Estimateurs de variance issus d'un plan produit pour l'enquête Elfe

Hélène JUILLARD[1](*), Guillaume CHAUVET[2](**), Anne RUIZ-GAZEN[3](***)

(*) Ined

(**) Crest/Ensai

(***) Toulouse School of Economics

Résumé

L'Etude Longitudinale Française depuis l'Enfance (Elfe), démarrée en 2011, compte plus de 18 300 nourrissons recrutés en maternité. Dans chacune des maternités tirées aléatoirement, tous les nourrissons de la population cible, nés durant l'un des 25 jours répartis parmi les 4 saisons de l'année 2011, ont été sélectionnés. L'échantillon initial est le résultat d'un plan de sondage non standard que nous appelons plan *produit* (encore appelé *cross-classified sampling*, voir Ohlsson 1996). Il se présente pour cette enquête sous la forme du croisement de deux échantillonnages indépendants : celui des maternités et celui des jours. Elfe est la première cohorte française consacrée au suivi des enfants, de la naissance à l'âge adulte, qui aborde les multiples aspects de la vie de l'enfant sous l'angle des sciences sociales, de la santé et de l'environnement lié à la santé. La dimension temporelle du plan ne pourra être négligée si les estimations recherchées sont susceptibles de variations journalières ou saisonnières. Dans cet article, la variance issue d'un plan produit est présentée et comparée à celle issue d'un plan classique à deux degrés. De plus, la non-réponse est prise en compte dans le calcul de l'estimateur de cette variance. Pour l'enquête Elfe, plusieurs modélisations du plan sont proposées et les estimateurs de variance associés sont dérivés. Enfin, pour l'utilisateur, plusieurs estimateurs simplifiés facilement mis en œuvre grâce à des procédures logicielles existantes (R/SAS/Stata) sont comparés et illustrés sur données Elfe.

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 (twenty five days), representing each of the four seasons in 2011, were selected. The initial sample is the result of a non-standard sampling design, called *cross-classified sampling* (Ohlsson, 1996). It consists of the crossing of two independent

1. helene.juillard@ined.fr
2. guillaume.chauvet@ensai.fr
3. anne.ruiz-gazen@tse-fr.eu

samplings : the population of maternities and the population of days. ELFE is the first longitudinal study of its kind in France, tracking children from birth to adulthood. In this paper, the variance of this type of design is presented and variance estimators are derived, taking account of the nonresponse. Some simplified variance estimators are proposed and illustrated with ELFE data.

Mots-clés

Estimateurs simplifiés de variance, plan de sondage à deux degrés, plan stratifié, variance anticipée, variance due à la non-réponse.

Introduction

Chaque échantillonnage conduit à une variance. Cette variance est une mesure d'incertitude (ou de précision) relative au fait de sélectionner un échantillon et reflète la façon dont l'échantillon a été tiré. Dans le cas d'un recensement, cette variance est nulle. Après déroulement d'une enquête, les informations relatives à un seul échantillon sont connues et les calculs du paramètre estimé $\hat{\theta}$ et de sa variance estimée $\hat{\mathbf{V}}(\hat{\theta})$ sont possibles. De ces calculs dépendront les intervalles de confiance associés à chaque paramètre estimé. Dans ce document, nous considérons uniquement des paramètres θ en population finie (totaux, ratios, coefficients de corrélation...) et nous supposons que l'aléa provient du tirage de l'échantillon (inférence basée sur le plan, voir Särndal, Swensson et Wretman, 1992).

Le plan utilisé pour l'enquête Elfe n'est pas standard. Il s'agit du produit de deux échantillonnages indépendants. Le calcul de l'estimateur de variance n'est pas directement disponible dans la littérature. Ce document propose de détailler le calcul et l'estimation de la variance dans le cas spécifique de l'enquête Elfe.

L'échantillonnage produit est présenté et comparé à celui plus classique à deux degrés d'échantillonnage. On peut notamment montrer que pour un estimateur de type Horvitz-Thompson la variance anticipée issue d'un plan produit est plus grande que celle issue d'un plan à deux degrés. Dans un second temps, un estimateur sans biais de variance est dérivé pour traiter ce type de plan de sondage et différentes modélisations possibles pour l'enquête Elfe sont comparées. Dans une troisième partie, la phase de non-réponse est prise en compte dans le calcul de l'estimateur de variance. L'estimateur sans biais de variance s'avère pouvoir prendre des valeurs négatives et se présente sous une forme relativement complexe. Pour ces raisons, plusieurs estimateurs simplifiés, prenant en compte les procédures logicielles déjà existantes (R, SAS, Stata), sont définis et mis en œuvre sur des données Elfe. Un des objectifs poursuivis est d'aiguiller l'utilisateur pour l'estimation de la variance en présence d'un tel plan de sondage.
Les détails des calculs présentés dans cet article sont donnés dans Juillard, Chauvet et Ruiz-Gazen (2015).

1 Enquête Elfe et échantillonnage produit

Dans cette partie, sont présentées les deux sélections (maternités et jours) réalisées et croisées pour obtenir l'échantillon de nourrissons Elfe. Le plan produit est défini dans un cadre théorique, puis comparé à d'autres plans. Plus particulièrement, la différence en terme de variance entre un plan produit et un plan à deux degrés est formulée et illustrée.

1.1 Contexte et modélisations du plan de sondage

La population d'inférence est celle des nourrissons [4] nés durant l'année 2011 en France métropolitaine. Toutes les familles sélectionnées ont été enquêtées peu de temps après l'accouchement dans certaines maternités métropolitaines et durant certains jours de l'année.

4. Les nourrissons éligibles sont ceux issus d'un accouchement au plus gémellaire, hors grands prématurés, ayant une mère majeure et en mesure de donner un consentement éclairé notamment dans l'une des langues proposées (français, anglais, arabe ou turc), nés dans une maternité métropolitaine et dont les parents ne résidaient pas temporairement en métropole.

Le plan de sondage pour les maternités est un plan probabiliste (voir Figure 1). Concernant les jours, 25 ont été choisis durant quatre périodes (appelées vagues) couvrant les quatre saisons de l'année (dont la moitié devait coïncider avec l'échantillon démographique permanent E.D.P.) : du 1er au 4 avril, les 27 et 28 juin, du 1er au 4 juillet, du 27 au 29 septembre, du 1er au 4 octobre, du 28 au 30 novembre et du 1er au 5 décembre. Il est important de noter que les deux échantillons (maternités et jours) ont été sélectionnés indépendamment.

FIGURE 1 – Représentation schématique du plan de sondage utilisé pour l'enquête Elfe

L'échantillonnage probabiliste des maternités correspond à un plan stratifié : cinq strates à effectifs égaux avec tirages à allocation proportionnelle au nombre de naissances recensées en 2008. Il s'agissait d'un tirage systématique avec pour variable de stratification implicite le statut juridique de la maternité, le niveau de médicalisation et la région en cinq postes. Par la suite, on supposera être dans le cas d'un plan stratifié avec plan SI [5] dans chaque strate (plan STSI).

L'échantillonnage des jours n'est pas aléatoire, d'où la nécessité de le modéliser. Nous proposons trois modélisations. La première consiste en un plan STSI avec quatre strates (saisons) et tirage SI à l'intérieur de chaque strate de respectivement 4, 6, 7 et 8 jours. Cette modélisation permet de représenter l'effet saisonnier du plan mais néglige l'effet grappe (jours presque consécutifs sélectionnés durant chaque saison). La deuxième modélisation considère le plan des jours comme un tirage SI de 25 jours parmi 365. La troisième modélisation considère le plan des jours comme un plan stratifié avec plan SIC [6] dans chaque strate (plan STSIC) à deux strates (semestres) dans lesquelles deux grappes de jours auraient été tirées (ce qui permet de prendre en compte l'effet grappe mais pas celui des saisons). Les estimateurs de variance issus de ces trois plans STSI, SI et STSIC sont formulés et illustrés dans la suite du document.

1.2 Définition du plan produit

Notons U_M la population des maternités de taille N_M et U_D la population des jours de taille N_D. Les indices i et j sont utilisés pour les maternités et les indices k et l pour les jours. On considère un plan de sondage p_M dans la population U_M menant à un échantillon S_M de taille (moyenne) n_M et un plan de sondage p_D dans la population U_D menant à un échantillon S_D de taille (moyenne) n_D (voir Figure 2). On suppose que les deux plans sont indépendants.

Le plan produit $p(\cdot)$ est alors défini sur la population produit $U = U_M \times U_D$ par

$$p(s) = p_M(s_M) \times p_D(s_D) \quad \text{pour tout} \quad s = s_M \times s_D \subset U_M \times U_D. \tag{1}$$

5. Tirage aléatoire simple sans remise.
6. Tirage aléatoire de grappes (clusters) d'unités, à probabilités égales, sans remise.

