CALCULATION OF GINI-COEFFICIENT FOR A STATISTICAL POPULATION

Daniela DAMIAN*, Nicoleta DAVID*

Department of Mathematics, Informatics and Socio-Human Sciences
*"George Barițiu" University of Brasov, Romania

Abstract. *The tasks and problem definitions which are treated in descriptive statistics are various. An important area is the representation of statistical material, frequent of figures. We can speak about concentration, if a relatively small number/quota of "n" quantities has a high contribution to the total sum.*

Keywords: *concentration, relative frequency distribution, Lorenz curve, Gini-coefficient.*

1. INTRODUCTION

In this article we shall deal with the distribution of the sum of n results of a sample to be examined for the different characteristics. In case all values of the sample are ≥ 0, we can talk about a concentration if a small quota of the n values has a high quota in the total sum.

Definition 1

Let consider an ordered sample $x_i > 0$, $1 \le i \le n$.

For all x_i, $1 \le N_i \le n$, we define the following numbers:

$$f_i = \frac{N_i}{n} \quad \text{and} \quad p_i = \frac{\sum_{i=1}^{N_i} x_i}{\sum_{i=1}^{n} x_i} \tag{01}$$

Where: f_i describes the relative quota of the N_i indices related to all n indices, and f_i the relative quota of the first N_i sum value related to the total sum.

We can now connect the n + 1 points in a two-dimensional coordinate system

$(0,0),(f_1,p_1),\ldots,(f_{n-1},p_{n-1}),(1,1)$

The curve resulted is the Lorenz curve.

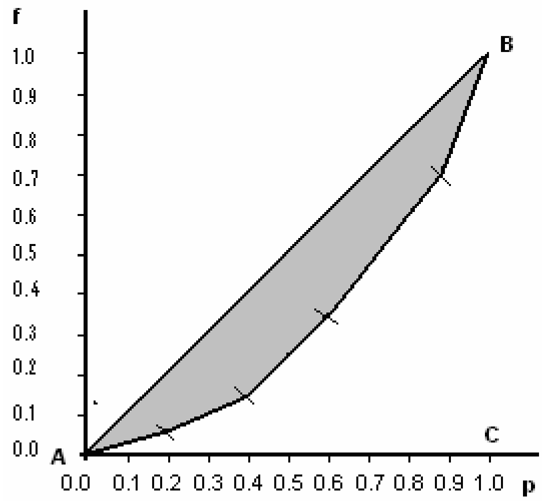

Fig. 1 Lorenz curve

The Lorenz curve remains the same even if a different coordinate system is used, namely the cumulative relative frequency and the cumulative grouping.

These are defined as follows:

Definition 2

Let consider an ordered sample with $x_i > 0$; $h(x_i)$ is the absolute frequency distribution that means $h(x_i)$ is the number of all equal x_i.

For all x_i, $1 \le N_i \le n$ we define the numbers u_i and v_i as follows:

$$u_i = \frac{1}{n} \cdot \sum_{i=1}^{N_i} h(x_i)$$

and

$$v_i = \frac{\sum_{i=1}^{N_i} h(x_i) \cdot x_i}{\sum_{i=1}^{n} h(x_i) \cdot x_i} \qquad (02)$$

Where u_i denote the cumulative relative frequency and v_i denote the cumulative grouping.

In essence, the Gini-Coefficient is the proportion between the surface marked off by the diagonal AB and the Lorenz curve, and the triangle surface ABC.

From this definition we conclude that the Gini-Coefficient takes the value 0 if all points of the Lorenz curve are located on the diagonal AB that means we speak about a uniform concentration [6,7].

Also, the Gini-Coefficient takes the value 1 if the surface between the diagonal AB and the Lorenz curve is identical with the triangle surface ABC. In this case we speak about a maximal concentration.

In statistics we have two types of populations.

The individual population and the population split in intervals. Depending on the type of used population we have two formulas to calculate the Gini-Coefficient.

2. INDIVIDUAL STATISTICAL POPULATION

In this case x_i are the values the population can take. N_i is the number of elements that compound the population.

The calculation formula is [2]:

$$I_g = \frac{\sum_{i=1}^{n} (x_i - x_{i-1}) \cdot N_i \cdot (n - N_i)}{n^2 \overline{x}} \qquad (03)$$

Remark:

To calculate the Gini-Coefficient with the help of working-tables it is necessary to order increasing the values of x_i.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (04)$$

Example 1

Table 1 presents the dates referring to the international trade of goods - monthly series. The source is the *Bulletin of International Trade Statistics 2008,* page 16 [3].

Table 1

Population	Jun. 2007	Jul. 2007	Aug. 2007
x_i	8114,0	8333,6	7157,5
Population	Sept. 2007	Oct. 2007	Nov. 2007
x_i	8273,8	9473,5	9518,4
Population	Dec. 2007	Ian. 2008	Feb. 2008
x_i	7992,5	9099,2	10415,7
Population	Mar. 2008	Apr. 2008	Mai 2008
x_i	10009,0	9990,0	10781,8
Population	Jun. 2008		
x_i	11076,0		

Next we will draw the working table [5]. The table will contain the values x_i in increasing order.

