QUANTITATIVE METHODS IN ECONOMICS
Volume XVIII, No. 3, 2017, pp. 472 – 491

Received: 09.09.2017
Accepted: 05.10.2017

DECOMPOSITION AND NORMALIZATION OF ABSOLUTE DIFFERENCES, WHEN POSITIVE AND NEGATIVE VALUES ARE CONSIDERED: APPLICATIONS TO THE GINI COEFFICIENT

Katarzyna Ostasiewicz
Faculty of Management, Informatics and Finances
Wroclaw University of Economics, Poland
e-mail: katarzyna.ostasiewicz@ue.wroc.pl
Achille Vernizzi (ORCID: 0000-0002-1641-5003)
Department of Economics, Management and Quantitative Methods
Università degli Studi di Milano, Italy
e-mail: achille.vernizzi@unimi.it

Abstract: We show how the absolute differences approach is particularly effective to interpret the Gini coefficient (G) when a distribution includes both positive and negative values. Either in erasing units having negative values, or in transforming negative values into zero, a significant variability fraction can be lost. When including negative values, instead of correcting G, to maintain it lower than 1, the standard G should be kept to compare the variability among different situations; a recent normalization, G_p, can be associated to G, to evaluate the variability percentage inside each situation.

Keywords: absolute difference components, negative values, normalization of Gini based coefficients

INTRODUCTION

The Gini coefficient is normally used in presence of non negative values, so that, when the distribution at stake includes negative values, it is common practice either excluding units with negative values, or transforming negative values into zero, with the latter suggested by OECD [2014]. Many transferable variables can take on negative values in their distributions. When dealing with monetary variables, e.g., there could be several reasons for an income unit to have negative net income, at least in terms of a particular source. For example, when assessing income units and financial assets such as capital gains, negative values can be

DOI: 10.22630/MIBE.2017.18.3.44

observed. Negative values can also be seen dealing with self-employed workers' incomes, if losses are greater than gains; the same money transfers are positive revenues - taking into consideration persons who receive them – and negative revenues – while considering the other persons. Another example is tax systems that admit negative income taxes, which can originate, for instance, from child allowances.

The most frequently used single measure of income inequality is the Gini coefficient of concentration. However, when a distribution includes negative values, as Castellano [1937] observes, the Lorenz curve lays under the x-axis (here we suppose that the average of the variable is positive) and the Gini coefficient can assume values greater than one, as it is observed by Hagerbaumer [1977], Pyatt et al. [1980], Lambert and Yitzhaki [2013]. In eliminating the observations with negative values or in converting them into zero, this outcome is avoided.

However, this approach may neglect a significant proportion of variability and, as a consequence, can lead to unreliable comparisons among distributions.

In order to restrict the Gini coefficient to the range 0-1, Chen et al. [1982] modify the normalizing factor by adding a certain component. This component depends on the distribution of negative values and of such proportion of the smallest positive values, which are enough to compensate for the former. The authors' proposition is in fact not a normalization but rather an ad hoc correction, as it depends on the particular form of the compensating area at stake (Chen et al.'s method was subsequently completed by Berrebi and Silber [1985], who provided a correct expression for the general case - when the fractional number of smallest positive units compensates for the sum of the negative ones).

Chen et al.'s correction has the advantage of making the modified Gini coefficient decrease for any equalitarian redistribution. However, Chen et al.'s coefficient becomes less and less sensitive as the concentration increases. Raffinetti et al. [2015] provide several examples on this point and suggest a normalization that keeps into account the potential maximum Gini mean difference. The authors formulate certain conditions for the application of their normalization.

In this paper we attempt to better understand the behaviour and the meaning of the Gini coefficient, of its modifications presented in the literature and its practical adaptations when negative values are observed. Here, we consider, together with the standard Gini coefficient, the coefficient G_p, introduced by Raffinetti et al., and the correction introduced by Chen et al. including Berrebi and Silber's completion. The behaviours of these indexes are tested when compensative transfers occur between units with positive values and units with negative values of the variable, so that the negative values are transformed into zero, thanks to transfers from units with positive values. The paper is organized as follows. The next section examines the components of the standard Gini coefficient when it is calculated either by including units that have negative values or excluding these units or turning their values into zero. The section which follows, is an overview of several adjustments proposed in the literature on the calculation of the Gini coefficient. In particular, the section provides a deepening on the Chen et al. correction and shows what Raffinetti et al.'s normalization means and how it should be used. The section entitled "Compensative redistributions" considers the behaviour of the indexes previously introduced under the compensative equalitarian redistribution. The section which comes next, provides a numerical example, which illustrates the theoretical behaviours described in the previous sections; this section shows also how the standard Gini coefficient should be interpreted, with the information provided by G_p. The last section offers a conclusion.

THE GINI COEFFICIENT IN THE PRESENCE OF NEGATIVE VALUES

Let's consider a variable that takes on negative values, units arranged in a non-decreasing order $\left[x_1, x_2, ..., x_N, x_{N+1}, x_{N+2}, ..., x_M\right]$.

We suppose that the first N units, $i = 1, 2, \dots N$, have negative values, while the remaining units, $i = N + 1, N + 2, \dots M$) are non-negative. We assume that the sum of the non-negative values, $\sum_{i=N+1}^{M} x_i = T_a$, is greater than the absolute sum of the negative values, $\sum_{i=1}^{N} |x_i| = T_n$, i.e., $\sum_{i=1}^{M} x_i = T_a - T_n > 0$.

If we split the whole distribution into two groups, the former containing the negative values and the latter the non-negative values, we can write the sum of the absolute differences as:

$$S = \sum_{i=1}^{M} \sum_{j=1}^{M} |x_i - x_j| = S_n + 2\left[NT_a + (M-N)T_n\right] + S_a . \tag{1}$$

In (1), $S_n = \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|$ is the within group component for the units with negative values, $S_a = \sum_{i=N+1}^{M} \sum_{j=N+1}^{M} |x_i - x_j|$ is the within group component for the units with non-negative values and $NT_a + (M-N)T_n = \sum_{i=1}^{N} \sum_{j=N+1}^{M} |x_i - x_j|$ is the between-group component, which in S appears twice (see appendix for further details).

If we apply the well-known results concerning the sum of absolute differences (see [Gini 1930; Castellano, 1937]), under the condition that the total amount both of absolute negative values, T_n, and of positive values, T_a, remain constant, we can see that $\max\{S_n\} = 2(N-1)T_n$; $\max\{S_a\} = 2(M-N-1)T_a$. Therefore, it follows that

$$\max\{S\} = \max\{S_n\} + 2\left[NT_a + (M-N)T_n\right] + \max\{S_a\}$$
$$= 2(N-1)T_n + 2\left[NT_a + (M-N)T_n\right] + 2(M-N-1)T_a$$

$$=2(M-1)(T_a+T_n), \tag{2}$$

which illustrates what Raffinetti et al. [2015] report in their expression (5).

