as follow

$$S_\eta = \frac{2}{n} \max \left(\sum_{j=1}^{s-1} N_j ; \sum_{s+2}^{k} R_j \right) \ ,$$

where N_j and R_j, $j = 1, \ldots, k$, are respectively the cumulative and the retro-cumulative frequencies of the observed X, and s is the integer part of μ defined in (2).

PROOF
Let $D_j = \left| N_j - \tilde{N}_j \right|$ denote the absolute difference between N_j and the cumulative frequency of the bipolar mean $\tilde{N}_j$, $j = 1, \ldots, k$. Since $D_k = n - n = 0$, we can omit the k-th value in the sum in (3). Then, for the calculation of S_η, we can consider only the elements $D_1, \ldots, D_{k-1}$ whose can also be expressed as follows

$$\begin{aligned} D_j &= \left| N_j - \tilde{N}_j + n - n \right| \\ &= \left| R_{j+1} - \tilde{R}_{j+1} \right| \qquad j = 1, \ldots, k-1 \ , \end{aligned} \tag{4}$$

where $\tilde{R}_2, \ldots, \tilde{R}_k$ denote the retro-cumulative frequencies of the bipolar mean η.

Using relation (4) in (3) and remembering that $R_1 - \tilde{R}_1 = n - n = 0$, it results that

$$\begin{aligned} nS_\eta &= \sum_{j=1}^{k} \left| R_j - \tilde{R}_j \right| \\ &= \sum_{j=1}^{s} \left| R_j - \tilde{R}_j \right| + \left| R_{s+1} - \tilde{R}_{s+1} \right| + \sum_{j=s+2}^{k} \left| R_j - \tilde{R}_j \right| \ , \end{aligned} \tag{5}$$

where s, as previously underlined, is the integer part of μ. By definition, we have that

$$\begin{aligned} \tilde{R}_j &= n \qquad && \text{for} \quad j < s+1 \\ \tilde{R}_j &= n(\mu - s) \qquad && \text{for} \quad j = s+1 \\ \tilde{R}_j &= 0 \qquad && \text{for} \quad j > s+1 \ . \end{aligned}$$

Then, the relation (5) becomes

$$\begin{aligned} nS_\eta &= \sum_{j=1}^{s} |R_j - n| + |R_{s+1} - n(\mu - s)| + \sum_{j=s+2}^{k} R_j \\ &= s \cdot n - \sum_{j=1}^{s} R_j + |R_{s+1} - n(\mu - s)| + \sum_{j=s+2}^{k} R_j \ . \end{aligned} \tag{6}$$

By adding and subtracting the quantity $\sum_{j=s+1}^{k} R_j$ on the right-hand side of equation (6) we obtain:

$$\begin{aligned} nS_\eta &= s \cdot n - \sum_{j=1}^{s} R_j + |R_{s+1} - n(\mu - s)| + \sum_{j=s+2}^{k} R_j + \sum_{j=s+1}^{k} R_j - \sum_{j=s+1}^{k} R_j \\ &= s \cdot n - n \cdot \mu + R_{s+1} + 2 \sum_{j=s+2}^{k} R_j + |R_{s+1} - n(\mu - s)| \\ &= 2 \sum_{j=s+2}^{k} R_j + R_{s+1} - n(\mu - s) + |R_{s+1} - n(\mu - s)| \ . \end{aligned} \tag{7}$$

Now let $A = R_{s+1} - n(\mu - s)$.

- If $A \leq 0$ we have $A + |A| = 0$, then

$$nS_\eta = 2 \sum_{j=s+2}^{k} R_j \ . \tag{8}$$

- If $A > 0$ we have $A + |A| = 2A$, then

$$\begin{aligned} nS_\eta &= 2 \sum_{j=s+2}^{k} R_j + 2R_{s+1} - 2n(\mu - s) \\ &= -2 \sum_{j=1}^{s} R_j + 2ns \qquad (9) \\ &= 2 \sum_{j=1}^{s-1} N_j \qquad (10) \end{aligned}$$

