CALCUL DE PRECISION DANS LE CAS D'ECHANTILLONS ROTATIFS : L'EXEMPLE DES STATISTIQUES EU-SILC AU LUXEMBOURG

Guillaume Osier[1]

[1] *Institut National de la Statistique et des Etudes Economiques du Grand-Duché de Luxembourg (STATEC) - Guillaume.Osier@statec.etat.lu*

Résumé. Cette contribution propose une formule d'estimation permettant le calcul des marges d'erreurs dans le cas d'échantillons rotatifs. Cette formule est à la fois justifiée sur le plan théorique et facile à mettre en pratique à l'aide des logiciels standards d'analyse statistique comme SAS, Stata, SPSS ou R. Elle peut être étendue aux indicateurs longitudinaux, aux indicateurs de changement ainsi qu'aux statistiques non linéaires *via* la technique de linéarisation. Des résultats numériques obtenus à partir des statistiques européennes sur les revenus et les conditions de vie (EU-SILC) sont fournis à la fin de la contribution.

Mots-clés. Enquêtes sociales, calcul de variance, marge d'erreur, intervalle de confiance

Abstract. This contribution proposes a variance estimator in case of rotating panels. The formula is both theoretically justified and easy to implement using standard statistical software such as SAS, Stata, SPSS or R. The estimator can be extended to deal with longitudinal indicators, indicators of net changes and also non-linear indicators through the linearisation technique. Numerical results based on the European Statistics on Income and Living Conditions (EU-SILC) are given at the end of the paper.

Keywords. Social surveys, variance estimation, margin of error, confidence interval

1. Introduction

EU-SILC est aujourd'hui la source de référence pour des micro-données comparables au niveau européen sur les revenus et les conditions de vie des ménages et des personnes. L'enquête EU-SILC au Luxembourg (ci-après dénommée "LU-SILC") est conduite chaque année par le STATEC en collaboration avec le LISER (*Luxembourg Institute of Socio-Economic Research*). LU-SILC permet de construire les indicateurs de référence sur la pauvreté, les inégalités et l'exclusion sociale au Luxembourg.

En 2016, un échantillon aléatoire simple stratifié de 5500 personnes âgées de 18 ans ou plus a été tiré dans le Registre National des Personnes Physiques (RNPP). Les strates étaient constituées par les 12 cantons qui partagent le territoire du Grand-Duché, le canton de Luxembourg étant lui-même subdivisé entre la ville de Luxembourg ("Luxembourg-ville") et le reste du canton ("Luxembourg-campagne"). Les personnes sélectionnées ont ensuite été contactées et leur ménage ainsi que l'ensemble de ses membres ont été interrogés dans le cadre de la collecte. Une procédure analogue a été utilisée au cours des années précédentes pour sélectionner l'échantillon LU-SILC, à la différence près que les fichiers de l'Inspection Générale de la Sécurité Sociale (IGSS) ont servi de base de sondage pour le tirage à la place du RNPP. Les unités d'échantillonnage étaient alors constituées de ménages "fiscaux" rassemblant les assurés à la Sécurité Sociale ainsi que leurs "co-assurés".

Suivant en cela les règlements européens, un panel rotatif sur 4 années a été mis en place pour la collecte des données LU-SILC. Cela implique que tous les membres des ménages qui ont été tirés dans l'échantillon sont ensuite suivis et réinterrogés au cours des trois vagues (années) suivantes avant d'être finalement retirés et remplacés par un nouveau sous-échantillon. Les individus qui sont tirés pour la première fois constituent le sous-échantillon entrant, tandis que ceux qui sont réinterrogés forment la composante "panel" de l'échantillon. De cette façon, 25% de l'échantillon est renouvelé à chaque nouvelle année.

2. L'approche générale

2.1. Définitions et notations

Soit $s_\tau^{P,i}$ l'échantillon d'individus "panel" qui a été séléctionné l'année τ dans le sous-échantillon i $(i = 1 \dots 4)$ et soit $s_\tau^{M,i}$ l'échantillon des ménages privés dont au moins un des membres appartient à l'échantillon d'individus "panel" $s_\tau^{P,i}$. En outre, on désigne par $s_\tau^{A,i}$ l'ensemble de tous les membres des ménages de l'échantillon $s_\tau^{M,i}$. Sur la base de ces notations, l'échantillon individuel transversal à l'année τ est donné par[1] $\tilde{s}_\tau = \cup_{i=1}^{4} s_\tau^{A,i}$. On note aussi $\tilde{s}_\tau^{M} = \cup_{i=1}^{4} s_\tau^{M,i}$ et $\tilde{s}_\tau^{P} = \cup_{i=1}^{4} s_\tau^{P,i}$.

