Echantillonnage produit : une application à l'Etude Longitudinale Française depuis l'Enfance

Hélène Juillard[1], Guillaume Chauvet[2] & Anne Ruiz-Gazen[3]

[1] *Ined - France, email : helene.juillard@ined.fr*
[2] *Ensai (Crest), Campus de Ker Lann, Bruz - France*
email : guillaume.chauvet@ensai.fr
[3] *Toulouse School of Economics - France*
email : anne.ruiz-gazen@tse-fr.eu

Résumé. L'Etude Longitudinale Française depuis l'Enfance (Elfe), démarrée en 2011, compte plus de 18 300 nourrissons dont les parents ont consenti à leur inclusion en maternité. Cette cohorte, consacrée au suivi des enfants, de la naissance à l'âge adulte, aborde les multiples aspects de la vie de l'enfant sous l'angle des sciences sociales, de la santé et de la santé-environnement. Dans chacune des maternités tirées aléatoirement, tous les nourrissons de la population cible, nés durant l'un des 25 jours répartis parmi les quatre saisons, ont été sélectionnés. Cet échantillon est le résultat d'un plan de sondage non standard que nous appellerons échantillonnage produit. Il se présente pour cette enquête sous la forme du croisement de deux échantillonnages indépendants : celui des maternités et celui des jours. Si l'on peut facilement imaginer un effet grappe dû à l'échantillonnage de maternités, on peut symétriquement imaginer un effet grappe dû à l'échantillonnage des jours. La dimension temporelle du plan ne pourra alors être négligée si les estimations recherchées sont susceptibles de variations journalières ou saisonnières. Seront proposés des estimateurs de variance adaptés à ce plan de sondage. Une étude par simulations illustrera nos propos.

Mots-clés. Enquête épidémiologique, estimation de variance, plan à plusieurs degrés

Abstract. The 2011 ELFE cohort comprises more than 18,000 children whose parents consented to their inclusion. In each of the selected maternity units, targeted babies born during four specific periods, representing each of the four seasons in 2011, were selected. ELFE is the first longitudinal study of its kind in France, tracking children from birth to adulthood. It will examine different aspects of these children's lives from the perspectives of health, social sciences and environmental health. The ELFE cohort was selected through a non-standard sampling design that we call product sampling, with independent selections of the sample of maternity units and of the sample of days. There may exist a maternity cluster effect, but there may exist a day cluster effect as well. The temporal dimension of the design can not be neglected when the desired estimates are subject to daily or seasonal variations. In this work, we propose variance estimators to

handle this type of sampling designs, and we derive specific variance estimators adapted to the ELFE case. A small simulation study supports our findings.

Keywords. Epidemiological survey, multistage sampling, variance estimation

1 Introduction

Elfe est une enquête longitudinale de type cohorte, comprenant 18 300 nourrissons à l'inclusion. Elle est consacrée au suivi des enfants, de la naissance à l'âge adulte, et aborde les multiples aspects de la vie de l'enfant sous l'angle des sciences sociales, de la santé et de la santé-environnement (Pirus et al., 2010). Cette étude est originale de par notamment sa pluridisciplinarité, la participation des deux parents, mais aussi son plan de sondage. Les nourrissons inclus dans la cohorte sont issus de deux échantillonnages : leur date de naissance fait partie d'un échantillon de jours de l'année 2011, et leur lieu de naissance appartient à un échantillon de maternités en France métropolitaine. L'échantillon des jours étant le même pour chaque maternité sélectionnée (ou vice versa, l'échantillon des maternités étant le même pour chaque jour sélectionné), on ne peut considérer ce plan comme un tirage à deux degrés classique, c'est-à-dire vérifiant l'hypothèse standard d'indépendance entre les tirages d'unités secondaires relatifs à chaque unité primaire. L'échantillon final se forme au croisement de lieux sélectionnés et de temps choisis : il résulte du produit de deux échantillonnages. Dans l'enquête Elfe, 349 maternités parmi 544 ont été tirées au sort pour participer à l'enquête. Par ailleurs quatre périodes de l'année 2011 ont été sélectionnées pour représenter chaque saison : du 1er avril au 4 avril, du 27 juin au 4 juillet, du 27 septembre au 4 octobre et enfin du 28 novembre au 5 décembre. Tous les enfants nés pendant ces périodes dans l'une des maternités métropolitaines associées à Elfe, ont pu participer à l'étude.

2 Echantillonnage produit indépendant

Considérons un plan de sondage $p_M(\cdot)$ sur une population U_M (de maternités) conduisant à un échantillon s_M. Nous utiliserons les indices i et j pour les individus de cette population. Soient π_i^M (> 0) et π_{ij}^M, les probabilités d'inclusion d'ordres un et deux et $\Delta_{ij}^M = \pi_{ij}^M - \pi_i^M \pi_j^M$. Considérons un autre plan de sondage $p_D(\cdot)$ sur une population U_D (de jours) conduisant à un échantillon s_D. Nous utiliserons les indices k et l pour les individus de cette population. Soient π_k^D (> 0) et π_{kl}^D, les probabilités d'inclusion d'ordres un et deux et $\Delta_{kl}^D = \pi_{kl}^D - \pi_k^D \pi_l^D$.

