Survey Research Methods (2009)
Vol.3, No.3, pp. 167-195
ISSN 1864-3361
http://www.surveymethods.org

Variance estimation for complex indicators of poverty and inequality using linearization techniques

Guillaume Osier
Statistical Office of the European Communities (EUROSTAT), Luxembourg

The paper presents the Eurostat experience in calculating measures of precision, including standard errors, confidence intervals and design effect coefficients – the ratio of the variance of a statistic with the actual sample design to the variance of that statistic with a simple random sample of same size – for the "Laeken" indicators, that is, a set of complex indicators of poverty and inequality which had been set out in the framework of the EU-SILC project (European Statistics on Income and Living Conditions).
The Taylor linearization method (Tepping, 1968; Woodruff, 1971; Wolter, 1985; Tillé, 2000) is actually a well-established method to obtain variance estimators for nonlinear statistics such as ratios, correlation or regression coefficients. It consists of approximating a nonlinear statistic with a linear function of the observations by using first-order Taylor Series expansions. Then, an easily found variance estimator of the linear approximation is used as an estimator of the variance of the nonlinear statistic.
Although the Taylor linearization method handles all the nonlinear statistics which can be expressed as a smooth function of estimated totals, the approach fails to encompass the "Laeken" indicators since the latter are having more complex mathematical expressions. Consequently, a generalized linearization method (Deville, 1999), which relies on the concept of influence function (Hampel, Ronchetti, Rousseeuw and Stahel, 1986), has been implemented.
After presenting the EU-SILC instrument and the main target indicators for which variance estimates are needed, the paper elaborates on the main features of the linearization approach based on influence functions. Ultimately, estimated standard errors, confidence intervals and design effect coefficients obtained from this approach are presented and discussed.
Keywords: EU-SILC survey, non-linear statistics, influence function, standard error, confidence interval, design effect coefficient

1 The EU-SILC instrument

1.1 Introduction

European Statistics on Income and Living Conditions (EU-SILC) is actually the main instrument for the compilation of comparable indicators on social cohesion in the European Union (EU).[1] It consists of a series of national sample surveys that collect on an annual basis comparable multidimensional micro-data on income, poverty, social exclusion and living conditions. Every year, both cross-sectional data (pertaining to a given time or a certain time period) and longitudinal data (pertaining to individual-level changes over time) are collected over representative samples of households and individuals aged 16 or more.

The EU-SILC was launched under a gentleman's agreement with six EU-15 countries plus Norway in 2003 and relaunched in 2004 under a European Regulation with twelve EU-15 countries (Belgium, Denmark, Greece, Spain, France, Ireland, Italy, Luxembourg, Austria, Portugal, Finland and Sweden) plus Estonia, Norway and Iceland. In 2005, the rest of the EU-25 countries joined the project, whereas Bulgaria, Romania, Turkey and Switzerland commenced EU-SILC in 2006.

1.2 Policy context

EU-SILC is actually a key monitoring tool at EU level for the so-called "Lisbon Strategy", a pillar of which is to build a more inclusive European Union by making a notable impact on eradicating poverty by 2010. In particular, the Laeken European Summit in December 2001 formally adopted a list of outcome indicators of poverty and social exclusion. Those "Laeken" indicators allow monitoring in a comparable way of Member States' progress towards agreed EU objectives regarding fight against poverty and social exclusion.[2]

The above indicators are extracted from different EU data sources:

Contact information: Guillaume Osier, Statistical Office of the European Communities (EUROSTAT), 5 Rue Alphonse Weicker, L-2721 Luxembourg, Grand-Duchy of Luxembourg, e-mail: Guillaume.Osier@statec.etat.lu

[1] See also Clemenceau and Museux (2006) for a presentation of the instrument

[2] Further information regarding the "Laeken" indicators as well as the EU Social Inclusion Process is available on European Commission, Directorate General Employment, Social Affairs and Equal Opportunities: http://ec.europa.eu/social

Table 1: The "Laeken" indicators

(L1)	At-risk-of-poverty rate by various classifications
(L2)	Inequality of income distribution: S80/S20 quintile share ratio
(L3)	At-persistent-risk-of-poverty rate by gender (60% median)
(L4)	Relative median at-risk-of-poverty gap
(L5)	Regional cohesion (dispersion of regional employment rates)
(L6)	Long term unemployment rate
(L7)	Persons living in jobless households
(L8)	Early school leavers not in education or training
(L9)	Life expectancy at birth
(L10)	Self-defined health status by income level
(L11)	Dispersion around the at-risk-of-poverty threshold
(L12)	At-risk-of-poverty rate anchored at a moment in time
(L13)	At-risk-of-poverty rate before social transfers by gender
(L14)	Inequality of income distribution: Gini coefficient
(L15)	At-persistent-risk-of-poverty rate by gender (50% median)
(L16)	Long term unemployment share
(L17)	Very long term unemployment rate
(L18)	Persons with low educational attainment

- EU Labour Force Survey (EU-LFS) for *(L5)*, *(L6)*, *(L7)*, *(L8)*, *(L16)*, *(L17)* and *(L18)*
- Demographic sources for the indicator *(L9)*

With regard to the other indicators – *(L1)*, *(L2)*, *(L3)*, *(L4)*, *(L10)*, *(L11)*, *(L12)*, *(L13)*, *(L14)* and *(L15)* – they are derived from EU Statistics on Income and Living Conditions (EU-SILC).

The EU-SILC "Laeken" indicators, henceforth referred to as "the EU-SILC indicators", comprise both income-poverty and income-inequality measures. Together with the other indicators, they cover four important dimensions of social inclusion (financial poverty, employment, health and education), thus highlighting the "multidimensionality" of the phenomenon of social exclusion.

1.3 The EU-SILC indicators

In this section precise definitions of the EU-SILC "Laeken" indicators are given. After we remove the indicator "Self-defined health status by income level" *(L10)* which, in the absence of agreed methodology, is not being produced, there are nine of them. Those can be divided into so-called "Poverty" and "Inequality" measures:

(1) "Poverty" measures
   (1) The at-risk-of-poverty rate: *(L1)*, *(L11)*, *(L12)* and *(L13)*
   (2) The at-persistent-risk-of-poverty rate: *(L3)* and *(L15)*
   (3) The relative median at-risk-of-poverty gap: *(L4)*
(2) "Inequality" measures
   (1) The S80/S20 quintile share ratio: *(L2)*
   (2) The Gini coefficient: *(L14)*

1.3.1 The at-risk-of-poverty rate. This is the share of persons with an income below the so-called "at-risk-of poverty threshold". The latter is defined as a given percentage of the median income:

- For *(L1)*, *(L12)* and *(L13)*: 60% of the median income
- For *(L11)*: 40%, 50% and 70% of the median income

Either the income after social transfers (Indicators *L1*, *L11* and *L12*) or the income before social transfers (Indicator *L13* – at-risk-of-poverty rate before social transfers) is used in calculations. Besides, the at-risk-of-poverty threshold can be anchored at a given moment in time and then the poverty rate is calculated for any subsequent time using a fixed threshold (Indicator *L12* – at-risk-of-poverty rate anchored at a moment in time). Another approach (Indicators *L1*, *L11* and *L13*) consists of re-calculating the at-risk-of-poverty threshold for each year rather than using a fixed value.

The at-risk-of-poverty rate can also be broken down according to household or personal characteristics (e.g., age group, gender, NUTS2[3] geographical region or most frequent activity status). However, it must be kept in mind that the at-risk-of-poverty threshold which is used in breakdowns keeps being calculated over the total population, and not over the sub-population which is considered.

1.3.2 The at-persistent-risk-of-poverty rate. The at-persistent-risk-of-poverty rate is actually the core EU-SILC longitudinal indicator.[4] For a four-year panel, it is defined as the share of persons who are at-risk-of-poverty at the fourth wave of the panel and at two of the three preceding waves. The at-risk-of-poverty threshold is set at 60% of the median income.

1.3.3 The relative median at-risk-of-poverty gap. The relative median at-risk-of-poverty gap is the difference between the median income of persons below the at-risk-of-

[3] Nomenclature of Territorial Units for Statistics

[4] As a longitudinal indicator, the at-persistent-risk-of-poverty rate is not included in the present study, the paper rather focusing on cross-sectional variance calculation

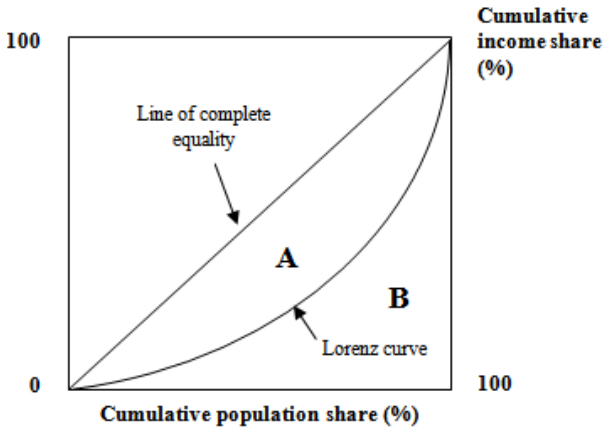


Figure 1. The Lorenz curve

poverty threshold and the at-risk-of-poverty threshold, expressed as a percentage of the latter. The at-risk-of-poverty threshold is set at 60% of the median income.

As in the at-risk-of-poverty rate, the median income of persons below the at-risk-of-poverty threshold can be broken down according to household or personal characteristics. However, the at-risk-of-poverty threshold keeps being calculated over the total population, and not over the subpopulation which is considered.

1.3.4 The S80/S20 income quintile share ratio. The income quintile share ratio (S80/S20) measures the degree of income inequality in a population. It is defined as the ratio of the total income of the persons in the top income quintile (i.e. 20% of the population with the highest income) over that of the persons in the bottom income quintile (i.e. 20% of the population with the lowest income).

1.3.5 The Gini coefficient. The Gini coefficient is another popular measure of income inequality. It ranges from 0, which reflects complete equality (i.e. all the persons receive the same income), and 1, which indicates complete inequality (i.e. one person has all the income, all others have none).

Let U be a population of size N and let $y = \{y_i, i \in U\}$ be an income distribution over the population U. Let r_i be the rank of i in the distribution y after we sort it in ascending income. The Gini coefficient G can be written as:

$$1 + G = \frac{2 \cdot \sum_{i \in U} (r_i - 1) \cdot y_i}{\left(\sum_{i \in U} 1\right) \cdot \left(\sum_{i \in U} y_i\right)} \quad (1)$$

The coefficient can also be easily represented by the area between the Lorenz curve and the line of complete equality.

The Lorenz curve maps the cumulative income share against the distribution of the population (see Figure 1).

The Gini coefficient is calculated as the area A between the Lorenz curve and the line of complete equality divided by the sum of areas A and B. If each individual had the same income (i.e. complete equality), the Lorenz curve and the line of total equality would be merged and the Gini coefficient is zero. On the other hand, if one individual received all the income (i.e. complete inequality), the Lorenz curve would pass through the points (0, 0); (100, 0) and (100, 100), and the surfaces A and B would be similar, leading to a value of one for the Gini coefficient.

1.4 Variance estimation for the EU-SILC indicators

Actually, all the EU-SILC indicators experience sampling errors in that they are derived from sample surveys rather than censuses. Thus, the results obtained for any single sample would be likely to vary slightly from the true values for the population. With the intent to describe the sample-to-sample variation of the main EU-SILC target indicators, Eurostat has developed a streamlined approach to assess their level of precision. It consists of systematically estimating their standard errors as well as confidence intervals for all the participating countries.

An important feature of the EU-SILC samples is that they are not actually selected with equal probability, which affects their precision and has to be taken into account in calculations. Actually, most of the EU-SILC samples have been stratified by geographical region (Nomenclature NUTS2[5]), which generally makes the accuracy better. Besides, many of them have been clustered by, for instance, so-called "Census Areas". Although clustering reduces data-collection costs, it also tends to decrease the precision of estimates because the population elements in a cluster are likely to be more similar (more homogeneous) to each other than elements of a simple random sample. Finally, the effects of variable sample weights on accuracy need also to be assessed. The Design Effect (*Deff*), that is, the ratio of the variance of a statistic with a complex sample design to the variance that would be obtained with a simple random sample of same size, is a valuable tool for measuring the combined effect of design components like stratification, clustering or unequal weights. *Deff* values for the EU-SILC indicators have been estimated in addition to standard errors and confidence intervals.

