Revista Electrónica de Comunicaciones y Trabajos de ASEPUMA. **Rect@**
Volumen 21(2020). Páginas 1 a 7
DOI: 10.24309/recta.2020.21.1.01

THE GINI INDEX: A PROPOSAL FOR REVISION

URSICINO CARRASCAL ARRANZ
Ursicino.Carrascal@uva.es
Universidad de Valladolid, Departamento de Economía Aplicada
Avenida Valle Esgueva 6 47011- Valladolid (España)

Recibido (08/11/2019)
Revisado (28/03/2020)
Aceptado (31/03/2020)

RESUMEN: El índice de Gini se usa ampliamente en estadística para el estudio de la equidad en la distribución de una variable. Sin embargo, su definición tiene varias formulaciones y su uso a veces tiene problemas ocultos que pueden llevar a conclusiones incorrectas. Esa es la razón por la cual es necesario hacer algunos comentarios al respecto e incluso formular algunas propuestas para aclarar errores en su definición y uso. Presentamos una alternativa para resolver este error en una de las fórmulas más populares del índice de Gini.

Palabras Clave: Índice de Gini, desigualdad, distribución, equidad, concentración.

ABSTRACT: The Gini index is widely used in statistics for the study of equity in the distribution of a variable. However, its definition has several formulations and its use sometimes has hidden problems that can lead to incorrect conclusions. That is the reason why it is necessary to make some remarks this regard and even formulate some proposals to clarify some errors on its definition and use. We present an alternative to solve this error in one of the most popular formulas of the Gini index.

Keywords: Gini index, inequality, distribution, equity, concentration.

1. Introduction

The Gini index is widely used in the economic literature for the analysis of inequality of a distribution, being the most typical case the study of equity in the distribution of income in a country or a region[1]. Already, Gini himself proposes several alternative ways of measuring it, and this formulation has been expanded since then ([1] summarizes this situation in the title of his article: “More than a dozen ways of spelling Gini”). Some authors have even proposed the use of mechanical methods, using a planimeter to calculate the area between the equity line and the Lorenz curve, or using grid paper ([2], p. 97).

The basic lines of the definition are proposed by Gini himself ([3]) in the two formulas that we will see here, that appear in the descriptive statistical manuals such as [4], [5], [6] or [7] or in works such as those from [1], [8], [9], [10] or [11], to give just a few examples. However, we think that at this point there are still certain aspects that we need to point out or need to be considered.

2. Definition of the Gini index or concentration ratio

The Gini index, R, or concentration ratio as it is called by Gini, is expressed by [3], p. 213-4, in two alternative ways that do not provide the same result, although the interpretation of both of them leads to similar conclusions.

In [3], p. 213, appears a first formulation of the concentration ratio that depends on the differences between p_i and q_i:

$$R = \frac{\sum_{i=1}^{N-1}(p_i - q_i)}{\sum_{i=1}^{N-1} p_i} \tag{1}$$

where p_i is the cumulative relative frequency[2] of each element or individual (N in total) and q_i is the proportion of the variable cumulated to that element or individual with respect from the total of the variable.

$$q_i = \frac{\sum_{j=1}^{i} x_j}{\sum_{j=1}^{N} x_j} \tag{2}$$

By definition p_N and q_N are both 1, so their difference is zero and it is not necessary to include those terms in the formula of equation (1).

If we were in a situation of total equity, it would be fulfilled in all the cases that $p_i = q_i$, so that the index would be equal to zero. A Gini index close to zero would, therefore, be understood as a situation close to equity.

On the other hand, if the variable were concentrated in a single individual or element, it would occur that $q_1 = q_2 = \cdots = q_{N-1} = 0$ so the numerator and denominator of R would coincide and the quotient would equal one. A Gini index value close to one would, therefore, be understood as a situation close to total concentration.

Thus, the index would be limited between 0 and 1, providing a clear understanding of the situation of equity or concentration that occurs in a distribution, and allowing at the same time the comparison with results of alternative distributions, being a non-dimensional indicator that does not depend on the units of measurement of the data at hand.