It follows that, when the distribution of the variable at stake includes negative values, the Gini coefficient

$$G=\frac{S}{2(M-1)(T_a-T_n)} \text{ lies in the interval } 0\le G\le\frac{T_a+T_n}{T_a-T_n}. \tag{3}$$

The upper bound holds if the total positive amount is possessed by one unit, the total loss is suffered by another single unit, and x_i's are equal to 0 for the remaining (M - 2) units.

We stress that in (3) for the maximum to remain unchanged, it is enough that the ratio (T_n/T_a) remains constant.

When the number of units is large enough, G is approximated by $G = \Delta/2\mu$, with $\Delta = S/M^2$ and $\mu = (T_a - T_n)/M$. Analogous simplifications apply to other indexes considered in this paper, whenever deviations are substituted by ratios of averages.

As mentioned above, the majority of researchers either erase the units with negative values or convert the negative values into zero. These procedures should be adopted when both the sum of negative values and the number of units with negative values and are negligible.

The Gini coefficient erasing negative values

If the negative values are erased, the Gini coefficient becomes

$$G_a=\frac{S_a}{2(M-N-1)T_a}. \tag{4}$$

G_a excludes from its numerator both the variability within the units with negative values, S_n, and the variability between these units and those with non-negative values, $\left[NT_a+(M-N)T_n\right]$.

The Gini coefficient while turning negative values into zeros

When the negative values are turned into zero, the Gini coefficient becomes

$$G_{za}=\frac{2NT_a+S_a}{2(M-1)T_a}. \tag{5}$$

In expression (5), the component $2NT_a$ expresses the differences between the first N units (i = 1, 2,…, N), which are set as equal to zero and the units that maintain their original non-negative values ($i = N + 1, N + 2,…, M$) (see appendix for further details).

G_{za} excludes S_n and part of the between-group variability, i.e., $(M-N)T_n$.

If we rewrite the denominator of (5) as $2(M-1)T_a=2(M-N-1)T_a+2NT_a$ and then compare G_{za} with G_a, we see that both the numerator and the denominator of the former differ from the numerator and the denominator of the latter by the same quantity, $2NT_a$. Then, as $S_a \leq 2(M - N - 1)T_a$, we are able to conclude that $G_a \leq G_{za}$.

It should be noted that $G_{za} \leq G$, as the denominator of (3) is smaller than that of G_{za}, while the numerator of G is greater than that of G_{za}. Then, *a fortiori*, $G_a \leq G$.

NEGATIVE VALUES AND ADJUSTMENTS IN THE CALCULATION OF THE GINI COEFFICIENT

Chen et al. [1982] (henceforth CTR) suggest a correction that, on the one hand, allows preservation of the whole variability in S and, on the other hand, keeps the modified Gini coefficient within the range [0; 1]. Another treatment is proposed by Raffinetti et al. [2015]. Basing on (3), they suggest dividing S (as calculated with formula (1)) by $2(M-1)(T_a+T_n)$, i.e., dividing G by its upper bound $(T_a+T_n)/(T_a-T_n)$.

The CTR correction

The authors' correction is obtained by "freezing" the ratio between the average of the net available amount, and the average of absolute differences, calculated within a particular subset of the distribution: the subset which includes all the negative values and the smallest positive values. Even if CTR and BS start from absolute differences, the authors' methodology is eventually conducted and interpreted in terms of areas bounded by the Lorenz curve: consequently, according to the authors' approach, what is "frozen" is the area which lies below the *x*-axis. The CTR correction was completed by Berrebi and Silber [1985] (henceforth BS).

Here, we shall consider the CTR-BS correction entirely under the approach of absolute differences, as do Raffinetti et al. [2015].

In order to understand the rationale of the formula, we introduce some further pieces of notation. Having ordered the units in non-decreasing order with respect to the values of the variable, we suppose that $\sum_{i=1}^{K} x_i \leq 0$, and that $\sum_{i=1}^{K+1} x_i > 0$. Indeed, as BS observe, the sum of negative values is not necessarily compensated by an exact (integer) number of non-negative values; we can write that $\sum_{i=1}^{K} x_i + \eta x_{K+1} = 0$ and $(1-\eta)x_{K+1} + \sum_{i=K+2}^{M} x_i = T_a - T_n$, with $\eta = \left|\sum_{i=1}^{K} x_i\right| / x_{K+1}$ (or $\eta = -\sum_{i=1}^{K} x_i / x_{K+1}$).

We can now represent the distribution of the variable as

$$\left[x_1, x_2, ..., x_N, x_{N+1}, x_{N+2}, ..., x_K, (x_{K+1})_1, (x_{K+1})_2, x_{K+2} ...., x_M\right] \tag{6}$$

In (6), x_i, $i = 1, 2,\ldots,N$, are the units with a negative value of the variable and, for the remaining units, $i = N + 1, 2,\ldots,M$, $x_i \geq 0$. In (6), x_{K+1} is split into two sub-units: $(x_{K+1})_1$ with weight η and $(x_{K+1})_2$ with weight $(1 - \eta)$, respectively, $\eta \leq 1$. It follows that

$$2\sum_{i=1}^{M}\left|x_{K+1}-x_i\right| = 2\sum_{i=1}^{K}\left(x_{K+1}-x_i\right)\eta + 2\sum_{i=1}^{K}\left(x_{K+1}-x_i\right)(1-\eta)$$
$$+2\sum_{i=K+2}^{M}\left(x_i-x_{K+1}\right)\eta + 2\sum_{i=K+2}^{M}\left(x_i-x_{K+1}\right)(1-\eta);$$

$(x_{K+1})_1$ will be regarded as belonging to the "lower" set in (6), and $(x_{K+1})_2$ as belonging to the "upper" set of (6).