L'unité finale d'échantillonnage qui nous intéresse est caractérisée par un couple d'éléments (i, k), avec $i \in U_M$ et $k \in U_D$. On s'intéresse à une variable Y prenant la valeur Y_{ik} dans la maternité i, le jour k . Dans l'enquête Elfe, on se référera donc à l'élément "grappe des nourrissons nés dans la même maternité i, le même jour k".

Chaque unité finale appartient à une population U, définie par le produit des deux populations sources :

$$U = U_M \times U_D$$

Nous définissons l'**échantillon produit** par :

$$s = s_M \times s_D.$$

Dans ce cadre général, le plan produit $p(s)$ peut prendre différentes formes, chacun des tirages pouvant être effectué conditionnellement ou pas à l'issue de l'autre tirage. Le travail présenté ci-après considère un cas particulier du **plan produit**, cas dans lequel les deux échantillonnages se font indépendamment l'un de l'autre :

$$p(s) = p_M(s_M) \times p_D(s_D).$$

Remarquons que cette **hypothèse d'indépendance** entre deux tirages est analogue à l'hypothèse d'invariance utilisée dans l'échantillonnage classique à deux degrés (Särndal, Swensson et Wretman, 1992, page 134).

Sous ces conditions, on peut alors facilement définir les probabilités d'inclusion d'ordres un et deux et les covariances Γ_{ijkl} relatives au plan produit à partir de celles de chaque plan source. Pour toutes les unités $i, j \in U_M$ et $k, l \in U_D$:

$$\begin{aligned}
\mathbf{E}\left(\mathbf{1}_{\{(i,k)\in s\}}\right) &= \pi_i^M \pi_k^D, \\
\mathbf{E}\left(\mathbf{1}_{\{(i,k)\in s\}}\mathbf{1}_{\{(j,l)\in s\}}\right) &= \pi_{ij}^M \pi_{kl}^D, \\
\Gamma_{ijkl} \equiv \mathbf{Cov}\left(\mathbf{1}_{\{(i,k)\in s\}}, \mathbf{1}_{\{(j,l)\in s\}}\right) &= \pi_{ij}^M \pi_{kl}^D - \pi_i^M \pi_j^M \pi_k^D \pi_l^D. \qquad (1)
\end{aligned}$$

On s'intéresse au total $t_Y = \sum_{i\in U_M} \sum_{k\in U_D} Y_{ik}$ estimé sans biais par

$$\hat{t}_Y = \sum_{i\in S_M} \sum_{k\in S_D} \frac{Y_{ik}}{\pi_i^M \pi_k^D} = \sum_{i\in S_M} \frac{\hat{Y}_{i\bullet}}{\pi_i^M} = \sum_{k\in S_D} \frac{\hat{Y}_{\bullet k}}{\pi_k^D}$$

avec $\hat{Y}_{i\bullet}$, l'estimateur du total sur la maternité i et $\hat{Y}_{\bullet k}$, l'estimateur du total sur le jour k. La variance de l'estimateur $\hat{t}_Y$ peut alors s'écrire :

$$\mathrm{V}\left(\hat{t}_Y\right) = \sum_{i,j\in U_M} \sum_{k,l\in U_D} \Gamma_{ijkl} \frac{Y_{ik}}{\pi_i^M \pi_k^D} \frac{Y_{jl}}{\pi_j^M \pi_l^D}. \qquad (2)$$

Alternativement, on pourra aussi utiliser la décomposition de la variance pour écrire :

$$\begin{aligned}
\mathrm{V}\left(\hat{t}_Y\right) &= \mathbf{E}\left\{\mathrm{V}\left(\hat{t}_Y|S_M\right)\right\} + \mathrm{V}\left\{\mathbf{E}\left(\hat{t}_Y|S_M\right)\right\} \qquad (3) \\
&= \mathbf{E}\left\{\mathrm{V}\left(\hat{t}_Y|S_D\right)\right\} + \mathrm{V}\left\{\mathbf{E}\left(\hat{t}_Y|S_D\right)\right\}. \qquad (4)
\end{aligned}$$

3 Estimation de variance

Un estimateur de $\mathrm{V}\left(\hat{t}_Y\right)$ est :

$$\hat{\mathrm{V}}_{HT}\left(\hat{t}_Y\right) = \sum_{i,j\in S_M}\sum_{k,l\in S_D}\frac{\Gamma_{ijkl}}{\pi^M_{ij}\pi^D_{kl}}\frac{Y_{ik}}{\pi^M_i\pi^D_k}\frac{Y_{jl}}{\pi^M_j\pi^D_l}.$$

Cet estimateur est sans biais si tous les π^M_{ij} et tous les π^D_{kl} sont strictement positifs, pour tous $(i,j)\in U^2_M$, $(k,l)\in U^2_D$.