Soit $Y = \sum_{k \in U} y_k$ le total de la variable y sur la population U, de taille N. Sauf mention contraire, U désigne l'ensemble de la population résidente au Grand-Duché à la date de l'enquête. Y représente le paramètre d'intérêt que l'on cherche à estimer.

Sur la base de l'échantillon transversal d'individus $\tilde{s}_\tau$, le total Y de y à l'année τ est estimé en prenant la somme pondérée des valeurs de y obtenues sur l'échantillon:

$$\hat{Y}_\tau = \sum_{k \in \tilde{s}_\tau} \omega_k \, y_k \qquad (1)$$

où ω_k désigne le coefficient de pondération de l'individu $k \in \tilde{s}_\tau$. Les "poids" ω_k permettent d'extrapoler à la population totale les données partielles qui ont été observées sur un échantillon. Pour cela, ces coefficients doivent tenir compte des probabilités de sélection des individus dans l'échantillon, être ajustés pour tenir compte de la non-réponse et calés afin de reproduire des agrégats tirés de sources externes (Särndal et al, 1992).

2.2. Lemme préparatoire

Dans l'objectif de produire un estimateur de la variance de $\hat{Y}_\tau$, cette formule doit être réécrite en tenant compte de la façon dont les pondérations ω_k ont été calculées.

1) Agrégation au niveau ménage

[1] On fait l'hypothèse ici qu'il n'y a pas de doublons entre les sous-échantillons

Comme les pondérations des individus sont identiques pour tous les individus d'un même ménage, la somme pondérée au niveau individuel peut être réécrite au niveau ménage:

$$\hat{Y}_\tau = \sum_{k \in \tilde{s}_\tau} \omega_k \, y_k = \sum_{h \in \tilde{s}_\tau^M} \tilde{\omega}_h \, Y_h \qquad (2)$$

où $Y_h = \sum_{k \in h} y_k$ désigne le total de y calculé sur l'ensemble des individus du ménage h. Y_h devient ainsi la nouvelle variable d'intérêt.

2) Application de la méthode du partage des poids

La méthode du partage des poids (Lavallée, 2002) a été appliquée pour le calcul des pondérations des ménages à partir des pondérations des individus "panel". Pour rappel, les individus "panel" correspondent, pour le sous-échantillon entrant, à l'ensemble des individus qui ont répondu à l'enquête. Ces individus sont suivis au cours des vagues subséquentes et permettent de définir l'échantillon de ménages à interroger: il s'agit des ménages dont au moins un des membres appartient à l'échantillon des individus "panel". La relation entre les individus "panel" et les ménages est complexe dans la mesure où plusieurs individus peuvent conduire à sélectionner les mêmes ménages (relation *many-to-one*). La méthode généralisée du partage des poids (MGPP) va permettre de construire des pondérations pour les ménages à partir des pondérations des individus "panel" en prenant en considération les liens complexes qui existent entre les deux populations.

Si $\tilde{s}_\tau^M$ désigne l'échantillon de ménages et $\tilde{s}_\tau^P$ celui des individus "panel" dont il est tiré, la relation entre les pondérations individuelles p_i et ménage $\tilde{\omega}_h$ telle que donnée par la MGPP s'écrit:

$$\tilde{\omega}_h = \frac{\sum_{i \in h} \omega'_{hi}}{\sum_{i \in h} L_{hi}} \qquad (3)$$

où pour tout individu $i \in h$:

$$\omega'_{hi} = \sum_{j \in \tilde{s}_\tau^P} p_j \, 1_{j=(h,i)} \qquad (4)$$

$1_{j=(h,i)}$ vaut 1 si l'individu "panel" j correspond à l'individu (h,i). Dans le cas contraire, il est égal à 0. Le dénominateur L_{hi} correspond au nombre de "liens", c'est à dire le nombre de bases de sondages dans lesquelles l'individu (h,i) aurait pu être tiré.

