ARGUMENTA OECONOMICA
No 1 (42) 2019
PL ISSN 1233-5835

*Katarzyna Ostasiewicz**, *Edyta Mazurek**

ON THE ACCURACY OF INEQUALITY MEASURES CALCULATED FROM AGGREGATED DATA

Our aim in this study is to compare the typically used methods of calculating inequality measures for aggregated data with the modified approaches by the relative error. We start with the individual observations (obtained from the tax office) and calculate the exact values of a few inequality measures. Then, we construct the interval distribution in the form that is usually available in the statistical yearbooks. The standard approach is to treat all observations from a given class as if concentrated in the exact middle of this class. We have investigated a few other approaches, including the one that assumes the knowledge of the accurate mean values of subsequent classes (although these exact means are usually not available).The accuracy of each method is rated by comparing the results obtained by different methods with the exact ones. The inequality measures which are the matter of interest here are the Gini index which is the most commonly used, and also the Theil and Atkinson indexes.

Keywords: the Gini index, the Theil index, the Atkinson index, aggregated data
JEL Classifications: C4, C8, J3
DOI: 10.15611/aoe.2019.1.04

1. INTRODUCTION

Measuring the inequalities of income distribution is an important issue nowadays and the accuracy of the calculations of the measures has become a challenge to be faced (Jędrzejczak, 2012). The problem of the measurement of inequalities of income distribution is relevant to the measurement of the justice and progression of the taxation (Monti, Pellegrino and Vernizzi 2015; Mazurek, Pellegrino and Vernizzi 2015; Pellegrino and Vernizzi 2013). Besides the problems with data reliability, there is the issue of the form in which they are most commonly available. Usually one has to deal with data grouped as the frequency distribution. For comparing the tax systems in many countries, we have to work on the aggregated data as individual data is not available. A common approach is to use a linear interpolation, that is – to assume that all observations within

* Faculty of Management, Computer Science and Finance, Department of Statistics, Wrocław University of Economics.

a given interval are equal. However, this implies zero inequality within each class interval, which significantly lowers the overall inequality. To avoid this underestimation, several approaches have been proposed that are based on fitting either a probability density function or the Lorenz curve, and either a single function to the entire range or piecewise functions (Kakwani; Budd, 1970; Fisk, 1961; Kakwani, Podder; Aitchison, Brown, 1954). However, these methods have been developed for data that is grouped in a specific way, when the mean value for each class interval is available. A significant improvement in estimating inequality measures was made by Gastwirth (1972), who obtained lower and upper bounds of estimation. The applicability of this method also depends on the knowledge of the actual mean values of the observations belonging to each interval class. However, in most cases, as usual in statistical yearbooks, only the width and numbers of observations within each class are provided.

The aim of this paper is to check the accuracy of the traditional approach that treats all observations within a given class as being concentrated in the middle of it and to compare the modifications of this approach, in order to investigate whether the accuracy of the calculations of inequality measures that are based on the observed income distribution may be improved. We will compare the relative error methods of calculating these measures with the exact results, obtained from individual observations, and with the approximation that is based on the knowledge of the actual means within each range. The analysis is based on the individual data from the tax office. The data set concerns the year 2007. The inequality measures which are the matter of interest of this paper are: the Gini index, which is the most commonly used, as well as the Theil and the Atkinson indexes. The problem of measuring the inequality is particularly important for income distribution, when it is known that the distribution is extremely skewed.

The paper is organized as follows. In the next section we describe the source of individual data and the way of constructing a frequency distribution. The inequality measures are briefly defined in Sections 3 and 4 and the simplest and most common approach of calculating these measures for grouped data is presented. These sections also consider some simple modifications of this most popular, traditional approach. In Sections 5 and 6 we discuss some possible extensions based on the density function or the cumulative income distribution function. Section 7 presents the conclusions.

2. THE DATA

The analysis in this paper is based on the individual data from the Wroclaw tax office from the fiscal year of 2007. To be exact, this data set contains information on gross income for taxpayers (households) that file their tax return in the Municipality of Wroclaw, in the Fabryczna Tax Office (district identification). In this analysis, households are equated with couples of taxpayers who take advantage of joint taxation using PIT-37. After deleting observations with non-positive gross income, the whole population consists of 19,487 households. The analyses were performed by the authors' own programs, written in the *R* language and in Mathematica 7.

The population of 19,487 households is divided into five subpopulations with respect to the number of dependent children. The subpopulations are as follows:

- C– families without children: 10,625 households,
- C+1– families with one child: 5,458 households,
- C+2– families with two children: 2,935 households,
- C+3– families with three or more children: 469 families,
- ALL– the whole population with each family type (C and C+1 and C+2 and C+3).

The values of inequality indexes calculated for individual data will be regarded as exact values of these indexes and serve as a reference point for comparison with the values obtained on the basis of grouped data.

Let $x_1, x_2, \ldots, x_n$ be incomes of n income units (households) of the pre-specified subpopulation. Then, the Gini index for the individual data is defined as follows:

$$G = \frac{1}{2n^2\bar{x}} \sum_{i,j=1}^{n} \left| x_i - x_j \right|, \tag{1}$$

where $\bar{x}$ denotes the average income. Table 1 presents the Gini indexes for all five subpopulations for individual datasets. These values will be treated as exact values of the Gini index.

Table 1

The Gini index for the individual data

Family type				
C	**C+1**	**C+2**	**C+3**	**ALL**
0.37178	0.34647	0.34651	0.38701	0.36650

Source: own calculations.

Table 2

Frequency distributions of income for the analyzed population

i	**Income interval (PLN)** $(x_{i-1} - x_i)$	**Number of households –** n_i			
		Family type			
		ALL	**C**	**C+1**	**C+2**
1	0 – 15000	908	714	122	58
2	15000 – 30000	2785	1789	672	259
3	30000 – 45000	3597	2247	874	402
4	45000 – 60000	3355	1879	902	496
5	60000 – 75000	2633	1305	845	409
6	75000 – 90000	2010	905	658	398
7	90000 – 105000	1290	560	425	274
8	105000 – 125000	1072	474	356	221
9	125000 – 145000	610	263	212	115
10	145000 – 165000	396	151	116	112
11	165000 – 195000	325	130	108	82
12	195000 – 255000	273	116	94	53
13	255000 – 315000	233	92	74	29
14	above 315000	-	-	-	27
	Total	**19487**	**10625**	**5458**	**2935**

Source: own calculations.

Table 3

The frequency distribution of income for families with three or more children

i	**Income interval (PLN)** $(x_{i-1} - x_i)$	**Number of households** n_i
1	0 – 15000	14
2	15000 – 30000	65
3	30000 – 45000	74
4	45000 – 60000	78
5	60000 – 75000	74
6	75000 – 90000	49
7	90000 – 105000	31
8	105000 – 135000	30
9	135000 – 165000	28
10	165000 – 225000	11
11	225000 – 285000	7
12	2855000 – 405000	5
13	above 405000	3
	Total	**469**

Source: own calculations.

Usually the data concerning incomes is presented in aggregated form which implies the loss of information. In order to examine how much this loss influences the indexes of inequalities, the individual data for all five subpopulations is aggregated by standard rules of aggregating and presented in Tables 2 and 3. The cumulative frequency distribution tables were constructed according to theory. The individual data is grouped according to gross income. Table 2 presents aggregated data for all population and for three family types: without children, with one child and for families with two children.

An analogous frequency income distribution (with different class intervals) is created for families with three or more children and presented in Table 3.

The created distributions will be used in the next section to illustrate the calculations of the inequality measures.

3. THE GINI INDEX FOR SOME SIMPLE ALTERNATIVES OF TREATING THE AGGREGATED DATA

Suppose that n individual observations are grouped into c classes of interval form, $(x_{i-1} - x_i)$, $i = 1,\ldots,c$. Assume that n_i denotes the number of families belonging to class i, furthermore suppose that $\dot{x}_i$ denotes a mid-value of this class:

$$\left(\dot{x}_i = \frac{x_{i-1} + x_i}{2} \right),$$

and $\overline{x}'$ is the average income:

$$\overline{x}' = \frac{1}{n} \sum_{i=1}^{c} \dot{x}_i n_i .$$

Five simple alternatives for the calculation of the Gini index for the aggregated data are defined below. First, assume that all n_i observations are concentrated in the middle of the interval. Graphically this situation is presented in Figure 1a.