Moreover, condition $A \leq 0$ can also be expressed as

$$N_s \geq \tilde{N}_s \ , \tag{11}$$

or alternatively as

$$\sum_{s+2}^{k} R_j \geq n \cdot s - \sum_{j=1}^{s} R_j \ . \tag{12}$$

Using the results (8), (10) and (11), S_η can be defined in the following way

$$S_\eta = \begin{cases} \frac{2}{n} \sum_{j=s+2}^{k} R_j & \text{if } N_s \geq \tilde{N}_s \\ \frac{2}{n} \sum_{j=1}^{s-1} N_j & \text{otherwise .} \end{cases} \tag{13}$$

So S_η can be obtained using the sum of the first $s-1$ cumulative frequencies $N_1, \ldots, N_{s-1}$ (if $N_s < \tilde{N}_s$) or using the sum of the last $k-s-1$ retro-cumulative frequencies $R_{s+2}, \ldots, R_k$ (if $N_s \geq \tilde{N}_s$). In order to obtain S_η we can use, instead of condition (11), condition (12) where frequencies related to the bipolar mean are not directly needed. In particular, by comparing (9) and (10), it is easy to note that

$$\sum_{j=1}^{s-1} N_j = n \cdot s - \sum_{j=1}^{s} R_j \ ,$$

then, using condition (12) instead of (11) in formula (13) we obtain

$$S_\eta = \begin{cases} \frac{2}{n} \sum_{j=s+2}^{k} R_j & \text{if } \sum_{s+2}^{k} R_j \geq n \cdot s - \sum_{j=1}^{s} R_j \\ \frac{2}{n} \left(n \cdot s - \sum_{j=1}^{s} R_j \right) & \text{otherwise ,} \end{cases} \tag{14}$$

which is equal to

$$S_\eta = \frac{2}{n} \max \left(\sum_{s+2}^{k} R_j ; n \cdot s - \sum_{j=1}^{s} R_j \right) \ , \tag{15}$$

and equivalently to

$$S_\eta = \frac{2}{n} \max \left(\sum_{j=1}^{s-1} N_j ; \sum_{s+2}^{k} R_j \right) \ . \tag{16}$$

◇

The main feature of this formula, together to its simplicity, is that no frequency distribution associated to the bipolar mean is needed.

3. THE BIPOLAR MEAN-INDEPENDENCE

In this chapter, the concept of the bipolar mean-independence is introduced. Let X at least an ordinal variable observed in correspondence of g different categories of the variable Y. Suppose we are interested in studying the influence of Y on the behavior of X. If X was quantitative there would be many technique

anyone can use. For example, the well known correlation ratio of Pearson can be adopted in order to evaluate how the arithmetic mean of X varies conditionally to Y. This coefficient is usually used for judging non-linear relationships between the two variables. Naturally, if X is a qualitative ordinal variable, this road can not be traveled since any conditional arithmetic mean can be calculated. The idea is to give a new definition of the "mean dependence" so that it can be used also in the case of qualitative ordinal variables. The following subsection introduces a new concept of dependence which is similar to that measured by the correlation ratio of Pearson but it exploits the conditional bipolar means, instead of the simple conditional means. After this definition, a normalized measure for evaluate the strength of the bipolar mean-dependence is provided.

3.1 DEFINITION OF BIPOLAR MEAN-INDEPENDENCE

Let the n statistical units of a given population be classified according to the variables X and Y which have finite number of categories (or values) denoted respectively by $(b_1, b_2, \ldots, b_k)$ and $(a_1, a_2, \ldots, a_g)$ and where X is at least ordinal (see Table 1). Then, for the simplicity sake, n_{ij} is the number of times that (a_i, b_j) is observed.