- Si l'individu était présent à chacune des quatre années alors $L_{hi} = 4$.
- Si l'individu était présent trois années sur quatre alors $L_{hi} = 3$.
- Si l'individu était présent deux années sur quatre alors $L_{hi} = 2$.
- Si l'individu était présent une année sur quatre alors $L_{hi} = 1$.

Les pondérations individuelles p_j correspondent aux poids de tirage des individus. Pour les sous-échantillons qui forment la composante "panel", ces pondérations doivent être ajustées afin de

corriger le biais causé par l'attrition. Si $\tilde{p}_j$ désigne le poids de tirage de l'individu j et r_j la probabilité de "réponse" de j, alors on a: $p_j = \frac{\tilde{p}_j}{r_j}$

En appliquant la formule de la MGPP, l'estimateur $\hat{Y}_\tau = \sum_{h \in \tilde{s}_\tau^M} \tilde{\omega}_h Y_h$ peut encore s'écrire:

$$\hat{Y}_\tau = \sum_{h \in \tilde{s}_\tau^M} \left(\frac{\sum_{i \in h} \omega'_{hi}}{\sum_{i \in h} L_{hi}} \right) Y_h = \sum_{h \in \tilde{s}_\tau^M} \frac{Y_h}{L_h} \sum_{i \in h} \omega'_{hi} = \sum_{h \in \tilde{s}_\tau^M} \sum_{i \in h} \omega'_{hi} z_{hi} \tag{5}$$

où pour tout ménage $h \in \tilde{s}_\tau^M$ et tout individu $i \in h$ on a:

$$L_h = \sum_{i \in h} L_{hi} \quad \text{et} \quad z_{hi} = \frac{Y_h}{L_h}$$

Finalement, si l'on remplace ω'_{hi} par son expression (cf. précédemment), alors on obtient:

$$\begin{aligned} \hat{Y}_\tau = \sum_{h \in \tilde{s}_\tau^M} \sum_{i \in h} \omega'_{hi} z_{hi} = \sum_{h \in \tilde{s}_\tau^M} \sum_{i \in h} \left(\sum_{j \in \tilde{s}_\tau^P} p_j \, 1_{j=(h,i)} \right) z_{hi} &= \sum_{h \in \tilde{s}_\tau^M} \sum_{i \in h} \sum_{j \in \tilde{s}_\tau^P} p_j \, 1_{j=(h,i)} z_{hi} \\ &= \sum_{j \in \tilde{s}_\tau^P} p_j \left(\sum_{h \in \tilde{s}_\tau^M} \sum_{i \in h} 1_{j=(h,i)} \, z_{hi} \right) \end{aligned} \tag{6}$$

Soit U_τ^M la population des ménages privés résidant au Grand-Duché. En posant pour un individu j, $Z_j = \sum_{h \in U_\tau^M} \sum_{i \in h} 1_{j=(hi)} \, z_{hi}$ alors on obtient le résultat suivant:

$$\begin{aligned} \hat{Y}_\tau &= \sum_{j \in \tilde{s}_\tau^P} p_j \left(\sum_{h \in \tilde{s}_\tau^M} \sum_{i \in h} 1_{j=(h,i)} \, z_{hi} \right) \\ &= \sum_{j \in \tilde{s}_\tau^P} p_j \left(\sum_{h \in U_\tau^M} \sum_{i \in h} 1_{j=(h,i)} \, z_{hi} \right) \\ &= \sum_{j \in \tilde{s}_\tau^P} p_j \, Z_j \end{aligned} \tag{7}$$

En conclusion l'estimateur $\hat{Y}_\tau$ du total Y_τ peut s'écrire comme un estimateur linéaire construit à partir de l'échantillon des individus "panel" en utilisant les pondérations correspondantes. L'intérêt de cette formule[2] est de ramener le calcul au niveau des individus, c'est à dire des unités d'échantillonnage.