Graphical representation of pairs(p_i, q_i), and the union of these points provides what is known as the Lorenz curve or concentration curve. From this representation, [3], p. 214, proposes "a discrete default approximation of *R*" that compares, on the one hand, the area contained between the Lorenz curve and the

[1] Although the distribution of income is the most commonly used, it is not the only one, since it can be applied to any variable that can be distributed in alternative ways. It has even been applied even to analyse the distribution of playing time among the players of a football team, so that an attempt was made to identify whether there were the typical "rotations" if the distribution of time turned out to be equitable or if, on the contrary, there were no rotations, but rather a group of incumbent players and another group of substitute players if the distribution was not equitable.

[2] $p_i = i/N$ for individual data. If we use a statistical table defined by intervals, p_i is the total elements up to that level divided by *N*.

equity line and on the other hand the area of the triangle with vertices in $(0,0)$, $(1,0)$ and $(1,1)$, which would be the area between the equity line and the Lorenz curve in the case of total concentration as it appears in equation (3). The reading of the results of this index would be the same as the one seen in the previous formulation.

$$R = \frac{\frac{1}{2} - \sum_{i=1}^{N} \frac{(p_i - p_{i-1})(q_i + q_{i-1})}{2}}{\frac{1}{2}} = 1 - \sum_{i=1}^{N} (p_i - p_{i-1})(q_i + q_{i-1}) \quad (3)$$

3. The problem is the definition of the Gini index as a comparison of areas

The main problem of the previous definition is that in the situation of total concentration of a variable, the Lorenz curve will not coincide with the cathetus of that triangle with vertices in $(0,0)$, $(1,0)$ and $(1,1)$; this implies that the Gini index calculated as a relation of areas will not be able to reach the maximum of 1, complicating both the interpretation of the ratio and the comparison with results of other distributions of different sample size.

To visualize this difficulty let us take a limit distribution in which we study the situation of equity between two individuals, where one has nothing and the other has everything (table 1 and figure 1 represent this situation). In this case, the dashed line represents the polygonal joining the points $(0,0)$, $(0{,}5,0)$ and $(1,1)$ and indicates the situation of maximum concentration, which however does not coincide with the cathetus of the lower triangle represented by dots[3]. Thus, the Gini index would be $0{,}5$ calculated with the formula of equation (3), which would be far from the situation of total concentration that it is supposed to represent.

Table 1. Limit distribution between two individuals

Individual	**Variable**	p_i	q_i
1	0	0,5	0
2	1	1	1
Sum	1		

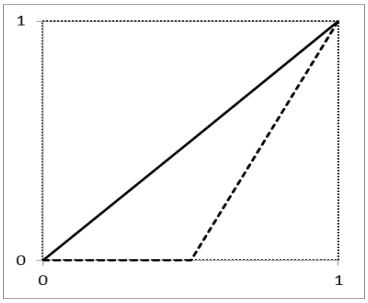

Figure 1. Lorenz curve limit distribution between two individuals

It is possible to think that the problem is reduced when increasing the number of observations, bringing the corner of our triangle closer to the point $(1,0)$, but without ever reaching that point with exactitude, so we would never say *stricto sensu* that the value corresponding to the total concentration is reached attending to the Gini index. Table 2 shows the maximum value of the Gini index for samples of different size, such as the one in the previous example, in which we increase the number of zeros and there is only a 1 in the data of the variable, and we observe how, effectively, this upper limit is not reached.

[3] This same problem is already evident, for example, in [9], p.27.

Table 2. Maximum value of the Gini index for samples of different size

Total data	0	1	*R*, Gini index maximum
10	9	1	0,9
20	19	1	0,95
50	49	1	0,98
100	99	1	0,99
200	199	1	0,995
500	499	1	0,998
1000	999	1	0,999
2000	1999	1	0,9995
5000	4999	1	0,9998
10000	9999	1	0,9999

4. A solution for the calculation of the Gini index as a comparison of areas

To solve this calculation problem[4] that leads to a proper interpretation of the index we should not divide by ½ in equation (3) but by the area corresponding to the case of total concentration. In this case of total concentration, the previous index would be

$$R = \frac{\frac{1}{2} - \frac{(1-p_{N-1})(1)}{2}}{\frac{1}{2}} = 1 - (1 - p_{N-1}) = p_{N-1} \tag{4}$$

And dividing R by the quantity obtained in (4) we would have an alternative ratio

$$R' = \frac{1-\sum_{i=1}^{N}(p_i - p_{i-1})(q_i + q_{i-1})}{p_{N-1}} = \frac{1-\sum_{i=1}^{N} f_i(q_i + q_{i-1})}{p_{N-1}} \tag{5}$$

where f_i is the relative frequency and p_i is the cumulated relative frequency. Thus, we guarantee that this index also varies between 0 and 1, being able to reach both extremes, as does the first version of the concentration ratio (equation (1)).

