INEQUALITY MEASURES FOR HISTOGRAMS

Benito V. Frosini

1. INTRODUCTION AND SUMMARY

Theoretical researches on concentration or inequality measures are mainly concerned with coherence properties (such as the Pigou-Dalton property) addressed to transfers between individuals, and with control of these properties on particular indices which are functions of all individual observations. On the other hand, economic applications are typically concerned with large masses of data, which are summarized by means of a distribution of frequencies and quantities over k classes, i.e. histograms. Two main problems are worth considering: (1) the goodness of approximation - to indices computed on individual data - of the same indices worked out on histograms; (2) the meaning and properties of inequality indices that are functions of the only frequencies and quantities pertaining to the k classes. These two kinds of investigation will be restricted to Gini and Pietra-Ricci indices, both for the approximation problem and for the derivation of similar indices to be applied in the case of histograms (for the original papers see Gini, 1914; Pietra, 1915; Ricci, 1916).

The following formal references and symbols will be used.

X will be a non-negative statistical variable, giving rise to values $x_1, \dots, x_n$ on n individuals. In order to give the concentration or inequality for the distribution of X a practical meaning, the phenomenon must be *transferable*, i.e. it must make sense to reduce the quantity pertaining to individual I, and correspondingly to augment the quantity of another individual J. The identification of individuals is assumed irrelevant; thus, for ease of reference, the statistical variable X will be defined by increasing (or not decreasing) values, namely $x_1 \le x_2 \le \dots \le x_n$.. With reference to all these n values we can define:

T = total = $\sum x_i$
m = mean = T/n
the individual share q_i is defined as $q_i = x_i/T \quad (i = 1, \dots, n)$;
$F_i = i/n$ = cumulative relative frequency ($i = 1, \dots, n$);
$Q_i = (x_1 + \dots + x_i)/T$ = cumulative relative quantity ($i = 1, \dots, n$).

With reference to successive groups $B_1, \dots, B_j, \dots, B_k$ of individuals, obtained by aggregating neighbouring values, we define:

$n_1, \ldots, n_k$ are the absolute frequencies ($\sum n_j = n$);
$N_1, \ldots, N_k$ are the cumulative absolute frequencies;
$f_1, \ldots, f_k$ are the relative frequencies ($\sum f_j = 1$);
$F_1, \ldots, F_k$ are the cumulative relative frequencies;
x_{jr} ($j = 1, \ldots, k$; $r = 1, \ldots, n_j$) are the x values organized in the k groups;
m_j = mean of x values in group B_j;
$q_1, \ldots, q_k$ are the shares of the groups ($q_j = m_j n_j / T, j = 1, \ldots, k$);
$Q_1, \ldots, Q_k$ are the cumulative shares;
$F_j \geq Q_j$ for $j = 1, \ldots, k - 1$; $F_k = Q_k = 1$.

When the successive groups are identified by classes for the variable X, such classes are defined by $C_j = (a_j, b_j]$, so that $a_j < x_{jr} \leq b_j$, ($j = 1, \ldots, k$; $r = 1, \ldots, n_j$); $x_j = (a_j + b_j)/2$ is the central point of class C_j.

2. THE GINI INDEX APPLIED TO HISTOGRAMS

It is well known that Gini's approach to measuring concentration or inequality is based on the comparison of cumulative shares of individuals (frequencies) and cumulative shares of income, being the individuals ordered by the amount of income:

$$A = \sum_{i=1}^{n-1} (F_i - Q_i) ; \tag{1}$$

as $\sum_1^{n-1} F_i = \sum_1^{n-1} i / n = (n - 1)/2$, exact normalization of A turns out as the Gini *concentration ratio* $R = 2A/(n - 1)$, also obtainable from the Lorenz plot as the normalized concentration area.

For easier reference and further elaboration, in the sequel we shall take into consideration the *Gini index*

$$G = \frac{n-1}{n} R = \frac{2}{n} \sum_{i=1}^{n} i\, q_i - \frac{n+1}{n} \tag{2}$$

(see e.g. Frosini, 1987, p. 193), which practically coincides with R in most applications (n sufficiently large), and is independent of n (it is invariant with respect to mixtures of distributions which are replicas of a given one).