TABLE 1: Contingency table related of the n statistical units classified according with variables X and Y

Y \ X	b_1	$\ldots$	b_j	$\ldots$	b_k	
a_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1k}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$
a_i	n_{i1}		n_{ij}		n_{ik}	$n_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$	$\vdots$
a_g	n_{g1}	$\ldots$	n_{gj}	$\ldots$	n_{gk}	$n_{g\cdot}$
	$n_{\cdot 1}$	$\ldots$	$n_{\cdot j}$	$\ldots$	$n_{\cdot k}$	n

From Table 1 we can observe g conditional distributions of X related with the values $a_1, \ldots, a_g$ respectively. Let ${}_iR_1, \ldots, {}_iR_k$ and ${}_i\tilde{R}_1, \ldots, {}_i\tilde{R}_k$, $i = 1, \ldots, g$, the retro-cumulative frequencies respectively of the i-th conditional distribution of X and of the i-th conditional bipolar mean η_i. Similarly let ${}_iN_1, \ldots, {}_iN_k$ and ${}_i\tilde{N}_1, \ldots, {}_i\tilde{N}_k$, $i = 1, \ldots, g$, the cumulative frequencies respectively of the i-th conditional distribution of X and of the i-th conditional bipolar mean η_i. Suppose now to consider the distribution obtained through the sum of the g conditional bipolar means. In particular, if we denote with $\tilde{R}_1^*, \ldots, \tilde{R}_k^*$ the

retro-cumulative frequencies of such a distribution, we have

$$\begin{aligned}\sum_{j=1}^{k} \tilde{R}_j^* &= \sum_{j=1}^{k} {}_1\tilde{R}_j + \sum_{j=1}^{k} {}_2\tilde{R}_j + \cdots + \sum_{j=1}^{k} {}_g\tilde{R}_j \\ &= n_{1\cdot}\mu_1 + n_{2\cdot}\mu_2 + \cdots + n_{g\cdot}\mu_g \\ &= n\mu \ , \end{aligned} \tag{17}$$

where

$$\begin{aligned}\mu_i &= \frac{1}{n_{i\cdot}} \sum_{j=1}^{k} {}_i\tilde{R}_j \\ &= \frac{1}{n_{i\cdot}} \sum_{j=1}^{k} j \cdot n_{ij} \quad i = 1, \ldots, g \ , \end{aligned} \tag{18}$$

as in formula (2). The main result is that the bipolar mean associated to the distribution derived from the sum of the conditional bipolar means is equal to the bipolar mean of the marginal distribution of X. This results will be use in the next section to build a measure for evaluating the departure from the bipolar mean-independence defined in what follows.

DEFINITION (Bipolar mean-independence)
Let X at least an ordinal variable and Y of any type. We say that X is bipolar mean-independent of Y if the following result holds:

$$\mu_i = \mu \quad \text{for all } i = 1, \ldots, g \ ,$$

where μ_i, $i = 1, \ldots, g$, are defined as in (18) and μ is obtained from the marginal distribution of X as in (2).

Remember that the quantities μ_i, $i = 1, \ldots, g$, are not necessary conditional arithmetic means since X could be only ordinal. Nevertheless, as we will seen in the next subsection there is a clear parallelism between mean-dependence and bipolar mean-dependence.

3.1 MEASURE THE BIPOLAR MEAN-DEPENDENCE

Denote with $\tilde{n}_{s+1}$ the frequency in position $s+1$ of the (marginal) bipolar mean η of X. This quantity can be expressed in terms of the frequencies associated to the conditional bipolar means. In fact, using the result in (17), we

obtain

$$\begin{aligned}
\tilde{n}_{s+1} &= n(\mu - s) \\
&= n\mu - ns \\
&= \sum_{j=1}^{k} \tilde{R}_j^* - ns \\
&= \sum_{l=1}^{g}\sum_{j=1}^{k} {}_l\tilde{R}_j - \sum_{l=1}^{g} n_{l\cdot} \cdot s \\
&= \sum_{l=1}^{g}\sum_{j=1}^{k} \left({}_l\tilde{R}_j - n_{lj} \cdot s\right) \ . \qquad (19)
\end{aligned}$$

