On Corrado Gini's 1932 paper *"Intorno alle curve di concentrazione"*. A selection of translated excerpts*

Giovanni Maria Giorgi, Stefania Gubbiotti

Abstract

Metron, continuing its editorial policy of diffusion of the main scientific results reached by the so-called Italian Statistical School in the first half of the XXth century, publishes an English synthetic version of the article "Intorno alle curve di concentrazione" written by Corrado Gini in 1932, in the occasion of the 50th anniversary of his death.

The main focus of the paper is the study of the concentration curve, with special emphasis on its fundamental features and properties and on the relationship with other relevant curves. In particular, the Author aimed at investigating on the goodness of the approximation of the concentration area, and therefore of the concentration ratio, both from a methodological and an empirical point of view, in some specific cases in which Cotes quadrature formula can be applied. One of the most innovative contributions is the alternative analytical representation of the concentration curves in a coordinate system which assumes the so-called equidistribution line as x-axis and its perpendicular line as y-axis. Furthermore, the impact of the presence of a superior and/or inferior limit in the variable of interest on the maximum concentration triangle is examined: suitable correction coefficients are derived for computing the corresponding concentration ratio, that take into account these restrictions.

Keywords Concentration Curve, Gini Concentration Ratio, Equitension, Equiconcentration.

1 Introduction

Corrado Gini, the founder of *Metron*, passed away on March 13, 1965. To celebrate the 50th anniversary of his death, *Metron* continues its editorial policy of promoting the diffusion of the main results achieved by scholars belonging to the Italian statistical tradition in the first half of the XXth century. This occasion allows us to follow the advice of Camilo Dagum (1925-2005) who, immediately after the "International Conference in Memory of two Eminent Social Scientists: C. Gini and M.O. Lorenz" held in Siena on May 23-26, 2005, suggested in a private correspondence to Giovanni Maria Giorgi (at that time Editor of *Metron*) to publish an English translation of the article "Intorno alle curve di concentrazione" (1932) by Gini. Unfortunately, the sudden and unexpected death of Camilo Dagum and the overlap of a considerable amount of commitments of G.M. Giorgi did not allow to take into account his suggestions in due time. Dagum, who had the opportunity of knowing Gini personally when he visited the Institute of Statistics in Rome, thought that some results of Italian statisticians should have been highlighted at international level in order to prevent their rediscovery. A brief synthesis of the topic is now in order, to let the reader understand Dagum's point. He had noticed, indeed, that the analytical representation of the concentration curves derived by Gini (1932) had been proposed decades later[1]. In particular, as he stressed in his review of Kakwani's book (1980) published in Journal of Business & Economic Statistics (1986, vol.4, n. 3, p.391), "the new coordinate

*The original complete version was published on *Metron*: Gini C., *Intorno alle curve di concentrazione*, Metron, **9**(3-4), 3–76 (1932).

†**G.M. Giorgi** - email: giovanni.giorgi@uniroma1.it

‡**S. Gubbiotti** (corresponding author) - email: stefania.gubbiotti@uniroma1.it - Dipartimento di Scienze Statistiche, Sapienza Università di Roma - Piazzale Aldo Moro n. 5, 00185 Roma, Italia

[1]It is worth to remark that Gini and his collaborators indifferently refer to the *concentration curve* or *Lorenz curve*, since, to quote Pietra (1937, p.315) it is "the same curve arranged this way or that", depending on the choice of axes.

system for the Lorenz curve was first introduced by Gini in a contribution to the XXth Session of the International Statistical Institute and further developed by Gini (1932), Galvani (1932), and other Italian scholars". Then, Kakwani and Podder (1976) came to the same result, of course independently of Gini, and Kakwani (1980, ch.7) analyzed in depth all the consequences of the new representation. In our opinion, Kakwani and Podder still have the great merit of letting a wide audience of scholars know and take advantage of this result both in further research and applications.

Our purpose is to make a selection of translated excerpts from Gini's work finally available to the scientific community, in order to acknowledge the authorship of his genuine idea and to avoid further possible misunderstandings and inaccuracies. In this way we celebrate the anniversary of Gini and, at the same time, we accept the suggestion of Camilo Dagum, who contributed in a significant way to the innovation, extension and diffusion of Gini's scientific work (see, e.g., [5, 7, 8, 9]).

Editorial requirements led us to consider the salient aspects of Gini's article, given the extension of the original version (73 pages), which is available upon request for those readers who may be interested in further details. Here we try to find a compromise between the faith to Gini's text and the effort in improving the readability for a modern scholar. Specifically, we keep the genuine notation[2] and, in the footnotes, we refer to the original paragraph numbering, while providing additional section titles to make the paper easier to follow. Most of the original figures are preserved and reproduced, with few exceptions that will be highlighted in the text, when necessary: for the sake of brevity we omitted, in fact, some redundant figures that we considered not fundamental to the overall comprehension of the paper.

Before undertaking his broad overview on the concentration curves, Gini recalls that one of the variability measures he previously introduced (see [14]), the mean difference, was gaining increasing popularity among statisticians. Many authors, indeed, consider it to be the most suitable variability index (see, for instance, [1, 2, 3, 4, 10, 12, 19, 20, 21, 25, 26, 27, 28, 31]). Moreover, the relationships between the mean difference and the concentration curve, first analyzed in [15], was further highlighted and discussed in [2, 10, 19, 20].

2 Relationships between the concentration curves and the other distribution curves

According to Gini, a fundamental key to understand the concepts of variability and concentration is the analysis of the basic relationships between the concentration curves and the other distribution curves[3].

Let v be a generic value of the variable of interest, f_v the number of occurrences of v, p_v the number of occurrences of values not larger than v, q_v the corresponding cumulative quantity, and $n = \sum f_v$ the total number of observations.

It is possible to distinguish the following three categories of distribution curves of a variable.

a. **Graduation curves.**[4] Given a Cartesian coordinate system, a *graduation curve* is obtained by representing v (on the y-axis) with respect to p_v (on the x-axis). For example, a graduation curve, that Galton previously called *ogive* (see [11]), can be described by sorting a group of people by height and drawing the ideal curve connecting their heads profile. Moreover, a 90 degrees left rotation of the graduation curve yields the Pareto curve of incomes, which usually represents the values v of unit income on the x-axis and $n - p_v$, i.e. the number of people with a unit income larger than v, on the y-axis.

b. **Frequency curves.** A *frequency curve* represents f_v as a function of v.

c. **Concentration curves.** Finally, in a *concentration curve* the values q_v are plotted against p_v.

[2] Here a brief reminder of Gini's notation is in order: A denotes the arithmetic mean, M is the median, ${}^1S_A$ and ${}^1S_M$ are the simple mean deviations from the mean and the median respectively, Δ is the mean difference without repetition, R is the concentration ratio. Other symbols will be introduced through the text.

[3] This section corresponds to paragraph 2 of the original paper.

[4] According to the glossary *A Dictionary of Statistical terms* by M.G. Kendall and W.R. Buckland [24], apart from a change of axes, the *graduation curve* (in Italian, *curva di graduazione*) is equivalent to the distribution curve.

By integrating a frequency curve from right to left and by rotating it 90 degrees left, one retrieves the corresponding graduation curve. Then, a subsequent integration of the graduation curve from left to right yields the related concentration curve. Figure 1 reproduces a very effective plot that Gini conceived to summarize these relationships between the three categories. Here, the first figures of the original paper representing each separated curve are skipped.

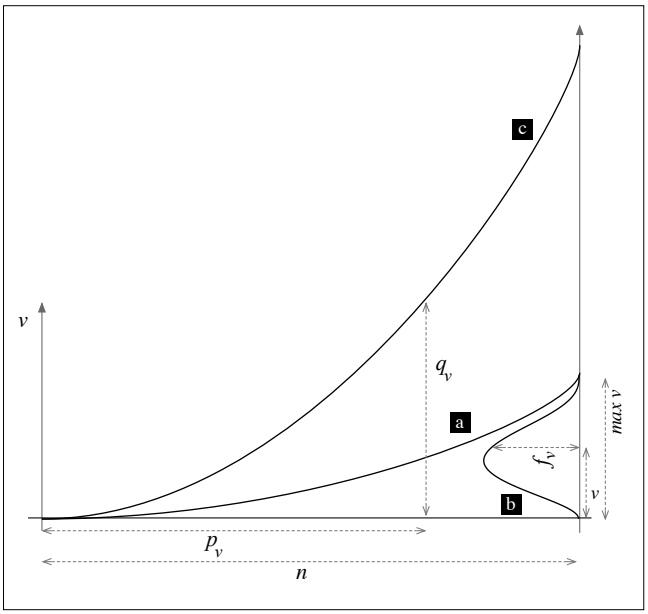

Figure 1: Summary of the relationship between the three distribution curves: a) graduation curve, b) frequency curve, c) concentration curve.

As a consequence of the above mentioned property, if a frequency curve is symmetric with respect to (w.r.t.) the median (e.g. the normal distribution), the corresponding graduation curve is symmetric as well w.r.t. its median point: starting from this point, indeed, the values v on the y-axis decrease (respectively increase) towards the left (respectively right) in a given order. In fact, it is straightforward to show that symmetry of the frequency curve implies the following relations:

$$\int_{-\infty}^{o-h} f_v dv = \int_{-\infty}^{o} f_v dv - \int_{h}^{o} f_v dv,$$

$$\int_{-\infty}^{o+h} f_v dv = \int_{-\infty}^{o} f_v dv + \int_{h}^{o} f_v dv,$$

where o is the median of the frequency curve and $h > 0$. The two integrals on the left side of the above equations represent the values v on the y-axis of the graduation curve that are equally distant from its median point which corresponds to a y-value $\int_{-\infty}^{o} f_v dv$.