Re-sampling methods like Bootstrap or Jackknife were deliberately ruled out as their implementation happens to be time-consuming, especially with EU-SILC-like sampling designs involving several selection stages with unequal selection probabilities at each stage as well as weight adjustments (non-response correction, calibration to external data sources...). Besides, re-sampling methods are not easily reproducible at Eurostat level where nearly thirty countries are being dealt with.

Consequently, Eurostat opted for an "analytic" approach to variance estimation. The idea was to apply ad-hoc variance estimation formulae intended to reflect the main features of the sample design, including weight adjustments for non-response and calibration to external data sources. However, since the EU-SILC indicators are nonlinear statistics, they had to be "linearized" so to make variance calculations

[5] See footnote 4

tractable. The linearization technique for variance estimation consists of approximating a nonlinear statistic with a linear function of the observations. Then, an easily found variance estimator of the linear approximation is used as an estimator of the variance of the nonlinear statistic.

Finally, the SAS software Poulpe, developed by the French National Statistics Office (INSEE), was used to work out variance estimates for the "linearized" EU-SILC indicators. The present paper does not intend to describe the software and its statistical capabilities. For information, the reader can refer to Ardilly and Osier (2007).

2 The Taylor linearization method for variance estimation

2.1 Introduction

The Taylor linearization method (Tepping, 1968; Woodruff, 1971; Wolter, 1985; Tillé, 2000) yields an approximate estimator of the variance of a nonlinear statistic by using first-order Taylor Series expansions. The idea is to approximate a non-linear statistic with a linear function of estimated totals. Then, a variance estimator of the nonlinear statistic is given by a variance estimator of its linear approximation, which can be easily worked out.

Intuitively, the linearization approach rests on the assumption that the sample-to-sample variation of a non-linear statistic around its expected value is small enough to be considered linear. The latter assumption is particularly correct with large samples; although there is no definite evidence how large a sample should be for the linear approximation to be valid. Since most of the EU-SILC samples comprise thousands of individuals, there should be no problem in applying the linearization method to the EU-SILC data. On the other hand, one should be more careful when applying the method to smaller samples (e.g., for small domain estimation).

2.2. Linearization of a smooth function of population totals

Let θ denote a population parameter which can be expressed as a nonlinear function f of p population totals $Y_1, Y_2 \dots Y_p$:

$$\theta = f\left(Y_1, Y_2 \dots Y_p\right) \tag{2}$$

The function f is supposed to possess continuous derivatives up to order two. A natural way to estimate θ from a sample s of the population consists of plugging estimators $\hat{Y}_i$ of the totals Y_i into the expression (2):

$$\hat{\theta} = f\left(\hat{Y}_1, \hat{Y}_2 \dots \hat{Y}_p\right) \tag{3}$$

In most cases, the estimators $\hat{Y}_i$ are either "Π-expanded" Horvitz-Thompson estimators or generalized regression (GREG) estimators:

$$\hat{Y}_i = \sum_{k \in s} w_k(s) \cdot y_{ik} \tag{4}$$

where the sample weight $w_k(s)$ is either the design weight of k, that is, the reciprocal of the inclusion probability or the so-called g-weight (Särndal, Swensson and Wretman, 1992).

Let N be the size of the target population, and let us assume that $N^{-\alpha} f(.)$ tends to a limit for some $\alpha \geq 0$ (α is generally referred to as the "degree" of a statistic). For example, $\alpha = 1$ for a population total and $\alpha = 0$ for a ratio of totals.

The first step of the Taylor linearization method is to approximate the statistic $\hat{\theta}$ by a linear function of the totals $\hat{Y}_i$ by using a first-order Taylor series approximation of the function f around the point $(Y_1, Y_2 \dots Y_p)$:

$$\begin{aligned} & N^{-\alpha}\hat{\theta} \\ & = N^{-\alpha} f(\hat{Y}_1, \hat{Y}_2 \dots Y_p) \\ & = N^{-\alpha} f(Y_1, Y_2, \dots Y_p) \\ & + N^{-\alpha} \textstyle\sum_{i=1}^{p} [\frac{\partial f}{\partial y_i}(Y_1, Y_2, \dots Y_p)](\hat{Y}_i - Y_i) + R_n \\ & = N^{-\alpha}\theta + N^{-\alpha} \textstyle\sum_{i=1}^{p} c_i(\hat{Y}_i - Y_i) + R_n \end{aligned} \tag{5}$$

where for all i from 1 to p:

- $c_i = \frac{\partial f}{\partial y_i}(Y_1, Y_2, \dots Y_p)$ (6)
- The term R_n is a remainder term.

Under mild asymptotic assumptions (Tillé, 2000), the remainder term R_n is of order $1/n$ in probability, where n is the sample size, so R_n can be neglected as n grows.

Hence, the statistic $N^{-\alpha}\hat{\theta}$ can be written as follows:

$$N^{-\alpha}\hat{\theta} = N^{-\alpha}\left(\theta - \sum_{i=1}^{p} c_i Y_i\right) + N^{-\alpha} \sum_{i=1}^{p} c_i \hat{Y}_i + O_p\left(\frac{1}{n}\right) \tag{7}$$

Using the previous notations, we have:

$$\begin{aligned} & \textstyle\sum_{i=1}^{p} c_i \hat{Y}_i = \sum_{i=1}^{p} c_i \left[\sum_{k \in s} w_k(s)\, y_{ik}\right] \\ & = \textstyle\sum_{k \in s} w_k(s) \left[\sum_{i=1}^{p} c_i y_{ik}\right] = \\ & \textstyle\sum_{k \in s} w_k(s)\, z_k = \hat{Z} \end{aligned} \tag{8}$$

As a result, the statistic $N^{-\alpha}\hat{\theta}$ can be written as the sum of a constant term C, a linear function of the estimated totals $\hat{Y}_i$ and a remainder of order $1/n$ in probability:

$$N^{-\alpha}\hat{\theta} = C + N^{-\alpha}\hat{Z} + O_P\left(\frac{1}{n}\right) \tag{9}$$

The main result of the Taylor linearization method states that the variance of the non-linear statistic $\hat{\theta}$ can be approximated by the variance of the linear statistic $\hat{Z}$, in the sense that:

$$Var\left(N^{-\alpha}\hat{\theta}\right) = Var\left(N^{-\alpha}\hat{Z}\right) + O\left(\frac{1}{n^{3/2}}\right) \tag{10}$$

Following (8), we define the "linearized" variable of $\hat{\theta}$ at k as:

$$z_k = \sum_{i=1}^{p} c_i y_{ik} = \sum_{i=1}^{p} \left[\frac{\partial f}{\partial y_i}\left(Y_1, Y_2 \dots Y_p\right)\right] y_{ik} \tag{11}$$

Notwithstanding all this, one final hurdle remains in that the expression (11) of the "linearized" variable cannot always be implemented for it contains population quantities, namely the p totals $Y_1, Y_2 \dots Y_p$, the values of which are unknown when we observe only a sample of the population. In practice, estimators $\hat{Y}_i$ of the totals Y_i are just plugged into (11). Thus, the estimated "linearized" variable of $\hat{\theta}$ at k is given by:

$$\hat{z}_k = \sum_{i=1}^{p} \hat{c}_i y_{ik} = \sum_{i=1}^{p} \left[\frac{\partial f}{\partial y_i} \left(\hat{Y}_1, \hat{Y}_2 \dots \hat{Y}_p \right) \right] y_{ik} \qquad (12)$$

Although the above expression contains estimated quantities, those are treated as if they were exact, that is, their randomness is not taken into account in variance calculations. Actually, the error introduced can be considered negligible as long as the sample size is large enough (Deville, 1999).

2.3 Limitations of the Taylor linearization method

The Taylor linearization (TLM) method for variance estimation handles all the nonlinear statistics which can be expressed as a regular function (i.e., continuously differentiable up to order two) of estimated totals, for instance, ratios, correlation or regression coefficients. However, the TLM cannot deal with all nonlinear statistics and one has to accept that there are statistics for which the method cannot be used. Thus, in EU-SILC, we come across indicators which due to their complexity cannot be handled through the TLM. For instance, the Gini coefficient is based on rank statistics and the at-risk-of-poverty threshold is based on income quantiles (median). Regarding the at-risk-of-poverty rate, it is calculated on the basis of a poverty line which is estimated itself from sample observations, thus making the indicator more complex than a mere proportion. Therefore, variance estimation for the at-risk-of-poverty rate should take into account both the randomness which is brought by the at-risk-of-poverty threshold and that of the estimated proportion of "poor" persons given the poverty threshold. Besides, there is some degree of covariance between the at-risk-of-poverty threshold and the at-risk-of-poverty rate which should be accounted for.

3 The generalized linearization method for variance estimation

In order to deal with nonlinear statistics for which the Taylor method cannot be used, Deville (1999) presented a generalized linearization method based on the concept of influence function. The concept of influence function was first introduced in Robust Statistics (Hampel, Ronchetti, Rousseeuw and Stahel, 1986). In addition to encompassing more nonlinear statistics than the Taylor method, the linearization based on influence functions does not involve more calculations. In fact, as we'll see in this section, the derivation rules for influence functions are similar to the rules for computing the derivative of a function in standard differential calculus.

3.1 Definitions and notations

Let U denote a population of size N and let M be the measure which allocates a unit mass to each of the units i in U:

$$M(i) = M_i = 1 \qquad (13)$$

We seek to estimate a population parameter θ which can be expressed as a functional T of the measure M:

$$\theta = T(M) \qquad (14)$$

As a matter of fact, many parameters can be expressed in the form of (14), for instance:

- The population total Y of a variable y:
  $Y = \sum_{i \in U} y_i = \sum_{i \in U} y_i \times M(i) = \int y dM = T(M)$ (15)
- The ratio R of two population totals X and Y:
  $R = \frac{Y}{X} = \frac{\int y dM}{\int x dM} = T(M)$ (16)
- The cumulative distribution function: let $\{inc_i\}_i$ be an income distribution over the population U. The cumulative distribution function F at x is the share of population elements whose income is lower than x:
  $F(x) = \frac{\sum_{i \in U} 1(inc_i \leq x)}{N} = \frac{\int 1(inc \leq x) dM}{\int dM} = T(M)$ (17)
  where the function $1(inc_\bullet \leq x)$ is equal to 1 for all i whose inc_i is lower than x, and 0 otherwise.
- The at-risk-of-poverty threshold, that is, 60% of the median income:
  $ARPT = 0.6 \times MED(M) = T(M)$ (18)
  where the median income $MED(M)$ splits the income distribution into halves: $F[M, MED(M)] = 0.5$, where $F(M, .)$ designates the cumulative income distribution function (17).

A natural way to estimate (14) from a sample s of the population consists of plugging an estimated measure $\hat{M}$ of M into (14):

$$\hat{\theta} = T\left(\hat{M}\right) \qquad (19)$$

The estimated measure $\hat{M}$ allocates the sample weight $w_i(s)$ for all units i in s, and 0 otherwise:

$$\hat{M}(i) = \hat{M}_i = \begin{cases} w_i(s) \, for \, i \in s \\ 0 \, for \, i \notin s \end{cases} \qquad (20)$$

For instance, if we consider the population total (15) of a variable y, the so-called "plug-in" estimator (19) can be written as:

$$\hat{Y} = T\left(\hat{M}\right) = \int y d\hat{M} = \sum_{i \in s} w_i(s) \, y_i \qquad (21)$$

Likewise, if we consider the ratio (16) of two population totals X and Y, we got:

$$\hat{R} = T\left(\hat{M}\right) = \frac{\int y d\hat{M}}{\int x d\hat{M}} = \frac{\sum_{i \in s} w_i(s) \, y_i}{\sum_{i \in s} w_i(s) \, x_i} \qquad (22)$$

When it comes to the cumulative income distribution function (17), we obtain:

$$\hat{F}(x) = T\left(\hat{M}\right) = \frac{\int 1(inc \leq x)\, d\hat{M}}{\int d\hat{M}} = \frac{\sum\limits_{i \in s} w_i(s)\, 1(inc_i \leq x)}{\sum\limits_{i \in s} w_i(s)} \tag{23}$$

Finally, regarding the at-risk-of-poverty threshold (18), we got the following estimator:

$$\hat{ARPT} = T\left(\hat{M}\right) = 0.6 \times MED\left(\hat{M}\right) \tag{24}$$

where the estimated median income $MED\left(\hat{M}\right)$ satisfies $F\left[\hat{M}, MED\left(\hat{M}\right)\right] = 0.5$, where:

$$F\left(\hat{M}, x\right) = \frac{\sum\limits_{i \in s} w_i(s) \times 1(inc_i \leq x)}{\sum\limits_{i \in s} w_i(s)}$$

3.2 Main results

The linearization technique relies on asymptotic assumptions which hold provided the sample size is large enough. Deville (1999) described the asymptotic framework within which the linearization method works. As in Isaki and Fuller (1982), a sequence of populations (U_σ) of increasing size (N_σ) is considered. Let (s_σ) be a sequence of samples of increasing size (n_σ) which are selected from the populations (U_σ) with a probability sampling design $p_\sigma(.)$ Let X be a vector of population totals estimated by $\hat{X}$. To simplify the notations, the subscript σ is dropped. Then, the three following assumptions are postulated:

(i) $N^{-1}X$ has a limit as σ tends to infinity

(ii) The sequence $N^{-1}\left(\hat{X} - X\right)$ converges in probability to 0

(iii) The sequence $n^{1/2}N^{-1}\left(\hat{X} - X\right)$ converges to a multidimensional normal distribution

Within this framework, the main result of this generalized linearization theory can be stated as follows: under the same definitions and notations as in the previous section, the variance of the so-called "plug-in" estimator (19) can be approximated with that of a linear statistic:

$$Var\left[T\left(\hat{M}\right)\right] \cong Var\left[\sum_{i \in s} w_i(s)\, z_i\right] \tag{25}$$

where the "linearized" variable z at k is given by the following functional derivative:

$$z_k = Lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t} = IT_k(M) \tag{26}$$

δ_k is the Dirac measure at k : $\delta_k(i) = 1$ if and only if $i = k$.

