Variance estimation for complex survey data and microsimulation

Tim Goedemé
Lorena Zardo Trindade
Herman Deleeck Centre for Social Policy

18 January 2018
EUROMOD Winter School, University of Antwerp

Universiteit Antwerpen

Conclusion

- Statistics & samples are a powerful tool
  - Need limited number of observations
  - Point estimate *and* estimate of precision

However, without an estimate of its precision, a point estimate is pointless…

- … at least for evidence-based policy-making

Universiteit Antwerpen

2

Requirements

- Estimating the sampling variance requires:

  - Sound sample designs

  - Good documentation of the sample (design)

  - Access to high-quality microdata with sufficient information on weighting, imputation and sample design

  - Adequate and consistent sample design variables

  - Adequate software, estimation methods, skills and expertise

Universiteit Antwerpen

3

Communication

- To researchers

- To policy-makers and politicians

- To the wider public

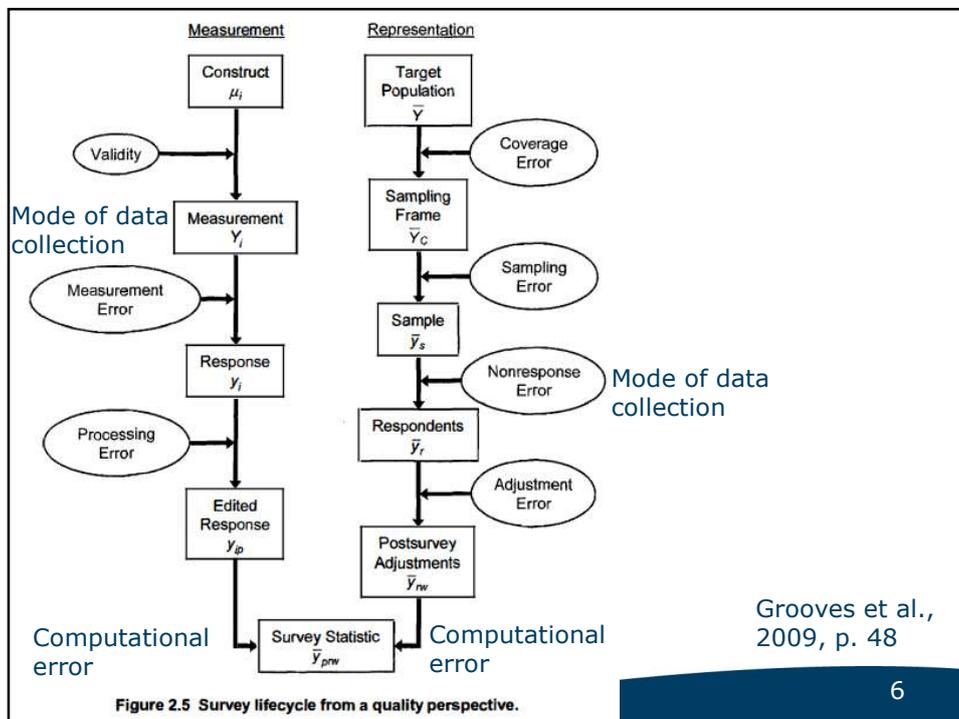- Improve awareness of both sampling and non-sampling errors

Universiteit Antwerpen

4

Communication

Measures of statistical reliability

Confidence interval > standard error

Standard error > degrees of freedom

Degrees of freedom > number of observations

Universiteit Antwerpen

5

Figure 2.5 Survey lifecycle from a quality perspective.