[2] L'ouvrage de Lavallée parle de formule de dualité

2.3. Formule finale

En faisant l'hypothèse que les sous-échantillons sont deux à deux indépendants, la variance de l'estimateur $\hat{Y}_\tau = \sum_{j \in \tilde{s}_\tau^P} p_j Z_j$ peut s'écrire comme la somme des variances relatives à chaque sous-échantillon:

$$V\left(\hat{Y}_\tau\right) = V\left(\sum_{j \in \tilde{s}_\tau^P} p_j Z_j\right) = \sum_{i=1}^{4} V\left(\sum_{k \in s_\tau^{P,i}} p_j Z_j\right) = \sum_{i=1}^{4} V_i \qquad (8)$$

Il reste alors à estimer chacun des termes V_i. Pour cela, on va décomposer la variance totale en une variance liée au tirage et une variance liée à la non-réponse. Une démarche analogue se trouve dans Massiani (2013).

1) Cas des nouveaux entrants dans l'échantillon

A partir de 2016, le sous-échantillon d'individus entrants est sélectionné selon une procédure "indirecte" qui consiste à tirer d'abord un échantillon aléatoire simple stratifié d'individus âgés de 18 ans ou plus et à interroger l'ensemble des membres des ménages dont au moins un des membres fait partie de l'échantillon. Cette procédure de tirage est tout à fait analogue à celle qui est utilisée pour sélectionner les ménages à interroger à partir des individus "panel". La méthode du partage des poids s'applique donc là aussi. Dans ces conditions, la variance s'écrit:

$$V_i = V\left(\sum_{k \in s_\tau^{P,i}} p_j Z_j\right) = V\left(\sum_{k \in r} d_j K_j\right) \qquad (9)$$

où r désigne l'ensemble des individus de 18 ans ou plus qui ont été tirés dans l'échantillon et dont le ménage a participé à l'enquête. Le terme K_j correspond à la variable d'intérêt:

$$K_j = \sum_{h \in U_\tau^M} \sum_{i \in h} 1_{j=(h,i)} \frac{\tilde{Z}_h}{L_h} = \sum_{h \in \tilde{s}_\tau^M} \sum_{i \in h} 1_{j=(h,i)} \frac{\tilde{Z}_h}{L_h} \qquad (10)$$

où L_h désigne le nombre de personnes âgées de 18 ans ou plus dans le ménage h et $\tilde{Z}_h = \sum_{k \in h} Z_k$ est le total de la variable Z sur l'ensemble des individus du ménage h. d_j est le coefficient de pondération de l'individu j: il s'agit du poids de tirage $\tilde{d}_j$ (*design weight* en anglais) de l'individu divisé par la probabilité de réponse θ_j:

Le calcul de la variance repose ensuite sur le conditionnement par rapport à l'échantillon initial s des individus âgés de 18 ans ou plus qui ont été sélectionnés dans le registre de la population:

$$V_i = V\left(\sum_{k \in r} d_j K_j\right) = V_s\left[E\left(\sum_{k \in r} d_j K_j | s\right)\right] + E_s\left[V\left(\sum_{k \in r} d_j K_j | s\right)\right] = A + B \qquad (11)$$

A représente la variance liée au tirage de l'échantillon et B celle liée à la non-réponse. Sous l'hypothèse que la probabilité de réponse θ_k de l'individu k est connue, on peut écrire:

$$A = V_s\left[E\left(\sum_{k\in r} d_j\, K_j|s\right)\right] = V_s\left(\sum_{k\in s} \tilde{d}_j\, K_j\right) = \sum_{l=1}^{L}(1-f_l)\frac{S_l^2}{n_l} \qquad (12)$$

où S_l^2 est la dispersion de la variable d'intérêt K_j calculée sur la population U_l de la strate l (Pour rappel, les strates sont formées des 12 cantons du Grand-Duché, avec une subdivision du canton de Luxembourg entre Luxembourg-ville et Luxembourg-campagne):

$$S_l^2 = \frac{1}{N_l}\sum_{j\in U_l}\left(K_j - \overline{K}\right)^2 \qquad (13)$$

n_l est le nombre total d'individus qui a été tiré dans la strate l. Enfin, f_l correspond à ce qu'on appelle le taux de sondage, c'est à dire le rapport entre la taille de l'échantillon dans sur la strate et la taille de la population correspondante: $f_l = \frac{n_l}{N_l}$

