ESTIMATION OF INCOME INEQUALITY MEASURES BY REGIONSON THE BASIS OF POLISH HBS

ALINA JĘDRZEJCZAK
University of Lodz
jedrzej@uni.lodz.pl

ABSTRACT

In the analysis of income and wage size distributions statistical measures of concentration are often used. They usually become basic tools in the investigations concerning poverty and social welfare issues. They can also be helpful to analyze the efficiency of a tax policy or to measure the level of social stratification and polarization. Among many income inequality measures the Gini and Zenga coefficients are of greatest importance. Unfortunately the standard errors of these measures, being actually sample statistics, are rarely reported in practice.

Estimators of many concentration coefficients are nonlinear functions of sample observations thus their standard errors cannot be obtained easily. The methods of variance estimation that can solve this problem include: various replication techniques as jackknife, bootstrap and BRR methods, Taylor expansion technique and some parametric procedures based on income distribution models.

In the paper some estimation methods for Gini and Zenga concentration measures are presented together with their application to the analysis of income distributions in Poland. This effort was made to compare the NUTS1 regions in Poland from the point of view of income inequality. The basis for the calculations was individual data coming from the Household Budget Survey conducted by Polish Central Statistical Office in the years 2006-2008. The variance estimates were obtained by means of the bootstrap and the parametric approach based on the Dagum type-I model.

Key words: income distribution, income equality, standard error stimation

ESTIMATORS OF INCOME INEQUALITY MEASURES

Among many income inequality (concentration) measures the ***Gini index*** is the most popular. It is mainly due to its good statistical properties and straightforward economic interpretation. Gini index of inequality can be defined as double the area between the Lorenz curve and „the line of equal shares", that is the line describing perfect equality. The Gini index can be expressed as follows:

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

where: $p = F(y)$ is a cumulative distribution function of income, $L(p)$- the Lorenz function given by the following formula:

$$L(p) = \mu^{-1}\int_0^p F^{-1}(t)dt\,, \qquad (2)$$

where μ denotes the expected value of a random variable Y and $F^{-1}(p)$ is the p^{th} quantile.

One can estimate the value of the Gini index from the survey data using the following formula:

$$\hat{G} = \frac{2\sum_{i=1}^{n}(w_i y_{(i)} \sum_{j=1}^{i} w_j) - \sum_{i=1}^{n} w_i y_{(i)}}{(\sum_{i=1}^{n} w_i)\sum_{i=1}^{n} w_i y_{(i)}} - 1 \qquad (3)$$

where: $y_{(i)}$- household incomes in a non-descending order,

w_i- survey weight for i-th economic units,

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

Another interesting measure of income inequality based on a concentration curve was proposed by Zenga (1984). It is called "***point concentration measure***", being sensitive to changes of inequality in each part (point) of a population. ***Zenga synthetic inequality index*** can be expressed as the area below the Zenga curve Z_p which is based on the relation between income and population quantiles:

$$Z_p = [y_p^* - y_p]/y_p^*, \qquad (4)$$

where y_p denotes the population p^{th} quantile and y_p^* is the corresponding income quantile defined as follows;

$$y_p^* = Q^{-1}(p) \qquad (5)$$

The function $Q(p)$ is usually called *first moment distribution function* and can be interpreted as cumulative income share related to the mean income.

Zenga synthetic inequality measures takes the form;

$$Z = \int_0^1 Z_p dp \qquad (6)$$

The commonly used nonparametric estimator of the Zenga index (6) was introduced by Aly and Hervas (1999) and can expressed by the following equation:

$$\hat{Z} = 1 - \frac{1}{n\bar{y}} \left\{ y_{1:n} + \sum_{j=1}^{n-1} y \left\langle \frac{\sum_{i=1}^{j} y_{i:n}}{\bar{y}} \right\rangle_{:n} \right\} \qquad (7)$$

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

INEQUALITY DECOMPOSITION BY SUBPOPULATIONS

The Gini index, being the function of both: individual income and the ranking of economic units, cannot be decomposed easily by population subgroups as well as by factor components. Regardless of these difficulties, a great effort has been made to specify the conditions under which the decomposition of the Gini coefficient by subgroups and by income components is feasible. Lerman and Yitzhaki (1984) introduced a clear decomposition of the Gini index by income components based on their covariance formula, providing a useful tool for income inequality analysis. The decomposition by subpopulations, however, proved to be far more complicated. Since Shorrocks (1984) characterized the class of inequality indices that are decomposable by population subgroups, the Gini index has been considered to be decomposable only when the subpopulations do not overlap. In fact, when the distributions overlap the third component called "overlapping" or "interaction term", rather difficult to interpret, has to be taken into consideration. That "third component" was discussed by Pyatt (1976), Silber (1989), Yitzhaki (1994), Deutsch and Silber (1999), to name only a few, what resulted in some interesting decomposition formulas. Unfortunately, they are computationally cumbersome and it is not always clear what meaningful interpretation each of the components has.

