Gini Coefficient

Nicola Altobel, Matteo Golinelli, Giulio Schincariol

April 2025

1 Introduction

The Gini Coefficient, or Gini Index, is a valuable tool used to describe the concentration of a quantitative characteristic within a given population. This index was developed by Professor Corrado Gini during the early decades of the 20th century, with the aim of finding an efficient way to represent the variability and concentration of such characteristics, and to make meaningful comparisons between them.

1.1 Definition of the Index

After introducing the arithmetic mean of differences as an index of variability [3], Professor Corrado Gini further expanded on this concept in his 1914 article "Sulla Misura della Concentrazione e Variabilità dei Caratteri". In this work, he proposed an index capable of producing results independently of the distribution curve of the studied characteristic. At the time, the most widely used index was obtained from an ordered set $\{a_1, a_2, ..., a_n\}$ computing the values given by the formulas (1) for each i (where $i < n$). The inequality $p_i > q_i$ served as an indicator of concentration: the greater the inequality, the more concentrated the characteristic was deemed to be.

$$q_i = \frac{\sum_{k=1}^{i} a_k}{\sum_{k=1}^{n} a_k} \qquad\qquad p_i = \frac{i}{n} \tag{1}$$

Building on this proposition, Gini suggests using the quotients $R_i = \frac{p_i - q_i}{p_i}$ instead. The values of R_i range between 1 (representing perfect concentration) and 0 (corresponding to the cases of equidistribution of the characteristic). As a measure of overall concentration, Gini proposes calculating the weighted mean of the n - 1 values of R_i obtained through equation (2).

$$R = \frac{\sum_{i=1}^{n-1} (p_i - q_i)}{\sum_{i=1}^{n-1} p_i} \tag{2}$$

The index R satisfies the following properties:

a R increases as the concentration of the characteristic increases.

b $R = 1$ in cases of perfect concentration.

c $R = 0$ in cases of minimum concentration (i.e., perfect equality).

d The product $R \cdot p_i$ corresponds to the expected value of $(p_i - q_i)$ in cases where the characteristic falls below a certain threshold.

In his article, Professor Gini emphasizes that R is a robust and well-defined measure that can be applied to any quantitative characteristic and used to compare concentration levels across different distributions. Moreover, the index can be approximated with reasonable accuracy by expressing it as the ratio of the mean difference to the value that the mean difference would assume under conditions of minimal inequality. Additionally, the index can be derived from the graphical method proposed by Lorenz (1905), Chatelain (1907), and Séailles (1910), which is commonly used to depict the distribution of wealth.

2 How to compute the Gini Index

The Gini coefficient can be calculated using two primary methods. Although both yield the same result, they offer different perspectives on what the coefficient represents.

2.1 The "mean" method:

The Gini index as written in the introduction is an indicator that is normally used expresses the expected absolute gap between people's incomes relative to the mean income in the population. This way of calculating it is more intuitive and is well explained in the article [4], reporting the example:

> *"Imagine that our population now consists of just two people who met on the street, and their total income equals 100$. The average income between the two people is 50$. If inequality is at its highest, so that one person has 100$, and the other has 0$, the gap between them will be 100$, which is twice the mean income."*

By understanding this simple problem, we can understand how the formula works for upscaled cases. In fact, the Gini index (G) is determined mathematically as the average of the absolute value of the relative mean difference in incomes between all possible pairs of individuals:

$$G = \frac{1}{\bar{y}(n-1)} \sum_{i \neq j}^{n} \frac{1}{n} \sum_{j}^{n} \frac{|y_i - y_j|}{2} \tag{3}$$

where: $\bar{y}$ = mean of the incomes, n = number of individuals, y_i and y_j the income of individuals i and j.

2.2 The "Lorenz curve" method

In order to understand this way of computing the Gini index, let us define the *Lorenz curve*: it is a graphical representation of the distribution of income or of wealth. It was developed by Max O. Lorenz in 1905 to represent inequality of wealth distribution [8].

