Paper Code : ES1 – 44

REGIONAL DIFFERENCES IN INCOME DISTRIBUTION IN THE SLOVAK REPUBLIC

Terek Milan
Muchova Eva
University of Economics in Bratislava, Slovak republic

Abstract

The paper deals with the regional incomes structure analysis in Slovak republic by median, medial and a measure of concentration based on the European Union statistics on income and living conditions in Slovak republic data (EU-SILC 2014). The survey containing more components such as random sampling, stratification, clustering and so on is obviously called complex survey. EU-SILC data are the data from complex survey. In such cases the sampling weights are commonly used to provide the correct results in statistical inference. The using sampling weights in construction of empirical probability mass function and empirical cumulative distribution function will be described. The estimation of the population median with aid of empirical cumulative distribution function will be described and the relations enabling to estimate population medial with aid of empirical cumulative distribution function will be formulate. The population median, medial and a measure of concentration estimation of the whole gross household incomes for the whole Slovak republic and separately for eight Slovak regions will be realized. The regional results will be compared mutually as well as with the results for the whole Slovak republic.

Keywords: regional incomes structure, sampling weights, empirical probability mass function, empirical cumulative distribution function, measures of concentration

Introduction

The level and structure of household incomes significantly affect the behavior of microeconomic entities as consumers, savers, owners of production factors, and consequently investors. The income structures also determine macroeconomic indicators such as consumption, savings and investment of household into human and physical capital. The regional structure of incomes in Slovak republic by median, medial and the measure of concentration based on medial will be analyzed on the basis of data from the European Union Statistics on Income and Living Conditions (EU-SILC) realized in Slovak republic in the year 2014. EU-SILC is an instrument aiming at collecting timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions. This instrument is anchored in the

European Statistical System[1]. In general EU-SILC data are the data from complex survey.

The survey containing more components such as random sampling, stratification, clustering and so on is obviously called complex survey. A survey may be stratified with several stages of clustering and rely on ratio and regression estimation. In these cases sampling weights are commonly used to provide the correct results.

The analysis of the regional structure of incomes in Slovak republic by median and medial based on using sampling weights will be studied in the paper. The using sampling weights in construction of empirical probability mass function and empirical cumulative distribution function will be described. On the basis of these functions the estimation of median and medial of the whole gross household incomes for the whole Slovak republic and separately for eight domains (subpopulation) – Slovak regions is realized. The concentration of the whole gross household incomes will be also measured.

Methods

Distributions of incomes or wages are obviously skewed and outliers[2] are present. Then, the interpretation power of the mean is very small[3]. In general in such distributions the mean is not considered as appropriate measure of central tendency. The mean income is not convenient measure of "typical" income. The median is generally considered as good measure of central tendency in such distributions because of its stability and robustness toward outliers. Alternatively some non-traditional measures of location can be also interesting as good measures of central tendency for such distributions. The using of the trimmed mean (Piegorsch, 2015, p. 55), Winsorized mean or M-estimators is recommended[4]. Interesting results provides also traditional measures of central tendency applied on the data set from which the outliers were removed[5].

Sometimes the standard statistical methods supposing the independence and identical distribution of observations are applied to the data from complex surveys. In Lohr (2010, pp. 287 – 288) is stated: "When you read the paper or book in which the

[1] For more details, see: European Union Statistics on Income and Living Conditions (EU-SILC). Retrieved from http://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions.

[2] We will define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data (Barnett & Lewis, 1994, p. 7).

[3] For more details see: Halley (2004, pp. 39-52).

authors analyze data from the complex survey, see whether they accounted for the data structure in the analysis, or whether they simply ran the raw data through non-survey statistical package procedure and reported the results. If the latter, their inferential results must be viewed with suspicion".

Sampling Weights

The sampling (design) weights allow to construct an empirical distribution for the population (in fact it is an empirical distribution of the observation from the population). The estimation of some population quantities based on this distribution is possible.

Suppose the size *N* of finite population *U* is known. Symbol *x* denotes variable under study and also its values, $U = \{1,2,...N\}$ is the set of unit indexes in the population.