5. Equivalence between the two formulations

It can be verified that equations (1) and (5) are equivalent if we work with original data, not grouped in tables. In this case, it is evident that it is fulfilled that $f_i=1/N$, $p_i=i/N$ and $p_{N-1}=(N-1)/N$, and by equalizing both expressions we have:

$$R = \frac{\sum_{i=1}^{N-1}(p_i - q_i)}{\sum_{i=1}^{N-1} p_i} = \frac{1-\sum_{i=1}^{N} f_i(q_i + q_{i-1})}{p_{N-1}} \tag{6}$$

We start by substituting $f_i = 1/N$, $p_i = i/N$ and $p_{N-1} = (N-1)/N$ in (6), so that it results in:

$$\frac{\sum_{i=1}^{N-1}\left(\frac{i}{N} - q_i\right)}{\sum_{i=1}^{N-1}\frac{i}{N}} = \frac{1-\sum_{i=1}^{N}\frac{1}{N}(q_i + q_{i-1})}{\frac{N-1}{N}} \tag{7}$$

[4] Concerning the formula of equation (3), [3], p.214, shows an example with a result that "...is certainly approximate by default, since the fractioned curve is internal to the effective concentration curve. Some other values of *R* more approximate can be obtained through different formulas." It should also be added, as we have seen, that it is also smaller because it is divided by ½, which is a higher area than the one that should really be compared to.

In this expression the following addition is in both terms,

$$\sum_{i=1}^{N-1}\frac{i}{N}=\frac{1}{N}\sum_{i=1}^{N-1}i=\frac{1}{N}\frac{1+(N-1)}{2}(N-1)=\frac{1}{N}\frac{N}{2}(N-1)=\frac{N-1}{2} \tag{8}$$

So that it can be written:

$$\frac{\frac{N-1}{2}-\sum_{i=1}^{N-1}q_i}{\frac{N-1}{2}}=\frac{1-\sum_{i=1}^{N}\frac{1}{N}(q_i+q_{i-1})}{\frac{N-1}{N}} \tag{9}$$

We multiply by 2 both the numerator and the denominator in the term on the left and by N both the numerator and the denominator in the term on the right, so we have:

$$\frac{N-1-2\sum_{i=1}^{N-1}q_i}{N-1}=\frac{N-\sum_{i=1}^{N}(q_i+q_{i-1})}{N-1} \tag{10}$$

We can eliminate the denominators and we get:

$$N-1-2\sum_{i=1}^{N-1}q_i=N-\sum_{i=1}^{N}(q_i+q_{i-1}) \tag{11}$$

And as

$$\begin{gathered}\sum_{i=1}^{N}(q_i+q_{i-1})=\sum_{i=1}^{N}q_i+\sum_{i=1}^{N}q_{i-1}=\\ =(q_1+q_2+\cdots+q_{N-1}+q_N)+(q_0+q_1+q_2+\cdots+q_{N-1})=\\ =q_N+2\sum_{i=1}^{N-1}q_i\end{gathered} \tag{12}$$

since $q_0=0$. Replacing the result of (12) in (11) we have:

$$N-1-2\sum_{i=1}^{N-1}q_i=N-q_N-2\sum_{i=1}^{N-1}q_i \tag{13}$$

Which is always true because $q_N=1$. It is, therefore, demonstrated that both formulations are equivalent when working with original data.

6. Does (sample) size matter?

Looking at figure 2 it seems that the size does matter, since the dashed line curve is completely internal to solid line curve and it is, therefore, closer to the line of equity, so that it can be understood that the distribution is more equitable. However, the Gini index corresponding to both distributions (using either of the two formulations we are handling) is the same and equal to $1/3$.