Taking into account the organization of data x_{jr} into k groups or classes, we can write

$$G = \frac{2}{n} \sum_{j=1}^{k} \sum_{r=1}^{n_j} (N_{j-1} + r) q_{jr} - \frac{n+1}{n} ;$$

calling $$G_j = \frac{2}{n_j}\left\{\sum_{r=1}^{n_j} r\frac{q_{jr}}{q_j} - \frac{n_j+1}{n_j}\right\}$$

the Gini index computed on the values inside the group B_j, after some algebra one obtains (Frosini, 1987, p. 207):

$$G = \sum_{j=1}^{k} q_j(2F_j - f_j) - 1 + \sum_{j=1}^{k} q_j f_j G_j \tag{3}$$

The Gini index obviously coincides with

$$G_M = \sum_{j=1}^{k} q_j(2F_j - f_j) - 1 \tag{4}$$

if the successive groups are made of identical values, i.e. if G is computed on the frequency distribution (m_1,n_1; ... ; m_k,n_k), or - in other words - if there is no variability around m_j in each group. Another way of expressing the same thing is that (4) is based only on the knowledge of the frequencies f_j and shares q_j of each group. With respect to G_M, G is augmented by a term containing the group concentration indices G_j; when the groups are at least of the order of ten, as G_j is multiplied by "small" values $q_j f_j$, the increase represented by this term is practically irrelevant.

Some computations of this kind have been made in two papers, having as the main objective the comparability of a number of inequality measures computed on different distributions, empirical and theoretical; in order to ensure an exact formal comparability, the number of *aggregate units* for each distribution was fixed in advance as 10, or 25, or 50, or 100, each aggregate unit comprising the same number of individuals (f_j = 1/10, or 1/25 etc.). As expected, with k = 100 or 50 the values of G (indicated by R^* in the cited papers) turned out as practically identical - to the third decimal place - to the population values; but also with k = 25 and even k = 10 the approximation is satisfactory, if one excepts the case of some Pareto distributions with very large inequality.

The recourse to aggregate units may be evaluated according to two viewpoints: (a) the same number of aggregate units ensures comparability between different populations or distributions; (b) the computation of an index with reference to k aggregate units - or k classes - is made for achieving an "estimate" of the true value, i.e. the index computed on the whole population or distribution. On account of this last purpose, it is generally acknowledged that the concentration of several distributions is ordered according to their Lorenz curves - and consequently according to their concentration areas. Now, if we consider the Gini index G, it depends only on the concentration area, being equivalent to this area divided by ½. Quite to the contrary, the concentration ratio R is obtained by division by (k - 1)/2k; what happens when passing from k to h aggregate units, $h < k$, is that the concentration area can be slightly reduced - thus showing *less* concen-

tration - but the reduction in the denominator can yield - as an overall effect - an increase in the index R (see Frosini, 1985 *a*, § 5). This fact explains the contrasting behaviour of R and G, displayed in the tables in Frosini (1985 *b*, 1989); of course, only the behaviour of G is coherent with the Lorenz ordering.

When - as usual in most applications - the available frequency distribution is based on classes $C_j = (a_j,b_j]$, it is possible to establish the maximum of G, or an upper bound very near to the maximum. This is achieved by computing the maximum G_j under the hypothesis $C_j = [a_j,b_j]$, thus determining - without imposing integer values for the absolute frequencies - a distribution in C_j with frequency

$n_j(b_j - m_j)/(b_j - a_j)$ for a_j, and frequency $n_j(m_j - a_j)/(b_j - a_j)$ for b_j

(Frosini, 1984, p. 386; 1987, p. 180). By applying (4) to the frequency distribution with values a_j and b_j, and corresponding shares - concerning frequencies and quantities - f_{ja} and q_{ja} for a_j, f_{jb} and q_{jb} for b_j, one obtains:

$$G_j = q_{ja}(2F_{ja} - f_{ja}) + q_{jb}(2F_{jb} - f_{jb}) - 1$$

and after some algebra, taking into account that $f_{ja} + f_{jb} = q_{ja} + q_{jb} = 1$ and the above result concerning the frequency of a_j:

$$G_j = f_{ja} - q_{ja} = f_{ja}(1 - a_j/m_j) = \frac{b_j - m_j}{b_j - a_j} \times \frac{m_j - a_j}{m_j}. \tag{5}$$

This expression can be used as an upper bound for G_j, when $m_j = mq_j/f_j$ is known.

As sometimes happens, only the frequencies f_j for class C_j can be known. In this case, the following results may reveal useful. Calling $A_j = b_j - a_j$ the class width, if we assume a uniform distribution in each class, namely

$$x_{jr} = a_j + (2r - 1)A_j/2n_j \qquad r = 1, \ldots, n_j \tag{6}$$

the class mean coincides with the central point ($m_j = x_j = a_j + A_j/2$), and the Gini index for class C_j can be written (Frosini, 1987, p. 208):

$$G_j = \frac{n_j + 1}{n_j} \times \frac{A_j (n_j - 1)}{6 x_j n_j}; \tag{7}$$

thus the last term in (3) can be written as:

$$\sum_{j=1}^{k} q_j f_j G_j = \frac{1}{6nT} \sum_{j=1}^{k} (n_j + 1)(n_j - 1) A_j \tag{8}$$

and for large values of all n_j we can approximate by

$$\sum_{j=1}^{k} q_j f_j G_j \quad \approx \quad \frac{1}{6m} \sum_{j=1}^{k} f_j^2 A_j \; ; \tag{9}$$

problems can arise for the determination of the last width A_k, as the last class is usually open. These expressions may be used as sensible approximations of the Gini index.

