# Variance estimation for complex survey data and microsimulation

Tim Goedemé
Lorena Zardo Trindade
Herman Deleeck Centre for Social Policy

19 January 2018
EUROMOD Winter School, University of Antwerp

# Introduction to exercises

Variance estimation for EU-SILC

- Run the do-file "**SvysetEUSILC-Year-version"** according to the year and data version of your choice [available on https://timgoedeme.com/eu-silc-standard-errors/]. Do not forget to replace the original EU-SILC D-file by the new one, containing the constructed sample design variables.
  - Merge the new file for household register (D-file), constructed in a, with personal register (R-file), household data (D-file) and personal data (P-file)
- Or, merge csv files containing the new standard design variables (*strata1* and *psu1*)

- There are three key survey design variables weight, psu1 and strata1

# Variance estimation for EU-SILC

```
.
.  * Case 1: Assuming a simple random sample of persons
.
.         svyset _n [pw=rb050]

      pweight: rb050
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>

.         svy, subpop(if country=="ES"): prop dep4
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =       1      Number of obs    =      108,241
Number of PSUs   = 108,241      Population size =  122,979,427
                               Subpop. no. obs =       33,573
                               Subpop. size    = 46,354,779.1
                               Design df       =      108,240

--------------------------------------------------------------
             |             Linearized
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
dep4         |
           0 |   .9415797   .0019045      .9377331    .9452026
           1 |   .0584203   .0019045      .0547974    .0622669
--------------------------------------------------------------
```

```
.
. * Case 4: Assuming a stratified sample of clusters
.
.         svyset psu1 [pw=rb050], strata(strata1)

      pweight: rb050
          VCE: linearized
  Single unit: missing
     Strata 1: strata1
         SU 1: psu1
        FPC 1: <zero>

.         svy, subpop(if country=="ES"): prop dep4
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      18      Number of obs    =      33,573
Number of PSUs   =   1,996      Population size =  46,354,779
                                Subpop. no. obs =      33,573
                                Subpop. size    =  46,354,779
                                Design df       =       1,978

--------------------------------------------------------------
             |             Linearized
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
dep4         |
           0 |   .9415797   .0037197      .9338431    .9484615
           1 |   .0584203   .0037197      .0515385    .0661569
--------------------------------------------------------------
Note: 117 strata omitted because they contain no subpopulation
      members.
```

# Introduction to exercises

- Data (dta + csv)
  - *import delimited [include file path]*
  - Synthetic data on BE, RO, HU and IT

- DASP

# Exercise 1

Whether you use survey settings, or not, can make a tremendous difference on the estimated sampling variance, confidence intervals, etc. With this exercise, you will use *svyset* to declare survey settings, *svydescribe* to check the structure of the sample and will compare estimated standard errors for several survey design settings.

a. Using data for each country in turn, assume the survey sample designs listed below and calculate the standard errors ratio for the proportion and the total severely materially deprived (*dep4*). Please note that before survey commands can be used, the survey design must be declared by using the *svyset* command (*svyset [ psu ] [ weight ] [ , design_options options ]).* For design option pleas use *help svyset*
   - simple random sample of persons
   - simple random sample of households
   - simple random sample of ultimate clusters [use *psu1*]
   - stratified sample of clusters [use *psu1* and *strata1*]

b. Use the *svydescribe* command to describe the survey design of each country in the dataset. How does it change with variations in survey settings as defined above?

c. (Additional) Explore Stata *estat effects*

# Exercise 2

Data is often weighted if the purpose is to infer to the broader population. In this exercise, you will compare weighted with unweighted results. Some regression commands offer the possibility to use 'robust standard errors'. What are the differences with using full sample design features?

Using data for each country in turn:

a. compare the unweighted mean of the household disposable income (*hydisp*) and its standard error, ignoring survey design, with the weighted mean and standard error allowing for all features of the survey design (*svy*);
b. do a similar comparison for an ordinary least-squares regression of *hydisp* on gender (*female)* and adult status *(adults)*;
c. regress *hydisp* on gender (*female)* and adult status *(adults)* using weights and cluster-robust standard errors, and compare with the previous results.
d. Using appropriate survey settings, estimate the ratio of Family/Children related allowances (*fambens*) and household disposable income (*hydisp*). How does this differ from the average share of Familly/Children related allowances in total disposable household income? Why?