Having defined

$$S_0 = \sum_{i=1}^{K}\sum_{j=1}^{K}\left|x_i-x_j\right| + 2\sum_{i=1}^{K}\left(x_{K+1}-x_i\right)\eta \tag{7}$$

, which is the sum of absolute differences within the subset $\left[x_1, x_2, \ldots, x_N, x_{N+1}, x_{N+2}, \ldots, x_K, \left(x_{K+1}\right)_1\right]$, and referring to Raffinetti et al. [2015], formula 3 and the proof reported below the formula, the CTR-BS Gini coefficient can be written as:

$$G_{C-S} = \frac{S}{S_0 + 2(M-1)\left(T_a - T_n\right)}. \tag{8}$$

As the appendix shows (formulae A6 ÷ A11), an alternative decomposition for S is

$$S = S_0 + 2(K+\eta)\left(T_a - T_n\right) + S_u. \tag{9}$$

In (9)

$$S_u = \sum_{i=K+2}^{M}\sum_{j=K+2}^{M}\left|x_i-x_j\right| + 2\sum_{i=K+2}^{M}\left(x_i-x_{K+1}\right)(1-\eta) \tag{10}$$

is the sum of absolute differences among units in the subset $\left[\left(x_{K+1}\right)_2, x_{K+2}, \ldots, x_M\right]$; $(K+\eta)\left(T_a - T_n\right)$ is the sum of absolute differences between these units in subset $\left[x_1, x_2, \ldots, x_N, x_{N+1}, x_{N+2}, \ldots, x_K, \left(x_{K+1}\right)_1\right]$ and those in subset $\left[\left(x_{K+1}\right)_2, x_{K+2}, \ldots, x_M\right]$

If we focus just on S_u, by applying the usual results, S_u yields its maximum when $x_M = \left(T_a - T_n\right)$ and the remaining x_i ($i=K+1$, $x_{K+2},\ldots x_{M-1}$) are zero; if this is the case,

$$\max\left\{S_u\right\} = 2(M-K-2)\left(T_a - T_n\right) + 2\left(T_a - T_n\right)(1-\eta) \tag{11}$$

and consequently

$$2(K+\eta)\left(T_a - T_n\right) + \max\left\{S_u\right\} = 2(M-1)\left(T_a - T_n\right). \tag{12}$$

Therefore, as $2(K+\eta)\left(T_a - T_n\right) + S_u \leq 2(M-1)\left(T_a - T_n\right)$, G_{C-S} cannot be greater than 1, as we have assumed that the net amount of the variable is positive, $0 \leq G_{C-S} \leq 1$. Obviously, G_{C-S} is zero if all x_i ($i = 1, 2, \ldots, M$) are equal, in which case all the three components in (9) are zero.

We can observe that the CTR-BS correction does not refer to a theoretical extreme situation: it adds to the denominator a quantity, S_0, which is present in the numerator: consequently it is an ad hoc procedure. Moreover, even if G_{C-S} is a direct function of S_0, Raffinetti et al. [2015] observe that the more S_u approaches $\max\{S_u\}$, the less sensitive G_{C-S} is to what exists inside this set of units.

Reconsidering the CTR correction

We will now add several further considerations and introduce a revision of the CTR approach. Having accepted that the subset $\left[x_1, x_2, ..., x_N, x_{N+1}, x_{N+2}, ..., x_K, \left(x_{K+1}\right)_1\right]$ is kept constant, for a given net amount $\left(T_a - T_n\right)$ the maximum S is generated by the set

$$\left[x_1, x_2, ..., x_N, x_{N+1}, x_{N+2}, ..., x_K, \left(x_{K+1}\right)_1, (0)_2, 0...., \left(T_a - T_n\right)\right]. \tag{13}$$

In the appendix, formula (A20) shows that the overall sum of absolute differences for the elements in set (13) is

$$S^* = S_0 + 2\left(M-1\right)\left(T_a - T_n\right) + 4(M - K - 1 - \eta)T_n. \tag{14}$$

In (14), the component $4(M - K - 1 - \eta)T_n$ is the so-called transvariation term (see Dagum's terminology, 1997): it arises because the two subsets

$$\left[x_1, x_2, ..., x_N, x_{N+1}, x_{N+2}, ..., x_K, \left(x_{K+1}\right)_1\right] \text{ and } \left[(0)_2, 0...., \left(T_a - T_n\right)\right] \tag{15}$$

now overlap, as, within the first subset, at least one x_i is greater than zero.

So, if we normalize by the maximum S (maximum - keeping the lower subset constant, as in (13)) we yield

$$G^*_{C-S} = \frac{S_0 + 2\left(K + \eta\right)\left(T_a - T_n\right) + S_u}{S_0 + 2\left(M-1\right)\left(T_a - T_n\right) + 4(M - K - 1 - \eta)T_n}. \tag{16}$$

We observe that if, instead of calculating the absolute differences $\sum_{i=1}^{M}\sum_{j=1}^{i}\left|x_i - x_j\right|$, among the elements of the set (13), we calculate the simple differences $\sum_{i=1}^{M}\sum_{j=1}^{i}\left(x_i - x_j\right)$, (see [Lambert 2001], Ch. 2), we yield $S_0 + 2\left(M-1\right)\left(T_a - T_n\right)$, which is the correction adopted by CTR. Note that S_0 would coincide with S^*; only if $x_i \geq x_j$, for all $i > j$ (i = 1, 2, …, M). This condition is not fulfilled in (13), then, being $S_0 < S^*$, we have that $G^*_{C-S} < G_{C-S}$.

The Raffinetti, Siletti and Vernizzi normalization

If we normalize G, taking into account its upper bound (as in (3)), we yield

$$G_p = G\frac{T_a - T_n}{T_a + T_n}, \tag{17}$$

G_p is the index suggested by Raffinetti et al. [2015]: the maximum for G_p is 1.

COMPENSATIVE REDISTRIBUTIONS

Any equitable transfer lowers the standard Gini coefficient, as defined by expression (3). If we consider a redistribution that compensates negative values into non-negative values, by subtracting the overall amount T_n from units having positive values, after the compensation, all the indexes, introduced above, coincide with the standard Gini coefficient. However, even if such a redistribution is performed by equitable transfers, after this redistribution, the Gini coefficient may be greater than G_a, G_{za}, G^*_{C-S}, and G_p, calculated for the distribution before these transfers . The only exception is G_{C-S}.

As an example, let's consider an equalitarian compensation, achieved at the expense of the units with the smallest positive values. This compensation acts inside the subset $\left[x_1, x_2, ..., x_N, x_{N+1}, x_{N+2}, ..., x_K, \left(x_{K+1}\right)_1\right]$ and transforms all the x_i's within the subset into zeros. Notice that both before and after the compensation, $\sum_{i=1}^{K} x_i + \eta x_{K+1} = 0$. The subset $\left[\left(x_{K+1}\right)_2, x_{K+2} ...., x_M\right]$ remains unchanged. We label this redistribution "minimal compensation".

After such a redistribution, all the Gini indexes introduced in the previous sections (G_a, G_{za}, G_{C-S}, G^*_{C-S}, and G_p) can be reduced to the expression[1]

$$G = \frac{2(K+\eta)(T_a - T_n) + S_u}{2(M-1)(T_a - T_n)}. \tag{18}$$

Needless to say, for $M \to \infty$, when dividing the numerator and the denominator by M^2, the final result is practically the same if we leave (18) unchanged.