Concernant la composante B (variance liée à la non-réponse), on doit faire l'hypothèse que la phase de non-réponse correspond à un tirage de Poisson. Cela implique chaque individu dans l'échantillon initial répond indépendamment des autres individus. Si l'on note θ_k la probabilité de réponse de l'individu k, que l'on suppose connue, on obtient alors:

$$B = E_s\left[V\left(\sum_{k\in r} d_j\, K_j|s\right)\right] = E_s\left(\sum_{k\in s}\tilde{d}_j^2\, K_j^2\,\frac{1-\theta_j}{\theta_j}\right) = \sum_{k\in U}\tilde{d}_j\, K_j^2\,\frac{1-\theta_j}{\theta_j} \qquad (14)$$

Chacun des termes A et B peut alors être estimé à partir des données observées sur l'échantillon:

$$\hat{A} = \sum_{l=1}^{L}(1-f_l)\frac{s_{l,d}^2}{n_l} \qquad (15)$$

$s_{l,d}^2$ correspond à la dispersion de la variable d'intérêt calculée sur l'échantillon (et non sur la population). L'indice d signifie que le calcul a été pondéré en utilisant les poids individuels d_j pour les individus $j \in r$.

$$s_{l,d}^2 = \frac{1}{\sum_{j\in r_l} d_j - 1}\sum_{j\in r_l} d_j\left(K_j - \overline{K}_{rl}\right)^2 \qquad (16)$$

Quant au terme de non-réponse B, on peut l'estimer par:

$$\hat{B} = \sum_{k \in r} \tilde{d}_j^2 K_j^2 \frac{1 - \hat{\theta}_j}{\hat{\theta}_j^2} = \sum_{k \in r} d_j^2 K_j^2 \left(1 - \hat{\theta}_j\right) \quad (17)$$

La variance pour le sous-échantillon entrant s'obtient donc par: $\hat{V}_i = \hat{A} + \hat{B}$.

$\hat{\theta}_j$ correspond à l'estimation de la probabilité de réponse de l'individu j. L'estimation de ces probabilités repose traditionnellement sur une modélisation logistique s'appuyant sur des prédicteurs dont les valeurs sont observées à la fois sur les individus répondants et les individus non-répondants.

2) Cas de la partie "panel" de l'échantillon

Pour les trois sous-échantillons qui sont réinterrogés, il faut ajouter une phase supplémentaire d'attrition. Le tirage des individus "panel" constitue la première phase de tirage, qui conduit à un échantillon initial $\tilde{s}_\tau^{P,i}$ d'individus. Ensuite, l'attrition conduit à la perte d'une partie des individus de l'échantillon initial. L'échantillon final d'individus "panel" tenant compte de l'attrition est $s_\tau^{P,i}$. En utilisant la même approche que précédemment, la variance peut être décomposée en une variance de tirage et une variance supplémentaire liée à l'attrition:

$$V_i = V\left(\sum_{j \in s_\tau^{P,i}} p_j Z_j\right) = V_{\tilde{s}_\tau^{P,i}}\left[E\left(\sum_{j \in s_\tau^{P,i}} p_j Z_j | \tilde{s}_\tau^{P,i}\right)\right] + E_{\tilde{s}_\tau^{P,i}}\left[V\left(\sum_{j \in s_\tau^{P,i}} p_j Z_j | \tilde{s}_\tau^{P,i}\right)\right] = V_I + V_{II} \quad (18)$$

V_I est la variance liée au tirage et V_{II} est la variance liée à l'attrition. Notons r_k la probabilité pour l'individu "panel" k de continuer à faire partie du panel sachant qu'il a été tiré à la première vague et faisons l'hypothèse supplémentaire que l'attrition suit un tirage de Poisson. La variance liée au tirage s'écrit:

$$V_I = V_{\tilde{s}_\tau^{P,i}}\left[E\left(\sum_{j \in s_\tau^{P,i}} p_j Z_j | \tilde{s}_\tau^{P,i}\right)\right] = V_{\tilde{s}_\tau^{P,i}}\left(\sum_{j \in \tilde{s}_\tau^{P,i}} \tilde{p}_j Z_j\right) \quad (19)$$

On se ramène ainsi à un calcul de variance sur un sous-échantillon entrant. On peut alors appliquer la formule de décomposition présentée dans la section précédente:

$$\hat{V}_I = \hat{A} + \hat{B} \quad (20)$$

où $\hat{A}$ et $\hat{B}$ sont respectivement les variances d'échantillonnage et de non-réponse.

