The decomposition by subgroups of the inequality curve $Z(p)$ and the inequality index ξ

La scomposizione in gruppi della curva di diseguaglianza $Z(p)$ e dell'indice ξ

Francesco Porro and Michele Mario Zenga

Abstract This paper provides a procedure to decompose by subgroups the inequality index ξ, proposed by Zenga in [3]. Through this decomposition, such index can be seen as the weighted average of two terms, which represent the "within" and the "between" components. This decomposition arises from the decomposition by subgroups of the inequality $Z(p)$ curve, that originates the index itself.
Abstract In questo lavoro viene presentata una procedura per la scomposizione dell'indice di diseguaglianza ξ proposto da Zenga in [3]. In questa scomposizione, tale indice risulta essere uguale alla media aritmetica ponderata di due termini che rappresentano la componente "within" e la componente "between". La presente scomposizione viene effettuata a partire dalla scomposizione tra gruppi della curva di diseguaglianza $Z(p)$, dalla quale trae origine l'indice stesso.

Key words: inequality, subgroups decomposition, inequality curve $Z(p)$, inequality index ξ, cograduation table

1 Introduction

A very important characteristic of an inequality index is its suitability to be decomposed by subgroups and by sources. This paper proposes a decomposition by subgroups of the inequality index ξ introduced by Zenga in 1984 [3]. The approach is similar to the one used by Radaelli in [1] and it is based on the decomposition of the pointwise inequality index $Z(p)$. The paper is organized as follows: in the next section the definition of the $Z(p)$ curve and of the inequality index ξ are provided.

Francesco Porro
Universitá degli Studi di Milano-Bicocca, Milano e-mail: francesco.porro1@unimib.it

Michele Mario Zenga
Universitá degli Studi di Milano-Bicocca, Milano e-mail: michele.zenga@unimib.it

The following subsection describes the procedure to fill the cograduation table. In section 3, the proposed decomposition by subgroups is illustrated, using an example.

2 Notation and definitions

First of all, it is useful to recall the following definition (see [3]).

Definition 1. Let Y be a non-negative continuous random variable with probability density function f, strictly increasing distribution function F, and positive finite expectation μ. Let $Q(y)$ be the first incomplete moment of Y: $Q(y) = \frac{1}{\mu}\int_0^y t f(t)dt$. Then the inequality curve $Z(p)$ of Y is defined as:

$$Z(p) = \frac{y^*_{(p)} - y_{(p)}}{y^*_{(p)}} = 1 - \frac{y_{(p)}}{y^*_{(p)}}, \qquad p \in [0,1], \tag{1}$$

where $y_{(p)}$ and $y^*_{(p)}$ denote the inverse functions of F and Q. The inequality index ξ of the random variable Y is defined as

$$\xi = \int_0^1 Z(p)dp = 1 - \int_0^1 \frac{y_{(p)}}{y^*_{(p)}} dp.$$

Note that the $Z(p)$ curve assumes values in the interval $[0,1]$, therefore the index ξ is a normalized inequality index.

The $Z(p)$ curve and the index ξ can also be defined for discrete variates. Let $\{(y_j, n_j); j = 1,2,...k\}$ be the non-negative ordered values of the variate Y and their frequencies, observed on $n = \sum_{j=1}^k n_j$ units of a finite population. Let: $f_r(y_j) = n_j/n$; $T = \sum_{j=1}^k y_j n_j$; $M = T/n$ be the positive mean of Y; $s_j = s(y_j) = y_j n_j/Mn$ be the share of y_j; $F(y) = \sum_{y_j \le y} n_j/n$, and $Q(y) = \sum_{y_j \le y} s(y_j)$. In the case of discrete variates, the inverse function of F and Q are given by

$$y_{(p)} = F^{-1}(p) = \begin{cases} \inf\{y : F(y) \ge p\} \text{ if } p \in]0,1] \\ \inf\{y : F(y) > 0\} \text{ if } p = 0 \end{cases}$$

$$y^*_{(p)} = Q^{-1}(p) = \begin{cases} \inf\{y : Q(y) \ge p\} \text{ if } p \in]0,1] \\ \inf\{y : Q(y) > 0\} \text{ if } p = 0, \end{cases}$$

consequently the definition of $Z(p)$ is the same one provided in (1) for the continuous case. The index ξ is then defined as the area under the $Z(p)$ curve in the interval $[0,1]$. In the discrete case, the function $Z(p)$ is a step function: the steps can be at the points where $y_{(p)}$ or $y^*_{(p)}$ (or both) change value. Such points have to be found in the following set:

$$J = \left\{ F(y_j) = \sum_{i=1}^{j} f_r(y_i); j = 1,...,k \right\} \cup \left\{ Q(y_j) = \sum_{i=1}^{j} s(y_i); j = 1,...,k \right\}. \tag{2}$$