The Gini index for this case is calculated according to the formula:

$$G^{(m)} = \frac{1}{2n^2 \overline{x}'} \sum_{i,j=1}^{c} \left| \dot{x}_i - \dot{x}_j \right| n_i n_j . \tag{2}$$

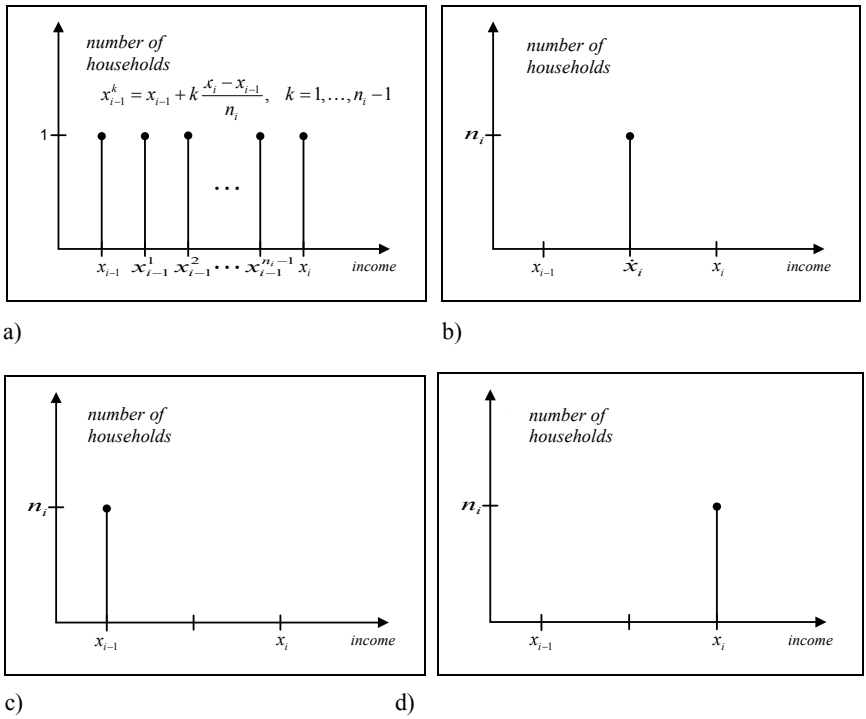

Fig. 1. Graphical presentation of representatives of *i*th income class
Source: own presentation.

In the second and third cases the n_i families are concentrated on the lower and the upper boundary of the *i*th interval, respectively (see Figures 1b and 1c). The corresponding formulae for the Gini index are the following:

$$G^{(l)} = \frac{1}{2n^2\bar{x}'}\sum_{i,j=1}^{c}\left|x_{i-1} - x_{j-1}\right| n_i n_j , \tag{3}$$

$$G^{(u)} = \frac{1}{2n^2\bar{x}'}\sum_{i,j=1}^{c}\left|x_i - x_j\right| n_i n_j . \tag{4}$$

The fourth possibility considered here consists in the assumption that all observations are distributed evenly in the interval (see Figure 1d). For each of *c* classes, the evenly distributed incomes are obtained by the formula:

$$y_{ik} = x_{i-1} + k\frac{x_i - x_{i-1}}{n_i}, \qquad k = 1,2,\ldots,n_i \; i = 1,\ldots,c .$$

In this way individual statistical series $\left(y_1, y_2, \ldots, y_n\right) = \left(y_{11}, y_{12}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}, \ldots, y_{cn_c}\right)$ could be reconstructed from aggregated data.

The Gini index for the reconstructed data is calculated using formula (1):

$$G^{(n)} = \frac{1}{2n^2\overline{y}} \sum_{i,j=1}^{n} \left| y_i - y_j \right|. \tag{5}$$

The last possibility considered here assumes that all n_i observations are distributed randomly within the interval (x_{i-1}, x_i). The uniform distribution, $U(x_{i-1}, x_i)$, is assumed – for each of c classes, the n_i incomes are reconstructed by random choosing n_i values from the distribution $U(x_{i-1}, x_i)$. For individual statistical series $\left(z_1, z_2, \ldots, z_n\right)$ obtained in such a way the measure of inequality is calculated using formula (1) and reads:

$$G^{(r)} = \frac{1}{2n^2\overline{z}} \sum_{i,j=1}^{n} \left| z_i - z_j \right|. \tag{6}$$

It may be expected that value $G^{(l)}$ obtained from formula (3) would be greater than value $G^{(u)}$ obtained with the use of upper bounds (formula (4)), while the value of $G^{(m)}$ will have the value somewhere between. This is unavoidable in the case of equally spaced intervals: for such intervals the sum of the differences in the formula for the Gini index is exactly the same in all three cases, however, the average value is the lowest for the lower boundaries of intervals and it is the greatest for the upper boundaries. As the average value is in a denominator in the expression for the Gini index, it turns out that $G^{(l)} > G^{(m)} > G^{(u)}$. In general, class intervals do not have to be equally spaced, however, it turns out that the inequality for $G^{(l)}, G^{(m)}$ and $G^{(u)}$ still holds in such cases, which will be shown afterwards.

It could be expected that inequality indexes calculated with assumptions four and five would be greater than in the case of concentrating all observations in the middle of the interval, as such approaches do not imply zero inequality within classes. One can also expect that values obtained at assuming uniform spacing of the observations and those obtained with the assumption of random spacing would converge for high frequencies.

Using these five formulas (2), (3), (4), (5) and (6), the measures of inequality are calculated for each data set presented in Section 2 (for each family type, for five empirical income distributions). The results for the Gini index are presented in Table 4.

Table 4

The Gini index for various methods of calculations

Method	Family type					Mean abs. dev.
	C	C+1	C+2	C+3	ALL	
$G^{(m)}$	0.36967 (-0.00211)	0.34151 (-0.00496)	0.33826 (-0.00825)	0.38284 (-0.00417)	0.36180 (-0.00470)	0.004838
$G^{(l)}$	0.41214 (0.04035)	0.36707 (0.02060)	0.36015 (0.01365)	0.40238 (0.01538)	0.39504 (0.02854)	0.023702
$G^{(u)}$	0.33774 (-0.03404)	0.32157 (-0.02490)	0.32083 (-0.02567)	0.36807 (-0.01893)	0.33628 (-0.03022)	0.026752
$G^{(n)}$	0.37546 (0.00368)	0.34576 **(-0.00071)**	0.34187 **(-0.00463)**	0.38712 **(0.00011)**	0.36656 **(0.00006)**	**0.001838**
$G^{(r)}$	0.37505 (0.00327)	0.34579 **(-0.00068)**	0.34109 **(-0.00542)**	0.38590 **(-0.00110)**	0.36662 **(0.00012)**	**0.002116**

Source: own calculations.

Each cell of the table contains two values (except for the last column). One is the value of the appropriate Gini index and the second (below, in the brackets) is the difference between this Gini index and the exact value of the Gini index given in Table 3. For example, for a family without children, $G^{(m)} = 0.36967$ and the difference between the exact Gini index $G = 0.37178$ is equal to $0.36967 - 0.37178 = -0.00211$. Negative values mean that the estimated Gini index has underestimated the real inequality. Bold font identifies the values which deviate in absolute value from the exact value less than the results obtained within a traditional approach (the first row, treating all values as concentrated in the centers of the income intervals). Additionally, the last column presents the average (over five cases) of the absolute deviations for each method. Bold font in this last column identifies the results that are better than those of the traditional approach.

The values of the Gini index are visualized also in Figure 2.

The values of the Gini index for the given method and different family types are joined with a line only for the legibility.

