Empir Econ (2012) 43:175–197
DOI 10.1007/s00181-011-0464-1

Approximation and decomposition of Gini, Pietra–Ricci and Theil inequality measures

Benito V. Frosini

Received: 28 January 2010 / Accepted: 15 February 2011 / Published online: 24 March 2011

Abstract Two related problems are dealt with in this article, concerning some popular inequality indices proposed by Gini, Pietra–Ricci and Theil: (1) the calculation of the index when only a frequency distribution is available, thus needing some kind of approximation; and (2) a reasonable decomposition of the index calculated for a mixture, with components related to 'within' and 'between' inequalities, and possibly to the separate contributions of each group to the overall inequality. Beside the proposals arising from the specific structure of each inequality index, a general approach for identifying the within component is utilized, which is based on the fixation of a given number of fictitious individuals (called *aggregate units*), common to every group. Regarding the Gini index, a general expression is obtained for the approximation problem, while the within inequality is more easily managed by the recourse to aggregate units. The decomposition of the Pietra–Ricci index displays three components, clearly ascribable to within inequality, to a mixture effect and to a mean effect. Regarding the Theil index, some simple and very accurate approximation formulae are obtained. An application of all the indices and their decompositions has been made for the 2004 income distribution for Italy (Bank of Italy Survey).

Keywords Income inequality · Gini, Pietra–Ricci and Theil indices · Decomposition of inequality measures

JEL Classification C46 · D31

B. V. Frosini (✉)
Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore,
Largo Gemelli, 1, 20123 Milan, Italy
e-mail: benito.frosini@unicatt.it

1 A short outline of income inequality measures and group decompositions

The motivation for this article stems from recent renewed interest in the proposal of specialized measures of income (or wealth) inequality, and in their decomposition according to groups of individuals (or households). Amongst the latest methodological proposals, at least Zenga's new inequality index based on the ratios between lower and upper arithmetic means, and its group decomposition obtained by Radaelli, are worth mentioning (Zenga 2007; Radaelli 2008), amongst the most meaningful applications and discussions of group decomposition, at least the research by Quintano et al. (2009) on the evolution of income inequality in Italy deserves careful attention for the achieved results and insightful suggestions. Another motivation is related to a decomposition of Pietra–Ricci index, obtained by the author, starting from a proposal by Frosini (2005), aimed at providing a new interpretation for this index, as well as approximation formulae for the common situation of using a frequency, or a frequency–quantity, distribution over k classes.

Actually, approximation problems and group decompositions reveal an important overlapping: in fact, the k classes of a frequency distribution can be viewed (formally) as k groups; an index approximation obtained by the distribution of partial means can be assimilated to a special case of *between* inequality; and the improvement to this first approximation, taking into account of the dispersion of values within each class, is tantamount to add a measure for *within* inequality. Anyway, both problems, of approximation and decomposition, will be distinctly treated for three indices, perhaps the best known and generally applied amongst the many available: Gini, Pietra–Ricci and Theil indices.

The study of group decomposition of inequality measures has followed two main approaches. A first approach has tried to identify those inequality indices I whose functional form (defined with respect to individual incomes) can be split—when applied to a set S of individuals, constituted by k groups $S_j (j = 1, \ldots, k)$—as

$$I = I_B + I_W \tag{1}$$

where I_B and I_W maintain the same functional form as I, I_B being interpreted as a function of the inequality *between* groups (applied to the distribution of partial means, not necessarily arithmetic), and I_W being some kind of average of the inequality *within* groups (see Theil 1967; Bourguignon 1979; Cowell 1980; Shorrocks 1980). The indices thus obtained pertain to a very restricted set; Theil's index will be examined in the sequel.

It is easily acknowledged that the above additive decomposition cannot be a property of normalized indices, such as those of Gini and Pietra–Ricci (Zenga 1987, p. 336; Frosini 1989, p. 350). A second approach can thus be specialized in one of the following four sub-approaches: (a) to compute an index I_B with the above meaning, leaving I_W as a residual; (b) to compute I_W with the above meaning, leaving I_B as a residual; (c) to jointly compute I_B and I_W, however without imposing any functional form, nor a mathematical expression in closed form; and (d) to give up the above meanings, established both for I_B and I_W, and to suggest an interpretation for a group

decomposition derived by a clever rewriting of I as the sum of two or more terms, with appropriate meanings.

The (a) approach was advocated by Bhattacharya and Mahalanobis (1967, p. 150; Zagier 1983, p. 104) for the Gini ratio, although strongly opposed by Dagum (2001, pp. 31–32). The (b) approach was propounded by Frosini (1985, 1989), within a general proposal for the computation of any inequality measure based on a fixed number of units, the so-called aggregate units (or AUs), to be defined shortly. The (c) approach has found to be a very neat proposal by means of the Shapley decomposition, a general procedure which allows to additively disentangle the contributions of one or more classification factors (e.g. stratification by geographic districts, age, sex, ethnicity, etc.) to total inequality; the basic reference concerning the Shapley decomposition is Shorrocks (1999) (see also Deutsch and Silber 2007, 2008). Amongst the contributions to the (d) approach, applied to the decomposition of the Gini index, at least those of Silber (1989), Yitzhaki and Lerman (1991), Dagum (1997a,b, 2001) and Okamoto (2009), are worth mentioning. By a recourse to a new linear operator, called the G-matrix, Silber displays a three-term decomposition, where the interaction term (besides the 'within' and 'between' terms) 'is a measure of the intensity of the permutations which occur when instead of ranking all the individual shares by decreasing income shares, one ranks them, firstly by decreasing value of the average income of the population subgroup to which they belong, and secondly, within each subgroup, by decreasing individual income share' (Silber 1989, p. 112). By starting with a simple expression of the Gini index, based on the covariance between income values and their ranks, Yitzhaki and Lerman have obtained a three-term decomposition, where the interaction term 'reflects the impact of stratification, or of intra-group variability in overall ranks' (Yitzhaki and Lerman 1991, p. 322). The authors show that the 'between' term—depending on a covariance—could take negative values, although in the presence of quite heterogeneous groups. The three term decomposition by Dagum, and the two-term decomposition by Okamoto, will be commented on at the next section, being more relevant for the topics there presented (and also to avoid repetitions).

Frosini's proposal of AUs was motivated by the well-known fact that the computation of (practically) any inequality measure on theoretical distributions, or on observed distributions of several thousands of individuals, gives about the same result—even to two significant figures—as the same index computed on few dozens of appropriately spaced values (Frosini 1985, p. 308). Of course, by suppressing the dispersion inside each class leads to a reduction of total inequality (for a simple example, concerning the Gini index, see Fei et al. 1979, pp. 403–405). However, in the cases just outlined, the negative bias is usually quite negligible (see Frosini 1985, for many computations concerning real and theoretical distributions, and number of classes $n = 10, 25, 50$). This fact affords the possibility of making use of a fixed number of AUs, e.g. 50 or 100, when several distributions are jointly examined. If one disposes of the Lorenz function L for a given population and wants to determine the unitary shares for m AUs, then the m income shares are simply obtained as

$$q_r = L\,(r/m) - L\,((r-1)/m) \quad r = 1, \ldots, m \tag{2}$$

If T is the total income, then the corresponding absolute share is $x_r = q_r T$.

An immediate application of AUs is related to the case of a population composed of k groups (usually with different sizes). In this case, the overall computation of the AUs gives rise to the following matrix:

$$\begin{matrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{k1} & x_{k2} & \dots & x_{km} \end{matrix} \tag{3}$$

having the k groups (all with m units) on the rows, and the successive shares on the columns. When the k groups are not—as usual—of the same size, and have indeed sizes $n_1, \dots, n_k (\sum n_j = n)$, with total income T_j for group S_j, this same reference can be maintained by providing n_j replicas for group $j (j = 1, \dots, k)$, so scaled as to yield the total $\mu_j = T_j / n_j$ for each replica concerning the group S_j. Take note that all indices here examined are scale independent, and also independent of the size of population (or ISP), in the sense that an index I remains unchanged when applied to a mixture of k identical populations. Note also that the device of replicas is just aimed at ensuring the same group weights $f_1, \dots, f_k$ in matrix (3) as those existing in the reference population.