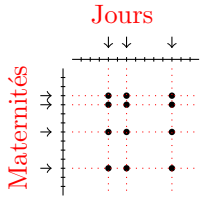

FIGURE 2 – Echantillonnage de maternités et de jours pour un plan produit

Soient π_i^M (> 0) et π_{ij}^M, les probabilités d'inclusion d'ordres un et deux pour le plan p_M et $\Delta_{ij}^M = \pi_{ij}^M - \pi_i^M \pi_j^M$. Soient π_k^D (> 0) et π_{kl}^D, les probabilités d'inclusion d'ordres un et deux pour le plan p_D 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 maternité $\times$ jour (i, k), avec $i \in U_M$ et $k \in U_D$.

Grâce à l'hypothèse d'indépendance, on peut alors facilement calculer 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 $k, l \in U_D$:

$$\begin{aligned} \mathbf{E}\left(\mathbf{1}_{\{(i,k)\in s\}}\right) &= \pi_i^M \pi_k^D, & (2)\\ \mathbf{E}\left(\mathbf{1}_{\{(i,k)\in s\}}\mathbf{1}_{\{(j,l)\in s\}}\right) &= \pi_{ij}^M \pi_{kl}^D, & (3)\\ \Gamma_{ijkl} \equiv \mathbf{Cov}\left(\mathbf{1}_{\{(i,k)\in s\}}, \mathbf{1}_{\{(j,l)\in s\}}\right) &= \pi_{ij}^M \pi_{kl}^D - \pi_i^M \pi_j^M \pi_k^D \pi_l^D & (4) \end{aligned}$$

où $\mathbf{1}_{\{\cdot\}}$ est la fonction indicatrice.

Notre variable d'intérêt Y prend la valeur Y_{ik} pour la maternité i et le jour k. On s'intéresse au total $t_Y = \sum_{i\in U_M} \sum_{k\in U_D} Y_{ik}$ estimé sans biais par

$$\hat{t}_Y = \sum_{i\in S_M} \sum_{k\in S_D} \frac{Y_{ik}}{\pi_i^M \pi_k^D} = \sum_{i\in S_M} \frac{\hat{Y}_{i\bullet}}{\pi_i^M} = \sum_{k\in S_D} \frac{\hat{Y}_{\bullet k}}{\pi_k^D} \quad (5)$$

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 :

$$\mathrm{V}\left(\hat{t}_Y\right) = \sum_{i,j\in U_M} \sum_{k,l\in U_D} \Gamma_{ijkl} \frac{Y_{ik}}{\pi_i^M \pi_k^D} \frac{Y_{jl}}{\pi_j^M \pi_l^D}. \quad (6)$$

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

$$\hat{\mathrm{V}}_{HT}\left(\hat{t}_Y\right) = \sum_{i,j\in S_M} \sum_{k,l\in S_D} \frac{\Gamma_{ijkl}}{\pi_{ij}^M \pi_{kl}^D} \frac{Y_{ik}}{\pi_i^M \pi_k^D} \frac{Y_{jl}}{\pi_j^M \pi_l^D}. \quad (7)$$

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

1.3 Comparaison avec un plan à deux degrés

Le plan produit est un plan particulier avec deux phases d'échantillonnage. Il diffère de façon évidente d'un plan à une seule phase, directement dans la population produit. Dans

ce qui suit, nous comparons le plan produit et un plan à deux degrés. Les formules de variance sont présentées, puis comparées sous un modèle.

1.3.1 Présentation des plans comparés

Si le plan produit est bien un plan dans la population produit $U_M \times U_D$, il est caractérisé par deux plans sources (tirage de i dans U_M, tirage de k dans U_D), et diffère d'un plan de sondage direct dans cette population, c'est-à-dire qui tirerait directement des unités (i, k) dans $U_M \times U_D$ comme illustré en Figure 3.

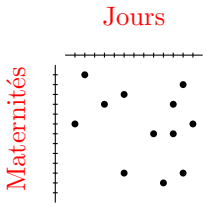

FIGURE 3 – Echantillonnage de maternités et de jours pour un tirage direct dans la population produit

Pour l'enquête Elfe, on discerne deux phases d'échantillonnage : celle sur les jours et celle sur les maternités. Le plan Elfe peut-il être considéré comme un plan classique à deux degrés avec au premier degré un échantillonnage de maternités et au second degré un échantillonnage de jours (Figure 4) ? Ou symétriquement, un plan classique à deux degrés avec au premier degré un échantillonnage de jours et au second degré un échantillonnage de maternités ?

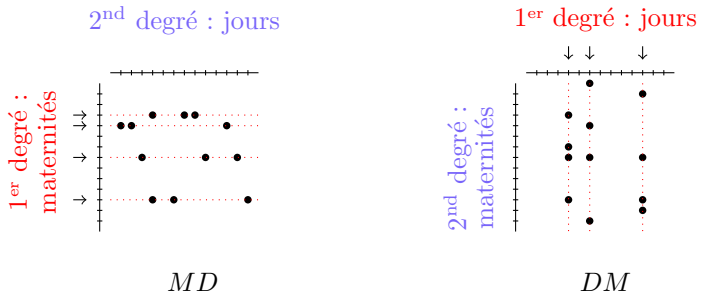

FIGURE 4 – Echantillonnage de maternités et de jours pour un plan à deux degrés, avec tirage de maternités au premier degré (à gauche) ou tirage de jours au premier degré (à droite)

Un plan classique à deux degrés requiert deux hypothèses : l'indépendance entre les tirages effectués à chaque degré, encore appelée propriété d'invariance ; l'indépendance entre les différents tirages effectués au second degré, conditionnellement au premier degré de tirage. Pour un plan produit, la première hypothèse est vérifiée (indépendance entre l'échantillon de maternités et l'échantillon de jours) mais la seconde ne l'est pas (le même échantillon de jours est utilisé pour chaque maternité).

On considère le cas particulier où p_D et p_M sont des plans SI. On notera par la suite ce plan SI $\times$ SI. La variance donnée en formule (6) peut se réécrire sous la forme :

$$\begin{aligned} \mathrm{V}_{prod}\left(\hat{t}_Y\right) &= N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)S^2_{Y_{\bullet\circ}} + N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)S^2_{Y_{\circ\bullet}} \\ &+ N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)S^2 \end{aligned} \tag{8}$$

où

$$S^2_{Y_{\bullet\circ}} = \frac{1}{N_D-1}\sum_{k\in U_D}\left(Y_{\bullet k}-\frac{1}{N_D}\sum_{l\in U_D}Y_{\bullet l}\right)^2, \tag{9}$$

$$S^2_{Y_{\circ\bullet}} = \frac{1}{N_M-1}\sum_{i\in U_M}\left(Y_{i\bullet}-\frac{1}{N_M}\sum_{j\in U_M}Y_{j\bullet}\right)^2, \tag{10}$$

$$S^2 = \frac{1}{(N_D-1)(N_M-1)}\sum_{k\in U_D}\sum_{i\in U_M}\left(Y_{ik}-\bar{Y}_{i\bullet}-\bar{Y}_{\bullet k}+\bar{\bar{Y}}_{\bullet\bullet}\right)^2, \tag{11}$$

avec

$$Y_{i\bullet} = \sum_{l\in U_D}Y_{il}, \qquad Y_{\bullet k}=\sum_{j\in U_M}Y_{jk}, \tag{12}$$

$$\bar{Y}_{i\bullet} = \frac{1}{N_D}Y_{i\bullet}, \qquad \bar{Y}_{\bullet k}=\frac{1}{N_M}Y_{\bullet k}, \tag{13}$$

$$\bar{\bar{Y}}_{\bullet\bullet} = \frac{1}{N_D}\frac{1}{N_M}\sum_{l\in U_D}\sum_{j\in U_M}Y_{jl}. \tag{14}$$

Considérons le cas d'un plan de sondage à deux degrés où on sélectionne au premier degré un échantillon S_M de taille n_M dans U_M, puis dans chaque unité primaire i de S_M on sélectionne indépendamment un échantillon S_i d'unités secondaires dans U_D. On considère le cas particulier où tous les échantillons S_i sont sélectionnés avec la même taille n_D. On note dans la suite MD ce plan de sondage et V_{MD} la variance correspondante. Dans le cas d'un sondage aléatoire simple sans remise à chaque degré, noté {SI,SI}, on obtient :

$$\mathrm{V}_{MD}\left(\hat{t}_Y\right) = N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)S^2_{Y_{\bullet\circ}} + \frac{N_M}{n_M}N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)\sum_{i\in U_M}S^2_{Y_{i\circ}} \tag{15}$$

où

$$S^2_{Y_{i\circ}} = \frac{1}{N_D-1}\sum_{k\in U_D}\left(Y_{ik}-\frac{1}{N_D}\sum_{l\in U_D}Y_{il}\right)^2. \tag{16}$$

Le cas d'un plan de sondage à deux degrés noté DM est obtenu de façon analogue en considérant la population U_D au premier degré. On note V_{DM}, la variance correspondant à ce plan de sondage. Dans le cas {SI,SI}, on obtient :

$$\mathrm{V}_{DM}\left(\hat{t}_Y\right) = N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)S^2_{Y_{\circ\bullet}} + \frac{N_D}{n_D}N_M^2\left(\frac{1}{n_M}-\frac{1}{N_M}\right)\sum_{i\in U_D}S^2_{Y_{\circ k}} \tag{17}$$

où

$$S^2_{Y_{\circ k}} = \frac{1}{N_M-1}\sum_{i\in U_M}\left(Y_{ik}-\frac{1}{N_M}\sum_{j\in U_M}Y_{jk}\right)^2. \tag{18}$$

