KM 26(1987)
pag 61 - 74

ON THE COMPUTATION OF INEQUALITY MEASURES FROM GROUPED INCOME DATA[1]

J.G. Odink[2] and E. van Imhoff[3]

Abstract

This paper deals with the underestimation error that results if indexes of income inequality are computed from grouped data. On the basis of decomposition formulas we propose a method for approximating the true Theil and Gini index from grouped data. Special attention is paid to the highest income bracket, which lacks a finite upper limit. An empirical application indicates that the proposed method approximates the true indexes with 5-digit accuracy.

---

[1]The authors are indebted to Professor J.S. Cramer and Professor J. Hartog for helpful comments.

[2]Department of Microeconomics, University of Amsterdam, Jodenbreestraat 23, 1011 NH Amsterdam, The Netherlands. Phone: 020 - 5254255.

[3]Fiscal-Economic Institute, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands. Phone: 010 - 4081498.

ON THE COMPUTATION OF INEQUALITY MEASURES FROM GROUPED INCOME DATA

1. Introduction

The various measures of income inequality that have been used in the past years have been designed for application to individual income data, at least in principle. Most published data however, particularly the official income statistics of national statistical institutes, concern income data that have been grouped in brackets. It is well known that measures of inequality computed from such grouped data underestimate the "true" value of the measure, i.e. the value as computed from the underlying individual incomes. The measures as computed from grouped data thus constitute lower bounds to the true measures. Upper bounds can be obtained by assuming that the incomes within each bracket are maximum unequally distributed.

In a previous paper (Odink & Van Imhoff, 1984) we developed a method for approximating the true Theil index of income inequality from grouped income data. Our calculations showed that the approximation error was only .1% for the Dutch Central Bureau of Statistics (CBS) frequency distribution.

The purpose of our present paper is threefold. In the first place we extend our approximation method to the Gini index. Second, we investigate various methods for handling the highest income bracket which lacks a finite upper limit. Finally we apply the theoretical results to a comprehensive dataset, consisting of a subsample of some 29,000 wages of employees from the Dutch CBS "Loonstructuuronderzoek 1979".

The plan of this paper is as follows. Section 2 gives the decomposition formulas for the Theil index (T) and the Gini index (G). Section 3 discusses methods for the computation of income inequality within closed brackets while section 4 deals with the highest, open bracket. In section 5 we present the results of our empirical analysis. The final section summarizes the major conclusions.

2. Decomposition formulas and the underestimation error

The effect of grouping on observed income inequality can be derived from the decomposition formula for the inequality index under consideration. Both T and G are decomposable, although the decomposability of G holds only if the grouping is ordered. Such a restriction does not exist for the decomposition of T.

The formulas for the decomposition of the total inequality into inequality between groups and inequality within groups for both T and G are given in Odink & Van Imhoff (1987). The value of the inequality index computed from grouped data equals the inequality between groups. Grouping results in the disregarding of the inequality within the groups.

The relative importance of the underestimation error due to grouping depends on the underlying income distribution, the way in which the grouping has been performed, and the inequality index under consideration. In Gastwirth (1972) formulas are given for strict lower and upper bound to the true Gini index for grouped data, as well as for modified bounds under the assumption of a unimodal probability density function. Similar results hold for the Theil index (Theil, 1967; Gastwirth, 1975).

In this paper we are not very much interested in theoretical bounds within which the true inequality index must lie. Rather we look for a computation method that approximates the true inequality index as closely as possible and investigate the order of magnitude of the remaining approximation error. In choosing such an approximation method one should distinguish between closed brackets with finite lower and upper limits, and the highest, open income bracket. Both types of brackets will be discussed in turn in the following sections.

3. Income inequality within closed brackets

Consider a closed bracket [a,b) with $0 \leq a < b < \infty$, containing all incomes X_i, $i=1,\ldots,N$, for which $a \leq X_i < b$. The mean income in bracket [a,b) is assumed to be known and equal to m.

The distribution of the incomes within the bracket can be described by means of a probability density function (pdf). The form of this pdf, which cannot be observed if published data are in grouped form, determines the income inequality within the bracket. In principle any pdf can be chosen as long as its mean is equal to m.

Three extreme forms of possible pdf's are given in Figures 1-3. Figure 1 corresponds to the unambiguous lower bound to within-bracket inequality, viz. zero. Figure 2 illustrates the case of maximum inequality, with all incomes concentrated at the bracket's lower and upper limit. Figure 3 corresponds to Gastwirth's (1972; 1975) "sharpened upper bound". Here income inequality is maximum, subject to the restriction that the pdf is non-increasing (a similar figure can be drawn for non-decreasing pdf).