An upper bound for G_j in this case (of ignoring m_j) may be derived from (5); for an easier reference, if we consider the function

$$g(x) = \frac{b-x}{b-a} \times \frac{x-a}{x} \quad 0 < a \le x \le b,$$

it is easy to show that it is maximized for $x = \sqrt{ab}$ (geometric mean of the class limits). Thus an upper bound for (5), allowing variation in m_j between a_j and b_j, is

$$G_j \quad = \quad \frac{b_j - \sqrt{a_j b_j}}{b_j - a_j} \times \frac{\sqrt{a_j b_j} - a_j}{\sqrt{a_j b_j}} \tag{10}$$

A simple example, which mimics the example displayed in Frosini (1987, p. 208), however with a lower limit of the first class greater than zero, exploits the following distribution of 15 subjects in four classes:

C_j	n_j	x_j	m_j	G_j (7)	G_j (5)	G_j (10)
2 ⊣ 10	2	6	4	0.1667	0.3750	0.3820
10 ⊣ 22	6	16	15	0.1215	0.1944	0.1946
22 ⊣ 34	4	28	30	0.0670	0.0889	0.1084
34 ⊣ 52	3	43	48	0.0620	0.0648	0.1058

By applying (3) and (8), the evaluation of G under the assumption of uniform distribution within the classes gives G = 1.2812 - 1 + 0.0244 = 0.3056. Under the hypothesis that the class means m_j are the values listed above, application of (3) and (5) gives: $\sup_I G$ = 1.3267 - 1 + 0.0335 = 0.3602. Under the hypothesis that the class means are unknown, application of (3) and (10) gives, by assuming m_j as the geometric mean of a_j and b_j:

$\sup_{II} G$ = 1.2998 - 1 + 0.0396 = 0.3394.

It must be noted that the last value $\sup_{II} G$ is determined by maximizing the intra-class concentration; but this simple fact cannot entail an overall maximization - for the given frequency distribution onto the k classes - as actually happens in the example.

3. THE PIETRA-RICCI INDEX APPLIED TO HISTOGRAMS

Another well known inequality measure is Pietra-Ricci index (cfr. Frosini, 1989, pp. 352-7; 2001, p. 148); by its definition with respect to n x_i values, it can be written:

$$P = \frac{1}{2T}\sum_{i=1}^{n}\left| x_i - m \right| \tag{11}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\left| q_i - \tfrac{1}{n} \right|$$

$$= \frac{1}{T}\sum_{x_i > m}(x_i - m) = \frac{1}{T}\sum_{x_i < m}(m - x_i) \ .$$

When a distribution in k groups or classes is considered, putting as before m_j and n_j the group means and absolute frequencies, $q_j = m_j n_j / T$, $f_j = n_j / n$, a very good approximation to (11) - for the reasons to be explained shortly - is usually given by

$$P_H = \frac{1}{2T}\sum_{j=1}^{k}\left| m_j - m \right| n_j = \frac{1}{2}\sum_{j=1}^{k}\left| q_j - f_j \right| . \tag{12}$$

While the Gini index does not possess any intrinsic or substantial significance, aside from a formal meaning as a normalized area in the Lorenz plot, the index P is recognized as the share of the total T that should be redistributed by the people possessing more than the mean towards the people possessing less than the mean, in order to reach perfect equality. On the other hand, although P satisfies the Pigou-Dalton criterion, it is less sensitive than G (or R); in fact, G always decreases by applying an egalitarian transfer between two individuals (Frosini, 1987, p. 190), while P is bound to remain unchanged when the egalitarian transfer takes place between individuals either over the mean or under the mean, without moving their position with respect to the mean after the transfer. Anyway, as a judgment about inequality of distributions is actually an *overall* judgment, being sensitive to *any* transfer seems - in itself - a doubtful property of an inequality measure; having a concrete and simple interpretation seems more important for such a measure.

