

## Variance Estimation in EU-SILC Survey

Mārtiņš Liberts  
■ Central Statistical Bureau of Latvia

## Task

- To estimate sampling error for Gini coefficient estimated from social sample surveys (EU-SILC)
- Estimation of sampling errors for totals and ratios was also analysed

8 November 2006

EU-SILC Methodological Workshop

2

## Gini coefficient

- The Gini coefficient is a measure of inequality of a distribution, defined as the ratio of area between the Lorenz curve of the distribution and the curve of the uniform distribution, to the area under the uniform distribution
- It is often used to measure income inequality
- It is a number between 0 and 1, where 0 corresponds to perfect equality (e.g. everyone has the same income) and 1 corresponds to perfect inequality (e.g. one person has all the income, and everyone else has zero income)

8 November 2006

EU-SILC Methodological Workshop

3

## Estimates of Gini Index

- Income inequality
  - by Gini index
  - Estimated from HBS

1996	1997	1998	1999	2000	2001	2002	2003	2004
31	31	32	31	34	NA	34	36	36

8 November 2006

EU-SILC Methodological Workshop

4

## EU-SILC

- Target population – population of Latvia living in private households
- Main variables of interest – income at household and individual level
- Yearly survey – organised once in a year
- Started in 2005 in Latvia

8 November 2006

EU-SILC Methodological  
Workshop

5

## Sampling Design

- Two-stage sampling design
- The first stage – stratified systematic pps sampling of census counting areas
- The second stage – simple random sampling of households (addresses)
- All individuals belonging to selected household are surveyed

8 November 2006

EU-SILC Methodological  
Workshop

6

## First stage sampling

- The list of census counting areas was created for the last population census in 2000 (4279 areas)
- Census counting area is relatively small geographical area
- The size of area is defined by number of households in area
- Areas are stratified in four strata by urbanisation degree (Riga, 6 other cities, towns and rural areas)

8 November 2006

EU-SILC Methodological  
Workshop

7

## Estimation of sampling errors

- Re-sampling methods are used
  - Dependent random group (DRG) method
  - Jackknife method
- Methods are used at the PSU level
- Both methods use the same resampling mechanisms – by dividing the whole sample in non-overlapping sub-samples

8 November 2006

EU-SILC Methodological  
Workshop

8

## DRG and Jackknife

- From each sub-sample the parameter  $\theta$  is estimated
- The estimate of variance
- The parameter  $\theta$  is estimated by deleting each sub-sample
- The estimate of variance (in case of stratified sampling)

$$\hat{V}_{DRG}(\theta) = \frac{1}{A(A-1)} \sum_{k=1}^A (\theta_k - \theta)^2$$

$$\hat{V}_{JK}(\theta) = \frac{A-1}{A} \sum_{k=1}^A (\theta_k - \theta)^2$$

8 November 2006

EU-SILC Methodological  
Workshop

9

## Resampling

- Resampling can be done in two ways:
  - Using the same sampling scheme
  - Using randomisation

8 November 2006

EU-SILC Methodological  
Workshop

10

## Randomisation

- PSUs can be grouped in sub-samples in random order
  - In this case variance estimate will differ each time the variance estimator is applied

8 November 2006

EU-SILC Methodological  
Workshop

11

## Linearization

- In case of complex statistics the variance estimator becomes also more complex compared to variance estimator of total
- The linearization technique can be applied for complex statistics to get approximate variance estimate
- The goal of linearization is to find  $z_i$  for each unit in the sample so that variance of estimate of parameter  $\theta$  could be approximated by

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

8 November 2006

EU-SILC Methodological  
Workshop

12

## Linearization

- In case of differentiable functions (for example Ratio of two totals) the expansion of estimator in Taylor series can be applied to linearize the estimator
- In case of non-differentiable functions the expanded theory (Deville, 1999) can be used

8 November 2006

EU-SILC Methodological  
Workshop

13

## Linearization of Gini Index

- Gini Index can be linearized by

$$\Pi_i(M) = \frac{2x_i \left( \sum_{j \in U} 1(x_j \leq x_i) \right) + 2 \sum_{j \in U} x_j 1(x_j \geq x_i) - x_i - (T(M) + 1) \left( \sum_{j \in U} x_j + Nx_i \right)}{N \sum_{j \in U} x_j}$$

8 November 2006

EU-SILC Methodological  
Workshop

14

## Program

- The program is written in SPSS mainly using the macro commands
- It is possible to estimate sampling error for arbitrary two stage sampling design using DRG un Jackknife methods at PSU level
- Sampling error is estimated for totals (SUM), ratios of two totals (RATIO) and Gini index (GINI)