Cette formule n'est valable pour les sous-échantillons qui ont été tirés à partir de 2016. Pour les individus qui ont été tirés avant 2016, elle n'est plus valable, puisque le mode de tirage a changé entre 2015 et 2016. Comme alternative, on peut appliquer la formule suivante (Osier et al., 2013):

$$\hat{V}_{\tilde{s}_\tau^{P,i}}\left(\sum_{j\in\tilde{s}_\tau^{P,i}}\tilde{p}_j\,Z_j\right)=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(Z_{hi\cdot}-Z_{h\cdot\cdot})^2 \qquad (21)$$

Cette formule repose sur l'hypothèse que les ménages ont été tirés avec remise à l'intérieur de chaque strate, ce qui simplifie considérablement les calculs. La variance est alors estimée en calculant les agrégats de la variable d'intérêt sur chaque ménage, et en prenant la dispersion de ces agrégats entre les ménages.

Concernant la variance V_{II} due à l'attrition, elle peut être estimée en faisant l'hypothèse que l'attrition correspond à une phase supplémentaire de tirage et qu'elle suit un modèle de Poisson. On obtient alors:

$$V_{II}=E_{\tilde{s}_\tau^{P,i}}\left[V\left(\sum_{j\in s_\tau^{P,i}}p_j\,Z_j|\,\tilde{s}_\tau^{P,i}\right)\right]=E_{\tilde{s}_\tau^{P,i}}\left(\sum_{j\in\tilde{s}_\tau^{P,i}}\tilde{p}_j^2\,Z_j^2\frac{1-r_j}{r_j}\right)=\sum_{j\in U}\tilde{p}_j\,Z_j^2\frac{1-r_j}{r_j} \qquad (22)$$

La variance est alors estimée par:

$$\hat{V}_{II}=\sum_{j\in s_\tau^{P,i}}\tilde{p}_j\,Z_j^2\frac{1-\hat{r}_j}{\hat{r}_j^2}=\sum_{j\in s_\tau^{P,i}}p_j\,Z_j^2\left(1-\hat{r}_j\right) \qquad (23)$$

Les probabilités r_j peuvent être estimées à partir d'un modèle logistique basé sur un ensemble de prédicteurs dont les valeurs sont observées à la fois sur les individus répondants et non-répondants.

3. Application numérique (LU-SILC 2016)

Les résultats numériques qui suivent ont été obtenus à partir des données LU-SILC 2016 en appliquant la méthode décrite dans les sections précédentes. Des estimations de variances et de marges d'erreurs sont présentées pour quelques indicateurs clés sur la pauvreté et l'inégalité. L'approche proposée permet de traiter aussi bien des statistiques définies sur l'ensemble de la population que des ventilations définies sur des sous-populations d'intérêt comme les classes d'âge, les femmes ou les hommes.

Les programmes de calcul d'indicateurs et de variance ont été écrits à l'aide du logiciel R.

Tableau 1 : Intervalles de confiance et marges d'erreurs, 2016

		Valeur estimée	Intervalle de confiance (90%)		Marge d'erreur (%)	Intervalle de confiance (95%)		Marge d'erreur (%)	Intervalle de confiance (99%)		Marge d'erreur (%)
			Inf	*Sup*		*Inf*	*Sup*		*Inf*	*Sup*	
Seuil de risque de pauvreté (60% du revenu médian)	Total	20314	19640	20988	3,3	19533	21095	3,8	19286	21342	5,1
	Hommes	20612	19841	21382	3,7	19718	21505	4,3	19436	21788	5,7
	Femmes	19978	19307	20650	3,4	19200	20757	3,9	18953	21004	5,1
	0-17	17565	16656	18473	5,2	16511	18618	6,0	16177	18952	7,9
	18-29	18470	17724	19216	4,0	17605	19335	4,7	17331	19609	6,2
	30-49	20130	19150	21109	4,9	18994	21266	5,6	18634	21625	7,4
	50-64	22628	21637	23619	4,4	21479	23777	5,1	21115	24141	6,7
	>64	23815	22872	24758	4,0	22721	24908	4,6	22375	25254	6,0