What the eye does not see is that as they are not referred to samples of the same size, the drawing of the maximum concentration in each one of them does not coincide, so the Lorenz curves cannot coincide even representing equivalent situations: the dashed line curve refers to a sample of 4 data so that the curve of maximum concentration would have its corner in the point $(0{,}75, 0)$ while the solid line curve refers to a sample with an equivalent distribution between 20 observations so that the corner of the curve of maximum concentration would be in the point $(0{,}95, 0)$.

The conclusion would be that to compare Gini index corresponding to different distributions, the size of the sample does not always matter if we use the formulations we have presented here (equations (1) or (5)), which makes it interesting again to practice the correction we have proposed herein equation (5). If this is not the case, we can arrive at results that do not correspond to reality.

At the same time, it is useful to also qualify the exploitation of the Lorenz curves in the sense that the graphical representation of two curves only allows to deduce which of the two is more equitable if the sample size is the same for both of them (and if the curves are not crossed); but if the sample

size is not the same or if the different tables do not have the same number of intervals, the curves are not valid to draw conclusions.

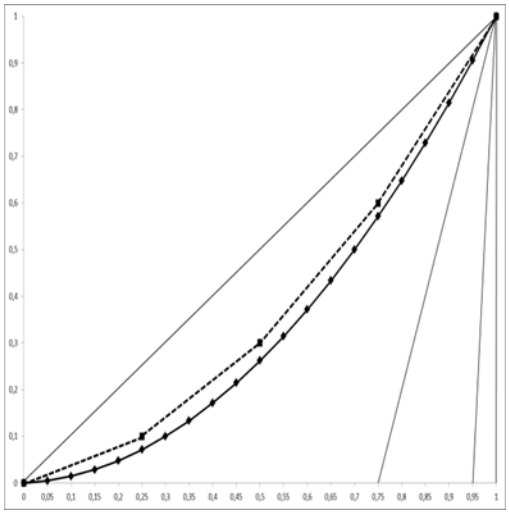

Figure 2. The problem of size

7. Conclusions

Several important questions about the definition and use of the Gini index are therefore evident:

That the Gini index defined as equation (3) does not usually reach the maximum of the interval [0, 1], and therefore the correction in (5) is recommended.

For individual data, the Gini index defined as the ratio of the sum of the differences between p_i and q_i split by the sum of the pi is equivalent to that which relates the area generated between the equity line and the Lorenz curve divided by the cumulative relative frequency of the penultimate data. (Equations (1) and (5)).

That the sample size can influence the shape of the Lorenz curve, but it does not influence the index defined according to equations (1) and (5).

Acknowledgments

I thank the referees for their helpful comments and suggestions that have improved the paper.

References

1. S. Yitzhaki, More than a dozen ways of spelling Gini, *Research on Economic Inequality*, **8**, (1998) 31-38.
2. G. Calot, *Curso de estadística descriptiva*. (Editorial Paraninfo; Madrid, 1985).
3. C. Gini, *Curso de estadística*. (Editorial Labor, Madrid, 1953).
4. A. Alcaide, *Estadística aplicada a las ciencias sociales*. (Editorial Pirámide, Madrid, 1976).
5. M.P. Martín-Guzmán and F.J. Martín Pliego, *Curso básico de estadística económica. 2º ed*.. (Editorial AC, Madrid, 1987).
6. J.M. Sarabia, *Curso práctico de estadística. 2ª ed*. (Civitas ediciones, Madrid, 2000).
7. U. Carrascal, *Estadística descriptiva con Microsoft Excel 2010* (Ra-Ma Editorial, Madrid, (2011).
8. E. Ferreira and A. Garín. Una nota sobre el cálculo del índice de Gini, *Estadística Española*, Vol. **39**, nº 142, (1997) 207-218.
9. J. M. Montero, Sobre concentración económica: Índice E para colectivos discretos. *Estadística Española* Vol. **45**, nº 152, (2003) 22-54
10. L. Ceriani and P. Verme, The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini *The Journal of Economic Inequality* **10** (2012) 421-443.

11. S. Yitzhaki and E. Schechtman, *The Gini Methodology: A Primer on a Statistical Methodology* (Springer, New York, 2013).