Sometimes the direct computation of $Z(p)$ as a function of $y_{(p)}$ and $y^*_{(p)}$ can be quite complicated. Using the so-called cograduation table as described in [4], such computation becomes easier.

2.1 The cograduation table

The cograduation table has been introduced by Salvemini in [2]. This table allows to make explicit an association rule between two discrete variates (X,W). Here it is used for the variates Y and Y^*. To illustrate how the cograduation table between Y and Y^* has to be completed, the following example from [4] is considered. Let $y_1 = 5$, $y_2 = 8$, $y_3 = 17$, $y_4 = 30$ and $y_5 = 40$ be the values observed of the variate Y. The starting point is a table with in the first column the different values y^*_i and in the last column the corresponding shares s_i. The bottom row is filled with the values y_j, and the first row with the corresponding $f_r(y_j)$. The inner cells are filled using the following procedure. First of all, the cell in the lower left-hand corner of the table, corresponding to the couple $(y_1 = 5, y^*_1 = 5)$ is filled with s_1. After that, the second cell to be completed is the one corresponding to the couple $(y_1 = 5, y^*_2 = 8)$. There, it is allocated 0.08, which is the minimun between $s_2 = 0.08$ and the "remaining frequency" of y_1 $(0.20 - 0.05 = 0.15)$. Then, the above cell is filled with 0.07, since it is the minimun between $s_3 = 0.17$ and $0.20 - 0.08 - 0.05 = 0.07$. At this point the first column is saturated, since all the first relative frequency $f_r(y_1) = 0.20$ has been used. Then the procedure moves to the adjacent cell, corresponding to the couple $(y_2 = 8, y^*_3 = 17)$. Using the same rule as before, the cell is filled with 0.10 then the procedure moves to the above cell $(y_2 = 8, y^*_4 = 30)$, and so on. The procedure stops when the upper right-hand cell is filled with the last frequency $f_r(y_5) = 0.20$. The result is the following table in which a weight is associated to each couple (y_j, y^*_i).

		$f_r(y_j)$					
		0.20	0.20	0.20	0.20	0.20	1
	40	-	-	-	0.20	0.20	0.40
	30	-	0.10	0.20	-	-	0.30
y^*_i	17	0.07	0.10	-	-	-	0.17 s_i
	8	0.08	-	-	-	-	0.08
	5	0.05	-	-	-	-	0.05
		5	8	17	30	40	
				y_j			

There are 8 cells with non-null weight. The inequality index ξ can be calculated, using the Table 1, where:

- r counts the cells with non-zero weight;
- v_r is the weight associated to the r-th cell;
- $p_r = \sum_{m=1}^{r} v_m$ (these values coincide with the elements of set J defined in (2));
- $y_{(p_r)}$ and $y^*_{(p_r)}$ are the values of Y and Y^* associated to the r-th cell;
- $Z(p_r)$ is calculated through formula (1).

The sum of the values in the last column gives the index $\xi = \sum_{r=1}^{8} Z(p_r)v_r = 0.342$.

Table 1 Calculus of the values $Z(p)$ and ξ.

r	v_r	p_r	$y_{(p_r)}$	$y^*_{(p_r)}$	$Z(p_r)$	$Z(p_r)v_r$
1	0.05	0.05	5	5	0	0
2	0.08	0.13	5	8	0.375	0.030
3	0.07	0.20	5	17	0.706	0.049
4	0.10	0.30	8	17	0.529	0.053
5	0.10	0.40	8	30	0.733	0.073
6	0.20	0.60	17	30	0.433	0.087
7	0.20	0.80	30	40	0.250	0.050
8	0.20	1	40	40	0	0
	1					0.342