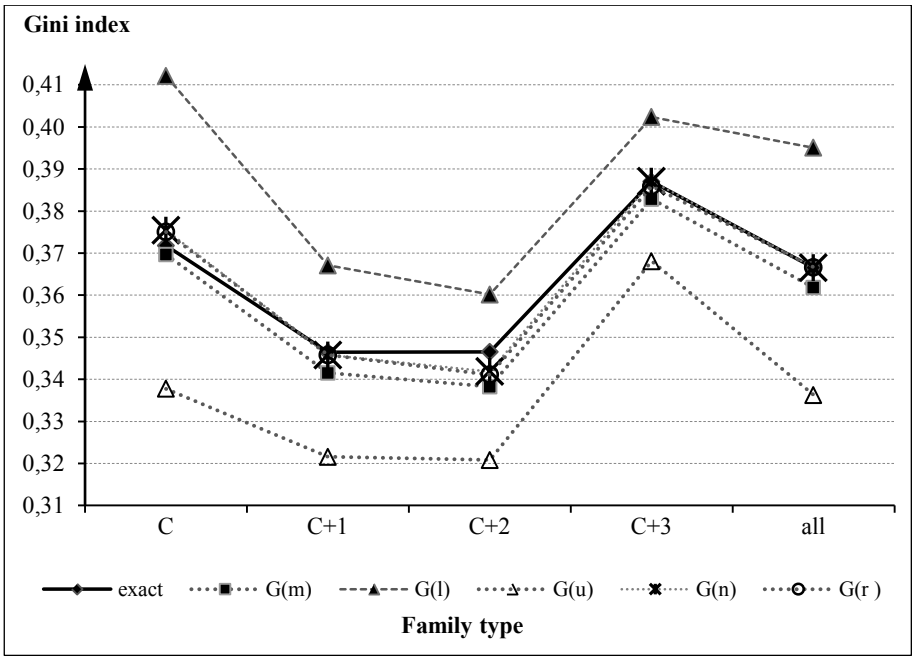

Fig. 2. The Gini index estimated from grouped data

Source: own presentation based on results from Table 4.

It may be seen that for the Gini index both methods in which all values are concentrated in the middle and at the upper limit of the given income interval, systematically underestimate the value of inequality measures. That was expected, as these methods assume zero inequality within each interval. In addition, the deviation from the exact values in the approach with the upper limit of income intervals is always larger than in the traditional approach. As for the method with all values within a given interval concentrated at the lower limit of it, in spite of the fact that it assumes zero inequality within intervals as well, the underestimation of the mean value overwhelms this effect and it systematically overestimates inequality measures. Both the methods with lower and upper limits of the interval as representatives of it give results worse than the traditional approach, as may be expected.

The methods with values within a given interval equality spaced between its boundaries $\left(G^{(n)}\right)$ and with random values within interval $\left(G^{(r)}\right)$ are both better than the traditional approach and neither of them seems to have any systematical bias (towards underestimation or overestimation) of the real value.

4. THE THEIL AND THE ATKINSON INDEXES

Analogous comparisons are performed for two other familiar indices: the Theil index and the Atkinson index.

The Theil index for individual observations is defined as (Theil, 1967):

$$T = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i}{\overline{x}}\ln\frac{x_i}{\overline{x}}, \tag{7}$$

where x_i, $\overline{x}$ and n denote, as above, an individual observation, the average value and the number of observations. For grouped data it may be approximated by:

$$T^{(m)} = \frac{1}{n}\sum_{i=1}^{c}\frac{\dot{x}_i}{\overline{x}'}\ln\frac{\dot{x}_i}{\overline{x}'}n_i, \tag{8}$$

with the same notation as above. Table 5 presents the results for the Theil index.

Table 5

The Theil index for various methods of calculations

Method	Family type					Mean abs. dev.
	C	C+1	C+2	C+3	ALL	
Exact T	0.242836	0.211639	0.221661	0.288899	0.23853	
$T^{(m)}$	0.237908 (-0.00493)	0.200026 (-0.01161)	0.197274 (-0.02439)	0.265867 (-0.02303)	0.225593 (-0.01294)	0.015379
$T^{(u)}$	0.203459 (-0.03938)	0.182354 (-0.02929)	0.181136 (-0.04053)	0.254942 (-0.03396)	0.199817 (-0.03871)	0.036372
$T^{(n)}$	0.243259 **(0.000423)**	0.20347 **(-0.00817)**	0.200083 **(-0.02158)**	0.26957 **(-0.01933)**	0.229919 **(-0.00861)**	**0.011622**
$T^{(r)}$	0.243258 **(0.000422)**	0.203514 **(-0.00813)**	0.198603 **(-0.02306)**	0.265361 (-0.02354)	0.230141 **(-0.00839)**	**0.012707**

Source: own calculations.

The last three rows of Table 5 present modifications of the calculations of the Theil index analogous with modifications of the calculations of the Gini index. For $T^{(u)}$ all values within a given class are assumed to be concentrated at the upper limit of it. For the Theil index it is not possible to apply the method which would treat all values as concentrated at the lower

limit of a given interval, because in this case it would require dealing with data equal to zero, leading to $\ln 0$. Indexes $T^{(n)}$ and $T^{(r)}$ are calculated from equation (7) assuming all values within a given class equally spaced between its boundaries or randomly taken within this class.

The last column presents the average of absolute deviations for each method. Again, the bold font identifies the results that are better than in the traditional approach (the method presented in the second row in Table 5).

The discussion and conclusions expressed for the Gini index also hold for the Theil index. Methods with values within a given interval equally spaced between its boundaries $\left(T^{(n)}\right)$ and with random values within the interval $\left(T^{(r)}\right)$ are both better than the traditional approach $\left(T^{(m)}\right)$. Both methods that assume all values to be concentrated either in the middle or at the upper limit of the given income interval, systematically underestimate the exact value of the Theil index (*T*).

Finally, the results for the Atkinson index will be presented in the same way.

The Atkinson index is defined as (Atkinson, 1970):

$$A_{\varepsilon} = 1 - \frac{1}{\bar{x}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i^{1-\varepsilon}\right]^{1/(1-\varepsilon)}, \qquad \varepsilon > 0, \quad \varepsilon \neq 1, \tag{9}$$

where x_i, $\bar{x}$ and n denote individual observations, the average value and the number of observations while ε is a parameter called "inequality aversion" or a sensitivity parameter. The higher the value of this parameter, the more sensitive the Atkinson index becomes to inequalities at the bottom of the income distribution. For aggregated data it may be approximated by:

$$A_{\varepsilon}^{(m)} = 1 - \frac{1}{\bar{x}'}\left[\frac{1}{n}\sum_{i=1}^{c} \dot{x}_i^{1-\varepsilon} n_i\right]^{1/(1-\varepsilon)}, \tag{10}$$

with these same symbols as in the previous formulae.

Four modifications of the Atkinson index $A_{\varepsilon}^{(l)}, A_{\varepsilon}^{(u)}, A_{\varepsilon}^{(n)}, A_{\varepsilon}^{(r)}$ are defined analogously to the modifications in calculations of the Gini index. Tables 6 and 7 present the results for the Atkinson index with $\varepsilon = 0.1$ and $\varepsilon = 0.5,$ respectively.

Table 6

The Atkinson index, $\varepsilon = 0.1$ for various methods of calculations

Method	Family type					Mean abs. dev.
	C	C+1	C+2	C+3	ALL	
Exact $A_{0.1}$	0.023984	0.02087	0.021701	0.02806	0.023506	
$A_{0.1}^{(m)}$	0.023628 (-0.00036)	0.019868 (-0.001)	0.019584 (-0.00212)	0.02615 (-0.00191)	0.02242 (-0.00109)	0.001294
$A_{0.1}^{(l)}$	0.030394 (0.00641)	0.023094 (0.002224)	0.022361 **(0.00066)**	0.02874 **(0.00068)**	0.027339 (0.003833)	0.002762
$A_{0.1}^{(u)}$	0.020055 (-0.00393)	0.017994 (-0.00288)	0.017882 (-0.00382)	0.024891 (-0.00317)	0.019714 (-0.00379)	0.003517
$A_{0.1}^{(n)}$	0.024214 **(0.000229)**	0.02023 **(-0.00064)**	0.01988 **(-0.00182)**	0.026537 **(-0.00152)**	0.022886 **(-0.00062)**	**0.000966**
$A_{0.1}^{(r)}$	0.024204 **(0.000219)**	0.020235 **(-0.00063)**	0.019741 **(-0.00196)**	0.026173 **(-0.00189)**	0.022904 **(-0.0006)**	**0.00106**

Source: own calculations.

As in the previous tables, below are the values calculated using a given method, showing the differences between these values and the exact values of the Atkinson index (calculated for individual data). Bold font identifies the values which deviate in the absolute value from the exact value less than the results obtained within the traditional approach from formulae $A_{0.1}^{(m)}$ for Table 6 and from $A_{0.5}^{(m)}$ for Table 7. The last column presents the average of the absolute deviations for each method. Bold font identifies results that are better (in the average sense) than those in the traditional approach.

Table 7

The Atkinson index, $\varepsilon = 0.5$ for various methods of calculations

Method	Family type					Mean abs. dev.
	C	C+1	C+2	C+3	ALL	
Exact $A_{0.5}$	0.114967	0.099267	0.10115	0.126592	0.111851	
$A_{0.5}^{(m)}$	0.115395 (0.000427)	0.096946 (-0.00232)	0.095455 (-0.00569)	0.123003 (-0.00359)	0.109712 (-0.00214)	0.002834
$A_{0.5}^{(l)}$	0.170195 (0.055228)	0.121767 (0.0225)	0.116823 (0.015673)	0.147654 (0.021061)	0.149742 (0.037891)	0.030471
$A_{0.5}^{(u)}$	0.094985 (-0.01998)	0.085503 (-0.01376)	0.085142 (-0.01601)	0.113631 (-0.01296)	0.093648 (-0.0182)	0.016184
$A_{0.5}^{(n)}$	0.120017 (0.005049)	0.09939 **(0.000123)**	0.097469 **(-0.00368)**	0.12555 **(-0.00104)**	0.113239 **(0.001388)**	**0.002257**
$A_{0.5}^{(r)}$	0.119798 (0.004831)	0.099461 **(0.000194)**	0.096873 **(-0.00428)**	0.124851 **(-0.00174)**	0.113224 **(0.001373)**	**0.002483**

Source: own calculations.

For the Atkinson measures the following inequality also holds: $A_\varepsilon^{(l)} > A_\varepsilon^{(m)} > A_\varepsilon^{(u)}$. Both the methods with the lower and upper limit of intervals as representatives of them (referring to intervals) give results worse than in the standard approach. Again, methods with values within a given interval equally spaced between its boundaries and with random values within the interval are better than the traditional approach. It may be expected that the methods for all the presented measures of inequalities should be convergent while increasing the numbers, as equally spaced values and a uniform distribution over a given interval should converge in the limit of a large number of values. Indeed, the largest differences between these two methods may be observed for empirical income distributions for families with two children (C+2) and three children (C+3), which are the least numerous of all five cases. However, it is not guaranteed that these two methods will always give better estimations of inequality measures than the standard one, as there are cases in which the standard approach is better to approximate exact values, nevertheless they are better in the great majority of the cases.

5. METHODS BASED ON A DENSITY FUNCTION

Besides the simple methods described above, one may try some more advanced approaches to calculate inequality measures based on limited (grouped) data. One possibility can be methods based on a probability density function.

Let X denote the income of a member of the population. Assume that X is a random variable with probability density function $g(x)$, and corresponding cumulative distribution function $F(x)$. Continuous counterparts of the measures of inequality defined by formulae (1), (7) and (9) are the following:
the Gini index:

$$G = \frac{1}{\mu}\int_{-\infty}^{\infty} F(x)\left[1 - F(x)\right]dx, \tag{11}$$

the Theil index:

$$T = \int_{-\infty}^{\infty} g(x)\frac{x}{\mu}\ln\frac{x}{\mu}dx, \tag{12}$$

the Atkinson index:

$$A_\varepsilon = 1 - \frac{1}{\mu}\left[\int_{-\infty}^{\infty} x^{1-\varepsilon} g(x)\,dx\right]^{1/(1-\varepsilon)}, \quad \varepsilon > 0, \quad \varepsilon \neq 1, \tag{13}$$

where μ is the expected value of the income distribution.

To use the above formulae, in the first step it is necessary to perform the estimation of probability density function $g(x)$, given only observations at a discrete set of points. We will apply six methods to approximate the probability density function based on known frequencies, f_i's, $f_i = \frac{n_i}{n}$ (if needed, scaled according to the different widths of intervals, f_i^{sc}). The methods are described below.

METHOD 1

First, we will adopt the method described by Kakwani (Kakwani, 1976), based on piecewise linear approximations of a probability density function. The method relies on fitting a piecewise linear function normalized up to the frequency in each given interval. Within the original approach described by Kakwani (see also below), we need also to know the mean value within each interval. For the data usually provided in statistical yearbooks we do not have such information. Thus, we have to follow another rule that will determine the slope of the line within each interval. There are two simple possibilities that are based on the differences of frequencies of neighboring intervals.

The first is that the slope is proportional to the difference between the next and the current intervals, which is equivalent to connecting the left upper corners of histogram rectangles, i.e. points (x_{i-1}, f_i^{sc}). After normalization we get a piecewise linear function given by:

$$f_i^{(l\_lin)}(x) = a_i^{(l\_lin)} x + b_i^{(l\_lin)} \quad \text{for } x \in (x_{i-1}, x_i], \tag{14}$$

where

$$a_i^{(l\_lin)} = \frac{2f_i}{(x_i - x_{i-1})^2}\frac{f_{i+1}^{sc} - f_i^{sc}}{f_{i+1}^{sc} + f_i^{sc}},$$

$$b_i^{(l\_lin)} = \frac{2f_i}{(x_i - x_{i-1})^2}\frac{f_i^{sc} x_i - f_{i+1}^{sc} x_{i-1}}{f_{i+1}^{sc} + f_i^{sc}}$$

and a corresponding cumulative distribution:

$$F_i^{(l\_lin)}(x) = \sum_{j=1}^{i-1} f_i + \frac{a_i^{(l\_lin)}}{2}\left(x^2 - x_{i-1}^2\right) + b_i^{(l\_lin)}\left(x - x_{i-1}\right).$$

from which we can calculate inequality measures according to formulae (11) to (13).

METHOD 2

Within the second approach to this question the slope is proportional to the difference between frequencies of the current and the previous intervals, which is equivalent to connecting the right upper corners of histogram rectangles, that is, points $\left(x_i, f_i^{sc}\right)$.

Similarly to *METHOD 1*, we get a piecewise linear function given by:

$$f_i^{(u\_lin)}(x) = a_i^{(u\_lin)} x + b_i^{(u\_lin)} \text{ for } x \in \left(x_{i-1}, x_i\right], \tag{15}$$

where

$$a_i^{(u\_lin)} = \frac{2f_i}{\left(x_i - x_{i-1}\right)^2}\frac{f_i^{sc} - f_{i-1}^{sc}}{f_i^{sc} + f_{i-1}^{sc}},$$

$$b_i^{(u\_lin)} = \frac{2f_i}{\left(x_i - x_{i-1}\right)^2}\frac{f_{i-1}^{sc} x_i - f_i^{sc} x_{i-1}}{f_{i+1}^{sc} + f_i^{sc}}$$

and corresponding cumulative distribution:

$$F_i^{(u\_lin)}(x) = \sum_{j=1}^{i-1} f_i + \frac{a_i^{(u\_lin)}}{2}\left(x^2 - x_{i-1}^2\right) + b_i^{(u\_lin)}\left(x - x_{i-1}\right),$$

from which we can again calculate inequality measures inserting these functions into the proper formulae.

It is worth mentioning that functions $f^{(l\_lin)}$ and $f^{(u\_lin)}$ are in general discontinuous. However this is not a shortcoming of this approach, inasmuch as other approaches, such as the one described by Kakwani, not to mention the standard approach (which may be considered as taking the density function in the form of peaks in the centers of intervals), are also discontinuous.

