# 2. Variance estimation approach

The computation of standard errors for EU-SILC estimates is confronted with several challenges:

- complex sample designs involving stratification, geographical clustering, unequal probabilities of selection for the sample units and post-survey weighting adjustments (re-weighting for unit non-response and calibration to external data sources)
- rotating samples
- problems with quality, documentation and availability of sample design variables
- complex non-linear indicators, longitudinal indicators and indicators of net changes
- different methods of imputation used across countries
- confidentiality issues
- limited resources in terms of budget, staff and time both at national and EU level

Over the past few years, several projects, working groups, task forces and individual authors have addressed one or more of these challenges. However, the knowledge remains rather scattered and is not very accessible for National Statistical Institutes (NSIs) and the wider research community, especially for the non-statistician researchers.

Standard error estimates should reflect as much as possible the sample design, weighting procedures, imputation and the characteristics of the indicators of interest. Otherwise they may be severely biased. On the other hand, the increased complexity of EU-SILC, the widening of the user community and the increased reliance on EU-SILC for policy targeting and evaluation have enhanced the need for comparable, accurate as well as workable solutions for the estimation of standard errors and confidence intervals for EU-SILC based indicators. Therefore, we need an approach making a trade-off between statistical accuracy and operational efficiency. The proposed approach is general enough to be valid under most of the EU-SILC sampling designs, which is actually a challenge considering the important differences in sampling design between countries. The approach is also simple and easy to implement using standard statistical software, such as SAS, SPSS or R, and should require minimal computing power.

Re-sampling methods like Bootstrap or Jackknife are flexible enough to be applicable to the sampling designs and the target indicators used in EU-SILC, no matter their complexity (Verma and Betti, 2011). On the other hand, the computational effort may be considerable, which is not desirable when standard error estimates need to be produced quickly for a large number of target indicators, including breakdowns. That is why we have proposed to use direct variance estimators (Berger, 2004). The main assumption underlying such estimators is that sample units have been selected with replacement, which considerably simplifies the estimation of the variance. If sample units are selected without replacement, then this approach will lead to conservative estimates. The overestimation is negligible as long as the sampling fraction is close to zero. Note that this is nearly always the case with the EU-SILC sampling designs. Furthermore, those direct estimators can be easily extended to cover multi-stage designs by using the well-known 'ultimate cluster' approximation (e.g. Särndal et al., 1992) and to deal with complex non-linear indicators on the basis of the linearisation procedure (e.g. Deville (1999); Wolter (2007); Osier (2009)). In what follows, we further explain this approach to variance estimation in some detail. We first discuss the case of linear indicators before elaborating on the case of non-linear indicators. Subsequently we explain how multivariate linear regression offers an easy tool for estimating the variance both of linear and non-linear indicators. Finally, we elaborate on calibration, imputation and measures of net changes over time.

## 2.1 Case of linear indicators

Linear indicators are means, totals or proportions. The estimation of the variance of linear indicators is rather straightforward, and is covered in most textbooks on Statistics (e.g., Särndal et al (1992)). Consider a population $U$ composed of $N$ identifiable units (households or individuals). Let $s$ denote a sample of size $n$ drawn from $U$ using a probabilistic design so that every unit $k$ is having its own, known inclusion probability $\pi_k$. For example, in case of simple random sampling without replacement, the inclusion probability is $\pi_k = n / N$ for each $k$.

Suppose we wish to estimate the total $\theta = \sum_{k \in U} y_k$ , where $y_k$ is the value of a study variable $y$ for $k$. $y$ can be a continuous variable (e.g., household income), or a dummy variable for a population category (e.g., 1 if the person is unemployed, 0 otherwise). If $y$ is a dummy, then $\theta$ is a count (e.g., the total number of unemployed in the population). Let $\hat{\theta} = \sum_h \sum_i \sum_j \omega_{hij} y_{hij}$ be an estimator of $\theta$, for which an estimate of the standard error is required. We propose the following variance estimator:

$$\hat{V}(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum (y_{hi\bullet} - \bar{y}_{h\bullet\bullet})^2, \qquad (1)$$

where $y_{hi\bullet} = \sum_{j=1}^{m_{hi}} \omega_{hij} \bullet y_{hij}$ and $\bar{y}_{h\bullet\bullet} = \left( \sum_{i=1}^{n_h} y_{hi\bullet} \right) \Big/ n_h$