Grooves et al., 2009, p. 48

6

Comparability of income variables

- MetaSILC 2015: An assessment of the content and cross-country comparability of the EU-SILC benefit variables

- EU-SILC 2015 (and smaller database for 2010)

- Funding: Net-SILC 3

Universiteit Antwerpen

7

MetaSILC

- Knowledge of content (aggregation) and comparability of income variables is key

- Description of target variables in Doc065 & Quality reports not sufficiently detailed for:

  - Identifying exact classification of all income components in all countries

  - Evaluating level of cross-country comparability ('correct' classification) depends also on question)

Universiteit Antwerpen

8

Survey

- Online questionnaire among NSIs in 2 rounds

- Questions on all 34 income variables

- For each of the income components
  - official name (national language) and the equivalent name in English
  - the target variable code and name
  - the source of the income information used (register data, questionnaire, imputation)
  - information on gross-net collection
  - changes between wave 2010 and wave 2015
  - changes planned for future waves
  - additional questions on data processing of specific variables (HY030, PY050, PY021)

Universiteit Antwerpen

9

MetaSILC

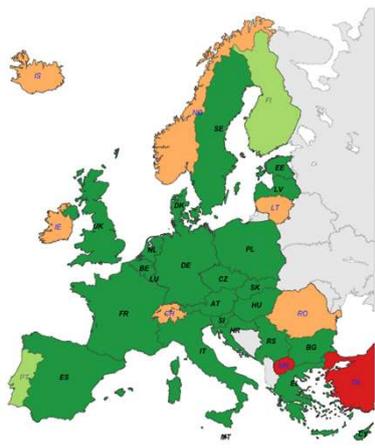- Excel database

- Detailed report

- Summary paper

Available: end 2018

Universiteit Antwerpen

10

Database

- Excel file with information on 26 countries

- The exact composition of all income variables of EU-SILC cross-sectional 2015 wave
  - 34 variables, over 2000 income components

- Latvia, Poland and Sweden
  - Income from benefits with information only on "mixed components"

Information provided
Information partly provided
No Information, data for EU-SILC 2015 available
No Information, data for EU-SILC 2015 not available

Universiteit Antwerpen

11

Findings

- Definition of variables & 'borderline cases'

- Non-compliance with Eurostat guidelines

- Level of detail of data collection

- Most survey data, but significant amount of register

- Net vs gross incomes & net-to-gross procedures are not consistent across countries

- Differences in imputation procedures

- Comparability issues across time

- Difference between information in MetaSILC 2015 and other sources (Euromod Reports, MISSOC, Quality reports)

Universiteit Antwerpen

12

MetaSILC

- Net-SILC 3 (2016-2020)

  - + information on health, housing and production for own consumption

  - + outlier treatment; imputation

Universiteit Antwerpen

13

Conclusion

- The sampling variance is an important challenge to indicators for evidence-based policy-making

- Increases awareness of both sampling and non-sampling errors

Universiteit Antwerpen

14

Conclusion

Key messages

1. If estimates are based on samples -> estimate and report SEs, CIs & p-values

2. Always take as much as possible account of sample design when estimating SEs, CIs & p-values

3. Never delete observations from the dataset

4. Never simply compare confidence intervals

Universiteit Antwerpen

15

Literature

- Goedemé, T. (2013) 'How much confidence can we have in EU-SILC?', *Social indicators research*, 110(1): 89-110, doi:10.1007/s11205-011-9918-2

- Heeringa, S. G., West, B. T. and Berglund, P. A. (2010), Applied Survey Data Analysis, Boca Raton: Chapman & Hall/CRC, 467p.

- Wolter, K. M. (2007), Introduction to Variance Estimation, New York: Springer, 447p.

- https://timgoedeme.com/eu-silc-standard-errors/.

Universiteit Antwerpen

17

Resources

Background materials

- Handouts

- Do-files & exercises

- https://timgoedeme.com/eu-silc-standard-errors/ (papers, do-files, csv-files)

- Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). Applied Survey Data Analysis. Boca Raton: Chapman & Hall/CRC

Universiteit Antwerpen

18

Contact details

- Tim.Goedeme@uantwerpen.be

- Lorena.ZardoTrindade@uantwerpen.be

Universiteit Antwerpen

19