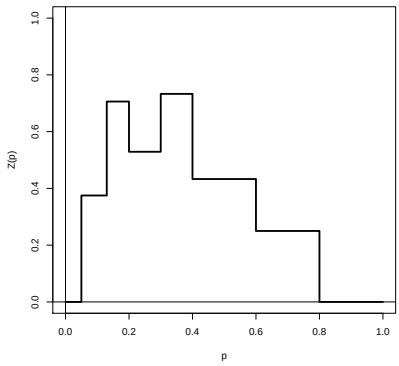

Fig. 1 Graph of the $Z(p)$ function. The area below it can be decomposed in eight rectangles: each of them has base given by v_r, height equal to $Z(p_r)$ and area equal to $Z(p_r)v_r$. Consequently the area under $Z(p)$, gives ξ.

3 The decomposition by subgroups

In order to semplify the explanation, a set of $n = 10$ statistical units is now considered. The units are divided into two disjoint groups, say G_1 and G_2, with cardinality $n_1 = 4$ and $n_2 = 6$. The values that Y assumes for the group G_1 are $\{5, 11, 14, 20\}$, and for the group G_2 are $\{2, 11, 23, 24, 42, 48\}$. First of all, the cograduation table has to be filled, by using all the nine different values of Y. It is worth remarking that the values of Y are nine, since the value 11 is repeated in the both groups. Following the procedure of subsection 2.1, the following cograduation table can be obtained.

	$f_r(y_j)$										
	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	1	
48	-	-	-	-	-	-	0.04	0.1	0.1	0.24	
42	-	-	-	-	0.05	0.1	0.06	-	-	0.21	
24	-	-	-	0.07	0.05	-	-	-	-	0.12	
23	-	-	0.085	0.03	-	-	-	-	-	0.115	
y_i^* 20	-	-	0.1	-	-	-	-	-	-	0.1	s_i
14	-	0.055	0.015	-	-	-	-	-	-	0.07	
11	0.065	0.045	-	-	-	-	-	-	-	0.11	
5	0.025	-	-	-	-	-	-	-	-	0.025	
2	0.01	-	-	-	-	-	-	-	-	0.01	
	2	5	11	14	20	23	24	42	48		
	y_j										

Now, as described in the previous section, the values of the couples (y_j, y_i^*) with non-zero weight has to be reported in a table. There are 17 cells with non-zero weight. Moreover here, it is needed to recognize whether the contribute to the inequality measure $Z(p)$ of each couple belongs to the within or to the between component. The idea is to assign the weight v_r to the within component if both the values of the related couple belong to the same group, otherwise it is assigned to the between component. For example, for $r = 1$, since the weight $v_1 = 0.01$ is related to the couple $(y_1 = 2, y_1^* = 2)$, consisting of two values of the same group (G_2), it is charged to the within component (see Table 2). The second one, $v_2 = 0.025$ is charged to the between component, since the corresponding values $y_1 = 2$ and $y_2^* = 5$ lay in two different groups. The assignment of the third weight needs further details. Such value is related to a couple containing the value 11, which belongs to both the groups. For this reason, the weight $v_3 = 0.065$ is divided into two equal parts: one is attributed to the within component, the other one to the between component. For applying this splitting, some columns and some rows have been added in Table 2. There, t counts the cells with non-zero weight (included the splitted ones), $\tilde{v}_t$ denotes the corresponding weight, and $\tilde{p}_t = \sum_{m=1}^{t} \tilde{v}_m$. In other words, v_3 is divided in two equal weights $\tilde{v}_3$ and $\tilde{v}_4$, both equal to $0.065/2 = 0.0325$. In such way the inequality related to the couple $y_1 = 2$ and $y_3^* = 11$ is divided in equal parts between the two components. By repeating this procedure, all the weights can be assigned, as reported in Table 2. After that, the contributes to the inequality index ξ can be evaluated by the products $Z(\tilde{p}_t)\tilde{v}_t$ (see the last two columns in Table 2). It arises that the overall weight which the decomposition assigns to the within component is 0.62, and the weight to the between component is 0.38. Then the index ξ can be described as the weighted average of the two components with exactly such weights. In other words, it holds the decomposition:

$$\xi = \xi_B \cdot 0.38 + \xi_W \cdot 0.62 = 0.3996,$$

$$\text{where} \qquad \xi_B = \frac{0.1756}{0.38} = 0.4621 \qquad \text{and} \qquad \xi_W = \frac{0.2240}{0.62} = 0.3613$$

are the contributions to the inequality of the between and the within components.