For what concerns the behaviour of G_{za}, having labelled S_c the sum of absolute differences within the subset $\left[x_{N+1}, x_{N+2}, ..., x_K, \left(x_{K+1}\right)_1\right]$ and using the decomposition (see appendix, formula A14)

$$S_a = S_c + 2(K - N + \eta)(T_a - T_n) - 2(M - K - \eta)T_n + S_u, \tag{19}$$

[1] It should be noted that, if the compensation takes place involving the highest value, i.e., including the share η of x_M, than the denominator of (18) should be replaced by $2(M-1+\eta)(T_a - T_n)$.

G_{za} can be expressed as

$$G_{za} = \frac{2NT_a + S_a}{2(M-1)T_a} = \frac{S_c - 2(M-K-\eta-N)T_n + 2(K+\eta)(T_a - T_n) + S_u}{2(M-1)(T_a - T_n) + 2(M-1)T_n}. \quad (20)$$

It can be verified that after the compensation, the Gini coefficient, calculated by (18), is greater than the before-compensation G_{za}. In fact, after some manipulations we yield that inequality (20) is verified if

$$\frac{S_c - 2(M-K-\eta-N)T_n}{2T_n} < \frac{2(K+\eta)(T_a - T_n) + S_u}{2(T_a - T_n)}.$$

In the l.h.s. of the above expression, the maximum is reached when $S_c = 2(K+\eta-N-1)T_n$, whilst in the r.h.s. the minimum is reached when $S_u = 0$. When both circumstances are verified, after elementary simplifications, the inequality becomes $2(K+\eta)-1-M < K+\eta$, from which we yield $K+\eta-1 < M$, which is trivially verified.

As $G_{za} \geq G_a$, a fortiori, G_a is lower than the Gini coefficient in (18).

Let's now compare the after-compensation Gini coefficient (18) with G_p, which can be written as

$$G_p = \frac{S_0 + 2(K+\eta)(T_a - T_n) + S_u}{2(M-1)(T_a - T_n) + 4(M-1)T_n}. \quad (21)$$

Keeping in mind G_p, as in expression (21), and G, as in (18), let us investigate conditions under which it will happen, that: $G_p \geq G$. After some algebraic exercises we can see that it is equivalent to:

$$\frac{S_0}{4(M-1)T_n} \geq \frac{2(K+\eta)(T_a - T_n) + S_u}{2(M-1)(T_a - T_n)}. \quad (22)$$

However, inequality (22) does not hold, even when the left-hand side is maximum and the right-hand side is minimum. Indeed, in (22), the right-hand side is minimum when S_u is zero: in this case it reduces to $(K+\eta)/(M-1)$. The maximum for the left-hand side is obviously obtained when S_0 is maximum: as in the subset $\left[x_1, x_2, \ldots, x_N, x_{N+1}, x_{N+2}, \ldots, x_K, (x_{K+1})_1\right]$ both the sum of absolute negative values and that of positive values is T_n, we have that $\max\{S_0\} = 4(K+\eta-1)T_n$. Consequently, the maximum for the left-hand side of (22) is $(K+\eta-1)/(M-1)$. Thus, (22) never holds.

Conversely, before the compensation, G_{C-S} is greater than the after-compensation Gini coefficient (18). In fact, after the compensation, S_0 becomes zero; when this component disappears, the numerator of (8), expressed by (9), decreases proportionally more than its denominator.

The same does not happen for what concerns G^*_{C-S}. By comparing expression (16) with (18), we see that the former is greater than the latter if

$$\frac{S_0}{S_0+4(M-K-1-\eta)T_n} \geq \frac{2(K+\eta)(T_a-T_n)+S_u}{2(M-1)(T_a-T_n)}. \tag{23}$$

Ceteris paribus, the maximum for the left-hand side is reached when S_0 is maximum; that is, for $S_0 = 4(K+\eta-1)T_n$. In this case the inequality (23) is

$$\frac{(K+\eta-1)}{(M-2)} \geq \frac{(K+\eta)}{(M-1)} + \frac{S_u}{2(M-1)(T_a-T_n)}.$$

If both M and K are large enough, the left-hand side term and the first addend on the right-hand side are almost equal; thus, generally, inequality (23) does not hold. It follows that G^*_{C-S} cannot be greater than the Gini coefficient after the minimal compensation.

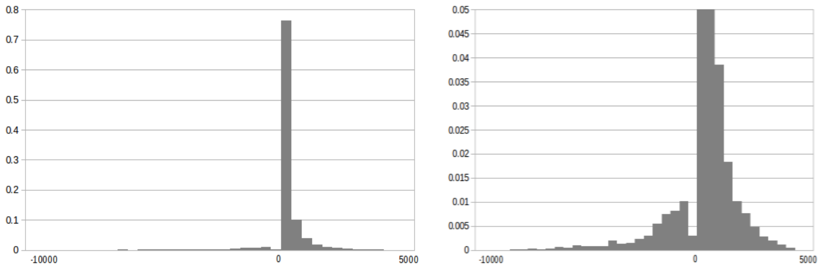
We conclude that only two of the indexes considered here always decrease when negative values are transformed into zero when compensated by an equalitarian redistribution from positive values: the usual Gini coefficient (as defined by expression (3)) and G_{C-S}. In adopting G we have to accept that it can be greater than 1. If we adopt G_{C-S}, we have to be aware that in the denominator it presents an ad hoc correction: due to this ad hoc correction, comparisons among G_{C-S} related to different situations, should be done only if the ratio between $(M-1)(T_a-T_n)$ and S_0 remains constant.

NUMERICAL EXAMPLE

In this section we will examine the measures discussed in the previous sections, as applied to the data generated from log-normal distribution. We will deal with two set of data, both consisting from 10,000 numbers. The first set includes 500 negative numbers generated from log-normal distribution with the parameters: (7.528; 0.812) (and with the sign inverted to negative), 1,500 zero's and 8,000 positive numbers generated from log-normal distribution with parameters (5.428; 1.262). The second set consists from 1000 negative numbers generated from log-normal distribution with the parameters: (7.528; 0.812) (and with the sign inverted to negative), 1,500 zero's and 7,500 positive numbers generated from log-normal distribution with parameters (5.278; 1.376). The relative sizes of negative, zero and positive samples were chosen as to mimic some known properties of empirical distributions of net incomes of Italian households. It is known, that the share of negative values varies over time, while the share of zero incomes remains relatively constant. Moreover, the parameters of the log-normal distributions were chosen to ensure realistic values of skewness and kurtosis for both sets of data. The histograms of relative frequencies for set 1 and set 2 are presented in Figures 1 and 2, respectively.