Table 2

x_i	N_i	$x_i - x_{i-1}$	$n - N_i$	$(x_i - x_{i-1}) \cdot N_i \cdot (n - N_i)$
7157,5	1	7157,5	12	85890
7992,5	2	835	11	18370
8114,0	3	121,5	10	3645
8273,8	4	159,8	9	5752,8
8333,6	5	59,8	8	2392
9099,2	6	765,6	7	32155,2
9473,5	7	374,3	6	20960,8
9518,4	8	44,5	5	1780
9990,0	9	471,6	4	16977,6
10009,0	10	19,0	3	570
10415,7	11	406,7	2	8947,4
10781,8	12	366,10	1	4393,2
11076,0	13	294,2	0	0
Σ=120235	Σ=91			Σ=201834

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{120235}{13} = 9248{,}85$$

Hence

$$I_g = \frac{201834}{13^2 \cdot 9248{,}85} = 0{,}1291$$

3. POPULATION SPLIT IN INTERVALS

We consider the same population, presented in the application above. This time we will split the population in intervals of equal size. So the value x_i is not a precise value, it becomes an interval.

Because we cannot calculate the Gini-Coefficient using intervals we will consider for each interval the mean value, which will be the new value of x_i.

In this case the Gini-Coefficient is calculated with following formula [1]:

$$I_g = 2 \cdot \left(\frac{1}{2} - \sum_{i=1}^{n} \frac{1}{2} \cdot (u_i - u_{i-1})(v_i + v_{i-1}) \right) \quad (05)$$

Where u_i and v_i were presented in (2) and $h(x_i)$ is the number of appeared values in each interval.

Example 2

Hence the population became:

Table 3

x_i	$h(x_i)$	Mean Value of Interval $= x_i$
7001-7500	1	7250
7501-8000	1	7750
8001-8500	3	8250
8501-9000	0	8750
9001-9500	2	9250
9501-10000	2	9750
10001-10500	2	10250
10501-11000	1	10750
11001-11500	1	11250
	$\Sigma = 13$	

The working table is:

Table 4

$\sum_{i=1}^{N_i} h(x_i)$	u_i	$h(x_i) \cdot x_i$	$\sum_{i=1}^{N_i} h(x_i) \cdot x_i$
1	0,077	7250	7250
2	0,154	7750	15000
5	0,385	24750	39750
5	0,385	0	39750
7	0,538	18500	58250
9	0,692	19500	77750
11	0,846	20500	98250
12	0,923	10750	109000
13	1	11250	120250

v_i	$u_i - u_{i-1}$	$v_i + v_{i-1}$	$(u_i - u_{i-1}) \cdot (v_i + v_{i-1})$
0,0603	0,077	0,0603	0,0046
0,1247	0,077	0,185	0,0142
0,3306	0,231	0,4553	0,0601
0,3306	0	0,6612	0
0,4844	0,153	0,815	0,1247
0,6466	0,154	1,131	0,1742
0,817	0,154	1,04636	0,2254
0,9064	0,077	1,7234	0,1327
1	0,077	1,9064	0,1468
			$\Sigma = 0{,}8827$

Hence,

$$I_g = 2 \cdot \left(\frac{1}{2} - \frac{1}{2} \cdot 0{,}8827 \right) = 0{,}1172$$

Although in the two examples we have used the same population the final results are different. This difference comes from the different values of x_i.

The low value of the Gini-Coefficient comes from the relatively uniformly distributed values of x_i.

4. CONCLUSIONS

A distribution series representing the mean income per person shows us the existence of certain segments of population with different buying power [4]. The statistics offers us scientific methods for research complex processes and hence to prevent the lack of balance in the society.

The concentration is also used for example to describe the competition condition in a domain of activity. For the concentration two extremes are define. The absence of concentration leads us to perfect competition, the ideal situation which we can not find in a real market economy. On the other hand maximum of concentration leads us to the other extreme that means to monopole, an unwished situation in a democratic country and avoid in practice by the antimonopoly legislation. Hence, we can conclude, that the concentration processes (of individual income, finance power etc.) are specific for a market economy.

In practice, the most used statistic instrument for measuring the concentration is the Gini-Coefficient. Depending of the type of population (individual population or population split in intervals), we have two formulae to calculate the Gini-Coefficient. Because in the paper, we have considered for both cases the same dates, the two values for the Gini-Coefficient, we have obtain, are very close. The two values are not identical because for the individual population the Gini-Coefficient has a more precisely value as in the case of split population in intervals.

As we can see, the two values are very close to zero. That means that the values of the international trade of goods - monthly are very uniform distributed.

REFERENCES

1. Gert, H., *Grundlagen der Mathematik, der Statistik und des Operationsresearch fur Wirtschaftswissenschaftler*, Oldenbourg Verlag Munchen, Wien, 1994;
2. Bogdan, V., Harlab, A., *Statistica manageriala*, Editura Fundatiei AXIS, IASI, 2004;
3. * * * *Bulletin of International Trade Statistics 2008,* page 16 www.insse.ro/cms/rw/pages/index.ro.do;
4. Backhaus, K., Erichson, B., Plinke, W., Weiber, R., *Multivariate Analysemethoden*, Springer, New York, 2003, pp. 45-155;
5. Brosius, G., Brosius, F., *Excel 2000 professionell - Tabellenkalkulation mit Office 2000*, Addison-Wesley, 1999, pp. 217-223;
6. Cismaşiu, C., *Complemente de matematici pentru economşti*, Reprografia Universității Transilvania, Braşov, 1996;
7. Cocan, M., Orman, G.V., Cocan, A., *Curs de matematică pentru economişti*, Reprografia Universității Transilvania, Braşov, 1992.