Figure 1: pdf for complete equality

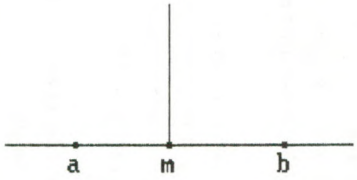


Figure 2: pdf for maximum inequality

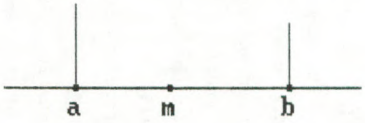


Figure 3: non-increasing pdf for maximum inequality

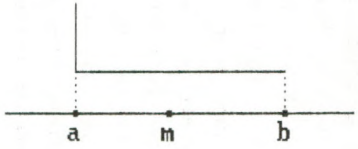
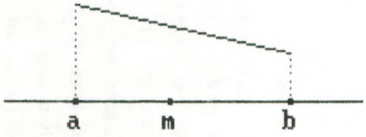


Figure 4: linear pdf



We propose (cf. Odink & Van Imhoff, 1984) to approximate the inequality within the bracket by assuming the pdf to be linear (see Figure 4):

$$f(z) = \begin{cases} 0 & \text{for } z < a \\ Az + B & \text{for } a \leq z < b \\ 0 & \text{for } z \geq b \end{cases} \qquad (1)$$

Since the following two conditions must hold:

$$\int_{-\infty}^{+\infty} f(z)\, dz = 1 \qquad (2)$$

$$E_f(z) = \int_{-\infty}^{+\infty} z \cdot f(z)\, dz = m \qquad (3)$$

A and B can be obtained from (1)-(3):

$$A = \frac{12m - 6(b+a)}{(b-a)^3} \qquad (4)$$

$$B = \frac{(b-a)^2 + 3(b+a)^2 - 6m(b+a)}{(b-a)^3} \qquad (5)$$

Thus, the slope of the linear pdf is uniquely determined by the bracket's limits a and b, and the mean income m.
The Theil index for any continuous income distribution equals (cf. Theil, 1967, p. 96):

$$T_f = \int_{-\infty}^{+\infty} (z/m) \cdot \log(z/m) \cdot f(z)\, dz \qquad (6)$$

For the linear pdf defined by (5) the Theil index can be shown to equal:

$$T_{lin} = \frac{B}{m} \cdot \left[\frac{1}{2} \cdot b^2 \cdot \log(b/m) - \frac{1}{2} \cdot a^2 \cdot \log(a/m) - \frac{1}{4} \cdot b^2 + \frac{1}{4} \cdot a^2\right] +$$

$$+ \frac{A}{m} \cdot \left[\frac{1}{3} \cdot b^3 \cdot \log(b/m) - \frac{1}{3} \cdot a^3 \cdot \log(a/m) - \frac{1}{9} \cdot b^3 + \frac{1}{9} \cdot a^3\right] \qquad (7)$$

Similarly, the Gini index for any continuous distribution equals (cf. Gastwirth, 1972):

$$G_f = \frac{1}{m} \cdot \int_{-\infty}^{+\infty} F(z) \cdot [1-F(z)]\, dz \tag{8}$$

where F(z) is the cumulative distribution function defined by

$$F(z) = \int_{-\infty}^{z} f(u)\, du \tag{9}$$

For the linear pdf (from (9) and (1)):

$$F(z) = \frac{1}{2} \cdot Az^2 + Bz + C \tag{10}$$

where from the condition

$$F(a) = 0 \tag{11}$$

C can be seen to equal

$$C = -\frac{1}{2} \cdot Aa^2 - Ba \tag{12}$$

The Gini index for the linear pdf equals:

$$G_{lin} = \frac{1}{m} \cdot \left[-\frac{A^2}{20} \cdot (b^5 - a^5) - \frac{AB}{4} \cdot (b^4 - a^4) + \frac{A/2 - AC - B^2}{3} \cdot (b^3 - a^3) + \right.$$
$$\left. + \frac{B - 2BC}{2} \cdot (b^2 - a^2) + C \cdot (1-C) \cdot (b-a) \right] \tag{13}$$