1.3.2 Comparaison des variances anticipées

La différence entre la variance issue d'un plan produit SI × SI donnée en formule (8), d'une part, et la variance issue d'un plan {SI,SI} donnée en formule (15) ou (17), d'autre part, n'est pas nécessairement positive. Nous considérons le modèle de comportement

$$m : Y_{ik} = \mu + \sigma_1 U_i + \sigma_2 V_k + \sigma_3 W_{ik} \tag{19}$$

avec $U_i, V_k, W_{ik} \sim \mathcal{N}(0,1)$ et $\sigma_1, \sigma_2, \sigma_3 \in \mathbb{R}^+$, où σ_1 représente un effet maternité, σ_2 un effet jour, et σ_3 un effet résiduel. Sous le modèle (19), on peut montrer que la variance anticipée du plan produit est toujours plus grande que celle du plan à deux degrés considéré. En effet, si on note E_m l'espérance sous le modèle (19),

$$E_m\left[\mathrm{V}_{prod}\left(\hat{t}_Y\right) - \mathrm{V}_{MD}\left(\hat{t}_Y\right)\right] = N_M^2 N_D^2 \frac{n_M - 1}{n_M}\left(\frac{1}{n_D} - \frac{1}{N_D}\right)\sigma_2^2, \tag{20}$$

$$E_m\left[\mathrm{V}_{prod}\left(\hat{t}_Y\right) - \mathrm{V}_{DM}\left(\hat{t}_Y\right)\right] = N_M^2 N_D^2 \frac{n_D - 1}{n_D}\left(\frac{1}{n_M} - \frac{1}{N_M}\right)\sigma_1^2. \tag{21}$$

Cette différence dépend du second degré d'échantillonnage : plus la taille des échantillons du second degré est grande, plus les deux variances se rapprochent. A l'inverse, la variance V_{prod} est d'autant plus grande devant V_{MD} que la variabilité inter-jours σ_2^2 est grande. De façon analogue, la variance V_{prod} est d'autant plus grande devant V_{DM} que la variabilité inter-maternités σ_1^2 est grande.

Cette comparaison peut se faire de façon relative en calculant le ratio des variances anticipées. L'espérance sous le modèle (19) de la variance issue d'un plan produit peut par exemple s'écrire :

$$\begin{aligned} E_m\left[\mathrm{V}_{prod}\left(\hat{t}_Y\right)\right] &= N_M^2 N_D^2 \left(\frac{1}{n_M} - \frac{1}{N_M}\right)\sigma_1^2 + N_M^2 N_D^2 \left(\frac{1}{n_D} - \frac{1}{N_D}\right)\sigma_2^2 \\ &+ N_M^2 N_D^2 \left(\frac{1}{n_M n_D} - \frac{1}{N_M N_D}\right)\sigma_3^2 \end{aligned} \tag{22}$$

En Figure 5, nous considérons le ratio des variance en générant les données sous le modèle (19) en fixant μ à 200 et σ_1, σ_2 et σ_3 à 5. Chaque population contient $N_M = 1000$ maternités et $N_D = 1000$ jours. Les tailles d'échantillon n_M et n_D varient de 10 à 500. On constate que quand la taille d'échantillon n_D augmente, le ratio $\mathrm{V}_{MD}\ /\mathrm{V}_{prod}$ se rapproche de 1. A l'inverse, lorsque la taille n_M augmente, le ratio se rapproche de 0.

Remarquons que le plan produit a aussi été étudié et comparé au plan à deux degrés dans Vos (1964) et dans Bellhouse (1981). Vos (1964) considère une famille générale de plans de sondages adaptés au tirage de lignes et de colonnes sur une grille rectangulaire. Le plan produit, qui correspond à des tirages indépendants de lignes et de colonnes, est un cas particulier de cette famille de plans. Vos exhibe notamment des conditions, dont on peut penser qu'elles sont la plupart du temps vérifiées, sous lesquelles le plan produit conduit à une variance plus grande que pour le plan à deux degrés. Cette comparaison est faite dans le cas particulier de plans aléatoires simples sans remise des lignes et des colonnes. Bellhouse (1981) s'intéresse à des données spatiales et compare différents plans de sondages permettant aussi d'échantillonner des lignes et des colonnes sur une grille rectangulaire. Le plan produit ainsi que le plan a deux degrés sont notamment envisagés

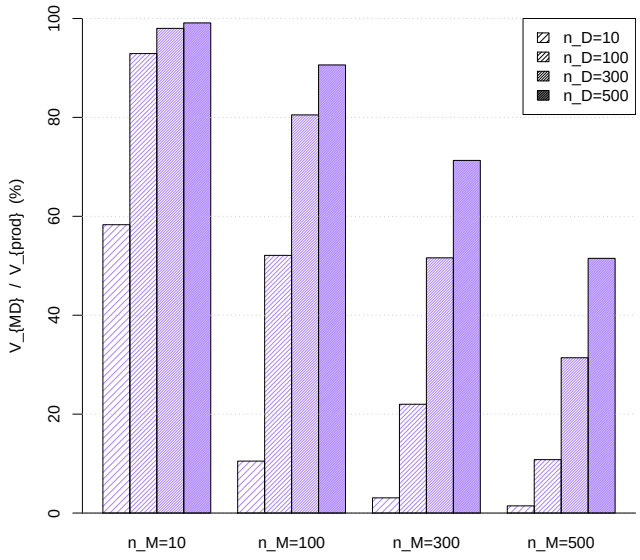

FIGURE 5 – Ratio V_{MD} /V_{prod} (%) en fonction de n_M et n_D

avec et sans stratification. Les variances anticipées sont calculées pour ces plans en particulier sous un modèle de super-population prenant en compte la tendance spatiale. Dans ce contexte, Bellhouse montre aussi que le plan produit conduit à une variance anticipée supérieure à celle du plan à deux degrés.

2 Estimation de la variance issue du plan Elfe

Dans cette partie, les trois modélisations présentées en sous-section 1.1 sont formulées et comparées. Pour chacune de ces trois options, le plan sur les maternités, p_M, est un plan STSI (stratifié avec tirage SI de maternités à l'intérieur de chaque strate). Pour le plan sur les jours, p_D, on considère trois possibilités :

- un plan STSI (stratifié avec tirage SI de jours à l'intérieur de chaque strate) ;
- un plan SI ;
- un plan STSIC (stratifié avec tirage SI de grappes d'unités jours à l'intérieur de chaque strate).

2.1 Modélisation STSI $\times$ STSI

On considère le plan de sondage pour lequel p_M est un plan aléatoire simple stratifié de taille n_{Mg} à l'intérieur de chaque strate U_{Mg} de taille N_{Mg} avec $g = 1, ..., G$ (voir le Tableau 1), et où p_D est un plan aléatoire simple stratifié de taille n_{Dh} à l'intérieur de chaque strate U_{Dh} de taille N_{Dh} avec $h = 1, ..., H$ (voir Tableau 2). Un estimateur sans biais de la variance de $\hat{t}_Y$ est donné par :

$$\hat{\mathbf{V}}_{prod}\left(\hat{t}_Y\right) = \hat{\mathbf{V}}_D\left(\hat{t}_Y\right) + \hat{\mathbf{V}}_M\left(\hat{t}_Y\right) - \hat{\mathbf{V}}_E\left(\hat{t}_Y\right) \tag{23}$$

avec

$$\hat{\mathbf{V}}_D\left(\hat{t}_Y\right) = \sum_{h=1}^{H} N_{Dh}^2 \left(\frac{1}{n_{Dh}} - \frac{1}{N_{Dh}}\right) s^2_{\hat{Y}_{\bullet\circ,h}}, \tag{24}$$

$$\hat{\mathbf{V}}_M\left(\hat{t}_Y\right) = \sum_{g=1}^{G} N_{Mg}^2 \left(\frac{1}{n_{Mg}} - \frac{1}{N_{Mg}}\right) s^2_{\hat{Y}_{\circ\bullet,g}}, \tag{25}$$

$$\hat{\mathbf{V}}_E\left(\hat{t}_Y\right) = \sum_{g=1}^{G} N_{Mg}^2 \left(\frac{1}{n_{Mg}} - \frac{1}{N_{Mg}}\right) \sum_{h=1}^{H} N_{Dh}^2 \left(\frac{1}{n_{Dh}} - \frac{1}{N_{Dh}}\right) \frac{1}{(n_{Mg}-1)(n_{Dh}-1)} s_{E,hg}, \tag{26}$$

où

$$s^2_{\hat{Y}_{\bullet\circ,h}} = \frac{1}{n_{Dh}-1} \sum_{k\in S_{Dh}} \left(\hat{Y}_{\bullet k} - \frac{1}{n_{Dh}} \sum_{l\in S_{Dh}} \hat{Y}_{\bullet l}\right)^2, \tag{27}$$

$$s^2_{\hat{Y}_{\circ\bullet,g}} = \frac{1}{n_{Mg}-1} \sum_{i\in S_{Mg}} \left(\hat{Y}_{i\bullet} - \frac{1}{n_{Mg}} \sum_{j\in S_{Mg}} \hat{Y}_{j\bullet}\right)^2, \tag{28}$$

$$s_{E,hg} = \sum_{i\in S_{Mg}} \sum_{k\in S_{Dh}} \left[Y_{ik} - \frac{1}{n_{Mg}} \sum_{j\in S_{Mg}} Y_{jk} - \frac{1}{n_{Dh}} \sum_{l\in S_{Dh}} Y_{il} + \frac{1}{n_{Mg}} \frac{1}{n_{Dh}} \sum_{j\in S_{Mg}} \sum_{l\in S_{Dh}} Y_{jl}\right]^2 \tag{29}$$

L'estimateur de variance se décompose en trois termes : $\hat{\mathbf{V}}_D\left(\hat{t}_Y\right)$ qui représente un effet inter-jours, $\hat{\mathbf{V}}_M\left(\hat{t}_Y\right)$ qui représente un effet inter-maternités, $\hat{\mathbf{V}}_E\left(\hat{t}_Y\right)$ qui représente un effet résiduel.

