Estimation of Concentration Measures and Their Standard Errors for Income Distributions in Poland

Alina Jędrzejczak

Published online: 24 May 2012

© The Author(s) 2012. This article is published with open access at Springerlink.com

Abstract Measures of concentration (inequality) are often used in the analysis of income and wage size distributions. Among, them the Gini and Zenga coefficients are of greatest importance. It is well known that income inequality in Poland increased significantly in the period of transformation from a centrally planned economy to a market economy. High income inequality can be a source of serious problems, such as increasing poverty, social stratification, and polarization. Therefore, it seems especially important to present reliable estimates of income inequality measures for a population of households in Poland in different divisions. In this paper, some estimation methods for Gini and Zenga concentration measures are presented together with their application to the analysis of income distributions in Poland by socio-economic groups. The basis for the calculations was individual data coming from the Polish Household Budget Survey conducted by the Central Statistical Office. The standard errors of Gini and Zenga coefficients were estimated by means of the bootstrap and the parametric approach based on the Dagum model.

Keywords Income distribution · Income inequality · Variance estimation

JEL C10 · J30

Introduction

Measures of inequality are widely used to study income, welfare, and poverty issues. They can also be helpful to analyze the efficiency of a tax policy or to measure the level of social stratification and polarization. They are most

A. Jędrzejczak (✉)

Chair of Statistical Methods, Faculty of Economics, University of Lodz, Lodz, Poland
e-mail: jedrzej@uni.lodz.pl

frequently applied to dynamic comparisons (comparing inequality across time). The Gini concentration coefficient based on the Lorenz curve is the most widely used measure of income inequality. The Zenga point concentration measure, based on the Zenga curve, has recently received some attention in the literature.

The true values of income inequality coefficients are usually unknown and they can only be estimated on the basis of sample data coming from household budget surveys. Estimators of concentration coefficients are usually nonlinear, thus their standard errors cannot be obtained easily. The methods of variance estimation that can solve this problem include: various replication techniques, Taylor expansion, and parametric procedures based on income distribution models.

The main objective of the paper is to use survey data to analyze income inequality in Poland by socio-economic groups by means of selected concentration measures and their decomposition. This approach can further be used to assess relative economic affluence of one subpopulation with respect to another and to estimate stratification indices. To complete the analysis, some variance estimation techniques that can be used to estimate the standard errors of Gini and Zenga inequality measures should also be presented and applied.

Estimation of Income Concentration Measures

Out of the many income concentration measures, the Gini index is the most popular, mainly due to its good statistical properties and straightforward economic interpretation. The Gini index of inequality can be defined as double the area between the Lorenz curve and the line of equal shares. The Lorenz curve is expressed in Fig. 1. The line at 45° represents perfect equality of incomes and the

Fig. 1 Lorenz concentration curve

area between this line and the Lorenz curve is called concentration area. The Gini index can be expressed as follows:

$$G = 2 \int_0^1 (p - L(p)) dp \quad (1)$$

where:

$L(p)$ the Lorenz function

$p=F(y)$ cumulative distribution function of income.

The Gini index can also be interpreted as an average gain to be expected if an economic unit had a choice between its own income and that of another economic unit selected at random, relative to the mean income. One can estimate the value of the Gini index from the survey data using the following formula:

$$\hat{G} = \frac{2 \sum_{i=1}^n \left(w_i y_{(i)} \sum_{j=1}^i w_j \right) - \sum_{i=1}^n w_i y_{(i)}}{\left(\sum_{i=1}^n w_i \right) \sum_{i=1}^n w_i y_{(i)}} - 1 \quad (2)$$

where:

$y_{(i)}$ household incomes in a non-decending order,

w_i survey weight for i -th economic units, and

$\sum_{j=1}^i w_j$ rank of j -th economic unit in n -element sample.