METHOD 3

So far we have described linear approaches that are linear counterparts of "point" approaches (all observations within a given interval concentrated at one value), which are based on taking the lower or upper limits of intervals. However, in the case of "point" approaches there is also the third, the most intuitive and the most commonly used approach, that is based on taking the middles of intervals. Thus, here we would like to investigate the approach which consists in connecting not the lower and not the upper limits, but the centers of intervals, that is, points $\left(\left(x_{i-1}+x_i\right)/2, f_i^{sc}\right) \equiv \left(\dot{x}_i, f_i^{sc}\right)$. In that way we will get a standard relative frequency polygon, however, it is not necessarily normalized. To be specific, the area below such a curve equals

$$A = \sum_{i=1}^{c-1}\left(\dot{x}_{i+1}-\dot{x}_i\right)\left(\min\left(f_i^{sc}, f_{i+1}^{sc}\right)+0.5\left|f_i^{sc}-f_{i+1}^{sc}\right|\right)+ \\ 0.5\left[\left(x_1-x_0\right)f_1^{sc}+\left(x_c-x_{c-1}\right)f_c^{sc}\right],$$

where $\min\left(\cdot,\cdot\right)$ denotes the smaller of the two values. We have to normalize the plot, thus obtaining the piecewise linear curve in the form:

$$fp^{(norm)}\left(x\right) = \frac{1}{A}\left(\frac{f_{i+1}^{sc}-f_i^{sc}}{\dot{x}_{i+1}-\dot{x}_i}x + \frac{f_i^{sc}\dot{x}_{i+1}-f_{i+1}^{sc}\dot{x}_i}{\dot{x}_{i+1}-\dot{x}_i}\right),$$

for $x \in \left[\dot{x}_i, \dot{x}_{i+1}\right]$, $i = 0,\ldots,c$,

where $\dot{x}_0 = \frac{3x_0 - x_1}{2}$, $\dot{x}_{c+1} = \frac{3x_c - x_{c-1}}{2}$, $f_0^{sc} = f_{c+1}^{sc} = 0$. Note that as we normalize the frequency polygon as a whole, then thus obtained normalized frequency polygon, $fp^{(norm)}$ is a continuous function.

METHOD 4

Finally, let us briefly recall here a method developed half a century ago to be applied to cases when the mean values for each interval are known. In spite of the fact that we are not able to apply it in cases of the observed income distribution considered in this paper, we will still try to make use of this method. This method, described by Kakwani (Kakwani, 1976), is based on approximating the probability density function by a piecewise linear function within all ranges but the first and last (open) ones, while for the

latter two the Pareto functions are used. There are two conditions for function $g_i(x)$ within each interval: $\int_{x=x_{i-1}}^{x_i} g_i(x) = f_i$ and $\int_{x=x_{i-1}}^{x_i} x g_i(x) = \mu_i$, where μ_i denotes the mean value for *i.* interval. Thus we get:

$$g_i(x) = \frac{2f_i}{x_i - x_{i-1}}\left(2 - 3\frac{\mu_i - x_{i-1}}{x_i - x_{i-1}}\right) + \frac{6f_i}{\left(x_i - x_{i-1}\right)^2}\left(2\frac{\mu_i - x_{i-1}}{x_i - x_{i-1}} - 1\right)\left(x - x_{i-1}\right) \quad (16)$$

for $i = 2,\ldots,c-1$, and for the first and the last income classes the functions are defined as the Pareto functions in the following way:

$$g_1(x) = \frac{\mu_1 f_1}{x_1\left(x_1 - \mu_1\right)}\left(\frac{x}{x_1}\right)^{\frac{2\mu_1 - x_1}{x_1 - \mu_1}}, \quad (17)$$

$$g_c(x) = \frac{\mu_c f_c}{x_{c-1}\left(\mu_c - x_{c-1}\right)}\left(\frac{x_{c-1}}{x}\right)^{\frac{2\mu_c - x_{c-1}}{\mu_c - x_{c-1}}}. \quad (18)$$

As was mentioned before, we are not able to apply this method directly, as we deal with a standard income distribution with the knowledge of neither mean values for particular intervals nor an exact total mean value. However we will use this method by putting the middle of an interval for the exact mean value. It may be noticed that in that way we obtain a step function in the majority of the whole range: the slope of function ρ_i depends on the difference between the middle of the class and the actual mean value for this class. If this difference is zero the slope is zero as well. Hence we regain an approximation highly similar to the uniform distribution of the values within each interval. However, there appear differences for the first and last (open) classes where the distribution of values is modeled in a different way.

METHOD 5

Going beyond linear approximations, we have also tried to approximate the probability density function by a polynomial function. We have used the Hermite interpolation method implemented in program *Mathematica*. In the case of no given values of derivatives this method boils down to the Lagrange interpolation. For the set of c given points, $\left(\left(x_{i-1} + x_i\right)/2, f_i^{sc}\right) \equiv \left(\dot{x}_i, f_i^{sc}\right)$, $i = 1,\ldots,c$, they are interpolated by the polynomial with a degree at most $c-1$ given in the form:

$$G_{c-1}(x) = \sum_{i=1}^{n} l_i(x) f_i^{sc},$$

where $l_i(x) = \dfrac{(x-\dot{x}_1)\ldots(x-\dot{x}_{i-1})(x-\dot{x}_{i+1})\ldots(x-\dot{x}_n)}{(x_i-\dot{x}_1)\ldots(x_i-\dot{x}_{i-1})(x_i-\dot{x}_{i+1})\ldots(x_i-\dot{x}_n)}$.

The results for the estimation of income inequality measures for grouped observations that are using methods based on a probability density function are presented in Table A1 in the Appendix. We will also compare these results with the values that would be obtained if the real means for particular intervals of income were known, to check how much such knowledge, which is usually not available in Polish statistical yearbooks, would improve the estimations of inequality measures. These results that use the real means are based on a method that approximates the probability density function by the piecewise linear function described above (see expressions (16)-(18)) and will be denoted in what follows as ***METHOD 6***.

The results for the Gini index are presented in Figure 3. The analogous figures for other inequality measures look alike. The values of the Gini index for the given method and a different family type are connected with a line only for better legibility.

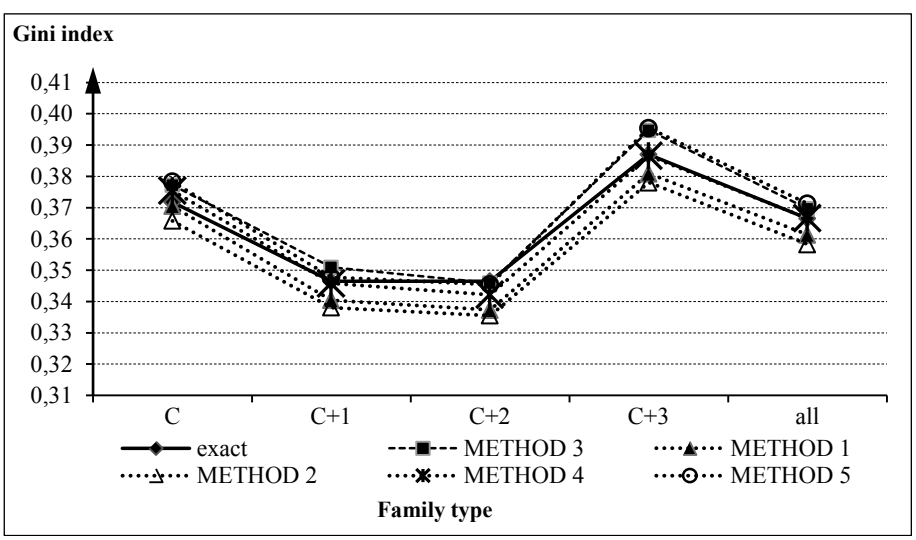

Fig. 3. The Gini index estimated by methods based on probability density functions for income.

Source: own presentation based on results from Table 8.

The linear approximations of a probability density function using the lower/upper limit of the intervals to fix the points to be fitted in analogy to the methods described in Section 2 (with all values within an interval concentrated at the lower/upper limit of it), in most cases underestimate/overestimate the real values and are not good approaches, as might be expected. METHOD 1 and METHOD 2 always underestimate the value of inequality measures. Figure 2 presents the results of these methods for the Gini index. As for fitting a piecewise linear function to points $\left(\left(x_{i-1}+x_i\right)/2, f_i^{sc}\right)$ (METHOD 3), it turns out to give a slightly better estimation, on average, than the standard approach for all indexes: G, T, $A_{0.1}$ and $\mathrm{A}_{0.5}$.