Taux de risque de pauvreté (%)	Total	16,5	14,9	18,0	9,6	14,6	18,3	11,1	14,1	18,9	14,6
	Hommes	16,1	14,4	17,8	10,6	14,1	18,1	12,3	13,5	18,7	16,3
	Femmes	16,7	15,0	18,4	10,0	14,8	18,6	11,6	14,1	19,3	15,3
	0-17	14,1	11,6	16,6	18,0	11,2	17,0	20,8	10,2	18,0	27,4
	18-29	15,8	13,3	18,3	15,9	12,9	18,7	18,4	12,0	19,6	24,2
	30-49	15,7	13,5	17,8	13,9	13,1	18,2	16,1	12,3	19,0	21,2
	50-64	18,0	15,8	20,1	12,0	15,5	20,5	13,9	14,7	21,2	18,3
	>64	13,0	10,6	15,3	18,2	10,2	15,7	21,2	9,4	16,6	27,8
Ratio interquintiles Q5/Q1	Total	2,4	2,3	2,5	3,9	2,3	2,5	4,5	2,2	2,5	5,9
	Hommes	2,3	2,2	2,4	4,1	2,2	2,4	4,8	2,2	2,5	6,3
	Femmes	2,4	2,3	2,5	4,1	2,3	2,5	4,8	2,3	2,6	6,3
	0-17	2,4	2,2	2,6	7,5	2,2	2,6	8,7	2,1	2,7	11,4
	18-29	2,4	2,2	2,6	9,0	2,1	2,6	10,4	2,1	2,7	13,7
	30-49	2,4	2,3	2,5	5,0	2,2	2,5	5,7	2,2	2,6	7,6
	50-64	2,4	2,3	2,6	6,0	2,2	2,6	6,9	2,2	2,6	9,1
	>64	2,2	2,1	2,4	7,0	2,1	2,4	8,1	2,0	2,5	10,7
Indice Foster-Greer-Thorbecke (alpha=1)	Total	4,8	4,2	5,3	11,6	4,1	5,4	13,5	3,9	5,6	17,8
	Hommes	4,6	4,0	5,2	12,7	3,9	5,3	14,8	3,7	5,5	19,5
	Femmes	4,9	4,3	5,6	13,3	4,2	5,7	15,4	3,9	5,9	20,3
	0-17	3,6	2,6	4,5	25,8	2,5	4,6	29,9	2,2	5,0	39,4
	18-29	4,3	3,4	5,2	21,3	3,3	5,4	24,6	2,9	5,7	32,4
	30-49	4,2	3,5	4,8	15,3	3,4	4,9	17,7	3,2	5,2	23,4
	50-64	5,9	5,0	6,8	15,9	4,8	7,0	18,4	4,5	7,3	24,2
	>64	4,7	3,4	5,9	26,2	3,2	6,1	30,4	2,8	6,5	40,1
Coefficient de Gini	Total	31,0	30,2	31,8	2,6	30,1	31,9	3,0	29,8	32,2	4,0
	Hommes	30,4	29,1	31,8	4,4	28,9	32,0	5,1	28,4	32,5	6,7
	Femmes	31,6	30,3	32,9	4,1	30,0	33,1	4,8	29,6	33,5	6,3
	0-17	30,3	28,9	31,7	4,6	28,7	31,9	5,3	28,2	32,4	7,0
	18-29	30,3	28,7	32,0	5,4	28,5	32,2	6,2	27,9	32,8	8,2
	30-49	31,4	30,1	32,6	4,1	29,9	32,8	4,7	29,4	33,3	6,2
	50-64	30,8	29,3	32,2	4,7	29,1	32,5	5,4	28,6	33,0	7,1
	>64	29,4	27,8	31,0	5,5	27,5	31,3	6,4	26,9	31,9	8,5
Ratio interquintiles des revenus S80/S20	Total	4,9	4,6	5,2	5,9	4,6	5,3	6,9	4,5	5,4	9,1
	Hommes	4,8	4,5	5,1	6,1	4,4	5,1	7,0	4,3	5,2	9,3
	Femmes	5,1	4,7	5,4	6,9	4,7	5,5	8,0	4,5	5,6	10,6
	0-17	4,6	4,2	5,0	9,2	4,1	5,1	10,7	3,9	5,2	14,1
	18-29	4,8	4,3	5,3	10,8	4,2	5,4	12,5	4,0	5,6	16,4
	30-49	4,8	4,4	5,2	8,7	4,3	5,3	10,0	4,2	5,5	13,2
	50-64	5,1	4,6	5,5	8,6	4,6	5,6	10,0	4,4	5,7	13,1
	>64	4,6	4,1	5,1	11,0	4,0	5,2	12,7	3,8	5,4	16,8