As regards the computation of P on k groups of observations, the value obtained by application of (12) is usually a very precise "estimate" of the population value. It is immediately seen that, if m coincides with a class limit, the sum in the numerator of (11) remains unchanged when computed on the k classes; for example, let $m = b_h = a_{h+1}$ (upper limit of h-th class and lower limit of $(h + 1)$-th class); P can be written

$$P = \frac{1}{2T}\left\{\sum_{j=1}^{b}(m - m_j)n_j + \sum_{j=b+1}^{k}(m_j - m)n_j\right\}$$

$$= \sum_{j=1}^{b}(f_j - q_j) = \sum_{j=b+1}^{k}(q_j - f_j) \quad . \tag{13}$$

In the hypothesized situation, the computation of P on the k groups or classes yields exactly the same value computable when knowledge of all the n observations is available. It is also evident that when a reasonable number of classes is available ($k \geq 10$), a very good approximation is always achieved by computing P on the frequency-quantity distribution (f_1,q_1; ... ; f_k,q_k) by means of one of the formulas

$$P = \sum_{f_j > q_j}(f_j - q_j) = \sum_{f_j < q_j}(q_j - f_j) \tag{14}$$

which are exact formulas when there is no variation inside the values x of each group or class.

In the tables 1 and 2 in Frosini (1989) all the P values computed for $k = 50$ and 100 aggregate units are exactly equal - to the sixth significant figure - to the theoretical P values computed on a wide range of Pareto and Lognormal distributions.

A very interesting graph can be displayed, which shows the meaning of P as an area between the two relevant distributions, of frequencies and quantities; the reference is to Figure 1, showing the histograms of these distributions, for a survey on 3,000 families made by Banca d'Italia in 1980, derived from the data in Table 1.

TABLE 1

Distribution of 3,000 families in 15 income classes (Banca d'Italia Survey, 1980).
(Mean income m = 12.856 millions Lire)

Classes (Millions Lire)	% Families $100f_j$	% Income $100q_j$	$100F_j$	$100Q_j$
0 ⊣ 2	1.3	0.2	1.3	0.2
2 ⊣ 4	6.9	1.7	8.2	1.9
4 ⊣ 6	12.4	5.0	20.6	6.9
6 ⊣ 8	15.1	8.3	35.7	15.2
8 ⊣ 10	13.4	9.4	49.1	24.6
10 ⊣ 12	11.2	9.6	60.3	34.2
12 ⊣ 14	9.9	10.1	70.2	44.3
14 ⊣ 16	7.8	9.1	78.0	53.4
16 ⊣ 18	6.4	8.5	84.4	61.9
18 ⊣ 20	3.9	5.7	88.3	67.6
20 ⊣ 22	2.8	4.6	91.1	72.2
22 ⊣ 25	2.5	4.5	93.6	76.7
25 ⊣ 30	2.4	5.1	96.0	81.8
30 ⊣ 40	2.4	6.3	98.4	88.1
40 - ∞	1.6	11.9	100	100

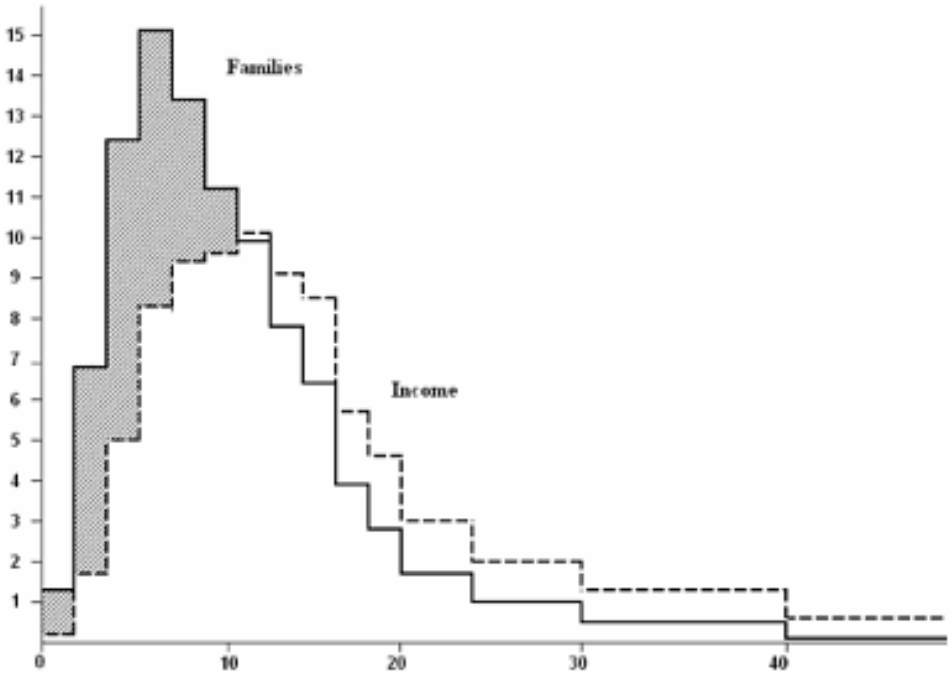

Figure 1 – Histograms for the distributions of frequencies and quantities, from data in Table 1.