For later reference we also give the formulas corresponding to maximum inequality within the bracket (Figure 2). For the Theil index (Theil, 1967, p. 132):

$$T_{max} = \left[\frac{b-m}{b-a}\right] \cdot \frac{a}{m} \cdot \log\left[\frac{a}{m}\right] + \left[\frac{m-a}{b-a}\right] \cdot \frac{b}{m} \cdot \log\left[\frac{b}{m}\right] \tag{14}$$

For Gini (Gastwirth, 1972):

$$G_{max} = \frac{(m-a) \cdot (b-m)}{m \cdot (b-a)} \tag{15}$$

4. Income inequality within the highest, open bracket

The crucial characteristic of the highest income bracket is that its upper limit is infinite: the highest bracket is a limiting case of the closed bracket for $b \to \infty$.

For the expression for the maximum Gini index (15) this property of the highest bracket does not lead to any particular problems since its limiting value is finite:

$$\lim_{b \to \infty} G_{max} = \frac{m-a}{m} \tag{16}$$

(cf. Gastwirth, 1972). However, for the Theil index the corresponding expression in (14) breaks down if $b \to \infty$. For this reason Theil (1967, p. 98) proposed for computing the inequality within the highest bracket the assumption that the individual incomes follow the Pareto distribution:

$$f(z) = \alpha \cdot z^{-\alpha-1} \cdot a^{\alpha} \qquad \text{for } z \geq a \tag{17}$$

The parameter α can be obtained from the condition

$$\int_a^{\infty} z \cdot f(z)\, dz = m \tag{18}$$

from which

$$\alpha = \frac{m/a}{m/a - 1} = \frac{m}{m-a} \tag{19}$$

In this case the Theil index equals:

$$T_{Pareto} = \frac{1}{\alpha-1} - \log \frac{\alpha}{\alpha-1} \tag{20}$$

Similarly, for G we have:

$$G_{Pareto} = \frac{1}{2\alpha-1} \tag{21}$$

(cf. Gastwirth, 1972).

Theil (1967) used approximation (20) in combination with (14) for closed brackets to compute an upper bound to the true Theil index. In a previous

paper (Odink & Van Imhoff, 1984) we adopted a similar strategy, combining (20) with (7) to approximate the true Theil index.

Here we wish to propose an alternative approximation procedure for the highest bracket. This method consists of putting an artificial upper limit to the values that individual incomes within the highest bracket can take. This artificial upper limit is obtained by postulating the pdf within the bracket to be linear with mean m (see Figure 5).

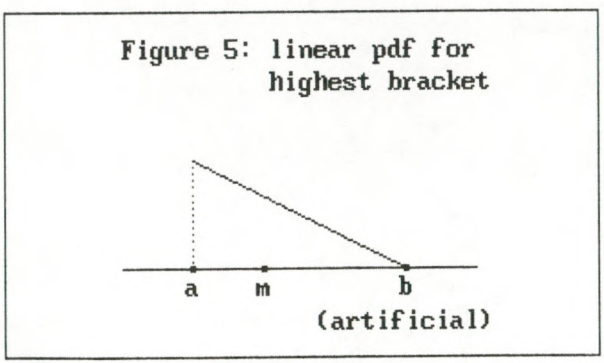
Starting from (1), (4) and (5), with b now unknown, b can be obtained by requiring that

$$\lim_{z \to b} f(z) = f(b) = 0 \qquad (22)$$

This condition implies that

$$b = 3m - 2a \qquad (23)$$

Postulating (23), all formulas of the previous section can be applied to the highest bracket which has now in fact become a closed bracket.



5. Empirical analysis

To illustrate the various methods discussed in the previous sections we have applied our formulas to a subsample from the CBS "Loonstructuuronderzoek 1979" (CBS, 1983). The sample consists of 29,271 weekly wages of employees, ranging from 6 to 2,635 guilders. Since the CBS brackets (CBS, 1984) have

been devised for the classification of yearly incomes we have multiplied each wage with a scale factor 56 (52 weeks per year plus four weeks bonus). The grouped income data are given in the Appendix to Odink & Van Imhoff (1987); a plot of the linear approximation of the pdf for these data is given in Figure 6.

A slight disadvantage of the dataset used is that it does not include any very high wages: the questionnaires used by the CBS refer to wages below a certain maximum level only. As a consequence the highest three brackets (as well as the lowest bracket, consisting of negative incomes) of the original 32 brackets classification remain empty.

Table 1 gives the values of T and G for the individual wages and for grouped data.