Let s_i the integer part of μ_i, $i = 1, \ldots, g$, and consider only the first term ($l = 1$) of the sum on the right-hand side of equation (19). We have that

$$\begin{aligned}
\sum_{j=1}^{k} \left({}_1\tilde{R}_j - n_{1j}s\right) &= \sum_{j=1}^{s_1} (n_{1\cdot} - n_{1j}s) + n_{1\cdot}(\mu_1 - s_1) - n_{1s_1+1} \cdot s + \sum_{j=s_1+2}^{k} n_{1j} \cdot s \\
&= -s \cdot {}_1N_{s_1} + n_{1\cdot}\mu_1 - n_{1s_1+1} \cdot s - s \cdot {}_1R_{s_1+2} \\
&= n_{1\cdot}\mu_1 - s\left({}_1N_{s_1} + {}_1R_{s_1+2}\right) - n_{1s_1+1} \cdot s \ . \qquad (20)
\end{aligned}$$

Since ${}_1N_{s_1} + {}_1R_{s_1+2} = n_{1\cdot} - n_{1s_1+1}$, the following results hold

$$\begin{aligned}
\sum_{j=1}^{k} \left({}_1\tilde{R}_j - n_{1j}s\right) &= n_{1\cdot}\mu_1 - n_{1\cdot} \cdot s \\
&= n_{1\cdot}(\mu_1 - s) \qquad (21) \\
&= n_{1\cdot}(\mu_1 - s) + n_{1\cdot} \cdot s_1 - n_{1\cdot} \cdot s_1 \\
&= n_{1\cdot}(\mu_1 - s_1) + n_{1\cdot}(s_1 - s) \\
&= \tilde{n}_{1s_1+1} + n_{1\cdot}(s_1 - s) \ . \qquad (22)
\end{aligned}$$

Adopting the same procedure to the other terms ($l = 2, \ldots, g$) in the sum on the right-hand side of (19), equation (19) becomes

$$\tilde{n}_s = \sum_{i=1}^{g} \left[\tilde{n}_{is_i} - n_{i\cdot}(s_i - s)\right] \qquad (23)$$

The same simple algebra can be used to define $\tilde{n}_{s+1}$ in terms of the frequencies associated to the conditional bipolar means:

$$\tilde{n}_{s+1} = \sum_{i=1}^{g} \left\{\tilde{n}_{is_i+1} + n_{i\cdot}(s_i - s)\right\} \qquad (24)$$

Formula (23) and (24) informs that the frequencies of the marginal bipolar mean can be obtained using frequencies associated to the conditional bipolar means and "correction factors" whose depend on the distances between the s_i, $i = 1, \ldots, g$, and s.

Consider now Table 2(a) which reports two prospects whose elements consitute the basis for calculating (23) and (24). Let

$$n_i^- = [\tilde{n}_{is_i} - n_{i\cdot}(s_i - s)] \ , \quad n_i^+ = [\tilde{n}_{is_i+1} + n_{i\cdot}(s_i - s)] \ ,$$

and denote with $\overset{*}{n}_i^-$ and $\overset{*}{n}_i^+$ the frequencies associated to the i-th bipolar mean (respectively in position s_i and position $s_i + 1$) in the case of bipolar mean-independence.

TABLE 2: *Conditional and marginal bipolar means and correction factors*

(a) Bipolar means

$\tilde{n}_{1s_1}$	$\tilde{n}_{1s_1+1}$	n_1
$\vdots$	$\vdots$	$\vdots$
$\tilde{n}_{is_i}$	$\tilde{n}_{is_i+1}$	n_i
$\vdots$	$\vdots$	$\vdots$
$\tilde{n}_{gs_1}$	$\tilde{n}_{gs_g+1}$	n_g
$\tilde{n}_s$	$\tilde{n}_{s+1}$	n

(b) Correction factors

$n_{1\cdot}(s_1 - s)$
$\vdots$
$n_{i\cdot}(s_i - s)$
$\vdots$
$n_{g\cdot}(s_g - s)$

By definition, in this case we would have $\mu_1 = \ldots = \mu_g = \mu$ and

$$\overset{*}{n}_i^- = n_{i\cdot}(s + 1 - \mu) \quad \text{and} \quad \overset{*}{n}_i^+ = n_{i\cdot}(\mu - s) \ , \quad i = 1, \ldots g \ .$$