Source : Statec, EU-SILC

4. Conclusion

L'approche qui est proposée dans ce rapport pour l'estimation des marges d'erreurs dans LU-SILC est intéressante car elle est à la fois fondée théoriquement et facile à programmer à partir des logiciels statistiques "traditionnels" comme SAS, SPSS, STATA ou R. Elle peut être mise en œuvre pour produire rapidement des estimations de précision pour un très grand nombre de statistiques, ce qui est un avantage si l'on doit publier des résultats dans un délai très bref. Elle permet de traiter des indicateurs transversaux, des indicateurs longitudinaux et des indicateurs de changements entre deux années. Elle permet également de traiter le cas des statistiques non linéaires *via* la technique de linéarisation (Wolter, 2007)

A l'opposé, une faille de la méthode est que les calculs n'englobent pas l'effet de l'imputation des revenus manquants, ce qui conduit généralement à sous-estimer la précision. Une autre limite de la méthode concerne la taille de l'échantillon, qui doit être suffisamment importante pour pouvoir appliquer les résultats théoriques sur la linéarisation. Cela peut poser problème si l'on travaille sur des sous-populations de très petite taille.

Bibliographie

- Ardilly P. (2006). *Les techniques de sondage*, Technip.
- Beaumont J.-F. et Bissonnette J. (2011). "Variance Estimation under Composite Imputation: The Methodology Behind SEVANI". Survey Methodology, vol. 37, pp. 171-179.
- Berger Y. G. et Priam R. (2013). "A Simple Variance Estimator of Change for Rotating Repeated Surveys: an Application to the EU-SILC Household Surveys", Southampton: Southampton Statistical Sciences Research Institute. http://eprints.soton.ac.uk/347142/.
- Deville J-C. et Särndal C-E. (1992). "Calibration estimators in survey sampling". Journal of the American Statistical Association, pp. 376-382.
- Deville J-C. et Särndal C-E. (1994). "Variance Estimation for the Regression Imputed Horvitz-Thompson Estimator". Journal of Official Statistics, 10, 4, 381-394.
- Eurostat (2013). "Handbook on precision requirements and variance estimation for ESS households surveys", Methodologies and Working Papers. http://ec.europa.eu/eurostat/documents/3859598/5927001/KS-RA-13-029-EN.PDF
- Groves R., Fowler F., Couper M., Singer E. and Tourangeau R. (2004). *Survey Methodology*, Wiley.
- Kish L. (1965). *Survey sampling*, John Wiley & Sons.
- Lavallée P. (2002). *Le sondage indirect ou la méthode généralisée du partage des poids*, Ellipses.
- Massiani A. (2013). "Estimation de la variance d'indicateurs transversaux pour l'enquête SILC en Suisse", Techniques d'enquête, Vol.39, N°1, pp.139-167.
- Osier G. (2009)."Variance estimation for complex indicators of poverty and inequality using linearization techniques", Surveyy Research Methods (SRM), Vol.3, N°3, pp. 167-195. https://ojs.ub.uni-konstanz.de/srm/
- Osier G., Berger Y.G. et Goedemé T. (2013). "Standard error estimation for the EU-SILC indicators of poverty and social exclusion". Eurostat, Methodologies and Working Papers. http://ec.europa.eu/eurostat/documents/3888793/5855973/KS-RA-13-024-EN.PDF
- Rao J.N.K. et Wu C.F.J. (1988). "Re-Sampling Inference With Complex Survey Data". Journal of the American Statistical Association, Vol.83, N°401, pp. 231-241.
- Särndal C-E., Swensson B. and Wretman J. (1992). *Model Assisted Survey Sampling*, Springer.
- Tillé Y. (2000). *Echantillonnage et estimation en populations finies*, Dunod.
- Wolter K. M. (2007). *Introduction to variance estimation*, Springer.