Once the above matrix is available, a single distribution of values

$$y_r = (x_{1r} + \dots + x_{kr}) / k \tag{4}$$

in case of equally sized groups, or of values

$$y_r = \frac{1}{n} \sum_{j=1}^{k} x_{jr} n_j \tag{5}$$

in case of groups with different sizes, provides a distribution suitable for the appreciation of the comprehensive *within* inequality, thus leaving I_B in (1) as a residual. The distribution of these values—the so-called percentile distribution—or the set of these values, will be called S_P. Some justifications, and several applications of this procedure, are provided by Frosini in the articles of 1989, 1990a and 2003. The special applications to Gini, Pietra–Ricci and Theil indices are presented at Sect. 2.

The calculation of an inequality index with respect to the distribution of partial means (of classes constituting a single frequency distribution) is a special case of 'between inequality'; however, to avoid confusion, we will use a suffix M, as in G_M, for an approximation of an inequality index G based on the distribution of h partial means, and a suffix B, as in G_B, for the evaluation of the 'between inequality' existing in case of a mixture constituted of k groups (or distributions).

Most symbols about the distribution of values $0 \le x_1 \le x_2 \le \dots \le x_n$ in the population, and of values $0 \le x_{j1} \le x_{j2} \le \dots \le x_{jn_j}$ for the jth group, are usefully listed as follows:

$T = \sum x_i$ = total of population S (or mixture S); $T_j = \sum_r x_{jr}$ = total of group S_j
n = total size of population S; n_j = size of group S_j

$\mu = T/n =$ general mean; $\mu_j = T_j/n_j =$ mean of group S_j
$q_i = x_i/T =$ share of ith individual in the population
$q_{jr} = x_{jr}/T_j =$ share of rth individual of group S_j
$f_j = n_j/n =$ frequency weight of group S_j; $q_j = T_j/T =$ quantity share of group S_j.

As the reader has probably surmised, owing to the many quotations of this same author appearing above, this is only the last investigation about inequality measures, following many others, which started on studies about formal properties of variability, inequality and diversity measures (only slightly touched in this article, and thus without specific quotations). The previous proposals and results, resumed in this article, are concerned with (a) the group decomposition of normalized indices, (b) the computation of inequality indices through the definition of a fixed number of AU, and (c) the consequent proposal of deriving the 'within inequality' from the distribution of AUs in the groups; (d) the special application of this procedure to particular indices (some cases are reported at Sect. 2); and (e) the special and simple characterization of Pietra–Ricci index, both for its algebraical and graphical representations (at Sect. 2.2). The novelty of this article concerns with (1) the comparative examination of several recent proposals about decomposition, (2) some procedures for accurate approximations of indices based on frequency–quantity distributions, and mostly (3) the three-term decomposition of Pietra–Ricci inequality measure; and (4) moreover, an application of the above procedures is implemented to the 2004 Bank of Italy Survey on household incomes.

2 Computation and decomposition of Gini, Theil and Pietra–Ricci indices

2.1 The case of Gini index

Amongst the many computation formulae, devised by Gini and other authors, for the Gini index, we only quote the following ones, as they are functional with respect to subsequent considerations:

$$G = \Delta/2\mu \quad \text{with } \Delta = \sum_{i,j=1}^{n} |x_i - x_j|/n^2 \tag{6}$$

being Δ the mean absolute difference (with repetition) of the statistical variable, or random variable, which describes the incomes in the reference population;

$$G = \frac{2}{nT}\sum_{i=1}^{n} i x_i - \frac{n+1}{n} = \frac{2}{n}\sum_{i=1}^{n} i q_i - \frac{n+1}{n} \tag{7}$$

A relevant graphical correspondence, widely used in the literature, is the double of the concentration area in the representation of the Lorenz curve in a square of side one.

Actually, the original definition of Gini's concentration index R (Gini 1914) was slightly different, being based on the mean absolute difference *without* repetition; with

this definition, the index R has the advantage of being exactly normalized between zero and one, with $R = 1$ achieved when one individual amongst n gets all the income T (all other $n - 1$ values being zero). However, (a) G is practically equivalent to R in most applications, as $G = (n - 1)R/n$, and mostly (b) G has the ISP property, while R maintains some dependence on the size n of population, thus being unsuitable to manage mixture distributions (Frosini 1987, p. 192).

The computation of (6) and (7) (as well as for other equivalent formulae) is simple and direct, thus no problem can arise if we dispose of all the original values x_i. However, when applying to official statistics, usually summarizing several hundreds or thousands of individual values in few classes, we can only manage the frequencies (and sometimes the quantities too) pertaining to the given classes. In this case, we must get the information concerning the accuracy of the approximation, for a given inequality index, based on the available frequency (or frequency–quantity) distribution, over h classes. Luckily enough, this information is rather simple and useful in the case of the Gini index. If the data x_i are organized in h groups or classes, so that $x_{jr} (j = 1, \ldots, h; r = 1, \ldots, n_j)$ is the rth ordered value in the class $C_j = (a_j, b_j]$, then $q_{jr} = x_{jr}/T$, and

$$G_j = \frac{2}{n_j} \left\{ \sum_{r=1}^{n_j} r \frac{q_{jr}}{q_j} - \frac{n_j + 1}{n_j} \right\} \tag{8}$$

is the Gini index for the values inside the class C_j, after some algebra one obtains

$$G = \sum_{j=1}^{h} q_j \left(2F_j - f_j\right) - 1 + \sum_{j=1}^{h} q_j f_j G_j = G_{\mathrm{M}} + G_{\mathrm{C}} \tag{9}$$

where F_j is the jth cumulative frequency for the population (Frosini 1987, p. 207; for different but equivalent expressions for G_{C} see Gastwirth 1972, p. 309; Kakwani 1980, p. 100). The second summation in this formula

$$G_{\mathrm{C}} = \sum_{j=1}^{h} q_j f_j G_j \tag{10}$$

is the required 'correction' to the Gini index

$$G_{\mathrm{M}} = \sum_{j=1}^{h} q_j (2F_j - f_j) - 1 \tag{11}$$

computed on (q_j, f_j) values of the h classes, with the implicit assumption of no dispersion—within every class C_j—around its mean value μ_j (practically, G_{M} is the Gini index computed on the frequency distribution $(\mu_1, n_1; \ldots; \mu_h, n_h)$. The above correction G_{C} is a linear combination of the Gini indices computed on the h classes, with weights $q_j f_j$ (not summing to one); actually, when h is as low as ten, these

weights are rather low, and the increase to G_{M} represented by this term is practically irrelevant (cf. Frosini 1985, 1989). Anyway, something more can be said, in general, on the computation of this correction. If the average $\mu_j = \mu q_j / f_j$ of class $C_j = (a_j, b_j]$ is known, then an upper bound for G_j (with respect to distributions inside the class C_j with the same mean μ_j) can be written

$$G_j^B = \frac{b_j - \mu_j}{b_j - a_j} \times \frac{\mu_j - a_j}{\mu_j} \tag{12}$$