A natural measure to evaluate the departure from the case of independence can be constructed starting from the following summation:

$$\sum_{i=1}^{g} \left| n_i^- - \overset{*}{n}_i^- \right| + \sum_{i=1}^{g} \left| n_i^+ - \overset{*}{n}_i^+ \right| \ .$$

After some simple algebra we observe that

$$\left| n_i^- - \overset{*}{n}_i^- \right| = \left| n_i^- - \overset{*}{n}_i^+ \right| ,$$

for all $i = 1, \ldots, g$, then we can consider, for simplicity, only the quantity

$$\delta_\eta = \sum_{i=1}^{g} \left| n_i^+ - \overset{*}{n}_i^+ \right| \ .$$

in our investigation.

Naturally, it results that $\delta_\eta = 0$ only in the case of bipolar mean-independence. Suppose now that holds $s_1 = \ldots = s_g = s$ but not $\mu_1 = \ldots = \mu_g = \mu$. In this case, as in the case of independence, all the values in Table 2(b) are equal to zero but here one (or more) of the n_i^+ is different from $\overset{*}{n}_i^+$. The quantity

$\left|n_i^+ - \overset{*}{n}_i^+\right|$ measures the distance, in terms of frequency difference, between η_i and η. This is easy to understand since $\eta_1, \ldots, \eta_g$, and η have the same positions s and $s+1$. Otherwise, if the bipolar means are very different among them, the generic element $n_{i\cdot}(s_i - s)$ in Table 2(b) could be significantly different from zero by influencing the quantity $\left|n_i^+ - \overset{*}{n}_i^+\right|$. The higher the difference $(s_i - s)$, the more n_i^+ will be far from $\overset{*}{n}_i^+$. Note that the quantity $\bar{s}_i = (s_i - s)$ is an integer. Then, for a unit increase of $|\bar{s}_i|$, the quantity n_i^+ moves away from $\overset{*}{n}_i^+$ for a quantity equal to $n_{i\cdot}$.

Summarizing, the differences $\left|n_i^+ - \overset{*}{n}_i^+\right|$, $i = 1, \ldots, g$, which compose δ_η give information about the departure from bipolar mean-independence both in terms of frequencies of the bipolar means (elements in Table 2(a)) and in terms of distance of the bipolar means (elements in Table 2(b)). It is interesting to note that δ_η can be expressed also in the following way

$$\delta_\eta = \sum_{i=1}^{g} |\mu_i - \mu| \, n_{i\cdot} \, . \tag{25}$$

Formula (25) highlights, as indeed was to be expected, that the distance from the case of bipolar mean-independence evaluated using elements (23), can be also obtained exploiting the quantities μ_i, $i = 1, \ldots, g$, which remember that are not necessary means.

Since by definition $\delta_\eta \geq 0$, by maximizing δ_η we can obtain a normalized index included into the interval $[0, 1]$ which assumes 1 in the case of maximum bipolar mean dependence and zero in the case of bipolar mean-independence. The goal is then find the frequency configuration of Table 1 (fixing the marginal distributions) which allows to maximize quantity (25). A program which provide all the possible configuration such that proposed by Greselin (2003) could be used. By calculating the indexes for each configuration the maximum value of (25) could be identified. In what follows we propose a very simple procedure that provides the interested configuration in the most of the case. However, also when the procedure fails, it provides configurations with indexes very near to the maximum one. The following steps compose the procedure:

1. choose the most frequent value a_i and allocate part (or all) of the frequencies $n_{i\cdot}$ (according to the marginal distribution of X) in the i-th row in correspondence on the more distance value (of X) from the value μ. Allocate part (or all according to the marginal distribution of X) of the remaining frequencies in correspondence of the nearest value of X previously found. Repeat this last step until the frequencies $n_{i\cdot}$ are exhausted.

2. choose the second most frequent value of Y and follow the same procedure provided in the previous step considering that the frequencies of the second most frequent value must be allocate coherently with the frequencies allocated in step 1.

3. repeating step 2. for the third most frequent value of Y (allocating the frequencies coherently with the previous steps), for the fourth most frequent value of Y and so on until the less frequent value of Y is considered (and hence until the frequencies of X have been exhausted).