The operation of introducing a special function for open classes (METHOD 4) reduces slightly the relative error of the estimation compared to the best modification of a standard approach, that is to the method described in Section 2, which is a relevant reference point here (remember that approximating the function within METHOD 4 and the best modification of the standard approach are identical except from the open classes), for G, T, $A_{0.1}$ but not for $A_{0.5}$. Fitting the probability density function of incomes with a polynomial function in fact does not improve the obtained results. Although this approach has been marked with a bold font in Table 8, because the average result is slightly better than in the standard approach (with the middles of intervals as representatives of them), this improvement is almost zero and probably would vanish when examining more examples.

In Table A1 in the Appendix there are presented the results for METHOD 6 that are in principle not available within the scope of the task that we face here, as this method requires the knowledge of the actual mean values for each interval of income for the observed income distribution which is not accessible in normal circumstances (that is why they are printed in italics). One can see that for the Gini index this method is much better, by two orders of magnitude, than the best of the methods that are not using the knowledge of intervals' means. For other measures it is still very efficient, however one may notice that it is surprisingly not the best method in the case of $A_{0.1}$.

6. APPROACHES BASED ON THE CUMULATIVE DISTRIBUTION FUNCTION

As mentioned above, approaches that are based on approximating the probability density function may be considered as suffering a serious imperfection, as we approximate here with a function the points' coordinates

which are already approximated. However this is not the case when we are dealing with a cumulative income distribution function, as we have some fixed points which definitely belong to the real cumulative, i.e. points $\left(x_i, f_i^c\right)$, $i=0,\ldots,c$, where $f_i^c=\sum_{j=1}^{i} f_i$, $f_0^c=0$, $x_0=0$.

To approximate the cumulative income distribution function we will use:

1. piecewise linear function,
2. log-logistic distribution,
3. log-normal distribution,
4. polynomial function as a cumulative distribution.

The simplest method that could be used to approximate the cumulative density function would be to fit it with a piecewise linear function, connecting points $\left(x_i, f_i^c\right)$. One can notice that this approach is very similar to the modification of the traditional approach (described in Section 3) for which all values within an interval are assumed to be equally spaced between its boundaries. This is because for the uniform (and continuous) distribution of the values within each interval, the cumulative function increases linearly within the whole range of this interval. Within the modification of the standard approach we assume a finite number of observations equally spaced, thus the cumulative distribution will differ from a piecewise linear form (being a discontinuous step function), tending to such linear form (smoothing the steps out) with the number of observations tending to infinity. The results obtained within the piecewise linear fitting of the cumulative distribution, LINEAR, reflect this similarity, as the results for these two methods are almost identical, which will be seen in what follows.

Now we proceed to more advanced methods of fitting the cumulative distribution of incomes. We want to approximate the cumulative distribution function with one of the functions that are widely used to describe the distributions of income. We have decided to use two distributions. The first of them and the most commonly used is a log-normal distribution. The density function of it is given by:

$$PDF_{ln}(x)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{\left(\ln x-\mu\right)^2}{2\sigma^2}\right],\quad x>0, \tag{19}$$

where μ is a scale parameter and σ – a shape parameter.

The cumulative distribution cannot be expressed in terms of elementary functions and has the form:

$$CDF_{ln}(x) = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left[\frac{\ln x - \mu}{\sigma\sqrt{2}}\right], \quad x > 0, \tag{20}$$

where $erf(y) = \frac{2}{\pi}\int_0^y \exp(-t^2)dt$ denotes the error function.

The second distribution considered here is a log-logistic distribution, recently claimed as the one that may appear to be the better choice, at least in some applications (Kot and Adamkiewicz-Drwiłło, 2013). The density of it has the form:

$$PDF_{ll} = \frac{(a/b)(x/b)^{a-1}}{\left[1+(x/b)^a\right]^2}, \quad x > 0. \tag{21}$$

In contrast to the log-normal distribution, the cumulative distribution function of the log-logistic distribution is given in a simple analytical form, namely:

$$CDF_{ll}(x) = \left[1+\left(\frac{x}{b}\right)^{-a}\right]^{-1}, \quad x > 0. \tag{22}$$

We have applied the least squares method in cases of both distributions: to estimate the parameters μ and σ in the case of a log-normal distribution and a and b in the case of a log-logistic one. That is, we have minimized the sum $\sum_{i=0}^{c}\left(CDF_{ln}(x_i;\mu,\sigma) - f_i^c\right)^2$ with respect to μ and σ, and the sum $\sum_{i=0}^{c}\left(CDF_{ll}(x_i;a,b) - f_i^c\right)^2$ with respect to a and b. We have obtained the values of the parameters for which this theoretical form of a distribution is best fitted to the empirical data. The case of fitting the log-normal distribution will be denoted in what follows as *LOG_NORMAL* and fitting the log-logistic distribution case will be denoted as *LOG_LOGISTIC*. For the Gini index, in the case of the log-logistic distribution, we will not have to use formula (9), as this inequality measure is expressed in terms of one of the parameters of CDF_{ll}: $G = 1/a$. For other inequality measures, the Theil index and the Atkinson indexes, and for all indexes for the log-normal distribution, we still use the formulae that integrate over the whole range of incomes, equations (10) and (11).

Finally, we have approximated the cumulative distribution function of income by a polynomial function, using the Hermite interpolation. The results of this approach will be denoted by *POLYNOMIAL*.

Table A2 in the Appendix presents the results of the calculations of the inequality measures based on the methods described in this section.

Approximating the cumulative density function with a piecewise linear function gives better estimations of $G, T, A_{0.1}$ and $A_{0.5}$ than the traditional approach, and the results are largely similar to the best modification of the standard approach, with the assumption of equally spaced values within the given interval (see Tables 5 to 8). Figure 4 presents the Gini index for the methods based on a cumulative density function.

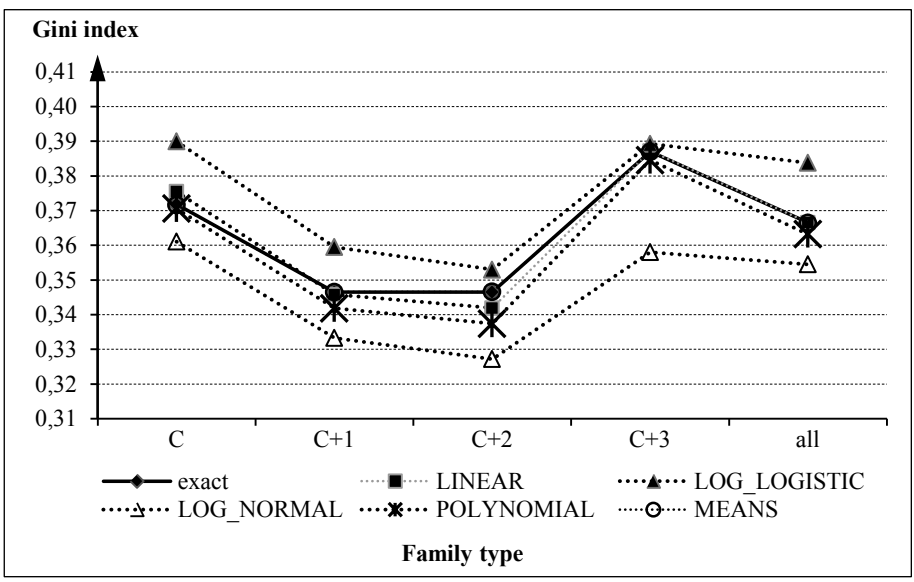

Fig. 4. The Gini index estimated by methods based on cumulative density functions for income

Source: own presentation based on the results from Table 4.