The total Gini ratio calculated for a population of size n divided into k subpopulations can be decomposed as follows (Dagum 1997):

$$\begin{aligned} G &= G_w + G_b + G_t \\ G_w &= \sum_{j=1}^k G_j p_j s_j \\ G_b &= \sum_{j=2}^k \sum_{h=1}^{j-1} G_{jh} (p_j s_h + p_h s_j) D_{jh} \\ G_t &= \sum_{j=2}^k \sum_{h=1}^{j-1} G_{jh} (p_j s_h + p_h s_j) (1 - D_{jh}) \end{aligned} \quad (3)$$

The component G_w is the contribution of within-groups' inequality to the Gini index and G_b is the contribution of net between-groups' inequality, while G_t denotes the contribution of populations overlapping, also called transvariation. The terms p_j and s_j denote the population and income shares of the j -th subpopulation, respectively. The term D_{jh} , called either economic distance ratio or REA, plays a crucial role in the decomposition (3), and can be regarded as the measure of relative economic affluence of the j -th subpopulation with respect to the h -th subpopulation:

Another interesting measure of income inequality based on a concentration curve was proposed by Zenga (1990). It is called point concentration measure, being sensitive to changes of inequality in each part (point) of a population (Kleiber and Kotz 2003).

The Zenga synthetic inequality index Z can be expressed as the area below the Zenga curve Z_p , which is based on the relation between income and population quantiles (see: Fig. 2):

$$Z = \int_0^1 Z_p dp \quad (4)$$

The area below the Z_p curve representing the concentration area is equal to 1 in the case of perfect concentration, and takes value 0 when all incomes are equal. The Zenga curve does not represent the forced behavior, as does the Lorenz curve, so it can take various shapes depending on the underlying income distribution model.

The commonly used nonparametric estimator of the Zenga index (4) was introduced by Aly and Hervas (1999) and can be expressed by the following equation:

$$\hat{Z} = 1 - \frac{1}{ny} \left[y_{1:n} + \sum_{j=1}^{n-1} y_j \left\langle \frac{\sum_{i=1}^j y_{i:n}}{y} \right\rangle_n \right] \quad (5)$$

where:

$y_{i:n}$ i -th order statistics in n -element sample based on weighted data, and
 y sample arithmetic mean.

Fig. 2 Zenga concentration curve

The estimator (5) has been proven to be consistent and asymptotically normally distributed.

Estimation of Standard Errors

The precision of estimation seems very important from the point of view of statistical inference that should be the part of each statistical analysis based on random samples. The reliable values of estimation errors are necessary to conduct statistical inference methods, in particular to verify statistical hypothesis and construct confidence intervals. The following remarks can be helpful to realize the importance of the problem to be discussed in this subsection:

- The precision of an estimator T_n is usually discussed in terms of its sampling variance $D^2(T_n)$ or its standard error being simply the square root of the variance.
- In many cases, the exact value of sampling variance is unknown, because it depends on unknown population quantities
- After survey data have been obtained, however, an estimate of the variance $\widehat{D}^2(\widehat{\theta})$ can be calculated.
- For most income concentration measures (Gini and Zenga indices among them), explicit variance estimators are theoretically complicated—it is hard to derive general mathematical formulas for nonlinear statistics, especially when the sampling design is complex.

Thus, many approximate techniques for variance estimation can be used to obtain standard errors of income inequality measures, including: (Wolter 2003)

- Taylor linearization technique,
- Random groups method,
- Balanced Half Samples (BHS), also called Balanced Repeated Replication (BRR),
- Jackknife technique,
- Bootstrapping,
- Parametric approach based on maximum likelihood theory,
- Generalized Variance Function (GVF)- first applied in Current Population Survey CPS in 1947.

In the context of inequality measures Taylor linearization, the jackknife, and the bootstrap are the methods of variance estimation most often applied (see: Verma and Betti 2005; Davidson 2009; Kordos and Zięba 2010).

The Taylor linearization technique approximates the nonlinear estimator T_n by a pseudoestimator $g(Y)$ which is a linear function of sample observations. It is based on the first-order Taylor expansion around a parameter θ and neglecting the remainder term:

$$g(Y) \approx g(\theta) + \sum_{i=1}^k g'_i(\theta) (Y_i - \theta_i) \quad (6)$$

We often use the variance of the linearized statistic $g(Y)$ as the approximation of the true estimator variance:

$$D^2[g(Y)] \approx \sum_{i=1}^k g'_i(\theta) V(Y_i) + \sum_{i>j} g'_i(\theta) g'_j(\theta) \text{Cov}(Y_i, Y_j) \quad (7)$$

where:

$g'(\theta)$ first derivative of a function $g(\theta)$
 $V(Y_i)$ variance of a random variable Y_i
 $\text{cov}(Y_i, Y_j)$ covariance between variables Y_i and Y_j .