Symbol *S* denotes sample from the population – subset containing *n* units from *U*. Let's π_i be the probability that unit $i \in U$ will be in random sample. Sampling weights for any sampling design are defined as follows

$$w_i = \frac{1}{\pi_i} \qquad (1)$$

Sampling weight of unit *i* can be interpreted as number of units in the population represented by unit *i*. These weights can be modified with regard to nonresponse and coverage error (for more details see, for example, Levy & Lemeshow (2008)). When sampling weights for all observations units in the sample are equal the sample is called self-weighting. Each observed unit represents the same number of unobserved units in the population. When sampling weights are not equal for all observations units in the sample, the sample is called non-self-weighting. If the sample is non-self-weighting, point estimates of population quantities produced by standard statistical software supposing statistical independence and the same distribution of observations, will be biased. It is also the case in the above mentioned application. The EU-SILC sample is non-self-weighting. The capturing the structure of data is necessary in point estimation of population quantities. The use of sampling weights is needed.

Estimating an Empirical Probability Mass Function and Empirical Cumulative Distribution Function

Suppose the values for the entire population of *N* units are known. A value of probability mass function in *x* is

$$p(x) = \frac{N_{(x)}}{N} \qquad (2)$$

where $N_{(x)}$ is number of units whose value is *x*. A value of cumulative distribution

function in *x* is

$$F(x) = \sum_{y \le x} p(y) \qquad (3)$$

Note that it is probability mass function and cumulative distribution function of observation from the population because the „model-free" or „distribution-free" approach to sample survey is under consideration[6].

Sampling weights allow to construct empirical probability mass function and empirical cumulative distribution function. Empirical probability mass function $\hat{p}(x)$ is defined by the sum of weights for all observations taking on the value *x* divided by the sum of all the weights:

$$\hat{p}(x) = \frac{\sum_{i \in S; x_i = x} w_i}{\sum_{i \in S} w_i} \qquad (4)$$

Empirical cumulative distribution function $\hat{F}(x)$ is

$$\hat{F}(x) = \sum_{y \le x} \hat{p}(y) \qquad (5)$$

Estimating of Median and Medial

Any population quantity can be estimated from the empirical probability mass function $\hat{p}(x)$ or from empirical cumulative distribution function $\hat{F}(x)$. Population quantiles will be estimated as follows. Since the empirical cumulative distribution function $\hat{F}$ is a step function, the interpolation is usually needed to find a unique value for the quantile. Let y_1 be the largest value in the sample for which $\hat{F}(y_1) \le p$ and let y_2 is smallest value in the sample for which $\hat{F}(y_2) \ge p$. Then population *p* quantile estimate is

$$\hat{Q}_p = y_1 + \frac{p - \hat{F}(y_1)}{\hat{F}(y_2) - \hat{F}(y_1)}(y_2 - y_1) \qquad (6)$$

[6] For more details see: Cochran, 1977, pp. 8-9.

Population median will be estimated according to (6), for p = 0.5.

Medial (*Ml*) is such value for which the sum of variable values less or equal to *Ml* is equal to the half of variable total. We will formulate the relations enabling to estimate medial with aid of sampling weights. It can be proven that if all values of variable are nonnegative then: *Ml* ≥ *Me* (Dagnelie, 1998, p. 81). The sum of variable values for all observations taking on the value *x* we will call the class total. The medial is calculated as median but on the basis of class totals instead of frequencies. Empirical probability mass function $\hat{p}_{Ml}(x)$ in this case can be defined as:

$$\hat{p}_{Ml}(x) = \frac{\sum_{i \in S; x_i = x} w_i x_i}{\sum_{i \in S} w_i x_i} \tag{7}$$

Empirical cumulative distribution function $\hat{F}_{Ml}(x)$ is then

$$\hat{F}_{Ml}(x) = \sum_{y \le x} \hat{p}_{Ml}(y) \tag{8}$$

Let y_1 be the largest value in the sample for which $\hat{F}_{Ml}(y_1) \le 0{,}5$ and let y_2 is smallest value in the sample for which $\hat{F}_{Ml}(y_2) \ge 0{,}5$. Then the medial can be estimated by

$$\hat{Ml} = y_1 + \frac{0{,}5 - \hat{F}_{Ml}(y_1)}{\hat{F}_{Ml}(y_2) - \hat{F}_{Ml}(y_1)}(y_2 - y_1) \tag{9}$$

The medial provides in some application areas very interesting interpretation possibilities.