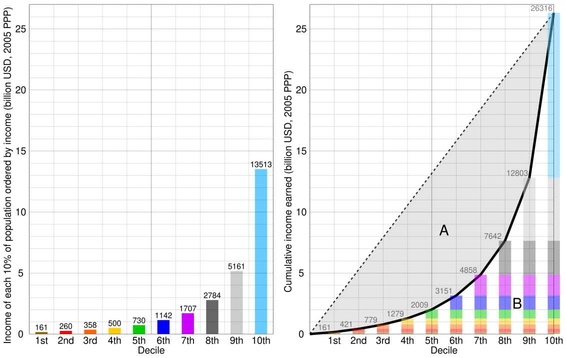


Figure 1: Lorenz curve

The Lorenz curve $L(p)$, as we can see in Figure 1, shows that the distribution of wealth in the population is not omogeneous otherwise it would have followed the *"line of equality"* $(y = x)$, with this method the Gini coefficient highlights how far the Lorenz curve falls from this line by comparing the areas A and B:

$$G = \frac{A}{A+B} = \frac{\frac{1}{2} - \int_0^1 L(p)dp}{\frac{1}{2}} = 1 - 2\int_0^1 L(p)dp \tag{4}$$

Where $A + B = \int_0^1 x dx = \frac{1}{2}$

3 Relating to probability

3.1 A relation with the expected value

The Gini index, being a measure of inequality within a population, is strongly related to probability and statistics. Consider the following experiment:

> *Pick a dollar earned by a US. household at random, assuming that every dollar is equally likely to be chosen. Record the value of the percentile variable, p, of the person who earned that dollar.*

As shown in [2], the random variable p has probability distribution function (PDF):

$$s(p) = \frac{d}{dp}L(p)$$

where $L(p)$ is the Lorenz curve associated with p. The PDF $s(p)$ is called the *share distribution* function, since it tells us what share of the whole is owned by the portion of population in the percentile p. It is proven in [2] that the Gini index and the expected value of p, denoted as $\bar{p}$ are related by:

$$G = 2\bar{p} - 1 \quad and \quad \bar{p} = \frac{G+1}{2} \tag{5}$$

Consider the experiment explained above. The Gini coefficient of household income in the US in 2006 was 0.47. This means that the average dollar was earned at a percentile level of $\bar{p} = \frac{0.47+1}{2} = 0.735$, or above the 73rd percentile.

3.2 The lower of two incomes

Furthermore, another interesting relation involving the Gini index is presented in [2]. Consider this new experiment:

Pick two households in the US and record the lower of their incomes
as Y, a random value that takes values in $[0, +\infty)$

It can be proven that the expected value of Y, in ratio to the mean income $\bar{X}$, is equal to the complement of the Gini index:

$$\frac{\bar{Y}}{\bar{X}} = 1 - G \tag{6}$$

Considering again 0.47 as the Gini index, we can conclude that the lower income of two random households is $53\% = 1 - G$ of the mean income. This means that, on average, the poorer of two randomly chosen families in the US earns only about half the national mean.

4 Examples

4.1 Frequency of words usage in English

Considering the Gini index as a valuable measure of concentration, we can apply this tool to various cases involving the distribution of characteristics. A simple example can be found by analyzing the frequency of word usage in average English communication.

We can reasonably expect a high Gini value in this context: the most frequently used words typically serve to support and structure the communication, providing logical flow to the actual content. These common words act as connectors between less frequently used, meaning-bearing terms.

By examining the 500 most commonly used words [1], we already gain a useful perspective on the distribution (Figure 2). Since we are not considering the entire corpus of English words, we must normalize the observed frequencies. This normalization allows us to obtain the desired distribution for analysis.

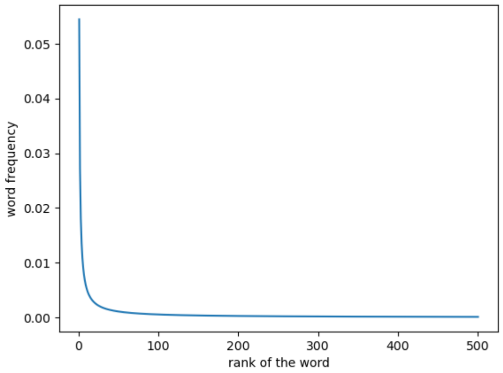


Figure 2: Representation of the distribution.

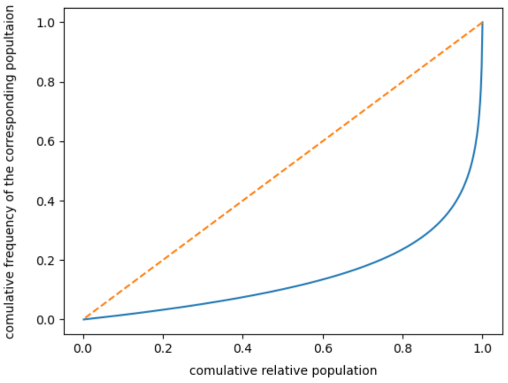


Figure 3: Representation of the Lorenz curve

To compute the Gini index, the data must first be sorted in increasing order, after which the cumulative frequencies are calculated. Using a simple program, these values can be efficiently computed, resulting in the corresponding Lorenz curve (Figure 3). With this refined dataset, we can proceed to calculate the Gini index as described in the previous section. The result is a value of 0.7135, which supports our hypothesis. Notably, since the analysis is based solely on the 500 most frequently used words, the index is expected to assume even higher values when a larger corpus is considered. This is due to the fact that each additional word typically appears with a progressively lower frequency, causing the Lorenz curve to become even steeper compared to the case studied.