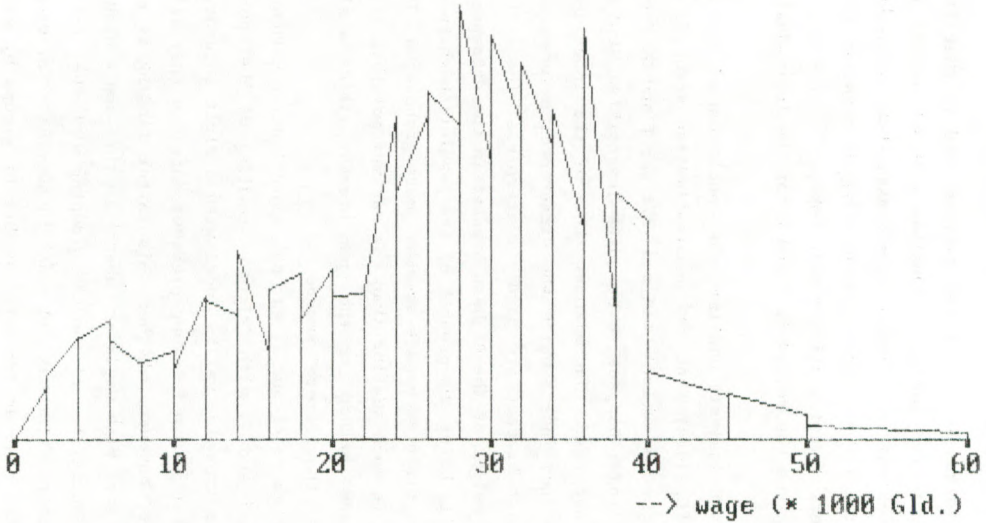
The present analysis confirms the conclusion of our earlier paper (Odink & Van Imhoff, 1984) that the underestimation error is almost negligible for the CBS classification: .45% and .27% for T and G, respectively.

The Gini index happens to be quite insensitive with respect to the correction method used. The maximum value of the index is only slightly larger than the true index, while the differences between the various methods for the highest bracket are hardly discernible. The latter is due to the fact that the weight of the highest bracket in the decomposition formula for G is very small, being the product of its population share and its income share. For T the differences are somewhat more pronounced. The actual error due to grouping is much smaller than the maximum possible error. Not surprisingly, the "closed" method for the open bracket yields a slightly higher maximum error than the "Pareto" method.

The performance of our linear pdf approximation method is surprisingly good: for both T and G, with "closed" handling of the open bracket, it approximates the true inequality index with 5-digit accuracy. The Pareto-approximation that we used in our previous paper is only slightly inferior. It is not clear, however, whether this latter finding is partly due to the fact that the very high wages are absent in the sample used.

Finally, comparison of the CBS grouping with the "fractile" groupings, with equal number of incomes and equal income share for each group, respectively, illustrates that an identical number of groups by no means guarantees that the underestimation errors are of the same order of magnitude. The CBS classification is preferable to fractiles as far as the computation of inequality measures is concerned.

Figure 6: Linear approximation of pdf for grouped data



Source: CBS (1983)

Table 1: Theil and Gini indexes for 29,271 individual wages - 28-group classification

THEIL-index:	correction	index	dev. in %
true	-	0.13688	-
28 CBS brackets:			
no correction	-	0.13626	- 0.45
maximum + Pareto	0.00188	0.13814	+ 0.92
maximum + closed	0.00202	0.13828	+ 1.02
linear + Pareto	0.00076	0.13702	+ 0.10
linear + closed	0.00062	0.13688	0.00
28 population fractiles	-	0.13464	- 1.64
28 income fractiles	-	0.13201	- 3.56
GINI-index:	**correction**	**index**	**dev. in %**
true	-	0.27872	-
28 CBS brackets:			
no correction	-	0.27796	- 0.27
maximum + true	0.00115	0.27911	+ 0.14
maximum + Pareto	0.00114	0.27910	+ 0.14
maximum + closed	0.00114	0.27910	+ 0.14
linear + Pareto	0.00076	0.27872	0.00
linear + closed	0.00075	0.27872	0.00
28 population fractiles	-	0.27795	- 0.28
28 income fractiles	-	0.27704	- 0.60

The differences between the various methods in Table 1 discussed above are admittedly not very spectacular. Indeed, one could question the relevance of the whole exercise, given that the maximum underestimation error never exceeds 3.5%.

