Variance estimation for the main EU indicators of poverty and social exclusion: First results of the Net-SILC2 project

Guillaume OSIER (Statistics Luxembourg & Luxembourg Income Study)
Tim GOEDEME (University of Antwerp, Belgium)
Yves BERGER (University of Southampton, UK)

Meeting of country correspondents for the OECD income distribution database
Paris, 21-22 February 2013

Context – Net-SILC2 research project

- Research project funded by Eurostat, aiming at:
  - carrying out in-depth methodological work and comparative socio-economic research using EU Statistics on Income and Living Conditions (EU-SILC) data,
  - developing common tools and approaches regarding various aspects of data production,
  - managing the scientific organisation of international conferences on EU-SILC.
- Net-SILC2 brings together expertise from 16 European partners: the Luxembourg-based CEPS/INSTEAD Research Institute (Net-SILC2 coordinator), six National Statistical Institutes (from Austria, Finland, France, Luxembourg, Norway and the UK), the Bank of Italy, and academics from 8 research bodies.
- The project is divided into 26 work packages (WP) covering key methodological topics such as the use of income registers, the measurement of material deprivation in the EU or the implications of the EU-SILC following rules for longitudinal analysis.
- One WP dealing with standard error estimation and other sampling issues.

Aims of the WP

- The main objective of the WP is to develop a practicable set of recommendations on standard error estimation in EU-SILC both for data producers (NSIs) and data users:

  (a) suggestions concerning the concrete implementation procedures for computing standard errors at NSI's level (production database) and at database users level, i.e. non-NSI's level;

  (b) concrete recommendations for better recording of sampling design variables (e.g. suitable documentation and metadata), after reviewing the current practices on micro-data for the sample design variables (Goedemé, 2010)

- In addition, computer codes based on existing statistical packages (e.g. SAS, SPSS…) should be developed for computation of standard errors for the main "Europe 2020" indicators of poverty and social exclusion.

Key outputs so far

> Workshop on Standard error estimation and related sampling issues (Eurostat, 29-30th March 2012)

http://www.cros-portal.eu/content/workshop-2930-march-2012

Draft papers:

- BERGER, Y. "*Variance estimation for longitudinal measures for two stage cluster sampling designs: an application to the at-persistent-risk-of-poverty rate*"
- BERGER, Y. "*Variance estimation for measures of change for two stage cluster sampling designs: an application to the EU-SILC survey*"
- GOEDEMÉ, T. "*The EU-SILC sample design variables: critical review and recommendations*"
- OSIER, G. "*Design effect (deff) estimation*"
- OSIER, G. "Standard error estimation from user's perspective: main approaches"
- OSIER, G. & DI MEGLIO, E. "*The linearisation approach implemented by Eurostat for the first wave of EU-SILC: What could be done from the second wave onwards?*"

> Net-SILC2 Conference (Vienna, 6-7th December 2012)

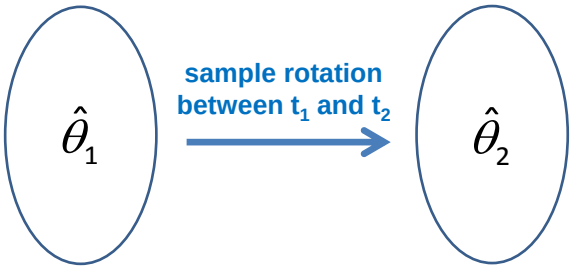
- BERGER, Y. ; GOEDEMÉ, T. & OSIER, G. "*Standard error estimation and related sampling issues*"

http://www.statistik.at/web_en/about_us/events/eu_silc_conference_2012/conference/index.html

Variance estimation methodology

- We have proposed to use direct variance formulas (Berger 2003) as a compromise solution between statistical accuracy and operational efficiency. This approach consists of a quick and easy algorithm to compute the variance of both cross-sectional and longitudinal measures using basic statistical techniques (linear regression)
- This regression-based approach can also be extended to cope with estimators of net change between two cross-sectional waves (Berger and Priam 2013)

Estimators of net changes



$$V\left(\hat{\theta}_2 - \hat{\theta}_1\right) = V\left(\hat{\theta}_2\right) + V\left(\hat{\theta}_1\right) - 2 \cdot Cov\left(\hat{\theta}_2, \hat{\theta}_1\right)$$

$$= V\left(\hat{\theta}_2\right) + V\left(\hat{\theta}_1\right) - 2\sqrt{V\left(\hat{\theta}_2\right)} \cdot \sqrt{V\left(\hat{\theta}_1\right)} \cdot \underbrace{Corr\left(\hat{\theta}_2, \hat{\theta}_1\right)}_{>0,\ ???}$$