The functional derivative (26) is called the influence function. The concept of influence function, which was first introduced in the field of Robust Statistics (Hampel, Ronchetti, Rousseeuw and Stahel, 1986), aims to grasp the effect of an infinitesimal contamination in the observations.

The expression (26) cannot be calculated when we observe only a sample of the population because it involves unknown population quantities. In practice, the sample measure $\hat{M}$ as defined in (20) is plugged into (26). Thus, the estimated "linearized" variable at k is given by:

$$\hat{z}_k = IT_k\left(\hat{M}\right) = Lim_{t \to 0} \frac{T\left(\hat{M} + t\delta_k\right) - T\left(\hat{M}\right)}{t} \tag{27}$$

Despite the expression of the "linearized" variable contains estimated quantities, those are treated as if they were exact, that is, their randomness is not taken into account in variance calculations. The error introduced by this approximation can be considered negligible as long as the sample size is large enough (Deville, 1999).

3.3 Derivation rules for influence functions

In many cases, complicated limit calculations by direct application of (26) can be avoided using derivation rules for influence functions. Some of the most basic rules are given next. See Deville (1999) for more details. Actually, the derivation rules for influence functions are similar to the rules for computing the derivative of a function, which makes it easy to handle.

- Constant:
  Let T be a constant functional of the measure M, that is $T(M) = c$. The influence function of T at k is equal to 0:
  $$I(c)_k(M) = 0 \tag{28}$$
- Linear combination of two functionals:
  Let T and S be two functionals of M and let a and b be two constants. The influence function of the linear combination $a \cdot T + b \cdot S$ is given by:
  $$I(a \cdot T + b \cdot S)_k(M) = a \cdot IT_k(M) + b \cdot IS_k(M) \tag{29}$$
- Product of two functionals:
  Let T and S be two functionals of M. The influence function of the product of T and S is given by:
  $$I(T \times S)_k(M) = T(M) \times IS_k(M) + S(M) \times IT_k(M) \tag{30}$$
- Ratio of two functionals:
  Let T and S be two functionals of M. The influence function of the ratio T/S is given by:
  $$I\left(\frac{T}{S}\right)_k(M) = \frac{S(M) \times IT_k(M) - T(M) \times IS_k(M)}{S(M)^2} \tag{31}$$
- Composition of two functionals:
  Let T and S be two functionals of M. The influence function of the composition of T and S, that is, the functional $M \to T[S(M)]$ is given by:
  $$I[T(S)]_k(M) = IT_k[S(M)] \times IS_k(M) \tag{32}$$

- Functional with a parameter:
  Consider the functional $M \rightarrow T[M, S(M)]$, where $S(M)$ represents a scalar parameter of T that is calculated from the observations. For instance, the at-risk-of-poverty rate $ARPR$ can be written in this way, for we have: $ARPR(M) = F[M, ARPT(M)]$, where $ARPT(M)$ is the at-risk-of-poverty threshold (60% of the median income) and $F(M,.)$ designates the cumulative income distribution function (17). The influence function can be expressed as a sum of two terms: the first one is the influence function of T with respect to M holding the parameter $S(M)$ constant, while the other one accounts for the influence function of $S(M)$:

$$\begin{aligned} & I[T(M, S(M))]_k \\ = & I[T(M, S(M) | S(M) \ fixed]_k + \\ & \frac{dT(M,x)}{dx}|_{x=S(M)} \times IS_k(M) \end{aligned} \tag{33}$$

3.4 Examples

3.4.1 Example. Influence function of a population total

Consider the population total Y as defined in (15):

$$Y = \sum_{i \in U} y_i = \int y \cdot dM = T(M)$$

The influence function of T at k is given by:

$$\begin{aligned} & IT_k(M) \\ = & Lim_{t \rightarrow O} \frac{T(M + t\delta_k) - T(M)}{t} \\ = & Lim_{t \rightarrow O} \frac{\int y \cdot d(M + t\delta_k) - \int y \cdot dM}{t} \\ = & Lim_{t \rightarrow O} \frac{\int y \cdot d(t\delta_k)}{t} = y_k \end{aligned} \tag{34}$$

3.4.2 Example. Influence function of a ratio of two population totals

Now, let R be the ratio of two population totals X and Y:

$$R = \frac{Y}{X} = \frac{\int y \cdot dM}{\int x \cdot dM} = \frac{U(M)}{V(M)} = T(M)$$

By using the derivation rule (31) for a ratio of two functionals, we obtain:

$$IR_k(M) = I\left(\frac{U}{V}\right)_k(M) = \frac{V(M) \times IU_k(M) - U(M) \times IV_k(M)}{V(M)^2}$$

According to (34), the influence functions of U and V are given by:

$$\begin{cases} IU_k(M) = y_k \\ IV_k(M) = x_k \end{cases}$$

Consequently, the influence function of the ratio R is:

$$IR_k(M) = \frac{X \times y_k - Y \times x_k}{X^2} = \frac{1}{X}(y_k - R \cdot x_k) \tag{35}$$

The values of the total X and the ratio R are unknown when a sample of the population is observed, so they are replaced by estimated values $\hat{X}$ and $\hat{R}$. Finally, the estimated influence function is:

$$I\hat{R}_k(M) = \frac{1}{\hat{X}}\left(y_k - \hat{R}x_k\right) \tag{36}$$

The cumulative distribution function at x as defined in (17) is a particular case of ratio, therefore, its influence function is:

$$IF(x)_k(M) = \frac{1}{N}[1(inc_k \leq x) - F(x)] \tag{37}$$

4 Application of the generalized linearization method to the EU-SILC indicators

In addition to being easy to implement, another major advantage of the linearization based on influence functions is that it can deal with statistics for which the Taylor linearization cannot be used. In particular, the EU-SILC indicators fall into this category (see section 2.3). Actually, the minimum requirement for a statistical functional is to be Hadamard differentiable (Van der Vaart, 1998). The concept of Hadamard differentiability is in fact less stringent than the more restrictive Fréchet derivatives, as used in the Taylor linearization method: though they are not Fréchet differentiable, many well-known statistical functionals (e.g., quantiles) are Hadamard differentiable (Reeds, 1976; Van der Vaart, 1998).

4.1 Linearization of the at-risk-of-poverty rate

This section presents the expression of the influence function of the at-risk-of-poverty rate, which is actually the core EU-SILC indicator. The step-by-step calculations are presented in Appendix 1, as well as the influence functions of the other EU-SILC indicators[6].

With the same notations as in the section 3.1, the influence function at k of the at-risk-of-poverty rate ($ARPR$) is given by:

$$\begin{aligned} IARPR_k(M) = & \frac{1}{N}[1(inc_k \leq ARPT(M)) - ARPR(M)] - \\ & \frac{0.6}{N} \times \frac{\tilde{F}'[ARPT(M)]}{\tilde{F}'[MED(M)]} \times [1(inc_k \leq MED(M)) - 0.5] \end{aligned} \tag{38}$$

[6] With the exception of the influence function of the at-persistent-risk-of-poverty rate (see footnote 4)

where $\tilde{F}'[ARPT(M)]$ and $\tilde{F}'[MED(M)]$ are the values of the derivative of the cumulative income distribution function $\tilde{F}$ at the points $ARPT(M)$ (at-risk-of-poverty threshold) and $MED(M)$ (median income), respectively. These two quantities can be interpreted as the "income densities" at $ARPT(M)$ and $MED(M)$.

The influence function (38) can be regarded as a sum of two terms: the first term is the influence function that would be obtained assuming the at-risk-of-poverty threshold $ARPT(M)$ is constant, while the second term is a "correction" which takes into account the fact that the at-risk-of-poverty threshold is estimated from sample observations.

An important issue is that the derivative of an empirical cumulative distribution function is always 0 or not defined. Let approximate $\tilde{F}$ by the following convolution product:

$$\tilde{F}_K(x) = \int \tilde{F}(t) \cdot K(x,t) \cdot dt \tag{39}$$

where the two variable function $K(.,.)$ is a Gaussian kernel:

$$K(x,t) = \frac{1}{h\sqrt{2\pi}} exp\left[-\frac{(x-t)^2}{2h^2}\right] \tag{40}$$

It can be easily seen that the function $\tilde{F}_K$ is differentiable, and we have for all x:

$$\tilde{F}'_K(x) = \frac{1}{h\sqrt{2\pi}} \cdot \frac{1}{N} \cdot \sum_{k \in U} exp\left[-\frac{(x-x_k)^2}{2h^2}\right] \tag{41}$$

The derivative $\tilde{F}'_K$ exists and is strictly non-negative. The smoothing parameter h can be estimated using the "plug-in" estimator (Silverman, 1986):

$$\hat{h} = \frac{\hat{\sigma}}{N^{-1/5}} \tag{42}$$

where $\hat{\sigma}$ is the estimated standard deviation of the income distribution.

It is worth mentioning however that the estimated income density function happens to be quite sensitive to the choice of the bandwith parameter h (Verma and Betti, 2005).

Finally, the influence function of the at-risk-of-poverty rate at k is:

$$\begin{aligned} IARPR_k(M) &= \frac{1}{N}\left[1(inc_k \le ARPT(M)) - ARPR(M)\right] \\ &- \frac{0.6}{N} \times \frac{\tilde{F}'_K[ARPT(M)]}{\tilde{F}'_K[MED(M)]} \times \left[1(inc_k \le MED(M)) - 0.5\right] \end{aligned} \tag{43}$$

4.2 Estimation over subpopulations

The EU-SILC indicators happen to be broken down by household or individual characteristics. For instance, one may want to know the at-risk-of-poverty rate not only for the total population, but also for the following age groups: 15-24, 25-34, 35-44, 45-54, 55-64 and 65+ year-olds. Other domains of study are also considered, for example, the subpopulations defined by NUTS2 region, household tenure status (owner/tenant) or dwelling type (house/apartment).

Domain estimation is actually a long-established theory (Särndal, Swensson and Wretman, 1992). By introducing the domain membership indicator variable, which is equal to 1 for all the units in the domain and 0 otherwise, no extra difficulty should be expected with calculations. For example, let consider the at-risk-of-poverty rate $ARPR_C$ for a given subpopulation C of size N_C. Let 1_C be the variable which is equal to 1 for all the units in C and 0 otherwise:

$$ARPR_C = G[M, ARPT(M)] = T(M) \tag{44}$$

where:

- The at-risk-of-poverty threshold $ARPT(M)$ is defined as 60% of the median income $MED(M)$ of the total population
- G is the cumulative income distribution function over the subpopulation C:

$$G(M,x) = \frac{1}{N_C}\sum_{i \in C} 1(inc_i \le x) = \frac{\int 1_C \cdot 1(inc \le x)\, dM}{\int 1_C \cdot dM} \tag{45}$$

The influence function of (44) can be written as:

$$\begin{aligned} &IARPR_{C,k}(M) \\ &= \tfrac{1_C(k)}{N_C} \times \left[1(inc_k \le ARPT(M)) - ARPR_C(M)\right] \\ &- \tfrac{0.6}{N} \times \tfrac{\tilde{G}'_K[ARPT(M)]}{\tilde{G}'_K[MED(M)]} \times \left[1(inc_k \le MED(M)) - 0.5\right] \end{aligned} \tag{46}$$

The proof of (46) is actually the same as that of the at-risk-of-poverty rate for the total population (see Appendix 1, section 2.2).

