STATEC

Variance estimation for rotating samples: the case of the EU-SILC survey in Luxembourg

Guillaume Osier

ITACOSM, Florence, 5-7 June 2019

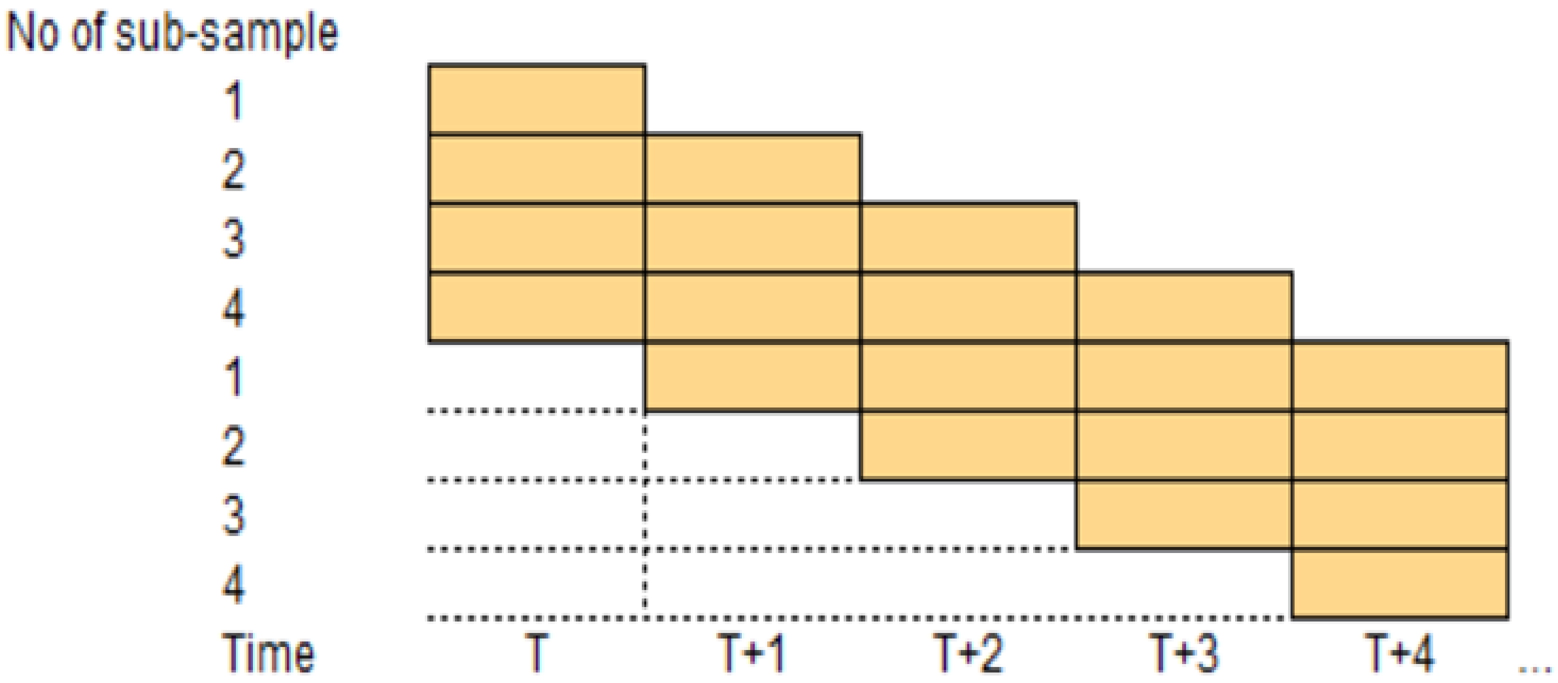
EU-SILC

- Reference database at EU level for comparable micro-data on income and living conditions
- Collected on a yearly basis since 2004
- Cross-sectional and longitudinal micro-data
- Strong legal basis at EU level
- Enable to calculate key poverty and inequality indicators (at-risk-of-poverty rate, income quantile ratio, Gini coefficient etc.)