Strate g	Taille de la strate N_{Mg}	Taille de l'échantillon n_{Mg}
1	108	28
2	108	47
3	109	66
4	108	97
5	111	111

TABLEAU 1 – Tailles des strates et des échantillons dans chaque strate pour le plan de sondage p_M

2.2 Modélisation STSI × SI

Nous considérons ici une modélisation du plan p_D sous la forme d'un tirage SI. L'estimateur de variance issu du plan STSI × SI correspondant se retrouve en utilisant H=1 dans

Strate h	Taille de la strate N_{Dh}	Taille de l'échantillon n_{Dh}
1	91	4
2	91	6
3	91	7
4	91	8

TABLEAU 2 – Tailles des strates et des échantillons dans chaque strate pour une modélisation du plan de sondage p_D sous la forme d'un plan STSI

la formule (23). On obtient :

$$\hat{\mathbf{V}}_D\left(\hat{t}_Y\right) = N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)s^2_{\hat{Y}_{\bullet\circ}} \tag{30}$$

$$\hat{\mathbf{V}}_M\left(\hat{t}_Y\right) = \sum_{g=1}^{G} N_{Mg}^2\left(\frac{1}{n_{Mg}}-\frac{1}{N_{Mg}}\right)s^2_{\hat{Y}_{\circ\bullet,g}} \tag{31}$$

$$\hat{\mathbf{V}}_E\left(\hat{t}_Y\right) = \sum_{g=1}^{G} N_{Mg}^2\left(\frac{1}{n_{Mg}}-\frac{1}{N_{Mg}}\right)N_D^2\left(\frac{1}{n_D}-\frac{1}{N_D}\right)\frac{1}{(n_{Mg}-1)(n_D-1)}s_{E,g}. \tag{32}$$

Pour l'enquête Elfe, selon cette modélisation, on dispose d'un échantillon de 25 jours parmi les 365 de la population.

2.3 Modélisation STSI × STSIC

Les jours choisis au sein de chaque saison étant consécutifs, on peut imaginer un tirage par grappes de jours. La modélisation proposée repose sur deux strates (semestres) à l'intérieur desquelles deux grappes de jours sont tirées à probabilités égales (SIC) [7]. L'estimateur de variance issu d'un plan STSI × STSIC se retrouve en utilisant la formule (23) avec H=2 et en remplaçant les Y_{ik} par :

$$Y_{ik} = \text{total pour la maternité } i\text{, la grappe } k \text{ de jours} \tag{33}$$

$$= \sum_{d=1}^{D_k} Y_{ik_d} \quad \text{avec } D_k = \text{nombre de jours de la grappe } k$$

$$\text{où } Y_{ik_d} = \text{total pour la maternité } i\text{, le jour } d \text{ dans la grappe } k. \tag{34}$$

En effet, ici les populations des strates sont des populations de grappes de jours, alors que pour les deux modélisations précédentes, il s'agissait de populations de jours. Les tailles d'échantillon et les tailles de strates correspondant à cette modélisation du plan p_D sont données en Tableau 3.

2.4 Comparaisons entre les différentes modélisations

Les trois modélisations du plan p_D sont illustrées dans le Tableau 4 à partir de trois variables issues de la base de données Elfe. Il est important de noter que les résultats

7. On considère que dans la première strate, il y a 4 jours choisis au premier trimestre, 6 jours au second, donc en moyenne 5 jours. Il y a donc environ $182/5 \approx 36$ grappes de 5 jours dans ce semestre, et on en a tiré 2 par plan SI : $\pi_k^D = 1/18$. On considère que dans la seconde strate il y a 7 jours choisis au troisième semestre, 8 jours au quatrième, donc 7,5 jours en moyenne, arrondi à 7 jours. Il y a donc environ $182/7 \approx 26$ grappes de 7 jours dans ce semestre, et on en a tiré 2 par plan SI : $\pi_k^D = 1/13$.

Strate h	Taille de la strate N_{Dh}	Taille de l'échantillon n_{Dh}
1	36	2
2	26	2

TABLEAU 3 – Tailles des strates et des échantillons dans chaque strate pour une modélisation du plan de sondage p_D sous la forme d'un plan STSIC

affichés ne prennent en compte ni la phase de non-réponse (abordée dans la section suivante) ni l'étape de calage. Ils ne sont présentés ici qu'à titre d'illustration pour comparer les différentes modélisations.

La première variable *Nombre de naissances* est une variable présentant à la fois une variabilité entre les jours (moins de naissances les samedis et dimanches) et une variabilité entre les maternités. Pour la seconde variable *Nombre de naissances sous césarienne*, la variabilité due au jour de naissance est très grande (césariennes programmées en semaine), alors que la dernière variable *Nombre de mères suivies par une sage-femme* présente une variabilité importante entre les différentes maternités.

Pour chacune de ces variables, on calcule le total estimé donné en (5) pour le plan STSI $\times$ STSI et pour le plan STSI $\times$ SI. Dans le cas du plan STSI $\times$ STSIC, la variance est très forte en raison notamment de la variation des tailles de grappe ; on calcule donc dans ce cas l'estimateur par le ratio, en redressant sur le nombre total de jours.

On calcule les estimateurs de variance ainsi que leurs composantes $\hat{\mathbf{V}}_D$, $\hat{\mathbf{V}}_M$ et $\hat{\mathbf{V}}_E$ associés à chacune des trois modélisations données respectivement en sections 2.1, 2.2 et 2.3. On calcule également le coefficient de variation estimé :

$$\hat{\mathrm{CV}}\left(\hat{t}_Y\right) = \frac{\sqrt{\hat{\mathbf{V}}_X\left(\hat{t}_Y\right)}}{\hat{t}_Y}.$$

Dans le Tableau 4, les trois modélisations donnent des estimations comparables pour les totaux respectifs des trois variables. La modélisation STSI $\times$ STSIC conduit à une légère surestimation. Concernant la variance, les estimations sont aussi comparables. Notons que le coefficient de variation estimé est assez élevé pour la variable *Nombre de naissances sous césarienne*, mais le calcul ne prend pas en compte l'étape de calage.

Dans la suite de ce document nous retiendrons la modélisation STSI $\times$ STSI pour l'application à l'enquête Elfe, modélisation permettant de considérer des effets saisonniers. Cependant la modélisation STSI $\times$ STSIC prenant en compte l'effet grappe de jours n'est pas à écarter.

3 Estimation de la variance issue du plan Elfe avec prise en compte de la non-réponse

Le traitement de la non-réponse de l'enquête Elfe est présenté. L'estimateur de variance prenant en compte l'échantillonnage produit mais aussi la non-réponse est calculé et illustré par simulations.