Note that estimators constructed using this method are not necessarily unbiased or numerically stable. Despite of it, the statistics calculated using weights are much closer to the population quantities as in not weighting case (Lohr, 2010, p. 293).

Measures of concentration

Concentration on the broadest sense means the accumulation of the objects, for example income units, to subjects, for example households (Marfels, 1971, p. 753). The measures of concentration are used mainly for measurement of distribution of the wages or incomes total. If for example some percentage of households obtains the same total of incomes, the distribution of incomes is perfectly uniform–the concentration is

null. If one household obtains the whole income total, the concentration is maximal. In general, three measures of concentration are known– Lorentz curve, Gini index and the measure of concentration based on medial (Coeurjolly, 2015). The last one will be used in the application under consideration:

$$\Delta = \frac{Ml - Me}{R} \qquad (10)$$

where *R* is the range.
It will be estimated by

$$\hat{\Delta} = \frac{\hat{M}l - \hat{M}e}{R_S} \qquad (11)$$

where R_S is the sample range.

Because of skewing of the distribution of incomes, the outliers will be determined and removed from the data set in R_S calculation process. The method of outliers detection of Tukey, applied also in box plots will be used. That is the robust method of outliers detection – not itself influenced by outliers. Values more than 1.5 of inter-quartile range[7] below the first quartile and 1.5 of inter-quartile range above third quartile are considered to be outliers. The quartiles needed for inter-quartile range calculation will be estimated according to (6), for $p = 0.25$ and $p = 0.75$.

Analysis of Regional Structure of Incomes by Median, Medial and Measure of Concentration Based on Medial

The structure of incomes analysis in the Slovak regions was effectuated using the data from the survey EU-SILC executed in the Slovak Republic in 2014. In the Slovak Republic the stratified two-stage survey design is regularly used. The stratification is effected with two stratification variables – region and settlement size. There are eight regions in the Slovak Republic: Bratislava, Trnava, Trencin and Nitra in western Slovakia, Zilina and Banska Bystrica in central Slovakia, Kosice and Presov in eastern Slovakia. The EU-SILC 2014 survey was executed on the sample of 6,010 households, 5,490 households and 13,433 individuals 16+ old were included in the database. Sampling weights were calculated and modified with respect to nonresponse. These weights can be used to inference about the population of Slovak households. EU-SILC sample is non-

[7] The inter-quartile range is the difference between the third and first quartile.

self-weighting.

The data from EU-SILC 2014 survey are concentrated in many sets. Each household is identified by one identification number. The structure of the whole gross household incomes in eight domains – the Slovak regions – was analyzed. The above mentioned regions correspond with the categories of one from stratification variables. Firstly, the matching of needed data – sampling weights and whole gross household incomes was effected according to household numbers. Then, the matched data were distributed according to regions. Eight sets of data were obtained, one for each region. Each region was analyzed separately.

The values of the empirical probability mass function were calculated according to (4) and on the basis of that the values of the empirical cumulative distribution function were calculated by relation (5) for the whole Slovak republic and separately for each region.

The estimate of median whole gross household income was calculated according to relation (6) for the whole Slovak republic and separately for each region[8]. The estimate of population median whole gross household income for the whole Slovak republic in the year 2014 equals 13,305.83 euros. The obtained results for regions shows table no. 1. The table no. 1 presents also the ordering of regions according to the median whole gross household income.

Table no. 1 – Regional structure of median whole gross household income in 2014

Region number	Region name	Estimate of median whole gross household income in 2014 (Euros)	Order of region according to median whole gross household income
1	Bratislava	14,491.37	1.
2	Trnava	13,969.12	4.
3	Trencin	14,368.47	2.
4	Nitra	12,379.67	7.
5	Zilina	14,054.85	3.
6	Banska Bystrica	11,746.41	8.
7	Presov	13,595.22	5.
8	Kosice	13,118.16	6.