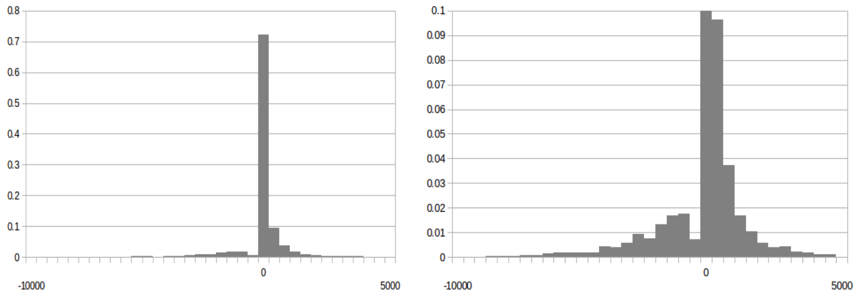
The main descriptive statistics of the data are summarized in the rows 2-14 of Table 1. The minimal compensation described in the above sections for non-decreasing series occurs at 8,568 and 9,767 positions for set 1 and set 2 respectively, see row 15 of Table 1.

Figure 1. Histogram of relative frequencies for random numbers constituting set 1. The right-hand-side picture is the same distribution but with truncated vertical axis, for better visualization of small relative frequencies for values far from 0



Source: own preparation

Figure 2. Histogram of relative frequencies for random numbers constituting set 2. The right-hand-side picture is the same distribution but with truncated vertical axis, for better visualization of small relative frequencies for values far from 0



Source: own preparation

Considering the simplest ways of dealing with negative values – erasing them – one can see, that in this way we omit 35% of variability for the set 1 (see: $S_a/S = 0.650$) and as much as 53.3% of overall variability for set 2 (as: $S_a/S = 0.467$). Moreover, as erasing negative values make the overall average greater than the real average, the value of the Gini index calculated over such treatment will capture even less of inequality than the fraction of variability captured suggest. Indeed, the ratio G_a/G equals to 0.461 for the set 1 (almost 54% missing) and 0.164 for the set 2 (as much as almost 84% missing). Surely, it doesn't seem to be the proper way of dealing with data with this order of number of negative subset and with the overall average so influenced by the presence of negative values. Note, that it would suffice only one of these factors (strong underestimation of variability/strong overestimation of average value – of course, they are not independent) for the Gini value to be strongly influenced by such an artificial treatment of negative values.

Table 1. Descriptive statistics

	set number 1	set number 2
number of positive values	8,000	7,500
number of negative values	500	1,000
number of zeros	1,500	1,500
minumum value	-15,643.30	-28,121.30
maximum value	24,614.40	57,079.00
total amount of positive values	3,872,080.98	3,850,597.15
mean for positive values	484.01	513.41
total amount of absolute negative values	1,262,786.83	2,636,622.28
mean for negative values	-2,525.57	-2,636.62
overall mean	260.93	121.40
coefficient of variation	4.80	14.01
skewness	1.70	3.25
kurtosis	65	197
the lowest rank of the value for which the cumulative sum of ordered distribution is positive	8,568	9,767

Source: own calculations

The second simple way of dealing with negative values – turning them into zeros – gives a similar picture. The fractions of overall variability captured in this treatment increase – it is 69.8% for set 1 (= 0.698, which is an increase of 4.5 percentage points as compared to the previous treatment of erasing zeros) and 53.9% for set 2 (0.539, which is an increase of 7.2 percentage points as compared to the previous treatment of erasing zeros). Still, the fraction of the value of Gini index calculated for whole sets captured within this treatment is smaller than the fraction of overall variability captured, and is equal to 0.47 for set 1 and 0.17 for set 2, what is – for both set 1 and set 2 – higher fraction than within treatment of just erasing negative values.

If we consider the relevant share of variability not taken into account by G_a and G_{za}, we should conclude that these two indexes do not represent the actual variability, and consequently, they systematically underestimate the inequality. Moreover, there are problems in comparing distributions either with different percentages of units with negative values or with different ratios T_n/T_a.

As it was shown in the previous sections, values G_p, G_{za}, G_a, and G^*_{C-S} are always smaller than the value of Gini index after minimal compensation and it indeed holds for both set 1 and set 2 (see Table 2). Moreover, simple geometrical interpretation shows, that $G_{C\text{-}S}$ has always to be smaller than Gini index after minimal compensation. Indeed, minimal compensation just turns the negative part of the Lorenz curve into zero. As $G_{C\text{-}S}$ is equal to: $2(A + B)/(1 + 2A)$, while Gini after minimal compensation is equal to $2B$, (A – denoting the area between negative part of the Lorenz curve and horizontal axis, while B – the area between positive part of Lorenz curve, horizontal axis and the line of equal share) it turns out that if for $A > 0$ Gini after minimal compensation will be smaller than $G_{C\text{-}S}$ for $B < 1/2$, that is, always.

Table 2. Values of different measures of inequality discussed in the text

	set number 1	set number 2
G	1.558	4.453
upper bound for G	1.968	5.344
$G_p=G/G_{max}$	0.792	0.833
S_a/S	0.650	0.467
G_a	0.719	0.729
G_a/G	0.461	0.164
(S_a+2NT_a)/S	0.698	0.539
G_{za}	0.733	0.756
G_{za}/G	0.470	0.170
$G_{c\text{-}s}$	0.947	0.996
G^*_{C-S}	0.874	0.974
G after min. compensation	0.913	0.984

Source: own calculations

However if we look at the two $G_{C\text{-}S}$ indexes, the effect of the minimal compensation does not appear to be so relevant as it is detected by the standard Gini coefficients. Indeed, due to the compensation, the Gini coefficient lowers from 1.558 to 0.913 in data set 1, and from 4.453 to 0.984 in data set 2.. Conversely, the decrease of the two $G_{C\text{-}S}$ indexes appears much smaller in both data sets, as, before the compensation, the two indexes are 0.947 and 0.996, respectively (after the minimal compensation $G_{C\text{-}S}$ amd G coincide):

On the basis of Frosini's ([1984], p. 274) observation that the term concentration should be applied only when non-negative values are considered, we should keep in mind that, when negative values are considered, the Gini coefficient is no longer a concentration coefficient, it is just a relative variability index. By looking to the standard Gini coefficients, in Table 2, we can say that in the second data set the relative variability is 2.9 times greater than in the first one. After the minimal compensative equalitarian redistributions, the relative inequality decreases to 0.913 and to 0.984 in the two data sets, respectively. Moreover, as after the compensations the negative values have been raised to zero, the two Gini coefficients can be considered concentration indexes. If we look at the G_p indexes, we can add that in the first data set, the relative variability is the 79.2 % of its potential maximum, whilst in the second it is the 83.3% of its potential maximum. After the compensative redistribution, even if the relative variability has decreased, the Gini coefficients are closer to their potential maximum, which is now 1, than they were before the compensation.

CONCLUSIONS

The purpose of this research was to indicate a valid operating procedure to calculate inequality when a distribution includes negative values. Generally, in overall income distributions only a few units have negative values. However, when we disaggregate overall income distributions into their sources, units having negative values can no longer be considered a negligible phenomenon. Another situation where many units with negative values can be observed is given by tax systems, which introduce family allowances through the form of negative income taxes.