4.2 Gini index and Olympics medals

The Gini index is widely used in sports to measure inequality in salaries or performance. In this example, we will compute the Gini index of the all-time Olympics games medal table to understand whether the distribution between countries is equal or not.
Firstly, we computed the Gini index of the Olympic medal table of all time by country.

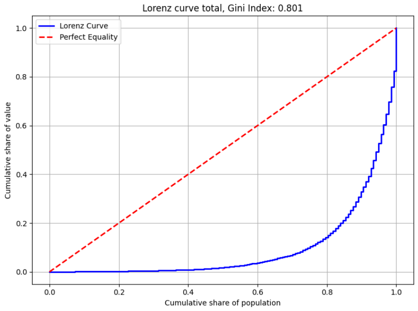


Figure 4: Lorenz curve of all-time medals. Data from [5]

This resulted in a Gini index $G = 0.801$, which means that the concentration is strongly inequal, as a little amount of countries have the most amount of medals.

In our second computation we considered to normalize the amount of medals won by the population of the country, in order to get a more meaningful result, since it is clear that a bigger country statistically wins more medals than a smaller one. This, though, resulted in an even higher Gini index (0.845):

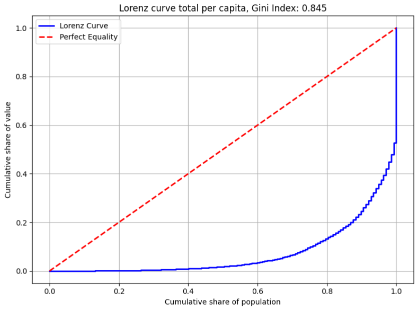
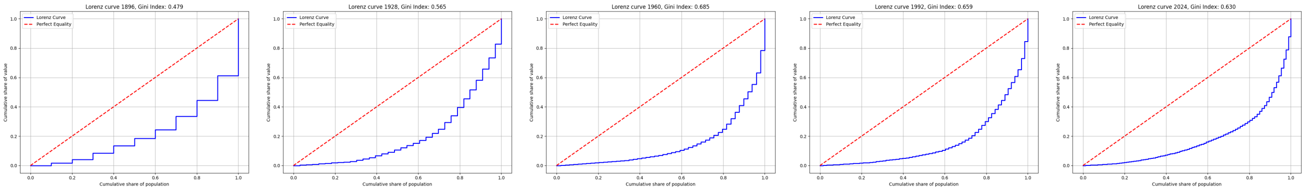


Figure 5: Lorenz curve of medals per capita. Data from [5]

This happens because there are small countries that won a little amount of medals, but have such a small population that their medals to inhabitants ratio is very high. San Marino, for example, has won only three medals, but its medals-to-habitant ratio is 10 times bigger than any other country. Another approach, which we did not consider, is to normalize the number of medals by the GDP of each country.

In this scenario, inequality is not necessarily unfair, since as we said bigger Countries are expected to win more medals than small ones, but the Gini index still gives us an insight in how concentrated Olympic success is between Countries.

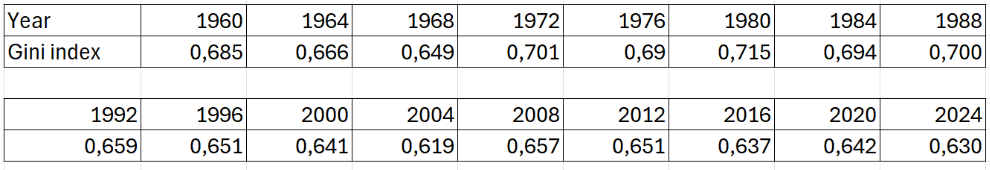
An interesting observation is to analyze whether the concentration increased or decreased over time. We considered data from 1896, 1928 and from 1960 to 2024.