Preliminary results

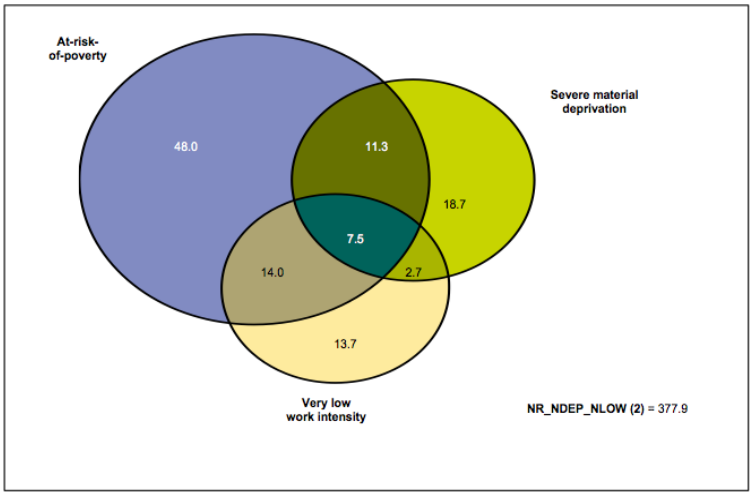
1) Standard errors for cross-sectional measures: at-risk-of-poverty or social exclusion indicator (AROPE), 2010
2) Standard errors for longitudinal measures: at-persistent-risk-of-poverty rate, 2006-2009
3) Standard errors for estimators of net changes: AROPE, 2009-2010

Assumptions:

- NUTS2 Region as proxy for stratification
- Ignore calibration and imputation

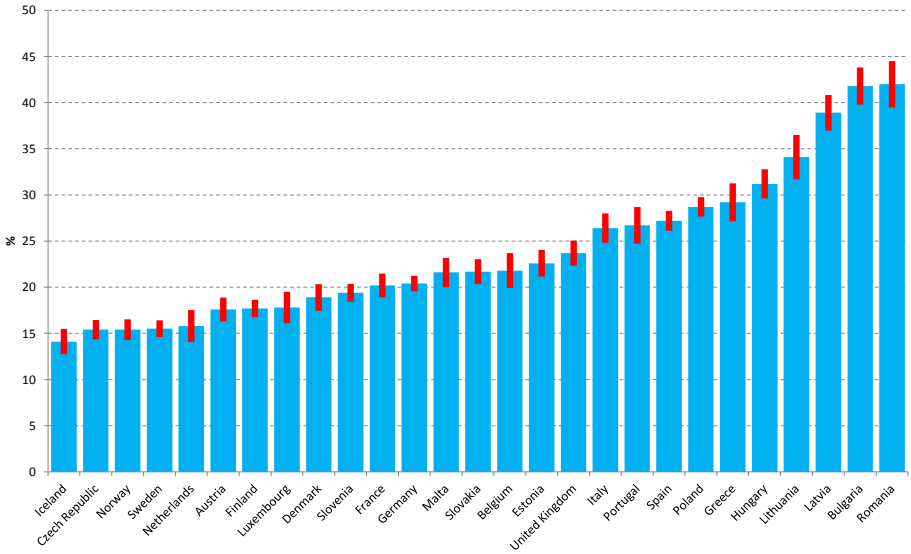
!!! Preliminary results: depend on the quality of the design variables !!!

The AROPE indicator

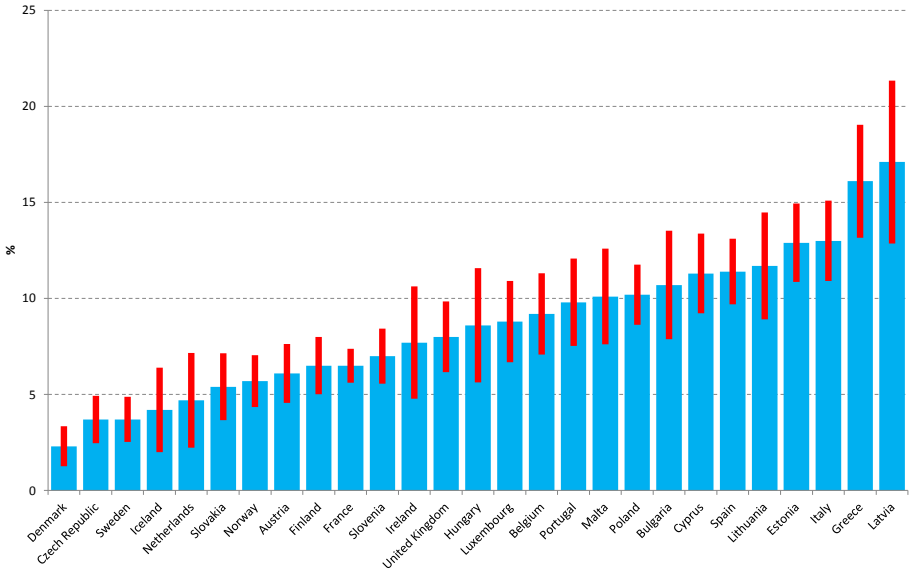


(1) The sum of the data for the seven groups at-risk-of-poverty or social exclusion differs slightly from the total (published elsewhere) due to rounding.
(2) Population neither at-risk-of-poverty, nor severely deprived nor living in households with very low work intensity
Source: Eurostat (online data code: ilc_ pees01)