EU-SILC in Luxembourg

- Since 2016, the survey has been conducted every year by Luxembourg's Statistical Institute (STATEC) in collaboration with LISER (*Luxembourg Institute of Socio-Economic Research*)
- Stratified simple random sample of 5000 individuals aged 18+ drawn from the National Population Register
- The individuals selected in the sample are contacted and their entire households are interviewed
- The sampled individuals are followed-up over four consecutive years (rotating panel)

Figure 2 - Le panel rotatif dans LU-SILC



Variance estimation

- Key to assess data quality
- Different types of indicators :
  - Cross-sectional (at-risk-of-poverty rate, Gini coefficient)
  - Longitudinal (persistence in poverty)
  - Net changes
- « Complex » sampling design (indirect sampling, non-response + calibration to external sources)

General approach

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

- $\tilde{s}_\tau = \cup_{i=1}^{4} s_\tau^{A,i}$ = cross-sectional sample of individuals
- ω_k = individual weight
- y_k = study variable
- $\tilde{s}_\tau^M = \cup_{i=1}^{4} s_\tau^{M,i}$ = cross-sectional sample of households
- $\widetilde{\omega}_h$ = household weight
- Y_h = total of the study variable at household level

Lemma

$$\hat{Y}_\tau = \sum_{j \in \tilde{s}_\tau^P} p_j Z_j$$

- $\tilde{s}_\tau^P = \cup_{i=1}^{4} s_\tau^{P,i}$ = sample of « panel » persons
- p_j = « panel » person weight
- $Z_j = \sum_{h \in U_\tau^M} \sum_{i \in h} 1_{j=(h,i)} \frac{Y_h}{L_h}$

This formula is a direct outcome from the Weight Share Method (Lavallée, 2007)

Analytical formula

- Assuming the sub-samples are independent

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

- **<u>Case 1</u>:** First-wave sub-sample

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

Design weight of individuals adjusted from unit non-response

First-wave sub-sample

$$Vsamp = \sum_{l=1}^{L} N_l^2 (1 - f_l) \frac{S_l^2}{n_l}$$

$$\hat{V}sampl = \sum_{l=1}^{L} N_l^2 (1 - f_l) \frac{s_l^2}{n_l}$$

$$Vrep = \sum_{k \in U} \tilde{d}_j \, K_j^2 \frac{1 - \theta_j}{\theta_j}$$

$$V\hat{r}ep = \sum_{k \in r} d_j^2 \, K_j^2 \left(1 - \hat{\theta}_j\right)$$

Panel component

$$V_i = V\left(\sum_{j \in s_\tau^{P,i}} p_j\, Z_j\right) = V_I + V_{II}$$

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

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

= Variance second phase (attrition)

Panel component (next)

The first phase variance is estimated the same way as the variance of first-wave samples. As regards the taking into account of attrition (second phase), we have:

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

Weighting adjustment factor for attrition

Extending the general approach

- Longitudinal indicators (persistence of poverty)
- Net changes ➔ Estimating the correlation matrix of regression residuals (Berger et al., 2012)
- Taking calibration into account ➔ The « residual » trick

Numerical results (LU-SILC 2016)

		Value	Confidence interval (90%)		Margin of error (%)	Confidence interval (95%)		Margin of error (%)	Confidence interval (99%)		Margin of error (%)
			Inf	Sup		Inf	Sup		Inf	Sup	
At-risk-of-poverty rate (%)	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
Gini coefficient	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
Inter-quintile share ratio 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

Tableau 3 : Significativité des changements entre 2015 et 2016

		Valeur estimée (2016)	Valeur estimée (2015)	Corrélation entre 2015 et 2016	p-value (%)
Seuil de risque de pauvreté (60% du revenu médian)	Total	20314	21162	0,39	0,969
	Hommes	20612	21364	0,39	0,910
	Femmes	19978	21080	0,37	0,994
	0-17	17565	18991	0,42	0,988
	18-29	18470	20186	0,24	0,993
	30-49	20130	21030	0,38	0,886
	50-64	22628	22687	0,36	0,079
	>64	23815	22515	0,28	0,974
Taux de risque de pauvreté (%)	Total	16,5	15,3	0,25	0,718
	Hommes	16,1	15,3	0,27	0,518
	Femmes	16,7	15,5	0,24	0,693
	0-17	14,1	15,2	0,29	0,444
	18-29	15,8	15,0	0,23	0,356
	30-49	15,7	14,9	0,27	0,402
	50-64	18,0	16,7	0,21	0,577
	>64	13,0	9,6	0,11	0,955

Conclusion

- The approach proposed is both theoretically justified and easy to implement using standard software tools (SAS, STATA, R …)
- Can deal with most of the indicators commonly used in EU-SILC (cross-sectional, longitudinal, net changes)
- Can treat non-linear indicators through the linearisation technique
- Limitations: measurement errors, imputation

STATEC

Institut national de la statistique
et des études économiques

Thank you for your attention

13, rue Erasme
L-1468 Luxembourg

info@statec.etat.lu
statec.lu