

Variance estimation in EU-SILC survey

Mārtiņš Liberts*

* Mathematical Support Division, Central Statistical Bureau of Latvia
(Martins.Liberts@csb.gov.lv)

Abstract: The implementation of variance estimation methods – resampling methods (dependent random group method and jackknife) and linearization in EU-SILC survey are discussed in the paper. The paper focuses on estimation of variance for totals, ratios of two totals and Gini coefficient. However developed methodology could be used also for other statistics. The developed program (in SPSS®) for estimation of variance for arbitrary sample with possibilities of adjusting the parameters of methods applied will be also presented.

1 Introduction

Measurement of accuracy is important part in production of statistics based on survey sampling. The most common measure of accuracy is sampling error. The task of study is to develop methodology for estimation of sampling errors for complex (nonlinear) statistics and to apply it to Household Budget Survey (HBS) and EU-SILC (SILC).

2 Population Parameters

Three types of population parameters will be considered in the paper – total, the ratio of two totals and Gini coefficient.

Parameter	Population parameter	Estimate of parameter
Total	$X = \sum_{i=1}^N x_i$	$\hat{X} = \sum_{i=1}^n x_i w_i$
The ratio of two totals	$R = \frac{X}{Y}$	$\hat{R} = \frac{\hat{X}}{\hat{Y}}$
Gini index	$G = 100 \left(\frac{2 \sum_{i=1}^N x_i R_i - X}{NX} - 1 \right),$ $R_i = \sum_{j=1}^i 1 - \text{Rank of unit } i \text{ if sorted ascending by } x_i$	$\hat{G} = 100 \left(\frac{2 \sum_{i=1}^n x_i w_i \hat{R}_i - \hat{X}}{N\hat{X}} - 1 \right),$ $\hat{R}_i = \sum_{j=1}^i w_j - \text{Estimate of rank of unit } i \text{ if sorted ascending by } x_i$

3 Design of Surveys

Both surveys considered in the study share similar design. Households and individuals are survey units. Two-stage sampling is used for households; two-stage cluster sample is used for individuals.

Stratified systematic *pps* (sampling with probability proportional to size) sample of population census (2000) areas is used at the first stage. Stratification is made by degree of urbanisation – Riga, 6 other largest cities, towns and rural areas (four strata). PSUs are selected by several starting points (6 or 3 for HBS, 4 for SILC).

Simple random sampling of households is used at second stage.

All individuals from selected households are sampled – so households form clusters of individuals.

4 Estimation of Sampling Errors

It is hard to find direct estimators of sampling errors for estimates of complex statistics – especially in case of sampling design described in previous section. The approximation methods are used as alternative. Re-sampling methods (dependent random groups and jackknife) and linearization methods are considered in the paper.

4.1 Dependent Random Groups

The sample s from population U is divided in A non-overlapping subgroups s_1, \dots, s_A . The sample s should be divided so that all subgroups preserve the same sampling design as the sample s . The estimate of population parameter θ could be estimated as $\hat{\theta}_1, \dots, \hat{\theta}_A$. It is possible to estimate a variance of $\hat{\theta}$ by

$$\hat{V}_{DRG2} = \frac{1}{A(A-1)} \sum_{a=1}^A (\hat{\theta}_a - \hat{\theta})^2 \quad (1)$$

4.2 Jackknife

Similarly to dependent random groups technique the sample is divided in non-overlapping A sub-samples. The parameter θ is estimated from the sample s by deleting one of sub-sample for each $a = 1, \dots, A$. The resulting estimates $\hat{\theta}_{(a)}$ are used to estimate the variance of $\hat{\theta}$ by

$$\hat{V}(\hat{\theta}) = \frac{A-1}{A} \sum_{\alpha=1}^A (\hat{\theta}_{(\alpha)} - \hat{\theta})^2 \quad (2)$$

4.3 Linearization

The idea of linearization is to estimate a variance of complex statistics using the same estimator of variance as for totals. The goal of linearization is to find z_i for each unit in the sample so that variance of $\hat{\theta}$ could be approximated by

$$\hat{V}(\hat{\theta}) \approx \hat{V}\left(\sum_{i \in s} \frac{z_i}{\pi_i}\right) \quad (3)$$