Modélisations $p_M \times p_D$	Variables	Nombre de naissances	Nombre de naissances sous césarienne	Nombre de mères suivies par sage femme
STSI × STSI	$\hat{t}_Y$	368477	34431	42645
	$\hat{\mathbf{V}}_{prod}\left(\hat{t}_Y\right)$	6.7 10^7	1.4 10^7	3.2 10^6
	$\hat{\mathbf{V}}_M\left(\hat{t}_Y\right)$	1.8 10^7	4.9 10^5	2.7 10^6
	$\hat{\mathbf{V}}_D\left(\hat{t}_Y\right)$	5.2 10^7	1.4 10^7	8.9 10^5
	$\hat{\mathbf{V}}_E\left(\hat{t}_Y\right)$	3.7 10^6	3.5 10^5	3.7 10^5
	$\hat{\mathrm{CV}}\left(\hat{t}_Y\right)$	02.2 %	11.0 %	04.2 %
STSI × SI	$\hat{t}_Y$	370696	35435	42604
	$\hat{\mathbf{V}}_{prod}\left(\hat{t}_Y\right)$	6.5 10^7	1.3 10^7	3.1 10^6
	$\hat{\mathbf{V}}_M\left(\hat{t}_Y\right)$	1.9 10^7	5.3 10^5	2.6 10^6
	$\hat{\mathbf{V}}_D\left(\hat{t}_Y\right)$	4.9 10^7	1.3 10^7	8.7 10^4
	$\hat{\mathbf{V}}_E\left(\hat{t}_Y\right)$	3.5 10^6	3.7 10^5	3.5 10^5
	$\hat{\mathrm{CV}}\left(\hat{t}_Y\right)$	02.2 %	10.2 %	04.2 %
STSI × STSIC	$\hat{t}_Y$	381406	36231	43862
	$\hat{\mathbf{V}}_{prod}\left(\hat{t}_Y\right)$	6.9 10^7	8.6 10^6	3.5 10^6
	$\hat{\mathbf{V}}_M\left(\hat{t}_Y\right)$	1.7 10^7	5.4 10^5	2.8 10^6
	$\hat{\mathbf{V}}_D\left(\hat{t}_Y\right)$	5.6 10^7	8.4 10^6	1.1 10^6
	$\hat{\mathbf{V}}_E\left(\hat{t}_Y\right)$	5.2 10^6	4.0 10^5	4.3 10^5
	$\hat{\mathrm{CV}}\left(\hat{t}_Y\right)$	02.2 %	08.1 %	04.2 %

TABLEAU 4 – Comparaison entre les trois modélisations du plan $p_M \times p_D$

3.1 Phase de non-réponse

Durant l'enquête Elfe, 29 maternités parmi les 349 sélectionnées n'ont pas participé à l'enquête. Cette première étape de non-réponse a été traitée par la méthode des Groupes de Réponses Homogènes (G.R.H.). Ensuite, parmi ces 320 maternités, certaines n'ont pas participé à toutes les vagues d'enquête : 305 maternités en semestre 1, 312 en semestre 2, 311 en semestre 3 et 309 en semestre 4. Cette non-réponse a été traitée dans chaque strate de maternités en ajustant les probabilités d'inclusion par un quotient représentant le nombre de maternités participant au semestre sur le nombre de maternités attendues. Avec des taux de non-réponse relativement faibles pour les maternités (7 %) et pour les jours (3 % en moyenne), ces deux premières phases de non-réponse ne sont pas prises en compte dans le calcul de la variance de non-réponse mais traitées en ajustant simplement les probabilités d'inclusion.

Ensuite, il y a une phase de non-réponse au niveau nourrisson : 49 % des 36 000 familles approchées n'ont pas souhaité participer. La méthode des G.R.H. a de nouveau été utilisée pour traiter cette phase, puis, pour finir, un calage a été réalisé sur des variables socio-démographiques. Cette dernière phase de non-réponse est considérée dans le calcul de l'estimateur de variance qui suit mais l'étape de calage ne l'est pas.

Notre variable d'intérêt prend la valeur y_a pour le nourrisson a. Le total t_Y peut alors s'écrire

$$t_Y = \sum_{i \in U_M} \sum_{k \in U_D} Y_{ik} \quad \text{avec} \quad Y_{ik} = \sum_{a \in U_{ik}} y_a, \tag{35}$$

où U_{ik} représente la sous-population des nourrissons de la maternité i le jour k. On note $S_{R_{ik}}$ l'échantillon des répondants de la sous-population U_{ik}. La non-réponse est modélisée

par une seconde phase de tirage au sein de l'échantillon complet des nourrissons. Pour cela, on fait l'hypothèse qu'il existe des groupes homogènes de réponse, avec comportements de réponse indépendants dans ces G.R.H.. En se basant sur la méthode des scores (Eltinge et Yansaneh, 1997) afin d'estimer les probabilités de réponse, F groupes de réponses homogènes sont créés. On notera $\hat{p}_f$ la probabilité de réponse estimée pour le G.R.H. f, et S_{R_f} l'échantillon des n_{R_f} répondants du G.R.H. f. On a donc $\hat{p}_a = \hat{p}_f$ pour tout $a \in S_{R_f}$.
Dans ce cas, le total t_Y est estimé approximativement sans biais par

$$\begin{aligned}
\hat{t}_{Y\star} &= \sum_{i\in S_M}\sum_{k\in S_D}\frac{\hat{Y}_{ik}}{\pi_i^M\pi_k^D} \quad \text{avec} \quad \hat{Y}_{ik} = \sum_{a\in S_{R_{ik}}}\frac{y_a}{\hat{p}_a}, \\
&= \sum_{i\in S_M}\frac{\hat{\hat{Y}}_{i\bullet}}{\pi_i^M} \quad \text{avec} \quad \hat{\hat{Y}}_{i\bullet} = \sum_{k\in S_D}\frac{\hat{Y}_{ik}}{\pi_k^D}, \\
&= \sum_{k\in S_D}\frac{\hat{\hat{Y}}_{\bullet k}}{\pi_k^D} \quad \text{avec} \quad \hat{\hat{Y}}_{\bullet k} = \sum_{i\in S_M}\frac{\hat{Y}_{ik}}{\pi_i^M}.
\end{aligned} \tag{36}$$

3.2 Variance avec phase de non-réponse

Pour un plan produit et la phase de non-réponse présentée dans le paragraphe précédent, lorsqu'on utilise les estimations des probabilités de réponse issues de la méthode des scores, un estimateur approximativement sans biais de la variance peut être obtenu en adaptant le travail de Kim et Kim (2007). Cela conduit à :

$$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right) = \hat{\mathbf{V}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_{\text{NR}}\left(\hat{t}_{Y\star}\right) \tag{37}$$

où

$$\hat{\mathbf{V}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{i,j\in S_M}\sum_{k,l\in S_D}\sum_{a\in S_{R_{ik}}}\sum_{b\in S_{R_{jl}}}\frac{\Gamma_{ijkl}}{\pi_{ij}^M\pi_{kl}^D}\ \frac{1}{\hat{p}_{ab}}\ \frac{y_a}{\pi_i^M\pi_k^D}\frac{y_b}{\pi_j^M\pi_l^D}, \tag{38}$$

$$\hat{\mathbf{V}}_{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{f=1}^{F}\sum_{a\in S_{R_f}}\frac{1-\hat{p}_f}{\hat{p}_f^2}\left(\check{y}_a - \frac{1}{n_{R_f}}\sum_{b\in S_{R_f}}\check{y}_b\right)^2, \tag{39}$$

avec

$$\hat{p}_{ab} = \begin{cases} \hat{p}_a\hat{p}_b & \text{si} \quad a \neq b \\ \hat{p}_a & \text{sinon} \end{cases}, \tag{40}$$

$$\text{et} \quad \check{y}_a = \frac{y_a}{\pi_i^M\pi_k^D}. \tag{41}$$

La partie $\hat{\mathbf{V}}_{\text{NR}}\left(\hat{t}_{Y\star}\right)$ correspond à l'estimateur de la variance due à la non-réponse et $\hat{\mathbf{V}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right)$ correspond à l'estimateur de la variance due à l'échantillonnage. Ce dernier peut se décomposer sous la forme

$$\hat{\mathbf{V}}_{\text{ech}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \hat{\mathbf{V}}_{\text{ech1}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) - \hat{\mathbf{V}}_{\text{ech2}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) \tag{42}$$

avec

$$\hat{\mathbf{V}}_{\text{ech1}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{i,j\in S_M}\sum_{k,l\in S_D}\frac{\Gamma_{ijkl}}{\pi_{ij}^{M}\pi_{kl}^{D}}\frac{\hat{Y}_{ik}}{\pi_{i}^{M}\pi_{k}^{D}}\frac{\hat{Y}_{jl}}{\pi_{j}^{M}\pi_{l}^{D}}, \tag{43}$$

$$\hat{\mathbf{V}}_{\text{ech2}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{i\in S_M}\sum_{k\in S_D}\sum_{a\in S_{R_{ik}}}\left(1-\pi_{i}^{M}\pi_{k}^{D}\right)\left(\frac{y_a}{\pi_{i}^{M}\pi_{k}^{D}}\right)^2 \frac{(1-\hat{p}_a)}{\hat{p}_a^2}. \tag{44}$$