Table 2 Table for computing the Between and Within components of the inequality index ξ.

r	v_r	p_r	$y_{(p_r)}$	$y^*_{(p_r)}$	t	$\tilde{v}_t$	$\tilde{p}_t$	$y_{(\tilde{p}_t)}$	$y^*_{(\tilde{p}_t)}$	$Z(\tilde{p}_t)$	$\tilde{v}_t$		$Z(\tilde{p}_t)\tilde{v}_t$	
											Between	*Within*	*Between*	*Within*
1	0.01	0.01	2	2	1	0.01	0.01	2	2	0	-	0.01	0	-
2	0.025	0.035	2	5	2	0.025	0.035	2	5	0.6	0.025	-	0.015	-
3	0.065	0.1	2	11	3	0.0325	0.0675	2	11	0.8182	-	0.0325	-	0.0266
					4	0.0325	0.1	2	11	0.8182	0.0325	-	0.0266	-
4	0.045	0.145	5	11	5	0.0225	0.1225	5	11	0.5455	-	0.0225	-	0.0123
					6	0.0225	0.145	5	11	0.5455	0.0225	-	0.0123	-
5	0.055	0.2	5	14	7	0.055	0.2	5	14	0.6429	-	0.055	-	0.0354
6	0.015	0.215	11	14	8	0.0075	0.2075	11	14	0.2143	-	0.0075	-	0.0016
					9	0.0075	0.215	11	14	0.2143	0.0075	-	0.0016	-
7	0.1	0.315	11	20	10	0.05	0.265	11	20	0.45	-	0.05	-	0.0225
					11	0.05	0.315	11	20	0.45	0.05	-	0.0255	-
8	0.085	0.4	11	23	12	0.0425	0.3525	11	23	0.5217	-	0.0425	-	0.0222
					13	0.0425	0.4	11	23	0.5217	0.0425	-	0.0222	-
9	0.03	0.43	14	23	14	0.03	0.43	14	23	0.3913	0.03	-	0.0117	-
10	0.07	0.5	14	24	15	0.07	0.5	14	24	0.4167	0.07	-	0.0292	-
11	0.05	0.55	20	24	16	0.05	0.55	20	24	0.1667	0.05	-	0.0083	-
12	0.05	0.6	20	42	17	0.05	0.6	20	42	0.5238	0.05	-	0.0262	-
13	0.1	0.7	23	42	18	0.1	0.7	23	42	0.4524	-	0.1	-	0.0452
14	0.06	0.76	24	42	19	0.06	0.76	24	42	0.4286	-	0.06	-	0.0257
15	0.04	0.8	24	48	20	0.04	0.8	24	48	0.5	-	0.04	-	0.02
16	0.1	0.9	42	48	21	0.1	0.9	42	48	0.125	-	0.1	-	0.0125
17	0.1	1	48	48	22	0.1	1	48	48	0	-	0.1	-	0
	1					1					**0.38**	**0.62**	**0.1756**	**0.2240**

4 Conclusions

The proposed decomposition can be applied in any analysis, where it is useful to assess the Between and the Within components of inequality. For example in income analysis, if the incomes are grouped by area (geographical or sectorial), such decomposition shows how much inequality is due to the inequality in the areas and how much is related to the comparison of incomes coming from different areas. Since it is based on the decomposition of pointwise index $Z(p)$, it allows also the evaluation of the two components for different values of p_r.

References

1. Radaelli, P.: On the decomposition by subgroups of the Gini index and Zenga's uniformity and inequality indexes. International Statistical Review, **78**, 81-101 (2010)
2. Salvemini, T.: Sugli indici di omofilia. Suppl. Statistico Nuovi Problemi, **5**, 105-115 (1939)
3. Zenga, M.: Proposta per un indice di concentrazione basato sui rapporti fra quantili di popolazione e quantili di reddito. Giornale degli Economisti e Annali di Econ., **5-6**, 301-326 (1984)
4. Zenga, M.: Impiego delle tabelle di cograduazione per la determinazione dell'indice puntuale di concentrazione $Z(p)$. Statistica Applicata, **3**, 283-291 (1991)