In order to investigate the performance of the preferred method under somewhat grimmer circumstances we have repeated our calculations for two very bad groupings: 10 population fractiles and 10 income fractiles, respectively. The results are summarized in Table 2.

While the Gini index remains relatively insensitive with respect to the method used, for T the differences are indeed clearly discernible. The differences between the theoretical lower and upper bound to the true Theil index (no correction versus maximum underestimation error) has now become so large that they can hardly be used in making reliable statements about prevailing income inequality or its development over time.

Again, the proposed method (linear + closed) performs very satisfactorily. Its largest estimation error in absolute terms is only about 1.5%. With the exception of G for population fractiles, where the Pareto method yields a marginally smaller error, the method with linear interpolation of closed brackets and with an artificial upper limit to the highest, open bracket yields the best - and a very accurate - approximation.

6. Conclusions

Inequality measures computed from grouped income data underestimate the true value of the measure. In this paper we develop a method to approximate the true Theil (T) and Gini (G) indexes from grouped data. The method assumes the incomes within brackets to be distributed according to a linear probability density function.

For the highest income bracket, which lacks a finite upper limit, two alternative approximation methods are discussed. One method assumes a Pareto distribution; the other method assumes a decreasing linear probability density function and postulates an artificial upper limit to the values that incomes within the highest bracket can take.

Calculations based on a comprehensive dataset of individual wages yield the following conclusions:

- the underestimation error for the Dutch CBS classification is for both T and G almost negligible. For this specific classification one could well question the relevance of any approximation method.

Table 2: Theil and Gini indexes for 29,271 individual wages - 10-group classification

THEIL-index:	correction	index	dev. in %
true	-	0.13688	-
10 population fractiles:			
no correction	-	0.12768	- 6.72
maximum + Pareto	0.02324	0.15092	+ 10.26
maximum + closed	0.02605	0.15373	+ 12.31
linear + Pareto	0.01349	0.14117	+ 3.13
linear + closed	0.00718	0.13486	- 1.48
10 income fractiles:			
no correction	-	0.12266	- 10.39
maximum + Pareto	0.05374	0.17640	+ 28.87
maximum + closed	0.05486	0.17752	+ 29.69
linear + Pareto	0.01624	0.13890	+ 1.48
linear + closed	0.01450	0.13716	+ 0.20
GINI-index:	**correction**	**index**	**dev. in %**
true	-	0.27872	-
10 population fractiles:			
no correction	-	0.27350	- 1.87
maximum + true	0.00922	0.28272	+ 1.44
maximum + Pareto	0.00696	0.28045	+ 0.62
maximum + closed	0.00746	0.28096	+ 0.80
linear + Pareto	0.00562	0.27912	+ 0.14
linear + closed	0.00472	0.27822	- 0.18
10 income fractiles:			
no correction	-	0.27039	- 2.99
maximum + true	0.01318	0.28357	+ 1.74
maximum + Pareto	0.01284	0.28323	+ 1.62
maximum + closed	0.01293	0.28331	+ 1.65
linear + Pareto	0.00855	0.27894	+ 0.08
linear + closed	0.00843	0.27882	+ 0.04

- the proposed approximation method is accurate up to the fifth decimal inclusive for both the Theil and the Gini index;
- even with very bad initial classifications, the proposed correction method yields very satisfactory approximations for both T and G;
- classification in fractiles leads to much bigger underestimation errors than the CBS classification and is not to be preferred as far as the computation of inequality measures is concerned.

References

CBS (1983), Loonstructuuronderzoek 1979. The Hague: Staatsuitgeverij.

CBS (1984), De Personele Inkomensverdeling 1979. The Hague: Staatsuitgeverij.

Gastwirth, J.L. (1972), "The Estimation of the Lorenz Curve and Gini Index". Review of Economics and Statistics, vol. 54, pp. 306-322.

Gastwirth, J.L. (1975), "The Estimation of a Family of Measures of Economic Inequality". Journal of Econometrics, vol. 3, pp. 61-70.

Odink, J.G. & E. van Imhoff (1984), "True Versus Measured Theil Inequality". Statistica Neerlandica, vol. 38, pp. 219-232.

Odink, J.G. & E. van Imhoff (1987), "Gini, Theil, and the trade-off between optimal and efficient grouping of income data". Kwantitatieve Methoden, this issue.

Theil, H. (1967), Economics and Information Theory. Amsterdam: North Holland Publishing Company.

Ontvangen: 04-02-1987
Geaccepteerd: 04-09-1987