For the sake of the comparison between the different curves, it is common practice to eliminate the influence of the different number of cases, by letting the areas delimited by each curve and the x-axis being equal. Similarly, to make the comparison easier between the graduation curves, it is useful to consider the same number of cases that is represented on the x-axis. Moreover, for the concentration curves it is better to standardize both for the number of cases (x-axis) and for the amount of the variable (y-axis). Finally if p_v and q_v both range between 0 and 1, as shown for instance in Figure 1, they are interpreted as the proportions that the number of values not larger than v and the corresponding amount of the variable represent w.r.t. the total number of values and the total amount of the variable. In addition, it is worth noticing that the only required values to construct the concentration curves are the couples p_v, q_v; since it is not necessary to know the corresponding values v, in the rest of the paper Gini adopts the simpler notation p, q.

3 Approximation of the concentration area

In recalling some of the fundamental features of the concentration curves, Gini states that, for equally distributed variables, the concentration curve becomes a line that is called *equidistribution line*. In Figure 2(a) the equidistribution line is indicated by *ob*; the area delimited by *ob* and the concentration curve *oab* is called *concentration area*, and the *concentration ratio* is the ratio between the concentration area and its maximum value that is given by the area of the triangle *ocb*. In [15] Gini previously showed that the concentration ratio is equal to the ratio between the mean difference (without repetition) between the variable values and twice the arithmetic mean of the variable, in such a way that from the concentration curve one is allowed to determine the mean difference graphically. This feature is practically relevant when not all the values of p_v are known, or when the number of values of v is so large that it is necessary to group them in sufficiently large classes. The main issue, which is dealt with in the present section[5], is then to find the concentration area by knowing only some points of the concentration curve.

By referring to [15], Gini explains how to obtain a piecewise linear curve in place of a proper concentration curve by connecting the known points with linear segments. Again, he points out that the concentration area delimited by the piecewise linear curve would correspond to the mean difference in the ideal case in which all the values belonging to a given class were equal. Since the concentration curve is convex w.r.t. the x-axis, replacing the concentration curve by the piecewise linear curve provides an approximation from below for the concentration area. The Author says that, according to his empirical experience, when the number of classes is above ten and the classes are not too different, the approximation is satisfying. Of course the approximation improves as the number of classes increases. Moreover, to get better approximations, one can assume that the values in each class are not all equal, but they increase according to an arithmetic progression. In practice, this implies the introduction of an integration coefficient in the formulae, which is usually straightforward to be determined. Thanks to this coefficient, a relatively small number of classes (e.g. five) is sufficient to yield good approximations.

In summary, from a practical point of view, it is usually feasible to determine the concentration area even when the values are grouped in classes. However, Gini provides some useful remarks for a deeper analysis of the behavior of the concentration curve.

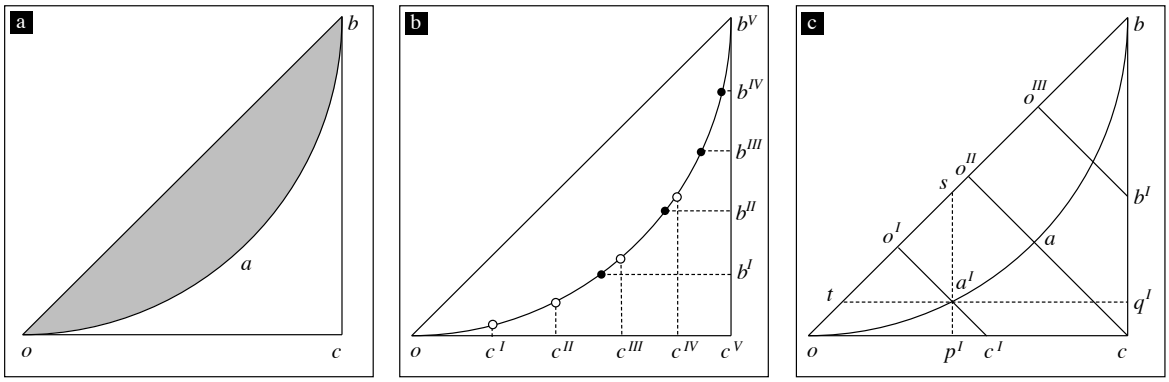

Figure 2: Approximation of the concentration area.

In Figure 2(b) Gini recalls the following procedure proposed by [2]. The range of q is split in five equal parts delimited by the points $b^I, b^{II}, b^{III}, b^{IV}, b^V$. Based on the corresponding points on the concentration curve (black bullets) the concentration area can be computed by adopting Cotes quadrature formula. This procedure is practically suitable only if the values are grouped in classes corresponding to equal amounts, which is often unrealistic. Otherwise, the required interpolations to determine the classes get more complicated, and they inevitably add a further element of uncertainty. Moreover, Gini argues that this method is also questionable from a theoretical point of view: equal intervals for the q values do not correspond to equal parts of the concentration curve. Due to the

[5]This section corresponds to paragraphs 3-7 of the original paper.

convexity of the curve, the first intervals always correspond to the longest parts of the concentration curve. However, it could also be possible to consider equally spaced values of p, instead of q, delimited by the points $c^I, c^{II}, c^{III}, c^{IV}, c^V$ in Figure 2(b). Consequently, the points on the concentration curve would be (systematically) different from the previous ones and the quadrature would result different. To overcome these difficulties, Gini suggests a reasonable alternative way to select the points on the concentration curve, that is to consider approximative equal parts on the curve, which is obtained by taking equal intervals on the equidistribution line (see Figure 2(c)). As a natural consequence, he introduced an unusual orientation of the concentration curve by letting the equidistribution line coincide with the x-axis, as illustrated in the following (see Figure 3(a)).

As mentioned before, here Gini presents an innovative representation of the concentration curve. Starting from Figure 2(c) he considers the segment $o^{II}c$ that divides the angle $\widehat{ocb}$ in two equal parts and that is perpendicular to ob in its median point o^{II}. The segment $o^{II}c$ intersects all possible concentration curves in points such that $p+q=1$. Similarly, the segment o^Ic^I, that is perpendicular to ob in o^I which delimits the segment $oo^I=\frac{1}{4}$ on the equidistribution line ob, intersects all possible concentration curves in points such that $p+q=\frac{1}{2}$. By extension, each line perpendicular to the equidistribution line in the generic point z intersects all possible concentration curves in points such that $p+q=k$, where k is twice the fraction of the length of the segment oz (denoted by $\overline{oz}$) with respect to the total length of ob (denoted by $\overline{ob}$). Since, $\overline{oc}=\overline{cb}=1$, it is $\overline{ob}=\sqrt{2}$ and therefore:

$$\overline{oz}=\frac{p+q}{\sqrt{2}}.$$

It is straightforward to find the distance of z from the corresponding point on the concentration curve, i.e.

$$\overline{za}=\frac{p-q}{\sqrt{2}}.$$

In summary, Gini obtains the representation in Figure 3(a) by considering on the x-axis the values $\frac{p+q}{\sqrt{2}}$ and on the y-axis the values of $\frac{p-q}{\sqrt{2}}$. Finally, by cancelling the constant $\frac{1}{\sqrt{2}}$, the concentration curve is expressed by taking the equidistribution line on the x-axis, with the values $p-q$ varying from 0 to 1, whereas on the y-axis the values $p+q$ vary from 0 to 2 (see Figure 3(b)).

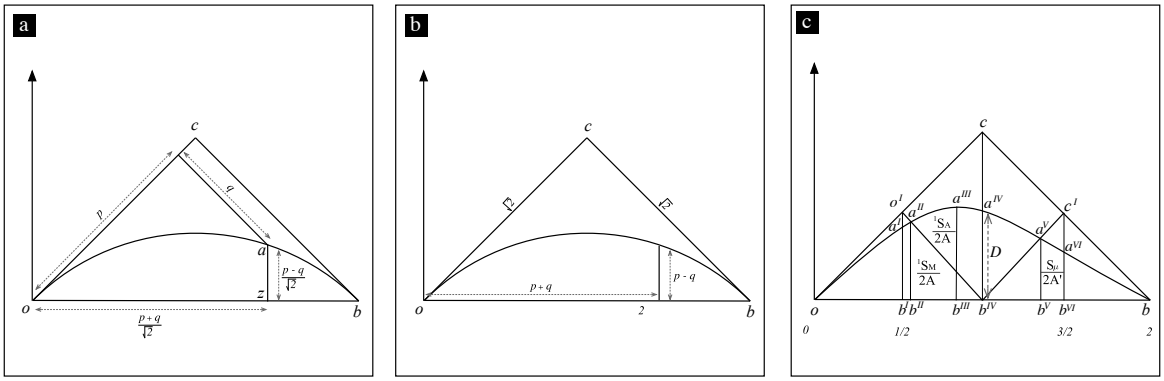

Figure 3: Representation of the concentration curve with the equidistribution line on the x-axis.

As a consequence of the previous arguments, in order to determine the concentration area Gini recommends to select equally spaced points of $p+q$, instead of equally spaced points of p or equally spaced points of q. Furthermore a comparison among these three alternative procedures is supported by two numerical examples; here we provide only one of them for the sake of brevity.