It may seem surprising that such a "coarse-grained" approach as fitting the cumulative density function with a function chosen beforehand, which in fact means in the case of both functions (17) and (19) estimating only two parameters, still gives some reasonable results. Although the average deviation from the exact values of the inequality indexes in both cases of LOG_LOGISTIC and LOG_NORMAL is larger than in the standard

approach, let us examine these results a bit closer as they seem to be very interesting, especially their comparison. Table 8 consists of the mean squared deviations for estimating distributions with log-normal and log-logistic distributions for five sets of data. It may be seen that in some cases the log-normal distribution is a better fit to the data and in the others, the log-logistic is better. This is in general reflected in the accuracy of the inequality measures, as for sets C+2 and C+3 the accuracy of the LOG_LOGISTIC method in estimating inequality is better for all four indexes, for sets C and A the accuracy of LOG_NORMAL method in inequality estimating is better for all four indexes, and for C+1 (LOG_NORMAL) is much better in two cases and very slightly worse in the remaining two cases. This last effect is opposite to the relative goodness-of-fit of the two distributions, but this inverse relationship is very slight. Therefore it is not conclusive which distribution is better to be fitted in the context of estimating inequality measures. However, a very interesting effect may be observed, namely fitting the log-logistic distribution always overestimates all the inequality measures examined here, while fitting the log-normal distribution always underestimates them. This effect does not seem to be an artifact of the choice of data, and probably deserves further investigation.

Table 8

Mean squared deviations for fitting aggregated data with log-normal and log-logistic distributions

Distribution	Family type				
	C	C+1	C+2	C+3	ALL
log-normal	$\mathbf{7.968 \cdot 10^{-5}}$	$\mathbf{10.657 \cdot 10^{-5}}$	$9.145 \cdot 10^{-5}$	$12.947 \cdot 10^{-5}$	$\mathbf{8.270 \cdot 10^{-5}}$
log-logistic	$8.690 \cdot 10^{-5}$	$11.587 \cdot 10^{-5}$	$\mathbf{7.583 \cdot 10^{-5}}$	$\mathbf{7.850 \cdot 10^{-5}}$	$10.180 \cdot 10^{-5}$

Source: own calculations.

Finally, interpolating the cumulative density function of income with a polynomial function does not improve the results, even if the numbers corresponding to this approach have been printed in bold for the Gini index and the Theil index. Formally, the average absolute deviation in these two cases is smaller than the one obtained using POLYNOMIAL, but the difference is so small, even on the scale of the differences we generally deal with here, that this approach may be regarded as obtaining the results of this same precision as the simplest standard approach, which is much more laborious.

SUMMARY AND CONCLUSIONS

Estimating inequality measures according to the simplest approach that treats all the values from a given interval as if concentrated in the center of this interval, underestimate inequalities as such an approach neglects inequality within each class. This underestimation increases as the number of classes decreases. In the examples considered in this paper where the number of classes equals 13 or 14, the relative error of estimation reached 2.4% for the Gini index, and as much as 11% for the Theil index. In spite of the fact that 2.4% (for the Gini index) may not seem to be so much in some issues, the differences in the Gini index of this level play their role, and the error of 11% (for the Theil index) always seems to be unacceptable. Thus, it would be beneficial if we had a method of the better estimation of inequality measures in cases of the need for better accuracy.

In this paper we have examined some simple methods that one can apply when dealing with a standard frequency distribution of income, i.e. knowing only the limits of intervals and their frequencies. We have shown that some of the methods considered here may improve the quality of estimations, sometimes as much as reducing the relative error a few times. However it seems that nothing can compare with the precision of the results of the method that deals with the known means of the intervals. Such considerations could serve as an argument in discussion over the form in which statistical offices provide the data. It also may be the argument in discussion over the inequality measure to be commonly used. Nowadays the most common one is the Gini index, however it has been often criticized and different measures are proposed, e.g. J. Galbraith very strongly recommends the use of the Theil index (Galbraith, 2009). In spite of the many advantages of this measure, we have to consider the problem of errors that are made when computing the Theil index on the basis of the standard frequency distribution of income. We have seen that even when applying to the Theil index the best of the methods investigated here, the relative error ranges from 1% to 7%, where this upper value still seems to be too great. If having a standard frequency income distribution, without knowledge of the intervals' means, the Gini index seems still to be the most reliable measure (among the measures examined here), and is also the one most liable to improve its accuracy while applying some more advanced methods.

REFERENCES

Aitchison, J., Brown, J., *On criteria for descriptions of income distribution*, "Metroeconomica" 6, pp. 88-107, 1954.

Aitkinson, A. B., *On the measurement of inequality*, "Journal of Economic Theory", 2 (3), pp. 244-263, 1970.

Budd, E. C., *Postwar changes in the size distribution of income in the U.S.*, "The American Economic Review" 60, pp. 247-260, 1970.

Fisk, P. R., *The graduation of income distributions*, "Econometrica" 29, pp. 171-185, 1961.

Galbraith, J. K., *Inequality, unemployment and growth: New measures for old controversies*, "The Journal of Economic Inequality" 7, pp. 189-206, 2009.

Gastwirth, J., *The estimation of the Lorenz curve and Gini index*, "The Review of Economics and Statistics", vol. 54, issue 3, pp. 306-16, 1972.

Jędrzejczak, A., *Income inequality analysis in Poland on the basis of household budget survey*, "Statistical Methods in Regional and Social Analyses under Integration and Globalization", pp. 131-149. Zakład Wydawnictw Statystycznych, Warsaw 2012.

Kakwani, N. C., *On the estimation of income inequality measures from grouped observations*, "The Review of Economic Studies" 43, pp. 483-492, 1976.

Kakwani, N. C., Podder, N., *Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations*, "Econometrica" 44, pp. 137-148, 1976.

Kakwani, N. C., Podder, N., *On the estimation of Lorenz curves from grouped observations*, "International Economic Review" 14, pp. 278-292, 1973.

Kot, S., Adamkiewicz-Drwiłło, H., *Rekonstrukcja światowego rozkładu dochodów na podstawie minimalnej informacji statystycznej* [*Reconstruction of world income distribution based on minimal statistical information*], „Śląski Przegląd Statystyczny", 11 (17), pp. 179-200, 2013.

Monti, M. G., Pellegrino, S., Vernizzi, A., *On Measuring Inequity in Taxation Among Groups of Income Units, Review of Income and Wealth*, "International Association for Research in Income and Wealth", vol. 61(1), pp. 43-58, 2015.

Mazurek, E., Pellegrino, S., Vernizzi, A., *Horizontal Inequity Estimation: The Issue of Close Equals Identification*, "Economia Politica" vol. XXX, no. 2, pp. 185-202, 2013.

Pellegrino, S., Vernizzi, A., *On Measuring Violations of the Progressive Principle in Income Tax Systems*, "Empirical Economics", 45, pp. 239-245, 2013.

Theil, H., *Economics and information theory*, vol. 7. North-Holland, Amsterdam 1967.

Received: February 2016, revised: December 2017

APPENDIX

Table A1

The inequality measures for methods of calculations based on the estimation of a probability density function of income