The jackknife technique was originally developed by Quenouille to reduce the bias of an estimator in a finite-population context. The jackknife method starts with partitioning the original sample into L dependent groups of equal size. Next, for each group, the estimator T_l (called pseudovalue) is calculated based on the data that remain after omitting the l -th group:

$$T_l = LT_n - (L - 1)T_{(l)} \quad (8)$$

The jackknife variance estimator is defined as:

$$\hat{D}_J^2(T_n) = \frac{1}{L(L-1)} \sum_{l=1}^L (T_l - T_Q)^2 \quad (9)$$

where:

$T_{(l)}$ the value of T based only on the data that remain after omitting the l -th group,
 T_Q jackknife estimator of θ defined as the simple arithmetic mean of pseudovalues, and
 L number of jackknife samples.

The bootstrap method, similar to the jackknife method, was introduced outside the field of survey sampling as a means of obtaining approximate variance estimates and confidence intervals. After drawing a series of N independent resamples (called bootstrap samples) by a design identical to the one by which the sample was drawn from the population, we calculate estimators T_k^* , $k=1 \dots N$. The bootstrap variance estimator is defined as:

$$\hat{D}_B^2(T_n) = \frac{1}{N-1} \sum_{k=1}^N (T_k^* - \bar{T}^*)^2. \quad (10)$$

Provided that the probability distribution of a variable of interest Y can be approximated by a theoretical distribution model, the method of variance estimation based on maximum likelihood theory can also be used. Let us assume that:

- an inequality measure of interest can be expressed as a function $g(\theta)$ of the parameters θ of an income distribution model given by a density function $f(y, \theta)$,
- the density function is well fitted to data, and
- the ML (maximum likelihood) estimates T_n of the parameters θ can be obtained.

According to the classical estimation theory, the ML estimators are asymptotically unbiased and normally distributed with variances given by the Cramer-Rao bound. The variance of the ML estimator of an inequality measure $g(\theta)$ takes the form:

$$D^2[g(\hat{\theta})] = \left[\frac{\partial g(\theta)}{\partial \theta} \right]^T \mathbf{I}_{\theta}^{-1} \left[\frac{\partial g(\theta)}{\partial \theta} \right] \quad (11)$$

where: $\mathbf{I}_{\theta}$ denotes the Fisher information matrix.

Application

The results of the calculations were obtained on the basis of the data coming from the Polish Household Budget Survey (HBS) for the years 2006 and 2008. In 2006 the randomly selected sample covered 37,508 households, i.e., approximately 0.3 % of the total number of households, while in 2006 the total sample size was 37,584. The samples were selected by two-stage stratified sampling with unequal inclusion probabilities for primary sampling units. In order to maintain the relation between the structure of the surveyed population and the socio-demographic structure of the total population, data obtained from the HBS were weighted with the structure of households by number of persons and class of locality coming from the Population and Housing Census 2002. The basic analysis presented in the paper was conducted after dividing the overall sample by socio-economic group, constructed according to the exclusive or main source of maintenance.

First, according to the formulas (2), (3), and (5), the estimates of Gini and Zenga inequality measures were calculated and the Gini index was decomposed into between and within-groups inequality. Then, the estimates of their standard errors were obtained using two variance estimation methods: bootstrapping and parametric approach. The estimation of Gini and Zenga coefficients for the entire population was also carried out. As a theoretical distribution model, the Dagum type-I function was used (see: Dagum 1977).

Table 1 shows the estimates of Gini and Zenga coefficients together with their standard errors calculated by means of the bootstrap method. The number of bootstrap replicates was $N=5000$. It can be easily noticed that the values of Zenga indices for socio-economic groups in Poland vary from 0.25 to 0.49, while the Gini coefficients take values from 0.29 to 0.43. Thus the Zenga coefficient seems to be more sensitive to differences between family incomes than the Gini one. The standard errors are significantly higher for the Zenga coefficient, being usually 3-6 % of the estimated values. The relative dispersion of the Gini index is usually 1-5 %. Additionally, Figs. 3 and 4 show that despite relatively small number of repetitions, the distributions of both inequality statistics can be approximated by the normal density curves.