> *Example 1.* Let us consider 100000 individual, whose income constitutes an arithmetic progression, for instance 1, 2, 3, ..., 100000. The total income is therefore $1+2+3+\ldots+100000=5000050000$, and the average income is $A=50000.5$. Moreover the mean difference without repetition is given by $\Delta=(n+1)/3=100001/3$, as shown in

[14] for arithmetic progressions. Finally, the concentration ratio is equal to the ratio between the mean difference and its maximum value $2A$, i.e.

$$R = \frac{\Delta}{2A} = \frac{100001}{3 \cdot 100001} = \frac{1}{3} = 0.333...$$

In this example, Gini aims at comparing this exact value with the approximated results summarized in Table 1, obtained by applying Cotes quadrature formula for interpolation with five prefixed equally spaced (i) values of p, (ii) values of q, (iii) values of $p+q$.[6] Row (iv) in Table 1 corresponds to the value of the concentration ratio computed by graphical tools.[7]

	prefixed values	resulting values	area	R
(i)	$p_1 = 0.2, p_2 = 0.4$	$q_1 = 0.04, q_2 = 0.16$	0.167	0.334
	$p_3 = 0.6, p_4 = 0.8$	$q_3 = 0.36, q_4 = 0.64$	0.167	0.334
(ii)	$q_1 = 0.2, q_2 = 0.4$	$p_1 = 0.447, p_2 = 0.632$	0.159	0.318
	$q_3 = 0.6, q_4 = 0.8$	$p_3 = 0.774, p_4 = 0.894$	0.159	0.318
(iii)	$p_1 + q_1 = 0.4, p_2 + q_2 = 0.8$	$p_1 - q_1 = 0.212, p_2 - q_2 = 0.249$	0.1655	0.331
	$p_3 + q_3 = 1.2, p_4 + q_4 = 1.6$	$p_3 - q_3 = 0.208, p_4 - q_4 = 0.120$	0.1655	0.331
(iv)			0.164	0.328

Table 1: Comparison of different procedures for approximated computation (using Cotes quadrature formula) of the concentration ratio, with $p_0 = 0, p_5 = 1, q_0 = 0, q_5 = 1$

In summary, in this case the best approximation is provided by procedure (ii), followed by (iv). Hence method (iii), proposed by Bortkiewicz in [2], turns out to be the least favorable. However, Gini reports that this rating of the considered procedures performances does not hold true in all applications. The advantage of procedure (ii) w.r.t. (iii) is the more direct determination of the values of q given prefixed p, than viceversa. As regards procedure (iv), on the one hand it allows to split the concentration curve into equal parts, on the other one it requires some additional complications in the computation, which may imply a less satisfactory approximation.

4 Fundamental features of the concentration curves

In this section[8], Gini examines the features of the concentration curves, by describing in details the information condensed in Figure 3(c). First of all, the maximum reachable value of a concentration curve, which is represented by the segment cb^{IV}, is half the basis ob. This suggests to adopt the ratio between its maximum $a^{III}b^{III}$ and the basis ob as a representative index for a concentration curve. Since the maximum value in a generic concentration curve could be reached on the segment cb^{IV}, it may be also reasonable to characterize the curve by $a^{IV}b^{IV}$ or by its complement ca^{IV}. Gini stresses that both choices can be supported by further considerations. The concentration area increases with the concentration and, consequently, the area delimited by the concentration curve and the segments oc and cb tends to be smaller. Hence $a^{III}b^{III}$ and ca^{IV} can be considered respectively as approximate indices of the two complementary areas. Moreover, these two ways of proceeding correspond to two different methods for measuring the concentration of variables. The former is to determine the ratio between the simple mean deviation from the arithmetic average w.r.t. its maximum value, $\frac{^1S_A}{2A}$, which was previously shown by [28] to be equal to the maximum difference between p and q, i.e. $a^{III}b^{III}$ in Figure 3(c). The latter actually consists in taking the difference, denoted by D in Figure 3(c), between the values p and q such that $p + q = 1$, setting the basis ob equal to 2. Another interesting characteristic of the concentration curve is represented by $a^{II}b^{II}$, where a^{II} is the point in which the concentration curve intersects the segment o^Ib^{IV} perpendicular to oc in its median point o^I. In [28] it was shown that $\overline{a^{II}b^{II}}$ measures the ratio between the simple mean deviation from the median

[6] In each case the *resulting values* of Table 1 are given by simple equations (details are provided in the original paper).
[7] In the original example Gini performed graphical computation (iv) by means of an Amsler planimeter.
[8] This section corresponds to paragraphs 8-9 of the original paper.

^1S_M and twice the arithmetic mean $2A$ and therefore it can be used as concentration index. If the distribution is symmetric, i.e. $A = M$, then the maximum value of the concentration curve is attained at the intersection with $o^I b^{IV}$ [9]. From Figure 3(c) Gini highlights that the simple mean deviation from the median corresponds to values of $p + q$ ranging from $1/2$ to 1 (corresponding to a value between b^I and b^{IV}) and its maximum value is $1/2$ (corresponding to the segment $o^I b^I$). In this case indeed $p = 1/2$ and q can vary in $[0, 1/2]$. The concentration curve would take value o^I in b^I, if all the variable was zero for the fist half of the terms. Furthermore, Gini considers $a^V b^V$, where a^V is the intersection point between the concentration curve and the segment $c^I b^{IV}$, that is symmetric w.r.t. $o^I b^{IV}$. Figure 3(c) shows that $a^V b^V$ reaches its maximum $1/2$ when it coincides with $c^I b^{VI}$ and corresponds to the values of $p + q$ belonging to $[1, 3/2]$. This also follows from $q = cc^I/cb = 1/2$, while p can take all values not smaller than q, i.e. $p \in [1/2, 1]$. In order to have the concentration curve reaching the maximum value in $c^I b^{VI}$, it would be necessary to have $p = 1$ and q simultaneously taking values 1 and $1/2$, which is not possible for an observed discrete distribution, but can only be considered as a limiting case, that is attained in a distribution in which one half of the total amount of the variable belongs to a single individual, as the number of observations tends to infinity. Finally, $a^V b^V$ has the interpretation of a simple mean deviation from the median of the antiseries, that is denoted by $^1S_\mu$, divided by $2A$.

At this point, the Author considers worthwhile a digression to illustrate further details about the antiseries, with special emphasis on the relationship between the series and the antiseries. Let v_s be the value of the $s-$th term of a given data set and let f_s be the associated frequency or weight. Then the antiseries is defined as the new series obtained by replacing each value v_s by its reciprocal $v_a = \frac{1}{v_s}$ and each frequency f_s by the product $f_a = f_s \cdot v_s$. Figure 4(a) simultaneously represents the frequency curves of the series and of the antiseries. Specifically, the points a and a', b and b' are corresponding values of the series and of the antiseries by inversion w.r.t. the disk of center 0 and radius 1, and therefore these values are such that $\overline{oa} \cdot \overline{oa'} = 1$, $\overline{ob} \cdot \overline{ob'} = 1$ and so on.

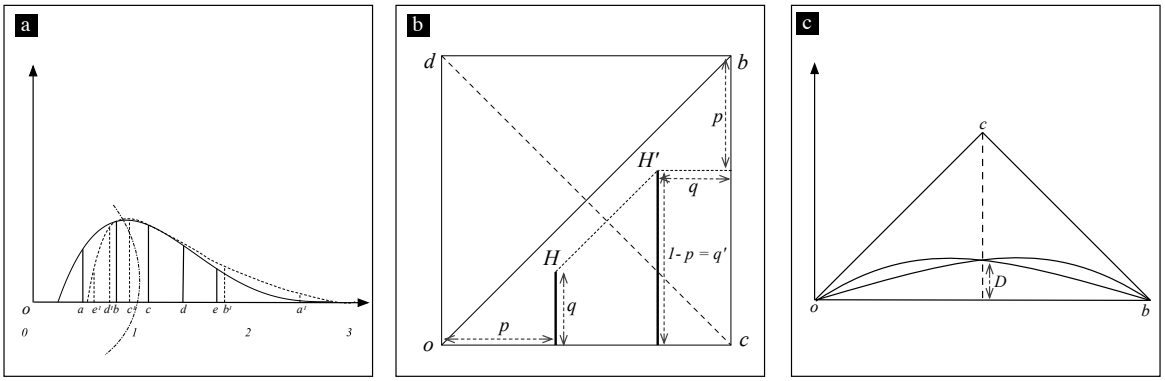

Figure 4: Series and antiseries.

Beyond its formal definition, Gini underlines that in many cases the antiseries has an intuitive interpretation. For instance, given the series of mortality coefficients (i.e. ratios between deaths and inhabitants) for certain geographical areas with weights proportional to the number of people living in that area, the antiseries consists of the ratios between inhabitants and deaths with weights proportional to the number of deaths. One of the most straightforward relationships between the series and the antiseries is the following

$$\frac{\sum f_s v_s}{\sum f_s} = \frac{1}{\frac{\sum f_s}{\sum f_s v_s}} = \frac{1}{\frac{\sum f_a v_a}{\sum f_a}},$$

that is the arithmetic mean of the antiseries is equal to the inverse of the arithmetic mean of the series. Table 2 exemplifies other interesting relationships.

[9] Here, there is probably a typo in the orginal paper, cb instead of $o^I b^{IV}$

	series	antiseries
Values	$l = v_1,\ v_2,\ v_3,\ v_4,\ v_5 = L$	$\frac{1}{L} = \frac{1}{v_5},\ \frac{1}{v_4},\ \frac{1}{v_3},\ \frac{1}{v_2},\ \frac{1}{v_1} = \frac{1}{l}$
Frequences	$f_1,\ f_2,\ f_3,\ f_4,\ f_5$	$f_5v_5,\ f_4v_4,\ f_3v_3,\ f_2v_2,\ f_1v_1$
Arithmetic mean	A	$1/A$
Total number of cases	$f_1 + f_2 + f_3 + f_4 + f_5 = n$	$f_5v_5 + f_4v_4 + f_3v_3 + f_2v_2 + f_1v_1 = nA$
Total amount of variable	nA	n
Incomplete moment of order 0	$p = \frac{f_1+f_2}{n}$	$p' = \frac{f_5v_5+f_4v_4+f_3v_3}{nA} = 1 - q$
Incomplete moment of order 1	$q = \frac{f_1v_1+f_2v_2}{nA}$	$q' = \frac{f_5+f_4+f_3}{A} = 1 - p$

Table 2: Relationships between series and antiseries (when the number of terms is equal to 5)).

By inspecting Figure 4(b) Gini points out that, since the point H of coordinates (p, q) and the point H' of coordinates $(p' = 1 - q, q' = 1 - p)$ are symmetric w.r.t. the line cd of equation $p + q = 1$, then the series and the antiseries are characterized by concentration curves that are symmetric w.r.t. cd (see Figure 4(c)). Consequently series and antiseries have in common:

- the same concentration area;
- the same concentration ratio R;
- the same maximum value for the difference $p-q$, and therefore the same ratio between the mean deviation from the arithmetic mean and $2A$ and also the same value of D (see Figure 4(c)).