(Frosini 2005, p. 30; for a different but equivalent expression see Gastwirth 1972, p. 308). Another approximation is offered by the assumption (requiring only the knowledge of the frequency distribution over the h classes) that the spread of the x_{jr} values of class C_j is uniform within the class—with width $A_j = b_j - a_j$—namely

$$x_{jr} = a_j + (2r - 1)\, A_j / 2n_j \quad r = 1, \ldots, n_j \tag{13}$$

(Frosini 2009, p. 92). In this case, the Gini index for the class C_j can be written as

$$G_j^U = \frac{n_j + 1}{n_j} \times \frac{A_j (n_j - 1)}{6 x_j n_j} \tag{14}$$

Thus, the correction term G_{C} can be written as

$$G^U = \frac{1}{6nT} \sum_{j=1}^{h} \left(n_j + 1\right) \left(n_j - 1\right) A_j. \tag{15}$$

For large values of n_j, G^U can be approximated by

$$G^A = \frac{1}{6\mu} \sum_{j=1}^{h} f_j^2 A_j. \tag{16}$$

This kind of approximation is usually better than the previous approximation, based on upper bounds, except for the last class (and sometimes, for the first class as well). A special approximation problem usually arises for the last class, be it closed, or (more often) open; it usually happens that the range of incomes in this class is much larger than the range in the preceding classes, and the uniformity assumption (13) is untenable (the frequencies decrease steadily as the income increases). A reasonable and easy device to manage this problem, and getting an accurate approximation, consists in assuming an interpolating Pareto distribution, with cumulative distribution function

$$F(x) = 1 - (\tau / x)^{\theta} \qquad x \geq \tau > 0; \quad \theta > 0$$

having mean $\tau\theta/(\theta-1)$ when $\theta > 1$. If one disposes of the moment estimator of θ, for $\theta > 1$ (when the frequency *and* the quantity of the last class C_h are available),

$$\hat{\theta} = \overline{X}/(\overline{X} - \tau) \tag{17}$$

with $\overline{X}$ the mean of the last class and τ its left endpoint, then the Gini index

$$G_h = 1/(2\hat{\theta} - 1) \tag{18}$$

for this class can thus be approximated (Frosini 1985, p. 311; 1990b, p. 234). Some examples of computations of the above approximations will be given at Sect. 3.

Turning to the general decomposition problem, the case for the Gini index appears rather awkward; the essential difficulty lies in the fact that G is a function of all the *ordered* values: a mixture of two (or more) *overlapping* distributions yields a general index G (for the mixture) with a complicated structure, which makes it difficult to disentangle the 'within' and 'between' contributions. Besides the results by Silber, and Yitzhaki and Lerman, summarized in the previous section, by a clever development of formula (6) Dagum (1997a,b, 2001) obtains a three-term decomposition

$$G = G_{\mathrm{w}} + G_{\mathrm{b}} + G_{\mathrm{t}} \tag{19}$$

where $G_{\mathrm{w}} = G_{\mathrm{C}}$ above (formula (10)), G_{b} is a term depending on the respective displacements of the partial distributions, and G_{t} is amenable to the transvariation between groups, or partial distributions (Dagum 1997a,b, 2001; Deutsch and Silber 1997; Quintano et al. 2009). If G_{w}, so defined, *could* be interpreted as 'la contributión a la desigualdad intra (*within*) grupos' (Dagum 2001, p. 38), then one could take the remaining sum $G_{\mathrm{b}} + G_{\mathrm{t}}$ as providing an overall contribution to the *between* inequality (differences between group means, as well differences in dispersion and skewness). However, the large contribution usually amenable to transvariation, and—mostly—the small contribution of G_{w} to the overall inequality, cast serious doubts on the interpretation of G_{w} as the *within* contribution, and—correspondingly—on the interpretation of $(G - G_{\mathrm{w}})$ as the *between* contribution (Quintano et al. 2009, pp. 439–440). As G_{w} in formula (19) corresponds to the correction term G_{C} in formula (9) *in all cases*, independently of any kind of overlapping between partial distributions, it could be reasonable to impute a (large? total?) part of G_{t} to within inequality.

Although acknowledging the insightful explanations—made by Dagum in the quoted articles—about the economic meaning of the various components of the sum (19), the within inequality $G_{\mathrm{w}} = G_{\mathrm{C}}$ appears largely unsatisfactory for several reasons: (a) as previously shown by Frosini (1987), and resumed above, G_{C} is the contribution to the overall G—beside the G_{M} contribution, only dependent on the spread of the partial means—of the inequality inside k *nonoverlapping* groups: how can we accept the *same* measure for the usual cases of largely overlapping groups, and even for the case of a mixture S with identical components S_j? For example, in a case of ten identical groups, each one with Gini index G, with $f_j = q_j = 1/10$, $G_{\mathrm{w}} = (1/10)G$, whereas a reasonable weighted average of the partial indices would result in $I_{\mathrm{W}} = G$ (cf. the following formula (20)). (b) G_{w} heavily depends on the number k of groups; as

k grows—other things being equal—G_{w} steadily diminishes; the reduction is usually heavy when passing from $k = 2$ to $k = 3, 4, 5$; from $k = 10$ onwards it reaches small values, however always decreasing towards zero—with what justification?

Turning to other proposals concerning the additive decomposition (1), let us pay our attention to the exclusive proposals: (a) define I_{B} by means of the given inequality measure applied to the distribution of partial means, leaving I_{W} as the residual $I_{\mathrm{W}} = I - I_{\mathrm{B}}$ (Bhattacharya and Mahalanobis 1967; Zagier 1983); (b) define I_{W} by means of the mixture of AUs, as suggested at Sect. 1, leaving I_{B} as the residual $I_{\mathrm{B}} = I - I_{\mathrm{W}}$; with this approach the Gini index applied to the AU distribution turns out amazingly simple and informative, namely the weighted mean of the partial Gini indices (Frosini 1989, p. 361):

$$I_{\mathrm{W}} = G\left(S_P\right) = \sum_{j=1}^{k} q_j G_j. \tag{20}$$

Interestingly enough, this simple and reasonable expression for I_{W} turns out to be the same obtained by Okamoto (2009, p. 155), who derived the following two-term decomposition of the Gini index, being F_j the distribution function of the jth group, and F the (mixture) distribution function for the entire population:

$$G = \sum_{j=1}^{k} q_j G_j + \sum_{j=1}^{k} f_j \frac{1}{n} \int \left(F_j(y) - F(y)\right)^2 \mathrm{d}y$$

As stressed by Okamoto, the second term in this formula is not only a residual, but is a clear expression of the between-group inequality; moreover, it is null if and only if all distributions $F_j (j = 1, \ldots, k)$ coincide.

2.2 The case of Pietra–Ricci index

The definition of the Pietra–Ricci index (Pietra 1915; Ricci 1916) is very simple, and immediately operational:

$$P = \frac{1}{2T} \sum_{i=1}^{n} |x_i - \mu| = \frac{1}{2} \sum_{i=1}^{n} |q_i - 1/n| \tag{21}$$

$$P = \frac{1}{T} \sum_{x_i > \mu} (x_i - \mu) = \frac{1}{T} \sum_{x_i < \mu} (\mu - x_i) \tag{22}$$

It is well known that this index satisfies the *weak principle of transfers* (WTP), and not the *strong principle of transfers* (SPT) as well (Marshall and Olkin 1979, p. 13; Castagnoli and Muliere 1990, pp. 174–175), because it does not change when non-egalitarian transfers take place only before the mean or after the mean. However, this fact should not be over-emphasized, because what is really decisive is an expected and reasonable behaviour of the index when applied to regular cases of real or theoretical