Differentiable parameters can be linearized by expansion in Taylor series. For ratio of

two totals $R = \frac{Y}{X}$ z_i can be expressed in form

$$z_i = \frac{1}{X}(y_k - Rx_k) \quad (4)$$

Broader class of parameters can be linearized using extended theory by J. C. Deville (1999). For example for Gini coefficient z_i can be expressed in form

$$z_i = \frac{2x_k \left(\sum_{i \in U} 1(x_i \leq x_k) \right) + 2 \sum_{i \in U} x_i 1(x_i \geq x_k) - x_k - (Gini + 1) \left(\sum_{i \in U} x_i + Nx_k \right)}{N \sum_{i \in U} x_i} \quad (5)$$

5 Software for Estimation of Sampling Errors

To apply the theory described in previous section software in SPSS macro language has been developed.

5.1 Possibilities of the Software

It is possible to use the software for both single stage and multi-stage sampling. In case of multi-stage sampling errors are estimated at the level of PSUs. Stratification is allowed at the first stage.

Design weights should be available for software. For estimation design weights are increased proportionally to ratio of full sample size and sub-sample size. It is possible to apply non-response correction for user defined response homogeneity groups and post-stratification by one variable.

It is possible to estimate sampling errors for totals (SUM), ratio of two totals (RATIO) and Gini coefficient (GINI). Linearization of RATIO and GINI is possible to speedup the execution of software.

It is possible to use two re-sampling methods for estimating of sampling errors – jackknife and dependent random groups technique. Methods are applied at the level of PSUs. Correction of finite population is applied at level of PSUs.

User can freely choose the number of sub-samples and how sub-samples are created. PSUs could be sub-grouped in random or user defined order. The grouping of sub-groups and sub-sampling of these groups is possible.

Sample units can be divided in sub-units by applying parameters of sample unit to corresponding sub-units. For example Gini coefficient has to be estimated at individual level by applying to each individual equalised income. The income of household is divided by equalised household size (according to modified OECD scale) and the result is applied to all household members. Household is sample unit and individuals are sub-units.

5.2 Base of the Software

The software is written in SPSS® syntax using macro commands. Currently it is based on six macro commands:

- !linrat – linearization of ratio;
- !lingini – linearization of Gini coefficient;
- !estim – estimator of indicator;
- !weight – weighting of sub-sample;
- !e_tion – estimation of indicator using estimator and weights;
- !proc – estimation of sampling error;
- !proc_u – main procedure.

User can control the software using several parameters. For example:

- File – survey data file (in SPSS format);
- Strata – variable of stratifications;
- Psu – variable of PSUs;
- Diz_sv – variable of design weights;
- Meth – method of resampling – dependent random groups or jackknife;
- E_tor – estimator;
- Lin – linearization (Yes/No);
- Div – number of sub-samples;
- And other parameters.

Example of execution of the software:

```
!proc_u
dir "C:\Darbs\Stockholm\DRG\files\SILC"
file "C:\Darbs\Stockholm\DRG\Data\SILC\SILC2005_data_ver02.sav"
p_file "C:\Darbs\Stockholm\DRG\Data\SILC\dem_info.sav"
strata=prl /
psu=atk iecirk /
pop_psu=4263
hh_id=db030 /
per_sk=per_sk
diz_sv=diz_sv
resp=resp
resp_gr=atk iecirk /
p_gr=prl /
p_var=per_sk
p_tot=iedz_sk
meth=DRG JACK /
rorder=0 /
repeat=1
psu_gr=sel_nr /
order=sel_nr /
div=4 /
e_tor=RATIO /
lin=0 /
level=H /
eqscale=per_sk /
var=hh07n hsl3n /
fast=1.
```

The software is good tool for research. It is possible to test different methods and parameters of methods for estimation of sampling error. The software has been used for estimation of sampling errors in EU-SILC and HBS surveys. It has been tested on different SPSS versions – SPSS 11.5, SPSS 12 and SPSS 14.

6 Results

The software has been used for estimation of sampling errors in EU-SILC 2005 survey. The next table shows results of sampling errors of two indicators – Lowest monthly income to make ends meet (X) and Total housing cost (Y).