Furthermore, with obvious notation, from

$$R' = \frac{\Delta'}{2A'} = \frac{\Delta' A}{2} = R, \qquad R = \frac{\Delta}{2A},$$

it follows that

$$\Delta' = \frac{\Delta}{A^2}, \qquad \Delta = \frac{\Delta'}{(A')^2},$$

and from

$$\frac{{}^1S_A}{2A} = \frac{{}^1S'_A}{2A'} = \frac{{}^1S'_A}{2},$$

it follows that

$${}^1S'_A = \frac{{}^1S_A}{A^2}, \qquad {}^1S_A = \frac{{}^1S'_A}{(A')^2}.$$

Moreover, Gini suggests a concrete interpretation of the point a^V which belongs to the segment $c^I b^{IV}$ in Figure 3(c). Due to the above mentioned symmetry, this point is indeed the intersection point between the concentration curve of the antiseries and the segment $o^I b^{IV}$ and therefore its coordinate on the y-axis is $\frac{{}^1S'_{M'}}{2A'}$.

In summary, the Author shows that the measures of concentration coincide for the series and the antiseries. Since in many practical situations it is possible to choose arbitrarily between a ratio or its inverse, he states that a suitable index of concentration or variability needs to satisfy the requirement of yielding consistent results for the series and the antiseries.

5 The case of variables with upper/lower limits

This Section takes into account variables that are characterized by an upper and/or a lower limit[10]. In his previous contribution [17] Gini showed that, when there is an upper limit L_s for a variable,

[10]This section corresponds to paragraphs 10-15 of the original paper.

a correction coefficient can be introduced in the concentration ratio formula. Specifically, in order to obtain the maximum value of the mean difference, he suggests to multiply $2A$ by $\frac{L_s - A_s}{L_s}$. Hence, the maximum concentration area is represented by the triangle *odb* (see Figure 5(a)), instead of *obc*, where $\frac{dc}{oc}$ is given by the ratio between the number of occurrences of L_s and the total number of observations and, therefore, it is $\frac{dc}{oc} = \frac{A_s}{L_s}$ and $\frac{odb}{ocb} = \frac{L_s - A_s}{L_s}$. Similarly, if there is a lower limit l_s for the value of the variable, the maximum concentration area is given by the triangle *oeb* in Figure 5(b), where *ce* corresponds to the proportion of the variable amount belonging to all the terms of the series but the last one, when each term has the value l_s and the last one has the residual amount; hence, $\frac{ce}{cb} = \frac{nl_s}{nA_s} = \frac{l_s}{A_s}$.

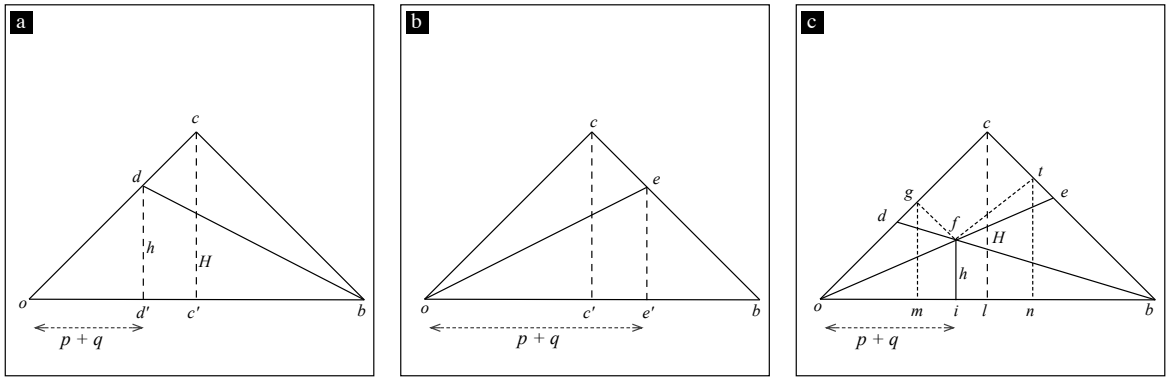

Figure 5: Maximum concentration area when the series has a) an upper limit, b) a lower limit, c) both.

Now, Gini remarks that an upper limit for the series corresponds to a lower limit for the antiseries (see again Table 2), since

$$L_s = \frac{1}{l_a}, \quad A_s = \frac{1}{A_a},$$

where l_a and A_a denote the lower limit and the arithmetic mean of the antiseries respectively, and therefore

$$\frac{L_s - A_s}{L_s} = \frac{A_a - l_a}{A_a}.$$

However, being the concentration ratio the same for the series and for the antiseries, the correction coefficient must be the same as well. Hence, the correction coefficient for the antiseries is $\frac{A_a - l_a}{A_a}$ when there is a lower limit l_a. This coefficient can be directly deduced by examining Figure 5(a) since the triangle *odb* which represents the maximum concentration area for the series with upper limit L_s can be regarded as the maximum concentration area for the antiserie with lower limit $l_a = \frac{1}{L_s}$, provided that the proportion of the total amount of the variable are represented on *co* and the proportions of the total number of terms are represented on *cb* (as in Figure 5(b)). Hence the following relationships hold:

$$\frac{dc}{oc} = \frac{A_s n l_a}{A_s n A_a} = \frac{l_a}{A_a}$$

and

$$\frac{dob}{cob} = \frac{l_a}{A_a} = \frac{A_a - l_a}{A_a}.$$

Moreover, it can be shown that $\frac{A_s - l_s}{A_s}$ is the correction coefficient to obtain the maximum mean difference by multiplying twice the arithmetic mean of the series, when there is a lower limit l_s; $\frac{L_a - A_a}{L_a}$ is the corresponding coefficient for the antiseries when there is an upper limit L_a. Finally, when l_a is the lower limit, the maximum mean difference is $2A_s \frac{A_s - l_s}{A_s} = 2(A_s - l_s)$, which can alternatively be obtained by subtracting l_s from all the values.

Here Gini takes also into account the case of a variable having both an upper and a lower limit. This yields a reduction in the maximum concentration area from *obc* to $ofb = obc - oce - efb$ as shown in Figure 5(c). Being the area of *obc* equal 1, it follows $oce = \frac{ce}{bc}$, $efb = \frac{eb}{bc}\frac{gc}{co}$, where *gc*

denotes the height of the triangle efb with respect to the base eb. As already shown, $\frac{ce}{bc} = \frac{l_s}{A_s}$ and $\frac{eb}{bc} = \frac{A_s - l_s}{A_s}$. Moreover, $\frac{gc}{oc}$ represents the proportion of cases necessary to distribute the remaining amount $A_s n - l_s n$ under the maximum concentration hypothesis. This proportion is $\frac{A_s - l_s}{L_s - l_s}$. Hence the area of ofb becomes

$$\text{Area}_{ofb} = 1 - \frac{l_s}{A_s} - \frac{A_s - l_s}{A_s}\frac{A_s - l_s}{L_s - l_s} = \frac{A_s - l_s}{A_s}\frac{L_s - A_s}{L_s - l_s}, \tag{1}$$

which is the correction coefficient to obtain the maximum mean difference, when there are both a lower limit l_s and an upper limit L_s. Moreover, Gini highlights that the coefficients related to the previously considered situations are obtained as special cases of Equation (1), letting respectively $l_s = 0$ (no lower limit) and $L_s = \infty$ (no upper limit). Finally, for $l_s = 0$ and $L_s = \infty$, the correction coefficient becomes 1. A similar relationship also holds for the antiserie.

In summary, in the presence of a lower limit l_s and an upper limit L_s, the maximum mean difference becomes

$$2(A_s - l_s)\frac{L_s - A_s}{L_s - l_s} \tag{2}$$

in place of $2A_s$, which can also be retrieved starting from the formula $2A_s\frac{L_s - A_s}{L_s}$ and subtracting l_s from the mean and from the upper limit.

Starting from Equation (2), Gini addresses an additional issue, that is finding the correction coefficient for the maximum mean difference of the complementary variable; the concentration of two complementary variables, indeed, has to coincide for the index to be considered satisfactory.

If the complementary variable is measured starting from L_s, the following relationships are in order: $A_c = L_s - A_s$, $L_c = L_s - l_s$ and $l_c = L_s - L_s = 0$, where A_c, L_c and l_c obviously denote the average, the upper limit and the lower limit of the complementary variable. Thanks to the above results, Equation (2) becomes

$$2\frac{(L_c - A_c)A_c}{L_c}, \tag{3}$$

that is the correction coefficient to be used for the arithmetic mean of the complementary variable, provided that its upper limit is L_c.

Conversely, if the complementary variable is measured starting from a value $K > L$, then a lower limit is also allowed. Therefore, it is $A_c = K - A_s$, $L_c = K - l_s$, $L_c = K - L_s$, and from Equation (2) Gini obtains an analogous correction coefficient for the complementary variable, i.e.

$$2\frac{(L_c - A_c)(A_c - l_c)}{L_c - l_c}. \tag{4}$$

Bearing in mind the characteristics of the concentration curves described in Section 4, Gini explaines that when the series has an upper limit and/or a lower limit, the maximum value of the difference $p - q$ does not correspond anymore to $p + q = 1$. Specifically, in the former case (see Figure 5(a)) this maximum value can be attained only for a value of $p + q$ smaller than 1, that is equal to the proportion of cases $\frac{od'}{oc'} = \frac{od}{oc} = \frac{L - A}{L}$ in which the value 0 can be assumed, and the maximum value is $L = \frac{L - A}{L}$. In the latter case (see Figure 5(b)) the maximum is a value of $p + q$ that exceeds 1 of a fraction $\frac{c'e'}{c'b'} = \frac{ce}{cb} = \frac{l}{A}$, which is due to the fact that all the cases have at least the minimum value l, and the maximum value is $L = \frac{A - l}{A}$. Finally, when there are an upper limit L and a lower limit l, the maximum value of $p - q$ can be reached in correspondence of $p + q = 1 - \frac{A - l}{L - l} + \frac{L - A}{L - l}\frac{l}{A} = \frac{L - A}{L - l}\frac{A + l}{A}$ which can be larger than, equal to or smaller than unit; the maximum value is $L = \frac{L - A}{L - l}\frac{A - l}{A}$.