Let us first consider k classes of equal width, which can be taken as unitary; both histograms for f_j and q_j comprise a total area of one; as

$$q_j = f_j\, m_j/m \quad ; \qquad q_j - f_j = n_j\,(m_j - m)/T,$$

for $m_j < m$ the inequality $q_j < f_j$ holds, whereas the opposite inequality $q_j > f_j$ takes place when $m_j > m$. As a result, when m coincides with a class limit, both sums (13) are equal to P, which means that the total area between frequencies and quantities - below or over the mean m - equals P. In case of classes of unequal width, after conventionally fixing the most frequent width equal to one, the ordinate for a class of width $w \neq 1$ must be calculated - as usual - by the division f_j/w, or q_j/w, thus ensuring a total area of one for the respective histogram.

Let us now assume that m is not a class limit (in which case $P_H = P$), and that m is included in the r-th class ($a_r < m < b_r$). Calling $\sum^r$ the sum extended to values in the r-th class, the r-th term of (12) can be written, when $m_r > m$:

$$D_r = q_r - f_r = \frac{1}{T}\sum{}^{r}(x_i - m) = \frac{1}{T}\left\{\sum\nolimits_{x_i<m}^{r}(x_i - m) + \sum\nolimits_{x_i>m}^{r}(x_i - m)\right\}$$

and when $m_r < m$:

$$D_r = f_r - q_r = \frac{1}{T}\sum{}^{r}(m - x_i) = \frac{1}{T}\left\{\sum\nolimits_{x_i<m}^{r}(m - x_i) + \sum\nolimits_{x_i>m}^{r}(m - x_i)\right\}$$

showing compensation between the two terms (one is negative, the other is positive). As a consequence, D_r is a lower bound of the exact contribution to P of the values in this class, which is given by

$$E_r = \frac{1}{T}\left\{\sum_{x_i<m}^{r}(m - x_i) + \sum_{x_i>m}^{r}(x_i - m)\right\},$$

whence the positive correction to D_r turns out, when $m_r > m$:

$$E_r - D_r = 2\sum_{x_i<m}^{r}\left(\frac{1}{n} - \frac{x_i}{T}\right) = 2[f^r(x_i<m) - q^r(x_i<m)] \tag{15}$$

being $f^r(\cdot)$ and $q^r(\cdot)$ - respectively - the relative frequencies and quantities in the r-th class; and when $m_r < m$:

$$E_r - D_r = 2\sum_{x_i>m}^{r}\left(\frac{x_i}{T} - \frac{1}{n}\right) = 2[q^r(x_i>m) - f^r(x_i>m)] \tag{16}$$

The worst case, in which D_r does not capture any contribution in E_r, happens when $q_r = f_r$, which is equivalent to $m_r = m$; in this case

$$E_r - D_r = 2\,[f^r(x_i < m) - q^r(x_i < m)] = 2\,[q^r(x_i > m) - f^r(x_i > m)]. \tag{17}$$

In order to get an upper bound for E_r (which usually turns out to be a good approximation), we can exploit a result from Frosini (1984, p. 392):

Among all distributions defined on the finite interval $[a,b]$ with a common mean $\mu = \gamma a + (1 - \gamma)b$, maximum concentration and maximum dispersion around the fixed percentile x_γ happen for the random variable which assumes the only values a and b, respectively with probabilities γ and $(1 - \gamma)$. Thus, from $\gamma a_r + (1 - \gamma)\, b_r = m_r$, one obtains

$$\gamma = (b_r - m_r)/(b_r - a_r). \tag{18}$$

From the data in Table 1, first of all we get the approximation to P given by the common value obtained by summation of all the positive differences $(f_j - q_j)$, or all the positive differences $(q_j - f_j)$, which is $P_H = 26.1\%$. Then, knowing that the general mean is $m = 12.856$ (included in the class 12 ⊣ 14), whence $m_r = q_r\, m/f_r = 10.1\times12.856/9.9 = 13.116$, from (18) we get $\gamma = 884/2000 = 0.442$; finally

$f^r(x_i < m) = 0.442\, f_r = 0.442\times0.099 = 0.04376$;
$q^r(x_i < m) = a_r\times0.04376/m = 0.0408$
$f^r(x_i < m) - q^r(x_i < m) = 0.003$;

the approximation thus obtained for P is then $P = 26.1\% + 0.3\% = 26.4\%$.