Method	Family type					Mean abs. dev.
	C	C+1	C+2	C+3	all	
1	2	3	4	5	6	7
GINI INDEX						
exact	0.37178	0.34647	0.34651	0.38701	0.36650	
METHOD 1	0.370491 **(-0.001289)**	0.340449 **(-0.00602)**	0.337316 (-0.00919)	0.380903 (-0.00611)	0.361354 (-0.00515)	0.005551
METHOD 2	0.365785 (-0.005995)	0.338089 (-0.00838)	0.335529 (-0.01098)	0.378041 (-0.00897)	0.35827 (-0.00823)	0.008511
METHOD 3	0.37721 (0.00543)	0.35091 **(0.00443)**	0.34592 **(-0.00059)**	0.39468 (0.00767)	0.36963 **(0.00313)**	**0.004250**
METHOD 4	0.37540 (0.00362)	0.34580 **(-0.00067)**	0.34219 **(-0.00432)**	0.38654 **(-0.00047)**	0.36659 **(0.00009)**	**0.001836**
METHOD 5	0.37828 (0.00650)	0.34776 **(0.00128)**	0.34546 **(-0.00105)**	0.39540 (0.00839)	0.37127 (0.00477)	**0.004398**
METHOD 6	*0.37161* ***(-0.00017)***	*0.34653* ***(0.00005)***	*0.34656* ***(0.00006)***	*0.38693* ***(-0.00008)***	*0.36647* ***(-0.00003)***	***0.000078***
THEIL INDEX						
exact	0.242836	0.211639	0.221661	0.288899	0.23853	
METHOD 1	0.234341 (-0.0085)	0.195051 (-0.01659)	0.193264 (-0.0284)	0.258313 (-0.03059)	0.220903 (-0.01763)	0.020339
METHOD 2	0.228805 (-0.01403)	0.193227 (-0.01841)	0.191783 (-0.02988)	0.255887 (-0.03301)	0.21784 (-0.02069)	0.023205
METHOD 3	0.24529 **(0.002454)**	0.209638 **(-0.002)**	0.204309 **(-0.01735)**	0.278992 **(-0.00991)**	0.233484 **(-0.00505)**	**0.007352**
METHOD 4	0.244092 **(0.001256)**	0.204544 **(-0.0071)**	0.20064 **(-0.02102)**	0.296962 **(0.008063)**	0.230932 **(-0.0076)**	**0.009007**
METHOD 5	0.247324 **(0.004488)**	0.205168 **(-0.00647)**	0.185343 (-0.03632)	0.278744 **(-0.01016)**	0.2363 **(-0.00223)**	**0.011933**
METHOD 6	*0.242968* ***(0.000132)***	*0.213758* ***(0.002119)***	*0.22024* ***(-0.00142)***	*0.296887* ***(0.007988)***	*0.238867* ***(0.000337)***	***0.002399***
ATKINSON INDEX, $\varepsilon = 0.1$						
exact	0.023984	0.02087	0.021701	0.02806	0.023506	
METHOD 1	0.0233184 -0.000666	0.0194002 -0.00147	0.019203 -0.0025	0.0254475 -0.00261	0.021989 -0.00152	0.001753
METHOD 2	0.0227038 (-0.00128)	0.0191821 (-0.00169)	0.019029 (-0.00267)	0.0251597 (-0.0029)	0.0216328 (-0.00187)	0.002083
METHOD 3	0.024363 (0.000379)	0.020825 **(-4.5E-05)**	0.02029 **(-0.00141)**	0.027453 **(-0.00061)**	0.023204 **(-0.0003)**	**0.000549**
METHOD 4	0.024283 **(0.000298)**	0.020322 **(-0.00055)**	0.019928 **(-0.00177)**	0.028649 **(0.000589)**	0.022971 **(-0.00053)**	**0.000749**
METHOD 5	0.024526 (0.000541)	0.020342 **(-0.00053)**	0.018634 (-0.00307)	0.027397 **(-0.00066)**	0.023435 **(-7E-05)**	**0.000974**
METHOD 6	*0.02573* *(0.001746)*	*0.023636* *(0.002766)*	*0.021583* ***(-0.00012)***	*0.028706* ***(0.000646)***	*0.023544* ***(3.78E-05)***	***0.001063***

1	2	3	4	5	6	7
ATKINSON INDEX, $\varepsilon = 0.5$						
exact	0.114967	0.099267	0.10115	0.126592	0.111851	
METHOD 1	0.115096 (0.000129)	0.0952264 (-0.00404)	0.0939841 (-0.00717)	0.12054 (-0.00605)	0.108501 (-0.00335)	0.004148
METHOD 2	0.110491 (-0.004476)	0.0933822 (-0.00588)	0.0925345 (-0.00862)	0.118179 (-0.00841)	0.105539 (-0.00631)	0.00674
METHOD 3	0.119383 (0.004416)	0.101811 (0.002544)	0.099145 **(-0.002)**	0.129501 **(0.002909)**	0.113822 **(0.001971)**	**0.002769**
METHOD 4	0.120185 (0.005218)	0.099636 **(0.000369)**	0.097624 **(-0.00353)**	0.12742 **(0.000828)**	0.113445 **(0.001594)**	**0.002307**
METHOD 5	0.119271 (0.004304)	0.098398 **(-0.00087)**	0.097998 **(-0.00315)**	0.128203 **(0.001611)**	0.113807 **(0.001956)**	**0.002378**
METHOD 6	*0.116122* *(0.001155)*	*0.100978* ***(0.001711)***	*0.100889* ***(-0.00026)***	*0.127471* ***(0.000879)***	*0.111904* ***(5.29E-05)***	***0.000812***

Source: own calculations.

Table A2

The inequality measures for methods of calculations based on a cumulative distribution function of income

Cumulative distribution function	**Family type**					**Mean abs. dev.**
	C	**C+1**	**C+2**	**C+3**	**all**	
1	2	3	4	5	6	7
GINI INDEX						
exact	0.37178	0.34647	0.34651	0.38701	0.36650	
LINEAR	0.37546 (0.00368)	0.34577 **(-0.00071)**	0.34189 **(-0.00462)**	0.38722 **(0.00021)**	0.36656 **(0.00006)**	**0.001856**
LOG_LOGISTIC	0.38996 (0.01817)	0.35947 (0.01299)	0.35299 **(0.00648)**	0.38926 **(0.00226)**	0.38377 (0.01726)	0.011434
LOG_NORMAL	0.361042 (-0.01074)	0.333285 (-0.01319)	0.327171 (-0.01934)	0.357994 (-0.02902)	0.354487 (-0.01201)	0.016858
POLYNOMIAL	0.37058 **(-0.00120)**	0.34180 **(-0.00468)**	0.33741 (-0.00909)	0.38457 **(-0.00244)**	0.36317 **(-0.00334)**	**0.004150**
THEIL INDEX						
exact	0.242836	0.211639	0.221661	0.288899	0.23853	
LINEAR	0.243278 **(0.000442)**	0.203501 **(-0.00814)**	0.200136 **(-0.02153)**	0.270105 **(-0.01879)**	0.229928 **(-0.0086)**	**0.0115**
LOG_LOGISTIC	0.294699 (0.051863)	0.243865 (0.032226)	0.233918 **(0.012257)**	0.293452 **(0.004553)**	0.283814 (0.045284)	0.029236
LOG_NORMAL	0.220108 (-0.02273)	0.185468 (-0.02617)	0.178311 (-0.04335)	0.216128 (-0.07277)	0.211604 (-0.02693)	0.038389
POLYNOMIAL	0.23703 (-0.00581)	0.198503 (-0.01314)	0.194088 (-0.02757)	0.271156 **(-0.01774)**	0.226239 **(-0.01229)**	**0.01531**

Table A2, cont.

1	2	3	4	5	6	**7**
ATKINSON INDEX, $\varepsilon = 0.1$						
exact	0.023984	0.02087	0.021701	0.02806	0.023506	
LINEAR	0.024216 **(0.000231)**	0.020234 **(-0.00064)**	0.019886 **(-0.00181)**	0.026588 **(-0.00147)**	0.022887 **(-0.00062)**	**0.000954**
LOG_LOGISTIC	0.028555 (0.004571)	0.023757 (0.002887)	0.022812 **(0.001111)**	0.028439 **(0.000379)**	0.027532 (0.004027)	0.002595
LOG_NORMAL	0.0217703 (-0.00221)	0.0183759 (-0.00249)	0.017673 (-0.00403)	0.0213809 (-0.00668)	0.0209381 (-0.00257)	0.003597
POLYNOMIAL	0.023486 **(-0.0005)**	0.019626 (-0.00124)	0.01923 (-0.00247)	0.0265 **(-0.00156)**	0.022407 (-0.0011)	0.001374
ATKINSON INDEX, $\varepsilon = 0.5$						
exact	0.114967	0.099267	0.10115	0.126592	0.111851	
LINEAR	0.120035 (0.005068)	0.099421 **(0.000154)**	0.097522 **(-0.00363)**	0.125883 **(-0.00071)**	0.113249 **(0.001398)**	**0.002191**
LOG_LOGISTIC	0.128306 (0.013339)	0.108603 (0.009336)	0.104642 **(0.003492)**	0.127838 **(0.001246)**	0.124163 (0.012312)	0.007945
LOG_NORMAL	0.104214 (-0.01075)	0.0885642 (-0.0107)	0.0852965 (-0.01585)	0.10243 (-0.02416)	0.100397 (-0.01145)	0.014585
POLYNOMIAL	0.113712 (-0.00126)	0.093474 (-0.00579)	0.092711 (-0.00844)	0.121135 (-0.00546)	0.108023 (-0.00383)	0.004955

Source: own calculations.