Tables 2 and 3 contain the results concerning Gini index decomposition by socio-economic groups. In 2008, the intragroup inequality (that is, the within-group

Table 1 Estimated values of Gini and Zenga inequality measures by socio-economic group and bootstrap estimates of their standard errors (first row-2006, second row- 2008)

Socio-economic group	Gini index $\hat{G}$	Standard error of $\hat{G}$	Coeff. of Variation CV	Zenga index $\hat{Z}$	Standard error of $\hat{Z}$	Coeff. of Variation CV
1. Employees	0.29	0.0043	0.0150	0.25	0.0073	0.0293
	0.29	0.0065	0.0222	0.26	0.0130	0.0504
2. Farmers	0.40	0.0145	0.0359	0.45	0.0266	0.0595
	0.43	0.0181	0.0423	0.49	0.0333	0.0674
3. Self-employed	0.36	0.0198	0.0551	0.38	0.0356	0.0938
	0.32	0.0132	0.0412	0.31	0.0224	0.0729
4. Retirees and pensioners	0.29	0.0038	0.0132	0.24	0.0062	0.0255
	0.30	0.0046	0.0154	0.25	0.0084	0.0327
5. Non-earned sources	0.36	0.0335	0.0928	0.38	0.0615	0.1595
	0.36	0.0185	0.0508	0.38	0.0324	0.0847
Total	0.34	0.0042	0.0124	0.33	0.0079	0.0240
	0.35	0.0045	0.0132	0.34	0.0093	0.0275

Source: Author's calculations

component G_w) accounted for 32 % of the overall inequality in Poland. The within-group component reflects the inner polarization of all the groups: what causes remarkable differentials in average income between managers and blue-collar workers within the group of employees, between entrepreneurs and the others within the group of self-employed, or between retirees and pensioners within the fourth group. Table 2 can also be helpful to answer the question to what extent particular groups contribute to the overall inequality. Because of very small income and population shares, the income disparities among the self-employed weigh only 0.6 % on the total inequality, while the contribution of farmers is even smaller being 0.5 %. The group with the highest share (24 %) in the overall Gini index is the group of employees.

Fig. 3 Bootstrap distribution of Gini index estimator ($N=5000$)

Fig. 4 Bootstrap distribution of Zenga index estimator ($N=5000$)

The net between-groups component G_b contributes 43 % of the total Gini coefficient. The highest value of economic distance ratio was observed between non-earned sources and self-employed ($D=0,88$)—the economic situation of self-employed is 88 % better than the non-earned sources (see: Table 3). The transvariation component G_t describing the overlapping of the subpopulations accounts for the remaining 24 % of the total income inequality in Poland.

In Table 4, the results of the estimation obtained using the parametric approach are presented. The theoretical income distribution model was Burr type-III function, also called the Dagum distribution. The model parameters λ , β , δ were estimated using the maximum likelihood method, demanding the solution of a nonlinear system of equations. The maximum likelihood estimates of Gini and Zenga coefficient and ML estimates of their corresponding standard errors follow the similar regularities as the ones observed for the bootstrap approach presented in Table 1.

Table 2 Income inequality decomposition by subpopulations in 2008 (socio-economic groups)

1. Between-group inequality G_b	0.1479 (43 %)
2. Within-group inequality G_w	0.1132 (32 %)
Contribution of	
– employees	0.0854 (24.0 %)
– farmers	0.0014 (0.5 %)
– self-employed	0.0021 (0.6 %)
– pensioners and retirees	0.0240 (7.0 %)
– non-earned sources	0.0003 (0.0 %)
3. Transvariation G_t	0.0829 (24 %)
4. Total income inequality G	0.3440 (100 %)

Source: Author's calculations

Table 3 Average family income and economic distance ratios for socio-economic groups in Poland in 2006

No. j	Socio-economic group	Mean income [PLN]	Economic distance ratio D_{jh}				
			1	2	3	4	5
1	Employees	2,944	0.00	0.34	0.42	0.65	0.78
2	Farmers	3,644	0.34	0.00	0.35	0.75	0.83
3	Self-employed	3,955	0.42	0.35	0.00	0.82	0.88
4	Pensioners, retirees	1,907	0.65	0.75	0.82	0.00	0.32
5	Non-earned sources	1,585	0.78	0.83	0.88	0.32	0.00

Source: Author's calculations

Concluding Remarks

The paper considered the problem of efficient estimation of inequality indices on the basis of random samples, including the measurement of inequality within and between subpopulations. Reliable estimates of inequality indices are usually available only on the national level, whereas in this paper, the detailed results for socio-economic groups were presented. They can be helpful to identify the sources of income inequality and poverty in Poland.