Another key indicator is the relative median at-risk-of-poverty gap, for which breakdowns by age group and gender are generally wanted. Let consider the relative median poverty gap $RMPG_C$ for a subpopulation C, that is, the relative difference between the at-risk-of-poverty threshold $ARPT$, that is, 60% of the median income of the total population, and the median income MED_C^p over the persons in C whose income is lower than $ARPT$:

$$RMPG_C = \frac{ARPT - MED_C^p}{ARPT} = 1 - \frac{MED_C^p}{ARPT} \tag{47}$$

The median income MED_C^p satisfies the following identity:

$$G\left(M, MED_C^p(M)\right) = \frac{1}{2} G[M, ARPT(M)] \tag{48}$$

where G is the cumulative income distribution function (45) over the subpopulation C. For the derivation of the influence function of (47), see Appendix 1.

4.3 Numerical results (EU-SILC 2004)

The next Tables contain estimated standard errors of the main EU-SILC indicators for five countries: Austria (AT),

Denmark (DK), Estonia (EE), Finland (FI) and Italy (IT). In addition, the following information is provided for each indicator:

- The estimated population value
- The achieved sample size
- The estimated lower and upper bounds of a 95% confidence interval: the linearized estimator $\hat{\theta}$ of θ is assumed to follow a normal distribution, thereby an estimated 95% confidence interval for θ is given by:

$$C\hat{I}(\theta) = \left[\hat{\theta} - 1.96 \times \sqrt{V\hat{a}r\left(\hat{\theta}\right)}, \hat{\theta} + 1.96 \times \sqrt{V\hat{a}r\left(\hat{\theta}\right)}\right] \quad (49)$$

- The estimated design effect ($Deff$): this is the ratio of the variance of the statistic with the actual sample design to the variance of that statistic with a simple random sample (SRS) of same size:

$$D\hat{e}ff = \frac{V\hat{a}r\left(\hat{\theta}\right)}{V\hat{a}r_{SRS}\left(\hat{\theta}_{SRS}\right)} \quad (50)$$

  The design effect measures the impact of design on sampling variability: it indicates how much precision have been lost by using a complex survey rather than a simple random sample.

The five countries which have been chosen (Austria, Denmark, Estonia, Finland and Italy) provide a "representative" sample of all the sampling designs implemented in the EU-SILC countries:

- Denmark and Finland collect income information from registers while Austria, Estonia and Italy collect this information using face-to-face interview.
- There are also differences between the three latter countries regarding the way they selected their samples: Austria performed a direct-element selection of addresses, whereas Italy selected a sample of addresses with a multistage sampling scheme. As to Estonia, a sample of households was selected using an "indirect" approach: they began by selecting a sample of persons aged 14 and over. Then, all the households the sampled persons belong to were eligible for inclusion in the sample.
- The sample sizes vary greatly from one country to another: Austria, Denmark and Estonia achieved sample sizes of 11550, 17290 and 11558 individuals, respectively, whereas Finland and Italy attained relatively larger samples (29070 and 61429 individuals, respectively).
- There are also important differences between the countries as regards the nature of the auxiliary information that was used to adjust the sample weights: Austria, Estonia and Italy adjusted their weights to demographic counts from census data (e.g., population totals broken down by age group and gender), whereas Denmark and Finland used updated register information. That way, the gain in accuracy should be important (see section 5.2)

The Tables 2 to 6 contain for these five sample countries the estimated variance figures which were obtained from the generalized linearization approach. For more numerical results relating to the EU-SILC 2004 operation, see Osier and Museux (2006).

5 Discussion

An interesting approach to validate the variance figures as presented in the Tables 2 to 6 is to examine the effect of design components which are known to play a key role in explaining the precision in estimates. This section will focus on the two following factors:

- The achieved sample size
- The weight adjustments to external data sources

5.1 Effect of the achieved sample size

Obviously, the level of sampling error depends on the achieved sample size: the higher the achieved sample size, the lower the level of sampling error. The Figure 2 presents for each of the five EU-SILC sample countries (Austria, Estonia, Denmark, Finland and Italy) the coefficients of variation (CV) of the following indicators:

- The at-risk-of-poverty rate (ARPR)
- The income quintile share ratio (S80/S20)
- The relative median poverty gap (RMPG)
- The Gini coefficient (Gini)

With the intent to show correlation between the coefficient of variation and the achieved sample size, the countries have been sorted in ascending sample size.

Looking at the Figure 2, the relative precision of estimates tends to be correlated with the sample size. If we consider the at-risk-of-poverty rate, the relative precision ranges from 4.6% in Austria and 3.2% in Estonia, which used samples of 11550 and 11558 individuals, respectively, to 1.6% in Italy, which used up to 6 times more individuals (61429). The values we obtained for Denmark (CV=0.5%) and Finland (CV=3.4%), although they both break the decreasing trend between the countries, can be explained by other factors than the sample size. The result for Denmark (0.5%) is caused by using auxiliary data on poverty to adjust the sample weights. When sample weights are adjusted to auxiliary variables which are correlated with the survey target variables, one can expect the accuracy to be better (Deville and Särndal, 1992). The Finnish case stems from the sample design itself in that it leads to over-representing upper income classes in the EU-SILC sample. This deteriorates the accuracy of the at-risk-of-poverty rate because upper income classes carry relatively less "information" about poverty than lower income classes. On the other hand, this feature of the Finnish EU-SILC design should improve the relative precision of the S80/S20 and Gini indicators (1.0% and 0.8%, respectively); which both measure income inequality rather than income poverty.

If we consider now the three other indicators, we can draw similar conclusions as in the previous case. As regards the

Table 2: Estimated standard errors SILC 2004 – Austria

Indicator	Value	Achieved sample size	Standard error	Confidence interval 95% Lower	Confidence interval 95% Upper	Standard error (%)	Deff
At-risk-of-poverty rate – total	12.8	11550	0.59	11.6	14.0	4.6	1.0
At-risk-of-poverty rate – male	11.3	5575	0.63	10.1	12.5	5.6	1.0
At-risk-of-poverty rate – female	14.2	5975	0.66	12.9	15.5	4.6	1.0
At-risk-of-poverty rate - 16-24 years	12.8	1279	1.24	10.4	15.2	9.7	1.0
At-risk-of-poverty rate - 25-49 years	11.2	4185	0.66	9.9	12.5	5.9	1.0
At-risk-of-poverty rate - 50-64 years	10.3	2188	0.88	8.6	12.0	8.5	1.0
At-risk-of-poverty rate - 65+ years	17.1	1611	1.16	14.8	19.4	6.8	1.0
At-risk-of-poverty rate - 16+ years	12.3	9263	0.52	11.3	13.3	4.2	1.0
At-risk-of-poverty rate - 16-64 years	11.2	7652	0.55	10.1	12.3	5.0	1.0
At-risk-of-poverty rate - 0-64 years	12.0	9913	0.64	10.7	13.3	5.3	1.0
At-risk-of-poverty threshold	10181.7	11550	86.39	10012.4	10351.0	0.8	1.0
S80/S20 income quintile share ratio	3.8	11550	0.08	3.6	3.9	2.2	1.0
Relative median poverty gap – total	20.0	1438	1.37	17.3	22.7	6.9	1.0
Relative median poverty gap – male	18.6	622	1.61	15.4	21.8	8.7	1.0
Relative median poverty gap – female	20.4	816	1.38	17.7	23.1	6.8	1.0
Relative median poverty gap - 16-64 years	20.4	832	1.62	17.2	23.6	7.9	1.0
Relative median poverty gap - 65+ years	20.6	268	1.79	17.1	24.1	8.7	1.0
Relative median poverty gap - 16+ years	20.6	1100	1.32	18.0	23.2	6.4	1.0
Gini coefficient	25.8	11550	0.44	24.9	26.6	1.7	1.0

Table 3: Estimated standard errors SILC 2004 – Denmark

Indicator	Value	Achieved sample size	Standard error	Confidence interval 95% Lower	Confidence interval 95% Upper	Standard error (%)	Deff
At-risk-of-poverty rate – total	11.0	17290	0.05	10.9	11.1	0.5	0.84
At-risk-of-poverty rate – male	10.7	8684	0.04	10.6	10.8	0.4	0.85
At-risk-of-poverty rate – female	11.2	8606	0.06	11.1	11.3	0.6	0.83
At-risk-of-poverty rate - 16-24 years	27.0	1872	0.07	26.9	27.1	0.3	0.84
At-risk-of-poverty rate - 25-49 years	8.8	6252	0.03	8.8	8.8	0.3	0.83
At-risk-of-poverty rate - 50-64 years	4.2	3612	0.09	4.0	4.4	2.1	0.84
At-risk-of-poverty rate - 65+ years	17.0	1848	0.19	16.6	17.4	1.1	0.81
At-risk-of-poverty rate - 16+ years	11.4	13584	0.06	11.3	11.5	0.5	0.83
At-risk-of-poverty rate - 16-64 years	10.2	11736	0.04	10.1	10.3	0.4	0.83
At-risk-of-poverty rate - 0-64 years	9.9	15442	0.04	9.8	10.0	0.4	0.85
At-risk-of-poverty threshold	12736.0	17290	13.07	12710.4	12761.6	0.1	0.85
S80/S20 income quintile share ratio	3.4	17290	0.06	3.3	3.5	1.8	0.94
Relative median poverty gap – total	19.0	982	0.77	17.5	20.5	4.1	0.84
Relative median poverty gap – male	21.8	461	0.99	19.9	23.7	4.5	0.85
Relative median poverty gap – female	18.1	521	0.84	16.4	19.8	4.7	0.83
Relative median poverty gap - 16-64 years	24.4	557	0.86	22.7	26.1	3.5	0.83
Relative median poverty gap - 65+ years	7.8	236	0.57	6.7	8.9	7.3	0.79
Relative median poverty gap - 16+ years	19.0	793	0.73	17.6	20.4	3.8	0.81
Gini coefficient	23.9	17290	0.41	23.1	24.7	1.7	0.96

Table 4: Estimated standard errors SILC 2004 – Estonia

Indicator	Value	Achieved sample size	Standard error	Confidence interval 95% Lower	Confidence interval 95% Upper	Standard error (%)	Deff
At-risk-of-poverty rate – total	20.2	11558	0.64	18.9	21.5	3.2	1.10
At-risk-of-poverty rate – male	19.5	5446	0.70	18.1	20.9	3.6	1.06
At-risk-of-poverty rate – female	20.8	6112	0.81	19.2	22.4	3.9	1.12
At-risk-of-poverty rate - 16-24 years	21.1	1949	1.19	18.8	23.4	5.6	1.10
At-risk-of-poverty rate - 25-49 years	18.8	3772	0.72	17.4	20.2	3.9	1.10
At-risk-of-poverty rate - 50-64 years	19.0	2054	1.29	16.5	21.5	6.8	1.08
At-risk-of-poverty rate - 65+ years	20.5	1565	1.84	16.9	24.1	9.0	1.13
At-risk-of-poverty rate - 16+ years	19.6	9340	0.65	18.3	20.9	3.3	1.12
At-risk-of-poverty rate - 16-64 years	19.3	7775	0.62	18.1	20.5	3.2	1.10
At-risk-of-poverty rate - 0-64 years	20.1	9951	0.65	18.8	21.4	3.2	1.06
At-risk-of-poverty threshold	1539.0	11558	18.96	1501.8	1576.2	1.2	1.08
S80/S20 income quintile share ratio	7.1	11558	0.23	6.6	7.6	3.3	1.15
Relative median poverty gap – total	26.3	2373	1.42	23.5	29.1	5.4	1.07
Relative median poverty gap – male	29.0	1096	1.79	25.5	32.5	6.2	1.06
Relative median poverty gap – female	22.5	1277	1.40	19.7	25.3	6.2	1.09
Relative median poverty gap - 16-64 years	30.2	1591	1.65	27.0	33.4	5.5	1.12
Relative median poverty gap - 65+ years	9.3	244	1.14	7.1	11.5	12.3	1.09
Relative median poverty gap - 16+ years	24.6	1835	1.30	22.1	27.1	5.3	1.10
Gini coefficient	37.4	11558	0.58	36.3	38.5	1.5	1.26