Anyway, as the term of (12) which needs to be approximated refers to the difference $|q_j - f_j|$ nearest to zero, a better approximation is practically of negligible interest.

TABLE 2

Deciles and income shares for tenths of families (Banca d'Italia Survey, 2002). (Mean income: Euros 27,868)

Classes (deciles)	Income share q_j	Cumulative Q_j	Mean income m_j
0 ⊣ 9,500	2.3	2.3	6,536
9,500 ⊣ 13,000	4.1	6.4	11,318
13,000 ⊣ 15,902	5.2	11.6	14.411
15,902 ⊣ 19,200	6.2	17.8	17,438
19,200 ⊣ 22.986	7.6	25.4	21,050
22,986 ⊣ 27,253	9.0	34.4	25,101
27,253 ⊣ 32,305	10.6	45.0	29,616
32,305 ⊣38,852	12.7	57.7	35,414
38,852 ⊣ 50,287	15.8	73.5	43.909
50,287 - ∞	26.5	100	73,831

A similar graphical and computational result can be achieved if we dispose of deciles, or tenths of both distributions - of frequencies and incomes. For example, from Table 2, by plotting on the abscissa the deciles, and comparing frequencies (always ten per cent) and income shares, a graph similar to Figure 1 could be obtained. Another possibility is displayed in Figure 2, where the abscissas do not take into account the deciles, but only the successive tenths of families (according to increasing incomes). Also in this plot, the total area for the tenths showing an inequality $f_j < q_j$ (which equals the total area of the tenths where $f_j > q_j$) is recognized to equal $P_H = 25.6\%$. In this case we know that $m = 27{,}868$ and $m_r = 29{,}616$, whence

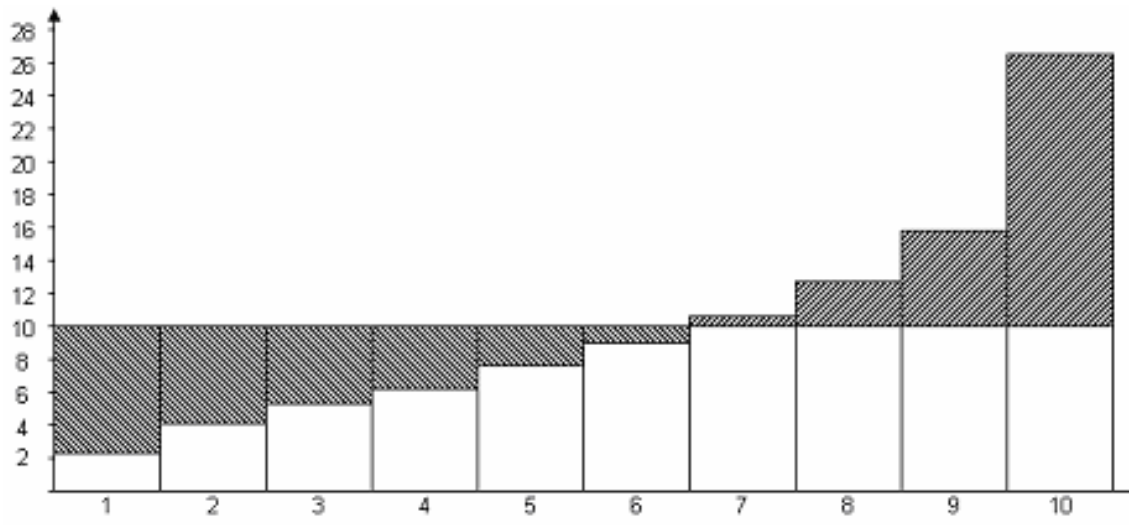

Figure 2 – Histograms for deciles in Table 2.

$$\gamma = (32{,}305 - 29{,}616)/(32{,}305 - 27{,}253) = 0.5323.$$

Finally,

$f^r(x_i < m) = 0.5323 \times 0.10 = 0.05323$; $q^r(x_i < m) = a_r \times 0.05323/m = 0.05206$;
$f^r(x_i < m) - q^r(x_i < m) = 0.00117$;

the approximation for P is then $P = 25.6\% + 0.1\% = 25.7\%$.