The results of the calculations presented in the paper reveal that the level of income inequality in Poland is high, as compared with many other European countries, especially for some socio-economic groups. The main component of income inequality in Poland, when measured by the Gini index, is economic disparity between socio-economic groups.

Table 4 Parametric estimates of the Gini and Zenga inequality measures and their standard errors based on the Dagum model parameters

Socio-economic group	Year	Parameters			Goodness-of-fit	Gini index	CV [%]	Zenga index	CV [%]
		λ	β	δ					
1. Employees	2006	27.3830	0.9572	3.4436	0.9753	0.2934	1.4	0.2594	2.5
	2008	63.5020	0.9445	3.4498	0.9704	0.2939	1.4	0.2601	2.5
2. Farmers	2006	21.2122	0.7441	2.5230	0.9441	0.4231	4.3	0.4872	6.8
	2008	359.5840	0.3681	3.5045	0.9543	0.3922	3.8	0.4320	4.9
3. Self-employed	2006	54.1232	0.8122	3.2129	0.9524	0.3275	3.8	0.3159	6.7
	2008	165.7337	0.7905	3.4738	0.9527	0.3058	3.6	0.2796	6.4
4. Pensioners and retirees	2006	4.6359	1.0756	3.2315	0.9402	0.3045	1.6	0.2776	3.0
	2008	5.3830	1.1699	3.0939	0.9240	0.3127	1.7	0.2916	3.1
5. Non-earned sources	2006	6.8157	0.5471	3.5911	0.9547	0.3322	4.0	0.3256	5.3
	2008	7.1906	0.6218	3.0583	0.9665	0.3697	5.3	0.3907	8.3
Total	2006	11.7510	0.9056	3.4928	0.9685	0.3407	1.0	0.3387	1.7
	2008	27.6980	0.7937	3.0316	0.9634	0.3524	1.6	0.3461	1.7

Source: Author's calculations

The high value of the overlapping component suggests that the socio-economic groups are not separated perfectly, so they cannot be regarded as strata.

In general, the inequality estimation was more efficient when the Gini index was applied, which resulted in fewer errors of estimates. On the other hand, the synthetic Zenga measure seemed more sensitive to slight changes of income inequality within the groups of households. Thus, it is clear that both inequality coefficients, accompanied by the measures of their precision, can be regarded as useful tools in income distribution analysis.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

References

- Aly, E. A., & Hervas, M. O. (1999). *Nonparametric Inference for Zenga's Measure of Income Inequality*, Metron LVII, 69–84.
- Dagum, C. (1977). A new model of personal income distribution. Specification and estimation. *Economie Appliquee*, XXX(3), 413–436.
- Dagum, C. (1997). A new approach to the decomposition of the Gini income inequality ratio. *Empirical Economics*, 22, 515–531.
- Davidson, R. (2009). Reliable inference for the Gini index. *Journal of Econometrics*, 150(1), 30–40.
- Kordos, J., & Zięba, A. (2010). Development of standard error estimation methods in complex household sample surveys in Poland. *Statistics in Transition*, 11(2), 231–252.
- Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. Hoboken New Jersey: Wiley.
- Verma, V., & Betti, G. (2005). *Sampling Errors for Measures of Inequality and Poverty*, Invited Paper in “Classification and Data Analysis Book of Short Papers”, CLADAG, Italian Statistical Society.
- Wolter, K. (2003). *Introduction to variance estimation*. New York: Springer.
- Zenga, M. (1990). *Concentration curves and concentration indices derived from them, w: income and wealth distribution, inequality and poverty* (pp. 94–110). Berlin: Springer.