Table 5: Estimated standard errors SILC 2004 – Finland

Indicator	Value	Achieved sample size	Standard error	Confidence interval 95% Lower	Confidence interval 95% Upper	Standard error (%)	Deff
At-risk-of-poverty rate – total	11.0	29070	0.38	10.3	11.7	3.4	1.40
At-risk-of-poverty rate – male	10.5	14736	0.41	9.7	11.3	3.9	1.41
At-risk-of-poverty rate – female	11.4	14334	0.46	10.5	12.3	4.0	1.42
At-risk-of-poverty rate - 16-24 years	19.5	3524	1.08	17.4	21.6	5.5	1.49
At-risk-of-poverty rate - 25-49 years	8.3	9489	0.41	7.5	9.1	4.9	1.37
At-risk-of-poverty rate - 50-64 years	7.8	6769	0.48	6.9	8.7	6.2	1.23
At-risk-of-poverty rate - 65+ years	16.6	2972	1.02	14.6	18.6	6.2	1.57
At-risk-of-poverty rate - 16+ years	11.3	22754	0.36	10.6	12.0	3.2	1.44
At-risk-of-poverty rate - 16-64 years	10.1	19782	0.34	9.4	10.8	3.4	1.37
At-risk-of-poverty rate - 0-64 years	9.9	26098	0.38	9.2	10.6	3.8	1.36
At-risk-of-poverty threshold	9984.0	29070	43.00	9899.7	10068.3	0.4	1.14
S80/S20 income quintile share ratio	3.5	29070	0.03	3.4	3.6	1.0	0.82
Relative median poverty gap – total	14.1	2746	0.68	12.8	15.4	4.8	1.42
Relative median poverty gap – male	14.9	1352	0.88	13.2	16.6	5.9	1.42
Relative median poverty gap - female	13.7	1394	0.76	12.2	15.2	5.6	1.44
Relative median poverty gap - 16-64 years	16.3	1793	0.84	14.7	17.9	5.2	1.38
Relative median poverty gap - 65+ years	9.4	323	0.91	7.6	11.2	9.7	1.77
Relative median poverty gap - 16+ years	14.1	2116	0.65	12.8	15.4	4.6	1.49
Gini coefficient	25.4	29070	0.20	25.0	25.8	0.8	0.78

Table 6: Estimated standard errors SILC 2004 – Italy

Indicator	Value	Achieved sample size	Standard error	Confidence interval 95% Lower	Confidence interval 95% Upper	Standard error (%)	Deff
At-risk-of-poverty rate – total	19.0	61429	0.30	18.4	19.6	1.6	1.50
At-risk-of-poverty rate – male	17.7	29757	0.34	17.0	18.4	1.9	1.67
At-risk-of-poverty rate – female	20.3	31672	0.32	19.7	20.9	1.6	1.31
At-risk-of-poverty rate - 16-24 years	22.9	5886	0.75	21.4	24.4	3.3	1.42
At-risk-of-poverty rate - 25-49 years	16.8	22679	0.35	16.1	17.5	2.1	1.62
At-risk-of-poverty rate - 50-64 years	14.7	11926	0.42	13.9	15.5	2.9	1.17
At-risk-of-poverty rate - 65+ years	21.1	11420	0.59	19.9	22.3	2.8	1.56
At-risk-of-poverty rate - 16+ years	18.0	51911	0.29	17.4	18.6	1.6	1.47
At-risk-of-poverty rate - 16-64 years	17.1	40491	0.32	16.5	17.7	1.9	1.53
At-risk-of-poverty rate - 0-64 years	18.5	49602	0.34	17.8	19.2	1.8	1.59
At-risk-of-poverty threshold	8129.0	61429	43.67	8043.4	8214.6	0.5	1.52
S80/S20 income quintile share ratio	5.7	61429	0.09	5.5	5.9	1.6	1.51
Relative median poverty gap – total	24.6	10125	0.68	23.3	25.9	2.8	1.42
Relative median poverty gap – male	24.9	4509	0.78	23.4	26.4	3.1	1.33
Relative median poverty gap – female	24.3	5616	0.71	22.9	25.7	2.9	1.51
Relative median poverty gap - 16-64 years	28.5	5910	0.81	26.9	30.1	2.8	1.42
Relative median poverty gap - 65+ years	15.7	2276	0.52	14.7	16.7	3.3	1.49
Relative median poverty gap - 16+ years	23.7	8186	0.63	22.5	24.9	2.7	1.38
Gini coefficient	33.1	61429	0.31	32.5	33.7	0.9	1.43

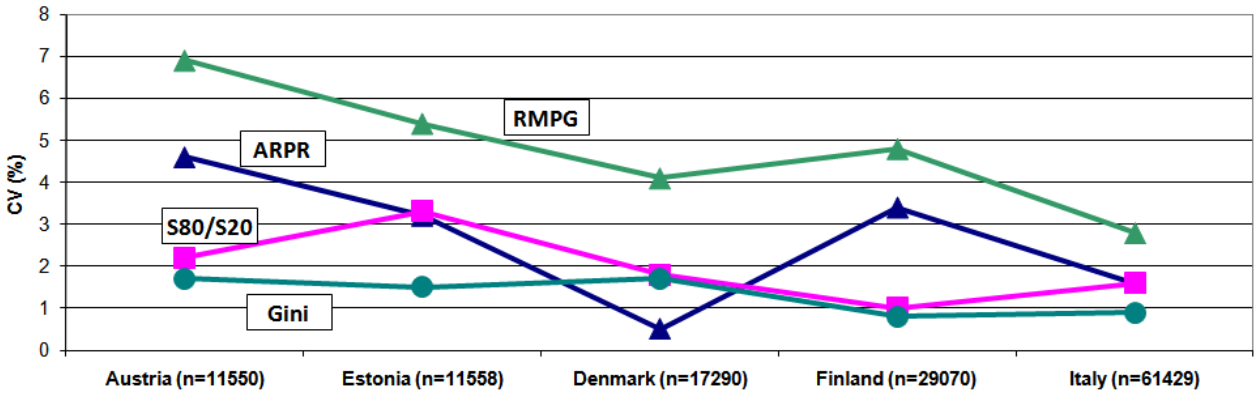


Figure 2. Estimated CVs and achieved sample sizes – EU-SILC 2004

relative median poverty gap (RMPG), the line is nearly parallel to the one we got for the at-risk-of-poverty rate. The breaks we observed for Denmark and Finland can be explained by the same reasons as previously. Concerning the S80/S20 and Gini indicators, we got nearly a decreasing trend among the countries, with only a break for the Gini coefficient in Estonia.

Finally, it should be noted that the achieved sample size is likely to explain most of the loss of efficiency for estimates over subpopulations, for instance, the at-risk-of-poverty rate broken down by age group and/or gender (see Tables 2 to 6).

5.2 Effect of weight adjustments to external data sources

Actually, most of the EU-SILC countries calibrated their sample weights to external data sources, thereby hoping to make their data more accurate. The so-called calibration technique (Deville and Särndal, 1992; Deville, Särndal and Sautory, 1993) consists of adjusting the sample weights in order to match population totals coming from external data sources. Thus, one can expect better accuracy in estimates, insofar as the survey target variables are correlated with the auxiliary variables the weights are calibrated to (Deville and Särndal, 1992). Actually, an interesting ability of the vari-

ance estimation software Poulpe is that it can estimate for a given statistic the standard error that would be achieved assuming the weights were not calibrated (Osier, 2003). In particular, this makes possible measuring the effect of calibration on the accuracy by computing standard errors before and after adjusting the weights. The Figure 3 presents the coefficients of variation (CV) of the following indicators before and after weight calibration:

- The at-risk-of-poverty threshold
- The at-risk-of-poverty rate
- The relative median poverty gap
- The income quintile share ratio
- The Gini coefficient

The impact of calibration in Austria, Estonia and Italy appears to be rather limited in that the coefficients of variation drop by around 0.1 percentage points. In fact, these countries calibrated their samples to demographic counts (e.g., population totals by age group and gender) coming from the last census data available or from another existing survey. Because demographic variables are not generally strongly correlated with the EU-SILC variables on income and poverty, one cannot expect the accuracy to be significantly better by using those data as calibration information.

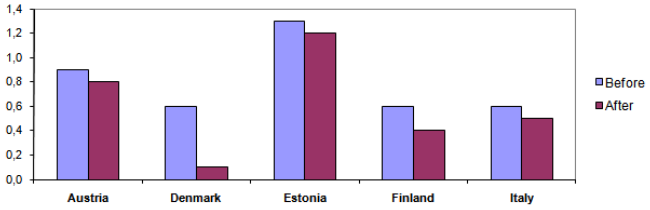
On the other hand, Finland calibrated their sample to population totals for certain income components, for example, work or pension income. This information is taken out from registers and, consequently, should be updated. As the Figure 3 shows, while calibration nearly halved the relative precision of the so-called "inequality" measures (S80/S20 and Gini – see §1.3), it had a rather limited impact on the "poverty" measures (at-risk-of-poverty threshold, at-risk-of-poverty rate and relative median poverty gap).

Finally, the situation in Denmark is somewhat at the opposite of the Finnish situation in that updated register information about poverty was used as calibration information (see section 5.1) As expected, whereas calibration made the poverty measures almost error-free, it had a relatively smaller impact on the accuracy of the inequality measures (S80/S20 and Gini).

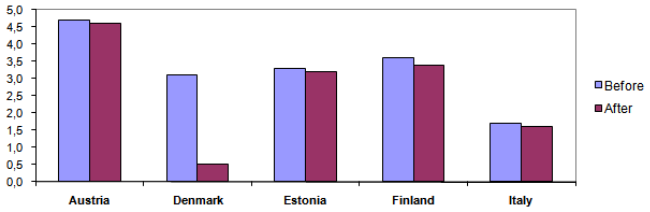
6 Conclusion

The linearization method based on influence functions coupled with an efficient software package for variance estimation, like Poulpe, has proved to be an easy and powerful solution to estimate the precision of complex non-linear statistics like the "Laeken" indicators:

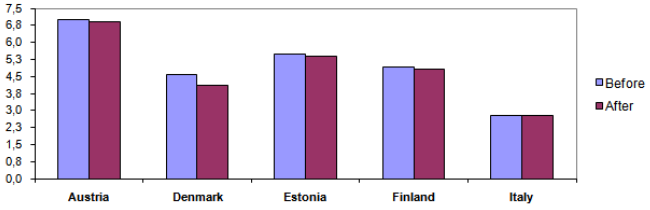
- EASY in the sense that the derivation rules for influence functions are similar to the rules for computing the derivative of a function in standard differential calculus. SAS macros to calculate influence functions can be easily written (see Appendix 2) and used as "black-box" programs by anyone who does not want to see their inner settings.
- POWERFUL in the sense that the approach encompasses more non-linear statistics than the linearization based on Taylor series.



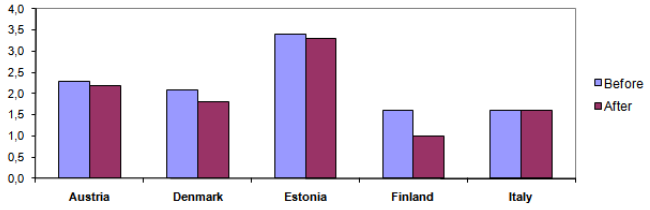
(a) CV (%) of the at-risk-of-poverty threshold before and after calibration, EU-SILC 2004



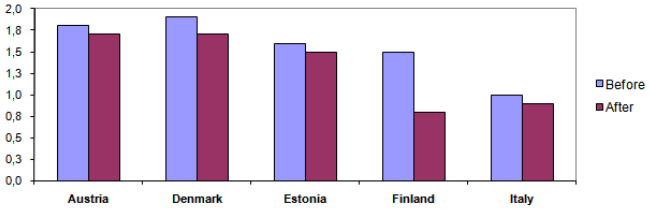
(b) CV (%) of the at-risk-of-poverty rate before and after calibration, EU-SILC 2004



(c) CV (%) of the relative median poverty gap before and after calibration, EU-SILC 2004



(d) CV (%) of the income quintile share ratio before and after calibration, EU-SILC 2004



(e) CV (%) of the Gini coefficient before and after calibration, EU-SILC 2004

Figure 3. CV before and after calibration

However, as indicated earlier, all linearization approaches rest on the asymptotic assumption that the sample size is large enough for the linear approximation to be valid. Although the assumption is likely to be correct when dealing with samples of thousands of units like, for instance, the samples which were selected for EU-SILC, one should be more careful when applying the method to smaller samples.

References

Ardilly, P., & Osier, G. (2007). Cross-sectional variance estimation for the French "Labour Force Survey". *Survey Research Methods*, *1*(2), 75-83.

Clemenceau, A., & Museux, J. M. (2006). *EU-SILC (Community Statistics on Income and Living Conditions): General presentation of the instrument.* Proceedings of the EU-SILC Conference (Helsinki, 6-8 November 2006), pp. 13-36. http://epp.eurostat.ec.europa.eu/.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, *25*, 193-203.

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376-382.

Deville, J. C., Särndal, C. E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, *88*(423), 1013-1020.

Hampel, F., Ronchetti, E., Rousseeuw, P., & Stahel, W. (1986). *Robust Statistics: the approach based on influence functions*. New York: John Wiley & Sons.

Isaki, C. T., & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, *77*, 89-96.

Osier, G. (2003). *Utilisation du logiciel poulpe pour des calculs de précision sur l'enquête emploi en continu.* Internal report, INSEE.

Osier, G., & Museux, J. M. (2006). *Variance estimation for EU-SILC complex poverty indicators using linearization techniques.* Proceedings of the European Conference on Quality in Survey Statistics (Q2006), Cardiff, 24-26 April 2006. http://www.statistics.gov.uk/events/q2006/.

Reeds, J. A. (1976). *On the Definition of Von Mises Functionals.* Ph.D. dissertation. Harvard University.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Series in Statistics.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. London: Chapman and Hall.

Tepping, B. J. (1968). *Variance Estimation in Complex Surveys.* Proceedings of the American Statistical Association, Social Statistics Section, pp. 11-18.

Tillé, Y. (2000). *Echantillonnage et estimation en populations finies*. Paris: Dunod.

Van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics.

Verma, V., & Betti, G. (2005). *Sampling errors and design effects for poverty measures and other complex statistics.* Working Paper, 53. Siena: Dipartimento di Metodi Quantitativi, Università degli Studi.

Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer.

Woodruff, R. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, *66*(334), 411-414.

Appendix 1: Linearization formulae

This Appendix contains the expressions of the influence functions of the main EU-SILC indicators:

- The at-risk-of-poverty threshold
- The at-risk-of-poverty rate
- The relative median poverty gap
- The income quintile share ratio (S80/S20)
- The Gini coefficient

The notations which are used here are similar to those introduced in the section 3.1 of the main document.

1. Linearization of the at-risk-of-poverty threshold (ARPT)

1.1 Expression of the indicator as a functional of M

The at-risk-of-poverty threshold is defined as 60% of the median income:

$$ARPT = 0.6 \times MED(M) = T(M)$$

where:

- $MED(M)$ is the median income, that is, the value which splits the income distribution in halves: $F[M, MED(M)] = 0.5$
- F designates the cumulative income distribution function: $F(M, x) = \frac{1}{N} \sum_{i \in U} 1(inc_i \leq x)$

1.2 Derivation of the influence function

Since, by definition, the median income satisfies the following identity: $F[M, MED(M)] = 0.5$, the corresponding influence function is equal to 0 for all k:

$$IF_k[M, MED(M)] = 0$$

On the other hand, the above influence function can be expanded using the derivation rule (33)

$$IF_k[M, MED(M) | MED(M)\, fixed] + \left[\frac{dF(M, x)}{dx} |_{x=MED(M)} \right] \times IMED_k(M) = 0$$

The influence function of F with respect to M holding $MED(M)$ constant is given in (37):

$$IF_k[M, MED(M) | MED(M)\, fixed] = \frac{1}{N} [1(inc_k \leq MED(M)) - 0.5]$$

Besides, let $\tilde{F}$ denote the function $x \mapsto F(M, x)$. Assuming the derivative $\tilde{F}'$ of $\tilde{F}$ exists and is strictly non-negative for all x, we have:

$$\frac{1}{N} [1(inc_k \leq MED(M)) - 0.5] + \tilde{F}'[MED(M)] \times IMED_k(M) = 0$$

Hence, the influence function of the median income is given by:

$$IMED_k(M) = -\frac{1}{N \cdot \tilde{F}'[MED(M)]} \cdot [1(inc_k \leq MED(M)) - 0.5]$$

and the influence function of the at-risk-of-poverty threshold at k is:

$$IARPT_k(M) = 0.6 \times IMED_k(M) = -\frac{0.6}{\tilde{F}'[MED(M)]} \times \frac{1}{N} \times [1(inc_k \leq MED(M)) - 0.5]$$

Unfortunately, the derivative of a cumulative distribution function is always 0 or not defined. In order to circumvent this problem, let approximate $\tilde{F}$ by the following convolution product:

$$\tilde{F}_K(x) = \int \tilde{F}(t) \cdot K(x, t) \cdot dt$$

where the two variable function $K(.,.)$ is a Gaussian kernel: $K(x,t) = \frac{1}{h\sqrt{2\pi}} \exp\left[-\frac{(x-t)^2}{2h^2}\right]$
It can be easily seen that the function $\tilde{F}_K$ is differentiable, and we have for all x:

$$\tilde{F}'_K(x) = \frac{1}{h\sqrt{2\pi}} \cdot \frac{1}{N} \cdot \sum_{k \in U} \exp\left[-\frac{(x-x_k)^2}{2h^2}\right]$$

The derivative $\tilde{F}'_K$ exists and is strictly non-negative. The smoothing parameter h can be estimated by: $\hat{h} = \frac{\hat{\sigma}}{N^{-1/5}}$, where $\hat{\sigma}$ is the estimated standard deviation of the income distribution.
Finally, the influence function of the at-risk-of-poverty threshold at k is given by:

$$IARPT_k(M) = -\frac{0.6}{\tilde{F}'_K[MED(M)]} \times \frac{1}{N} \times [1(inc_k \leq MED(M)) - 0.5]$$

2. Linearization of the at-risk-of-poverty rate (ARPR)

2.1 Expression of the indicator as a functional of M

The at-risk-of-poverty rate is the share of persons with an income below the at-risk-of-poverty threshold $ARPT(M)$:

$$ARPR = F[M, ARPT(M)] = T(M)$$

where:

- The at-risk-of-poverty threshold $ARPT(M)$ is defined as 60% of the median income $MED(M)$
- F is the cumulative income distribution function: $F(M,x) = \frac{1}{N} \sum_{i \in U} 1(inc_i \leq x)$

2.2 Derivation of the influence function

By using the derivation rule (33), we obtain:

$$IARPR_k(M) = IF_k[M, ARPT(M) | ARPT(M) \ \ fixed] + \left[\frac{dF(M,x)}{dx}\Big|_{x=ARPT(M)}\right] \times IARPT_k(M)$$

The expression of the influence function of F with respect to M holding $ARPT(M)$ constant is given in (37):

$$IF_k[M, ARPT(M) | ARPT(M) \, fixed] = \frac{1}{N}[1(inc_k \leq ARPT(M)) - ARPR(M)]$$

As to the influence function $IARPT_k(M)$ of the at-risk-of-poverty threshold, an expanded formula has just been worked out in the previous section:

$$IARPT_k(M) = -\frac{0.6}{\tilde{F}'_K[MED(M)]} \times \frac{1}{N} \times [1(inc_k \leq MED(M)) - 0.5]$$

Besides, let $\tilde{F}$ denote the function $x \mapsto F(M,x)$. Assuming the derivative $\tilde{F}'$ of $\tilde{F}$ exists and is strictly non-negative for all x, we have:

$$\begin{aligned}
&IARPR_k(M) \\
&= \tfrac{1}{N}[1(inc_k \leq ARPT(M)) - ARPR(M)] + \tilde{F}'[ARPT(M)] \times \left[-\tfrac{0.6}{\tilde{F}'[MED(M)]} \times \tfrac{1}{N} \times [1(inc_k \leq MED(M)) - 0.5]\right] \\
&= \tfrac{1}{N}[1(inc_k \leq ARPT(M)) - ARPR(M)] - \tfrac{0.6}{N} \times \tfrac{\tilde{F}'[ARPT(M)]}{\tilde{F}'[MED(M)]} \times [1(inc_k \leq MED(M)) - 0.5]
\end{aligned}$$

As with the previous case, the derivative of a cumulative distribution function is always 0 or not defined. In order to circumvent this problem, let approximate $\tilde{F}$ by the following convolution product:

$$\tilde{F}_K(x) = \int \tilde{F}(t) \cdot K(x,t) \cdot dt$$

where the function $K(.,.)$ is a Gaussian kernel: $K(x,t) = \frac{1}{h\sqrt{2\pi}} \exp\left[-\frac{(x-t)^2}{2h^2}\right]$
It can be easily seen that the function $\tilde{F}_K$ is differentiable, and we have for all x:

$$\tilde{F}'_K(x) = \frac{1}{h\sqrt{2\pi}} \cdot \frac{1}{N} \cdot \sum_{k \in U} \exp\left[-\frac{(x-x_k)^2}{2h^2}\right]$$

The derivative $\tilde{F}'_K$ exists and is strictly non-negative. The smoothing parameter h can be estimated by: $\hat{h} = \frac{\hat{\sigma}}{N^{-1/5}}$, where $\hat{\sigma}$ is the estimated standard deviation of the income distribution.
Finally, the influence function of the at-risk-of-poverty rate at k is given by:

$$IARPR_k(M) = \frac{1}{N}\left[1\left(inc_k \leq ARPT(M)\right) - ARPR(M)\right] - \frac{0.6}{N} \times \frac{\tilde{F}'_K\left[ARPT(M)\right]}{\tilde{F}'_K\left[MED(M)\right]} \times \left[1\left(inc_k \leq MED(M)\right) - 0.5\right]$$

3. Linearization of the relative median poverty gap

3.1 Expression of the indicator as a functional of M

The relative median poverty gap $RMPG$ is the difference between the at-risk-of-poverty threshold $ARPT$, that is, 60% of the median income, and the median income MED^p of the persons whose income is lower than $ARPT$, the difference being expressed relatively to $ARPT$:

$$RMPG = \frac{ARPT - MED^p}{ARPT} = 1 - \frac{MED^p}{ARPT}$$

1. The at-risk-of-poverty threshold $ARPT$ is set at 60% of the median income $MED(M)$:

$$ARPT = 0.6 \times MED(M) = T(M)$$

2. The median income MED^p of the persons whose income is below 60% of the median income actually satisfies the following identity:

$$F\left(M, MED^p(M)\right) = \frac{1}{2} F\left[M, ARPT(M)\right]$$

where F is the cumulative income distribution function: $F(M,x) = \frac{1}{N}\sum_{i \in U} 1\left(inc_i \leq x\right)$.

3.2 Derivation of the influence function

The idea is to use the derivation rule (31) for a ratio of two functionals. Thus, we obtain:

$$IRMPG_k(M) = -\frac{ARPT(M) \times IMED^p_k(M) - MED^p(M) \times IARPT_k(M)}{ARPT(M)^2}$$

As to the influence function $IARPT_k(M)$ of the at-risk-of-poverty threshold, an expanded formula has been worked out in the section (1.2) of this Appendix:

$$IARPT_k(M) = -\frac{0.6}{\tilde{F}'_K\left[MED(M)\right]} \times \frac{1}{N} \times \left[1\left(inc_k \leq MED(M)\right) - 0.5\right]$$

The only remaining issue is the derivation of the influence function $IMED^p_k(M)$ of the median income of the persons who are below 60% of the median income. By using the identity: $F\left(M, MED^p(M)\right) = \frac{1}{2} F\left[M, ARPT(M)\right]$, we can say that the corresponding influence functions are equal for all k:

$$IF_k\left[M, MED^p(M)\right] = \frac{1}{2} \cdot IF_k\left[M, ARPT(M)\right]$$

By using the derivation rule (33), we obtain:

$$\frac{1}{N} \cdot \left[1\left(inc_k \leq MED^p(M)\right) - F\left(M, MED^p(M)\right)\right] + \tilde{F}'_K\left(MED^p(M)\right) \times IMED^p_k(M)$$

$$= \frac{1}{2} \cdot \left[\frac{1}{N} \cdot \left[\ 1\left(inc_k \leq ARPT(M)\right) - F\left(M, ARPT(M)\right)\right] + \tilde{F}'_K\left(ARPT(M)\right) \times IARPT_k(M)\right]$$

The influence function $IMED^p_k(M)$ can thus be easily deduced from the above equation.

4. Linearization of the income quintile share ratio S80/S20

4.1 Expression of the indicator as a functional of M

Let $q_-(M)$ and $q_+(M)$ denote the bottom and the top income quintile, respectively. The income quintile share ratio S80/S20 can be written as:

$$S80/S20 = \frac{\int_{inc>q_+(M)} inc \cdot dM}{\int_{inc\leq q_-(M)} inc \cdot dM} = \frac{\int inc \cdot dM - \int_{inc\leq q_+(M)} inc \cdot dM}{\int_{inc\leq q_-(M)} inc \cdot dM} = \frac{R(M) - S\left[M, q_+(M)\right]}{S\left[M, q_-(M)\right]} = T(M)$$

where the functionals R and S are defined as:

$$R(M) = \int inc \cdot dM \quad and \quad S\left[M, x\right] = \int_{inc\leq x} inc \cdot dM$$

4.2 Derivation of the influence function

According to the previous section, the S80/S20 indicator can be expressed as a ratio of two functionals. The idea then is to apply the derivation rule (31):

$$IT_k(M) = \frac{S\left[M, q_-(M)\right] \times \left\{IR_k(M) - IS_k\left[M, q_+(M)\right]\right\} - \left\{R(M) - S\left[M, q_+(M)\right]\right\} \times IS_k\left[M, q_-(M)\right]}{S\left[M, q_-(M)\right]^2}$$

By using (34), we have: $IR_k(M) = inc_k$.
The last issue is the derivation of the influence function of $S\left[M, q(M)\right]$, where $q(M)$ stands for $q_-(M)$ (bottom income quintile) and $q_+(M)$ (top income quintile). By using the derivation rule (33), we got:

$$IS_k(M) = IS_k\left[M, q(M) | q(M)\ fixed\right] + \left[\frac{dS(M,x)}{dx}\Big|_{x=q(M)}\right] \times Iq_k(M)$$

1. The influence function of S holding the quintile value $q(M)$ constant is:

$$IS_k\left(M, q(M) | q(M)\ \ fixed\right)$$

$$= Lim_{t\to 0} \frac{S\left[M+t\delta_k, q(M)\right] - S\left[M, q(M)\right]}{t}$$

$$= Lim_{t\to 0} \frac{\int_{inc\leq q(M)} inc \cdot d(M+t\delta_k) - \int_{inc\leq q(M)} inc \cdot dM}{t}$$

$$= Lim_{t\to 0} \frac{\int_{inc\leq q(M)} inc \cdot d(t\delta_k)}{t} \quad = \quad \begin{cases} inc_k\ if\ inc_k \leq q(M) \\ 0\ otherwise \end{cases} \quad = inc_k \times 1\left[inc_k \leq q(M)\right]$$

2. The influence function $Iq_k(M)$ of the quintile is given by:

$$Iq_k(M) = -\frac{1}{N \cdot \tilde{F}'_K[q(M)]} \cdot [1(inc_k \leq q(M)) - \alpha]$$

where $\alpha \in [0,1]$ is the order of the quintile ($\alpha = 0.2$ for the bottom quintile, $\alpha = 0.8$ for the top quintile). The derivation is actually the same as for the influence function of the median income (see section 1.2).

3. Let $\tilde{S}$ denote the function $x \mapsto S(M,x)$. Considering the derivative of $\tilde{S}$ is always 0 or not defined, let approximate $\tilde{S}$ by the following convolution product:

$$\tilde{S}_K(x) = \int \tilde{S}(t) \cdot K(x,t) \cdot dt$$

where the function $K(.,.)$ is a Gaussian kernel: $K(x,t) = \frac{1}{h\sqrt{2\pi}} \exp\left[-\frac{(x-t)^2}{2h^2}\right]$

It can be easily seen that the derivative $\tilde{S}'_K$ of $\tilde{S}_K$ exists and is strictly non-negative.

Finally, the influence function of $S[M, q(M)]$ can be easily deduced from the three above results.

5. Linearization of the Gini coefficient

5.1 Expression of the indicator as a functional of M

Let U denote a population of size N and let $y = \{y_i, i \in U\}$ denote an income distribution over the population U. Let r_i be the rank of i in the distribution y after we sort it in ascending income. The Gini coefficient G is given by:

$$1 + G = \frac{2 \times \sum_{i \in U} r_i \cdot inc_i - \sum_{i \in U} inc_i}{N \cdot \sum_{i \in U} inc_i} = \frac{2 \times \int inc_i \cdot \left[\int 1\left(inc_j \leq inc_i\right) \cdot dM(j)\right] \cdot dM(i) \quad - \quad \int inc \cdot dM}{\left(\int dM\right) \cdot \left(\int inc \cdot dM\right)} \quad = T(M)$$

5.2 Derivation of the influence function

Let us denote:

- $T_1(M) = T_1 = \int inc_i \cdot \left[\int 1\left(inc_j \leq inc_i\right) \cdot dM(j)\right] \cdot dM(i)$
- $T_2(M) = \int inc \cdot dM = INC$
- $T_3(M) = \int dM = N$

The Gini coefficient can then be expressed as:

$$T(M) = \frac{2T_1(M) - T_2(M)}{T_2(M) \cdot T_3(M)}$$

By using the derivation rule (31), we obtain:

$$IT_k(M) = \frac{T_2(M) \cdot T_3(M) \cdot I(2T_1 - T_2)_k(M) - [2T_1(M) - T_2(M)] \cdot I(T_2T_3)_k(M)}{[T_2(M) \cdot T_3(M)]^2}$$

By using the derivation rule (29) as well as the result (34), the influence function of $(2T_1 - T_2)$ is given by:

$$I(2T_1 - T_2)_k(M) \quad = \quad 2 \times I(T_1)_k(M) - I(T_2)_k(M) \quad = \quad 2 \times I(T_1)_k(M) - inc_k$$

By using the derivation rule (30) and the result (34), the influence function of the product $(T_2 \cdot T_3)$ is given by:

$$I(T_2 \cdot T_3)_k(M) = T_2(M) \cdot I(T_3)_k(M) + T_3(M) \cdot I(T_2)_k(M) = INC \times 1 + N \times inc_k = INC + N \cdot inc_k$$

Consequently, the influence function of the Gini coefficient can be written as:

$$IT_k(M) = \frac{T_2(M) \cdot T_3(M) \cdot [2 \cdot I(T_1)_k(M) - inc_k] - [2T_1(M) - T_2(M)] \cdot (INC + N \cdot inc_k)}{[T_2(M) \cdot T_3(M)]^2}$$

$$= \frac{2 \cdot INC \cdot N \cdot I(T_1)_k(M) - INC \cdot N \cdot inc_k - \left[2 \times \int inc_i \left(\int 1\left(inc_j \leq inc_i\right) \cdot dM(j) \right) \cdot dM(i) - INC \right] \cdot (INC + N \cdot inc_k)}{(N \times INC)^2}$$

The last remaining hurdle is the derivation of the influence function of the functional $T_1(M)$. For all $t > 0$, we have:

$$\begin{aligned}
&T_1(M + t \cdot \delta_k) \\
&= \int inc_i \cdot \left[\int 1\left(inc_j \leq inc_i\right) \cdot d(M + t \cdot \delta_k)(j) \right] \cdot d(M + t \cdot \delta_k)(i) \\
&= \int inc_i \cdot \left[\int 1\left(inc_j \leq inc_i\right) \cdot dM(j) \right] \cdot dM(i) \\
&+ \int inc_i \cdot \left[t \cdot \int 1\left(inc_j \leq inc_i\right) \cdot d(\delta_k)(j) \right] \cdot dM(i) \\
&+ t \cdot \int inc_i \cdot \left[\int 1\left(inc_j \leq inc_i\right) \cdot dM(j) \right] \cdot d(\delta_k)(i) \\
&+ t \cdot \int inc_i \cdot \left[t \cdot \int 1\left(inc_j \leq inc_i\right) \cdot d(\delta_k)(j) \right] \cdot d(\delta_k)(i) \\
&= T_1(M) \\
&+ t \cdot \int inc_i \cdot [1(inc_k \leq inc_i)] \cdot dM(i) \\
&+ t \cdot inc_k \cdot \left[\int 1\left(inc_j \leq inc_k\right) \cdot dM(j) \right] \\
&+ t^2 \cdot inc_k
\end{aligned}$$

Consequently, we have for all $t > 0$:

$$\frac{T_1(M + t \cdot \delta_k) - T_1(M)}{t} = \int inc_i \cdot [1(inc_k \leq inc_i)] \cdot dM(i) + inc_k \cdot \left[\int 1\left(inc_j \leq inc_k\right) \cdot dM(j) \right] + t \cdot inc_k$$

So the influence function of T_1 is:

$$I(T_1)_k(M) = Lim_{t \to 0} \frac{T_1(M + t \cdot \delta_k) - T_1(M)}{t} = \int inc_i \cdot [1(inc_k \leq inc_i)] \cdot dM(i) + inc_k \cdot \left[\int 1\left(inc_j \leq inc_k\right) \cdot dM(j) \right]$$

Consequently, the influence function of the Gini coefficient can be easily obtained from the above expressions.

Appendix 2: SAS macros to compute influence functions

```
* ------------------------------------------------------------------------------
    SAS MACROS TO CALCULATE THE INFLUENCE FUNCTIONS OF THE EU-SILC INDICATORS:
       >>>  THE AT-RISK-OF-POVERTY THRESHOLD
       >>>  THE AT-RISK-OF-POVERTY RATE
       >>>  THE RELATIVE MEDIAN POVERTY GAP
       >>>  THE INCOME QUINTILE SHARE RATIO
       >>>  THE GINI COEFFICIENT
    AUTHOR: GUILLAUME OSIER

    ***** THESE MACROS HAVE BEEN ADAPTED FROM THE PROGRAMS THAT HAD BEEN DEVELOPED
   BY EUROSTAT, AND WHICH ARE AVAILABLE ON CIRCA *****
------------------------------------------------------------------------------- ;


* ------------------------------------------------
   MACRO 1 : LINEARIZATION OF THE AT-RISK-OF-POVERTY
                         THRESHOLD
 ------------------------------------------------- ;

%MACRO LIN_ARPT (DATA = , LIBRARY = , INCOME = , WEIGHT = , ORDER = 50, PERCENT = 60);
* ----------------------------------------------------------------
   >  DATA: SAS dataset
   >  LIBRARY: SAS library containing the dataset
   >  INCOME: income variable
   >  WEIGHT: weighting variable
   >  ORDER: Order of the income quantile (by default, 50%)
   >  PERCENT: Percentage of the income quantile (by default, 60%)
---------------------------------------------------------------- ;

data t;
 set &library..&data;
run;

proc univariate data=t noprint;
 var &income;
 weight &weight;
 output out=_out_ pctlpts=&order pctlpre=quant pctlname=ile;
run;

data _null_;
 set _out_;
 call symput('quant_val',quantile);
  /* The value of the median income is stored
     into the macro-variable &quant_val */
run;

proc iml;

 edit work.t;
 param={&income &weight};
 read all var param into mat;

 inc=mat[,1];
 wght=mat[,2];
```

```
 /* Population size */
 N=sum(wght);

 /* Bandwith parameter - h=S/N^(1/5) */
 h=sqrt((sum(wght#inc#inc)-sum(wght#inc)*sum(wght#inc)/sum(wght))/sum(wght))
 /exp(0.2*log(sum(wght)));

 /* Estimate of F'(quantile) */
 u=(&quant_val-inc)/h;
 vect_f=exp(-(u##2)/2)/sqrt(2*3.1415926536);
 f_quant=(vect_f'*wght)/(N*h);

 * ========== LINEARIZED VARIABLE  OF THE AT-RISK-OF POVERTY THRESHOLD ============== ;
           lin=-(&percent/100)#(1/N)#((inc<=&quant_val)-&order/100)/f_quant;

 create lin_var from lin[colname={linvar}];
 append from lin;

quit;

data &library..&data;
 merge &library..&data lin_var;
run;

%MEND LIN_ARPT;


* ------------------------------------------------------
    MACRO 2 : LINEARIZATION OF THE AT-RISK-OF-POVERTY
                          RATE
  ------------------------------------------------------ ;

%MACRO LIN_ARPR (DATA = , LIBRARY = , INCOME = , WEIGHT = , ORDER = 50 , PERCENT = 60);
* ---------------------------------------------------------------------
   >  DATA: SAS dataset
   >  LIBRARY: SAS library containing the dataset
   >  INCOME: income variable
   >  WEIGHT: weighting variable
   >  ORDER: Order of the income quantile (by default, 50%)
   >  PERCENT: Percentage of the income quantile (by default, 60%)
---------------------------------------------------------------------- ;

* === I. CALCULATION OF THE AT-RISK-OF-POVERTY THRESHOLD === ;

data t;
 set &library..&data;
run;

proc univariate data=t noprint;
 var &income;
 weight &weight;
 output out=_out_ pctlpts=&order pctlpre=quant pctlname=ile;
run;
```

```
data _out_;
 set _out_;
 threshold = (&percent/100)*quantile;  /* At-risk-of-poverty threshold */
run;

data _null_;
 set _out_;
 call symput('quant_val',quantile);
     /* Storage of the median income into the macro-variable &quant_val */
 call symput('thres_val',threshold);
     /* Storage of the poverty threshold into the macro-variable &thres_val */
run;

* === II. CALCULATION OF THE AT-RISK-OF-POVERTY RATE === ;

data t;
 set t;
 if &income<=&thres_val then poor=1; else poor=0;
run;

proc means data=t noprint;
 var poor;
 weight &weight;
 output out=_out_ mean(poor) = poor;  /* At-risk-of-poverty rate */
run;

data _null_;
 set _out_;
 call symput('rate_val',poor);
    /* Storage of the at-risk-of-poverty rate into the macro-variable &rate_val */
run;

* === III. LINEARIZATION OF THE AT-RISK-OF-POVERTY RATE === ;

proc iml;

 edit work.t;
 param={&income &weight};
 read all var param into mat;

 inc=mat[,1];
 wght=mat[,2];

 /* Population size */
 N=sum(wght);

 /* Bandwith parameter - h=S/N^(1/5) */
 h=sqrt((sum(wght#inc#inc)-sum(wght#inc)*sum(wght#inc)/sum(wght))
 /sum(wght))/exp(0.2*log(sum(wght)));

 /* Estimate of F'(quantile) */
 u1=(&quant_val-inc)/h;
 vect_f1=exp(-(u1##2)/2)/sqrt(2*3.1415926536);
 f_quant1=(vect_f1`*wght)/(N*h);
```

```
 /* Estimate of F'(beta*quantile) */
 u2=(&thres_val-inc)/h;
 vect_f2=exp(-(u2##2)/2)/sqrt(2*3.1415926536);
 f_quant2=(vect_f2`*wght)/(N*h);

/* Linearization of the at-risk-of-poverty threshold */
lin_thres=-(&percent/100)#(1/N)#((inc<=&quant_val)-&order/100)/f_quant1;

* ========== LINEARIZED VARIABLE OF THE AT-RISK-OF-POVERTY RATE ============== ;
       lin=100*((1/N)#((inc<=&thres_val)-&rate_val)+f_quant2*lin_thres);

 create lin_var from lin[colname={linvar}];
 append from lin;

quit;

data &library..&data;
 merge &library..&data lin_var;
run;


%MEND LIN_ARPR;


* ---------------------------------------------------
    MACRO 3 : LINEARIZATION OF THE RELATIVE MEDIAN
                AT-RISK-OF-POVERTY GAP
 --------------------------------------------------- ;

%MACRO LIN_RMPG (DATA = , LIBRARY = , INCOME = , WEIGHT = , ORDER = 50 , PERCENT = 60);
* ------------------------------------------------------------------
   >  DATA: SAS dataset
   >  LIBRARY: SAS library containing the dataset
   >  INCOME: income variable
   >  WEIGHT: weighting variable
   >  ORDER: Order of the income quantile (by default, 50%)
   >  PERCENT: Percentage of the income quantile (by default, 60%)
------------------------------------------------------------------- ;

* === I. CALCULATION OF THE AT-RISK-OF-POVERTY THRESHOLD === ;

data t;
 set &library..&data;
run;

proc univariate data=t noprint;
 var &income;
 weight &weight;
 output out=_out_ pctlpts=&order pctlpre=quant pctlname=ile;
run;

data _out_;
 set _out_;
 thres=(&percent/100)*quantile; /* At-risk-of-poverty threshold */
run;
```

```
data _null_;
 set _out_;
 call symput('quant_val',quantile);
     /* Storage of the median income into the macro-variable &quant_val */
 call symput('thres_val',thres);
     /* Storage of the poverty threshold into the macro-variable &thres_val */
run;

* === II. CALCULATION OF THE AT-RISK-OF-POVERTY RATE === ;

data t;
 set t;
 if &income<=&thres_val then poor=1; else poor=0;
run;

proc means data=t noprint;
 var poor;
 weight &weight;
 output out=_out_ mean(poor)=poor;
run;

data _null_;
 set _out_;

call symput('rate_val',poor);  /* At-risk-of-poverty rate */
run;

* === III. MEDIAN INCOME OF THE PERSONS WHOSE INCOME IS LOWER THAN THE POVERTY THRESHOLD === ;

proc summary data=t;
 where &income <= &thres_val;
 var &income;
 weight &weight;
 output out=_out_ median(&income)=mediane;
run;

data _null_;
 set _out_;
 call symput('median_poor',mediane);
run;

* === IV. LINEARIZATION OF THE RELATIVE MEDIAN AT-RISK-OF-POVERTY GAP === ;

proc iml;

 edit work.t;
 parametre={&income &weight};
 read all var parametre into mat;

 inc=mat[,1];\tab
 wght=mat[,2];

 /* Population size */
 N=sum(wght);
```

```
 /* Bandwith paramter - h=S/N^(1/5) */
 h=sqrt((sum(wght#inc#inc)-
 sum(wght#inc)*sum(wght#inc)/sum(wght))/sum(wght))/exp(0.2*log(sum(wght)));



 u1=(&quant_val-inc)/h;
 vect_f1=exp(-(u1##2)/2)/sqrt(2*3.1415926536);
 f_quant1=(vect_f1`*wght)/(N*h);

 u2=(&thres_val-inc)/h;
 vect_f2=exp(-(u2##2)/2)/sqrt(2*3.1415926536);
 f_quant2=(vect_f2`*wght)/(N*h);

u3=(&median_poor-inc)/h;
 vect_f3=exp(-(u3##2)/2)/sqrt(2*3.1415926536);
 f_quant3=(vect_f3`*wght)/(N*h);

 lin_thres=-(&percent/100)#(1/N)#((inc<=&quant_val)-&order/100)/f_quant1;
 lin_rate=(1/N)#((inc<=&thres_val)-&rate_val)+f_quant2*lin_thres;
 lin_median_poor=(0.5*lin_rate-(1/N)#((inc<=&median_poor)-0.5*&rate_val))/f_quant3;

* ==============  LINEARIZED VARIABLE OF THE RELATIVE MEDIAN POVERTY GAP ============= ;
    lin=100*(&median_poor*lin_thres/(&thres_val*&thres_val)-lin_median/&thres_val);

create lin_var from lin[colname={linvar}];
append from lin;

data &library..&data;
 merge &library..&data lin_var;
run;

%MEND LIN_RMPG;


* ------------------------------------------------------
   MACRO 4 : LINEARIZATION OF THE INCOME QUINTILE
                  SHARE RATIO S80/S20
 ------------------------------------------------------ ;

%MACRO LIN_IQR (DATA = , LIBRARY = , INCOME = , WEIGHT = , ALPHA = 20);
* ------------------------------------------------------------
   >  DATA: SAS dataset
   >  LIBRARY: SAS library containing the dataset
   >  INCOME: income variable
   >  WEIGHT: weighting variable
   >  ALPHA: Order of the income quantile (by default, 20%)
 ------------------------------------------------------------ ;

%let alpha2=%sysevalf(100-&alpha);

data t;
 set &library..&data;
run;
```

```
proc univariate data=t noprint;
 var &income;
 weight &weight;
 output out=_out_ pctlpts=&alpha &alpha2 pctlpre=quant pctlname=i1 i2;
run;

data _null_;
 set _out_;
 call symput ('quant_inf',quanti1); /* Bottom income quantile */
 call symput ('quant_sup',quanti2); /* Top income quantile */
run;

data t;
 set t;
 indinf = &income * (&income <= &quant_inf);
 indsup = &income * (&income > &quant_sup);
run;

proc means data=t noprint;
 var indinf indsup;
 weight &weight;
 output out=_out_ sum(indinf)=den /* Total income for the bottom quintile */
                  sum(indsup)=sup;/* Total income for the top quintile */
run;

data _null_;
 set _out_;
 call symput('num_val',num);
 call symput('den_val',den);
run;

proc iml;

edit work.t;
 param={&income &weight};
 read all var param into mat;

 inc=mat[,1];
 wght=mat[,2];
 v=wght#inc;

 /* Population size */
 N=sum(wght);

 /* Bandwith parameter - h=S/N^(1/5) */
 h=sqrt((sum(wght#inc#inc)-
 sum(wght#inc)*sum(wght#inc)/sum(wght))/sum(wght))/exp(0.2*log(sum(wght)));

 /*=====  1. Linearization of the bottom quantile =====*/

 u1=(&quant_inf-inc)/h;
 vect_f1=exp(-(u1##2)/2)/sqrt(2*3.1415926536);
 f_quant1=(vect_f1`*wght)/(N*h);

 lin_inf=-(1/N)#((inc<=&quant_inf)-&alpha/100)/f_quant1;
```

```
 /*=====  2. Linearization of the top quantile =====*/

 u2=(&quant_sup-inc)/h;
 vect_f2=exp(-(u2##2)/2)/sqrt(2*3.1415926536);
 f_quant2=(vect_f2`*wght)/(N*h);

 lin_sup=-(1/N)#((inc<=&quant_sup)-&alpha2/100)/f_quant2;

 /*=====  3. Linearization of the total income for the top quintile =====*/

 u3=(&quant_sup-inc)/h;
 vect_f3=exp(-(u3##2)/2)/sqrt(2*3.1415926536);
 f_quant3=(vect_f3`*v)/h;

 lin_num=inc-inc#(inc<=&quant_sup)-f_quant3#lin_sup;

/*=====  4. Linearization of the total income for the bottom quintile =====*/

 u4=(&quant_inf-inc)/h;
 vect_f4=exp(-(u4##2)/2)/sqrt(2*3.1415926536);
 f_quant4=(vect_f4`*v)/h;

 lin_den=inc#(inc<=&quant_inf)+f_quant4#lin_inf;

 /*========    LINEARIZED VARIABLE OF THE IQ SHARE RATIO     ========*/
 lin=((&den_val)#lin_num-(&num_val)#lin_den)/(&den_val*&den_val);

 create lin_var from lin[colname={linvar}];
 append from lin;

quit;

data &library..&data;
 merge &library..&data lin_var;
run;

%MEND LIN_IQR;

* ------------------------------------------------------
    MACRO 5 : LINEARIZATION OF THE GINI COEFFICIENT
-------------------------------------------------------;

%MACRO LIN_GINI (DATA= , LIBRARY = , INCOME = , WEIGHT = );
* ------------------------------------------------------
   >  DATA: SAS dataset
   >  LIBRARY: SAS library containing the dataset
   >  INCOME: income variable
   >  WEIGHT: weighting variable
------------------------------------------------------- ;

proc sort data=&library..&data;
 by &income;
run;

proc iml;
```

```
 edit &library..&data;
 param={&income &weight};
 read all var param into mat;

 taille=nrow(mat); /* Sample size */
 N=mat[+,2]; /* Population size */
 T=sum(mat[,1]#mat[,2]); /* Total income */

 r=j(taille,1,1);
 r[1,1]=mat[1,2];
 do i=2 to taille;
  r[i,1]=r[i-1,1]+mat[i,2]; /* r[i,1] is the cumulative weight of the person i */
 end;

 Num=sum((2*r[,1]-1)#(mat[,1]#mat[,2]));
 Den=N*T;

 /*** Gini coefficient ***/
     Gini=Num/Den-1;

 F=j(taille,1,1);
 F[1,1]=mat[1,2]/N;
 do i=2 to taille;
  F[i,1]=F[i-1,1]+mat[i,2]/N; /* Cumulative income distribution function */
 end;

 G=j(taille,1,1);
 G[1,1]=mat[1,1]*mat[1,2];
 do i=2 to taille;
  G[i,1]=G[i-1,1]+mat[i,1]*mat[i,2]; /* Weighted partial sum */
 end;

/*==========    LINEARIZED VARIABLE OF THE GINI COEFFICIENT    ==========*/
 lin=100*(2*(T-G[,1]+mat[,1]#mat[,2]+N*(mat[,1]#F[,1]))-mat[,1]-(Gini+1)*(T+N*mat[,1]))/(N*T);

 create lin_var from lin[colname={linvar}];
 append from lin;

quit;

data &library..&data;
 merge &library..&data lin_var;
run;

%MEND LIN_GINI;
```