3.3 Estimateur de variance dans le cas Elfe

En appliquant la formule (37) au cas particulier de l'enquête Elfe avec la modélisation STSI × STSI, on obtient :

$$\hat{\mathbf{V}}_{\text{ech1}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \hat{\mathbf{V}}_{D}^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_{M}^{\text{NR}}\left(\hat{t}_{Y\star}\right) - \hat{\mathbf{V}}_{E}^{\text{NR}}\left(\hat{t}_{Y\star}\right) \tag{45}$$

$$\hat{\mathbf{V}}_{\text{ech2}}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{g=1}^{G}\sum_{h=1}^{H}\sum_{f=1}^{F}\sum_{a\in S_{R_f}}\frac{N_{Mg}^2 N_{Dh}^2}{n_{Mg}n_{Dh}}\left(\frac{1}{n_{Mg}n_{Dh}}-\frac{1}{N_{Mg}N_{Dh}}\right)\frac{1-\hat{p}_f}{\hat{p}_f^2}{y_a}^2 \tag{46}$$

avec

$$\hat{\mathbf{V}}_{D}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{h=1}^{H}N_{Dh}^2\left(\frac{1}{n_{Dh}}-\frac{1}{N_{Dh}}\right)s^2_{\hat{\hat{Y}}_{\bullet\circ,h}} \tag{47}$$

$$\hat{\mathbf{V}}_{M}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{g=1}^{G}N_{Mg}^2\left(\frac{1}{n_{Mg}}-\frac{1}{N_{Mg}}\right)s^2_{\hat{\hat{Y}}_{\circ\bullet,g}} \tag{48}$$

$$\hat{\mathbf{V}}_{E}^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{g=1}^{G}N_{Mg}^2\left(\frac{1}{n_{Mg}}-\frac{1}{N_{Mg}}\right)\sum_{h=1}^{H}N_{Dh}^2\left(\frac{1}{n_{Dh}}-\frac{1}{N_{Dh}}\right)\frac{1}{(n_{Mg}-1)(n_{Dh}-1)}s_{\hat{E},hg} \tag{49}$$

et

$$s^2_{\hat{\hat{Y}}_{\bullet\circ,h}} = \frac{1}{n_{Dh}-1}\sum_{k\in S_{Dh}}\left(\hat{\hat{Y}}_{\bullet k}-\frac{1}{n_{Dh}}\sum_{l\in S_{Dh}}\hat{\hat{Y}}_{\bullet l}\right)^2, \tag{50}$$

$$s^2_{\hat{\hat{Y}}_{\circ\bullet,g}} = \frac{1}{n_{Mg}-1}\sum_{i\in S_{Mg}}\left(\hat{\hat{Y}}_{i\bullet}-\frac{1}{n_{Mg}}\sum_{j\in S_{Mg}}\hat{\hat{Y}}_{j\bullet}\right)^2, \tag{51}$$

$$s_{\hat{E},hg} = \sum_{i\in S_{Mg}}\sum_{k\in S_{Dh}}\left[\hat{Y}_{ik}-\frac{1}{n_{Mg}}\sum_{j\in S_{Mg}}\hat{Y}_{jk}-\frac{1}{n_{Dh}}\sum_{l\in S_{Dh}}\hat{Y}_{il}+\frac{1}{n_{Mg}}\frac{1}{n_{Dh}}\sum_{j\in S_{Mg}}\sum_{l\in S_{Dh}}\hat{Y}_{jl}\right]^2 \tag{52}$$

On retrouve dans $\hat{\mathbf{V}}_{\text{ech1}}^{\text{NR}}\left(\hat{t}_{Y\star}\right)$ les trois termes qui composaient la variance présentée en formule (23) (à la différence que les sous-totaux Y_{ik} sont ici estimés, prenant en compte l'ajustement de la non-réponse), auxquels on soustrait le terme $\hat{\mathbf{V}}_{\text{ech2}}^{\text{NR}}\left(\hat{t}_{Y\star}\right)$ afin d'obtenir un estimateur sans biais de la variance d'échantillonnage.

3.4 Simulations avec phase de non-réponse

On utilise $N_M = 500$ et $N_D = 500$, et pour chaque grappe (i, k), 30 valeurs associées aux nourrissons ont été générées selon le modèle :

$$y_{a_{ik}} = \nu + \sigma_1 u_i + \sigma_2 v_k + \sigma_3 w_{ik} + \sigma_4 \epsilon_{a_{ik}} \tag{53}$$

où ν est égal à 1000, $\sigma_1, \sigma_2, \sigma_3$ et σ_4 sont fixés à 5 et les u_i, v_k, w_{ik} et $\epsilon_{a_{ik}}$ sont générés indépendamment à partir d'une loi normale centrée réduite.

L'échantillonnage produit SI $\times$ SI est utilisé, en faisant varier les tailles d'échantillon ($n_M = n_D = 20$ et $n_M = n_D = 100$). Ensuite, un tirage de Bernoulli de paramètre p permet de sélectionner l'échantillon des répondants (un seul G.R.H.). Dans le Tableau 5, ce paramètre prend successivement deux valeurs : 0.9 et 0.7.

Chaque échantillonnage (tirage SI $\times$ SI, suivi du processus de non-réponse) est répété $B = 10000$ fois. On calcule le Biais Relatif Monte Carlo en pourcentage (BR), donné par

$$\mathrm{BR}_{\mathrm{MC}}(\hat{\mathbf{V}}) = 100 \times \frac{B^{-1} \sum_{b=1}^{B} \hat{\mathbf{V}}^{(b)} - \mathbf{V}}{\mathbf{V}}. \tag{54}$$

La vraie variance $\mathbf{V}$ est approximée par la variance Monte Carlo à partir d'un jeu indépendant de 50000 simulations. La variance due à la non-réponse $\mathbf{V}_{\mathrm{NR}}$ est obtenue à partir d'un autre jeu indépendant de 100×100 simulations, où on tire 100 échantillons selon un plan SI $\times$ SI, et pour chacun le mécanisme de réponse est répété 100 fois.

n_M	20	20	100	100
n_D	20	20	100	100
$\hat{p}$	0.9	0.7	0.9	0.7
$\mathbf{V}$	1.31E+14	1.33E+14	2.12E+13	2.11E+13
$\hat{\mathbf{V}}$	1.31E+14	1.46E+14	2.14E+13	2.13E+13
$\%\mathrm{BR}_{\mathrm{MC}}(\hat{\mathbf{V}})$	-0.08	9.16	0.88	0.71
$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{Y\star}\right)$	1.31E+14	1.45E+14	2.14E+13	2.12E+13
$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech1}}\left(\hat{t}_{Y\star}\right)$	6.51E+14	2.15E+15	4.14E+13	9.84E+13
$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech2}}\left(\hat{t}_{Y\star}\right)$	5.20E+14	2.01E+15	2.00E+13	7.72E+13
$\mathbf{V}_{\mathrm{NR}}$	4.98E+10	1.86E+11	2.05E+09	7.76E+09
$\hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{Y\star}\right)$	4.94E+10	1.90E+11	2.02E+09	7.79E+09
$\%\mathrm{BR}_{\mathrm{MC}}(\hat{\mathbf{V}}_{\mathrm{NR}})$	-0.94	2.11	-1.53	0.31

TABLEAU 5 – Simulations pour un plan de sondage SI $\times$ SI avec une phase de non-réponse

On constate que la variance due à la non-réponse est faible devant la variance d'échantillonnage. On remarque logiquement que la part de la variance due à la non-réponse $\hat{\mathbf{V}}_{\mathrm{NR}}$ augmente lorsque le taux de réponse diminue. On constate bien que les estimateurs $\hat{\mathbf{V}}_{\mathrm{NR}}$ et $\hat{\mathbf{V}}$ sont approximativement sans biais pour $\mathbf{V}_{\mathrm{NR}}$ et $\mathbf{V}$, respectivement.

4 A la recherche d'estimateurs simplifiés

Précédemment, un estimateur de la variance issu du plan de sondage Elfe a été présenté avec prise en compte de la non-réponse. Dans cette section, plusieurs estimateurs simplifiés sont proposés, pour différentes raisons :

– l'estimateur sans biais n'est programmé dans aucun logiciel à notre connaissance ;
– l'estimateur sans biais peut prendre des valeurs négatives, d'où la recherche d'estimateurs simplifiés, potentiellement biaisés mais positifs.

4.1 Estimateurs simplifiés

En prenant en compte les procédures logicielles existantes dans R, SAS et Stata, cinq estimateurs simplifiés ont été retenus :

- le premier estimateur correspond à une partie de l'estimateur sans biais, représentant la variance estimée inter-maternités en formule (48),

$$\hat{\mathbf{V}}_{\text{SIMP1}} \equiv \hat{\mathbf{V}}_M^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{g=1}^{G}(N_{Mg})^2\left(\frac{1}{n_{Mg}} - \frac{1}{N_{Mg}}\right) s^2_{\hat{Y}_{\circ\bullet,g}}, \tag{55}$$

- le deuxième estimateur correspond à une partie de l'estimateur sans biais, représentant la variance estimée inter-jours en formule (47),

$$\hat{\mathbf{V}}_{\text{SIMP2}} \equiv \hat{\mathbf{V}}_D^{\text{NR}}\left(\hat{t}_{Y\star}\right) = \sum_{h=1}^{H}(N_{Dh})^2\left(\frac{1}{n_{Dh}} - \frac{1}{N_{Dh}}\right) s^2_{\hat{Y}_{\bullet\circ,h}}, \tag{56}$$

- le troisième estimateur correspond à la somme des deux précédents estimateurs simplifiés,

$$\begin{aligned}\hat{\mathbf{V}}_{\text{SIMP3}} &\equiv \hat{\mathbf{V}}_{\text{SIMP1}} + \hat{\mathbf{V}}_{\text{SIMP2}} \\ &= \hat{\mathbf{V}}_D^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_M^{\text{NR}}\left(\hat{t}_{Y\star}\right),\end{aligned} \tag{57}$$