Using the table so obtained, the following quantity can be calculated:

$$\delta_\eta^M = \sum_{i=1}^{g} |\mu_i^M - \mu| \, n_{i\cdot} \, . \tag{26}$$

where μ_i^M denotes the quantities calculated as in (18). Then, the normalized index results

$$\Delta_\eta = \frac{\delta_\eta}{\delta_\eta^M} \tag{27}$$

which theoretically assumes 0 and 1 in presence of minumum and maximum bipolar mean dependence respectively.

Some considerations has to be made for the minimum value which can assume index (27). In particular we have that the smaller value of (27) could be higher then zero. In other words in many case we can have distribution which (fixing the marginal) can not allow to identify configuration such that $\Delta_\eta = 0$. However, this problem vanish when frequencies increasing and then we can neglect this issue whenever treating application with high data. Moreover, as for example the Cramer's one which measures the departure from independence in distribution, this index could attain values near its maximum very sporadically. The study of the behavior and of the properties of the proposed index remains an open issue which we shall pursue elsewhere.

4. Application

In this section we analyze the situation of the level of school attendance in the five italian macro-areas*. The *level of school attendance* is an ordinal variable with $k = 7$ categories. Table 1 reports its frequencies in the five macro-areas distributions. A more detailed analysis of the level of school attendance can be read in (Zenga, 2007), even if in that article the asymmetry situation of the distribution was studied.

In Table 4, the bipolar mean, and the deviation from the whole bipolar mean for the 5 macro-area are reported.

Looking at the Table 4, it is possible to assert that it exists a dependence in bipolar mean among the several macro-areas distributions, in fact the bipolar mean values are different each other and from the Italy distribution. Moreover, the bipolar means distributions it is splitted on the two categories Middle School and Secondary School, except for the Italian Islands: its consecutive categories are Primary Schhol and Middle School. In fact the Italian Islands is the

*Source: Italian Census of 2001

School attendance	N-W Italy	N-E Italy	Central Italy	South Italy	Italian Islands	Italy
Illiterate	80149	54551	90493	388711	160621	774525
Literate without Title	380182	402450	511650	897017	459784	2651083
Primary School	3418015	2474190	2248651	2577491	1258740	11977087
Middle School	4462158	3030269	2868596	3802267	1905328	16068618
Secondary School	3756195	2672599	2940365	3148284	1405923	13923366
University Diploma	160096	111076	126958	112362	51232	561724
University Degree	920910	613815	805853	780081	359876	3480535
Total Frequency	13177705	9358950	9592566	11706213	5601504	49436938

TABLE 3: Frequencies of the level of school attendance in the italian regional macro-area: inhabitants with at least 14 years old. (Source: 2001, Italian Census)

Italian Macro-area	n_j	μ_j	$\mu_j - \mu$	$\lvert\mu_j - \mu\rvert$	$\lvert\mu_j - \mu\rvert n_j$
N-W Italy	13177705	4.1837	0.0643	0.0643	847326.4315
N-E Italy	9358950	4.1382	0.0195	0.0195	182499.525
Center Italy	9592566	4.2156	0.0969	0.0969	929519.6454
South Italy	11706213	4.0150	-0.104	0.104	1217446.152
Italian Islands	5601504	3.9871	-0.132	0.132	739398.528
Italy	49436938	4.1191	-	-	-

TABLE 4: The bipolar means and deviations of the partial bipolar means

macroarea with the lowest level of attendance. As pointed previously, only two macro-areas show negative values for the deviation from total bipolar mean. In particular, the Italian Islands bipolar mean reports a lower value than the Italy bipolar mean, but it takes up different position from the Italy bipolar mean. The South Italy bipolar mean has a lower value than the Italy bipolar mean, but has the same position. The other macro-ares show positive deviation and the same position of the Italy bipolar mean. The index 25 is equal to $\delta_\eta = 3916190.282$, that is $\frac{\delta_\eta}{n} = 0.0792$, in average the deviation of the partial bipolar means from the whole bipolar mean is 0.0792, that is a very low value. Now it is possible to evaluate the value of 25 in case of the maximum dependence.