An interesting approach to the decomposition of the Gini index was proposed by Dagum (1997). It introduces the concept of economic distance between subpopulations as an important element in the Gini index decomposition by subpopulation groups. The interaction term is based on the concept of transvariation and can be viewed as a measure of distribution overlapping or the degree to which the incomes of different social groups cluster.

The inequality decomposition proposed by Dagum (1997) is based on the mean difference formula, expressing the Gini index as a relative dispersion measure. The mean difference Δ, being the absolute measure of dispersion, can be defined as the average absolute difference between all possible pairs of observations in a population of income receivers. Gini (1912) showed that the geometric approach, given by the formula (1), is related to the statistical approach via the concept of mean difference.

$$G = \frac{\Delta}{2\bar{Y}} = \sum_{r=1}^{n}\sum_{i=1}^{n} |Y_i - Y_r| \Big/ 2n^2\bar{Y} = \\ = \sum_{j=1}^{k}\sum_{h=1}^{k}\sum_{i=1}^{n_j}\sum_{r=1}^{n_h} |y_{ji} - y_{hr}| \Big/ 2n^2\bar{y} \qquad (8)$$

The last term of the formula given above represents the Gini index for a population divided into *k* subgroups (subpopulations). The Gini index for the sub-population *j* takes the form:

$$G_j = \frac{\Delta}{2\bar{Y}_j} = \frac{1}{2\bar{Y}_j}\sum_{r=1}^{n_h}\sum_{i=1}^{n_j} |y_{ji} - y_{jr}| \Big/ n_j^2 , \quad (9)$$

where: $\bar{Y}_j$ - the mean income in group *j*, n_j - frequency.

The Gini index expressed in terms of the Gini mean difference Δ can be also generalized for two-populations case, measuring the between-populations (or intra-groups) inequality. Thus the ***extended Gini index*** between groups *j* and *h* can be written as follows:

$$G_{jh} = \frac{\Delta_{jh}}{\bar{Y}_j + \bar{Y}_h} = \\ = \frac{1}{\bar{Y}_j + \bar{Y}_h}\sum_{i=1}^{n_j}\sum_{r=1}^{n_h} |y_{ji} - y_{hr}| \Big/ n_j n_h \qquad (10)$$

where Δ_{jh} denotes Gini mean difference modified for two income distributions.

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

$$G = G_w + G_b + G_t \qquad (11)$$

where:

$$G_w = \sum_{j=1}^{k} G_j p_j s_j$$

$$G_b = \sum_{j=1}^{k}\sum_{h=1}^{j-1} G_{jh}(p_j s_h + p_h s_j) D_{jh}$$

$$G_t = \sum_{j=2}^{k}\sum_{h=1}^{j-1} G_{jh}(p_j s_h + p_h s_j)(1 - D_{jh})$$

G_w – the contribution of ***within- groups inequality***,

G_b – the contribution of ***net between- groups inequality***,

G_t – the contribution of "***transvariation***" to the Gini index.

The term D_{jh}, called economic distance ratio or relative economic affluence (REA), is related to the normalized intensity of transvariation (which is simply 1- D_{jh}) and can be regarded as the measure of ***relative economic affluence*** of the *j*-th subpopulation with respect to the *h*-th subpopulation. It can be defined as the weighted average of income differences $y_{ji} - y_{hr}$, calculated for all the members belonging to the population *j*-th with incomes grater than those of the members belonging to the population *h*, given that $\bar{Y}_j > \bar{Y}_h$ (for details see: Dagum, 1980).

As it can be easily noticed the Gini index provides an unusual "between-group" component. It measures the income inequality between each and every pair of subpopulations, whereas entropy and most of between-groups inequality measures yield only the income inequalities between subpopulation means.

VARIANCE ESTIMATION METHODS

The precision of an estimator is usually discussed in terms of its variance. In many cases the exact value of this variance is unknown, because it depends on unknown population quantities. After survey data have been obtained, however, an appropriate estimate of estimator variance can be calculated.

Explicit variance estimators are often complicated and it is hard to derive their general mathematical formulas, especially for nonlinear estimators and complex sampling designs.

Most of income concentration measures are nonlinear functions of sampling observations so their standard errors are difficult to obtain and have rarely been reported in practice.

To solve this problem, some special approximate techniques for variance estimation can be used. They include:

- **Taylor linearization technique**, Wolter (1985);
- **Random groups method**, Mahalanobis (1944); Hansen, Hurwitz, Madow (1953);
- **Balanced Half Samples** (BHS), also called Balanced Repeated Replication (BRR), Mc Carthy (1969),
- **Jacknife** , Quenouille (1949), Durbin (1969)
- **Bootstrap** , Efron (1979)
- **Parametric approach** based on maximum likelihood theory.
- **Generalized Variance Function** (GVF)- first applied in Current Polpulation Survey CPS in 1947.

In the context of inequality measures Taylor linearization, jackknife, bootstrap and parametric approach are the methods of variance estimation most often used.

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

$$g(Y) \approx g(\theta) + \sum_{i=1}^{k} g_i^{'}(\theta)\,(Y_i - \theta_i) \quad (12)$$

The bootstrap method similarly to the jacknife was introduced outside the field of survey sampling as a means of obtaining approximate variance estimates and confidence intervals (see.: Efron, 1979). 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 R_k^* (k=1…N). The bootstrap variance estimator of a statistic R takes the form:

$$\hat{D}_B^2(R) = \frac{1}{N-1}\sum_{k=1}^{N}(R_k^* - \bar{R}^*)^2 \quad (13)$$

Provided that an empirical income distribution can be approximated by a theoretical model described by a probability density function f(y,**θ**), the method of variance estimation based on maximum likelihood theory can be used. ML estimators are asymptotically unbiased and normally distributed with variances given by the Cramer-Rao bound. Let us assume that an inequality measure of interest can be expressed as a function g(**θ**) of the model parameters **θ.**

The variance of the ML estimator of an inequality measure g(**θ**) takes the form:

$$\mathrm{D}^2[\mathrm{g}(\hat{\boldsymbol{\theta}})] = \left[\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^T \mathbf{I}_\theta^{-1}\left[\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \quad (14)$$

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

APPLICATION

The methods given above were applied to the analysis of income inequality in Poland. The basis for the calculations was the data coming from the Polish Household Budget Survey (HBS) conducted in the years 2006 and 2008. In 2006 the randomly selected sample covered 37508 households, i.e. approximately 0,3% of the total number of households, while in 2006 the total sample size was 37584. 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, the data obtained from the HBS were weighted with the structure of households by number of persons and class of locality coming from Population and Housing Census 2002. To obtain inequality coefficients we used the formulas (3) and (7). Next, the decomposition of income inequality in Poland by regions were done using equation (11). Finally, the estimates of standard errors were obtained using two variance estimation methods:

-bootstrapping,

- parametric approach.

The analysis was conducted after dividing the overall sample by region NUTS1 constructed according to the EUROSTAT classification. The estimates for the entire population of household were also calculated.

Table 1 depicts maximum likelihood estimates of the parameters of the Dagum type-I income distribution model. The values of overlap measure, reflecting the goodness-of-fit, are also reported in the table. Relatively large values of this measure (95% and more) confirm high consistency of the theoretical distributions that have been estimated to the corresponding empirical ones.

The results of parametric estimation for Gini and Zenga coefficients are presented in table 2. Estimated relative standard errors of the Gini index take value between 2% and 3%, while for the Zega index they vary from 3% to 5% with the most frequent value about 4%.

Table 1. Maximum likelihood estimates of the Dagum model parameters

Region	Year	Dagum model parameters			Overlap measure
		λ	β	δ	
1. Central	2006 2008	8,8473 23,0297	1,029? 0,8723	2,7084 2,7897	0,9618 0,9646
2. Southern	2006 2008	20,2299 42,6820	0,7837 0,7344	3,3689 3,3323	0,9622 0,9557
3. Eastern	2006 2008	12,0879 17,1291	0,8334 0,8219	3,1113 2,0331	0,9676 0,9582
4. North-western	2006 2008	14,2599 24,6655	0,8798 0,8873	3,0846 2,9967	0,9615 0,9581
5. South-western	2006 2008	15,0147 28,8773	0,8499 0,7551	3,2141 2,9976	0,9712 0,9587
6. Northern	2006 2008	6,8068 28,1621	1,0676 0,7984	2,7335 3,0851	0,9675 0,9543

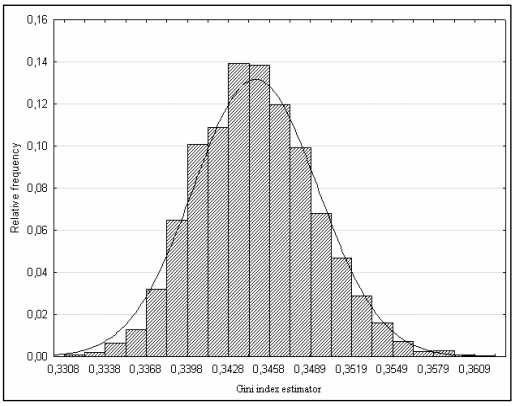
Source: author's calculations

Table 2. Parametric estimates of the Gini and Zenga inequality measures and their standard errors

Region	Year	Sample size	Gini index $\hat{G}$	$D(\hat{G})$ (CV w %)	Zenga index $\hat{Z}_1$	$D(\hat{Z})$ (CV w %)
1. Central	2006 2008	1972 2003	0,3671 0,3695	0,0081 (2,2) 0,0079 (1,9)	0,3859 0,3895	0,0149 (3,9) 0,0143 (3,7)
2. Southern	2006 2008	1916 1883	0,3157 0,3249	0,0066 (2,1) 0,0068 (2,1)	0,2961 0,3118	0,0109 (3,7) 0,0114 (3,6)
3. Eastern	2006 2008	1679 1651	0,3357 0,3568	0,0076 (2,3) 0,0083 (2,3)	0,3299 0,3668	0,0132 (4,0) 0,0146 (3,9)
4. North-western	2006 2008	1441 1438	0,3340 0,3429	0,0082 (2,5) 0,0086 (2,4)	0,3270 0,3425	0,0143 (4,4) 0,0150 (4,4)
5. South-western	2006 2008	999 1000	0,3236 0,3569	0,0095 (2,9) 0,0105 (2,9)	0,3092 0,3676	0,0161 (5,2) 0,0184 (5,0)
6. Northern	2006 2008	1339 1330	0,3610 0,3422	0,0098 (2,7) 0,0087 (2,5)	0,3751 0,3412	0,0178 (4,7) 0,0151 (4,4)

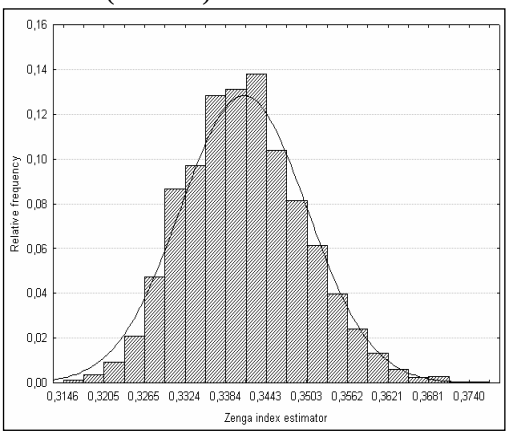
Source:author's calculations

Fig. 1. Bootstrap distribution of Gini index estimator (N=5000)



Source: author's calculations

Fig. 2. Bootstrap distribution of Zenga index estimator (N=5000)



Source: author's calculations

Table 3. Gini index decomposition by regions in 2006

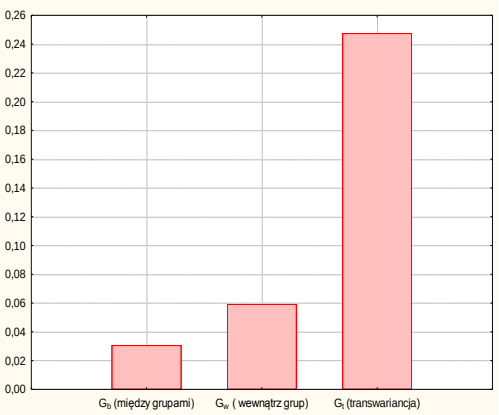
Between group inequality G_b		**0,0306** (9%)
Within-group inequality G_w		**0,0592** (18%)
contribution of- *central*	0,0182	
– *southern*	0,0125	
– *eastern*	0,0102	
– *north-eastern*	0,0079	
– *north-western*	0,0034	
– *northern*	0,0071	
Contribution of transvariation G_t		**0,2475** (73%)
Total income inequality G		**0,3373** (100%)

Table 4. Gini index decomposition by regions in 2008

Between group inequality G_b		**0,0424** (12%)
Within-group inequality G_w		**0,0600** (18%)
contribution of- *central*	0,0181	
– *southern*	0,0129	
– *eastern*	0,0097	
– *north-eastern*	0,0085	
– *north-western*	0,0039	
– *northern*	0,0069	
Contribution of transvariation G_t		**0,2416** (70%)
Total income inequality G		**0,3440** (100%)

Source: author's calculations

Fig.3. Gini index decomposition by region



The bootstrap and parametric variance estimators for the Gini index seem to be more stable than those of the Zenga measure. It can be concluded that in general the estimator of the Gini index is more efficient than the corresponding estimator of synthetic Zenga inequality measure.