- le quatrième estimateur correspond à l'estimateur de variance adapté à un plan classique à deux degrés, dans lequel les maternités constituent les Unités Primaires (UP) et les jours les Unités Secondaires (US),

$$\hat{\mathbf{V}}_{\text{SIMP4}} \equiv \hat{\mathbf{V}}_M^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \sum_{g=1}^{G}\frac{N_{Mg}}{n_{Mg}}\sum_{i\in S_{Mg}}\sum_{h=1}^{H}N_{Dh}^2\left(\frac{1}{n_{Dh}} - \frac{1}{N_{Dh}}\right) s^2_{\hat{Y}_{i\circ,h}} \tag{58}$$

$$\text{avec} \quad s^2_{Y_{i\circ,h}} = \frac{1}{n_{Dh}-1}\sum_{k\in S_{Dh}}(\hat{Y}_{ik} - \frac{1}{n_{Dh}}\sum_{l\in S_{Dh}}\hat{Y}_{il})^2, \tag{59}$$

- le cinquième estimateur correspond à l'estimateur de variance adapté à un plan classique à deux degrés, dans lequel les jours constituent les UP et les maternités les US,

$$\hat{\mathbf{V}}_{\text{SIMP5}} \equiv \hat{\mathbf{V}}_D^{\text{NR}}\left(\hat{t}_{Y\star}\right) + \sum_{h=1}^{H}\frac{N_{Dh}}{n_{Dh}}\sum_{k\in S_{Dh}}\sum_{g=1}^{G}N_{Mg}^2\left(\frac{1}{n_{Mg}} - \frac{1}{N_{Mg}}\right) s^2_{\hat{Y}_{\circ k,g}} \tag{60}$$

$$\text{avec} \quad s^2_{Y_{\circ k,g}} = \frac{1}{n_{Mg}-1}\sum_{i\in S_{Mg}}(\hat{Y}_{ik} - \frac{1}{n_{Mg}}\sum_{j\in S_{Mg}}\hat{Y}_{jk})^2. \tag{61}$$

Les estimateurs $\hat{\mathbf{V}}_{\text{SIMP1}}$ et $\hat{\mathbf{V}}_{\text{SIMP2}}$ peuvent se calculer à partir de la procédure *survey-means* du logiciel SAS (avec les options *cluster*, *strata* et *weight*) ou à partir des fonctions

svydesign et *svytotal* en utilisant le logiciel R (librairie *survey* et les paramètres *id* pour identifier l'UP, *strata*, *fpc* et *weights*) ou encore la procédure *svy* du logiciel Stata (en renseignant l'UP et les paramètres *strata*, *fpc* et *pweight*). L'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ est calculable en utilisant les procédures associées à $\hat{\mathbf{V}}_{\text{SIMP1}}$ et $\hat{\mathbf{V}}_{\text{SIMP2}}$ et en sommant les deux estimateurs. Pour les estimateurs $\hat{\mathbf{V}}_{\text{SIMP4}}$ et $\hat{\mathbf{V}}_{\text{SIMP5}}$, les logiciels R et Stata proposent des options pour prendre en compte le second degré d'échantillonnage (les mêmes procédures que celles citées précédemment). Concernant le logiciel SAS, ces deux estimateurs sont programmables en calculant séparément les deux termes.

Ces estimateurs simplifiés sont positifs et calculables à partir de procédures déjà programmées mais ne sont pas sans biais. Dans un contexte sans non-réponse, on a démontré que sous des conditions standard l'estimateur $\hat{\mathbf{V}}_{\text{SIMP1}}$ présente un biais proche de 0 lorsque la taille d'échantillon n_D devient grande et n_M est bornée. A l'inverse, l'estimateur $\hat{\mathbf{V}}_{\text{SIMP2}}$ présente un biais proche de zéro lorsque n_M devient grand et n_D est borné. Enfin, pour l'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$, lorsque n_M ou n_D devient grand, le biais est proche de zéro. Les détails sont donnés dans Juillard, Chauvet et Ruiz-Gazen (2015).

4.2 Etude par simulations du comportement des estimateurs simplifiés

Une étude par simulations a été effectuée afin d'évaluer les biais introduits en utilisant les différents estimateurs simplifiés. Dans cette partie, on ne considère pas de phase de non-réponse. Le modèle de superpopulation défini en (19) est utilisé avec $\mu = 1000$, $\sigma_3 = 5$, et en faisant varier σ_1 et σ_2.

Pour chacune des populations, le plan de sondage SI $\times$ SI est utilisé avec $n_M = 320$ et $n_D = 25$ dans une population produit de $N_M = 544$ maternités et $N_D = 365$ jours (tailles correspondant à celles de l'enquête Elfe).

La sélection de l'échantillon est répétée $B = 10,000$ fois pour chaque population. Dans chacun des $b = 1, \ldots, 10,000$ échantillons, on calcule l'estimateur sans biais de variance $\hat{\mathbf{V}}^{(b)}\left(\hat{t}_Y\right)$ et les estimateurs simplifiés $\hat{\mathbf{V}}^{(b)}_{\text{SIMP1}}$, $\hat{\mathbf{V}}^{(b)}_{\text{SIMP2}}$, $\hat{\mathbf{V}}^{(b)}_{\text{SIMP3}}$, $\hat{\mathbf{V}}^{(b)}_{\text{SIMP4}}$, $\hat{\mathbf{V}}^{(b)}_{\text{SIMP5}}$. Pour chaque estimateur $\hat{\mathbf{V}}_{\text{SIMP}}$, le Biais Relatif de Monte Carlo en pourcentage (BR) donné en (54) est calculé.

σ_1	0.5	0.5	0.5	5	5	5	50	50	50
σ_2	0.5	5	50	0.5	5	50	0.5	5	50
% BR ($\hat{\mathbf{V}}_{\text{SIMP1}}$)	-87.1	-99.9	-100.0	-24.0	-96.8	-99.1	-0.8	-19.8	-97.0
% BR ($\hat{\mathbf{V}}_{\text{SIMP2}}$)	-3.5	0.8	0.6	-72.4	-3.4	-0.9	-99.6	-80.1	-1.8
% BR ($\hat{\mathbf{V}}_{\text{SIMP3}}$)	9.4	0.9	0.6	3.6	-0.2	0.0	-0.4	0.1	1.2
% BR ($\hat{\mathbf{V}}_{\text{SIMP4}}$)	-73.7	-99.5	-99.8	-20.1	-96.4	-98.8	-0.8	-19.7	-96.8
% BR ($\hat{\mathbf{V}}_{\text{SIMP5}}$)	-2.8	0.8	0.6	-72.0	-3.4	-0.9	-99.4	-79.9	-1.8

TABLEAU 6 – Biais relatifs (en %) des estimateurs simplifiés

Dans le Tableau 6, seul l'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ est faiblement biaisé (< 10 %) quelle que soit l'importance des effets maternité (σ_1) et jour (σ_2). Les estimateurs $\hat{\mathbf{V}}_{\text{SIMP1}}$ et $\hat{\mathbf{V}}_{\text{SIMP4}}$ sont acceptables seulement lorsque σ_1 est grand (50) et σ_2 faible (0.5). Les estimateurs $\hat{\mathbf{V}}_{\text{SIMP2}}$ et $\hat{\mathbf{V}}_{\text{SIMP5}}$ présentent un biais négligeable lorsque σ_2 est égal ou supérieur à σ_1.

Les formules du biais relatif associé à chaque estimateur, sous l'espérance de ce modèle, dépendent aussi des tailles n_M et n_D et sont disponibles dans Juillard, Chauvet et Ruiz-Gazen (2015).

4.3 Comparaisons entre l'estimateur sans biais et les estimateurs simplifiés sur données Elfe

Dans cette partie, les résultats associés à l'estimateur $\hat{\mathbf{V}}$ (sans biais) ainsi qu'aux cinq estimateurs simplifiés présentés dans la section précédente sont illustrés sur données Elfe. Dans le Tableau 7, pour chacune des variables Elfe choisie, on calcule le total $\hat{t}_{Y\star}$ donné en formule (36), sa variance estimée $\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)$ donnée en formule (37), ainsi que chaque partie qui la compose : $\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech1}}\left(\hat{t}_{Y\star}\right)$ en (45), $\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech2}}\left(\hat{t}_{Y\star}\right)$ en (46) et $\hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{Y\star}\right)$ en (39). Il est à noter que les résultats affichés ne tiennent pas compte d'une possible étape de calage.