Source: own

Then the values of the empirical probability mass function $\hat{p}_{MI}(x)$ were calculated according to (7) and on the basis of that the values of the empirical cumulative

[8] See also in Terek, 2017.

distribution function $\hat{F}_{MI}(x)$ were calculated by relation (8) for the whole Slovak republic and separately for each region. The estimate of medial whole gross household income was calculated according to relation (9) for the whole Slovak republic and separately for each region[9]. The estimate of medial whole gross household income in 2014 for the whole Slovak republic was 20,355.80 euros and percentage of households having incomes less or equal to medial in Slovak republic was 74.15 %. The obtained results for regions shows the table no. 2. For example in the year 2014, in region Bratislava, the half of the incomes total was distributed among 76.71 % of "poorer" households (having incomes less or equal to 24,874.65 euros), the second half of incomes total was distributed among 23.29 % of "richer" households (having incomes greater or equal to 24,874.65 euros). For example in region Banska Bystrica, the half of the incomes total was distributed among 74.88 % of "poorer" households (having incomes less or equal to 18,301.86 euros), the second half of incomes total was distributed among 25.12 % of "richer" households (having incomes greater or equal to 18,301.86 euros).

Table no. 2 – Regional structure of medial whole gross household income in 2014

Region number	Region name	Estimate of medial whole gross household income in 2014 (Euros)	Order of region according to medial whole gross household income	Percentage of households having incomes less or equal to medial	Order of region according to percentage of households having incomes less or equal to medial
1	Bratislava	24,874.65	1.	76.71	1.
2	Trnava	20,331.04	5.	72.44	7.
3	Trencin	20,555.89	2.	73.40	5.
4	Nitra	19,364.40	6.	74.61	3.
5	Zilina	20,492.41	3.	73.88	4.
6	Banska Bystrica	18,301.86	7.	74.88	2.
7	Presov	20,489.51	4.	73.34	6.
8	Kosice	18,193.63	8.	71.90	8.

Source: own

The values of measure of concentration Δ were estimated by $\hat{\Delta}$ according to (11). The

[9] See also in Terek, 2017.

results are shown in table no. 3. The numbers of outliers removed from data sets in R_S calculation process are presented in the last column of table 3. The value of $\hat{\Delta}$ in the whole Slovak Republic was 17.46 %. In the calculation process of the sample range R_S for the whole Slovak republic, 157 outliers were removed.

The table no. 4 presents the ordering of regions according to all calculated indicators.

Table no. 3 – Regional structure of concentration of the gross household income in 2014

Region number	Region name	$\hat{\Delta}$ (%)	Order of region according to the value of $\hat{\Delta}$	The number of outliers removed from data set in R_S calculation process
1	Bratislava	21.90	1.	24
2	Trnava	15.86	6.	8
3	Trencin	15.51	7.	22
4	Nitra	17.98	3.	18
5	Zilina	15.97	5.	20
6	Banska Bystrica	18.45	2.	30
7	Presov	16.98	4.	15
8	Kosice	14.82	8.	18

Source: own

Table no. 4 – Ordering of regions according to all calculated indicators

Region number	**Region name**	**Order of region according to median whole gross household income**	**Order of region according to medial whole gross household income**	**Order of region according to percentage of households having incomes less or equal to medial**	**Order of region according to the value of $\hat{\Delta}$**
1	Bratislava	1.	1.	1.	1.
2	Trnava	4.	5.	7.	6.
3	Trencin	2.	2.	5.	7.
4	Nitra	7.	6.	3.	3.

5	Zilina	3.	3.	4.	5.
6	Banska Bystrica	8.	7.	2.	2.
7	Presov	5.	4.	6.	4.
8	Kosice	6.	8.	8.	8.

CONCLUSIONS

All calculations were realized in Excel 2013. The obtained ordering of regions according to median whole gross household income is very interesting. Obviously the great difference among Bratislava region with the capital of Slovakia Bratislava and the rest of Slovakia is expected. The analysis results show that the difference between first Bratislava and second Trencin regions is not very large. The median household incomes of the third Zilina, fourth Trnava and fifth Presov are also very close. There are bigger differences among last three regions. The median whole gross household income of Banska Bystrica is surprisingly low. The regions Bratislava, Trencin, Zilina, Trnava, Presov have the median whole gross household income greater and the regions Kosice, Nitra and Banska Bystrica less than in the whole Slovak republic.