Despite relatively small number of repetitions, the bootstrap distributions of both inequality statistics can be approximated by the normal density curves (see: fig. 1 and 2).

Analyzing the results of calculations presented in tables 3 and 4 one can easily notice that the regional income disparities in Poland are rather small- the between regions inequality is only 9% of the total Gini. The substantial contribution of transvariation G_t , equal to 74%, is an evidence of notable overlapping of income distributions for NUTS1. To analyze the problem more thoroughly one can observe the economic distance ratios D given by eq. (11), measuring the relative economic affluence of one region with respect to another. These values for all pairs of regions in Poland are rather small being approximately 0,05-0,1 and only *central* region is significantly more affluent than the others. As a result, the transvariation component is dominated mainly by the overlapping between the distributions of *central* region and the other regions. The highest value of D was observed for the regions: *central* and *eastern*. It is equal to 0,22 what means that the economic situation of *central* voivodeships in Poland is by 22% better than the situation of the eastern ones, taking into consideration the differences in mean incomes as well as in the shapes of the compared distributions.

The Gini ratios and means within regions do not differ significantly (table 2) so the contributions of particular subpopulations to the overall inequality are determined mainly by their sizes (table 2a). On the whole, the inequality within groups is responsible for only 17% of the total inequality.

CONCLUSIONS

The interesting results of the decomposition of income inequality in Poland, obtained on the basis of household budgets' data, suggest that this approach can be helpful for better understanding of the problem and can be used in many further economic analysis, including poverty and social welfare investigations. Decomposition of income inequality measures by sub-populations can be useful in comparing income distributions by assessing the contributions of between-group and within-group inequalities to the overall inequality of a population. It can also be useful in stratification and market segmentation by including the concept of overlapping.

REFERENCES

1. Aly E.A., M.O. Hervas (1999), Nonparametric Inference for Zenga's Measure of Income Inequality, *Metron* LVII: 69–84.
2. Gini C. (1912), Variabilita'e Mutabilita' *Studi Economicogiuridici Universita' di' Cagliari*, III, 2a, Bologna: 1-156.
3. Dagum C. (1977), A New Model of Personal Income Distribution. Specification and Estimation, *Economie Apliquee'*, XXX(3): 413-436.
4. Dagum C. (1980), Inequality Measures Between Income Distributions with Application. *Econometrica* 48: 1791–1803.

5. Dagum C. (1997), A New Approach to the Decomposition of the Gini Income Inequality Ratio, *Empirical Economics* 22: 515–531.
6. Deutsch I., J. Silber (1999), Inequality Decomposition by Population Subgroups and the Analysis of Interdistributional Inequality, in: J. Silber, *Handbook of Income Inequality Meosurement*,: 363–397.
7. Durbin J. (1959), A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika* 46: 477–480.
8. Efron B. (1979), Bootstrap Methods: Another Look at the Jacknife, *Annals of Statistics* 7: 1–26.
9. Lerman R., S. Yitzhaki (1984), A Note on the Calculation and Interpretation of the Gini Index, *Economic Letters* 15: 363–369.
10. McCarthy P. J. (1969), Pseudo-Replication Half-Samples, *Review of the International Statistical Institute* 37: 239–264.
11. Mahalanobis P. C. (1944), On Lagre-Scale Sample Surveys. *Philosophical Transactions of the Royal Society of London*, B, 231: 329–451.
12. Hansen M., W. Hurwitz, W. Madow (*1953*), Sample Survey Methods and Theory, Wiley.
13. Pyatt G. (1976), On the Interpretation and Disaggregation of Gini Coefficient, *The Economic Journal* 86: 243–255.
14. Quenouille M. H. (1949), Problems in Plane Sampling, *Annals of mathematical Statistics* 20: 355–375.
15. Shorrocks A. (1984), Inequality Decomposition by Population Subgroups, *Econometrica* 52, 1337–1339.
16. Silber J. (1989), Factor Components, Population Subgroups and the Computation of the Gini index of Inequality, *The Review of Economics and Statistics* 71: 107–115.
17. Wolter K.. (1985), Variance Estimation, Springer-Verlag, New York.
18. Yitzhaki S. (1994), Economic Distance and Overlapping of Distributions, *Journal of Econometrics* 61: 147–159.
19. Zenga M. (1984), Proposta per un indice di concentrazione basato sui rapporti fra quantili di popolazione e quantili reddito, *Giornale Degli Economisti ed Annali di Economia* 48: 301–326.