In this article we have shown that when a distribution includes negative values, neither dropping units with negative values nor transforming these values to zero are suitable practices. This should not be done if we do not want both to exclude a part of the variability that can be considerable and to make invalid comparisons among distributions, related either to different populations or to the same population in different periods. Even if the Chen et al. [1982] coefficient appears a feasible procedure that preserves the whole variability, it presents some limits: first, it is an ad hoc procedure and second, it presents several abnormal behaviours in some circumstances, as stressed by Raffinetti et al. [2015]. Moreover, even accepting Chen et al.'s idea of compensating the negative values with the lowest positive values and not caring about abnormal behaviours, Chen et al. correction should be amended, as we highlighted in section "Negative values and adjustments…". By applying the amendment, however, we have shown that the modified coefficient can increase even after an equalitarian redistribution. Instead of adopting ad hoc corrections, we suggest a procedure based on two instruments. In comparing inequality among different distributions, the standard Gini coefficient can be still conveniently used, even when dealing with negative values; G is no longer a concentration measure but just a relative measure of variability. By dividing the Gini coefficient by its upper limit, one yields the normalized index G_p, suggested by Raffinetti et al.. This normalized index is a measure of the percentage of the potential maximum variability, for each specific situation, keeping constant the sum of negative values and the sum of positive ones. G_p can be used unconditionally, in the cases which present the same ratio between the sum of absolute negative values and the sum of positive values, T_n/T_a.

ACKNOWLEDGEMENTS

The authors desire to thank Francesca De Battisti, Vittorio Frosini, Alina Jędrzejczak, Marek Kośny and an anonymous referee for fruitful discussions and useful suggestions. Needless to say, the authors are the only persons responsible for any deficiency or error in the article.

REFERENCES

Berrebi Z. M., Silber J. (1985) The Gini Coefficient and Negative Income: a Comment. Oxford Economic Papers, 37, 525-526.

Castellano V. (1937) Sugli indici relativi di variabilità e sulla concentrazione dei caratteri con segno. METRON, XIII, 31-50.

Chen C. N., Tsaur T. W., Rhai T. S. (1982) The Gini Coefficient and Negative Income. Oxford Economic Papers, 34, 473-478.

Dagum C. (1997) A New Approach to the Decomposition of the Gini Income Inequality Ratio. Empirical Economics, 22, 515-531.

Gini C. (1930) Sul massimo degli indici di variabilità assoluta e sulle sue applicazioni agli indici di variabilità relativa e al rapporto di concentrazione. METRON, VIII, 3-65.

Frosini B.V. (1984) Concentration, Dispersion and Spread: An Insight into Their Relationship, Statistica, 44, 373-394.

Hagerbaumer J. B. (1977) The Gini Concentration Ratio and the Minor Concentration Ratio: a Two-Parameter Index of Inequality. Review of Economics and Statistics, LIX, 377-379.

Lambert P.J. (2001) The Distribution and Redistribution of Income. Manchester University Press, Manchester and New York.

Lambert P. J., Yitzhaki S. (2013) The Inconsistency between Measurement and Policy Instruments in Family Income Taxation. FinanzArchiv: Public Finance Analysis, 69, 241-255.

Nygard F., Sandström A. (1981) Measuring Income Inequality. Acta Universitatis Stockholmiensis, Almqvist & Wiksell International, Stockholm.

OECD. Terms of Reference. OECD Project on the Distribution of Household Incomes, 2014. Available at http://www.oecd.org/statistics/data-collection/Income%20distribution_guidelines.pdf.

Pyatt G., Chen C., Fei J. (1980) The Distribution of Income by Factor Components. The Quarterly Journal of Economics, 94, 451-473.

Raffinetti E., Siletti E., Vernizzi A. (2015) On the Gini Coefficient Normalization when Attributes with Negative Values Are Considered. Statistical Methods & Applications, 24, 507-521.

APPENDIX A: THE DECOMPOSITION OF THE SUM OF ABSOLUTE DIFFERENCES

The distribution splits into the two subsets of negative and non-negative values

Consider the distribution

$$\begin{array}{l} \left[x_1, x_2, \ldots, x_N, x_{N+1}, x_{N+2}, \ldots, x_M\right] \\ x_i \le x_{i+h},\ \ h>1; \\ x_i<0,\ i=1,2,\ldots,N,\ \ \sum_{i=1}^{N}\left|x_i\right|=T_n; \\ x_i \ge 0,\ i=N+1,N+2,\ldots,M,\ \ \sum_{i=N+1}^{M} x_i=T_a;\ \ T_a-T_n>0. \end{array}, \qquad \text{(A1)}$$

We can split

$$S=\sum_{i=1}^{M}\sum_{j=1}^{M}\left|x_i-x_j\right|$$
$$=\sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_i-x_j\right|+2\sum_{i=1}^{N}\sum_{j=N+1}^{M}\left|x_i-x_j\right|+\sum_{i=N+1}^{M}\sum_{j=N+1}^{M}\left|x_i-x_j\right|.$$

In the terminology of Dagum [1997], the first and the third terms are within-group components and the second is the gross-between component, which corresponds to the between component because the two groups do not overlap, that is $x_j \ge x_i$, i=1, 2,…, K+1, j=K+1,.K+2,…,M. Indeed, in this case we can write $\sum_{i=1}^{N}\sum_{j=N+1}^{M}\left|x_i-x_j\right|$ as $\sum_{i=1}^{N}\sum_{j=N+1}^{M}\left(x_j-x_i\right)$. Keeping in mind that $x_i<0$ $\left(i=1,2,\ldots,N\right)$ and that $x_i \ge 0\left(i=N+1,N+2,\ldots,M\right)$, it is easy to show that

$$S_{n,a}=\sum_{i=1}^{N}\sum_{j=N+1}^{M}\left|x_i-x_j\right|=\sum_{i=1}^{N}\sum_{j=N+1}^{M}\left(x_j-x_i\right)=NT_a+\left(M-N\right)T_n \qquad \text{(A2)}$$

In this article we denote the two within-group components as:

$$S_n=\sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_i-x_j\right|, \text{ and } S_a=\sum_{i=N+1}^{M}\sum_{j=N+1}^{M}\left|x_i-x_j\right|. \qquad \text{(A3)}$$

If the x_i, i= 1, 2, …N, are set as equal to zero, we have:

$$S_{n,a}=\sum_{i=1}^{N}\sum_{j=N+1}^{M} x_j=NT_a.$$

The distribution splits into the subset of minimal compensation and the complementary subset