In the analysis based on medial whole gross household income, the results of regional ordering are not very different. The first three places are occupied by the same regions as according to median, the changes in the other places of ordering are only moderate. The regions Bratislava, Trencin, Zilina, Presov have the medial whole gross household income greater and the regions Trnava, Kosice, Nitra and Banska Bystrica less than in the whole Slovak republic. The finding that in all Slovak regions the half of the incomes total is distributed among 71.90 – 76.71 % of "poorer" households and the second half of incomes total is distributed among the rest of "richer" households is very interesting. The differences among Slovak regions in this indicator are only moderate. The regions Bratislava, Nitra and Banska Bystrica have percentage of households having incomes less or equal to medial greater and the regions Trnava, Trencin, Zilina, Presov and Kosice less than Slovak republic percentage.

The estimation of the measure of concentration Δ by $\hat{\Delta}$ is a few problematic because of estimation of the range *R* by the sample range R_S. On the other hand the outliers were removed from data sets before calculation of range, what could increase the numeric stability of this indicator. If we have data for more years in the past, the average sample range $\overline{R}$ could be better estimator[10]. Despite of it, some results are very interesting. The region Bratislava, first in median and medial whole gross household income has also the highest concentration of the whole gross household income, second place is occupied by region Banska Bystrica, taking the last places in median and

[10] The outliers would be removed in the calculation of each range entering to average calculation.

medial whole gross household incomes. Region Trencin occupying the second place in median and medial whole gross household income takes the seventh place in concentration of the whole gross household income.

The table no. 4 shows the ordering the regions according to all presented indicators. The differences between the region ordering according to median and medial whole gross household incomes are only moderate, the same only moderate difference is between the ordering according to percentage of households having incomes less or equal to medial and according to value of concentration measure $\hat{\Delta}$.

The application of correct methodology of estimation is very important in the context of the data from complex surveys analyses. It is clear that the estimates obtained with aid of finite weights which allow the used sample design, nonresponse and potentially also coverage error better reflect the reality.

Acknowledgements

The paper was supported by grant from Grant Agency of VEGA no. 1/0393/16 entitled „European Union in Post Crisis Period – Macro and Microeconomic Aspects".

References

Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data*. Hoboken: Wiley and Sons.

Cochran, W. G. (1977). *Sampling Techniques*. New York: J. Wiley and Sons.

Coeurjolly, J.-F. (2015). Chapitre 2. Caractéristiques des distributions à une variable quantitative. Retrieved from http://www-ljk.imag.fr/membres/Jean-Francois.Coeurjolly/documents/L1/chap2_print.pdf

Dagnelie, P. (1998). Statistique Théorique et Appliquée. Tom 1 – Statistique Descriptive et Bases de l` Inférence Statistique. Paris: DeBoeck and Larcier.

Halley, R. M. (2004). Measures of Central Tendency, Location, and Dispersion in Wage Survey Research. *Compensation and Benefits*, 2004/36, 39 (2004) pp. 39-52.

Levy, P. S. & Lemeshow, S. (2008). *Sampling of Populations. Methods and Applications. Fourth Edition.* Hoboken: Wiley and Sons.

Lohr, S. L. (2010). *Sampling: Design and Analysis. 2nd edition*. Boston: Brooks/Cole.

Piegorsch, W. W. (2015). *Statistical Data Analysis. Foundations for Data Mining, Informatics, and Knowledge Discovery.* Chichester: Wiley and Sons.

Marfels, Ch. (1971). *Absolute and Relative Measures of Concentration Reconsidered. Kyklos. International Review for Social Sciences*, Volume 24, Issue 4, pp. 753–766.

Terek, M. & Tibensky, M. (2014): Outliers and Some Non-Traditional Measures of Location in Analysis of Wages. *European Scientific Journal*, September 2014, Special Edition, Vol. 1, ISSN 1857 – 7881. Retrieved from http://eujournal.org/index.php/esj/article/view/4116

Terek, M. (2017). Regional Incomes Structure Analysis in Slovak Republic on the Basis of EU-SILC Data. *Scientific Annals of Economics and Business*, Vol. 64, No. 2. Retrieved from http://saeb.feaa.uaic.ro/index.php/saeb/issue/view/17/showToc

European Union Statistics on Income and Living Conditions (EU-SILC). Retrieved from http://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions.