

# Variance estimation: the linearization approach applied by Eurostat to the 2004 SILC operation

Guillaume Osier\*

\* Statistical Office of the European Communities (Eurostat), Luxembourg (guillaume.osier@ec.europa.eu)

**Abstract:** this paper intends to present the approach which was implemented by Eurostat for producing variance estimates for the EU-SILC social indicators (at-risk-of-poverty rate, at-risk-of-poverty threshold, Gini coefficient...). After linearizing the indicators, variance calculations were performed by the software Poulpe, a SAS macro-based application developed by France's National Statistical Office (INSEE). Variance estimates for target indicators are set out and some issues in connection with the method are raised.

## 1. The EU-SILC indicators

### 1.1. The "Laeken" indicators

In December 2001, the European Council of Laeken endorsed a first set of 18 common statistical indicators for social inclusion. They will allow monitoring in a comparable way Member States' progress towards agreed EU objectives regarding income, poverty and social exclusion.

### 1.2. Examples

#### *i) The at-risk-of-poverty threshold (ARPT)*

This is 60% of the median "equivalised" income. In EU-SILC, incomes are collected at household level and divided by the household size so as to convert them into individual incomes. The household size is calculated according to the modified OECD scale: weight 1 for the first adult aged 14 and over, 0.5 for any other adult and 0.3 for the children aged 13 or less. In particular all the members of a household receive the same income.

#### *ii) The at-risk-of-poverty rate (ARPR)*

This is the share of persons with an "equivalised" income below the at-risk-of-poverty threshold.

#### *iii) The income quintile share ratio ( $S_{80}/S_{20}$ )*

The income quintile share ratio is defined as the ratio of the total income of the persons above the top income quintile  $Q_{80}$  over that of the persons below the bottom income quintile  $Q_{20}$ :

$$\frac{S_{80}}{S_{20}} = \frac{\sum_{k|INC_k \geq Q_{80}} INC_k}{\sum_{k|INC_k \leq Q_{20}} INC_k} \quad (1)$$

iv) *The relative median poverty gap (RMPG)*

This is the relative difference between the at-risk-of-poverty threshold and the median income of the "poor" people:

$$RMPG = \frac{ARPT - MED_{POOR}}{ARPT} = 1 - \frac{MED_{POOR}}{ARPT} \quad (2)$$

v) *The Gini coefficient*

The Gini coefficient is a standard measure of the degree of inequality in an income distribution. It ranges from 0 and 1, with 0 representing perfect equality in the income distribution (all the units receive the same income) and 1 perfect inequality (one person has all the income). Let  $R_i$  denote the rank of a unit  $i$  in the ascending income sorted distribution. Then, we have:

$$1 + Gini = \frac{2 \sum_{i \in U} R_i \cdot INC_i - \sum_{i \in U} INC_i}{\left( \sum_{i \in U} 1 \right) \cdot \left( \sum_{i \in U} INC_i \right)} \quad (3)$$

### 1.3. Scope of the document

It aims to present the approach which was implemented by Eurostat for producing variance estimates for the "Laeken" indicators, taking into account their complex structure as well as the underlying sample design. This approach is clearly an "analytical" one. In particular, re-sampling techniques (Bootstrap, Jackknife...) were ruled out as the implementation may be cumbersome and not easily reproducible at European level.

## 2. The linearization technique

### 2.1. Basic idea

It consists of deriving from a "complex" non-linear statistic a linear statistic which has the same asymptotic variance:

$$Var(\hat{\theta}) \approx Var\left( \sum_{i \in s} \omega_{is} \times z_i \right) \quad (4)$$

Where:

- $s$  : effective sample
- $\omega_{is}$  : sample weight of  $i$  in  $s$
- $z_i$  : variable whose expression depends on  $\hat{\theta}$  ("linearized" variable)

Linearizing is the only way to make analytical variance calculations tractable as it solves the intrinsic problem of the complex structure of the "Laeken" indicators. Basically, two linearization frameworks were developed:

- A seminal framework (Taylor series)
- A generalized framework based on Influence functions

## 2.2. The seminal framework: Taylor series

This framework allows linearizing "weakly" non-linear estimators that can be expressed as differentiable functions of linear estimators:

$$\hat{\theta} = F(\hat{Y}_1, \hat{Y}_2 \dots \hat{Y}_p) \quad (5)$$

Where:

- $F$  is a differentiable function from  $R^p$  to  $R$
- $\hat{Y}_i$  is a linear estimator for the total  $Y_i$  of a variable  $y_i$

Under general assumptions, we have asymptotically:  $Var(\hat{\theta}) = Var\left(\sum_{i \in s} \omega_{is} \times z_i\right)$  with:

$$z_k = \sum_{j=1}^p \frac{\partial F}{\partial a_j}(Y_1, Y_2 \dots Y_p) \times y_{jk} \quad (6)$$

In this expression,  $z$  depends on the  $p$  totals  $Y_1, Y_2 \dots Y_p$ , which are unknown. In practice, estimated values are substituted for those quantities and actually calculations are carried out with the following pseudo-variable:

$$\hat{z}_k = \sum_{j=1}^p \frac{\partial F}{\partial a_j}(\hat{Y}_1, \hat{Y}_2 \dots \hat{Y}_p) \times y_{jk} \quad (7)$$

## 2.3. The generalized framework: Influence functions

In case of "strongly" non-linear estimators (like the EU-SILC indicators), i.e. non-differentiable functions of linear estimators, Taylor series cannot be used and the seminal framework no longer be applied. A generalized linearization framework covering a broader class of non-linear estimators was developed (Deville, 1999). The general expression of the target estimators is:

$$\hat{\theta} = F(\hat{M}) \quad (\text{"plug-in" estimators}) \quad (8)$$

Where:

- $F$  is a functional
- $\hat{M}$  is the "natural" measure on the sample  $s$ , i.e.  $\hat{M}(i) = \omega_{is}$  (sample weight) if  $i$  belongs to  $s$  and 0 otherwise

"Plug-in" estimators are used when one wants to estimate parameters  $\theta = F(M)$  where  $M$  is the "natural" measure on the target population  $U$ , i.e.  $M(i) = 1$  for all  $i$  in  $U$ .

**Example:** Let  $Y$  denote the total of a variable  $y$  over a population  $U$ . Let  $y_i$  the value of  $y$  on  $i$ . In this situation, the target parameter  $\theta$  and the corresponding "plug-in" estimator  $\hat{\theta}$  are:

$$* \quad \theta = \sum_{i \in U} y_i = \int_{k \in U} y_k dM(k) = F(M)$$

$$* \quad \hat{\theta} = F(\hat{M}) = \int_{k \in U} y_k d\hat{M}(k) = \sum_{k \in s} \omega_{ks} y_k$$

Under general assumptions, Deville (1999) shown that "plug-in" estimators can be linearized as well. A "linearized" variable at  $k$  is given by the **Influence function**:

$$z_k = IF_k(M) = \lim_{t \rightarrow 0} \frac{F(M + t\delta_k) - F(M)}{t} \quad (9)$$

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

In practice, Influence functions are unknown and have to be estimated by substituting the empirical measure  $\hat{M}$  for  $M$ :

$$\hat{z}_k = \hat{IF}_k(M) = \lim_{t \rightarrow 0} \frac{F(\hat{M} + t\delta_k) - F(\hat{M})}{t} \quad (10)$$

Deville (1999) shown that variance estimations based on estimated Influence functions are still valid so long as the sample size is big enough. Most of the commonly used estimators fit in with the generalized framework. In particular, this framework has been applied to the EU-SILC indicators.

### 3. The variance estimation software Poulpe

Poulpe is a SAS macro-based application developed by France's National Statistical Office (INSEE) for computing variance estimates for linear estimators with respect to a great number of sample designs. After linearizing the indicators, Eurostat has used Poulpe as software tool for variance estimation. For a given set of target linear(ized) estimators, Poulpe will estimate:

- The variances and standard errors
- The lower and upper bounds of the 95% confidence interval (based on a normal approximation):

$$CI(\theta, 95\%) = \left[ \hat{\theta} - 2\sqrt{\hat{V}(\hat{\theta})}, \hat{\theta} + 2\sqrt{\hat{V}(\hat{\theta})} \right] \quad (11)$$

- **The Design Effect.** This is the ratio of the variance with respect to the current sample design ( $P$ ) to the variance that would be obtained from a simple random sampling ( $SRS$ ) of same size and without replacement:

$$Deff = \frac{V_P(\hat{Y})}{V_{SRS}(N \cdot \bar{y})} \quad (12)$$

Poulpe can deal with total non-response by viewing it as an additional phase of selection and then applying variance formulas for multi-phase sampling (Särndal and al., 1992). In addition, the software can assess the impact of calibration weighting on variance by computing linear regression residuals (Deville and Särndal, 1992).

However, Poulpe cannot deal with item non-response (imputation).

#### 4. Numerical results (SILC 2004)

Variance estimates are available for 13 Countries:

- Austria (AT)
- Belgium (BE)
- Denmark (DK)
- Estonia (EE)
- Spain (ES)
- Finland (FI)
- France (FR)
- Greece (GR)
- Italy (IT)
- Ireland (IE)
- Norway (NO)
- Portugal (PT)
- Sweden (SE)

**Table 1: Estimated Coefficients of Variation CV (%)**

	AT	BE	DK	EE	ES	FI	FR	GR	IT	IE	NO	PT	SE
<b>Poverty threshold</b>	<b>0.9</b>	<b>1.0</b>	<b>0.1</b>	<b>1.2</b>	<b>0.8</b>	<b>0.4</b>	<b>0.6</b>	<b>1.0</b>	<b>0.6</b>	<b>0.9</b>	<b>0.5</b>	<b>1.5</b>	<b>0.6</b>
<b>Poverty rate</b>	<b>4.6</b>	<b>3.8</b>	<b>0.5</b>	<b>3.2</b>	<b>1.9</b>	<b>3.4</b>	<b>3.0</b>	<b>2.6</b>	<b>1.6</b>	<b>2.7</b>	<b>3.5</b>	<b>3.0</b>	<b>3.8</b>
Male	5.7	4.4	0.4	3.5	2.2	3.9	3.4	3.0	1.9	3.0	4.4	3.2	4.6
Male and 16-24 years	13.9	9.5	0.4	7.5	5.4	7.2	6.5	6.8	4.2	6.5	6.8	7.5	7.1
Male and 25-49 years	7.1	6.3	0.2	4.6	2.8	5.6	4.7	4.1	2.4	4.7	6.8	4.4	7.3
Male and 50-64 years	10.7	9.2	2.0	7.9	4.7	7.6	5.6	5.2	3.8	4.3	16.1	6.3	12.1
Male and 16-64 years	5.9	4.9	0.4	3.7	2.5	3.8	3.6	3.3	2.1	3.2	4.8	3.5	4.8
Male and 00-64 years	6.2	5.2	0.4	3.6	2.6	4.0	3.7	3.6	2.1	3.3	4.8	3.6	4.9
Female	4.6	4.0	0.5	3.8	2.0	4.0	3.1	2.6	1.7	2.9	4.1	3.3	4.5
Female and 16-24 years	11.3	9.4	0.2	7.5	5.1	6.8	5.8	6.6	3.7	5.7	6.6	7.3	7.5
Female and 25-49 years	6.4	5.9	0.4	4.8	2.7	6.8	4.2	3.7	2.1	4.0	8.2	4.2	7.3
Female and 50-64 years	9.6	7.7	3.4	9.1	4.6	8.9	5.7	4.9	3.4	4.5	14.7	6.0	15.3
Female and 16-64 years	5.3	4.5	0.5	4.0	2.4	4.4	3.4	3.1	1.8	3.0	5.1	3.6	5.1
Female and 00-64 years	5.7	4.8	0.5	3.9	2.5	4.6	3.5	3.2	1.8	3.2	5.3	3.7	5.3
<b>S<sub>80</sub>/S<sub>20</sub></b>	<b>2.1</b>	<b>2.9</b>	<b>1.8</b>	<b>3.2</b>	<b>1.6</b>	<b>1.0</b>	<b>1.9</b>	<b>3.7</b>	<b>1.6</b>	<b>2.0</b>	<b>4.2</b>	<b>3.5</b>	<b>1.5</b>
<b>Relative poverty gap</b>	<b>6.9</b>	<b>5.6</b>	<b>4.1</b>	<b>5.4</b>	<b>3.0</b>	<b>4.8</b>	<b>4.3</b>	<b>4.6</b>	<b>2.7</b>	<b>3.7</b>	<b>4.8</b>	<b>4.3</b>	<b>7.3</b>
<b>Gini coefficient</b>	<b>1.7</b>	<b>1.6</b>	<b>1.7</b>	<b>1.5</b>	<b>0.9</b>	<b>0.8</b>	<b>1.6</b>	<b>1.5</b>	<b>0.9</b>	<b>1.5</b>	<b>4.1</b>	<b>1.7</b>	<b>1.2</b>
<b>Mean income</b>	<b>0.9</b>	<b>1.0</b>	<b>0.6</b>	<b>1.3</b>	<b>0.7</b>	<b>0.2</b>	<b>0.8</b>	<b>1.0</b>	<b>0.6</b>	<b>0.9</b>	<b>1.5</b>	<b>1.9</b>	<b>0.5</b>

**Table 2: Estimated Design Effects**

	AT	BE	DK	EE	ES	FI	FR	GR	IT	IE	NO	PT	SE
<b>Poverty threshold</b>	<b>1.00</b>	<b>1.21</b>	<b>0.85</b>	<b>1.08</b>	<b>1.82</b>	<b>1.14</b>	<b>1.22</b>	<b>1.27</b>	<b>1.60</b>	<b>1.34</b>	<b>1.00</b>	<b>2.20</b>	<b>0.98</b>
<b>Poverty rate</b>	<b>1.00</b>	<b>1.04</b>	<b>0.84</b>	<b>1.10</b>	<b>1.43</b>	<b>1.40</b>	<b>1.13</b>	<b>1.15</b>	<b>1.41</b>	<b>1.30</b>	<b>1.00</b>	<b>1.22</b>	<b>0.96</b>
Male	1.00	1.06	0.85	1.06	1.51	1.41	1.09	1.21	1.55	1.28	1.00	1.39	0.97
Male and 16-24 years	1.00	0.90	0.85	1.07	1.39	1.50	0.99	1.11	1.57	1.39	1.00	1.06	0.98
Male and 25-49 years	1.00	1.10	0.82	1.10	1.45	1.37	1.15	1.14	1.56	1.31	1.00	1.18	1.00
Male and 50-64 years	1.00	1.01	0.87	1.08	1.39	1.23	0.98	1.03	1.36	1.22	1.00	1.07	1.02
Male and 16-64 years	1.00	0.99	0.84	1.09	1.45	1.36	1.03	1.12	1.63	1.33	1.00	1.05	0.98
Male and 00-64 years	1.00	1.07	0.86	1.06	1.52	1.37	1.04	1.27	1.71	1.28	1.00	1.10	0.99
Female	1.00	1.00	0.83	1.12	1.37	1.42	1.13	1.09	1.35	1.31	1.00	1.32	0.95
Female and 16-24 years	1.00	0.91	0.82	1.15	1.34	1.48	1.03	1.12	1.29	1.16	1.00	1.21	0.98
Female and 25-49 years	1.00	1.15	0.82	1.11	1.42	1.35	1.07	1.13	1.46	1.28	1.00	1.16	0.99
Female and 50-64 years	1.00	0.99	0.79	1.09	1.37	1.29	1.05	1.08	1.27	1.28	1.00	1.12	1.03
Female and 16-64 years	1.00	1.00	0.82	1.10	1.39	1.39	1.10	1.10	1.30	1.26	1.00	1.17	0.97
Female and 00-64 years	1.00	1.03	0.84	1.07	1.37	1.32	1.06	1.13	1.33	1.30	1.00	1.30	0.98
<b>S<sub>80</sub>/S<sub>20</sub></b>	<b>1.00</b>	<b>1.04</b>	<b>0.94</b>	<b>1.15</b>	<b>1.63</b>	<b>0.82</b>	<b>0.99</b>	<b>1.18</b>	<b>1.57</b>	<b>1.15</b>	<b>1.00</b>	<b>1.51</b>	<b>0.98</b>
<b>Relative poverty gap</b>	<b>1.00</b>	<b>1.07</b>	<b>0.84</b>	<b>1.07</b>	<b>1.47</b>	<b>1.42</b>	<b>1.12</b>	<b>1.35</b>	<b>1.58</b>	<b>1.29</b>	<b>1.00</b>	<b>0.98</b>	<b>0.96</b>
<b>Gini coefficient</b>	<b>1.00</b>	<b>1.03</b>	<b>0.96</b>	<b>1.26</b>	<b>1.69</b>	<b>0.78</b>	<b>0.98</b>	<b>1.23</b>	<b>1.53</b>	<b>1.10</b>	<b>1.00</b>	<b>1.36</b>	<b>0.99</b>
<b>Mean income</b>	<b>1.00</b>	<b>1.37</b>	<b>0.92</b>	<b>1.16</b>	<b>2.09</b>	<b>0.78</b>	<b>1.12</b>	<b>1.30</b>	<b>1.62</b>	<b>1.24</b>	<b>1.00</b>	<b>2.46</b>	<b>0.98</b>

Most of the results stated above can be expected considering the sampling designs implemented in each country. National differences can be explained by the effective sample sizes as well as most of the loss of accuracy for the domain estimates. However, it is worth pointing out other factors as they may have a strong impact on variance:

- **Clustering:** it will harm the quality of Belgium, Spain, France, Greece, Italy and Portugal's SILC data.
- **Calibration:** in our calculations, calibration always makes the accuracy better, but the impact depends on the calibration model. For example, the low CVs for Denmark are due to a powerful model (poverty data from registers).
- Other factors such as **random weighting**, which increases Deff (see Kish, 1965) or the **allocation of the sample** in case of stratified sampling. The latter is likely to explain the Deff values for Finland, and particularly the differences between the Gini coefficient and the mean income on one hand (low Deff values) and the poverty rates on the other (high Deff values).

## 5. Some issues

Even though the variance estimates set out in the previous table seem worthy of consideration, the approach has some drawbacks:

- The variance estimates may be sensitive to outliers in the income distributions
- The variance estimates may be sensitive to income density estimation
- Imputation is not taken into account

### 5.1. Impact of outliers

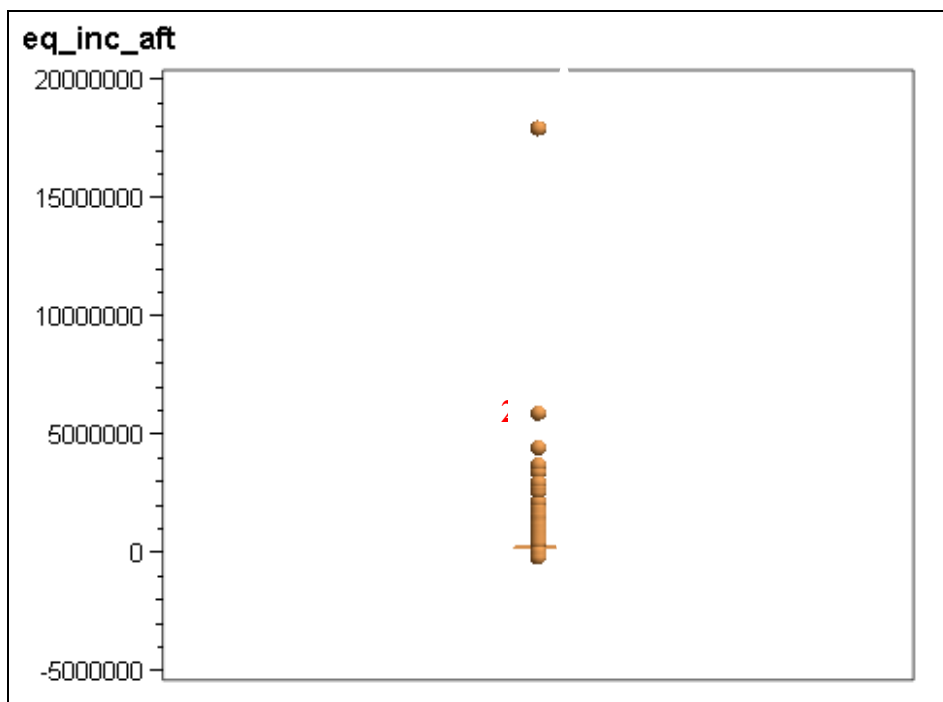
Outliers in income distributions can be problematic as they can make variance estimates less stable, especially for the Gini coefficient, the mean income and the income quintile share ratio  $S_{80}/S_{20}$ . The variance estimates for Norway are the most striking illustration of the impact outliers can have.

**Table 3: Summary statistics of Norway's income distribution**

N	Minimum	Maximum	Mean	Std Dev	CV(%)
15868	-197777.14	17939489.33	247440.29	363844.43	147

The income distribution has proved to have some outliers, as the next figure shows. The high value (147%!) for the Coefficient of variation is explained by those outlying observations.

**Figure 1: Boxplot of Norway's income distribution**



After recoding the three outliers and running variance estimation, the following results have been obtained:

**Table 4: Impact of recoding the outliers on CVs**

	<b>Before recoding</b>	<b>After recoding</b>
<b>Poverty threshold</b>	<b>0.5</b>	<b>0.5</b>
<b>Poverty rate</b>	<b>3.5</b>	<b>3.5</b>
<b>Relative Poverty Gap</b>	<b>4.8</b>	<b>4.7</b>
<b>Gini coefficient</b>	<b>4.1</b>	<b>2.2</b>
<b>Mean income</b>	<b>1.5</b>	<b>0.9</b>
<b>S<sub>80</sub>/S<sub>20</sub></b>	<b>4.2</b>	<b>2.3</b>

The impact of recoding appears to be not significant on the poverty threshold; the poverty rate and the relative median poverty gap whereas it is quite strong on the Gini coefficient, the mean income and the income quintile share ratio.

This result can be expected as the variances of these three indicators highly depend on the dispersion of the income variable. For instance, consider the mean income. Let  $y_i$  denote the income of a unit  $i$  and  $N$  the size of the target population. A "linearized" variable at  $k$  for the mean income  $\bar{Y}$  is:

$$z_k = \frac{1}{N} (y_k - \bar{Y}) \quad (13)$$

Under Simple Random Sampling without replacement of size  $n$ , the variance is:

$$Var(\hat{\bar{Y}}) \approx \left(1 - \frac{n}{N}\right) \frac{S^2}{n} \quad (14)$$

Where  $S^2$  is the dispersion of the income variable.

Variance formulas, highlighting the effect of the dispersion of the income variable, can be worked out for the Gini coefficient and the  $S_{80}/S_{20}$  as well.

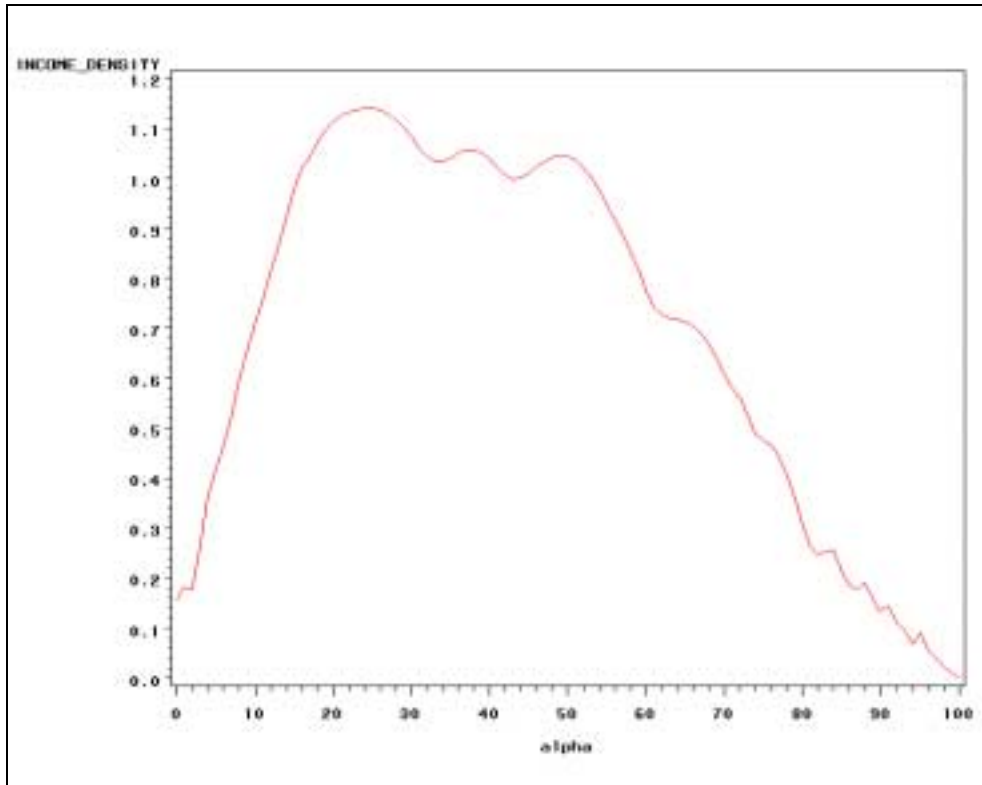
## 5.2. Impact of income density estimation

In many linearization formulas, income densities are needed. For instance, a "linearized" variable at  $k$  for the  $\alpha\%$  - income quantile  $q_\alpha$  is:

$$z_k = -\frac{1}{f(q_\alpha)} \times \frac{1}{N} \times \left[ I(INC_k \leq q_\alpha) - \frac{\alpha}{100} \right] \quad (13)$$

where  $f(q_\alpha)$  denotes the income density at  $q_\alpha$ . In practice, income densities have to be estimated. Estimated income densities can be pretty instable, especially on the tails of an income distribution, as the next figure shows.

**Figure 2: Income density plot (source: Portugal's EU-SILC data)**



Consequently, the impact of income density estimation may be strong on variance estimates for the income quintile share ratio  $S_{80}/S_{20}$  or the relative median poverty gap. Conversely, the impact ought to be smaller for the poverty threshold as a median income is used.

### 5.3. Imputation

Imputation has been resorted to in order to correct individual nonresponse in the income components. The variance calculations have been carried out without taking imputation into account. This naïve approach is clearly flawed as imputation has an impact on variance. Consider the **Imputed Horvitz-Thompson estimator** of the total  $t_y$  of a variable  $y$ :

$$\hat{t}_{y,imp} = \sum_{k \in s} \frac{y_{\bullet k}}{\pi_k} = \sum_{k \in s_r} \frac{y_k}{\pi_k} + \sum_{k \in s_m} \frac{\hat{y}_k}{\pi_k} \quad (14)$$

For a non-responding unit  $k$ , a value  $\hat{y}_k$  is calculated on the basis of an imputation model  $M$  and substituted for the exact value  $y_k$ . A (deterministic) imputation model can be stated in the following form:

$$(M): \hat{y}_k = \tilde{y}_k + u_k$$

$$\begin{cases} E_M(u_k) = 0 \\ V_M(u_k) = \sigma_k^2 \\ Cov_M(u_k, u_l) = 0 \end{cases} \quad (15)$$

In the following calculations, we assume that non-response is ignorable<sup>1</sup>

**\*\*\* Result 1 \*\*\***

Under a model-based approach, the variance of the Imputed Horvitz-Thompson estimator is given by:

$$Var(\hat{t}_{y,imp}) \approx Var\left(\sum_{k \in s_r} \frac{y_k}{\pi_k} + \sum_{k \in s_m} \frac{\tilde{y}_k}{\pi_k}\right) + (1 - \bar{\theta}) \cdot \sum_{i \in U} \frac{\sigma_i^2}{\pi_i} \quad (16)$$

Where  $\bar{\theta}$  denotes the mean response probability and  $\tilde{y}_k = E_M(\hat{y}_k)$

**\*\*\* Result 2 \*\*\***

Consider now the Imputed Horvitz-Thompson variance estimator:

$$\hat{V}_{HT,imp} = \sum_{i \in s} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_j} \cdot y_{\cdot i} \cdot y_{\cdot j} \quad (17)$$

This is the classical variance estimator that would be obtained considering the imputed values as the exact values. This estimator has been implicitly used in Eurostat's calculations as no specific procedure has been introduced for dealing with imputation. Then, we have:

$$\begin{aligned} E\hat{V}_{HT,imp} &\approx Var\left(\sum_{k \in s_r} \frac{y_k}{\pi_k} + \sum_{k \in s_m} \frac{\tilde{y}_k}{\pi_k}\right) + (1 - \bar{\theta}) \cdot \sum_{i \in U} \frac{(1 - \pi_i) \cdot \sigma_i^2}{\pi_i} \\ &\approx Var(\hat{t}_{y,imp}) \end{aligned} \quad (18)$$

As a conclusion, we can say the estimator (17) is almost unbiased for the variance of the Imputed Horvitz-Thompson estimator. In general, the variance (16) is greatly inferior to that of the Horvitz-Thompson estimator based on the exact values. This result is generally observed with deterministic imputation methods. To avoid it, one would prefer random imputation methods, which will reduce variance distortion but, on the other hand, may increase bias burden.

## 6. Conclusion

The approach for variance estimation that was presented in this paper has three main advantages:

<sup>1</sup> i.e. non-correlated with the target survey variables

- It is easily reproducible at European level
- Poulpe is able to deal with many EU-SILC sampling designs
- It has strong theoretical foundations

Nevertheless, despite all the possibilities Poulpe offers, it cannot easily deal with “highly” complex sample designs. For instance, variance estimation taking a rotational design into account (each year, one sub-sample is dropped and a new one is substituted for) is conceptually much more difficult to handle and Poulpe will rely on approximation formulas. The approach might become less appealing then.

## ***References***

Ardilly, P. (2006), *Les Techniques de Sondage*, 2<sup>nd</sup> Edition, Technip.

Ardilly, P. (2005), "Utilisation d'un modèle pour apprécier la pertinence d'une pondération", paper presented at the Colloque Francophone sur les Sondages, Québec, Canada.

Binder, D. A. and Patak, Z. (1994), "Use of Estimating Functions for Estimation From Complex Surveys", *Journal of the American Statistical Association*, Vol. 89, No. 427, pp. 1035-1043.

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

Deville, J. -C. (1999), "Variance estimation for complex statistics and estimators: linearization and residual techniques", *Survey Methodology*, Vol. 25, No. 02.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

Särndal, K. -E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer.

Tillé, Y. (2000), *Echantillonnage et Estimation en Populations Finies*, Dunod.

Woodruff, R. (1971), "A Simple Method for approximating the Variance of a Complicated Estimate", *Journal of the American Statistical Association*, Vol. 66, No. 334, pp. 411-414.

**Appendix: linearization formulas****1. At-risk-of-poverty threshold (ARPT)****1.1. Expression of the indicator as a functional of the measure M**

The at-risk-of-poverty threshold, which is 60% of the median "equivalised" income", can be expressed as:

$$\text{ARPT} = T(M) = 0.6 \times q(M)$$

Where  $q(M)$  stands for the median income.

Let  $C$  denote the subpopulation, of size  $N_C$ , in case of breaking down (If no breaking down, we have  $C=U$ =whole population). Let  $G(M,x)$  the following function:

$$G(M,x) = \frac{1}{N_C} \sum_{i \in U} 1(i \in C) \cdot 1(\text{INC}_i \leq x)$$

Then, the median "equivalised" income  $q(M)$  meets  $G[M, q(M)] = 0.5$

**1.2. Computation of the influence function**

Case 1 (ideal case): we assume that the function  $\tilde{G} : x \mapsto G(M,x)$  is derivable and strictly non-negative.

Then, we have:

$$G[M, q(M)] = 0.5 \Rightarrow IG_k[M, q(M)] = 0$$

$$\Rightarrow \left[ \frac{dG(M,x)}{dx} \Big|_{x=q(M)} \right] \times I_{q_k}(M) + IG_k[M, q(M) | q(M) \text{ fixed}] = 0$$

$$\Rightarrow \tilde{G}'[q(M)] \times I_{q_k}(M) + \frac{1(k \in C)}{N_C} [1(\text{INC}_k \leq q(M)) - 0.5] = 0$$

So, the influence function at  $k$  of  $q$  is equal to:

$$I_{q_k}(M) = -\frac{1}{\tilde{G}'[q(M)]} \times \frac{1(k \in C)}{N_C} \times [I(\text{INC}_k \leq q(M)) - 0.5]$$

Case 2: the function  $\tilde{G} : x \mapsto G(M, x)$  is not generally derivable so the previous formula cannot be applied. The solution consists in “regularizing”  $\tilde{G}$  through Gaussian kernel estimation:

$$\tilde{G}_k(x) = \int \tilde{G}(t) \cdot K(x, t) dt$$

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

It can be easily seen that  $\tilde{G}_k$  is derivable and:

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

In conclusion, the influence function at k of the functional q is equal to:

$$I_{q_k}(M) = -\frac{1}{\tilde{G}_k'[q(M)]} \times \frac{1(k \in C)}{N_C} \times [I(\text{INC}_k \leq q(M)) - 0.5]$$

Finally, the influence function at k for the at-risk-of-poverty threshold is given by:

$$\begin{aligned} \text{IARPT}_k(M) &= 0.6 \times I_{q_k}(M) \\ &= -\frac{0.6}{\tilde{G}_k'[q(M)]} \times \frac{1(k \in C)}{N_C} \times [I(\text{INC}_k \leq q(M)) - 0.5] \end{aligned}$$

## 2. At-risk-of-poverty rate (ARPR)

### 2.1. Expression of the indicator as a functional of M

Using the notations defined in the previous section, we have:

$$\text{ARPR} = T(M) = 100 \cdot G[M, \lambda(M)]$$

- $G(M, X) = \frac{1}{N_C} \sum_{i \in U} 1(i \in C) \cdot 1(\text{INC}_i \leq X)$
- $\lambda(M)$  is the poverty threshold (no breaking down)

## 2.2. Computation of the influence function

The method is the same as the poverty threshold's one.

If the function  $\tilde{G} : x \mapsto G(M, x)$  is derivable and strictly non-negative (ideal case), we have:

$$\begin{aligned} IT_k(M) &= 100 \cdot IG_k[M, \lambda(M)] \\ &= 100 \cdot \left( IG_k[M, \lambda(M) \mid \lambda(M) \text{ fixed}] + \tilde{G}'[\lambda(M)] \cdot I\lambda_k(M) \right) \\ &= 100 \cdot \left( \frac{1(k \in C)}{N_C} \cdot \{ 1[INC_k \leq \lambda(M)] - G[M, \lambda(M)] \} + \tilde{G}'[\lambda(M)] \cdot I\lambda_k(M) \right) \end{aligned}$$

The influence function  $I\lambda_k(M)$  of  $\lambda$  (poverty threshold) at  $k$  is computed according to the formula given in the previous paragraph.

Generally, the function  $\tilde{G} : x \mapsto G(M, x)$  is not derivable so should be “regularized” through Gaussian kernel estimation. After calculations and by replacing  $\tilde{G}$  by its corresponding “regularized” function  $\tilde{G}_k$ , we obtain the influence function at  $k$  for the at-risk-of-poverty rate:

$$IT_k(M) = 100 \cdot \frac{1(k \in C)}{N_C} \cdot \{ 1[INC_k \leq \lambda(M)] - (\text{poverty\_rate}) \} + 100 \cdot \tilde{G}_k'[\lambda(M)] \cdot I\lambda_k(M)$$

## 3. Income quantile share ratio (S80/S20)

### 3.1. Expression of the indicator as a functional of M

Let  $q_-(M)$  (respectively  $q_+(M)$ ) denotes the bottom quantile (respectively the top quantile). Then, the indicator is equal to:

$$\begin{aligned} S80/S20 = T(M) &= \frac{\int INC_k \cdot 1[INC_k > q_+(M)] \cdot dM(k)}{\int INC_k \cdot 1[INC_k \leq q_-(M)] \cdot dM(k)} \\ &= \frac{\int INC_k \cdot dM(k) - \int INC_k \cdot 1[INC_k \leq q_+(M)] \cdot dM(k)}{\int INC_k \cdot 1[INC_k \leq q_-(M)] \cdot dM(k)} \\ &= \frac{R(M) - S[M, q_+(M)]}{S[M, q_-(M)]} \end{aligned}$$

Where:

- $S(M, X) = \int INC_k \cdot 1[INC_k \leq X] \cdot dM(k)$
- $R(M) = \int INC_k \cdot dM(k)$

### 3.2. Computation of the influence function

$$\text{Let } A(M) = \frac{1}{S[M, q_-(M)]} \quad \text{and} \quad B(M) = R(M) - S[M, q_+(M)]$$

Then, the income quantile share ratio is  $\theta = T(M) = A(M) \cdot B(M)$ .

The corresponding influence function is:

$$\begin{aligned} IT_k(M) &= A(M) \cdot IB_k(M) + B(M) \cdot IA_k(M) \\ &= \frac{1}{S[M, q_-(M)]} \cdot [IR_k(M) - IS_k[M, q_+(M)]] - \frac{R(M) - S[M, q_+(M)]}{S^2[M, q_-(M)]} \cdot IS_k[M, q_-(M)] \\ &= \frac{INC_k}{S[M, q_-(M)]} - \frac{1}{S[M, q_-(M)]} \cdot IS_k[M, q_+(M)] - \frac{R(M) - S[M, q_+(M)]}{S^2[M, q_-(M)]} \cdot IS_k[M, q_-(M)] \end{aligned}$$

The main difficulty consists in linearizing the functional  $S(M)$ .

Case 1: the function  $\tilde{S} : x \mapsto S(M, x)$  is assumed to be derivable and strictly non-negative. Then we have:

$$\begin{aligned} &IS_k[M, q(M)] \\ &= IS_k[M, q(M) \mid q(M) \text{ fixed}] + \tilde{S}'[q(M)] \cdot Iq_k(M) \\ &= INC_k \cdot 1[INC_k \leq q(M)] + \tilde{S}'[q(M)] \cdot Iq_k(M) \end{aligned}$$

Case 2: the function  $\tilde{S} : x \mapsto S(M, x)$  is assumed not to be derivable so it should be “regularized” through Gaussian kernel estimation. We obtain after calculations:

$$\tilde{S}'_k(M, x) = \frac{1}{h\sqrt{2\pi}} \cdot \sum_{i \in U} INC_i \cdot \exp\left[-\frac{(INC_i - x)^2}{2h^2}\right]$$

In conclusion:

$$\boxed{IS_k[M, q(M)] = INC_k \cdot 1[INC_k \leq q(M)] + \tilde{S}'_k[q(M)] \cdot Iq_k(M)}$$

#### 4. Relative median at-risk-of-poverty gap (RMPG)

##### 4.1. Expression of the indicator as a functional of M

The indicator is defined as the difference between the at-risk-of-poverty threshold and the median income of the “poor” people, taken relatively to the at-risk-of-poverty threshold:

$$\begin{aligned}
 \text{RMPG} &= 100 \times \left( \frac{\text{ARPT} - \text{MEDIAN}_{\text{poor}}}{\text{ARPT}} \right) \\
 &= 100 \times \left( 1 - \frac{\text{MEDIAN}_{\text{poor}}}{\text{ARPT}} \right) \\
 &= 100 - 100 \times \frac{\text{MEDIAN}_{\text{poor}}}{\text{ARPT}} \\
 &= 100 - 100 \times \frac{A(M)}{B(M)} = T(M)
 \end{aligned}$$

##### 4.2. Computation of the influence function

$$\text{IT}_k(M) = -100 \times \left[ -\frac{A(M)}{B^2(M)} \text{IB}_k(M) + \frac{1}{B(M)} \text{IA}_k(M) \right]$$

The formula for computing  $\text{IB}_k(M)$  for all  $k$  is given in the first paragraph (linearization of the at-risk-of-poverty poverty threshold).

As regards the functional  $A(M)$ , it meets the following equality:

$$G[M, A(M)] = \frac{1}{2} \cdot G[M, B(M)]$$

Where:

- $G(M, X) = \frac{1}{N_C} \sum_{i \in U} 1(i \in C) \cdot 1(\text{INC}_i \leq X)$
- $C$  denotes the sub-population in case of breaking down of the median gap. If no breakdown,  $C=U$ =total population
- $N_C$  is the size of the subpopulation  $C$

So, we obtain as influence function for A:

$$IG_k[M, A(M)] = \frac{1}{2} IG_k[M, B(M)]$$

$$IG_k[M, A(M) | A(M) \text{ fixed}] + \tilde{G}'_k[A(M)] \times IA_k(M) = \frac{1}{2} \cdot \{ IG_k[M, B(M) | B(M) \text{ fixed}] + \tilde{G}'_k[B(M)] \times IB_k(M) \}$$

$$\Rightarrow IA_k(M) = \frac{1}{\tilde{G}'_k[A(M)]} \cdot \left\{ \frac{1}{2} \cdot [IG_k[M, B(M) | B(M) \text{ fixed}] + \tilde{G}'_k[B(M)] \times IB_k(M)] - IG_k[M, A(M) | A(M) \text{ fixed}] \right\}$$

All the terms in the above expression can easily be computed.

## 5. Gini index

### 5.1. Expression of the indicator as a functional of M

$$\begin{aligned} 1 + G &= \frac{2 \times \sum_{i \in U} r_i \cdot INC_i - \sum_{i \in U} INC_i}{N \cdot \sum_{i \in U} INC_i} \\ &= \frac{2 \times \int INC_i \cdot \left[ \int 1(INC_k \leq INC_i) \cdot dM(k) \right] \cdot dM(i) - \int INC_i \cdot dM(i)}{\int dM(i) \cdot \int INC_i \cdot dM(i)} \\ &= T(M) \end{aligned}$$

### 5.2. Computation of the influence function

Let us denote:

$$\begin{aligned} - \quad T_1(M) &= T_1 = \int_{k \in U} \left[ \int_{i \in U} 1(y_i \leq y_k) \cdot dM(i) \right] \cdot y_k \cdot dM(k) \\ - \quad T_2(M) &= Y = \int_{k \in U} y_k \cdot dM(k) \\ - \quad T_3(M) &= N = \int_{k \in U} dM(k) \end{aligned}$$

We can write then:  $T(M) = \frac{2T_1(M) - T_2(M)}{T_2(M) \cdot T_3(M)}$

The influence function of the functional T at k will be equal to:

$$\begin{aligned}
 IT(M, k) &= \frac{T_2(M) \cdot T_3(M) \cdot I(2T_1 - T_2) \cdot (k) - [2T_1(M) - T_2(M)] \cdot I(T_2 T_3) \cdot (k)}{[T_2(M) \cdot T_3(M)]^2} \\
 &= \frac{2NY \cdot IT_1(k) - NY \cdot IT_2(k) - (2T_1 - Y) \cdot [T_3(M) \cdot IT_2(k) + T_2(M) \cdot IT_3(k)]}{(NY)^2} \\
 &= \frac{2IT_1(k) - IT_2(k) - (G+1) \cdot [T_3(M) \cdot IT_2(k) + T_2(M) \cdot IT_3(k)]}{NY} \\
 &= \frac{2IT_1(k) - IT_2(k) - (G+1) \cdot [N \cdot IT_2(k) + Y \cdot IT_3(k)]}{NY} \quad (*)
 \end{aligned}$$

Let us now consider separately each functional  $T_1$ ,  $T_2$  and  $T_3$ . We easily verify whatever k:

- $IT_2(k) = y_k$
- $IT_3(k) = 1$

Substituting  $IT_2(k)$  and  $IT_3(k)$  in (\*) with the two previous values we have:

$$IT(M, k) = \frac{2IT_1(k) - y_k - (G+1) \cdot (Y + N \cdot y_k)}{NY}$$

The main remaining difficulty consists of calculating the influence function  $IT_1(k)$  of the functional  $T_1$

$$\begin{aligned}
 T_1(M + t\delta_k) &= \int_{i \in U} \left[ \int_{j \in U} 1(y_j \leq y_i) \cdot d(M + t\delta_k) \cdot (j) \right] \cdot y_i \cdot d(M + t\delta_k) \cdot (i) \\
 &= \int_{i \in U} \left[ \int_{j \in U} 1(y_j \leq y_i) \cdot dM(j) + \int_{j \in U} 1(y_j \leq y_i) \cdot d(t\delta_k) \cdot (j) \right] \cdot y_i \cdot d(M + t\delta_k) \cdot (i)
 \end{aligned}$$

Hence, we have:

$$\begin{aligned}
 & T_1(\mathbf{M} + t\delta_k) \\
 &= \int \left[ \int_{j \in U} 1(y_j \leq y_i) \cdot d\mathbf{M}(j) \right] \cdot y_i \cdot d(\mathbf{M} + t\delta_k) \cdot (i) + \int \left[ \int_{j \in U} 1(y_j \leq y_i) \cdot d(t\delta_k) \cdot (j) \right] \cdot y_i \cdot d(\mathbf{M} + t\delta_k) \cdot (i) \\
 &= T_1(\mathbf{M}) + \int \left[ \int_{j \in U} 1(y_j \leq y_i) \cdot d\mathbf{M}(j) \right] \cdot y_i \cdot d(t\delta_k) \cdot (i) + \int \left[ \int_{j \in U} 1(y_j \leq y_i) \cdot d(t\delta_k) \cdot (j) \right] \cdot y_i \cdot d\mathbf{M}(i) \\
 &\quad + \int \left[ \int_{j \in U} 1(y_j \leq y_i) \cdot d(t\delta_k) \cdot (j) \right] \cdot y_i \cdot d(t\delta_k) \cdot (i) \\
 &= T_1(\mathbf{M}) + t \cdot y_k \cdot \left[ \int_{j \in U} 1(y_j \leq y_k) \cdot d\mathbf{M}(j) \right] + t \cdot \int_{i \in U} 1(y_k \leq y_i) \cdot y_i \cdot d\mathbf{M}(i) + t^2 \cdot y_k
 \end{aligned}$$

Hence, we get the influence function for the functional  $T_1$ :

$$IT_1(k) = y_k \cdot \int_{j \in U} 1(y_j \leq y_k) \cdot d\mathbf{M}(j) + \int_{j \in U} y_j \cdot 1(y_k \leq y_j) \cdot d\mathbf{M}(j) = y_k \cdot \sum_{i \in U} 1(y_i \leq y_k) + \sum_{i \in U} y_i \cdot 1(y_i \geq y_k)$$

Finally, the influence function at  $k$  for the Gini index is equal to:

$$IT(\mathbf{M}, k) = \frac{2 \cdot \left[ y_k \cdot \sum_{i \in U} 1(y_i \leq y_k) + \sum_{i \in U} y_i \cdot 1(y_i \geq y_k) \right] - y_k - (G+1) \cdot (Y + N \cdot y_k)}{NY}$$