8 November 2006

EU-SILC Methodological  
Workshop

15

## Features of Program

1. The nonresponse correction at user defined groups
2. Poststratification of weights
3. Linearization of RATIO and GINI
4. Selection of number of subgroups
5. PSU ordering in random or user-defined order
6. PSU grouping
7. Estimation of parameters for sampling units or sublevel units

8 November 2006

EU-SILC Methodological  
Workshop

16

## Procedures in program

- !linrat – linearization of RATIO
- !lingini – linearization of GINI
- !estim – estimator
- !weight – weighting
- !e\_tion – estimation
- !proc – basic procedure
- !proc\_u – main procedure

8 November 2006

EU-SILC Methodological  
Workshop

17

## Parameters

- File – sample data file
- Strata – stratification variable
- Psu – PSU variables
- Diz\_sv – variable for design weights
- Meth – method of estimation of sampling errors
- E\_tor – estimators
- Lin – linearization
- Div – numbers of subgroups
- Other parameters

8 November 2006

EU-SILC Methodological  
Workshop

18

## Example

```
!proc_u
dir="C:\DRG\SILC\files"
file="C:\DRG\SILC\SILC2005_data_ver02.sav"
p_file="C:\DRG\SILC\dem_info.sav"
strata=pr1 / psu=atk iecirk / hh_id=db030 /
per_sk=per_sk diz_sv=diz_sv
resp=resp resp_gr=atk iecirk / p_gr=pr1 /
p_var=per_sk p_tot=iedz_sk
meth=drg jack / rorder=0 / repeat=1 psu_gr=sel_nr /
order=sel_nr /
div=2 4 8 12 16 / e_tor=sum ratio gini / lin=0 /
level=P / eqscale per_sk / var=hh07n hsl3n /
fast=1.
```

8 November 2006

EU-SILC Methodological  
Workshop

19

## Variables

- SILC
  - Total housing cost
  - Lowest monthly income to make ends meet

8 November 2006

EU-SILC Methodological  
Workshop

20

## Estimates of parameters

SILC (estimates)		
	Housing	Income
GINI	39.7	30.3
RATIO	0.117	8.566
SUM	39 351 775	337 079 686

8 November 2006

EU-SILC Methodological  
Workshop

21

## Results SILC (Gini)

SILC (CV, %)					
Estimator	Number of subgroups	Dependent Random		Jackknife	
		Housing	Income	Housing	Income
Gini	2	0.15	1.60	0.13	1.46
	4	1.96	1.11	1.73	0.95
	8	2.31	1.65	1.96	1.51
	12	1.79	1.40	1.55	1.41
	16	2.31	1.92	1.99	1.58
	20	2.01	1.88	1.68	1.51
	24	1.99	1.70	1.62	1.40
	28	1.98	2.04	1.74	1.61
	32	2.02	2.12	1.68	1.77

8 November 2006

EU-SILC Methodological  
Workshop

22

## Results SILC

Ordered by sel\_nr

SILC (mean CV, %)				
Estimator	Dependent Random Group		Jackknife	
	Housing	Income	Housing	Income
Gini	1.83	1.71	1.56	1.47
Ratio	2.40	2.29	2.10	2.11
Total	2.15	1.06	1.86	0.96

Ordered by sel\_nr3

SILC (mean CV, %)				
Estimator	Dependent Random Group		Jackknife	
	Housing	Income	Housing	Income
Gini	1.66	1.96	1.56	1.79
Ratio	2.00	1.87	1.67	1.67
Total	2.28	1.26	1.92	1.13

8 November 2006

EU-SILC Methodological  
Workshop

23

## Random ordering

- The stability (variance) of variance estimates was analysed
- Both methods works similarly and higher number of subgroups gives more stable results

8 November 2006

EU-SILC Methodological  
Workshop

24

## Linearization

		Estimate	Variance1	Varinace2	CV1	CV2
RATIO	DRG	1.160045	0.00019685	0.00018942	1.21	1.19
	JACK	1.160045	0.00019486	0.00019628	1.20	1.21
GINI	DRG	35.90869	0.908	0.925	2.65	2.68
	JACK	35.90869	1.206	1.198	3.06	3.05

8 November 2006

EU-SILC Methodological  
Workshop

25

## Conclusions

- The DRG and JACK methods gives similar results
- The variance estimates are dependent on ordering of PSUs
- Higher number of subgroups gives more stable results
- Linearization can be used to simplify variance estimation

8 November 2006

EU-SILC Methodological  
Workshop

26