Moreover, in comparing different concentration curves, Gini recommends to reduce to a constant quantity the maximum areas that the concentration curves may contain, namely the areas of the triangles obc, odb, oeb, ofb according to the different cases (no limits, upper limit only, lower limit only, both an upper and a lower limit, respectively).

The concentration curves satisfying this requirement are referred to as *reduced* concentration curves, in which, as a consequence, the value $p + q = 1$ will always correspond to the maximum possible difference $p - q$. The Author motivates this way of proceeding by several considerations.

First of all, to make different curves comparable, it is necessary that they live into the same area and reach the same limits. However, when the values of a variable have a lower limit, it is also possible to subtract the lower limit from each value and to compute the concentration based on these reduced values without restricting the maximum of the concentration area. As already remarked, the concentration ratio is the same in the two cases; moreover, it is convenient that the concentration curves coincide as well. Furthermore, the lower limit of a variable corresponds to the upper limit of the inverse and of the complementary variables: then, it is advisable to find a procedure which avoids the possibility of obtaining different concentration curves w.r.t. the original variable.

At this point, Gini provides an example to show how the reduced concentration curves can be obtained in the following three main cases:

(a): the variable v has both an upper limit L and a lower limit l;

(b): the variable v has a lower limit only;

(c): the variable v has an upper limit only.

Example 2. Let us consider the following distribution

v	f_v
$15 \vdash 25$	8
$25 \vdash 35$	11
$35 \vdash 45$	12
$45 \vdash 55$	30
$55 \vdash 65$	35
$65 \vdash 75$	16
$75 \vdash 85$	14
$85 \vdash 95$	6

where $n = 132$, $A = 55.68$.

(a) First of all let us assume that $l = 15$, $L = 95$. Under the assumption of uniform distribution within each class, the piecewise linear concentration curve is represented by the solid line in Figure 6(a). The maximum concentration case corresponds to the triangle obf, where f has coordinates $(p = 0.492; q = 0.133)$ determined for a number of terms ξ such that $15\xi + 95(132 - \xi) = 132A$ (maximum concentration distribution). By subtracting the constant $l = 15$ from the terms of the series, a new series is obtained with $l' = 0$ and with the same concentration ratio. The new curve, which passes by M' (instead of M), is represented by the dashed line in Figure 6(a). The related maximum concentration triangle is then ogb, where g has coordinates $(p = 0.507; q = 0)$. The vertices f and g substantially belong to the same parallel line of cb (this exactly happens when the starting distribution is continuous). This example shows how the general case (both upper limit and lower limit) reduces to the case in which only the upper limit exists.

(b) Let us start by the same distribution of the previous example and let us consider the lower limit only, i.e. $l = 15$ and $L = \infty$. The corresponding concentration curve is the same as before, but in Figure 6(b) the maximum concentration triangle is ofb, where f has coordinates $(p = 1; q = 0.234)$ (determined by assuming that the variable takes the minimum value 15 in all the 132 terms and therefore a global value 1980 which corresponds to a proportion of 0.234 out of the total amount). By subtracting the constant $l = 15$ from the terms of the series, a new series is obtained without limits, but with the same concentration ratio. The corresponding concentration curve (dashed line) is related to the whole triangle ocb as maximum concentration triangle. This example shows how the case in which there is a lower limit reduces to the unrestricted case.

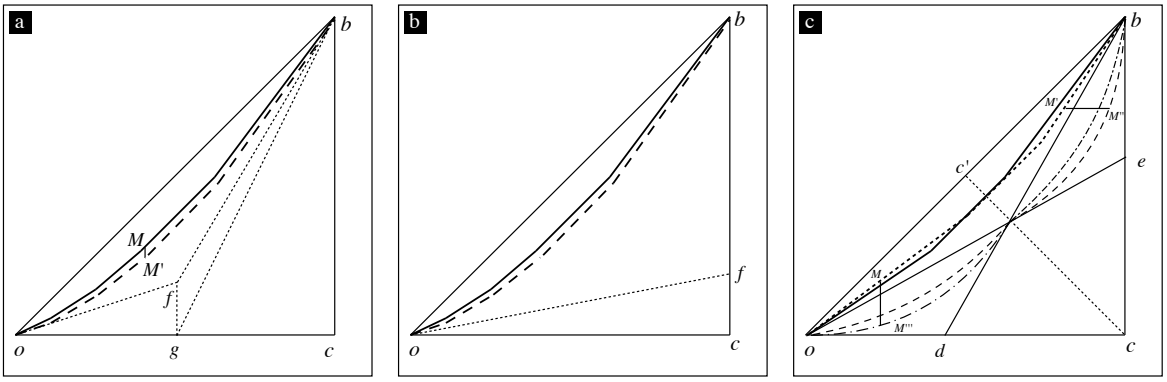

Figure 6: Illustration of the three cases of *Example 2*.

(c) Finally, let us assume $l = 0$ and $L = 95$ The corresponding concentration curve (solid line) is the same as before, but in Figure 6(c) it is referred to the maximum concentration triangle *odb*, in which the vertex d has coordinates $(p = 0.411; q = 0)$, obtained by assuming that L is assigned to the largest number of terms (indeed, $\frac{dc}{oc} = \frac{a}{l} = 0.589 = 1 - 0.411$). In order to derive the corresponding reduced concentration curve, one can resort to the antiseries. In the antiseries the values of the variable range from $l' = \frac{1}{L} = \frac{1}{95}$ to $L' = \infty$. As in case *(b)*, it is possible to derive a new auxiliary series such that $l'' = \frac{1}{95} - \frac{1}{95} = 0$ and $L'' = \infty$. Finally, the antiseries of this latter auxiliary series, which is such that $l''' = 0$ and $L''' = \infty$, yields the desired concentration curve.

In Section 3 Gini discusses the interpolation of the concentration curve by considering equidistant points on the equidistribution line. In the present section, instead, the focus is on the relevant points of the curve highlighted in Section 4, corresponding to the simple mean deviation from the arithmetic mean ${}^1S_A$, the difference $p - q$ on the maximum value D, the simple mean deviation from the median of the series ${}^1S_M$, the simple mean deviation from the median of the antiseries ${}^1S_\mu$. In order to find the concentration curve it is necessary to derive the corresponding values of p and q, as it is shown in the following example, which is based again on the data considered in *Example 1* of Section 3. Two additional examples of the original paper are omitted here for the sake of brevity.

Example 1 (cont.). Given $n = 100000$ individuals with income 1, 2, ..., 100000 respectively, we obtain $A = M = 50000.5$. It is then straightforward to find

$$ {}^1S_A = {}^1S_M = \frac{2(0.5 + 1.5 + \ldots + 49999.5)}{100000} = 25000 $$

and, therefore,

$$ \frac{{}^1S_A}{2A} = \frac{{}^1S_M}{2A} = 0.250. $$

In order to find $\frac{{}^1S_\mu}{2A'}$, it is necessary first of all to compute the median and the average of the antiseries, denoted by μ and A' respectively. The antiseries is given by the values $\frac{1}{100000}, \ldots, \frac{1}{3}, \frac{1}{2}, 1$, with associated frequencies $100000, \ldots, 3, 2, 1$. Then, the arithmetic mean of the antiseries is

$$ A' = \frac{1}{A} = \frac{1}{50000.5}. $$

As regards the median μ, since the number of terms is $n' = An = 5000050000$, the median term will correspond to the frequency z such that $1 + 2 + \ldots + z = 2500025000$, that is $z = 70710.53$ which is the positive root of the equation $z^2 + z - 5000050000$. Finally, $\mu = \frac{1}{70710.5}$. In order to find the simple mean deviation from μ, let us consider the sum of the deviations of the values smaller than μ, i.e.

$$ T_1 = \sum_{n=70711}^{100000} n \left(\mu - \frac{1}{n} \right) = 6066 $$

and the sum of the deviations of the values larger than μ, i.e.

$$T_2 = \sum_{n=1}^{70710} n\left(\frac{1}{n} - \mu\right) = 35355.$$

Hence,

$$'S_\mu = \frac{T_1 + T_2}{n'} = \frac{41421}{5000050000}$$

and therefore

$$\frac{'S_\mu}{2A'} = \frac{41421}{5000050000} \cdot \frac{50000.5}{2} = 0.207.$$

Finally, D is found as the value $p_t - q_t$ such that $p_t + q_t = 1$. By definition it is

$$p_t = \frac{t}{100000} \quad \text{and} \quad q_t = \frac{1 + 2 + \ldots + t}{1 + 2 + \ldots + 100000} = \frac{t^2 + t}{10000100000},$$

then it is straightforward to obtain $t = 61803.29$ and, consequently, $p_t = 0.618$, $q_t = 0.382$ and $D = 0.236$. In summary it is:

$$\frac{^1S_M}{2A} = 0.250;\ \frac{^1S_A}{2A} = 0.250;\ D = 0.236;\ \frac{^1S_\mu}{2A'} = 0.207,$$

that correspond to the values $p_1 + q_1, p_2 + q_2, p_3 + q_3$ to be determined as follows. Since $p_1 = 0.5$ and $p_1 - q_1 = 0.25$, it is $q_1 = 0.25$ and therefore $p_1 + q_1 = 0.75$. Moreover, due to the symmetry of the frequency curve, it is also $p_2 + q_2 = p_1 + q_1 = 0.75$. In order to find $p_3 + q_3$ it is convenient to refer to the antiseries, for which

$$p_1' = \frac{s}{n'} = \frac{s}{5000050000 = 0.5}.$$

This yields $s = 2500025000$ that belongs to the frequency class $z = 70710.53$, as derived before. Hence the corresponding value of q_s' is

$$q_s' = \frac{100000\frac{1}{100000} + 99000\frac{1}{99000} + \ldots + 70710\frac{1}{70710}}{100000} = 0.293$$

and therefore $p_1' + q_s' = 0.5 + 0.293 = 0.793$ and, back to the series,

$$p_3 + q_3 = 2 - 0.793 = 1.207,$$

which could also be obtained by adding 1 to $\frac{^1S_\mu}{2A'}$.

Finally, recalling that $x = p+q$ and $y = p-q$, the interpolation is based on the following five points

$$(x_1, y_1) = (0, 0);\ (x_2, y_2) = (0.75, 0.25);\ (x_3, y_3) = (1, 0.236);\ (x_4, y_4) = (1.207, 0.207);\ (x_5, y_5) = (2, 0).$$

By using the Lagrange method, it follows that

$$y = -0.0727x^4 + 0.4009x^3 - 0.9287x^2 + 0.8376x$$

and therefore

$$\int_0^2 y dx = 0.337.$$

Hence the resulting value 0.337 for the concentration ratio slightly exceeds the actual value 0.333.

With respect to the reduced curves introduced before, Gini illustrates several shapes of the concentration curves and introduces the following definitions. A concentration curve which reaches its maximum on the median axis, which means ${}^1S_A = D$, is called *culminating*. This type of curve can be either *symmetric* (see curve 1 in Figure 7(a)) or *asymmetric* (see curve 2 in Figure 7(a)). Asymmetric curves can also be *non culminating* and in this case, they can be distinguished in *right-asymmetric* (see curves 3 and 5 in Figure 7(a)) or *left-asymmetric* (see curves 4 and 6 in Figure 7(a)) if the value on the x-axis corresponding to ${}^1S_A$ is larger or smaller than that corresponding to D, respectively. Among the right-asymmetric curves a further distinction is in order: if the value on the x-axis corresponding to ${}^1S_A$ is smaller than that corresponding to ${}^1S_\mu$ the curve is called sub-right-asymmetric (curve 3), otherwise super-right-asymmetric (curve 5). Similarly, a sub-left-asymmetric curve (curve 4) is characterized by a value on the x-axis corresponding to ${}^1S_A$ smaller than ${}^1S_M$, whereas a super-left-asymmetric curve (curve 6) has a value on the x-axis corresponding to ${}^1S_A$ larger than ${}^1S_M$. The concept of symmetry, here defined for the concentration curves can be also applied to the other distribution curves: in frequency curves symmetry is considered w.r.t. the median value on the y-axis, whereas in graduation curves symmetry is considered w.r.t. the median point of the curve itself. However, a symmetric concentration curve corresponds to asymmetric graduation curves and viceversa. Indeed, the symmetry condition for frequency and graduation curves is $A = M$ and therefore ${}^1S_A = {}^1S_M$: this implies instead a left-asymmetric concentration curve, since the maximum value is reached at the intersection point between the concentration curve and the segment da.

Another interesting remark is that, for a given series and its associated antiseries with different graduation and frequency curves (see Figure 4(a)), the concentration curve is the same although it may be differently oriented. When the series and the antiseries have the same frequency and graduation curves (with different orientation), the concentration curve is symmetric. This happens when, given $p = x$ and $q = y$, it is $p = 1 - y$ and $q = 1 - x$, and therefore $p - q = x - y$ both for $p + q = x + y$ and for $p + q = 2 - x - y$: in other words, the values on the y-axis of the concentration curves are the same in two points of the equidistribution line that are symmetric w.r.t. the median point.

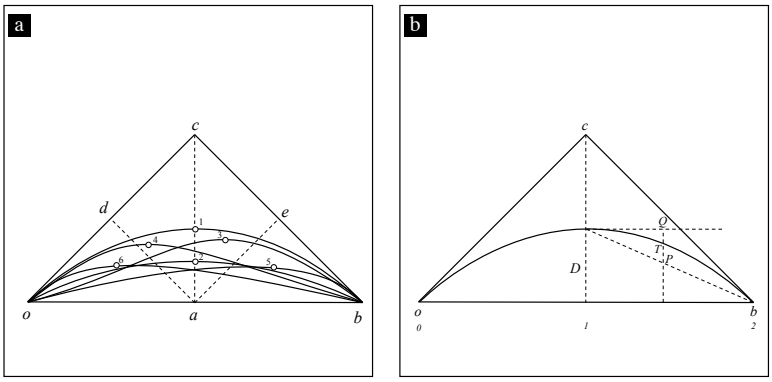

Figure 7: Shapes of the concentration curves.

6 Some relevant cases of concentration curves

At the time when Gini is writing his paper, he feels there is still space for further research and extensive study on the shape of the concentration curve, that can yield interesting results[11]. Apart from the opposite situations of maximum equality and maximum disequality, the other analytical expressions obtained for the concentration curves are those associated to some theoretical distributions (see [30], [19], [12] for linear distributions and [19] for exponential distributions), the curve related to global incomes, which also holds for labor incomes and rents (see [16] and [19]), and the one related to the number of children (see [16]). However, even in the aforementioned cases the main focus was on the measure of concentration level, rather than on the shape of the concentration curve. As regards

[11] This section corresponds to paragraphs 16-21 of the original paper.

the relationship between the concentration curve and the distribution curve, previous studies were restricted to the case of global incomes. Finally, for several distributions, such as for instance the maximum inequality distribution, the uniform distribution and other distributions (e.g. binomial, exponential, hyperbolic), the only available results were the values of the mean difference and of the concentration ratio.

Hence - Gini believes - a lot of aspects still deserve to be analyzed in further details. First of all, the concentration curves associated to the most important phenomena have to be studied, together with their features and their analytical representation. It is worth noticing that concentration curves are often simpler than frequency curves, especially for several economic and financial phenomena.

Another interesting point is to find the equations of the frequency curve and of the graduation curve starting from the equation of the concentration curve, and viceversa, which can be obtained through integration or differentiation respectively, as discussed in Section 2. When dealing with curves associated to theoretical distributions, the research has merely a theoretical dimension; but when the equations of interest apply to the description of certain concrete variables or phenomena, then the research acquires a practical dimension: the comparison between the observed data and those resulting from the different equations of the distribution curves allows to assess which equation ensures the best fit to the observed data. It is well known that, if different curves are linked by theoretical relationships, they do not represent the observed data equally well; this is due to two main reasons: on the one hand the theoretical relationships rely on assumptions not exactly correspondent to reality, on the other hand the approximations of one formula have a consequent impact on the other ones. Finally, once the analytical representations are determined, it is straightforward to find the value of the mean difference, of the concentration ratio and of other variability indices.

In order to obtain a suitable analytical representation of the concentration curves associated to concrete distributions, Gini describes some interpolating procedures to be used in the different cases considered below. Again some examples are omitted here for the sake of brevity.

First of all, he introduces further qualitative distinctions among the curves. Given the usual reference system $x = p + q$, $y = p - q$, a concentration curve is *asymmetric reducible* when it is possible to determine a function $\phi(x)$ such that the curve $y = f[\phi(x)]$ is symmetric, that is

$$f[\phi\{c + (x - c)\}] = f[\phi\{c - (x - c)\}],$$

where c is the value on the x-axis corresponding to the culminating point of the curve. Hence, a first type of asymmetric reducible curves will be that satisfying the following condition

$$f[c + (x - c)] = f[\omega\{c - (x - c)\}]$$

for a convenient constant ω; these curves are called reducible *for proportionality*. Conversely, curves such that

$$f[\log\{c + (x - c)\}] = f[\log\{\omega c - \omega(x - c)\}];$$

are called reducible *for logarithmic proportionality*.

(a) The interpolation of a *symmetric culminating* curve can be obtained through the following equation:

$$y = D - D|x - 1|^k, \tag{5}$$

where D is the maximum value on the y-axis (in this case the median value on the y-axis) and k is a parameter to be conveniently determined. Equation (5) clearly represents a curve that is symmetric w.r.t. the line $x = 1$, which reaches the value D for $x = 1$ and passes through the extremes of the concentration curve, i.e. $(0, 0)$ and $(2, 0)$. The subtracting term $D|x-1|^k$ represents the segment that is parallel to the y-axis, delimited by the point x and the tangent line to the curve itself passing by the culminating point. Moreover, since

$$\begin{aligned} y' &= -kD(x-1)^{k-1} \quad x > 1 \\ y' &= +kD(1-x)^{k-1} \quad x < 1, \end{aligned}$$

and

$$\begin{aligned} y'' &= -k(k-1)D(x-1)^{k-2} \quad x > 1 \\ y'' &= -k(k-1)D(1-x)^{k-2} \quad x < 1, \end{aligned}$$

the curve turns out to be concave with respect to the x-axis wherever $k(k-1) > 0$, that is $k > 1$. However, notice that for $x = 0$, $y'_{x=0} = kD$ and for $x = 2$, $y'_{x=2} = -kD$. Now, if $kD \leq 1$, the curve defined in Equation (5) is actually included in the maximum concentration triangle, otherwise the curve exceeds the triangle in the extreme parts. Hence, in order to let the curve (5) meet the general requirements of a concentration curve, it must be $1 < k \leq \frac{1}{D}$. Nonetheless, it may be that even for some values of k above $\frac{1}{D}$ the curve provides, from a practical point of view, a good representation of the concentration curve. As regards the actual determination of k, Cauchy method on the logarithm of $D|x-1|^k$ can be applied, based on some values measured on the one side or the other of the curve w.r.t. the maximum value.

(b) The interpolation of an *asymmetric culminating* curve can be obtained by adopting the same criteria described in paragraph *(a)*, but making a distinction between the left side and the right side of the curve w.r.t. the maximum value (median) and consequently by determining a suitable value of the parameters k or h respectively. The equation of the interpolating curve is therefore

$$y = D - \begin{cases} D(1-x)^k & 0 \leq x < 1 \\ D(x-1)^h & 1 \leq x \leq 2 \end{cases}, \tag{6}$$

Similar remarks about the behaviour of the interpolating curve in the extremes $x = 0$ and $x = 2$ also hold in this case.

(c) For a *non culminating concentration curve reducible for proportionality*, by denoting the maximum y-value with S and the corresponding x-value with M, an interpolating curve is given by

$$y = S - \begin{cases} S\left(\frac{M-x}{M}\right)^k & 0 \leq x < M \\ S\left(\frac{x-M}{2-M}\right)^k & M \leq x \leq 2 \end{cases}, \tag{7}$$

where the denominators M and $2 - M$ respectively imply the reduction to unity of the intervals determined by the culminating point. This reduction makes the curve symmetric by construction and therefore a common value of the parameter k can be found below and above M. As in case *(a)* it is straightforward to check that this curve is concave with respect to the x-axis for $k > 1$. Then, if we consider the derivative in $x = 0$, we have:

$$y'_{x=0} = \frac{Sk}{M}\left(\frac{M-x}{M}\right)^{k-1}|_{x=0} = \frac{Sk}{M}.$$

In order to have this extreme inside the maximum concentration triangle it must be $k \leq \frac{M}{S}$. Finally, k can be determined as explained in the previous paragraph *(b)*.

(d) A *non culminating concentration curve non reducible for proportionality* can be represented by an equation analogous to (7), but with a distinct parameter for each side of the curve, i.e.

$$y = S - \begin{cases} S\left(\frac{M-x}{M}\right)^k & 0 \leq x < M \\ S\left(\frac{x-M}{2-M}\right)^h & M \leq x \leq 2 \end{cases}. \tag{8}$$

(e) An alternative way of interpolating a culminating symmetric concentration curve may be as follows. Let us consider a generic $y = D - \overline{PQ} + \overline{PT}$, as in Figure 7(b). Now, we have $\overline{PQ} = D|x-1|$; as regards $\overline{PT}$, it may be considered as a function of x which equals 0 for $x = 0$, $x = 1$, $x = 2$ and is positive and concave w.r.t. each interval $[0, 1]$ and $[1, 2]$. Such a function can be obtained, for suitable choices of k and h in $(0, 1)$, as the product of the two functions $D|x-1|^k$ and $(1-D)|1-|x-1||^h$. Hence, the equation of the interpolating curve is as follows:

$$y = D - D|x-1| + D|x-1|^k(1-D)|1-|x-1||^h, \tag{9}$$

where k and h are determined with a procedure similar to the one described before. Notice that the third part of Equation (9) equals zero both for $D = 1$ (maximum concentration) and for $D = 0$ (null concentration), as required to let the curve coincide with one of the sides of the triangle.

(f) If the curve is culminating asymmetric, the procedure described in *(e)* needs to be applied for each part of the curve with respect to the culminating point, thus obtaining two couples of parameters k and h.

(g) By considering one of the two part of the interpolating curve, for instance, when $0 \leq x \leq 1$, and by taking the derivative one obtains

$$y' = D + D(1-D)(1-x)^{k-1}x^{h-1}\left\{-kx + h(1-x)\right\}.$$

Now, since $k-1 < 0$ and $h-1 < 0$, it is also

$$\lim_{x\to 0} y' = \lim_{x\to 1} y' = \infty.$$

This means that the part of the curve under consideration, although concave w.r.t. the x-axis, is characterized by a tangent line in the extremes that is parallel w.r.t. the y-axis, and therefore, it obviously exceeds the maximum concentration triangle. However, when the goal is to determine the concentration area, this interpolation can be still satisfying.

The following interpolation procedure is then described in details. Let us inscribe in a given concentration curve the triangle oP_1b with maximum height w.r.t. the base ob; moreover let us construct inside the curve on each of the segments oP_1 and P_1b the triangles oP_1P_2 and P_1bP_2'; and let us proceed in this way iteratively. Let us draw the tangent lines to the curve passing by the resulting vertices of these triangles $P_1, P_2, P_2', \ldots$, in such a way that new triangles arise containing the previous ones (like oP_1C_1 which contains oP_1P_2 in Figure 8(c)). In summary each triangle is contained by another one, but the triangle obP_1 is contained by the starting triangle obc, whose sides are not tangent w.r.t. the concentration curve.

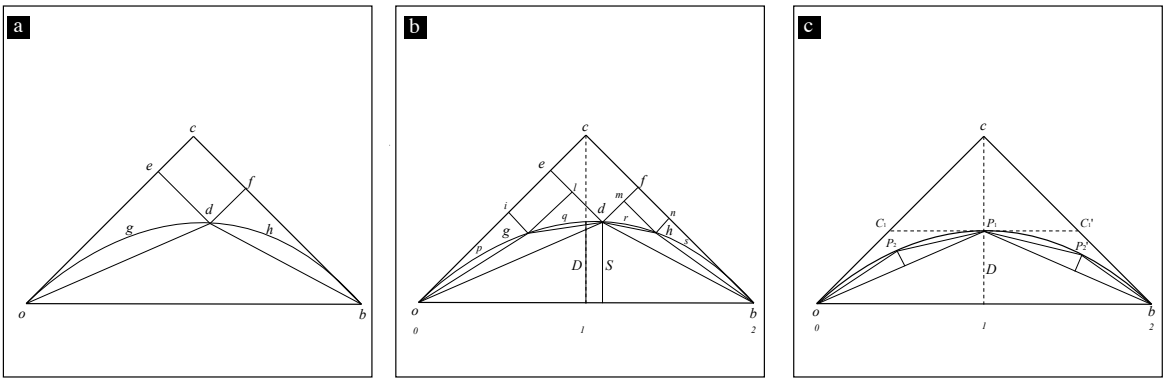

Figure 8: Illustration of the interpolation procedure for the concentration curve.

The situation in which the following proportionality relation

$$obP_1 : obc = oP_1P_2 : oP_1C_1 = P_1bP_2' : P_1bC_1' = \ldots$$

between subsequent couples of triangles holds, is defined *equitension*. Under this assumption, it is possible to proceed analytically to compute the area under the concentration curve, i.e. the concentration ratio R, by quadrature. Specifically, it is

$$R = \Delta_0 + (\Delta_1 + \Delta_1') + (\Delta_2 + \Delta_2' + \Delta_2'' + \Delta_2''') + \ldots,$$

where Δ_0 denotes the area of the first inscribed triangle, Δ_1 and Δ_1' the areas of the second iteration triangles, and so on with obvious notation.

(a) In the above setup, if the curve is culminating, it is $\text{Area}_{obc} = 1$ and $\Delta_0 = D$, so that the ratio between each triangle and its predecessor is given by D. Moreover the sum of the areas of the two triangles built at the second iteration, is given by:

$$oC_1P_1 + bP_1C_1' = D(1-D),$$

so that, by definition, it is

$$(\Delta_1 + \Delta_1') : D(1-D) = D \Rightarrow (\Delta_1 + \Delta_1') = DD(1-D)$$

and

$$(\Delta_1 + \Delta_1') : \Delta_0 = D(1-D).$$

Similarly, it is

$$(\Delta_2 + \Delta_2' + \Delta_2'' + \Delta_2''') : (\Delta_1 + \Delta_1') = D(1-D)$$

and so on. Consequently, the following formula expresses the concentration ratio in terms of the maximum y-value of the concentration curve D,

$$R = D + D^2(1-D) + D^3(1-D)^2 + \ldots = \frac{D}{1-D(1-D)}, \tag{10}$$

since $D(1-D) < 1$. Note that R reaches its maximum value only if $D = 1$. Furthermore, if the approximation is restricted to the $m-$th term, Equation (10) becomes

$$R = \frac{D - D^{m+1}(1-D)^m}{1-D(1-D)}$$

with absolute error $\frac{D^{m+1}(1-D)^m}{1-D(1-D)}$ and relative error $100 \cdot D^m(1-D)^m\%$.

(b) When the concentration curve is not culminating, similar steps can be followed as in the previous case, but replacing D by ${}^1S_A/2A$.

Formula (10) can be employed to determine an upper bound of the difference between R and D, that is

$$R - D = \frac{D^2(1-D)}{1-D(1-D)}$$

which achieves its maximum $4/3$ when $D(1-D)$ is maximum, that is for $D(1-D) = 1/4$. Hence, it is

$$R - D \leq \frac{4}{3}D^2(1-D).$$

Notice that similar considerations apply to case *(b)*.

In [15] Gini studies the impact on the concentration ratio of replacing a given distribution by a distribution in which the variable only takes values above a certain lower bound. This case is interesting both from a theoretical and a practical point of view, since it often happens that only a truncated distribution can be observed in place of the whole distribution. Specifically, he shows that if the value x is associated to a density Vx^{-h}, i.e. the number of occurrences of a value of the variable between x and $x + dx$ is given by $Vx^{-1}dx$, under some additional assumptions, then the concentration ratio for the whole distribution is approximately equal to the truncated distribution. A relevant example is that of the distribution of global incomes.

Moreover, Gini adds here some useful remarks: since the series and the antiseries have the same concentration curve, it is reasonable that if a series yields a concentration ratio equal to a partial series, obtained by truncation w.r.t. a lower bound, then the concentration ratio related to the corresponding antiseries will equal that of a partial antiseries, obtained by truncation w.r.t. an upper bound. Starting from these considerations, he investigates on the following point. He considers a whole distribution split in two parts by a suitable value that plays the role of a maximum value for the one part and a minimum value for the other one. Then the concentration ratio $R = \frac{c}{m}$, where c is the concentration area and m the area of the triangle corresponding to the maximum concentration case, can be expressed by the following formula

$$\frac{c_1 + c_2}{m_1 + m_2},$$

where c_1, m_1 and c_2, m_2 represent the concentration area and the maximum concentration area respectively for the two partial distributions. In other terms, given the two concentration ratios

$$R_1 = \frac{c_1}{m_1} \text{ and } R_2 = \frac{c_2}{m_2},$$

that are represented in Figure 8(a) by the ratios between the area of ogd w.r.t. the area of oed and the area of dhb and the area of dfb, we want to show whether the overall concentration ratio can be given by

$$R = \frac{\text{Area}_{ogp} + \text{Area}_{dhb}}{\text{Area}_{oed} + \text{Area}_{dfb}}.$$

In principle, this possibility cannot be excluded, at least for a suitable choice of d. Hence, Gini imagines to iterate the same construction of Figure 8(a) for each part of the distribution, and over and over again (see Figure 8(b)). Hence, for each iteration, he obtains a series of concentration areas denoted by $obd, odg, dbh; ogp, gdq, dhr, hbs; \ldots$, and the corresponding maximum concentration areas are: $obc; ode, dbf; ogi, gdl, dhm, hbn; \ldots$. Hence, he wants to investigate whether

$$obd \;:\; (odg + dbh) : (ogp + gdq + dhr + hbs) = \ldots = k \tag{11}$$

$$obc \;:\; (ode + dbf) : (ogc + gde + dhm + hbn) = \ldots = k \tag{12}$$

and therefore

$$\frac{obd}{obc} = \frac{odg + dbh}{ode + dbf} = \frac{ogp + gdq + dhr + hbs}{ogc + gde + dhm + hbn} = \ldots = R. \tag{13}$$

Equation (13) means that under the assumptions expressed by Equations (11) and (12) the sum of the concentration areas of a given step of the splitting procedure has constant ratio w.r.t. the corresponding sum of the maximum concentration areas. If Equation (13) does not hold, then Equations (11) and (12) cannot hold as well; but Equation (13) is not a sufficient condition for Equations (11) and (12). The assumptions (11) and (12) can be referred to as *uniform concentration* or *equiconcentration* hypotheses. At his point, Gini comments on the following intuitive considerations to show that these assumptions cannot hold. If one considers a polygonal line instead of a curve and adopts the same procedure described before, the culmination points are the vertices and therefore the concentration area is zero, but the corresponding sum of the maximum concentration area is not. Hence, the first ratio of (13) is different from 0, whereas the other is null. This shows that for a polygonal curve equiconcentration cannot hold. But this is also true in general, because the successive splitting of the curve makes the archs coincide with the chords and therefore the corresponding concentration areas tend to 0 faster than the respective maximum concentration areas. In practice, Condition (13) would imply that the single partial archs of the starting concentration curve maintain the same level of convexity with respect to their chords as the level of convexity of the total curve with respect to its chord, which is absurd for usual curves. However, the equiconcentration hypothesis, although unrealistic, can be used in order to determine an upper limit for the concentration ratio R. Let us consider a non culminating concentration curve as in Figure 8(b) and let $1 + \epsilon$ be the x-value of point d. It is straightforward to check that the area of the rectangle $edfc$ is

$$\frac{(S-1)^2 - \epsilon^2}{2},$$

while the area of the triangles odb and ocb are respectively S and 1, so that the total area of the two triangles ode and dbf is given by

$$1 - S - \frac{(S-1)^2 - \epsilon^2}{2} = \frac{1 - S^2 + \epsilon^2}{2}.$$

From the first part of Equation (13) it follows that

$$\frac{obd}{1} = \frac{odg + dbh}{\frac{1-S^2+\epsilon^2}{2}},$$

that is

$$\frac{odg + dbh}{odb} = \frac{1 - S^2 + \epsilon^2}{2}$$

and therefore, from Equation (11)

$$\frac{odg + dbh}{obd} = \frac{ogp + gdq + dhr + hbs}{odg + dbh} = \ldots = \frac{1 - S^2 + \epsilon^2}{2}. \tag{14}$$

Now, by replacing the concentration areas in Equation (15) by the areas of the inscribed triangles with maximum height we have approximately:

$$\frac{\Delta odg + \Delta dbh}{\Delta obd} \doteq \frac{\Delta ogp + \Delta gdq + \Delta dhr + \Delta hbs}{\Delta odg + \Delta dbh} \doteq \ldots \doteq \frac{1 - S^2 + \epsilon^2}{2} \tag{15}$$

and therefore the concentration area is given approximately by

$$\Delta odb + (\Delta odg + \Delta dbh) + (\Delta ogb + \Delta gdq + \Delta dhr + \Delta hbs) + \ldots$$

which finally yields

$$R \doteq S + S\frac{1 - S^2 + \epsilon^2}{2} + S\left(\frac{1 - S^2 + \epsilon^2}{2}\right)^2 + \ldots \doteq \frac{2S}{1 + S^2 - \epsilon^2}, \tag{16}$$

since $\Delta obd = S$ and $\frac{1-S^2+\epsilon^2}{2} < 1$; moreover, if the curve is culminating, i.e. $\epsilon = 0$, $S = D$, then it is

$$R \doteq \frac{2D}{1 + D^2}. \tag{17}$$

Finally, Gini remarks that Equations (16) and (17) correspond to the unrealistic hypothesis that the single parts of the concentration curve the same convexity holds as in the global curve. In practice, however, these formulae provide upper bounds for the concentration ratio, say R'', which can turn out to be helpful in some situations. On the other hand, Equation (10) is related to a plausible assumption, that is equitension, which provides instead a lower bound for the concentration ratio, say R'. In general, a concentration curve will be located in an intermediate position between these two extreme cases. In order to give more emphasis to these theoretical results, Gini verifies them on some numerical examples, which are not shown for the sake of conciseness.

7 Conclusions

The scope of this paper is to popularize a selection of translated excerpts from the 1932 paper "*Intorno alle curve di concentrazione*" in which Corrado Gini first developed some fundamental ideas about the concentration curve that constitute the basis of a whole branch of research. In particular, as we mentioned in the Introduction, Gini was the first to introduce a new coordinate system for the concentration curve later on independently rediscovered by [23]. Ensuing Camilo Dagum's advise of providing the scientific community with an English version of the original work, we celebrate the memory of the founding father of the so-called Italian Statistical School (and of *Metron*), by finally acknowledging the importance and the originality of his research on the concentration curves.

References

[1] Amoroso, L.: Ricerche intorno alla curva dei redditi. Annuali di Matematica pura ed applicata, serie IV **2**, 123–159 (1925)*

[2] Bortkiewicz, L.: Die Disparitätsmasse der Einkommens Statistik. XIX session de l'Institut International de Statistique, Tokio. **25**(3), 189–298 (1930)*

[3] Bowley, A.L.: Elements of Statistics. Fourth Edition, London, King & Son (1920)*

[4] Czuber, E.: Beitrag zur Theorie statistischer Reihen, Versicherungswissenschaftlichen Mitteilungen, Wien. **9**(2), 101—175 (1914)*

[5] Dagum, C.: The Generation and Distribution of Income, the Lorenz Curve and the Gini Ratio. Economie Appliquée. **23**(2), 327–367 (1980)

[6] Dagum, C.: Kakwani, N.C. (1980). Income inequality and poverty: methods of estimation and policy applications, Oxford University Press, New York. Journal of Business and Economic Statistics (Book Reviews). **4**(3), 391 (1986)

[7] Dagum, C.: On the Relationship between Income Inequality Measures and Social Welfare Functions. Journal of Econometrics. **43**(1-2), 91–102 (1990)

[8] Dagum, C.: The Social Welfare Bases of Gini and Other Income Inequality Measures. Statistica. **53**(1), 3–30 (1993)

[9] Dagum, C.: A New Approach to the Decomposition of the Gini Income Inequality Ratio. Empirical Economics. **22**(4), 515–531 (1997)

[10] Dalton, H.: The measurement of the inequality of incomes. The Economic Journal. **30**(119), 348–361 (1920)*

[11] Galton, F.R.S.: Inquiries into human faculty and its development. Macmillan, London. (1883)*

[12] Galvani, L.: Contributi alla determinazione degli indici di variabilità per alcuni tipi di distribuzioni. Metron. **9**(1), 3–45 (1931)*

[13] Galvani, L.: Sulle curve di concentrazione relative a caratteri limitati e non limitati. Metron. **10**(3), 61–73 (1932)

[14] Gini, C.: Variabilità e mutabilità. In: Studi Economico-Giuridici della Facoltà di Giurisprudenza della Regia Università di Cagliari, 3 (part 2), Cuppini, Bologna (1912)*

[15] Gini, C.: Sulla misura della concentrazione e della variabilità dei caratteri. Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti. **73**, 1203–1248 (1914)*

[16] Gini, C.: Indici di concentrazione e di dipendenza. Biblioteca dell'Economista, serie V, **20**. Utet, Torino (1911)*

[17] Gini, C.: Sul massimo degli indici di variabilità assoluta e sulle sue applicazioni agli indici di variabilità relativa e al rapporto di concentrazione. Metron. **7**(3), 3-65 (1930)*

[18] Gini, C.: Intorno alle curve di concentrazione. Metron. **9**(3-4), 3–76 (1932)

[19] Gumbel, E.J.: Ein Mass der Konzentration bei pekuniären Verteilungen. Archiv für Sozialwissenschaft und Sozialpolitik. **58**(1), 113–139 (1927)*

[20] Gumbel, E.J.: Das Konzentrationsmass. Mat. Sb. **35**(1), 65–86 (1928)*

[21] Julin, A.: Principes de Statistique théorique et appliquée (I). Statistique théorique. Paris (1921)*

[22] Kakwani, N.C.: Income inequality and poverty: methods of estimation and policy applications. Oxford University Press, New York. (1980)

[23] Kakwani, N.C., Podder, N.: Efficient Estimation of the Lorenz curve and associated inequality measures from grouped observations. Econometrica. **44**(1), 137–148 (1976).

[24] Kendall, M.G., Buckland, W.R.: A Dictionary of Statistical terms. IV edition, Ed. Longman (1982).

[25] March, L.: Les principes de la méthode statistique. Felix Alcan, Paris (1930)*

[26] Mortara, G.: Elementi di Statistica: appunti sulle lezioni di statistica metodologica, Istituto superiore di scienze economiche e commerciali di Roma. Athenaeum. (1917)*

[27] Mortara, G.: Lezioni di statistica metodologica: dettate nel R. Istituto superiore di scienze economiche e commerciali di Roma. Societa tipografica Leonardo da Vinci, Città di Castello (1922)*

[28] Pietra, G.: Delle relazioni tra gli indici di variabilità: nota I e nota II. Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti. **74**, 775-804 (1915)*

[29] Pietra, G.: Recenti pubblicazioni di metodologia statistica. Rivista Italiana di Sociologia **21**(2-3), 310–319 (1917)

[30] Ricci, U.: L'indice di variabilità e la curva dei redditi. Giornale degli Economisti e Rivista di Statistica, **27**(9), 177–228 (1916)*

[31] Savorgnan, F.: Intorno all'approssimazione di alcuni indici della distribuzione dei redditi. Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti.**74**, 837–893 (1915)*

Note: Starred references are reported from the original paper.