On calcule l'écart relatif entre $\hat{\mathbf{V}}_{\mathrm{SIMP}}$ et l'estimateur sans biais $\hat{\mathbf{V}}$ défini par :

$$ER = \frac{\hat{\mathbf{V}}_{\mathrm{SIMP}}\left(\hat{t}_{Y\star}\right) - \hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)}{\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right)}.$$

On constate à nouveau dans le Tableau 7 que la part de variance estimée due à la non-réponse $\hat{\mathbf{V}}_{\mathrm{NR}}$ est faible comparée à celle d'échantillonnage $\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{Y\star}\right)$. On rappelle qu'il s'agit d'une non-réponse non pas sur la première phase d'échantillonnage des unités groupées (i,k), mais à la seconde phase sur l'unité nourrisson.
Mis à part $\hat{\mathbf{V}}_{\mathrm{SIMP3}}$, tous les estimateurs simplifiés présentent des valeurs inférieures à l'estimateur sans biais. Rappelons que l'estimateur $\hat{\mathbf{V}}$ a déjà lui-même subi des simplifications (non prise en compte de la non-réponse au niveau maternité, ni celle au niveau jour) et présente des valeurs certainement plus petites qu'elles ne l'auraient été sans ces simplifications. L'estimateur $\hat{\mathbf{V}}_{\mathrm{SIMP3}}$ présente des ER relativement faibles et peu variables (entre 0 et 20 %, sauf pour la variable *Nombre de nourrissons ayant une mère âgée entre 18 et 25 ans* qui atteint 29 %). Tous les autres estimateurs présentent au moins un cas avec un ER supérieur à 45 % en valeur absolue. On observe que l'estimateur $\hat{\mathbf{V}}_{\mathrm{SIMP5}}$ s'avère intéressant dans plusieurs cas (parmi les dix variables étudiées, huit présentent un ER inférieur à 20 % en valeur absolue) mais extrêmement mauvais pour des variables présentant une variabilité inter-maternités importante (-47 % pour la variable *Nombre de nourrissons ayant une mère suivie par sage-femme*). Les estimateurs $\hat{\mathbf{V}}_{\mathrm{SIMP1}}$ et $\hat{\mathbf{V}}_{\mathrm{SIMP4}}$ s'avèrent inacceptables avec jusqu'à -95 % d'erreur relative.

L'estimateur $\hat{\mathbf{V}}_{\mathrm{SIMP3}}$ reste le seul estimateur simplifié acceptable quelle que soit la variable d'intérêt et pourra être recommandé aux utilisateurs des données Elfe.

Conclusion

Plusieurs modélisations du plan de sondage de l'enquête Elfe ont été proposées. Si la modélisation STSI pour le plan de sondage sur les jours permet de prendre en compte la saisonnalité, la modélisation STSIC n'est pas à écarter car elle tient compte de la structure par grappes de jours et de l'effet semestriel.
Un estimateur approximativement sans biais de la variance prenant en compte la non-réponse a été présenté. Comme sa forme (complexe) demande une programmation spécifique,

plusieurs estimateurs simplifiés et calculables avec des procédures logicielles déjà existantes ont été proposés. A partir de simulations et de variables issues de la base de données Elfe, ces estimateurs simplifiés facilement programmables ont été comparés. Les résultats montrent que l'estimateur $\hat{\mathbf{V}}_{\text{SIMP3}}$ pourrait être recommandé aux utilisateurs pour une estimation de la variance simple, et peu biaisée.
Dans un prochain travail, les calculs prendront en compte l'étape de calage.

Bibliographie

[1] Bellhouse, D. R. (1981). Spatial sampling in the presence of a trend. Journal of Statistical Planning and Inference, 5, 365-375.
[2] Eltinge, J. L. et Yansaneh, I. S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. Survey Methodology, 23, 33-40.
[3] Juillard, H., Chauvet, G. et Ruiz-Gazen, A. (2015). Variance estimation for cross-classified sampling. Document de travail.
[4] Kim, J. K. et Kim, J. J. (2007). Nonresponse Weighting Adjustment Using Estimated Response Probability. The Canadian Journal of Statistics, 35, 501-514.
[5] Ohlsson, E. (1996). Cross-Classified Sampling. Journal of Official Statistics, 12, No.3, 241-251.
[6] R Core Team (2012). *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
[7] Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). Model Assisted Survey Sampling. Springer-Verlag.
[8] SAS Institute (2010). *SAS/STAT® 9.22 User's Guide*. Cary : SAS Institute.
[9] StataCorp. 2013. *Stata : Release 13*. Statistical Software. College Station, TX : StataCorp LP.
[10] Vos, J. W. E. (1964). Sampling in space and time. Review of the International Statistical Institute, 32, No.3, 226-241.

Modélisation STSI × STSI, NR	Nombre de naissances	Nombre de naissances sous césarienne	Nombre de nourrissons ayant une mère suivie par sage-femme	Nombre de nourrissons ayant une mère primipare	Nombre de nourrissons ayant une mère mariée ou remariée
$\hat{t}_{Y\star}$	753342	73644	97775	330804	332504
$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right) = \hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{Y\star}\right)$	2.9 10^8	7.1 10^7	1.9 10^7	6.2 10^7	8.9 10^7
$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{Y\star}\right)$	2.8 10^8	6.7 10^7	1.3 10^7	5.2 10^7	7.8 10^7
$\hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{Y\star}\right)$	8.1 10^6	4.2 10^6	6.1 10^6	1.0 10^7	1.1 10^7
$\hat{\mathrm{CV}}\left(\hat{t}_{Y\star}\right)$	02.3 %	11.5 %	04.5 %	02.4 %	02.8 %
Logiciels					
$\hat{\mathbf{V}}_{\mathrm{SIMP1}}\left(\hat{t}_{Y\star}\right)$ (ER) SAS ou R ou Stata	8.9 10^7 (-69 %)	3.7 10^6 (-95 %)	1.4 10^7 (-29 %)	2.3 10^6 (-63 %)	2.4 10^7 (-73 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP2}}\left(\hat{t}_{Y\star}\right)$ (ER) SAS ou R ou Stata	2.4 10^8 (-16 %)	7.0 10^7 (-02 %)	8.7 10^6 (-55 %)	5.0 10^7 (-20 %)	7.7 10^7 (-14 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP3}}\left(\hat{t}_{Y\star}\right)$ (ER) SAS ou R ou Stata	3.3 10^8 (14 %)	7.3 10^7 (3.1 %)	2.2 10^7 (15 %)	7.3 10^7 (17 %)	1.0 10^8 (13 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP4}}\left(\hat{t}_{Y\star}\right)$ (ER) R ou Stata	1.2 10^8 (-58 %)	8.3 10^6 (-88 %)	1.9 10^7 (-01 %)	3.8 10^7 (-40 %)	4.0 10^7 (-55 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP5}}\left(\hat{t}_{Y\star}\right)$ (ER) R ou Stata	2.5 10^8 (-13 %)	7.1 10^7 (-0.3 %)	1.0 10^7 (-47 %)	5.4 10^7 (-13 %)	8.1 10^7 (-10 %)

	Nombre de nourrissons ayant une mère âgée entre 18 à 25 ans	Nombre de nourrissons ayant une mère avec un IMC supérieur à 30	Nombre de nourrissons ayant une mère ayant suivi des séances de préparation	Nombre de nourrissons ayant une mère étrangère ou apatride	Nombre de jumeaux
$\hat{t}_{Y\star}$	115987	80822	363664	92284	24519
$\hat{\mathbf{V}}\left(\hat{t}_{Y\star}\right) = \hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{Y\star}\right) + \hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{Y\star}\right)$	2.1 10^7	1.6 10^7	6.4 10^7	2.5 10^7	4.5 10^6
$\hat{\mathbf{V}}^{\mathrm{NR}}_{\mathrm{ech}}\left(\hat{t}_{Y\star}\right)$	1.4 10^7	1.1 10^7	5.6 10^7	1.8 10^7	3.4 10^6
$\hat{\mathbf{V}}_{\mathrm{NR}}\left(\hat{t}_{Y\star}\right)$	7.0 10^6	5.0 10^6	7.9 10^6	6.6 10^6	1.1 10^6
$\hat{\mathrm{CV}}\left(\hat{t}_{Y\star}\right)$	03.9 %	04.9 %	02.2 %	05.4 %	08.7 %
Logiciels					
$\hat{\mathbf{V}}_{\mathrm{SIMP1}}\left(\hat{t}_{Y\star}\right)$ (ER) SAS ou R ou Stata	1.0 10^7 (-50 %)	5.2 10^6 (-67 %)	2.8 10^7 (-57 %)	7.0 10^6 (-72 %)	1.4 10^6 (-70 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP2}}\left(\hat{t}_{Y\star}\right)$ (ER) SAS ou R ou Stata	1.6 10^7 (-21 %)	1.3 10^7 (-16 %)	4.5 10^7 (-30 %)	2.3 10^7 (-08 %)	4.0 10^6 (-12 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP3}}\left(\hat{t}_{Y\star}\right)$ (ER) SAS ou R ou Stata	2.7 10^7 (29 %)	1.9 10^7 (17 %)	7.3 10^7 (14 %)	2.9 10^7 (20 %)	5.3 10^6 (18 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP4}}\left(\hat{t}_{Y\star}\right)$ (ER) R ou Stata	1.9 10^7 (-09 %)	9.7 10^6 (-38 %)	4.0 10^7 (-37 %)	1.6 10^7 (-36 %)	3.9 10^6 (-14 %)
$\hat{\mathbf{V}}_{\mathrm{SIMP5}}\left(\hat{t}_{Y\star}\right)$ (ER) R ou Stata	1.9 10^7 (-08 %)	1.5 10^7 (-06 %)	4.9 10^7 (-24 %)	2.4 10^7 (-01 %)	4.4 10^6 (-02 %)

TABLEAU 7 – Comparaison entre différents estimateurs simplifiés et l'estimateur sans biais.