4. GINI-LIKE INDICES FOR GROUPS AND CLASSES

By reading the columns - in Table 1 - with cumulative percentages of families and incomes, and having in mind the original formula (1) devised by Gini (after normalization) for his concentration ratio, one could wonder whether a sum of the differences $(F_j - Q_j)$ (always ≥ 0 by construction) could yield some kind of concentration measure as well. A short investigation about this possibility will be now worked out, by separately examining the following cases:

(I) for comparison purposes (with other *comparable* distributions), the frequencies of the subsequent *groups* - ordered by income - are held fixed;

(II) always for comparison purposes, the income *classes* are held fixed.

In dealing with case I, the case of aggregate units - already pointed out above - is excluded, both for the large number of such units to be considered in practical applications, and mostly because the operational treatment of these units, by application of *any* concentration measure, exactly or approximately mirrors the treatment for individual units. Thus case I is described by the index

$$S = \sum_{j=1}^{k-1}(F_j - Q_j) = \sum_{j=1}^{k-1} F_j - \sum_{j=1}^{k-1} Q_j \tag{19}$$

where the frequencies F_j are fixed, while the cumulative income shares Q_j depend on the distribution of income among the k groups. Income transfers satisfy the Pigou-Dalton criterion, with no effect on (19) when increase and decrease of incomes take place within individuals of the same group, who remain in the same group also after the transfer. Normalization is simply obtained by division by ΣF_j:

$$T = 1 - \sum_{j=1}^{k-1} Q_j \Big/ \sum_{j=1}^{k-1} F_j \tag{20}$$

Simple calculations on the tables 1 and 2 will give an idea of the practical meaning, and the problems encountered, when using (20). In the case of Table 1, the relevant sums from 1 to $(k - 1)$ are: 875.2 for the cumulative frequencies F_j, 629 for the cumulative quantities Q_j, hence $T = 0.281$.

As the family shares in Table 1 are rather unequal, one could see what happens by aggregating some groups; if the first two classes are aggregated, and the same happens with the final classes, yielding new classes 16 ⊣ 20, 20 ⊣ 30, 30 - ∞, computation of (20) over the new ten classes yields

(1 - 329.9/506.4) = 0.349, quite far from the previous value 0.281, and near to the value of the Gini index G ($G = 0.372$ on the data of Table 1).

Concerning the data in Table 2, $T = 0.391$ (with ten classes), which raises to 0.418 with five classes, obtained by aggregation of neighbouring classes; the corresponding exact value for G (computed on individual data) is 0.359, whereas the approximation (4) gives 0.352.

As expected, the index T appears rather sensitive to the unequal distribution of the individuals among the groups; on the other hand, it can yield a reasonable approximation to the Gini index G when the number of groups (classes) is at least ten, and the shares of individuals in the k groups are not far from $1/k$.

Quite another approach - and meaning - ensues for case II, which demands that the k classes are defined and held fixed for all distributions to be compared. This case has been thoroughly examined by Frosini (1967) (where the index (19) is named g_2, and its normalization gives rise to the index φ_2). By a rearrangement of the terms, formula (19) yields:

$$S = \sum_{j=1}^{k-1}(k-j)\left(\frac{1}{n}-\frac{m_j}{T}\right)n_j \tag{21}$$

$$S = \sum_{j=1}^{k-1}(k-j)(f_j - q_j) \tag{22}$$

In a general setting, one could start with a maximum of constraints (concerning the set of comparable distributions), and then release one or more of them. With respect to (21) the initial set of constraints could be:

$n_1 + \ldots n_k = n$
$m_1 n_1 + \ldots + m_k n_k = T$
$0 \leq x_0 \leq m_1 < \ldots < m_k \leq x_\omega$
m_j = constant ($j = 1, \ldots, k$)
$n_j \geq 0$ ($j = 1, \ldots, k$).

The search for the minimum and the maximum of S under these constraints is a problem in linear programming; fortunately enough it is possible to get a general rule, which avoids formal recourse to linear programming (Frosini, 1967, §§ 4-6):
S is maximized when only the frequencies n_1 an n_k are positive; they are obtained from the system

$n_1 + n_k = n$
$m_1 n_1 + m_k n_k = T$;

min $S = 0$ when there exists m_j such that $m_j = m$; otherwise, S is minimized when the only frequencies n_h and n_l are positive, such that $m_h < m < m_l$, m_h being the greatest class mean lower than m, and m_l being the lowest class mean greater than m (n_h and n_l are obtainable by means of a system like the one above).

For the minimum we can always assume min $S = 0$, if we allow variation of m_j between the class limits, and the variable (e.g. income) is continuous or practically continuous. With a similar assumption for m_1 to lower as far as x_0, and for m_k to raise as

far as x_ω, the above system - with m_1 and m_k replaced, respectively, by x_0 and x_ω, can yield the required frequencies n_1 and n_k, when $x_\omega < T - (n - 1)x_0$; when $x_\omega \geq T - (n - 1)x_0$, x_ω is formally put equal to $T - (n - 1)x_0$, so that $n_1 = n - 1$ and $n_k = 1$.