- $h$ is the stratum label, with a total of $H$ strata. If no stratification, the whole target population $U$ can be regarded as one large stratum ($H = 1$)
- $i$ is the label of the primary sampling unit (PSU) within stratum $h$, with a total of $n_h$ PSUs
- $j$ is the household label within PSU $i$ of stratum $h$, with a total of $m_{hi}$ households. In case of a one-stage sampling design, each household is regarded as a PSU
- $\omega_{hij}$ is the sampling weight for household $j$ in PSU $i$ of stratum $h$. The weights $\omega_{hij}$ are used to make inference about the population. They are usually adjusted for unit non-response and calibration
- $y_{hij}$ is the value of the study variable $y$ for household $j$ in PSU $i$ of stratum $h$

If $n_h = 1$ for some strata, the estimator (1) cannot be used. A solution is to collapse strata to create "pseudo-strata" so that each pseudo-stratum has at least two PSUs. Common practice is to collapse a stratum with another one that is similar with regard to the target variables of the survey (Rust and Kalton (1987); Ardilly and Osier (2007)).

## 2.2 Case of non-linear indicators

The estimator (1) is valid for linear indicators, i.e. means, totals and proportions. However, most of the EU-SILC indicators are non-linear (e.g., the at-risk-of-poverty threshold, the at-risk-of-poverty rate, the income quintile share ratio or the Gini coefficient). In order to estimate the variance of non-linear indicators, the linearisation approach may be used (Kovacevic and Binder 1997, Deville 1999, Demnati and Rao 2004, Wolter 2007, Osier 2009). The principle is to approximate a non-linear indicator by a linear form by retaining only the first-order term of a Taylor expansion. The variance of the linear approximation can be used as an approximation of the variance of the non-linear indicator considered. The linearisation procedure is justified on the basis of asymptotic properties of large samples and populations (Verma and Betti, 2005).

Suppose $\theta$ is a complex non-linear indicator. The variance of an estimator $\hat{\theta}$ of $\theta$ is estimated by:

$$\hat{V}(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum (z_{hi\bullet} - \bar{z}_{h\bullet\bullet})^2 \ . \qquad (2)$$

This is exactly the same formula as (1), except that the study variable $y$ is replaced by the "linearised" variable $z$. For example, if $\theta = \left(\sum_{k \in U} y_k\right)\left(\sum_{k \in U} x_k\right)^{-1} = Y\, X^{-1}$ is the ratio of two population totals, then the "linearised" variable is: $z_k = X^{-1}\left(y_k - \theta \cdot x_k\right)$ for all $k$. More examples can be found in Osier (2009).

## 2.3 Interpretation in terms of regression residuals

The differences $(y_{hi\bullet} - \bar{y}_{h\bullet\bullet})$ in (1) and $(z_{hi\bullet} - \bar{z}_{h\bullet\bullet})$ in (2) can be seen as the residuals of the linear regression of the PSU aggregates $y_{hi\bullet}$ and $z_{hi\bullet}$ on the dummy variables for each stratum category (Berger, 2004). These dummy variables are equal to 1 if the $i$-th PSU belongs to the stratum $h$, 0 otherwise. This provides a quick and easy way to compute the variance of both cross-sectional and longitudinal measures using basic statistical techniques (multivariate linear regression).

## 2.4 Calibration and imputation

The approach proposed here reflects survey design features such as stratification, multi-stage selection, unequal probabilities of inclusion for the sample units and post-survey weighting adjustments for unit non-response. On the other hand, a specific approach is needed to measure how calibration weighting (Deville and Särndal 1992) affects the variance. The effect of calibration on variance is expected to be significant in the "Nordic" countries like Denmark or Finland in which powerful auxiliary information from income registers is used for calibration. As shown by Deville and Särndal (1992), the effect of re-weighting for calibration on variance estimation can be allowed for by replacing the study variable by the residuals of the regression on the calibration variables, and by calculating the variance assuming no calibration. Such an approach is easy to implement as long as the calibration variables are available as well as the initial weights before calibration or, equivalently, the calibration adjustment factors (also called *g-weights*). Up to now, the EU-SILC database does not contain this information.

A major shortcoming of the proposed approach is that it does not take the imputation variance into account. Actually, the EU-SILC income variables have been heavily imputed, with different imputation methods used across countries, as well as across different income components. For the sake of simplicity, imputed values have been treated as true values. Such an assumption may lead to severely under-estimating the variance, particularly when the proportion of imputed values is important (Rao and Shao, 1992). However, variance estimation under imputation is not an easy task. Direct estimation formulas are very complex (Deville and Särndal, 1994) and method-specific. Thus, though significant, it does not seem realistic to try to estimate the imputation variance on a streamline basis, even more so that the imputation methods used in the EU-SILC vary greatly from one country to another. Nevertheless, the imputation variance might be estimated occasionally using for instance the SAS software SEVANI developed by Statistics Canada (Beaumont and Bissonnette, 2011). For hot-deck imputation, Berger and Escobar (2012) proposed an approach to estimate the variance of change in the presence of imputed values.