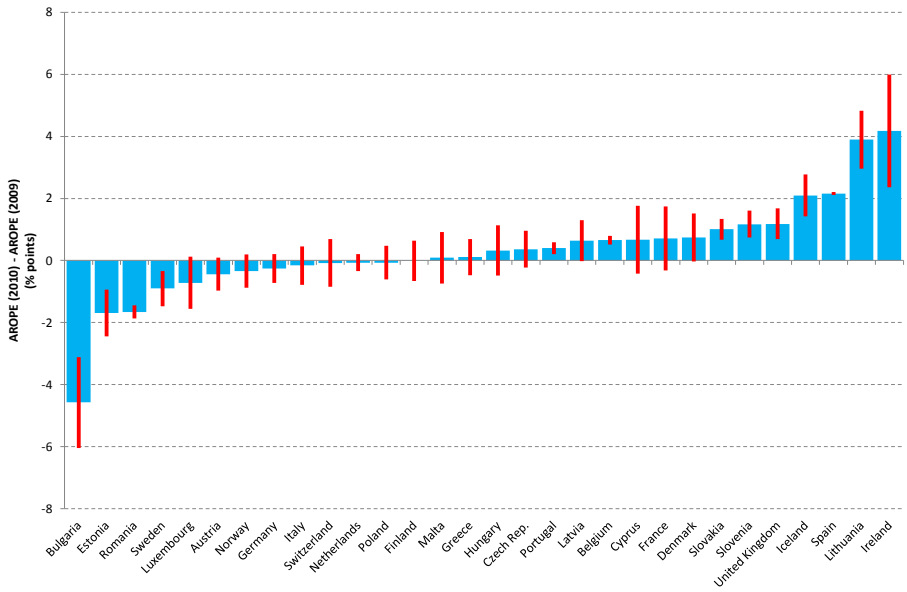
Cross-sectional std errors, AROPE, 2010



Persistent risk-of-poverty, 2006-2009



Standard errors for change, 2009-2010



User's perspective: main options

Standard error estimation from user's perspective (non-NSI's):

1. Release standard error estimates for a predefined list of target indicators
2. Include more sample design variables in the EU-SILC user data files
   - **DB060: PRIMARY SAMPLING UNITS (PSU)**
   - **DB062: SECONDARY SAMPLING UNITS (SSU)**
   - **DB070: ORDER OF SELECTION OF PSU**
   - **DB075: ROTATION GROUP**
   - **DB090: HOUSEHOLD CROSS-SECTIONAL WEIGHT**
   - **STRATUM CODE - CALIBRATION VARIABLES - CALIBRATION ADJUSTMENT FACTORS**
3. Use of replicate weights
4. Generalized Variance Functions: mathematical model which describes the relationship between a statistic and its variance

Variance Function: an example

- P_d = (estimated) at-risk-of-poverty rate for age group d
- n_d = sample size
- Variance model: $\mathrm{Re}l\mathrm{var}(P_d) = \frac{Var(P_d)}{P_d^2} = \alpha + \frac{\beta}{n_d}$

Based on sample observations, we estimate the model parameters α = 0.00146 and β = 11.106. Data users can henceforth derive the variance for any other at-risk-of-poverty rate P_d from the sample size n_d. For instance, for the 6-9 year-olds (n_d=592), we get:

$$CV(P_d) = \sqrt{0.00146 + \frac{11.106}{592}} = 14.2\%$$

Conclusion

- A simple approach for standard error estimation involving basic statistical techniques (linear regression)
- Take most of the characteristics of the survey into account: stratification, clustering, unequal probabilities of selection, reweighting for unit non-response and calibration to auxiliary data sources
- Deal with cross-sectional and longitudinal measures and can be adapted for measures of net change
- Can deal with complex parameters through the linearisation procedure (Deville 1999, Osier 2009)
- Need quality design variables

References

BEAUMONT, J.-F. & BISSONNETTE, J. (2011). Variance Estimation under Composite Imputation: The Methodology Behind SEVANI. *Survey Methodology*, vol. 37, pp. 171-179.

BERGER, Y. G. (2003). *A Simple Variance Estimator for Unequal Probability Sampling Without Replacement*. University of Southampton, Statistical Sciences Research Institute, Methodology Working paper M03/09. Available at: http://eprints.soton.ac.uk/7798/1/7798-01.pdf

BERGER, Y. G. & PRIAM, R. (2013). *A simple variance estimator of change for rotating repeated surveys: an application to the EU-SILC household surveys*.
University of Southampton, Statistical Sciences Research Institute. Available at: http://eprints.soton.ac.uk/347142/1/Berger_Priam_v1.pdf

DEVILLE, J-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, December 1999, vol. 25, no. 2, pp. 193-203.

GOEDEME, T. (2010). *The construction and use of sample design variables in EU-SILC. A user's perspective*. Report prepared for Eurostat. Available at:
http://www.centrumvoorsociaalbeleid.be/index.php?q=node/2157/en

OSIER, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, vol. 3, no. 3, pp. 167-195.
Available at: http://w4.ub.uni-konstanz.de/srm/article/view/369

SÄRNDAL, C-E. ; SWENSSON, B. & WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Thank you for your attention!