If a solution for the *absolute* frequencies n_1 and n_k is required, the property of *integer values* for every n_j must be assumed in the set of constraints; the linear programming problem can thus be solved by Gomory's method (see e.g. Dantzig, 1963, p. 514). Also for this problem the general solution can be found, which avoids a formal recourse to Gomory's method; this solution coincides with the distribution of absolute frequencies which maximizes the Gini index G (cf. Frosini, 1967, § 5; 1987, pp. 178-180).

When, in the expressions for max S, n and T are allowed to vary, a very simple result for max S is obtained: max $S = k - 1$ for all the above sub-cases. Remembering that, with k aggregate units, max $A = (k - 1)/2$ (see formula (1)), we can expect that the index obtained by normalizing S by division by $(k - 1)$ is roughly half of G. This really happens in most cases. For example, with reference to the reduction of Table 1 to ten classes (see above in this section), $100\ S = 506.4 - 329.9 = 176.5$, so that $T = 0.1765$, which is roughly half of g and of its approximation given by T.

Istituto di Statistica BENITO V. FROSINI
Università Cattolica del Sacro Cuore di Milano

REFERENCES

BANCA D'ITALIA (1981). I bilanci delle famiglie italiane nell'anno 1980. *Bollettino*, Anno XXXVI, Numero Unico, pp. 539-607.

BANCA D'ITALIA (2004). I bilanci delle famiglie italiane nell'anno 2002. *Supplementi al Bollettino statistico*.

G.B. DANTZIG (1963). *Linear Programming and Extensions*. Princeton University Press.

B.V. FROSINI (1967). Sul concetto e sulla misura della concentrazione statistica. *La Camera di Commercio di Milano*,n. 5, pp. 38-59; n. 6, pp. 23-47.

B.V. FROSINI (1984). Concentration, dispersion, and spread: an insight into their relationship. *Statistica*, 44, pp. 373-394.

B.V. FROSINI (1985 *a*). Sugli ordinamenti di concentrazione e di variabilità. *Rivista di Statistica Applicata*, 18, pp. 121-141.

B.V. FROSINI (1985 *b*). Comparing inequality measures. *Statistica*, 45, pp. 299-317.

B.V. FROSINI (1987). *Lezioni di Statistica. Parte prima* (2nd Edition). Vita e Pensiero, Milano.

B.V. FROSINI (1989). Aggregate units, within-group inequality, and the decomposition of inequality measures. *Statistica*, 49, pp. 349-369.

B.V. FROSINI (2001). *Metodi Statistici*. Carocci, Roma.

C. GINI (1914). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti Regio Istituto Veneto*, Tomo 73, Parte II, pp. 1203-1248.

G. PIETRA (1915). Delle relazioni tra gli indici di variabilità. Nota I. *Atti Regio Istituto Veneto*, Tomo 74, Parte II, pp. 775-792.

U. RICCI (1916). L'indice di variabilità e la curva dei redditi. *Giornale degli Economisti e Rivista di Statistica*, Serie Terza, 53, pp. 177-228.

RIASSUNTO

Misure di diseguaglianza applicate a distribuzioni di frequenza

Mentre le misure di diseguaglianza o concentrazione sono definite con riferimento alla distribuzione di un fenomeno non negativo fra n individui, le concrete applicazioni sono spesso effettuate su distribuzioni di frequenza con k classi, ovvero su istogrammi, se si guarda alla rappresentazione grafica. Riguardo a questo tipo di applicazioni questo articolo considera: (1) la bontà dell'approssimazione ottenuta, da indici calcolati su k classi, per gli indici corrispondenti calcolati su n individui; (2) il significato e le proprietà delle misure di diseguaglianza che sono funzione delle sole frequenze e quantità pertinenti alle k classi. Queste due indagini sono state applicate a classici indici di concentrazione proposti da Gini e da Pietra-Ricci.

SUMMARY

Inequality measures for histograms

While inequality or concentration measures are defined with reference to the distribution of a non-negative character among n individuals, most practical applications are effected on frequency distributions over k classes, namely on histograms, when thinking of the corresponding graphical representation. Concerning this type of applications, this paper examines: (1) the goodness of approximation – to indices computed on individual data – of the same indices worked out on histograms; (2) the meaning and properties of inequality indices that are functions of the only frequencies and quantities pertaining to the k classes. These two kinds of investigations have been addressed to classical concentration measures proposed by Gini and Pietra-Ricci.