(a) (b) (c) (d) (e)

Figure 6: (a) 1896, Gini:0.479 (b) 1928, Gini:0.565 (c) 1960, Gini:0.685 (d) 1992, Gini:0.659 (e) 2024, Gini: 0.630. Data from [6]

This shows that in 1896 medals were more equally spread, then the Gini index stabilized around 0.6, as also shown by the Gini indexes from 1960 to 2024:



Year	1960	1964	1968	1972	1976	1980	1984	1988	1992	1996	2000	2004	2008	2012	2016	2020	2024
Gini index	0,685	0,666	0,649	0,701	0,69	0,715	0,694	0,700	0,659	0,651	0,641	0,619	0,657	0,651	0,637	0,642	0,630

This shows that the distribution of medals per Country has been roughly the same for the past 65 years. A possible interpretation of this result is that the amount of medals won by a Country also depends on its historical sports and Olympics culture, and since this factor does not change rapidly in time, also the amount of medals won in each Olympics games are similar in a relatively short period of time, as the one we considered.
Interestingly, the Gini index of single years is consistently smaller than the one computed on the all-time medal tally. Our interpretation is that this happens because in the all-time medal table 137 countries are considered (all having won at least 1 medal), while in most Olympics the amount of countries is around 80. Since many countries have only won one medal, they make the all-time Gini-index considerably higher than the single Olympics ones.

4.3 Gini index: Evaluation of the Environmental Heterogeneity and Habitat Suitability

While learning and researching about the Gini index, we came across this Chinese research in which the index is utilized in a different environment. As we will see from this Chinese paper, the Gini index is not only used in the economic field; instead is, as we will see from this Chinese paper, also used as a measure of the heterogeneity of an habitat [7]. Giving a bit of context: the paper we're writing about is analysing different variables such as:

- SST (Sea Surface Temperature): It's a critical parameter in marine studies as it influences weather patterns, ocean currents, and the distribution of marine organisms.
- SSS (Sea Surface Salinity): Salinity affects seawater density, which in turn influences ocean circulation and climate.
- SDO (Dissolved Oxygen Saturation): Low SDO (or low DO in general) can indicate poor water quality and can stress or kill fish and invertebrates.
- SCHL (Chlorophyll-a Concentration): High values can indicate nutrient pollution and potential for harmful algal blooms

All this parameter affect the spawning and the survival of fish (in particular this research focused on estuarine gobies) and sea-life of the Yangtze Estuary. The paper is looking for heterogeneity characteristics of various environmental factors between 2018-2020 summer and spring.

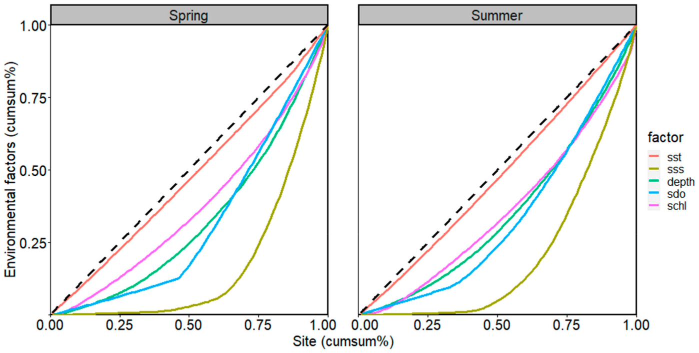


Figure 7: The Lorentz curve of environmental factors in the Yangtze Estuary in the spring and summer (2018–2020); the black dashed line represents the line of equality

As we know the colorful lines are the Lorenz curves, the greater the curvature of the curve, the more various the environmental factors are distributed spatially, the research indicated that the salinity of the YRE had high spatial heterogeneity (green curve Figure 7), with a Gini index of 0.62; the temperature had the lowest spatial heterogeneity (orange curve Figure 7), with a Gini index of 0.05, and there were significant seasonal changes. This permitted scientists to locate several best areas of the estuary and moments where the gobies could live and spawn better. This was an interesting use and example of the Gini index outside the field it was originally though for.

References

[1] American National Corpus. https://anc.org/data/anc-second-release/frequency-data/.

[2] F. A. Farris. The gini index and measures of inequality. *The American Mathematical Monthly*, 2010.

[3] C. Gini. Variabilità e mutabilità. *Tipografia di Paolo Cuppini, Bologna*, 1912.

[4] J. Hasell. Measuring inequality: what is the gini coefficient? *Our World in Data*, 2023. https://ourworldindata.org/what-is-the-gini-coefficient.

[5] Olympics all-time medals table. https://www.topendsports.com/events/summer/medal-tally/all-time-comparison-pop-all.htm.

[6] Olympics medals tables. https://www.olympic-museum.de/medal_table/olympic-games-medal-table-2020.php.

[7] Vv.Aa. Application of the gini index on the evaluation of the environmental heterogeneity and habitat suitability index for larval gobies. *Journal of Marine Science and Engineering*, 2023.

[8] Wikipedia. Lorenz curve — Wikipedia, the free encyclopedia, 2025.