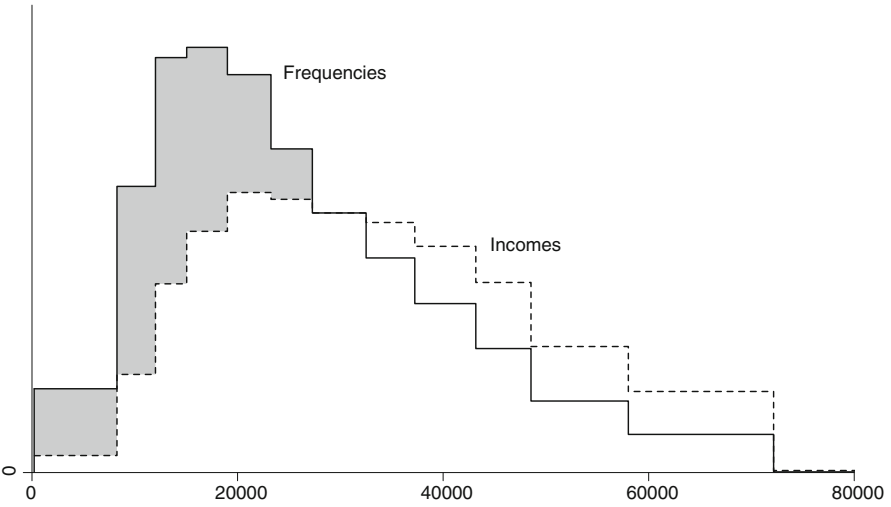

Fig. 1 Histograms for the distribution of frequencies and quantities, from data in Table 1 (Italy)

income distributions (e.g. Frosini 1989). Moreover, unlike the other main formulae for the measurement of income (or wealth) inequality, which lack any immediate economic meaning, the Pietra–Ricci index is clearly 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 achieve perfect equality. Besides, also the graphical meaning of the index P is more direct than the one relating to Gini's index: in fact, it is half the area between the two relevant distributions, of frequencies and quantities (Frosini 2005, pp. 33–36); an example is given in Fig. 1.

As to the operational application of P to frequency–quantity distributions, it is even simpler than the one exposed for the Gini index. In fact, the following equivalent formulae, which are based only on the frequencies and quantities of h classes (or non-overlapping groups),

$$P_H = \frac{1}{2T}\sum_{j=1}^{h}|\mu_j - \mu|n_j = \frac{1}{2}\sum_{j=1}^{h}|q_j - f_j| \tag{23}$$

$$P_H = \sum_{q_j > f_j}(q_j - f_j) = \sum_{f_j > q_j}(f_j - q_j) \tag{24}$$

give a very precise approximation of the population value, at least when (as usual) $h \geq 10$, with properly spaced classes. It is easily checked that a difference between the exact values (21)–(22) and the approximate values (23)–(24) is related only with the class containing the general mean μ (whereas there is coincidence when μ is equal to a class endpoint (Frosini 2005, pp. 32–35). An upper bound for the correction to be applied to formulae of P_H is provided by Frosini (2005, p. 35); however, making use of the assumption of uniform dispersion within the class containing μ, it is quite easy and immediate to achieve a very good approximation (as exemplified at Sect. 3).

Anyway, it must be admitted that P_H itself provides a sufficiently good approximation, when $h \geq 10$. Needless to say, the computation of P is wholly devoid of the problems which heavily affect most other inequality measures, concerning the approximations depending on the spread of incomes in the last class C_h.

Also the decomposition problem appears rather simple for the Pietra–Ricci index. Let us start with the simple case of two (non-degenerate) groups ($k = 2$), with respective means μ_1 and μ_2, and general mean μ. Necessarily $\mu_1 < \mu < \mu_2$, thus we are allowed to write, with respect to group values $x_{jr} (j = 1, 2; r = 1, \ldots, n_j)$:

$$
\begin{aligned}
TP &= \sum_{x_i > \mu} (x_i - \mu) = \sum_{x_{1r} > \mu} (x_{1r} - \mu) + \sum_{x_{2r} > \mu} (x_{2r} - \mu) \\
TP &= \sum_{x_{1r} > \mu} (x_{1r} - \mu_1) + \sum_{x_{1r} > \mu} (\mu_1 - \mu) + \sum_{x_{2r} > \mu} (x_{2r} - \mu_2) + \sum_{x_{2r} > \mu} (\mu_2 - \mu) \\
&= \sum_{x_{1r} > \mu_1} (x_{1r} - \mu_1) - \sum_{\mu_1 < x_{1r} \leq \mu} (x_{1r} - \mu_1) + \#(x_{1r} > \mu)(\mu_1 - \mu) \\
&\quad + \sum_{x_{2r} > \mu_2} (x_{2r} - \mu_2) + \sum_{\mu < x_{2r} \leq \mu_2} (x_{2r} - \mu_2) + \#(x_{2r} > \mu)(\mu_2 - \mu)
\end{aligned}
$$

giving rise to (letting P_j the corresponding index for group S_j, and $q_j = T_j / T$):

$$
\begin{aligned}
P = q_1 P_1 + q_2 P_2 - \frac{1}{T} \left\{ \sum_{\mu_1 < x_{1r} \leq \mu} (x_{1r} - \mu_1) + \sum_{\mu < x_{2r} \leq \mu_2} (\mu_2 - x_{2r}) \right\} + \\
+ \frac{1}{T} \left\{ \#\left(x_{1r} > \mu\right) \left(\mu_1 - \mu\right) + \#\left(x_{2r} > \mu\right) \left(\mu_2 - \mu\right) \right\}. \qquad (25)
\end{aligned}
$$

Some interesting features of this decomposition (anyway, a special case of the general formula (26) that will follow), are already worth mentioning. The first two terms of (25) give the contribution to P of the average group index, with weights given by the respective income shares; this sum is clearly interpretable as the 'within inequality':

$$
P_{\mathrm{W}} = q_1 P_1 + q_2 P_2.
$$

With this interpretation, the residual $P - P_{\mathrm{W}}$ is amenable to all other kinds of differences between groups; thus we could define the 'between inequality' by $P_{\mathrm{B}} = P - P_{\mathrm{W}}$.

In general, with k groups, the corresponding decomposition of TP, of the same kind of formula (25), turns out as follows:

$$
\begin{aligned}
TP &= \sum_{j=1}^{k} \sum_{x_{jr} > \mu} (x_{jr} - \mu) = \sum_{j=1}^{k} \left\{ \sum_{x_{jr} > \mu} (x_{jr} - \mu_j) + \sum_{x_{jr} > \mu} (\mu_j - \mu) \right\} \\
TP &= \sum_{\mu_j < \mu} \left\{ \sum_{x_{jr} > \mu_j} (x_{jr} - \mu_j) - \sum_{\mu_j < x_{jr} \leq \mu} (x_{jr} - \mu_j) + \sum_{x_{jr} > \mu} (\mu_j - \mu) \right\}
\end{aligned}
$$

$$+\sum_{\mu_j=\mu}\sum_{x_{jr}>\mu}(x_{jr}-\mu)$$

$$+\sum_{\mu_j>\mu}\left\{\sum_{x_{jr}>\mu_j}(x_{jr}-\mu_j)+\sum_{\mu<x_{jr}\le\mu_j}(x_{jr}-\mu_j)+\sum_{x_{jr}>\mu}(\mu_j-\mu)\right\}. \tag{26}$$