Consider the distribution of the variable as

$$\left[x_1, x_2, \ldots, x_N, x_{N+1}, x_{N+2}, \ldots, x_K, \left(x_{K+1}\right)_1, \left(x_{K+1}\right)_2, x_{K+2} \ldots, x_M\right] \qquad \text{(A4)}$$

In (A4), all values are ranked in non-decreasing order. As in (A1), x_i, i = 1, 2,…,N are the units with a negative value of the variable and, for the remaining units, $i = N$ + 1, 2,…,M, $x_i \ge 0$. In (A4), x_{K+1} appears twice: the former as $(x_{K+1})_1$, with weight $\eta = \left|\sum_{i=1}^{K} x_i\right| \Big/ x_{K+1} = -\sum_{i=1}^{K} x_i \Big/ x_{K+1}$ and the latter as $(x_{K+1})_2$, with weight $(1-\eta)$, so that $-\sum_{i=1}^{N} x_i = \sum_{i=N+1}^{K} x_i + x_{K+1}\eta = T_n$, $\sum_{i=1}^{K} x_i + \eta x_{K+1} = 0$ and $(1-\eta)x_{K+1} + \sum_{i=K+2}^{M} x_i = T_a - T_n$.

We can split (A4) into two subsets:

$$\left[x_1, x_2, \ldots, x_N, x_{N+1}, x_{N+2}, \ldots, x_K, (x_{K+1})_1\right] \text{ and } \left[(x_{K+1})_2, x_{K+2}, \ldots, x_M\right] \tag{A5}$$

and, coherently, we can split the sum of absolute differences as:

$$S = \sum_{i=1}^{M}\sum_{j=1}^{M}\left|x_i - x_j\right| = \left[\sum_{i=1}^{K}\sum_{j=1}^{K}\left|x_i - x_j\right| + 2\sum_{i=1}^{K}\left(x_{K+1} - x_i\right)\eta\right]$$
$$+2\left[\sum_{i=1}^{K}\sum_{j=K+2}^{M}\left(x_j - x_i\right) + \sum_{i=1}^{K}\left(x_{K+1} - x_i\right)(1-\eta) + \sum_{j=K+2}^{M}\left(x_j - x_{K+1}\right)\eta\right]$$
$$+\sum_{i=K+2}^{M}\sum_{j=K+2}^{M}\left|x_i - x_j\right| + 2\sum_{i=K+2}^{M}\left(x_i - x_{K+1}\right)(1-\eta). \tag{A6}$$

In (A6), the first addend is the sum of absolute differences within the first subset in (A5) and the third addend is the sum of absolute differences within the second subset. The second addend represents the sum of absolute differences between the elements of the two subsets: as the elements in the first subset are never greater than those in the second, all the differences are non-negative and the modulus symbol can be omitted.

We denote

$$S_0 = \sum_{i=1}^{K}\sum_{j=1}^{K}\left|x_i - x_j\right| + 2\sum_{i=1}^{K}\left(x_{K+1} - x_i\right)\eta \tag{A7}$$

and

$$S_u = \sum_{i=K+2}^{M}\sum_{j=K+2}^{M}\left|x_i - x_j\right| + 2\sum_{i=K+2}^{M}\left(x_i - x_{K+1}\right)(1-\eta) \tag{A8}$$

For what concerns the between-subset component,

$$S_{0,u} = \sum_{i=1}^{K}\sum_{j=K+2}^{M}\left(x_j - x_i\right) + \sum_{i=1}^{K}\left(x_{K+1} - x_i\right)(1-\eta) + \sum_{j=K+2}^{M}\left(x_j - x_{K+1}\right)\eta \tag{A9}$$

we can split and rearrange it as

$$S_{0,u} = K\sum_{j=K+2}^{M} x_j - (M-K-1)\sum_{i=1}^{K} x_i + Kx_{K+1}(1-\eta) - \sum_{i=1}^{K} x_i(1-\eta)$$
$$+\sum_{j=K+2}^{M} x_j\eta - (M-K-1)x_{K+1}\eta.$$

The six terms can now be conveniently combined as

- $K\sum_{j=K+2}^{M} x_j + Kx_{K+1}(1-\eta) = K(T_a - T_n)$;
- $-(M-K-1)\sum_{i=1}^{K} x_i - (M-K-1)x_{K+1}\eta = 0$;
- by adding and subtracting $x_{K+1}\eta$ to

$\left[-\sum_{i=1}^{K} x_i(1-\eta) + \sum_{j=K+2}^{M} x_j\eta\right]$, we yield:

$$-\sum_{i=1}^{K} x_i - x_{K+1}\eta + \left(\sum_{i=1}^{K} x_i + x_{K+1} + \sum_{j=K+2}^{M} x_j\right)\eta = 0 + \left(T_a - T_n\right)\eta .$$

The results here allow us to rewrite the between component simply as:

$$S_{0,u} = \left(K+\eta\right)\left(T_a - T_n\right). \tag{A10}$$

Therefore, we can represent (A6) in the form:

$$S = S_0 + 2\left(K+\eta\right)\left(T_a - T_n\right) + S_u . \tag{A11}$$

Define the sum of absolute values within the subset
$\left[x_{N+1}, x_{N+2}, ..., x_K, \left(x_{K+1}\right)_1\right]$:

$$S_c = \sum_{i=N+1}^{K}\sum_{N+1}^{K}\left|x_i - x_j\right| + 2\sum_{i=N+1}^{K}\left(x_{K+1} - x_i\right)\eta . \tag{A12}$$

Let's now consider the sum of absolute differences between this subset, which contains the lowest non-negative values, and $\left[\left(x_{K+1}\right)_2, x_{K+2} ...., x_M\right]$:

$$S_{c,u} = \sum_{i=N+1}^{K}\sum_{j=K+2}^{M}\left|x_i - x_j\right| + \sum_{i=K+2}^{M}\left|x_i - x_{K+1}\right|\eta$$
$$+\sum_{i=N+1}^{K}\left|x_{K+1} - x_i\right|\left(1-\eta\right) + \left|x_{K+1} - x_{K+1}\right|\eta\left(1-\eta\right)$$
$$= \left(K-N\right)\sum_{j=K+2}^{M} x_j - \left(M-K-1\right)\sum_{i=N+1}^{K} x_i + \sum_{j=K+2}^{M} x_j\eta - \left(M-K-1\right)x_{K+1}\eta$$
$$+\left(K-N\right)x_{K+1}\left(1-\eta\right) - \sum_{i=N+1}^{K} x_i\left(1-\eta\right);$$

by adding the first addend to the third and the second addend to the sixth, we yield

$$= \left(K-N+\eta\right)\sum_{j=K+2}^{M} x_j - \left(M-K-\eta\right)\sum_{i=N+1}^{K} x_i$$
$$-\left(M-K-1\right)x_{K+1}\eta + \left(K-N\right)x_{K+1}\left(1-\eta\right).$$

If we now add $\left(K-N+\eta\right)\left(1-\eta\right)x_{K+1}$ to the first addend and subtract it from the fourth, and we subtract $\left(M-K-\eta\right)\eta x_{K+1}$ from the second addend and we add it to the third, we yield

$$\left(K-N+\eta\right)\left(T_a - T_n\right) - \left(M-K-\eta\right)T_n - \eta\left(1-\eta\right)x_{K+1} + \eta\left(1-\eta\right)x_{K+1}, \tag{A13}$$

having used $T_a - T_n = \sum_{j=K+2}^{M} x_j + x_{K+1}\left(1-\eta\right)$ and $T_n = \sum_{i=N+1}^{K} x_i + x_{K+1}\eta$.