Table 1 Estimates of sampling errors in EU-SILC survey

Method	Estimator	Estimation	Estimation of variance	Coefficient of Variation (%)
Dependent Random Groups	SUM(X)	39 351 774.67	821 235 601 716	2.30
	SUM(Y)	337 079 686.10	14 605 983 870 634	1.13
	SUM(X)/SUM(Y)	0.12	0.000004037	1.72
	SUM(Y)/SUM(X)	8.57	0.021500621	1.71
	GINI(X)	39.71	0.475	1.74
	GINI(Y)	30.25	0.673	2.71
Jackknife	SUM(X)	39 351 774.67	831 832 862 430	2.32
	SUM(Y)	337 079 686.10	14 743 756 770 632	1.14
	SUM(X)/SUM(Y)	0.12	0.000003679	1.64
	SUM(Y)/SUM(X)	8.57	0.019817048	1.64
	GINI(X)	39.71	0.530	1.83
	GINI(Y)	30.25	0.667	2.70

Study about the linearization shows that it could be used to get faster estimates. In this case the estimates of sampling error are almost the same comparing estimates with and without linearization.

Table 2 Estimates of sampling errors using linearization for Gini coefficient

Method	Estimator	Number of sub-samples	Estimate of CV without linearization	Estimate of CV with linearization	Comparison
DRG	GINI	3	2.672	2.679	100.3%
DRG	GINI	4	2.284	2.277	99.7%
DRG	GINI	6	2.333	2.237	95.9%
DRG	GINI	12	1.923	1.849	96.2%
JACK	GINI	3	3.062	3.049	99.6%
JACK	GINI	4	1.999	1.999	100.0%
JACK	GINI	6	2.566	2.564	99.9%
JACK	GINI	12	2.011	2.010	100.0%

Estimates of sampling error are dependent on methodology of creating sub-samples (number of sub-samples, order of PSUs). The estimates of CV by different sub-sampling are varying. It can be seen in next table.

Table 3 Estimates of sampling errors by different sub-sampling

Nr	Method	Estimator	Number of sub-samples	Estimate of CV
1	DRG	GINI	52	2.224
2	JACK	GINI	52	2.680
3	JACK	GINI	26	2.649
4	DRG	GINI	34	2.449
5	DRG	GINI	42	2.312
1	JACK	RATIO	52	1.310
2	JACK	RATIO	42	1.284
3	DRG	RATIO	52	1.444
4	JACK	RATIO	20	1.350
5	DRG	RATIO	34	1.389

7 Conclusions

The software – created during the research is a good tool for using different methods of estimation of sampling errors. The software can be upgraded with additional methods or estimators of indicators. Analysis of linearization method shows that linearization is useful method in estimation of sampling errors. The analysis about the results of the survey will be continued.

References

A Publication

Central Statistical Bureau of Latvia (2005), "Mājsaimniecības budžets 2004. gadā", Rīga.

A Book

J. C. Deville (1999), "Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques", *Survey Methodology*, Statistics Canada, Vol. 25, No. 2, 193-203.

A Publication

European Commission, Eurostat, *The SAS macro for linearizing EU-SILC complex income indicators, User Guide*, Directorate F: Social Statistics and Information Society, Unit F-3: Living conditions and social protection statistics.

A Journal Article

J. Lapiņš, E. Vaskis, Z. Priede, S. Bāliņa (2002), "Household Sample Surveys in Latvia", *Statistics in Transition Journal of the Polish Statistical Association*, Volume 5, Number 4.

An Unpublished Paper

M. Liberts (2005), "Izlases apsekojumu teorija (Survey Sampling)", LU, Rīga.

An Unpublished Paper

M. Liberts (2004), "Prakses darba atskaite", LU, Rīga.

A Book

S. L. Lohr (1999), "Sampling: Design and Analysis", Brooks/Cole Publishing Company, Pacific Grove, Calif.

A Journal Article

A. Sandström, J. H. Wretman, B. Waldén (1988), "Variance Estimators of the Gini Coefficient – Probability Sampling", *Journal of Business & Economic Statistics*, Vol. 6, No. 1, American Statistical Association.

A Book

SPSS Inc (2002), "SPSS® Syntax Reference Guide".

A Book

C.-E. Särndal, B. Swensson, J. Wretman (1992), "Model Assisted Survey Sampling", Springer-Verlag, New York

A Website

Wikipedia, http://en.wikipedia.org/wiki/Gini_coefficient

A Book

K. M. Wolter (1985), "Introduction to Variance Estimation", Springer-Verlag