Before presenting another, more interesting decomposition expression, let us make some observations on formula (26). If all groups have the *same mean* μ (possibly with identical or different dispersion or skewness), then *TP* equals the second row of (26), and

$$P = P_{\mathrm{W}} = \sum_{j=1}^{k} q_j P_j$$

quite a reasonable result. On the other hand, having started with the operational formula for P given by the left hand side in (22), something unsatisfactory appears from the reading of the k rows (row j for group S_j) in the analytical development of formula (26). In fact, for $\mu_j < \mu$, the contribution of group S_j to the total *TP* is

$$T_j P_j - \sum_{\mu_j<x_{jr}\le\mu}(x_{jr}-\mu_j)+\sum_{x_{jr}>\mu}(\mu_j-\mu);$$

the first sum is positive and the second negative, thus—in any case—$T_j P_j$ is lessened, even to zero. This is algebraically reasonable, as some groups could make no contribution to values greater than the general mean μ; however, all reference to the group inequalities P_j has been lost, if one only looks at the group contributions in every row of (26). It is possible to maintain such a reference—and the corresponding information—if we start with the very definition of P, given by formula (21). Following the same kind of development which has led to formula (26), we can obtain the equivalent formula:

$$\begin{aligned}
2TP &= \sum_{j=1}^{k}\sum_{r=1}^{n_j}|x_{jr}-\mu| = \sum_{j=1}^{k}\left\{\sum_{x_{jr}>\mu}(x_{jr}-\mu)+\sum_{x_{jr}<\mu}(\mu-x_{jr})\right\}\\
&= \sum_{j=1}^{k}\left\{\sum_{x_{jr}>\mu}(x_{jr}-\mu_j)+\sum_{x_{jr}>\mu}(\mu_j-\mu)+\sum_{x_{jr}<\mu}(\mu-\mu_j)+\sum_{x_{jr}<\mu}(\mu_j-x_{jr})\right\}\\
&= \sum_{\mu_j<\mu}\left\{2T_jP_j-2\sum_{\mu_j<x_{jr}\le\mu}(x_{jr}-\mu_j)+\left[\#(x_{jr}<\mu)-\#(x_{jr}>\mu)\right](\mu-\mu_j)\right\}\\
&\quad+\sum_{\mu_j=\mu}2T_jP_j
\end{aligned}$$

$$+\sum_{\mu_j>\mu}\left\{2T_jP_j+2\sum_{\mu<x_{jr}\le\mu_j}(x_{jr}-\mu_j)\right.$$
$$\left.+\left[\#(x_{jr}>\mu)-\#(x_{jr}<\mu)\right](\mu_j-\mu)\right\}$$

and also

$$P=\sum_{j=1}^{k}q_jP_j+\sum_{\mu_j<\mu}\left\{-\frac{1}{T}\sum_{\mu_j<x_{jr}\le\mu}(x_{jr}-\mu_j)\right.$$
$$\left.+\frac{1}{2T}\left[\#(x_{jr}<\mu)-\#(x_{jr}>\mu)\right](\mu-\mu_j)\right\}$$
$$+\sum_{\mu_j>\mu}\left\{\frac{1}{T}\sum_{\mu<x_{jr}\le\mu_j}(x_{jr}-\mu_j)+\frac{1}{2T}\left[\#(x_{jr}>\mu)-\#(x_{jr}<\mu)\right](\mu_j-\mu)\right\} \tag{27}$$

Reconsidering the contribution to the inequality in the mixture S of group S_j, such contribution is equal to

$$D_j=q_jP_j-\frac{1}{T}\sum_{\mu_j<x_{jr}\le\mu}(x_{jr}-\mu_j)+\frac{1}{2T}\left[\#(x_{jr}<\mu)-\#(x_{jr}>\mu)\right](\mu-\mu_j) \tag{28}$$

if $\mu_j<\mu$;

$$D_j=q_jP_j \quad \text{if} \quad \mu_j=\mu \tag{29}$$
$$D_j=q_jP_j+\frac{1}{T}\sum_{\mu<x_{jr}\le\mu_j}(x_{jr}-\mu_j)+\frac{1}{2T}\left[\#(x_{jr}>\mu)-\#(x_{jr}<\mu)\right](\mu_j-\mu) \tag{30}$$

if $\mu_j>\mu$.

Notice that the median terms in (28) and (30) are negative, while the third terms can be $>=<0$. In conclusion, the decomposition of the Pietra–Ricci index P results as follows:

$$P=P_{\mathrm{W}}+P_{\mathrm{Bt}}+P_{\mathrm{Bm}} \tag{31}$$

where

$$P_{\mathrm{W}} = \sum_{j=1}^{k} q_j P_j \tag{32}$$

is the *within inequality*;

$$P_{\mathrm{Bt}} = \frac{1}{T}\left\{ \sum_{\mu_j > \mu} \sum_{\mu < x_{jr} \le \mu_j} (x_{jr} - \mu_j) - \sum_{\mu_j < \mu} \sum_{\mu_j < x_{jr} \le \mu} (x_{jr} - \mu_j) \right\} \tag{33}$$

is the *mixture effect*, that always lessens the overall inequality;

$$P_{\mathrm{Bm}} = \frac{1}{T} \sum_{j=1}^{k} \left[\#(x_{jr} > \mu) - \#(x_{jr} < \mu) \right] (\mu_j - \mu) \tag{34}$$

is the *mean effect*. The sum

$$P_{\mathrm{B}} = P_{\mathrm{Bt}} + P_{\mathrm{Bm}} \tag{35}$$

is the *between inequality*. A numerical application of these formulae will be presented at Sect. 3.

The within inequality $P(S_P)$, computed on the AUs, shows some reduction with respect to the average P_{W} in formula (32). For simplicity, let us assume k groups with same size $n_j = n$. Calling y_r the values of the percentile distribution (formulae (4)–(5)), the computation of P for this distribution gives rise to the following development (taking into account that the total of the AUs is $T_A = m\overline{y} = T/k$):

$$2TP/k = \sum_{r=1}^{m} |y_r - \overline{y}| = \sum_{r=1}^{m} \left| \frac{1}{k} \sum_{j=1}^{k} x_{jr} - \frac{1}{km} \sum_{j=1}^{k} \sum_{r=1}^{m} x_{jr} \right| = \frac{1}{k} \sum_{r=1}^{m} \left| \sum_{j=1}^{k} (x_{jr} - \mu_j) \right| \tag{36}$$

We know that

$$2T\,P = \sum_{j=1}^{k} \sum_{r=1}^{m} |x_{jr} - \mu| \tag{37}$$

and observe that some compensation may arise from negative and positive deviations in formula (36); if this happens, then (36) is smaller than (37) even when all the group means are equal ($\mu_j = \mu$); anyway, the absolute deviations $|x_{jr} - \mu_j|$ from the group means are generally expected to produce a smaller total than the absolute deviations $|x_{jr} - \mu|$ from the general mean. A numerical comparison of $P(S_P)$ and P_{W} will be made at Sect. 3.

2.3 The case of Theil index

Also the definition of the Theil index is very simple, and immediately usable when all individual incomes are available (Theil 1967, pp. 91–98):

$$\mathrm{Th} = \sum_{i=1}^{n} \frac{x_i}{T} \log\left(n\frac{x_i}{T}\right) = \sum_{i=1}^{n} q_i \log(nq_i) = \log n + \sum_{i=1}^{n} q_i \log q_i \tag{38}$$

where the logarithms are usually to the base e (as we shall make in the computations at Sect. 3). The range of values of Th is $[0, \log n]$, thus it is not normalized between 0 and 1, like—at least approximately—most other inequality measures; however, this weakness is counterbalanced by an easy and exact decomposition property. Also, the approximation problems can be usefully managed on the basis of this decomposition (Theil 1967, p. 95):

$$\mathrm{Th}(S) = \sum_{j=1}^{k} q_j \log\left(n\frac{q_j}{n_j}\right) + \sum_{j=1}^{k} q_j \sum_{r=1}^{n_j} \frac{q_{jr}}{q_j} \log\left(n_j\frac{q_{jr}}{q_j}\right) \tag{39}$$

where the first term is the between inequality

$$\mathrm{Th_B} = \sum_{j=1}^{k} q_j \log\left(n\frac{q_j}{n_j}\right) = \sum_{j=1}^{k} q_j \log\left(\frac{\mu_j}{\mu}\right) \tag{40}$$

and the second term is the within inequality

$$\mathrm{Th_W} = \sum_{j=1}^{k} q_j \sum_{r=1}^{n_j} \frac{q_{jr}}{q_j} \log\left(n_j\frac{q_{jr}}{q_j}\right) = \sum_{j=1}^{k} q_j \mathrm{Th}(S_j) \tag{41}$$

where

$$\mathrm{Th}\left(S_j\right) = \sum_{r=1}^{n_j} \frac{q_{jr}}{q_j} \log\left(n_j\frac{q_{jr}}{q_j}\right) = \sum_{r=1}^{n_j} \frac{x_{jr}}{T_j} \log\left(n_j\frac{x_{jr}}{T_j}\right), \tag{42}$$

thus, the within inequality is simply the weighted average of the group inequalities.

Although formula (39) displays all the group inequalities, if these are not really informative, then the computation of the between and within inequalities only requires the knowledge of the total and of the between inequality.

When we dispose of a frequency–quantity distribution $(n_1, T_1; \ldots; n_h, T_h)$ or $(f_1, q_1; \ldots; f_h, q_h)$, when $h \geq 10$ the first term $\mathrm{Th_B}$ in (39)—now called $\mathrm{Th_M}$ as the reference is a single frequency distribution—usually gives a good approximation of Th; anyway, the dispersion of incomes within each class is not accounted for. To do so, some devices are available for a reasonable approximation of each Th_j (= Theil

inequality within the class C_j). An upper bound for Th_j is (with the same symbols in formula (12), and following and analogous procedure)

$$\mathrm{Th}_j^B = \frac{(b_j - \mu_j)a_j}{(b_j - a_j)\mu_j} \log \frac{a_j}{\mu_j} + \frac{(\mu_j - a_j)b_j}{(b_j - a_j)\mu_j} \log \frac{b_j}{\mu_j} \tag{43}$$

(Theil 1967, p. 132). A more accurate approximation—perhaps excepting the last class C_h—is usually obtained by the assumption of a uniform spread of income values inside the class endpoints—see formula (13). In brief, $A_j = b_j - a_j$ and $p_{jr} = s_{jr}/s_j$, with

$$s_{jr} = a_j + (2r - 1)\, A_j/2n_j \tag{44}$$

$$s_j = \sum_{r=1}^{n_j} s_{jr} = n_j \left(a_j + b_j\right)/2 \tag{45}$$

the Theil index for the class C_j can be approximated by

$$\mathrm{Th}_j^A = \sum_{r=1}^{n_j} p_{jr} \log(n_j p_{jr}) = \log n_j + \sum_{r=1}^{n_j} p_{jr} \log p_{jr} \tag{46}$$

$$= \log n_j + \sum_{r=1}^{n_j} \frac{s_{jr}}{s_j} \log \frac{s_{jr}}{s_j}$$

$$= \frac{1}{s_j} \sum_{r=1}^{n_j} s_{jr} \log s_{jr} - \log \left(\frac{a_j + b_j}{2} \right) \tag{47}$$

If the sum in (47) is multiplied by the scanning interval A_j/n_j between successive values $s_{jr} (r = 1, \ldots, n_j)$, then it is recognized as the histogram approximating a known integral between the limits a_j and b_j:

$$\sum_{r=1}^{n_j} s_{jr} \log s_{jr} \times A_j/n_j \approx v_j = \int_{a_j}^{b_j} x \log x \mathrm{d}x = b_j^2(\log b_j - 1/2)/2$$

$$-a_j^2(\log a_j - 1/2)/2.$$

Thus,

$$T_j^A \approx \mathrm{Th}_j^C = \frac{2v_j}{A_j(a_j + b_j)} - \log \left(\frac{a_j + b_j}{2} \right). \tag{48}$$

Finally, we get the following kinds of approximations for the correction of formula (40)—now called $\mathrm{Th_M}$—applied to the frequency–quantity distribution over h classes:

$$\mathrm{Th}^B = \sum_{j=1}^{h} \mathrm{Th}_j^B q_j; \quad \mathrm{Th}^A = \sum_{j=1}^{h} \mathrm{Th}_j^A q_j; \quad \mathrm{Th}^C = \sum_{j=1}^{h} \mathrm{Th}_j^C q_j. \tag{49}$$

As already said, these approximations may be unsuitable for the last class C_h. As for the Gini index, after getting the estimator $\hat{\theta}$ of formula (17), we can apply the general formula for the Theil index in the case of Pareto distributions (Theil 1967, p. 98):

$$\mathrm{Th}_j = \frac{1}{\theta - 1} - \log \frac{\theta}{\theta - 1} \tag{50}$$

Some numerical comparisons of these approximations will be presented in the next section.

The general result concerning the *aggregate units* in case of k groups, with respective sizes n_j and total quantities T_j, turns out as follows. The matrix (3) is composed of n rows and m columns, as we consider n_j replicas of the AUs for group S_j. Each replica has a total $= \mu_j$; thus, having determined unitary shares for m AUs in the jth group, by means of differences between values of the Lorenz function

$$u_{jr} = L_j\,(r/m) - L_j\,((r-1)/m) \quad r = 1, \ldots, m,$$

the values x_{jr} of (3) are simply obtained as $x_{jr} = u_{jr}\mu_j$. Letting $x_{jr}^{(i)}$ the values of the ith replica ($i = 1, \ldots, n_j$) for group S_j, the total for the rth column in (3) is therefore

$$y_r = \sum_{j=1}^{k} \sum_{i=1}^{n_j} u_{jr}^{(i)} \mu_j = \sum_{j=1}^{k} u_{jr} T_j$$

(division by n is irrelevant, as we are using an ISP index); the sum of these AUs is

$$\sum_{r=1}^{m} y_r = \sum_{r=1}^{m} \sum_{j=1}^{k} u_{jr} T_j = \sum_{j=1}^{k} T_j \sum_{r=1}^{m} u_{jr} = T.$$

The Theil index applied to the m AUs $y_1, \ldots, y_m$ is therefore:

$$\begin{aligned} \mathrm{Th}(G_P) &= \sum_{r=1}^{m} \frac{y_r}{T} \log\left(m \frac{y_r}{T}\right) = \sum_{r=1}^{m} \sum_{j=1}^{k} u_{jr} q_j \log\left(m \sum_{j=1}^{k} u_{jr} q_j\right) \\ &= \sum_{j=1}^{k} q_j \sum_{r=1}^{m} u_{jr} \log\left(m \sum_{j=1}^{k} u_{jr} q_j\right) \end{aligned}$$

Table 1 Percent frequencies and quantities for 13 income classes, from the 2004 Bank of Italy Survey on household income and wealth (Banca d'Italia 2006)

Classes in Euros	North		Center		South		Italy	
C_j	f_j	q_j	f_j	q_j	f_j	q_j	f_j	q_j
250 −| 8274	2.666	0.485	2.247	0.415	10.061	2.759	5.007	1.035
8274 −| 12000	5.058	1.547	5.760	1.803	13.402	6.025	7.955	2.723
12000 −| 15040	7.642	3.069	7.316	3.029	13.250	7.988	9.416	4.285
15040 −| 19030	10.995	5.546	10.599	5.515	16.287	12.123	12.650	7.178
19030 −| 23230	11.820	7.359	13.767	8.915	12.528	11.568	12.475	8.777
23230 −| 27300	10.665	7.972	10.196	7.886	8.428	9.342	9.828	8.293
27300 −| 32540	11.325	10.024	11.636	10.623	7.555	9.926	10.152	10.142
32540 −| 37240	8.494	8.769	8.468	9.028	5.581	8.562	7.530	8.779
37240 −| 43190	9.126	10.831	8.986	11.038	4.252	7.488	7.493	10.047
43190 −| 48540	6.295	8.518	5.472	7.670	2.771	5.584	4.958	7.585
48540 −| 58030	6.432	10.069	6.221	10.149	2.354	5.488	5.045	8.947
58030 −| 72140	4.810	9.168	4.896	9.591	2.278	6.354	3.996	8.567
72140 −| 1022617	4.673	16.643	4.435	14.341	1.253	6.793	3.497	13.642

The comparison with (41) shows that $\mathrm{Th}_W = \mathrm{Th}(G_P)$ if and only if the rows in (3) are proportional, namely, if and only if the inequalities within each group are the same (cf. Frosini 1989, p. 362).

3 An application to the 2004 Bank of Italy Survey on household incomes

Many of the formulae previously exposed will be applied to some data coming from the 2004 Bank of Italy Survey on household income and wealth (Banca d'Italia 2006; Radaelli 2008). An essential abstract from these data is presented in Table 1, showing four frequency–quantity distributions (with relative frequencies and quantities), pertaining to Italy, as well as to the traditional geographical areas North, Center and South; the respective sample sizes and average incomes for households are: 8,008 units and 29,888 Euros for Italy, 3,638 units and 33,779 Euros for North, 1,736 units and 32,726 for Center, 2,634 units and 22,642 Euros for South; the respective weights concerning total income are: $q_N = 0.51345$, $q_C = 0.23737$, $q_S = 0.24918$. These data are immediately functional for the graphical representation of the two relevant histograms—for frequencies and quantities—reported in Fig. 1, and for the calculation of Pietra–Ricci's index.

Following the same order of Sect. 2, let us start with the computation about the Gini's index. Table 2 reports the main results relevant for the computation of the G index for geographical groups and Italy. Column (2) reports the index G_M computed from formula (11). The following four columns report some kinds of approximations for the inequalities within the classes: column (3) from formula (12); column (4) from formula (12) applied to the first twelve classes, and using the Pareto approximation (18) for the last class; column (5) from the uniform approximation (16); column (6)

Table 2 Computations for the Gini index, for the three main geographical areas and Italy (from data in Table 1 and text)

Groups (1)	G_{M} (2)	G^B (3)	G^B + Pareto (4)	G^A (5)	G^A + Pareto (6)	G (7) = (2) + (6)
North	33.89	0.61	0.51	1.24	0.41	34.30
Center	32.45	0.53	0.45	1.18	0.34	32.79
South	34.67	0.53	0.52	0.46	0.37	35.04
Italy	35.10	0.51	0.45	0.45	0.34	35.44

All original values are multiplied by 100

using this same approximation for the first twelve classes, and Pareto approximation (18) for the last class; and finally, column (7) provides the proposed approximation for G, obtained by summing G_{M} with the last approximation of the inequality within classes.

On the basis of these results, the within inequality computed by the average (20), obtained by means of the AUs approach, is $G(S_P) = 34.30 \times 0.51345 + 32.79 \times 0.23737 + 35.04 \times 0.24918 = 34.13\%$; the ensuing 'between inequality', obtained as a residual with respect to the overall inequality 35.44%, is 1.31%. Finally, the calculation of the Gini index for the three group means with formula (11) gives rise to another computation of the between inequality I_{B}; this kind of between inequality turns out to be 8.32%.

Turning to the Pietra–Ricci index, the necessary computations are easily derived from Table 1, as the differences $|q_j - f_j|$ for the h classes provide all the information required by formula (23). Figure 1 is related to frequency and quantity distributions for Italy, and gives an example of the graphical representation of P_H (the area between the two histograms, before or after the intersection point). The values of P_H, obtained by summing the positive differences $(f_j - q_j)$, or the positive differences $(q_j - f_j)$, in Table 1, are already good approximations for P; they are reported in column (2) of Table 3. As P_H exactly coincides with P when the mean μ falls just on a class endpoint, in order to get a better approximation, we should divide the class containing the mean in two sub-classes; the corresponding frequencies and quantities of these sub-classes are obtained by linear approximation (a very good approximation, since we are about the center of the distribution). As an example, consider the case of the North distribution; the mean μ_N =33,779 is included in the class 32,540 −| 37,240 of width 4,700, which must be split into the subclasses 32,540 −| 33,779 (of width 1,239) and 33,779 −| 37,240 (of width 3,461). The class frequency 309 can be proportionally divided according to the widths of the sub-classes; thus the frequency of the first sub-class is 309× 1,239/4,700 = 81 (while the other frequency is 309 − 81 = 228). Assuming that the mean of this sub-class is its central point 33,129, the positive difference $(f_{js} - q_{js})$—concerning relative frequencies and quantities of this sub-class—is (81/3,638 − 81× 33,129/122,889,326) = 0.00041; thus the percent estimated increase of P_H is 0.04, and the final approximation of P is 24.17 + 0.04 = 24.21; as expected, P_H itself is a very good approximation of P. The remaining approximations are also given in Table 3; notice that the second approximation is practically zero, owing to

Table 3 Computations of the Pietra–Ricci index, for geographical areas and Italy (from data in Table 1 and text)

Groups (1)	P_H (2)	Correction (3)	P (4)	Contribution to inequality (5)
North	24.17	0.040	24.21	11.918
Center	23.34	0.001	23.34	5.356
South	25.07	0.023	25.09	7.996
Italy	25.05	0.217	25.27	25.270

All values are multiplied by 100

the fact that the group mean $\mu_C = 32{,}726$ is very near to the left endpoint 32,540 of the class containing μ_C; on the contrary, the heaviest correction (although of limited impact) happens for Italy, as the general mean $\mu = 29{,}888$ is near the center of the class interval 27,300 −| 32,540.

The components in formulae (28)–(30), corresponding to the three groups, are as follows (where N = North, C = Center, S = South; these components obviously sum to the total inequality $P = 25.27$):

$$D_N = 12.428 - 0.248 - 0.262 = 11.918$$
$$D_C = 5.540 - 0.068 - 0.116 = 5.356$$
$$D_S = 6.252 - 0.463 + 2.207 = 7.996.$$

The components (32)–(34) of this decomposition of total inequality are as follows:

$P_W = 12.428 + 5.540 + 6.252 = 24.220$ is the within inequality;
$P_{Bt} = -0.248 - 0.068 - 0.463 = -0.779$ is the mixture effect;
$P_{Bm} = -0.262 - 0.116 + 2.207 = 1.829$ is the mean effect;

the between inequality is thus $P - P_W = P_{Bt} + P_{Bm} = 1.05$. Finally, the calculation of P on the group means (with corresponding group frequencies), as only the South group has a mean smaller than the general mean μ, is simply equal to (29,888 − 22,642)(1736/T) = 0.0526, or 5.26%. The computation of the within inequality $P(S_P)$, effected by the application of P to AUs, results in $P(S_P) = 24.089$, both for m (number of AUs) = 50 and = 100 (just a little less than $P_W = 24.220$, as expected). It is quite obvious that there is no convenience to make the calculation of the within inequality through the AUs, as the result is practically equivalent to the one provided by the decomposition (31).

As regards the computations for the Theil index, first of all we have computed the approximation (40)—called in this case Th_M—which uses only the distribution of partial means; these values, multiplied by 100 (as usual, although the Theil index is not normalized in the interval [0,1]), are reported in the second column of Table 4. As observed in Sect. 2.3, there are several types of approximations of the second term in (39), which has the meaning of 'within inequality'; the first type, based on the upper bound (43) for the inequality measure within the class $C_j (j = 1, \ldots, 13)$, has been

Table 4 Computations for the Theil index, for main geographical areas and Italy (from data in Table 1 and text)

Groups (1)	$\mathrm{Th_M}$ (2)	Th^B (3)	Th^B + Pareto (4)	Th^A and Th^C (5)	Th^A + Pareto (6)	Th (7) = (2) + (6)
North	19.995	10.966	3.069	2.496	2.796	22.791
Center	17.745	7.997	1.672	2.171	1.389	19.134
South	21.203	5.540	2.316	1.554	1.771	22.974
Italy	21.256	8.933	2.501	2.184	2.157	23.413

All original values are multiplied by 100

computed as the weighted mean of these bounds (see formula (49)). These approximations Th^B are reported in column (3) of Table 4; it is immediately evident that these 'corrections' are excessively large, owing to the exceptional wideness of the last class. Column (4) of Table 4 reports an analogous weighted mean, however replacing the (43) value for the last class with a Pareto estimation (formula (50)): the resulting approximations of the within inequalities (41) become much more reliable. The other kinds of approximation, based on the assumption of uniform spread within each class, yield values substantially comparable with the ones just mentioned: they are reported in column (5) of Table 4 as Th^A (see formulae (46) and (49)). The modification of such approximation by the recourse to a Pareto estimation (50) concerning the last class, does not substantially reduce the previous evaluations; actually, this happens only for the Center area, while for North and South they are even increased.

Although a good approximation of the integral to the histogram —resumed in formula (48)—was expected owing to the large frequencies of the classes, it was a really welcome result that the first four significant figures (all the figures reported in Table 4) coincide for the two kinds of approximation, thus rendering the integral approximation (48) very simple and reliable; that is the reason for the heading 'Th^A and Th^C' of the column (5) of Table 4. Finally, column (7) reports the proposed approximation for the Theil index, as the sum of values in columns (2) and (6).

The 'within' inequality (41) turns out to be $\mathrm{Th_W} = 22.791 \times 0.51345 + 19.134 \times 0.23737 + 22.974 \times 0.24918 = 21.969$; thus the between inequality $\mathrm{Th_B}$ is simply = 23.413—21.969 = 1.444%. The recourse to AUs (just to complete this application) changes these results; with $m = 50$ AUs, $\mathrm{Th}(G_P) = 19.293$; with $m = 100$, $\mathrm{Th}(G_P) = 19.377$ (instead of $\mathrm{Th_W} = 21.969$). For these choices of m, the evaluation of between inequality as a residual gives rise to values (23.413–19.293) = 4.120% for $m = 50$, and to (23.413–19.377) = 4.036% for $m = 100$.

4 Conclusions

The general problem of decomposition of inequality measures, so much debated in the literature, has been outlined at paragraph 1, where four approaches have been presented, with a short insight for each of them, also stressing the most recent achievements by Silber, Yitzhaki and Lerman, Dagum, Deutsch and Silber, Okamoto, and the

author's proposal. Some problems of computation, approximation and decomposition of Gini, Theil and Pietra–Ricci indices have been dealt with at Sects. 2.1, 2.2 and 2.3.

Amongst these indices, the most investigated in this article has been the less known and the more neglected one: the Pietra–Ricci index. Although satisfying only the weak principle of transfers, this index possesses an important and well-known economic property, whose graphical counterpart uses only the two basic histograms of frequencies and quantities. The problem of approximation—when only a frequency distribution is available—is practically non-existent, while the obtained three-term decomposition of this index appears operatively simple and highly informative. The application displayed at paragraph 3, concerning a recent large survey of household incomes for Italy, confirms the suitability of the Pietra–Ricci index, and—in any case—a coherent behaviour with respect to the other two indices.

Acknowledgments The author wishes to thank the three anonymous referees for their several, valuable comments and suggestions, leading to a more thoughtful and extensive coverage of the subject.

References

Banca d'Italia (2006) Household income and wealth in 2004. Supplement to the Statistical Bulletin—Sample Surveys, Year XVI, n 7

Bhattacharya N, Mahalanobis B (1967) Regional disparities in household consumption in India. J Am Stat Assoc 62:143–161

Bourguignon F (1979) Decomposable income inequality measures. Econometrica 47:901–920

Castagnoli E, Muliere P (1990) A note on inequality measures and the Pigou-Dalton principle of transfers. In: Dagum C, Zenga M (eds) Income and wealth distribution, inequality and poverty. Springer, Berlin, pp 171–182

Cowell FA (1980) On the structure of additive inequality measures. Rev Econ Stud 47:521–531

Dagum C (1997a) Scomposizione ed interpretazione delle misure di disuguaglianza di Gini e di entropia generalizzata. Statistica 57:295–308

Dagum C (1997b) A new approach to the decomposition of the Gini income inequality ratio. Empir Econ 22:515–531

Dagum C (2001) Desigualdad del rédito y bienestar social, descompositión, distancia direccional y distancia métrica entre distribuciones. Estud Econ Apl 17:5–52

Deutsch J, Silber J (1997) Gini's "transvariazione" and the measurement of distance between distributions. Empir Econ 22:547–554

Deutsch J, Silber J (2007) Decomposing income inequality by population subgroups: a generalization. In: Bishop J (ed) Inequality and poverty, research on economic inequality, vol 14. Elsevier, Amsterdam, pp 237–253

Deutsch J, Silber J (2008) On the Shapley value and the decomposition of inequality by population subgroups with special emphasis on the Gini index. In: Betti G, Lemmi A (eds) Advances on income inequality and concentration measures. Routledge, London pp 161–178

Fei JCH, Ranis G, Kuo SWY (1979) Growth with equity. The Taiwan case. Oxford University Press, New York

Frosini BV (1985) Comparing inequality measures. Statistica 45:299–317

Frosini BV (1987) Lezioni di statistica. Parte prima (Seconda edizione). Vita e Pensiero, Milano

Frosini BV (1989) Aggregate units, within-group inequality, and the decomposition of inequality measures. Statistica 49:349–369

Frosini BV (1990a) Ordinal decomposition of inequality measures in case of Dagum distributions. In: Dagum C, Zenga M (eds) Income and wealth distribution, inequality and poverty. Springer, Berlin

Frosini BV (1990b) Lezioni di statistica, Parte prima (Terza Edizione). Vita e Pensiero, Milano

Frosini BV (2003) Decomposition of inequality measures based on aggregate units. Estadistica 55:377–388

Frosini BV (2005) Inequality measures for histograms. Statistica 65:27–40

Frosini BV (2009) Metodi statistici (Seconda edizione). Carocci, Roma

Gastwirth JL (1972) The estimation of the Lorenz curve and Gini index. Rev Econ Stat 54:306–316
Gini C (1914) Sulla misura della concentrazione e della variabilità dei caratteri. Atti Regio Ist Veneto 73(Parte II):1203–1248
Kakwani NC (1980) Income inequality and poverty. Oxford University Press, London
Marshall AW, Olkin I (1979) Inequalities: theory of majorization and its applications. Academic Press, New York
Okamoto M (2009) Decomposition of Gini and multivariate Gini indices. J Econ Inequal 7:153–177
Pietra G (1915) Delle relazioni tra gli indici di variabilità Nota I. Atti Regio Ist Veneto 74(Parte II):775–792
Quintano C, Castellano R, Regoli A (2009) Evolution and decomposition of income inequality in Italy, 1991–2004. Stat Methods Appl 18:419–443
Radaelli P (2008) A subgroups decomposition of Zenga's uniformity and inequality indices. Stat Appl 6:117–136
Ricci U (1916) L'indice di variabilità e la curva dei redditi. Giornale degli Econ Riv Stat Serie Terza 53:177–228
Shorrocks AF (1980) The class of additively decomposable inequality measures. Econometrica 48:613–625
Shorrocks AF (1999) Decomposition procedures for distributional analysis: a unified framework based on the Shapley Value. Mimeo, University of Essex
Silber J (1989) Factor components, population subgroups and the computation of the Gini index of inequality. Rev Econ Stat 71:107–115
Theil H (1967) Economics and information theory. North Holland, Amsterdam
Yitzhaki S, Lerman RI (1991) Income stratification and income inequality. Rev Income Wealth 37:313–329
Zagier D (1983) Inequalities for the Gini coefficient of composite populations. J Math Econ 12:103–118
Zenga M (1987) Effetti della normalizzazione sul principio della somiglianza e sulla scomponibilità degli indici di concentrazione. In: Zenga M (ed) La distribuzione personale del reddito: problemi di formazione, di ripartizione e di misurazione. Vita e Pensiero, Milano
Zenga M (2007) Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. Stat Appl 6:137–151