*Note: use the option vce (cluster) for cluster-robust standard errors*

# Exercise 3

Using data for Hungary:

a. Use the *if* restriction and *subpop* options to estimate the proportion of healthy (*healthy*) adult (*adult14*) women (*female*) in the population who have been identified as being at risk of poverty (*arop60*). Compare the results and indicate which one is the most appropriate and why.
b. Using the at risk of poverty indicator (*arop60*), compare the sampling variance of the proportion of poor among individuals aged 25 to 60 (*ageC4==2| ageC4==3*) and the total population. Repeat the estimation for individuals aged 25 to 30 (*ageC4==2*).
c. Compare the sampling variance of the mean equivalised household disposable income (*hystd*) for women, when estimated using the *subpop* option, with the mean estimated using the *over* option. What if, instead of using a binary variable, we use a categorical variable, as marital status (*marital*). Which command option would be more efficient to estimate the mean for each category, *subpop* or *over*?
d. Estimate the share of the following income components in the total disposable household income (*hydisp*) (at the aggregate level) for individuals aged 20 to 60 years old, by tenure status (*tenure*)
    - Family/Children related allowances (*fambens*)
    - Tax on income and social contributions (*taxsc*)

*Note: use options subpop, over and commands prop, ratio, total, mean*

# Exercise 4

Using data for Italy:

a. Consider 2 regions as independent subsamples (IT1 and IT2) and verify if the difference and the ratio between the proportion of individuals living in a low work intensity household (*jobless*) is statistically different between both regions at 90, 95 and 99 percent level of significance.
b. Consider 2 age groups as subsamples (adults: 0=age<18; G2: 1=age>=18) and compare the proportion of individuals living in severe material deprivation (*dep4*) between both groups at 95 percent level of significance.
c. Estimate the following model and test if the coefficient associated to hours worked (*hourx*) is equal to the coefficient associated to permanent job (*permanentjob*).

$$lnwage = \beta_1\, lnhours + \beta_2\, permanentjob + \beta_3 age\ group + \varepsilon$$

d. Using the model estimated in **c**, test if the elasticity of wage to hours worked is equal to 1. What about 0.9?

*Note: use lincom and nlcom to test the significance of differences and the validity of linear and non-liner relations*

# Exercise 5

The Belgium government is willing to implement a family benefit over the next years. The new benefit will pay a grant of 280 euros to all children age 13 or younger (*hhnbr_child14=number of children per household*).

Using the equivalised household disposable income (*hystd*) as baseline income and the DASP package to estimate poverty and inequality measures:

a. indicate if the new benefit leads to a significant change in the FGT poverty indices when considering a **fixed poverty line**.
b. indicate if the new benefit leads to a significant change in the FGT poverty indices when considering a **floating poverty line**.
c. what happens to the inequality measure (Gini coefficient)?

*Note: Check thresh60 for poverty line values. Please ignore equivalence scales. Use lincom and nlcom to test the significance of differences*

# Exercise 6 (Extra)

New equivalence scales based on reference budgets are proposed for Belgium, Hungary, Italy, Romania.

Assuming the poverty threshold is exogenous, use the information on tables 1 and 2 to answer the questions below:

a. While keeping the original poverty threshold constant, estimate the proportion of individuals at risk of poverty on total population for each country.
b. When compared to the poverty measure estimated using the original equivalence scale:
   - Does the total proportion in poverty change significantly?
   - Do differences between adults and children change significantly?
   - Does the ratio of poverty among adults and children change significantly?
c. Please consider 95% of level of confidence and the following age groups: children (0-17); adult (18-64); old-age (65+) [see var *ageC3*].
d. Repeat exercises **a** and **b** allowing for changes in the poverty threshold, that is, recalculate the poverty threshold based on the modified equivalence scale.

*Note: Age groups are included in the data set.*

# Exercise 6 (Extra)

Table 1. Equivalence scale by country and school age group

| Population group | BE | HU | IT | RO |
|---|---|---|---|---|
| adults (excluding students) | 0.39 | 0.47 | 0.39 | 0.52 |
| children in primary school age | 0.43 | 0.51 | 0.36 | 0.5 |
| children in secondary school age + adults in tertiary education | 0.57 | 0.56 | 0.47 | 0.64 |

Table 2. Age groups and level of education

| Age group | Level of education |
|---|---|
| **0-5** | Pre-school |
| **6-11** | Primary school |
| **12-17** | Secondary school |
| **18-25** | Tertiary school (not mandatory) |
| **18+** | Adults excluding students |