En utilisant l'identité

$$\Gamma_{ijkl} = \pi^M_{ij}\Delta^D_{kl} + \pi^D_k\pi^D_l\Delta^M_{ij},$$

l'estimateur de variance peut alors se réécrire sous la forme :

$$\hat{\mathrm{V}}_{HT}\left(\hat{t}_Y\right) = \sum_{k,l\in S_D}\frac{\Delta^D_{kl}}{\pi^D_{kl}}\left(\frac{\hat{Y}_{\bullet k}}{\pi^D_k}\right)\left(\frac{\hat{Y}_{\bullet l}}{\pi^D_l}\right) + \sum_{i,j\in S_M}\frac{\Delta^M_{ij}}{\pi^M_{ij}\pi^M_i\pi^M_j}\sum_{k,l\in S_D}\frac{Y_{ik}Y_{jl}}{\pi^D_{kl}}.$$

Il s'agit d'un estimateur sans biais terme à terme de la décomposition de la variance donnée en (3). On reconnaîtra le premier terme comme la partie inter jours de l'estimateur de variance pour un plan de sondage à 2 degrés classique dans lequel les unités primaires seraient les jours et les unités secondaires, les maternités.

En utilisant l'identité

$$\Gamma_{ijkl} = \pi^D_{kl}\Delta^M_{ij} + \pi^M_i\pi^M_j\Delta^D_{kl},$$

l'estimateur de variance peut aussi se décomposer sous la forme :

$$\hat{\mathrm{V}}_{HT}\left(\hat{t}_Y\right) = \sum_{i,j\in S_M}\frac{\Delta^M_{ij}}{\pi^M_{ij}}\left(\frac{\hat{Y}_{i\bullet}}{\pi^M_i}\right)\left(\frac{\hat{Y}_{j\bullet}}{\pi^M_j}\right) + \sum_{k,l\in S_D}\frac{\Delta^D_{kl}}{\pi^D_{kl}\pi^D_k\pi^D_l}\sum_{i,j\in S_M}\frac{Y_{ik}Y_{jl}}{\pi^M_{ij}}.$$

Il s'agit d'un estimateur sans biais terme à terme de la décomposition de la variance donnée en (4). On reconnaîtra le premier terme comme la partie inter maternités de l'estimateur de variance pour un plan de sondage à 2 degrés classique dans lequel les unités primaires seraient les maternités et les unités secondaires, les jours.

Enfin, en utilisant l'identité

$$\Gamma_{ijkl} = \pi^M_{ij}\Delta^D_{kl} + \pi^D_{kl}\Delta^M_{ij} - \Delta^M_{ij}\Delta^D_{kl},$$

l'estimateur de variance peut aussi s'écrire sous une dernière forme, symétrique par rapport aux échantillons S_M et S_D :

$$\begin{aligned}\hat{\mathrm{V}}_{HT}\left(\hat{t}_Y\right) &= \sum_{k,l\in S_D}\frac{\Delta_{kl}^D}{\pi_{kl}^D}\left(\frac{\hat{Y}_{\bullet k}}{\pi_k^D}\right)\left(\frac{\hat{Y}_{\bullet l}}{\pi_l^D}\right)+\sum_{i,j\in S_M}\frac{\Delta_{ij}^M}{\pi_{ij}^M}\left(\frac{\hat{Y}_{i\bullet}}{\pi_i^M}\right)\left(\frac{\hat{Y}_{j\bullet}}{\pi_j^M}\right)\\ &- \sum_{i,j\in S_M}\sum_{k,l\in S_D}\frac{\Delta_{ij}^M}{\pi_{ij}^M}\frac{\Delta_{kl}^D}{\pi_{kl}^D}\frac{Y_{ik}}{\pi_i^M\pi_k^D}\frac{Y_{jl}}{\pi_j^M\pi_l^D}.\end{aligned}$$

Tout comme Beaumont, Béliveau et Haziza (2014) ont présenté des estimateurs simplifiés pour l'échantillonnage à deux phases, des simplifications utilisant les différentes formes de l'estimateur de variance seront proposées et les propriétés des différents estimateurs obtenus seront comparées. On pourra considérer l'estimateur de variance que l'on obtiendrait pour un plan à deux degrés avec hypothèse d'indépendance entre les tirages d'unités secondaires. Un des objectifs poursuivis est d'aiguiller l'utilisateur parmi les procédures existant dans les logiciels, en fonction de ses données et des hypothèses requises. Enfin, nous envisageons une estimation de variance par linéarisation ou par bootstrap pour des paramètres plus complexes qu'un total.

Bibliographie

[1] Beaumont, J.-F., Béliveau, A. et Haziza, D. (2014), Clarifying some aspects of variance estimation in two-phase sampling, preprint.
[2] Pirus, C., Bois, C., Dufourg, M.-N., Lanoë, J.-L., Vandentorren, S., Leridon, H. et l'équipe Elfe (2010), La construction d'une cohorte : l'expérience du projet français Elfe, Population 65(4) : 637-670.
[3] Särndal, C.-E., Swensson, B. et Wretman, J.H. (1992), Model Assisted Survey Sampling, Springer-Verlag.