

# ECHANTILLONNAGE PRODUIT : UNE APPLICATION À L'ÉTUDE LONGITUDINALE FRANÇAISE DEPUIS L'ENFANCE

Hélène Juillard<sup>1</sup>, Guillaume Chauvet<sup>2</sup> & Anne Ruiz-Gazen<sup>3</sup>

<sup>1</sup> *Ined - France, email : helene.juillard@ined.fr*

<sup>2</sup> *Ensaï (Crest), Campus de Ker Lann, Bruz - France  
email : guillaume.chauvet@ensai.fr*

<sup>3</sup> *Toulouse School of Economics - France  
email : anne.ruiz-gazen@tse-fr.eu*

**Résumé.** L'Étude 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 appelons 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

## 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  $\pi_i^M (> 0)$  et  $\pi_{ij}^M$ , 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  $\pi_k^D (> 0)$  et  $\pi_{kl}^D$ , 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  $p(\cdot)$  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 calculer les probabilités d'inclusion d'ordres un et deux et les covariances  $\Gamma_{ijkl}$  relatives au plan produit à partir de celles de chaque plan source. Pour toutes les unités  $i, j \in U_M$  et  $k, l \in U_D$  :

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

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

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

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

$$V(\hat{t}_Y) = \mathbf{E} \left\{ V(\hat{t}_Y | S_M) \right\} + V \left\{ \mathbf{E}(\hat{t}_Y | S_M) \right\} \quad (3)$$

$$= \mathbf{E} \left\{ V(\hat{t}_Y | S_D) \right\} + V \left\{ \mathbf{E}(\hat{t}_Y | S_D) \right\}. \quad (4)$$

### 3 Estimation de variance

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

$$\hat{V}_{HT}(\hat{t}_Y) = \sum_{i,j \in S_M} \sum_{k,l \in S_D} \frac{\Gamma_{ijkl}}{\pi_{ij}^M \pi_{kl}^D} \frac{Y_{ik}}{\pi_i^M \pi_k^D} \frac{Y_{jl}}{\pi_j^M \pi_l^D}.$$

Cet estimateur est sans biais si tous les  $\pi_{ij}^M$  et tous les  $\pi_{kl}^D$  sont strictement positifs, pour tous  $(i, j) \in U_M^2$ ,  $(k, l) \in U_D^2$ .

En utilisant l'identité

$$\Gamma_{ijkl} = \pi_{ij}^M \Delta_{kl}^D + \pi_k^D \pi_l^D \Delta_{ij}^M,$$

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

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

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 variance de premier degré 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

$$\Gamma_{ijkl} = \pi_{kl}^D \Delta_{ij}^M + \pi_i^M \pi_j^M \Delta_{kl}^D,$$

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

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

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 variance de premier degré 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

$$\Gamma_{ijkl} = \pi_{ij}^M \Delta_{kl}^D + \pi_{kl}^D \Delta_{ij}^M - \Delta_{ij}^M \Delta_{kl}^D,$$

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

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

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.