Keeping in mind (A12), (A13) and (A8), S_a can be written as

$$S_a = S_c + 2\left(K+\eta-N\right)\left(T_a - T_n\right) - 2\left(M-K-\eta\right)T_n + S_u . \tag{A14}$$

Let's now consider the distribution

$$\left[x_1, x_2, ..., x_N, x_{N+1}, x_{N+2}, ..., x_K, \left(x_{K+1}\right)_1, \left(0\right)_2, 0 ...., \left(T_a - T_n\right)\right]. \tag{A15}$$

In (A15), $(x_{K+1})_1$ has weight η and $(0)_2$ has weight $(1-\eta)$.

The within component S_0 remains unchanged as it was for (A7); conversely S_u becomes:

$$S_u = 2\sum_{i=K+2}^{M-1}\left|0 - \left(T_a - T_n\right)\right| + 2\left|0 - \left(T_a - T_n\right)\right|\left(1-\eta\right)$$

$$= 2(M-K-2)(T_a - T_n) + 2(T_a - T_n)(1-\eta)$$

$$= 2(M-K-1-\eta)(T_a - T_n). \quad \text{(A16)}$$

In distribution (A15), the two groups overlap; then we have to consider the gross-between component:

$$S_{0,u} = \sum_{i=1}^{K}\sum_{j=K+2}^{M}\left|x_j - x_i\right| + \sum_{i=1}^{K}\left|x_{K+1} - x_i\right|(1-\eta) + \sum_{j=K+2}^{M}\left|x_j - x_{K+1}\right|\eta. \quad \text{(A17)}$$

Also, in this case we can avoid modulus but we have to express $S_{0,u}$ by adding to (A10) the transvariation component:

$$S_{0,u}^{T} = 2\Big[\sum_{i=N+1}^{K}\sum_{j=K+2}^{M-1}(x_i - 0) + \sum_{i=N+1}^{K}(x_i - 0)(1-\eta)$$

$$+\sum_{j=K+2}^{M-1}(x_{K+1} - 0)\eta + x_{K+1}(1-\eta)\eta\Big]$$

$$= 2(M-K-2)\sum_{i=N+1}^{K} x_i + 2\sum_{i=N+1}^{K} x_i(1-\eta) +$$

$$+ 2(M-K-2)x_{K+1}\eta + x_{K+1}(1-\eta)\eta$$

$$= 2(M-K-2)\sum_{i=N+1}^{K}(x_i + x_{K+1}\eta) + 2\Big(\sum_{i=N+1}^{K} x_i + x_{K+1}\eta\Big)(1-\eta)$$

$$= 2(M-K-1-\eta)\sum_{i=N+1}^{K}(x_i + x_{K+1}\eta) = 2(M-K-1-\eta)T_n. \quad \text{(A18)}$$

Therefore, (A17) becomes

$$S_{0,u} = 2(K+\eta)(T_a - T_n) + 2S_{0,u}^{T} = 2(K+\eta)(T_a - T_n) + 4(M-K-1-\eta)T_n. \quad \text{(A19)}$$

Using (A16) and (A19), the overall sum of absolute differences becomes

$$S^* = S_0 + 2(K+\eta)(T_a - T_n) + 4(M-K-1-\eta)T_n + S_u$$

$$= S_0 + 2(M-1)(T_a - T_n) + 4(M-K-1-\eta)T_n. \quad \text{(A20)}$$

In (A20),

$$S_0 + 2(M-1)(T_a - T_n) = 2\sum_{i=1}^{M}\sum_{j=1}^{i}(x_i - x_j),$$

which would be the sum of absolute differences if the rank in (A15) are the same as in (A1). For more details on the information provided by the different ordering, see [Lambert, 2001 Ch. 2].

APPENDIX B: OPERATING FORMULAE

In order to simplify and fasten calculations, one can apply the operating formulae enlisted in this appendix.

$$S_a = \sum_{i=1}^{P}\sum_{j=1}^{P}\left|x_i - x_j\right| = 4\sum_{i=1}^{P} x_i i - 2\left(P+1\right)\sum_{i=1}^{P} x_i\,, \tag{B1}$$

$$S_n = \sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_i - x_j\right| = 4\sum_{i=1}^{N} x_i i - 2\left(N+1\right)\sum_{i=1}^{N} x_i\,. \tag{B2}$$

$$\begin{aligned} S_0 &= \sum_{i=1}^{K}\sum_{j=1}^{K}\left|x_i - x_j\right| + 2\sum_{i=1}^{K}\left(x_{K+1} - x_i\right)\eta \\ &= 4\sum_{i=1}^{K} x_i i - 2\left(2K+1\right)\sum_{i=1}^{K} x_i + x_{K+1}\eta^2\,, \end{aligned} \tag{B3}$$

where $\eta = \left|\sum_{i=1}^{K} x_i\right| \Big/ x_{K+1} = -\sum_{i=1}^{K} x_i \Big/ x_{K+1}$.

If $\sum_{i=1}^{K} x_i = 0$, then $S_0 = 4\sum_{i=1}^{K} x_i i$.

$$\begin{aligned} S_u &= \sum_{i=K+2}^{M}\sum_{j=K+2}^{M}\left|x_i - x_j\right| + 2\sum_{i=K+2}^{M}\left(x_i - x_{K+1}\right)\left(1-\eta\right) \\ &= 4\sum_{i=K+2}^{M} x_i i - 2\left(M+K+2\right)\sum_{i=K+2}^{M} x_i - 2\left(M-K-1\right)\left(\sum_{i=1}^{K+1} x_i\right) \\ &\quad + 2\left[\left(\sum_{i=1}^{K+1} x_i\right)\left(\sum_{i=K+2}^{M} x_i\right)\right] \Big/ x_{K+1}\,. \end{aligned} \tag{B4}$$

If $\sum_{i=1}^{K} x_i = 0$, then $S_u = 4\sum_{i=K+1}^{M} x_i i - 2\left(M+K+1\right)\sum_{i=K+1}^{M} x_i$.