The table of the maximum dependence is reported in Table 5 The value of the index in case of maximum dependence is equal to $\delta_\eta^M = 45961464.49$ and the normalized index is given by

$$\Delta_\eta = \frac{\delta_\eta}{\delta_\eta^M} = \frac{3916190.282}{45961464.49} = 0.0852.$$

According to the meaning of the 27, it seems to be reasonable to state that the education level is lightly affected by the macro-areas. In fact, the dependence level is equal to the 8.52% of the possible maximum level, that is a low level of dependence.

School attendance	N-W Italy	N-E Italy	Central Italy	South Italy	Italian Islands	Italy
Illiterate	774525	0	0	0	0	774525
Literate without Title	2651083	0	0	0	0	2651083
Primary School	9752097	0	2224990	0	0	11977087
Middle School	0	3099538	7367576	0	5601504	16068618
Secondary School	0	6259412	0	7663954	0	13923366
University Diploma	0	0	0	561724	0	561724
University Degree	0	0	0	3480535	0	3480535
Total Frequency	13177705	9358950	9592566	11706213	5601504	49436938
Bipolar Mean	2.6813	4.6688	3.7681	5.6426	4.0000	4.1191

TABLE 5: Frequencies in case of maximum dependece in bipolar mean of the level of school attendance in the italian regional macro-area

4. Conclusions

The analysis of the distributions with ordinal categories is typical of the social sciences. The bipolar mean is a tool to synthesize these distributions. In these work we defined a new formula for the bipolar mean. When more than a group is present for the same ordinal distribution, we introduced the theoretical concept of bipolar mean-independence and a measure of the bipolar mean-dependence. This measure assumes value equal to zero in case of every distribution has the same bipolar mean and maximum value when the distributions are very different each other and from the bipolar mean distribution. We provided, also, an empirical method to find that configuration that assumes the maximum value of the bipolar mean-dependence index. In this way we defined a normalized index of bipolar mean-independence. At the end, an example based on the real data about the level of school attendance in Italy is shown. The future works should be regard on the sample distribution of bipolar mean and consequently the sample distribution of the measure for the bipolar mean dependence.

RIASSUNTO

In questo lavoro viene presentato una ulteriore formula per determinare la media bipolare per variabili almeno ordinali. Viene inoltre introdotto il concetto di indipendenza in media bipolare. Tale concetto trae spunto dal concetto generale di indipendenza in media ma può essere applicato per variabili di tipo qualitativo su scala ordinale. Similarmente a quanto accade per i caratteri quantitativi, viene qui descritto anche l'indice normalizzato per la dipendenza in media bipolare. Infine il metodo viene applicato a dati reali per sottolineare la semplicità e l'utilità della procedura proposta.

REFERENCES

Brentari, E. Dancelli, L. Maffenini, W. (2009). Rapporto di ricerca del Dipartimento Metodi Quantitativi. Quaderno n. 338, Università degli Studi di Brescia.

Coombs, C.H. (1953). Theory and Methods of Social Measurement, pp 471-535 in L. Festinger and D. Katz, Research Methods in the Behavioral Sciences, New York, Dryden Press.

Tufte, E.R. (1970). Improving Data analysis in Political Science, in E.R. Tufte. The Quantitative Analysis of Social Problems, Reading, Addison-Wesley, 437-449.

Maffenini, W. and Zenga, M. (2005). Bipolar mean for ordinal variables. *Statistica & Applicazioni*, **3**(1), 3–18.

Maffenini, W. and Zenga, M. (2006). Bipolar mean and mean deviation about the bipolar mean for discrete quantitative variables. *Statistica & Applicazioni*, **4**(1), 35–53.

Greselin, F. (2003). Counting and enumerating frequency tables with given margins. *Statistica & Applicazioni*, **1**(2), 87–104.

Zenga, M. (2007). Asymmetry for ordinal variables. *Statistica & Applicazioni*, **5**(2), 205–221.